International Symposium on Chinese Spoken Language Processing, pp. 23-26, Oct. 13-15, 2000, Beijing

CONTEXT-INDEPENDENT CHINESE INITIAL-FINAL ACOUSTIC MODELING

LI Jing, ZHENG Fang, WU Wenhu

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084
lijing@sp.cs.tsinghua.edu.cn

ABSTRACT

In this paper, a method for the Context-Independent (CI) Chinese Initial-Final acoustic modeling for continuous speech recognition task is proposed. The initial-final (I/F) structure is a characteristic of Chinese language. Initials and finals are smaller units compared to syllables, the use of which is helpful to reduce the number of SRUs. Furthermore, it should be possible to build context-dependent (CD) models. In our experiments, we use knowledge-based criteria to define the CI initial-final units. There are four kinds of CI initial-final units in this paper. The experimental results show that the accuracy of the CI initial-final models is near to or lower than that of the CI syllable model, but the size of model is significantly reduced.

Keywords: SRUs, Context-Independent, Initial-Final Acoustic Modeling

1. INTRODUCTION

Normally, a large-vocabulary continuous speech recognition system has two primary parts: the acoustic model and the language model. The acoustic model is very important for the CSR system. If the acoustic model is not good enough, the performance of the system will not be well, even though the language model is very good. For acoustic model, appropriate SRUs must be chosen, which is a very important factor that may influence the performance of the acoustic model.

It is well known that Chinese is a syllabic language. There are just about 400 toneless Chinese syllables and about 1,300 toned Chinese Syllables. There are some advantages when taking the Chinese syllables as SRUs as some Large Vocabulary Chinese Speech Recognition Systems do [1]. On the other hand, shortcomings exist when taking syllables as SRUs. Firstly, the number of the syllables is relatively large, which will result in large storage requirement for acoustic models. For example, we use the Mixture Gaussian Continuous Probability Model (MGCPM) [2][3] to describe the Acoustic Model, every unit has 6 states, and every state includes 16 mixtures, the size of model is about 10MB. Secondly, the number of the units is too large for CD acoustic models to be built and hence for the co-articulation to be modeled. For example, we classify the left and right contexts between the syllables to 20 classes, respectively. For the CD syllable acoustic model, the number of units will be 167200 (20*418*20), which is too large for the acoustic model. So, it is necessary to choose another kind of SRU, the number of units must be fewer, and its performance should be comparable with the syllable units. Furthermore, it should be possible to build CD models.

Some kinds of subword units can be chosen: semi-syllable, initial-final, consonant-vowel, phoneme and so on. We choose the CI initial-final units as the SRUs in this paper, because this sort of units has some advantages for building the acoustic model. Firstly, the Initial-Final structure is a characteristic of the Chinese language. The initial part corresponds to a consonant, and the final part corresponds to a vowel. Hence a lot of linguistic knowledge can be used to define the initial-final units [4][5]. Secondly, the number of the units can be smaller, that is very helpful for building the CD models. Therefore, we choose the CI initial-final units in our experiments. In Section 2, we will give the definitions of the initial-final units. In Section 3, the experiment design is described. The next sections will show the experimental results and give the conclusions.

2. DEFINITION OF THE INITIAL-FINAL UNIS

In our definition of the initial-final units, the cross-syllable contexts are not considered, but the contexts between initial and final in the syllable are considered. It may be more accurate to call these SRUs "quasi-CD" or "pseudo-CD" initial-final units than "CI" initial-final units. But we call them "CI" initial-final units to emphasize that both for the CI syllable units and the initial-final units the cross-syllable context information has not been considered in this paper.

There are 418 CI syllable units in our system. We can simply divide the syllable into two parts according to the initial units and the final units. In this way, 418 initial units and 418 final units have been obtained (see IF_1 below). Further, we refine these units according to a predefined criterion. In our experiments, we refine the initials by their right contexts. This can be regarded as CD initial modeling. Because the vowel part is relatively stable and the recognition accuracy of the vowels is high, so the refinement of finals does not consider the contexts. The cross-syllable context information has not been considered either.

Four different definitions of the CI initial-final units provided in this paper are as follows.

IF_1 (418i+418f): As mentioned above, every CI syllable can be divided into two parts according to the initial unit and the

final unit. In this way, 418 initial units and 418 final units are generated from the 418 CI syllables.

IF_2 (418i+50f): Because the vowel part is relatively stable, so we classify the final units without considering the context. A new set of initial-final units IF_2 has been obtained after classifying the final units of the IF_1 to 50 classes. The number of final units is more than that of vowels of Chinese language (38), because some final units of the IF_1 have been regarded as different units although they are equivalent according to the vowel. For example, the final units of the toned syllable "de" and the neutral syllable "de" are different. In addition, some final units of the syllable with zero-initial are distinguished from other final units. So, the number of units is observably reduced.

IF_3 (173i+50f): We classify the initial units of the IF_2 by their right context, e.g. the left phonemes of the following final. For example, for the syllables {ba, bai, ban, bang, bao}, three initial units have been classified (e.g. {ba1, ba2 and ba3}). The symbol "ba1" denotes the initial of the syllable {ba}, and the symbol "ba2" denotes the initials of syllables {bai, ban}, while the "ba3" refer to the syllables {bao, bang}. So, we can get the new CI initial-final units set IF_3, in which the number of initial units is 173, and the number of final units is still 50.

IF_4 (87i+45f): This set of units is very similar to the IF_3. We refine the initial units of the IF_2 by their right context (e.g. the leftmost phoneme of the following final). In this way, the initials of the syllables {ba, bai, ban, bang, bao} will seemed to be a same unit "ba". In set IF_4, the number of initial units is 87. The initial units set is

{0a, 0e, 0y, 0o, 0w, 0v, ba, be, bi, bu, ca, ce, ci, cu, cha, che, chi, chu, da, de, de#, di, du, fa, fe, fi, fu, ga, ge, gu, ha, he, hu, ji, jv, ka, ke, ku, ma, me, mi, mu, la, le, li, lo, lu, lv, na, ne, ni, nu, nv, pa, pe, pi, pu, qi, qu, ra, re, ri, ru, sa, se, si, su, sha, she, shi, shu, ta, te, ti, tu, wa, we, xi, xv, za, ze, zi, zu, zha, zhe, zhi, zhu}

The first 6 initial units with prefix "0" are zero-initials. And the "de#" refer to the initial of the neutral syllable "de".

The number of final units in IF_4 is 45. The final units set is

{a, ai, an, ang, ao, e, e#, ei, en, eng, er, ia, ian, iang, iao, ie, in, iu, io, ieu, ing, iong, u, ua, uai, uan, uang, ui, un, uo, ong, v, van, ve, vn, o, ou, 0uai, 0uan, 0uang, 0uei, 0uen, i0, i1, i2}

The "e#" refers to the final of the neutral syllable "de".

All the three symbols i0, i1 and i2 denote the final units "i", where. "i0" is the final "i" following the initials {b, p, m, f, d, t, n, l, g, k, h, j, q, x}, "i1" the final unit "i" following the initials {z, c, s}, and i2 the final unit "i" following the initials {zh, ch, sh, r}.

The final units with prefix "0" denote the finals following the zero-initial "w". These final units have not been merged with other finals according to the same vowels, because their pronunciations are different from others.

3. EXPERIMENT DESIGN

In this section, we will introduce the design of our experiments. Subsection 3.1 describes the speech database used in our experiments and its labeling information. Subsection 3.2 introduces the features extracting and the acoustic modeling. The next two subsections present the details of the training and recognition procedure.

3.1 Speech database and labeling information

The speech database used in our experiments is taken from the "863 assessment" male speech database, of which all the sentences are uttered in standard Chinese with a little regional accent with some background noise. The database consists of 1,560 sentences, divided into three groups called A, B and C. All the speech data from Group A are used in this paper. There are 13 males' speech data and every male reads 521 sentences. We divide these data into two sets: the training set and the testing set. The training set contains 10 males' data (M00, M02, M03, M04, M05, M06, M18, M20, M21, M22), while the testing set contains the rest 3 males' data (M24, M25, M26).

We have the time boundary information in the syllable level generated first the Merging-Based Syllable Detection Automation (MBSDA) [6] and the manual adjustment. This is very benefit to the CI syllable model, for the starting position and the frame length of the syllables can be easily obtained from the information. But there is not any initial-final labeling information. An approximate segmentation method, such as the Non-linear Partition (NLP)[7] or the Linear Partition (LP) algorithm can be adopted to get the boundaries of the initial-final units. This will influences the performance of the CI initial-final acoustic model, for the segmentation method is not very exactitude.

3.2 Features extracting and AM description

The Mel-Frequency Cepstrum Coefficients (MFCCs) are used in our experiments. The feature is 30-demensional vector consisting of 10-demensional NMCEP as well as its first and second order differences. The following formulas are used to calculate the differences,

$$DMFCC(k) = MFCC(k+2) - MFCC(k-2) \quad (1)$$

$$DDMFCC(k) = DMFCC(k+2) - DMFCC(k-2) \quad (2)$$

DMFCC(k) denote the first order difference of the *MFCC* at the k-th frame, while *DDMFCC(k)* the second order. The 30-demensional features are generated by concatenating the *MFCC*, *DMFCC* and the *DDMFCC* coefficients.

We use the MGCPM to describe the acoustic model. The MGCPM refer to the simplified left-to-right HMMs that ignored the probability transition matrix. In the CI syllable acoustic model, every unit includes 6 states and every state is described by a Mixture Gaussian density (MGD), which is composed of 4 Gaussian distributions. In the CI initial-final acoustic model, each initial unit has 1 or 2 states (the zero-initial unit has only 1 state and any other initial unit has 2 states), and the final unit consists of 4 states

3.3 Training procedure

For the CI syllable acoustic modeling, the training procedure is as follows.

- Given the labeling information of the syllables, each observation feature sequence from training data is segmented into N segments (corresponding N states) using some segmentation method such as the NLP.
- For every unit, all of the features belonging to the same state are collected together and then grouped into M classes using some clustering algorithm such as LBG algorithm [8]. After estimating the MGD parameters, the initial CI syllable acoustic model is generalized.
- The Frame Synchronous Search (FSS)[9] algorithm used to decode the observation feature sequence from training data. And then the iterative CI syllable acoustic model can be trained out.

The training procedure of the CI initial-final acoustic model is very similar to the procedure of the CI syllable acoustic model. As mentioned in subsection 3.1, for the CI initial-final acoustic model, we should get the unit labeling information before the training procedure. The training procedure is as follows.

- For every syllable, the state number of the initial unit (N1) and that of the final unit (N2) can be obtained, sum of which is the total state number of the syllable (N=N1+N2). Segment the observation feature sequence into N segments using NLP algorithm. The first N1 states belong to the initial unit, while the rest states belong to the final unit.
- Same to the second step of the CI syllable training procedure. The CI initial-final acoustic model can be obtained after this step.
- For every syllable, the state number can be obtained by summing the state numbers of its corresponding initial and final units. And then we use certain search algorithm such as FSS for the decoding procedure and iterate the CI initial-final acoustic model. This step can be repeated. In this way, an improved CI initial-final acoustic model can be obtained.

Normally, we iterate the acoustic model twice in our experiments.

3.4 Recognition procedure

During the recognition procedure, the FSS algorithm is adopted. The input is the observation feature sequence of a syllable from the testing set and the output is a single syllable. We compute the top1 to top10 syllable accuracy.

For the CI initial-final acoustic model, an initial-final search tree (IFST) is created in order to conveniently process the search procedure. The IFST for the IF_4 is shown in Figure 1.

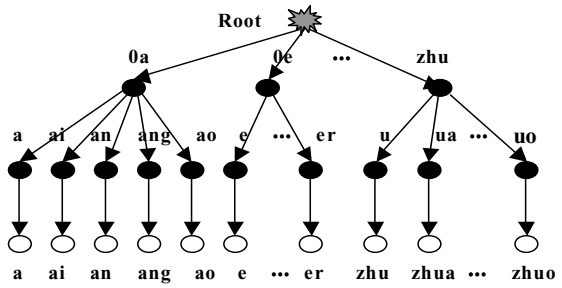

Figure 1. Initial-Final Search Tree for the IF_4

There are four kinds of nodes in the IFST. **The Root Node** is the topmost node of the IFST. It is a virtual node just containing pointers to its child nodes. **The Initial Node** is a node in the second level, which describes a certain initial unit. **The Final Node** is a node in the third level, which denotes a certain final unit. **The Leaf Node** is the bottommost node of the IFST. It contains the string of a syllable according to its parent nodes. During the recognition procedure, a syllable is generated after reaching one of the Leaf Nodes.

4. EXPERIMENT RESULT

In this section we will give the experimental results. It contains the results of the baseline and that of the CI initial-final acoustic model.

The baseline is the CI Syllable acoustic model, which use the toneless Chinese Syllables as the SRUs, the number of which is 418. Every unit has 6 states, and one state has 4 Gaussian mixtures. For the CI initial-final acoustic model, the zero-initial units have 1 state, and the other initial units have 2 states, while the final units have 4 states. Every state is described by 4 Gaussian mixtures. The recognition results are showed in Table 1.

From this table we can see that although the Top1 recognition rates of some CI initial-final acoustic model are lower than the baseline, but the number of units and the model size is significantly reduced, that is very helpful to the CD acoustic modeling.

Table 1. The recognition results of the CI syllable acoustic model and the CI initial-final acoustic model

SRUs of Acoustic Model	Top1	Top5	Top10	AM size (KB)
CI Syllable (418u)	71.2	94.8	96.1	2400
IF_1 (418i+418f)	71.8	94.1	96.7	2404
IF_2 (418i+ 50f)	68.7	93.4	96.6	1349
IF_3 (173i+ 50f)	61.2	90.0	95.2	645
IF_4 (87i+ 45f)	61.6	89.5	94.8	334

For the IF_1, the Top1 recognition rate is higher than the baseline, and the model sizes are almost equal. Actually, for the CI acoustic model, the initial-final acoustic model using the IF_1 units is equivalent to the CI syllable model. Because we can get the CI syllable model, through combining the models of the

initial units and its corresponding final units one by one. So, the IF_1 model can be regarded as the new baseline.

For the IF_2, the Top1 recognition rate is 3% lower than that of the IF_1, but the Top5 and Top10 recognition rates are near to the new baseline (the IF_1). At one time, the number of final units is reduced from 418 to 50, and the model size is decreased from 2,404KB to 1,349KB.

For the IF_3 and IF_4, the Top1 recognition rate is reduced by about 10%, and the Top 10 recognition rate is 1~2% lower than that of the new baseline. But the number of units is significantly reduced, and the model sizes are about 1/4 and 1/7 that of the IF_1.

5. CONCLUSION

In this paper, a method for the CI Chinese Initial-Final modeling in CSR systems is proposed where the initials and the finals are chosen as the units. We define four sets of initial-final units in this paper. The recognition rates of these CI initial-final acoustic models have been given. Via the comparison between the CI acoustic models and the CI syllable acoustic model, we conclude

- For the CI acoustic modeling, if we classify the final units (IF_2), the recognition rates of the initial-final acoustic model are about 3% lower than that of the baseline, but the model size is obviously reduced. If we further reduce the number of initial-final units (IF_3 & IF_4), the recognition rates are getting lower, but the number of units and the model size is significantly reduced. The decrease of the unit number and the model size is very helpful to the CD acoustic model.
- For the CD acoustic modeling, the initial-final units are very useful, even through the unit number has not been reduced. For example, we use the 418 CI syllable units and the IF_1 (418i+418f) to create the CD models. We assume there are 20 classes of the left and right cross-syllable contexts, respectively. For the CD initial-final acoustic model, the number of units will be 16,720 (20*418+418*20), much fewer than that of the CD syllable units (20*418*20 = 167,200), and besides, the CI initial-final unit is smaller than the CI syllable unit. Actually, the initial-final units have been shared by its corresponding syllables having same left or right cross-syllable contexts. The performances of both two CD models will be almost same. If we use other initial-final units (such as IF_2, IF_3 or IF_4), the number of units will much less.

In addition, the labeling information of the initial-final units do not exist for the initial-final acoustic modeling. The approximate approach is used to get the boundaries of the initial-final units in this paper, and this will influence the performance of the CI initial-final acoustic model.

6. ACKNOLEGEMENTS

We would like to thank Xu Mingxing, He Lei and Zhang Jiyong for providing some results and helpful discussions of this work.

7. REFERENCES

[1] Zheng F., Mou X. –L., Xu M. –X., Wu J., and Song Z. –J. "Studies and Implementation of the Techniques for Chinese Dictation Machines". *J. Software,* 10(4):436-444, 1999 (In Chinese)

[2] Zheng, F., Wu, W. -H., and Fang, D. –T. "Center-Distance Continuous Probability Models and The Distance Measure". *J. of Computer Sci. & Tech.*, 13(5):426-437, Sept.1998

[3] Zheng, F., Mou, X. -L., Wu, W. –H, and Fang D. –T. "On the Embedded Multiple-Model Scoring Scheme for Speech Recognition". *ISCSLP'98*, ASR-A3, pp.49-53, Dec.7-9, 1998, Singapore

[4] Wu Z. –J. "The Chinese Phonetics in 'Man-Machine Dialogue'", *Chinese Teaching In The World*, vol4, pp3-20, 1997 (In Chinese)

[5] Yang R. –L., Zhou Y. -M. *The Modern Chinese*, Publishing House of Electronics Industry, Beijing, 1995 (In Chinese)

[6] Zhang J. –Y., Zheng F., Du S., Song Z. –J., and Xu M. –X. "Merging-based Syllable Detection Automation in Continuous Speech Recognition". *J. of Software*, 10(11):1212-1215, 1999 (In Chinese)

[7] Xu M. -X. "Studies on Target Model Independent Segmental Acoustic Model". *Ph.D. thesis*, Tsinghua University, 1999 (In Chinese)

[8] Linde, Y., Buzo, A., Gray, R. M. "An Algorithm for Vector Quantization Design". *IEEE Trans.* O*n COM*-28(1), Jan., 1980, 28(1):84-95

[9] Lee C. -H. and Rabiner L. R. "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition". *IEEE Trans. On ASSP,* Nov. 1989, 37(11):1649-1658