An MCE based Classification Tree Using Hierarchical Feature-Weighting in Speech Recognition

Fan Wang, Fang Zheng, and Wenhu Wu

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China
{wangf, fzheng, wuwh}@sp.cs.tsinghua.edu.cn; http://sp.cs.tsinghua.edu.cn

Abstract

In this paper a hierarchical classification framework using the feature-weighting tree for the objective of applying diverse weighting to acoustic features is proposed for speech recognition. The hierarchical feature-weighting tree with a flexible structure complexity can be constructed optimally with the optimal splitting for the recognition confusion graph. Based on the minimum classification error principle, the subset-dependent training and the multi-level recognition method are proposed, where the feature weighting can be automatically trained without normalization in recognition. Both the mathematical analysis and the experimental results show that such a supervised hierarchical classification tree based on the feature weighting is efficient to reduce the speech recognition error.

1. Introduction

In current speech recognition research, feature extraction is one of the most important, and on the other hand, the most difficult issues. Actually, Feature extraction is an information mining procedure for both improving the discriminability and reducing the redundancy of the original input signal.

There are a number of different speech features for speech recognition currently, and mel-frequency cepstrum coefficients (MFCC) is one of the most popular ones [1]. For MFCC extraction, the parameters of the window size, the band-pass filters, and the feature dimension are all fixed.

Intuitively, it should be that different kinds of measurements are needed to make phonetic distinctions among different types of target speech signals, similar to human cognition. The measurements that are best to discriminate some classes of signals will most likely be sub-optimal for other classes. From this point of view, many researchers have achieved success on measurement optimization for feature extraction, such as using a massive parallel feature matrix that consists of a large number of different features, where the mixture weights can depend on the acoustic class and the relevant context estimated by such training criteria as the Maximum Likelihood (ML) criterion or the Minimum Classification Error (MCE) criterion [2].

Hallerstadt and Glass used multiple features with mixture weights by product rule for phonetic and word recognition [3]. Jiang and Huang developed a context-dependent feature and multiple-feature decoding by the maximum *a posterior* (MAP) principle driven by the training data [4]. However in [3], the mixture weights are uniform, in other words, the measurements for the features of all acoustic models are the same. And in [4] the principle is that the feature measurement is model-dependent, and the normalization of such diverse measurements is necessary in recognition. Without the normalization, the recognition scores generated by different measurements cannot be directly comparable due to their different dynamic ranges.

In the framework of the above methods for optimal measurement of acoustic feature, there exists a main conflict between the diverse measurements and their normalization for recognition. We want to use diverse measurements instead of uniform measurement for different models; however, there is not a satisfying solution to the normalization of these different measurements in recognition for the time being.

In this paper, a hierarchical classification framework is proposed to attempt to solve this conflict. Using the hierarchical measurements instead of the uniform measurement, the recognition is divided into several levels, where the task in each level is a subset classification with a uniform measurement of acoustic feature, in other words, the measurement is subset-dependent. All these measurements compose a feature-weighting tree for hierarchical classification. The construction of such a hierarchical classification tree is based on the minimum classification error (MCE) principle, using an optimal splitting method of the recognition confusion graph. The feature weighting in each level is obtained by the MCE training criterion along with the optimal sub-graph splitting. By using this method, the above conflict can be solved, where the classification is diverse feature weighting based and no normalization is needed because the feature weighting in classification for each level is uniform.

The experiments are done across a large vocabulary Chinese speech recognition task. The context-independent modeling for Chinese finals and initials is used. The results show that the use of MCE classification tree based training method could lead to a significant improvement of the recognition performance.

This paper is organized as follows. In Section 2, the feature weighting, the measurement of acoustic feature we focus on, with its application to speech recognition is introduced. The framework and mathematical analysis of the minimum-error hierarchical classification are given in Section 3. Some key components of the hierarchical

classification tree construction are discussed in Section 4. In the next section the database and our experimental results are showed with explanation. The conclusions are drawn in Section 6.

2. Feature weighting

The goal of the speech recognition for the acoustic model set $\{A^i, 1\le i\le I\}$ is to find the model A^* that maximizes the probability of A^i given the observation feature vector sequence $X=[x_1, x_2, \ldots, x_T]^T$ where $x_t= [x_{t1}, x_{t2}, \ldots, x_{tD}]^T$ is the D-dimensional feature vector at time frame t. Namely,

$$A^* = \arg\max_i P(A^i \mid X) \tag{1}$$

In the above criterion, there is no weighting for each dimension of the feature vectors, which can be regarded as being weighted using a unit vector. If a non-unit weighting vector, say $w=(w_1, w_2, \ldots, w_D)$, is used, the weighted feature vector sequence Y can be represented as $Y=[y_1, y_2, \ldots, y_T]^T$, where

$$y_t = W(w, x_t) \triangleq \left[w_1 x_{t1}, w_2 x_{t2}, \cdots, w_D x_{tD}\right] \tag{2}$$

If Hidden Markov Model (HMM) is used for the acoustic modeling, Equation (1) can be rewritten as

$$\begin{aligned} A^* &= \arg\max_i P(A^i \mid Y) \\ &= \arg\max_i \left(\sum_{t=1}^{T} \left[\log a^i_{q^i_{t-1} q^i_t} + \log b^i_{q^i_t}\left(W(w, x_t)\right) \right] + \log \pi^i_{q^i_0} \right) \end{aligned} \tag{3}$$

where $q^i = (q^i_0, q^i_1, \cdots, q^i_T)$ is the resulted optimal state transition sequence that maximizes probability $P(A^i|Y)$ given model A^i, and $(\pi^i, \{a^i\},\{b^i\})$ is the HMM parameter set of A^i.

In fact, the feature-weighting vector can be considered as a measurement for the corresponding feature vector. In our proposal, different subsets of models have different weighting vectors.

3. MCE based classification tree

Suppose $e(A^i, A^j|w)$ is the classification error count when the feature sequences belonging to model A^i are misclassified into model A^j using w as the weighting vector.

The total classification error count for any acoustic model subset $S=\{A^i\}$ given weighting w can be calculated using

$$E(S|w) = \sum_{\substack{i\ne j \\ A^i, A^j \in S}} e\left(A^i, A^j \mid w\right) \tag{4}$$

An optimal weighting vector for S can be obtained

$$w^*(S) = \arg\min_w E\left(S \mid w\right), \tag{5}$$

and accordingly the minimum classification error is

$$E^*(S) = E\left(S \mid w^*(S)\right) \tag{6}$$

Spilt S into two subset S_1 and S_2, where $S=S_1+S_2$ and $S_1 \cap S_2=\Phi$. From equation (6), we have

$$\begin{aligned} E^*(S) &= E\left(S \mid w^*(S)\right) \\ &= E\left(S_1 \mid w^*(S)\right) + E\left(S_2 \mid w^*(S)\right) + E\left(S1, S2 \mid w^*(S)\right) \end{aligned} \tag{7}$$

where

$$E(S_1, S_2 \mid w) = \sum_{A^i \in S_1, A^j \in S_2} e\left(A^i, A^j \mid w\right) \tag{8}$$

is the inter-class classification error between S_1 and S_2.

Though for most cases an optimal weighting vector over a model set will not be always the optimal one for its subsets, we can find optimal weighting vectors for its subsets such that

$$E^*(S_i) = E\left(S_i \mid w^*(S_i)\right) \le E\left(S_i \mid w^*(S)\right), \quad i = 1,2 \tag{9}$$

It is obvious that

$$E^*(S_1) + E^*(S_2) + E\left(S_1, S_2 \mid w^*(S)\right) \le E^*(S) \tag{10}$$

Hence, the classification error over a model set can be reduced if the sum of classification errors over its exclusive subsets could be reduced. Based on this idea, a hierarchical classification framework is developed, where the classifying is a multi-level procedure. The focus in each level is the classification of a specified subset with its own optimal feature weighting. All these hierarchical feature weighting could be represented by a tree structure illustrated in Figure 1.

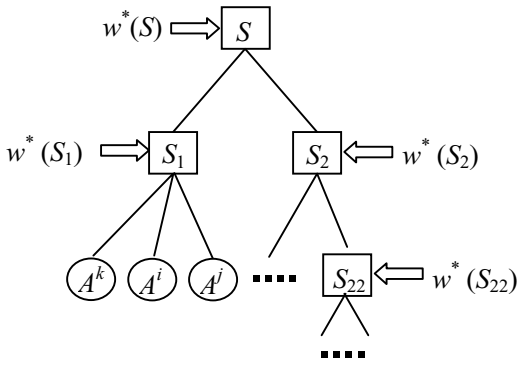

Figure 1. Illustration of the tree structure for hierarchical feature weighting

Such a classification tree for hierarchical feature weighting can be constructed using the above subset splitting procedure iteratively till no subset can be split, i.e. each terminal node in the tree corresponds to an acoustic model.

The recognition based on such a hierarchical classification tree is also a multi-level procedure. As shown in Figure.1, the input feature sequence X is firstly classified into one subset, say S_1, of S with the feature weighting vector $w^*(S)$ by Equation (3). And secondly a deeper level subset of S_1 will be found with weighting vector $w^*(S_1)$ for X. As the level goes deeper and deeper, a terminal node A^k will be reached finally. At this time, X is regarded to be classified into A^k.

4. Keys to classification tree construction

4.1. The way to optimally split S into S_1 and S_2 with $E(S_1,S_2|w^*(S))$ minimized

In order to split an acoustic model set into two subsets, a confusion graph is constructed, where each vertex represents one acoustic model and the edge between two vertexes means their inter-model classification error.

Actually, $E(S_1, S_2|w(S))$ is the sum of all edges in a cut set of the confusion graph given S_1 and S_2. Hence, the procedure for optimally splitting S into two sub-graphs S_1 and S_2, becomes the solving to the problem to find the minimum cut set of the given confusion graph of S. Such a search can be implemented using the Graph Maximum Flow algorithm [5].

4.2. Acceptable error threshold for subset splitting

In order to reduce the computation complexity, and most importantly, to avoid the over-fitting problem due to the mismatch between the training and the testing data, the subset splitting procedure will stop when the classification error of current subset is smaller than a predefined acceptable error threshold. When the splitting stops, all the models in the subset share the same feature-weighting vector. Because the training aims at minimizing the classification error for the training data, not for the testing data directly, the optimal parameters for the training data may not be optimal for the testing data. From the point of generalization view, defining such an acceptable error threshold will not only reduce the number of the nodes in the classification tree but also reduce the mismatch between the optimal weightings for the training data and those for the testing data. In recognition systems, the threshold can be either fixed or dynamic according to a certain criterion.

4.3. Feature-weighting training

The purpose of the feature-weighting training is to find an optimal feature-weighting vector for a subset in a certain level. It consists of two steps:

- ***Step*** 1: Inherit the weighting from its direct superset.
- ***Step***2: Update the weighting only by the data belonging to the most incorrect models based on the MCE criterion using the generalized probabilistic descent (GDP) algorithm [6]. The detailed equations are followed:

$$w_d = w_d - \varepsilon \frac{\partial}{\partial w_d} L(Y, A^i) \tag{11}$$

where

$$L(Y, A^i) = \frac{1}{1 + e^{-d_i(Y, A^i)}} \tag{12}$$

is the loss function of classification for model A^i given the feature sequence Y. And,

$$d_i(Y, A^i) = -P(Y \mid A^i) + \max_{j \neq i} P(Y \mid A^j) \tag{13}$$

is the disciminant function.

From Equations (3), (12), and (13), we have

$$\frac{\partial}{\partial w_d} L(Y, A^i) = \frac{\partial L(Y, A^i)}{\partial d_i(Y, A^i)} \cdot \frac{\partial d_i(Y, A^i)}{\partial w_d} \tag{14}$$

$$\frac{\partial L(Y, A^i)}{\partial d_i(Y, A^i)} = L(Y, A^i)\left(1 - L(Y, A^i)\right) \tag{15}$$

$$\frac{\partial d_i(Y, A^i)}{\partial w_d} = -\sum_{t=1}^{T} \frac{\partial \log b^i_{q^i_t}(y_t)}{\partial w_d} \tag{16}$$

If the output distribution is the mixed Gaussian distribution with diagonal covariance matrix, Equation (16) can be rewritten as

$$\frac{\partial \log b^i_{q^i_t}(y_t)}{\partial w_d} = -\frac{(2\pi)^{-D/2}}{b^i_{q^i_t}(y_t)} \times \sum_{k=1}^{K}\left[\frac{c^i_{q^i_t k}}{\sqrt{|R^i_{q^i_t k}|}} \cdot \sum_{d=1}^{D}\left(\frac{y_{td} - \mu^i_{q^i_t kd}}{\left(\sigma^i_{q^i_t kd}\right)^2} \cdot x_{td}\right) \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{y_{td} - \mu^i_{q^i_t kd}}{\sigma^i_{q^i_t kd}}\right)^2\right\}\right] \tag{17}$$

Where, c^i_{qk}, μ^i_{qkd} and $\left(\sigma^i_{qkd}\right)^2$ are the mixture gain, the mean and the covariance for dimension d, mixture k and state q of model A^i, and $R^i_{qk} = Diag\left[\sigma^i_{qk1}, \sigma^i_{qk2}, \cdots, \sigma^i_{qkD}\right]$.

5. Experimental Results

The experiments are done across the "863" Chinese continuous speech database [7]. The data of 10 males is taken as the training corpus while the data of another 2 males as the testing corpus, where the speech data of each male includes 521 utterances and each utterance contains a Chinese sentence of 6~10 syllables on an average.

The speech recognition unit set include 21 Chinese initials and 37 finals with final /*ueng*/ excluded, as listed in Table 1. There is also an extra garbage model for silence. The HMM has a left-to-right non-skip structure with three states for each unit and 16 mixed Gaussians for each state. For simplification, only context-independent modeling is considered for the time being.

Table 1: Speech recognition units in the experiments

Chinese Initial	Chinese Final
b,p,m,f,d,t,n,l,g,k,h,j,q, x,z,c,s,zh,ch,sh,r	a,ai,an,ang,ao,e,ei,en,eng,er,o, ong,ou,i,i1,i2,ia,ian,iang,iao,ie, in,ing,iong,iou,u,ua,uai,uan, uang,uei,uen,uo,v,van,ve,vn

The original feature is 34-dimentional, consisting of 16-dimentional MFCC and a frame energy as well as their first order derivatives.

The HMM models are trained by means of HTK [8]. Once trained, the HMM parameters, including mixture gains, mean vectors, covariance matrices, and initial and transition probability matrices, will remain the same.

The normal experiment using HTK is taken as the baseline. The Chinese initial and final boundary information will be kept for later use after the baseline experiment has been done. In the following experiments,

the feature weighting is being tested in an isolated word recognition manner.

In Table 2, the comparison in recognition performance between the baseline and the hierarchical classification tree based training criterion, where the acceptable error threshold is set to 5%.

Table 2: The recognition accuracy comparison (Acceptable error threshold is 5%)

Training Method	Training Corpus	Testing Corpus
Baseline (ML)	80.9%	70.1%
Classification Tree	89.9%	72.2%
Error Rate Reduction	*47.1%*	*7.0%*

As shown in Table 2, it is easy to see that the feature-weighting tree based hierarchical classification method is very efficient to increase the accuracy rate for the training set with even a 47% error rate reduction. Furthermore, the performance for the testing set is also improved, where the error rate reduction is 7%. The mismatch between the training corpus and the testing corpus is the main reason for the big difference in error rate reduction. It is obvious that bigger consistency between the training and testing corpora will always lead to bigger performance improvement of the recognition.

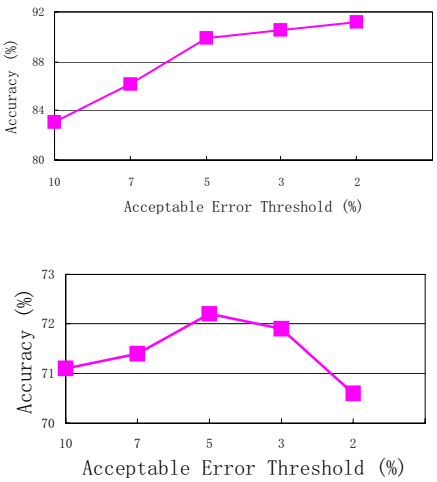

Figure 2: Recognition accuracy curves for feature-weighting tree based hierarchical classification over both the training set (upper) and the testing set (lower)

Figure 2 plots the curves of recognition accuracies for both the training and the testing corpora vs. different acceptable thresholds. For the training corpus, the best recognition accuracy is achieved when 2% is set as the threshold. Generally, the recognition error will be reduced with the decrease of the acceptable error for the training corpus. However, for testing corpus, the best recognition accuracy is not at the point of the minimum acceptable error. The reason of this result is due to the over-fitting mentioned in section 4.2. Hence, the selection of acceptable error is one of the most important components in this method.

6. Conclusions

A hierarchical classification framework based on the diverse measurements for training and recognition is proposed in this paper in order to reduce the overall classification error. The feature-weighting tree is an example of implementation of such diverse measurements for the hierarchical classification framework. Using the tree structure, which is constructed using the optimal splitting of the recognition confusion graph, an MCE based subset-dependent training method and a multi-level recognition procedure are proposed. In this data driven training scheme, the feature-weighting is automatically trained by the GDP algorithm to minimize empirical error counts for the given subset, and no normalization is required in recognition. Furthermore, the complexity of the hierarchical classification tree can be changed flexibly by setting the acceptable error threshold. The use of the threshold is also useful for avoiding the over-fitting.

The experimental results show that the error rate can be reduced significantly, by 47% and 7% for the training and the testing corpora, respectively, when using the feature-weighting tree based hierarchical classification method.

In this paper, the feature weighting is the unique measurement for the hierarchical classification. Actually, other parameters can be applied to such a hierarchical classification framework, such as the HMM parameters, feature transformation and so on. On the other hand, other training criterions could also be used for the framework, for example we can adopt a maximum mutual information (MMI) based hierarchical classification method.

7. References

[1] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, 28:357-366, 1980.

[2] Chou, Wu, "Discriminant-Function-Based minimum recognition error rate pattern-recognition approach speech recognition", *Proceedings of the IEEE*, 88(8): 1201-1223, 2000.

[3] Hallberstadt, A.K. and Glass, J.R., "Heterogeneous measurements and multiple classifiers for speech recognition", *Proceeding of ICSLP-98*, 1998.

[4] Jiang, Li and Huang, X.D., "Unified decoding and feature representation for improved speech recognition", *Proceeding of EuroSpeech-99*, 1999.

[5] Gross, J.L. and Yellen, J., *Graph Theory and Its Applications*, CRC Press, 1999.

[6] Juang, B.-H. and Katagiri, S., "Discriminative learning for minimum error classification", *IEEE Trans. Signal Processing*, 40:3043-3054, 1992.

[7] Zheng, F., Song, Z.-J., and Xu, M.-X., "EASYTALK: A large-vocabulary speaker-independent Chinese dictation machine", *Proceeding of EuroSpeech-99*, 2: 819-822, Budapest, Hungary, 1999.

[8] Young, S., "A review of large-vocabulary continuous speech recognition", *IEEE Signal Processing Magazine*, 13(5): 45-57, 1996.