6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

EASYTALK: A LARGE-VOCABULARY SPEAKER-INDEPENDENT CHINESE DICTATION MACHINE

Fang Zheng, Zhanjiang Song, Mingxing Xu, Jian Wu, Yinfei Huang, Wenhu Wu, Cheng Bi †
Speech Laboratory, Computer Science, Tsinghua University, Beijing, 100084, China
† Zhongshan Keysun Information Technology Co. Ltd., Guangdong, 528400, China
fzheng@sp.cs.tsinghua.edu.cn

ABSTRACT

The *EasyTalk* application is a large-vocabulary speaker-independent continuous Chinese speech recognition system, i.e. Chinese dictation machine (CDM), under the WINTEL environment. Addressed in this paper are a number of novel techniques adopted in the CDM engine which is the basis of *EasyTalk*, including the merging-based syllable detection automaton (MBSDA) and the statistical knowledge based frame synchronous search (SKB-FSS) algorithms in the acoustic processing stage, the percentage in critical area (CAP) and recognition score gap (RSG) methods for the acceptation and rejection decision, the word search tree (WST), the N-Gram, and the syllable synchronous network search (SSNS) algorithm in the language processing stage, the embedded multiple model sheme (EMM) and the fuzzy syllable set (FSS) for the robustness purpose.

Keywords: Chinese Dictation Machine (CDM), Word Search Tree (WST), Syllable-Synchronous Network Search (SSNS), Embedded Multiple Model (EMM), Fuzzy Syllable Set (FSS)

1. INTRODUCTION

Chinese language is hieroglyphic; there are over 6,700 frequently used characters. So two bytes are needed to represent each Chinese character in the computer. This makes the character-input more difficult than any other alphabet-based languages. The Chinese Dictation Machine (CDM) comes for this purpose. In this paper, a CDM Engine (CDME) and the techniques adopted in it will be presented.

CDME is a speaker-independent large-vocabulary continuous Chinese speech recognition engine; the block diagram of CDME is illustrated in Figure 1. The following technical aspects will be covered in details: (1) the acoustic modeling and the search strategies; (2) the acceptation / rejection methods; (3) the language modeling as well as the word decoding algorithm; (4) the solution to the robustness problem.

Based on the CDME, a Chinese dictating system *EasyTalk* and a voice command system *EasyCmd* are established.

2. THE ACOUSTIC MODELING

The training data for CDME are taken from the 863 Database jointly established by the University of Science and Technology of China, the Acoustics Institute of the Chinese Academy of Sciences, and the Linguistics Institute of the Chinese Academy of Social Sciences. The 863 Database consists of over 22 CDs of 16-bit wide mono-channel Chinese speech samples uttered by 100 males and 100 females at 16 kHz. Each 32-msec frame was represented by 16-order LPC cepstral coefficients and the regression analysis coefficient [1].

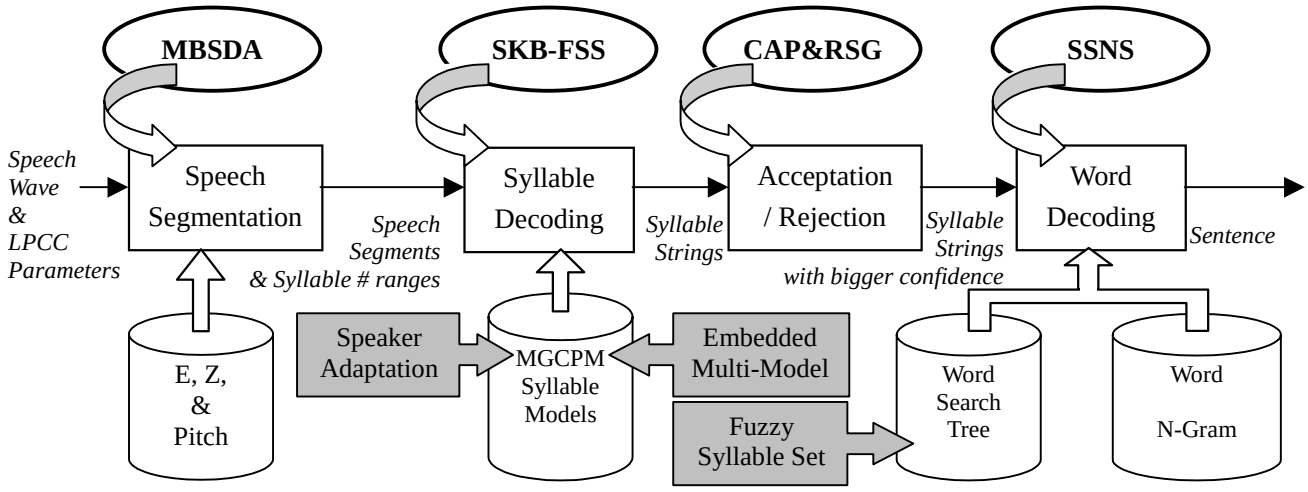

Figure 1. Functional block diagram of the Chinese Dictation Machine - *EasyTalk*

2.1 Segmental Probability Models (SPM)

Researches on HMM distance measures have showed that the probability transition matrix contributes not so much as the observation function matrix does to the performance [3][4][7][12], so a kind of SPM has been proposed based on the desertion of the HMM probability transition matrix. Two exampls are the center distance continuous probability model (CDCPM) and the mixed Gaussian continuous probability model (MGCPM) [12][13]. These continous SPMs adopt a left-to-right non-skipping topology. The intra-state feature space can be described by center distance normal (CDN) densities or mixed Gaussian densities (MGD). The state transition is controlled by the high robust Non-Linear Partion (NLP) [2] algorithm which is based on the equal feature variance sum (EFVS) criterion in the training procedure while the EFVS based search [9] or modified Viterbi algorithm [12] in the recognition procedure.

In CDME, the 6-state 16-MGD based MGCPM is adopted, which has been proved efficient. Choosing 419 toneless Chinese syllables as the speech recognition units (SRUs), MGCPM achives a satisfying top-10 isolated syllable recognition accuracy of 99.05% for 30-male training set and 95.65% for 8-male testing set across the 863 Database. The model is as good as the traditional HMM but much faster and smaller.

2.2 Search Strategies

As mentioned above, the EFVS-based search and Viterbi search algorithms can both be used for state decoding in continuous speech recognition. But the continuous Chinese speech is intermittent at boundaries of words (mostly syllables), so the intermission information can be utilized to lighten the searching loads. Two steps can be adopted: (1) try to find as many syllable separation points as possible; (2) perform efficient state decoding in each segment between the adjacent separation points.

2.2.1 MBSDA Algorithm

The merging-based syllable detection automaton (MBSDA) algorithm [10] utilizes the short-term frame energy, zero crossing rate, and pitch information to merge one or several adjacent highly similar frames into merged similar speech segments, which are regarded to belong to the same state of one syllable. A syllable detection automaton (SDA) is designed to consist of five states: silence/noise, Chinese initial, pseudo-silence (in-between the initial and final of certain Chinese syllables), Chinese final, and Chinese final tail. After processing the consecutive merged similar segments, the SDA will output a sequence of putative (syllable) separation points (PSPs), either true separation points (TSPs) or false separation points. It is possible to adjust the SDA parameters so that the outputed PSPs are TSPs. Each segment between the two adjacent segments is called a definite segment. The possible syllable number ranges in each definite segment are also given for future use in the searching procedure.

2.2.2 SKB-FSS Algorithm

A definite segment consists of one or several syllables, but the number of syllables will not be too big because of the intermission phenomenon. This forms the base of our quick search algorithm inside the definite segments.

When matching an unknown speech segment to any SRU's SPM, the traditional frame synchronous network search (FSS) algorithm [5] tries to find the best match. This may result in an undesired high matching score when a speech segment is being matched with a different model.

The NLP algorithm for the model training is based on the fact that the state dwells of adjacent states changes very slightly. So we proposed a statistical knowledge based frame synchronous search (SKB-FSS) algorithm using the differential state dwell distribution (DSDD) [11]. The SKB-FSS assigns a possible state dwell range $\left[d_{\min}^{(0)}, d_{\max}^{(0)}\right]$ to State 0 according to the syllable number range given by the MBSDA algorithm. For any other State *s*, the possible state dwell range is $\left[D^{(s-1)} + d_{\min}^{(s)}, D^{(s-1)} + d_{\max}^{(s)}\right]$, where $D^{(s-1)}$ is the average state dwell of all through-going states and $d_{\min}^{(s)} / d_{\max}^{(s)}$ are the DSDD information of the current state. In this algorithm, the state transition depends on not only the accumulated scores and the current observation as in the traditional algorithm but also the history of the state dwell distribution, so the scoring is made more robust especially for the speech speed. The DSDD information results in the performance improvement of 36.6% relatively.

The result of this step is a syllable string network that is made up of a catenation of parallel syllable string candidates for each definite segment.

2.3 Acceptation/Rejection Decision

An acceptation/rejection decision will be made to the results of the acoustic search stages, based on the percentage in critical area (CAP) and recognition score gap (RSG) [8], to reduce the acoustic candidate number. CAP is based on the fact that about 95% samples fall into the critical area of a normal distribution while RSG based on the statistical knowledge that the acoustic score differences of top candidates contain the information of the correct candidate. By combining CAP and RSG, we get a satisfying result of 3.46 average candidates with 99.73% total rejection accuracy.

3. THE LANGUAGE MODELING

3.1 Syllable Synchronous Network Search

Chinese is a syllabic language, a Chinese sentence is a string of Chinese words where each word consists of one or several Chinese characters, and mostly each character of a word with DEFINITE meaning is corresponding to a unique Chinese syllable in pronunciation. But there exists obstacles for Chinese speech recognition

compared to the western languages. (1) The homonym and homograph phenomena are rife. There are just about 418 toneless syllables and about 1,300 toned syllables but more than 6,700 frequently used characters. (2) The Chinese word boundaries are hard to determine in a given sentence. A multiple-character Chinese word can often be explained as a whole word or a catenation of sub-words, where the meaning may be the same or different. Thus, we modify the word decoding formula into

$$W^* = \underset{W}{argmax}\, P(A|W)P(W) = \underset{S=C(W)}{\underset{W}{argmax}}\, P(A|S)P(W),$$

and proposed a syllable-synchronous search (SSNS) algorithm [14], where $S=C(W)$ means S is the catenated syllable string of the word string W. SSNS algorithm forwards the acoustic syllable string network syllable-by-syllable along the vocabulary's word search tree and accumulates the word N-Gram scores of all competitive partial paths and finally takes the word string in the path ending at a word boundary and with the highest score as the final recognized sentence.

3.2 The Word Search Tree Structure

The word search tree (WST) is designed to reflect the relations among all the in-vocabulary words so that the redundancy for both the vocabulary storage and the acoustic searching consumption are reduced. In this tree, all nodes except the virtual root node and the leaf nodes are called syllable nodes (*SNode*), because they each contains the information of a syllable of a word. This tree is established recursively by this rule: all words whose first *n* syllables are exactly the same will share a unique *n*-th level *SNode* and the syllable stored in this *SNode* is exactly the *n*-th syllable of these words. The child node of the *n*-th level node (either the root node or any *SNode*) is one possible successive syllable of the current node to form the first (*n*+1) syllables of a word, it can be either an *SNode* (word length exceeding *n* syllables) or a Leaf Node (word length exactly *n* syllables). A *Leaf Node (LNode)* does not contain syllable information but the information of the word whose corresponding syllable string is exactly the same as the string of sequential syllables contained in the corresponding *SNodes* covered by the route travelling from the root node to its parent *SNode*. Because an *LNode* has no child node, reaching an *LNode* causes an accumulation of word N-Gram probabilities and an extended search from the *RNode* during the word decoding.

3.3 N-Gram Models

The used tri-grams are trained across about 40 Mega words' materials taken from the 1993 and 1994 *China Daily* and the *Turing*'s method [6] is used to smooth these zero N-Gram probabilities.

4. THE SOLUTION TO ROBUSTNESS

In Chinese, there are many kinds of regional accents all over China and overseas even the speakers themselves tend to speak in standard Chinese (Mandarin). Figure 2 shows two examples of accents, where *A* and *B* mean two different syllables. Case I is something like the different speaker issue, but Case II is possibly Chinese accent specific. In a certain accent for Case II, some syllables are mapped into quite different syllables (not just similar as in Case I), for example, a southern Chinese speaker may pronounce syllable 'zhi' completely into 'zi'. So we propose different approaches to the two different cases.

Figure 2. Two kinds of accent examples in Chinese

4.1 Embedded Multiple-Model Scheme

An embedded multiple-model (EMM) scheme tries to solve Case I problem in the acoustic processing layer [13]. EMM is based on the maximal (weighted or non-weighted) density instead of the mixed densities. It has been proved efficient.

4.2 Speaker Classification and Adaptation

For the first time a user starts the system, he is asked to say some predefined word sequence in order to be well classified according to the recognition results, and a suitable model will be selected. During the dictation, the accepted recognized syllable speech segment with sufficiently high confidence measure will be collected for later use of speaker acceptation based on a modified MAP method. This procedure is illustrated in Figure 3.

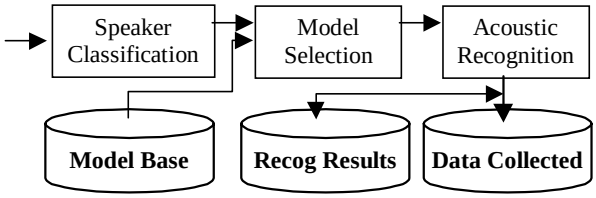

Figure 3. The data collecting in the pseudo-unsupervised speaker adaptation

4.3 Fuzzy Syllable Set

The Case II accent problem can not be solved in the acoustic layer. Our proposed solution is to apply the concept of the Chinese fuzzy syllable set to the SSNS algorithm in the language-processing layer.

Table 1. Examples of fuzzy syllable/initial/final mapping pairs

Whole Syllable	Initial	Final
ZHI → JI	Z ↔ ZH	AN ↔ ANG
CHI → QI	C ↔ CH	EN ↔ ENG
SHI → XI	S ↔ SH	IN ↔ ING
WANG ↔ HUANG	F ↔ H	
WEN ↔ WENG	N ↔ L	
GUO → GUI	W ↔ HU	
	Y ← R	

For clarification, we define *FSS*(*X*) the Fuzzy Syllable Set of a syllable *X* as a set of syllables that may be pronounced into *X* in a specific regional accent. By using the knowledge of regional accents, we can list a group of accent-syllable/initial/final to Mandarin-syllable/initial/final mapping pairs as shown in Table 1 for the speaker to check/uncheck one by one. Once the speaker checks any pairs of fuzzy syllables/initials/finials, the fuzzy syllable set generation processor will generate all the possible fuzzy syllable pairs. E.g., if "zhi→ji" is checked then *FSS*("ji") = {"zhi", "ji"}; if "z↔zh" is checked then *FSS*("zhe")={"zhe", "ze"}, *FSS*("za") = {"zha", "za"} and so on. The Case II accent problem can be solved by the arc-splitting technique in the SSNS algorithm.

5.APPLICATIONS OF THE CDM ENGINE

EasyTalk is one application using the CDME, it acts as a dictating text editor. There are 24,713 common used words, including 20.81% monosyllable words, 65.50% bi-syllable words, 7.36% thi-syllable words and 6.33% four-syllable words. Users can add new words/phrases easily and freely without any extra training. The user vocabulary size is limited only by the memory size of the personal computer. A satisfying result has been achievd, the word accuracy of *EasyTalk* is about 87.76% before adaptation and 91.8% after adaptation.

Another application is a Windows command voice navigating system named *EasyCmd*. *EasyCmd* uses most parts of the CDME except the language-processing module. Unlike the *EasyTalk*, a dynamic command set is maintained by the engine according to the active window. The dynamic command set is made up of two parts. The first part includes those dynamically retrieved from the menu, system menu and buttons of the focused window while the second part includes the fixed frequently used program links (such as 'my computer'), control keys (such as 'move left'), and some other special commands (such as 'show me the command set' and 'go to sleep'). User-defined new commands can be easily added into this part to extend the function of the system. *EasyCmd* achieves the accuracy rate of over 98% for in-vocabulary commands and true-rejection rate of over 95% for out-of-vocabulary (OOV) commands.

6. CONCLUSION

Though *EasyTalk* and *EasyCmd* are satisfying, there are still many points that should/can be improved.

For the acoustic modeling, the system can not adapt the accent well. This may be because of the training database and the modeling method. We also found that the cepstrum is not a best feature, how to find a kind of distinguishable feature remains a challenge for speech recognition.

In the language modeling, the *Turing* re-estimation method offers a good solution to the smooth of zero probabilities. But it always assigns a non-zero value to each zero N-Gram probability according to the corresponding (N-1)-Gram probabilities, which blurs the difference of different zero probabilities. Some zero-probabilities do be zero values but some are false zero just because of the lack of training materials. The semantics should play an important role in smoothing zero probabilities.

REFERENCES

[1] Furui, S. (1986), Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. on ASSP*, 34(1): 52-59, Feb. 1986.

[2] Jiang, L. (1989), The study on the methods and systems of speaker independent speech recognition based on the statistical probability models: [*Master Thesis*]. China: Tsinghua University, June 1989

[3] Juang, B.-H. & Rabiner, L. R. (1985), A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, Feb., 1985, 64(2): 391-408

[4] Lee, K.-F. (1988), Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system: [*Ph.D. dissertation*]. USA: CMU, Apr. 1988

[5] Lee, C.-H., Rabiner, L. R. (1989), A frame synchronous network search algorithm for connected word recognition. *IEEE Trans. On ASSP*, 37(11): 1649-1658, Nov. 1989

[6] Nadas, A (1985), On Turing's formula for word probabilities. *IEEE Trans. on ASSP*, 33(6), Dec. 1985

[7] Ney, H. (1993). Modeling and search in continuous speech recognition. *Euro. Conf. on Speech Tech.*, 1993, Berlin, 1: 491-498

[8] Xu, M.-X., Zheng, F., Wu, W.-H. (1997), Rejection in Speech Recognition Based on CDCPMs. *International Conference on Research on Computational Linguistics*, 412-419, Aug. 22-24, 1997, Taiwan

[9] Xu, M.-X., Zheng, F., Wu, W.-H. (1999), A fast and effective state decoding algorithm. *EuroSpeech*, 1999 (this conference)

[10] Zhang, J.-Y., Zheng, F., Du, S., Song, Z.-J., Xu, M.-X. (1999), The Merging-Based Syllable Detection Automaton in Continuous Chinese Speech Recognition. *J. of Software*, vol. 4, Apr. 1999 (in Chinese)

[11] Zheng, F., Xu, M.-X., and Wu, W.-H. (1998), The search strategies in continuous speech recognition. 5th *National Conference on Man - Machine Speech Communication* (*NCMMSC*-98), 138-143, Jul. 1998 (in Chinese)

[12] Zheng, F., Wu, W.-H., and Fang, D.-T. (1998), Center-distance continuous probability models and the distance measure. *J. of Computer Sci. & Tech.*, 13(5): 426-437, Sept. 1998

[13] Zheng, F., Mou, X.-L., Wu, W.-H., and Fang D.-T. (1998), On the embedded multiple-model scoring scheme for speech recognition. *International Symposium on Chinese Spoken Language Processing (ISCSLP'98)*, ASR-A3, pp.49-53, Dec.7-9, 1998, Singapore

[14] Zheng, F. (1999), A syllable-synchonous network search algorithm for word decoding in Chinese speech recognition. *IEEE ICASSP*-99, vol. 2, pp. 601-602, March 15-19, 1999