RESEARCH ON INTER-SYLLABLE CONTEXT-DEPENDENT ACOUSTIC UNIT FOR MANDARIN CONTINUOUS SPEECH RECOGNITION

Zhao Qingwei,   Wang Zuoying,   Lu Dajin
Department of Electronic Engineering, Tsinghua University, Beijing, 100084, P.R.C.
Tel: (086)(010)62781704   E-mail: zqw@thsp.ee.tsinghua.edu.cn

ABSTRACT

In this paper we present the algorithm on building inter-syllable context dependent unit for Mandarin continuous speech recognition, in order to avoid coarticulation effects. Firstly the clustering algorithm based on decision tree is introduced, which make full use of the phonological rules. On the basis of information theory, the splitting measurement of the clustering algorithm is studied, which is the difference in entropy result from model splitting. Then the recognition algorithm based on the inter-syllable context-dependent unit is presented, which is based on the principles of dynamic programming. At last, the speaker independent large vocabulary Mandarin continuous speech recognition experiment is discussed. It shows that, the error rate of the recognition system using the context-dependent unit is reduced more than 15% in contrast to the system using the context-independent unit, which reveals the good performance of the new unit.

1. INTRODUCTION

Due to the coarticulation phenomenon in continuous speech, the basic unit used in continuous speech recognition must be context-dependent. In English, the recognition unit often used in system is triphone, or sharing state. In this paper, we will study the problem of how to build and apply the context-depentdent unit for Mandarin continuous speech recognition.

Chinese language differ a lot from English, it consists of only single characters and each character is mono-syllablic. Each syllable includes two sub-syllable units, which can be defined as Initial and Final respectively, and be used as recognition unit. The coarticulation effects in Mandarin continuous speech include inter-syllable and intra-syllable two cases. For old acoustic unit, only the intra-syllable coarticullation is considered. In this paper, the new acoustic unit is built considering not only intra-syllable but also inter-syllable coarticulation effects.

For one sentence in Mandarin continuous speech, we can represented it as follows:

$$C_1V_1C_2V_2)C_3V_3\cdots C_nV_n$$

Considering the inter-syllable influence between Initial(C) and Final(V), the new initial unit can be denoted as $(V_{i-1})C_i$, that is, with regrard to the different V_{i-1}, the new initial unit is different. Similarly, the new Final unit can be denoted as $V_i(C_{i+1})$. However, if all the possible relations are considered, the number of inter-syllable context-dependent sub-syllable unit will be so much that there are not enough data to train the model. Hence the new unit must be clustered. The clustered sub-syllable unit, which is inter-syllable context dependent, is defined as *cdc* (initial) and *cdv* (final) respectively here.

For every recognition unit, the corresponding acoustic model must be built. Here we adopted the DDBHMM(Duration Distribution Based Hidden Markov Model)[1].

2.CLUSTERING ALGORITHM

Usually two type of clustering algorithm is studied. The first algorithm is agglomerative clustering[2], it is a conventional data-driven unsupervised clustering algorithm. Its main disadvantage is that when much new unit emerge that are not covered in training material, then it must be backed off to context independent unit, consequently the context dependency will be lost.

The second algorithm is clustering algorithm based on decision tree[3][4], which can deal with above defects. In fact, this algorithm is a kind of supervised clustering algorithm which make full use of the linguistic knowledge. Therefore, when many new units emerge that are not covered in the training material, the context dependency can still be processed.

2.1 Clustering based on decision tree

Here we introduced the clustering procedure for instance of the cdc unit, the cdv unit is the same.

For every inter-syllable context independent Initial C_i , regard to all of its possible context relations, a set of P={all of the possible $(V_{i-1})C_i$ unit } is built. And for every C_i , one decision tree is built The root node of the tree contents the data structure of set P, and each other node of the tree contains a sub set of P. In addition, every node of the tree include a binary question about the context of C_i . These questions are put forward according to phonological rules and considering the influence between the context unit.

For every not clustering Initial unit, its corresponding cdc unit can be easily found: from the root node of the decision tree, by answering each question associated with this node, the tree are traversed through, and finally one leave node is reached, which corresponds to a kind of unique cdc unit.

In this paper, we do not pay much attention to the clustering algorithm in detail, we only introduce the questions and the splitting measurement in building the tree.

Firstly we introduces the questions used in our clustering algorithm. The linguistic questions are categorical questions querying about the left or right context of current sub-syllable, such as “Is the right context of the current unit a fricative? ”. As we know, the sub-syllable in Chinese can be classified into vowels and consonants. The consonants consist of fricatives, stops, nasals, labials and so on. The vowel can be grouped into 9 classes, considering their ending phoneme as “a”,”o”,”e”, “i”, “u”,”v”, “n”, “ng”,”er” respectively. Certainly, the consonants or vowels can be classified according to more detailed features.

2.2 Splitting Measurement

In this sub section, we will discuss in detail the splitting measurement in building the decision tree. In the clustering procedure, when one node is split into two nodes, the distance measure is the difference in entropy result from model splitting. Only if the difference of the entropy exceeds one threshold, then the splitting of the node can proceed. The difference of the entropy can be computed as follows.

First the entropy of single model is discussed. As to model A, we suppose the data set used to train A includes N_a frames feature vectors, which is denoted as $X_a = \left\{ x_p \right\}_{p=1}^{N_a}$. And the probability distribution of model A is supposed to be represented by continuous gauss density, that is $f_a(x) = N(\mu_{a,}\ B_a)$. The dimension of the feature vector is n .

The entropy of the feature vector in model A can be represented as follows:

$$\begin{aligned} H_a &= -\int f_a(x)\log(f_a(x))dx \\ &= -E(\log(f_a(x)) \\ &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log(|B_a|) + \\ &\quad \frac{1}{2}E[(x-\mu_a)^\tau \cdot B_a^{-1} \cdot (x-\mu_a)] \end{aligned} \qquad (1)$$

It is easy to demonstrate:

$$E[(x-\mu_a)^\tau \cdot B_a^{-1} \cdot (x-\mu_a)] \quad = n \qquad (2)$$

then

$$H_a = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log(|B_a|) + \frac{n}{2} \qquad (3)$$

Suppose the N_a frames in the data set X_a is independent with each other, then the entropy of all of the vectors in X_a is the sum of their single entropy, which is $N_a H_a$. We adopted it as the entropy of model A, and defined as weighted entropy, that, is

$$\tilde{H}_a = N_a H_a \qquad (4)$$

when the model A is decomposed into model B and C ,

$$N_a = N_b + N_c$$

So the change in the weighted entropy is represented as:

$$\Delta \tilde{H} = \tilde{H}_a - \tilde{H}_b - \tilde{H}_c$$

$$= \frac{N_a}{2}\log(|B_a|) - \frac{N_b}{2}\log(|B_b|) - \frac{N_c}{2}\log(|B_c|) \qquad (5)$$

We can also deduce the above formula by another approach based on the maximum likelihood training algorithm which is used in DDBHMM[1].

First the likelihood of X_a generated by model A is denoted as L_a :

$$L_a = \log(f_a(X_a)) \qquad (6)$$

$$= \sum_{p=1}^{N_a} \log(f_a(x_p))$$

$$= N_a\left[-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|B_a|)\right]$$

$$-\frac{1}{2}\sum_{p=1}^{N_a}(x_p - \mu_a)^{\tau} \cdot B_a^{-1} \cdot (x_p - \mu_a) \quad (7)$$

since

$$\sum_{p=1}^{N_a}(x_p - \mu_a)^{\tau} \cdot B_a^{-1} \cdot (x_p - \mu_a)$$

$$= \sum_{p=1}^{N_a}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_{p_i} - \mu_i)(x_{p_j} - \mu_j)v_{ij}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}v_{ij}\left[\sum_{p=1}^{N_a}(x_{p_i} - \mu_i)(x_{p_j} - \mu_j)\right] \quad (8)$$

where v_{ij} represent the element of the matrix B_a^{-1}, which is the inverse matrix of covariance matrix B_a.

According to the maximum likelihood training algorithm, the covariance matrix is estimated as follows:

$$b_{ij} = \frac{1}{N_a}\sum_{p=1}^{N_a}(x_{p_i} - \mu_i)(x_{p_j} - \mu_j) \quad (9)$$

from (8) and (9), we can get:

$$\sum_{p=1}^{N_a}(x_p - \mu_a)^{\tau} \cdot B_a^{-1} \cdot (x_p - \mu_a)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}v_{ij}b_{ij}N_a$$

$$= N_a n \quad (10)$$

from (10) and (7), we can get

$$L_a = -N_a\left[\frac{n}{2}\log(2\pi) + \frac{1}{2}\log(|B_a|) + \frac{n}{2}\right] \quad (11)$$

then

$$L_a = -N_a H_a = -\tilde{H}_a \quad (12)$$

Therefore the relation of difference of weighted entropy with the likelihood is:

$$\Delta\tilde{H} = L_b + L_c - L_a \quad (13)$$

3. RECOGNITION ALGORITHM

Our Chinese continuous speech recognition system includes two passes, in the first pass the syllable candidates are recognized, and in the second pass the syllable candidates are translated into Chinese characters. To give enough information to the second pass, the first pass must provide syllable candidates in the form of syllable lattice. Therefore, when the new unit is applied, the recognition algorithm on building syllable lattice in the first pass must be rebuilt. The recognition principle based on the old unit can be seen from references[6].

The network graph must be firstly constructed when recognizing continuous speech using the new recognition unit.

3.1 Recognition network

The basic unit of the recognition network is syllable template. For each syllable W_i, its semi-syllable unit is C_p, V_q。Initial C_p is supposed to include K_p cdc units，which are $cdc_{p1}, cdc_{p2,}..., cdc_{pK_p}$ respectively。Similarly, Final V_q is suppose to include M_q cdv units ， which are $cdv_{q1}, cdv_{q2}, ..., cdv_{qM_q}$.

We can denote the template in a concise form, as showed in figure 1.

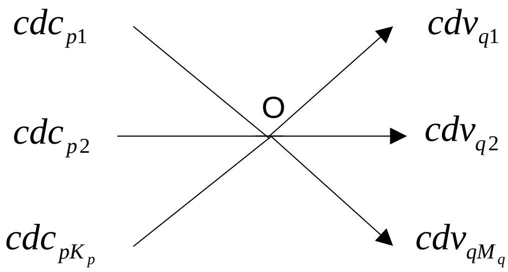

Figure1. the structure of syllable node

The last state of the unit cdc is connected in a point O , which is used as the start state of each unit cdv. That is, the optimal(minimum match distance) cdc unit is selected from the K_p cdc units, and then it can jump to the M_q cdv units. This structure is defined as syllable node, where cdc_{p1}、cdv_{q1} is defined as semi-syllable node.

The total network is composed of syllable node or semi-syllable node. The network graph is showed in figure 2.

3.2 Search Algorithm

The recognition process is to search an optimal path in the network, whose match distance is minimal. Here the distance refers to the negative log of the likelihood value. The search in the network is realized according to the principle of dynamic programming[5].

Now we introduce the procedure of the searching in the network.

3.2.1 match and jump in the syllable node

For each syllable node, as W_1， W_2， W_3 showed in figure 2, following steps are adopted：

1) Each cdc node is matched by Viterbi method.

2) The accumulative distance of the last state of each cdc node is compared, and the $cdc_{\min}$ node with minimum distance is selected.

3) Each cdv node is matched by Viterbi method. The first state of each cdv node must be compared with the last state of previous $cdc_{\min}$, to determine whether it needs updating.

During the matching procedure above, the variable for tracing back must be transferred。

3.2.2 construct auxiliary cdc node

When jumping between the syllable, the auxiliary cdc node should be built to store the intermediate result. Each cdc node at time $(t+1)$ may come from some（supposed to be N）cdv at time t，select the node with the minimum distance among the N cdv nodes, its accumulative distance is given to the auxiliary node cdc at time t. Then the auxiliary node can jump to the cdc nodes at time $(t+1)$.

3.2.3 jump between the syllable

The essence of the jump between the syllable node is the initialization of the cdc node. In figure 2, the connection line from the right dashed block to the left cdc node reflects this procedure. For each cdc in the syllable node w, if the unit no of the previous auxiliary cdc node is the same with w, then the accumulative distance of the first state of the cdc node is compared with the distance of the auxiliary node，if the latter is small, then the contents of the first state of the cdc node is updated by the auxiliary node. Consequently the jump between the syllable nodes is accomplished.

3.2.4 trace back and generate the syllable lattice

When the optimal path with the minimal accumulative distance is found in the recognition network according to the sequential order, the search space must be traced back in order to obtain the syllable node in the optimal path and the splitting point between them, at the same time, the lattice of the syllable candidate can also be obtained.

4. EXPERIMENT

At last, the experiment on speaker-independent large vocabulary Mandarin continuous speech recognition is presented.

The database is composed of continuous Chinese speech. The training database includes the speech data uttered by 82 speakers, each one speaks about 520 sentences. The test speech data includes 120 sentences spoken by six persons who are not within the 82 speakers. In addition, the test sentences are totally different from the training material, though both of them are extracted from the People's Daily in China.

Here we only concerns about the recognition result in the first pass of our system, that is, only the syllable not the Chinese character is recognized. The experiment result is showed in table1. It shows that the inter-syllable context dependent unit obtained from the decision-tree-based clustering algorithm reduces the error rate by more than 15%. In detail, The error rates of the 1-best, 2-best , 5-best, 25-best candidates of the system using the new unit are 17.14%, 8.59%, 3.47%, 1.21% respectively, the relative reduction from the baseline system is 15.0%, 20.1%, 28.2%, 26.7% respectively. This illustrates the good performance of the inter-syllable context-dependent unit.

Table1. Error rate of Mandarin continuous speech

	Top1	Top2	Top5	Top25
Base	20.16	10.75	4.83	1.65
New	17.14	8.59	3.47	1.21

5. CONCLUSION

In this paper we present the algorithm on building inter-syllable context dependent unit for Mandarin continuous speech recognition, in order to avoid coarticulation effects. Firstly the clustering algorithm based on decision tree is presented, which make full use of the phonological rules. This method demonstrates especially better when many new units emerge that are not covered in the training material. On the basis of information theory, the splitting measurement of the clustering algorithm is studied, which is the difference in entropy result from model splitting. Then the recognition algorithm based on the inter-syllable context-dependent unit is presented, which is based on the principles of dynamic programming. At last, the speaker independent large vocabulary Mandarin continuous speech recognition experiment is discussed. It shows that, the error rate of the recognition system using the context-dependent unit is reduced more than 15% in contrast to the system using the context-independent unit, which reveals the good performance of the new unit.

REFERANCES

[1] Wang Zuoying, “Improved hidden Markov Model in Speech Recognition”, 863 Smart Computer System Conference”, China, 1988

[2] Kai-Fu, Lee, “Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous peech Recognition”, IEEE Trans on ASSP, vol.38, No.4, Apr, 1990, pp509-609

2. [3] L.R.Bahl, etc, “Decision Tree for Phonological Rules in Continuous Speech”, ICASSP, 1991, pp185-188

[4] Mei-Yuh Hwang, etc, “Predicting Unseed triphones with Senones”, IEEE Trans. on SAP, Vol.4, No.6, Nov, 1996, pp412-419

[5] H.Ney, “the use of a one_stage dynamic programming algorithm for connected word recognition”, IEEE Trans on ASSP, vol.32, no.2, pp263-271, Apr, 1984

[6] Zhao Qingwei, Wang Zuoying, Lu Dajin, “Improved Algorithm with Duration Information for Continuous Speech Recognition”, published in *Journal of Tsinghua University*, vol.37, Oct, 1997.

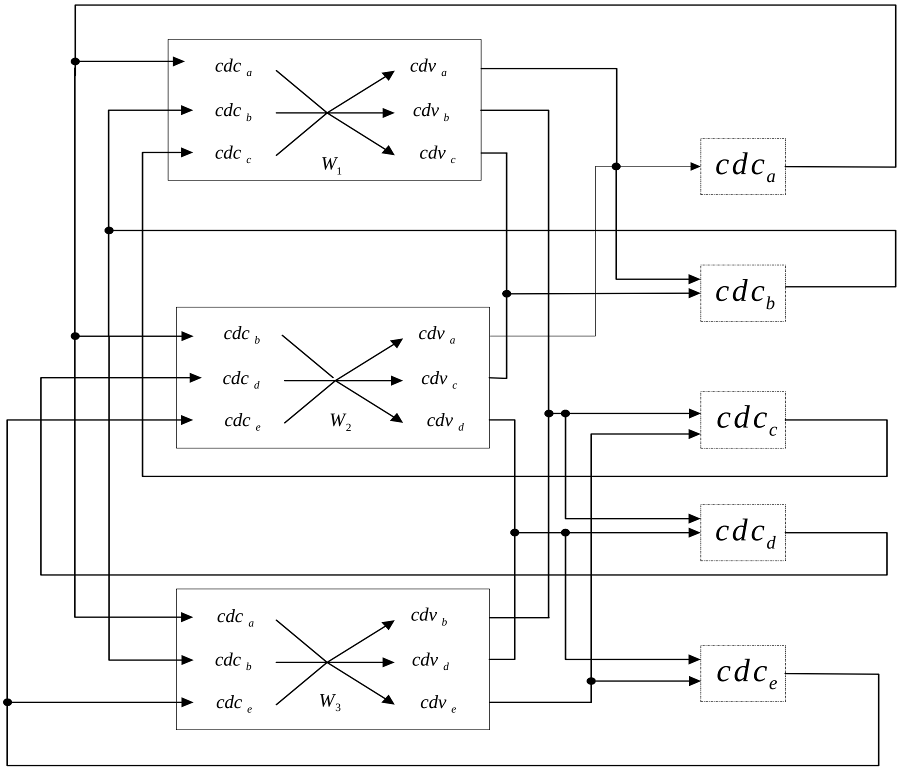

Figure2. The recognition network based on the inter-syllable context dependent unit