The First China-Japan Workshop on Spoken Language Processing (CJSLP'97), 224-228, Mar. 1997, Huangshan, P.R. China

A Real-World Large Vocabulary Speaker-Independent Speech Recognition System

Haixin Chai, Fang Zheng, Wenhu Wu, Ditang Fang
Speech Lab, Dept. of Computer Science and Technology, Tsinghua Univ.
Beijing, 100084, P.R.China
(chai@hs752.dcs.tsinghua.edu.cn, 86-10-62784141)

ABSTRACT

In this paper a real-world speaker-independent speech recognition system with 2000 Chinese phrases is introduced. Several new technologies, such as knowledge leading recognition strategy, the CDCPM acoustic model and a giant real-world speech database are used in this system, which help it gain very good performance and wide adaptability.

1 INTRODUCTION

This paper presents a real-world speaker-independent speech recognition system with 2000 Chinese phrases. The main goal of this system is to recognize phrases uttered naturally in continuous manner without any restriction. Furthermore, this system must be flexible in modifying vocabulary set.

The acoustic model of this system is based on syllable. It is natural to choose syllable as the recognition unit because Chinese is a syllable-based language. As the result, anyone who wants to modify or even replace the whole vocabulary set only needs to edit a text file, in which the Chinese character string and pronunciation of new phrases are listed. However traditional acoustic models are too complex to build a system based on syllable. We use a new model called CDCPM instead, which is somewhat a simplified case of HMM but reduces system complexity greatly, while still preserving good performance.

In continuous speech recognition the very natural thought is splitting input phonic stream into units corresponding to acoustic model, then recognizing them one by one.. Nevertheless there is not a successful system using this way till now. Because it is very hard even impossible to find a splitting method that can mark the starting and ending point of each unit exactly. In the other hand, there are many distinctive acoustic features known for segmentation. We studied this in detail, and developed a set of rules for splitting. In short, we do as much as we can, then leave unsure part to the following searching algorithm.

With limit of above splitting result, a frame-based searching algorithm is performed, which try to find a path with maximum likelihood corresponding to phonic stream. This procedure is very efficient using knowledge gained previously. For example, most boundaries between two connected syllables are determined, which can avoid many useless searchings. With no more than one-tenth time

expense to traditional frame-based searching algorithm, it can get satisfactory result.

As a real-world speech recognition system, it must be widely adaptable. We use a giant real-world speech database to accomplish this. The final system is very friendly to user.

This paper is organized as follows. In section 2, the CDCPM is described roughly. Section 3 gives our knowledge leading recognition strategy in detail. The restricted searching algorithm is also described in this section. Section 4 introduced our giant real-world speech database for training. The system performance and summary are given in section 5.

2 CDCPM

The center distance continuous probability model, which is called CDCPM, is somewhat a simplified case of HMM. It preserves only the B-matrix of HMM, and the observation output probability density function (PDF) is replaced with a one-dimensional (center-distance) probability density function (PDF). This replacement reduces time and space complexities to a great extent, while preserving good performance.

Denote the PDF of a random variable ξ With normal distribution as $\mathrm{N}(\boldsymbol{x};\mu_x,\sigma_x)$, where μ_x is its mean value and σ_x is its standard deviation. Define a new random variable $\eta = \left|\xi - \mu_x\right|$, the PDF of η is

$$p(y;\sigma_x) = \frac{2}{\sqrt{2\pi}\,\sigma_x}\exp(-y^2/2\sigma_x^2), \quad y \geq 0,$$

where the mean value μ_y of η can be calculated to be $\mu_y = \frac{2\sigma_x}{\sqrt{2\pi}}$. In fact, η is the distance between a normal variable ξ and its mean value μ_x, thus the defined distribution is a Center-Distance Normal (CDN) distribution. The CDN pseudo-PDF can be:

$$\mathrm{N}_{CD}(\boldsymbol{x};\mu_{\mathbf{x}},\mu_y) = \frac{2}{\pi\mu_y}\exp(-y^2(\boldsymbol{x},\mu_{\mathbf{x}})/\pi\mu_y^2)$$

D-dimensional case is similar to mono-dimensional case.

A mixture density CDCPM can be described by the following parameters:

(1) N: number of states per model;

(2) M: number of mixtures per state;

(3) D: number of dimensions per feature vector;

1) (4) $\vec{\mu}_{xnm} = (\mu_{xd}^{(nm)})$: mean vector of the *m*'th mixture component in *n*'th state;

2) (5) μ_{ynm} : mean center-distance of the *m*'th mixture component in *n*'th state;

3) (6) g_{nm} : mixture gain of *m*'th mixture component in *n*'th state.

4) In our system, the form of scoring using this observation PDF is given as,

5) $b_n(\vec{x}) = \max_{1\leq m\leq M} g_{nm}\,\mathrm{N}_{CD}(\vec{x};\vec{\mu}_{xnm},\mu_{ynm})$

6)

which is based on Nearest-Neighbor rule. The Bayesian learning method can be employed for a CDCPM in training. This kind of scoring strategy is useful and

proved efficient for isolated word recognition, while not easy to be used directly in continuous recognition scheme. Hence come the next recognition strategies.

3 KNOWLEDGE LEADING RECOGNITION STRATEGY

Frame-based searching algorithm is commonly used nowadays in continuous speech recognition. According to vocabulary set and structure of acoustic model, a searching accidence tree can be first developed. Then a searching procedure continues to try every possible path from root to leaf of the accidence tree, and chooses the path with maximum likelihood among them. The phrase corresponding to this path is the final result. For a large vocabulary recognition system it is impossible to try every path, so some pruning strategy must be performed. In fact, frame-based algorithm is also a Viterbi procedure by frames. Our first system is developed in this way with accuracy rate about 89.7%. However the expense in space and time is very high because many useless paths must be saved during recognition procedure. Then we tried another way.

In a sense, the result of searching algorithm can be regarded as a segmentation scheme with maximum likelihood. Finding the best segmentation method is the most natural thought in continuous speech recognition. The traditional segment procedure is, first splitting phonic wave into units corresponding to acoustic models, then scoring them respectively, choosing the best one of each unit, at last joining them together to get final result. In addition, it's well known that Chinese is a mono-syllable language system, and every syllable has the unique C-V structure, which makes segmentation easier than Latin-system language. Most boundaries between syllables can be marked by experienced researcher even only by viewing original speech wave. But it's still very hard to find a method to split phonic stream syllable by syllable exactly because of variety and complexity of speech. We studied this in detail, and developed our knowledge leading strategy.

Our main idea is to join these two strategies to get better performance. That is, first using acoustic knowledge to make marks on phonic stream as much as we can, trying to know more about input speech, then performing a searching procedure under knowledge guide to get correct result effectively. We studied the acoustic features, developed a set of certain rules. The total algorithm is described roughly as follows:

At first, CEP variances between two connected frames are calculated, then make original mark at point which CEP variance changes greatly. By carefully adjusting thresholds we are certain that all the starting and ending point of one syllable are included in these marks -- just increasing the number of them if not. Now we get primitive sections of phonic stream. The number of these sections is usually three or four times of syllable's number.

Next, let computer try to know more thing about these sections using our acoustic rules. It's well known that there is always (not all) a short silence segment between two connected syllables. The acoustic features such as energy and passing-zero rate of silence are unique to other sections. Use this knowledge we can point out most of these silent sections. Then we know that some Chinese consonants called unvoice consonants, also have distinctive phonic features. Our rules check features of every frame, and make marks only when they are sure about it.

Then, try to perform a frame-based searching procedure using knowledge gained above. A complete syllable in searching path must be connection of several primitive sections, or one section. While search begins, try to use knowledge as much as we can. For instance, if a silent section is detected, the section can not be part of a syllable. Another example, if an unvoice consonant detected, all the paths in searching must have a syllable beginning with this section. It is with such knowledge that searching becomes more efficient and saves a lot in time and space.

Here one thing must be pointed out that knowledge used in searching is not only for accelerating, but also necessary. In original frame-based searching, the pruning strategy is very important. If pruning too much the computer can seldom get right answer. On the other hand, if reserving too much the machine can not bear. Our experiments show that original frame-based searching also may get wrong answer even performing a complete search. Recognition in such way can avoid many useless trying, and help searching procedure abandon impossible path as early as possible.

4 TRAINGING DATABASE

The speech database used to train and test here is a giant Chinese database, uttered by 80 people aged 16 to 25 from all over the country. They all can speak Mandarin clearly maybe with a little accent. Speakers consist of 40 males and 40 females. The sub-vocabulary for each person includes mono-syllable word set (1100 Chinese words), bi-syllable word set (6300 words), tri-syllable word set (1100 words), quad-syllable word set (1000 words), penta-syllable word set (1 group by 76 words), hexa-syllable word set (1 group by 23 words) and hepta-syllable word set (1 group by 10 words). Five sub-vocabularies make up a complete vocabulary. In summary, the database uttered by 80 people is a 16 times' repetition of the vocabulary set. Finally, each speaker utters ten sentences.

Words or sentences are required to be uttered in Chinese Mandarin with a little local accent in the environment with some background noise, so that the obtained real-world speech database will be more available in practice. We think it is very important to develop a practicable speech recognition system.

Speech is first filtered to a bandwidth of 8KHz (cutoff frequency) and then digitized at 16KHz sampling rate. Such a giant database consists of 25GB speech data, about 230 hours' utterances.

5 SYSTEM PERFORMANCE AND SUMMARY

The vocabulary set of final system has 2000 phrase of 3 to 5 syllables. User can change the vocabulary set just by editing a text file. Table 1 lists recognition rates for training and testing sets. Further research is in progress to enlarge size of vocabulary set. All testing data used here are taken from the above giant speech database.

Table 1 *Performance of the 2000-phrase real-world system*

Training Set	Rate
M00	99.65%
M01	99.90%
M02	99.90%
M03	99.95%

Testing Set	Rate
M10	97.80%
M11	98.00%
M20	95.40%
M21	98.40%

We think our work can be summarized in a word that the more person tells computer, the better machine can do. Experiments show that

1) Acoustic knowledge is extremely useful in continuous speech recognition.
2) CDCPM is a successful simplified model of HMM.

REFERENCES:

[1] F. Zheng, W.-H. Wu, D.-T. Fang, "CDCPM with Its Applications to Speech Recognition," *Chinese J. of Advanced Software Research*, Supplement, 1996 (Accepted)

[2] J.R. Bellegarda, D. Nahamoo, "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. on ASSP*, vol.ASSP-38, No.12, Nov. 1990, pp.2033-2045

[3] X.D. Huang & M.A. Jack, "Semi-Continuous Hidden Markov Models for Speech Signals," *Computer Speech and Language* (1989), 3:239-251

[4] L. Jiang, W.-H. Wu, L.-H. Cai, and D.-T. Fang, "A Real-time Speaker-independent Speech Recognition System Based on SPM for 208 Chinese Words," in Proc. of *ICSP'90*, pp.473-476, 1990

[5] B.-H. Juang, L.R. Rabiner, "Mixture autoregresive hidden Markov Models for speech signals," *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. ASSP-33, 1985, pp.1404-1413

[6] C.-H. Lee, L.R. Rabiner, "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition," *IEEE Trans. on ASSP*, Vol. 37, No.11, Nov.1989, pp.1649-1658

[7] H. Ney, "Modeling and Search in Continuous Speech Recognition," in Proceedings of *European Conf. On Speech Technology*, Vol.1, Berlin, 1993, pp.491-498

[8] J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. *IEEE*, vol. 63, Apr. 1975, pp.562-580

[9] J. G. Wilpon, L. R. Rabiner, C.-H. Lee, E.R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Trans. on ASSP*, vol.38, No.11, pp.1870-1878, Nov. 1990