Toward On-line Learning of Chinese Continuous Speech Recognition System

Rong Zheng, Zuoying Wang

Department of Electronic Engineering, Tsinghua University
Beijing 100084, P.R. China

ABSTRACT

In this paper, we presented an integrated on-line learning scheme, which combined the state-of-art speaker normalization and adaptation techniques to improve the performance of our large vocabulary Chinese continuous speech recognition (CSR) system. We used VTLN to remove inter-speaker variation in both training and testing stage. To facilitate dynamic transformation scale determination, we devised a tree-based transformation method as the key component of our incremental adaptation. Experiments shows that the combined scheme of on-line learning (incremental & unsupervised) system, which gives approximately 22~26% error reduction rate, was proved to be better than either method when used separately at 18.34% and 2.7%..

1. INTRODUCTION

In the past two decades, considerable progress has been made in speech recognition technology. However, for speaker independent (SI) systems, their recognition accuracy is still worse than well-trained speaker dependent (SD) system and automatic speech recognition performance often degrades rapidly when there is a mismatch between the testing and the training conditions like outlier speakers. That is why there has been much interest in adaptation/normalization techniques for large vocabulary speech recognition recently. Current approaches to tackle this problem can be roughly categorized as [2], speaker normalization and speaker adaptation methods. Because SI systems are trained with a large amount of training data from many speakers to "remove" inter-speaker variety, there are two inherent weaknesses with it. (1) the resultant statistical models have to deal with wide range of variation caused by inter-speaker variability which will lead to reduced discriminatory capability and diffused acoustic models. (2) For outlier speakers who are not included in the training set, the performance will degrade sharply. Speaker normalization techniques like vocal tract length normalization [1](VTLN) and speaker adaptive training (SAT) tend to remove or alleviate inter-speaker variability thus result in a more "speaker-independent" model. In speaker adaptation (SA), the pre-trained speech recognition system is modified toward SD one by adapting SI codebooks by a few speech provided by the new speaker. According to the adaptation mode, SA methods can be (i) batch adaptation, where a limited amount of enrollment data are collected and are used to train the adaptation (ii) incremental adaptation, where the system adapts every section of utterances and uses the adapted model for the next few utterances and (iii) instantaneous adaptation, where attempts are made to improve recognition on the same data that are used to estimate the adaptation transformation. It is clear that unsupervised incremental adaptation is more attractive in both feasibility and computational cost. In this paper, we try to integrate speaker normalization and adaptation techniques in an incremental learning scheme to improve the performance our Chinese continuous speech recognition system given by fig. 1.

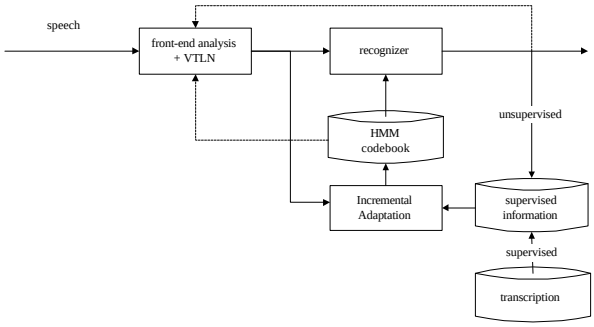

Figure 1: Framework of Incremental Learning of HMM

This paper is organized as follows. After the introductory part, VTLN is described in Section II. Motivation and realization of incremental transformation are presented in great detail in Sections III. In section IV, experiment results are given. Finally, this paper is concluded with section V.

2. VOCAL TRACT NORMALIZATION

From a physiological point of view, it is generally agreed that one of the major sources of inter-speaker variation is the vocal tract length (VTL). The primary effect of vocal tract length variation between speakers is a scaling of frequency. The aim of vocal tract length normalization is to estimate a frequency scaling factor for each speaker or utterance and then normalize the speech signal to an average vocal tract length so the parameterized speech is independent of this type of inter-speaker variation. The main issues to be addressed in the implementation of VTLN are (I) the estimation of normalization factor and (ii) how to do the frequency scaling.

The warping factor can be either estimated by searching a discrete set of possible scalings or using a more direct approach based on measuring formant frequencies. It has been showed [4] that formant approach is inferior because estimation of formant frequency is highly context-dependent and the criterion for calculating warping factor is not consistent with that of training stage. To do frequency scaling, there are also many choices. The scaling can be done in time domain, frequency domain or for filter-bank based front-end processing be integrated with filter-bank analysis [1]. The scaling can be linear or nonlinear. In this paper, we adopted a search approach based on ML criterion and filter bank based normalization following [4].

3. INCREMENTAL TRANSFORMATION

3.1. Motivation

In [3], Fan Zhang developed a block-wised transformation scheme. An initial set of SI models is adapted to the new speaker by transforming the mean parameters of models with a set of linear transformations. Same transformation is used across a number of distributions called "block". Assuming the codebook space is grouped into M blocks via K-mean clustering, for each block, let A_m be the corresponding transformation matrix and b_m be the bias, $m = 1,2,...M$; Given the HMM state $s_{i,m}$ in block m, the normalized codebook $\{\mu_{i,m}\}$ could be obtained through the transformation,

$$\hat{\mu}_{i,m} = A_m \mu_{i,m} + b_m \quad , \mathrm{I} = 1, 2, ., . k_m$$

k_m is the number of distributions in block m. Under this assumption, the probability density function(pdf) of speaker-adapted(SA) observations will have the form

$$P(O_t \mid S_{i,m}) \sim N(O_t : A_m \mu_{i,m} + b_m, \Sigma_{i,m})$$

where O_t stands for the observation vectors of HMM state $s_{i,m}$ from the adaptation data , N(;) denotes the Gaussian densities, A_m and b_m, m = 1,2,.M can be obtained by maximizing the a posterior probability of observtion O_t as given in,

$$\{\hat{A}_m, \hat{b}_m\} = \arg\max_{\{A_m, b_m\}} \{\prod_{m=1}^{M} \prod_{t=1}^{T} P(A_m \mu_{i,m} + b_m, \Sigma_{i,m} \mid O_t)$$

This maximization problem can be solved by E-M algorithm. The adaptation scheme is performed in batch mode for supervised adaptation over a fixed amount of data. The sentences to be adapted and the number of transformation blocks are determined beforehand by experiment. To extend this method to incremental adaptation, in which both the content and amount of adaptation data are not known, several important changes must be made.

3.2. Tree-based block transformation

The key idea of tree-based block transformation is as the amount of data for adaptation changes, the scale of transformation should change accordingly. For example, when there are only a small amount of data, the transformation should be very coase, while as more data are available there should be more blocks accordingly. To facilitate dynamic block determination, the acoustic models are first arranged into a tree in which each leaf node corresponds to one Gaussian distribution. A "cut" [6] denoted a set of nodes whose non-overlapping leaves comprise the whole model space as showed in fig. 2.

Let C denotes a cut., N_i is the ith node of cut C. All the leave nodes of the sub-tree with root node N_i share the same transformation. Therefore N_i here correspond to the "block" mentioned above while different cut corresponds to different block number, i.e. different model complexity. Cuts are determined by certain control strategy according to current amount of adaptation data. For example, in the beginning, when only a few data are provided, the cut may only contain the root node. Therefore a global transformation is implemented. As more data become available, the cut goes down the tree and thus the transformation follows a coarse-to-fine scheme. The major concerns of this method are (1) construction of model tree (2) control strategy.

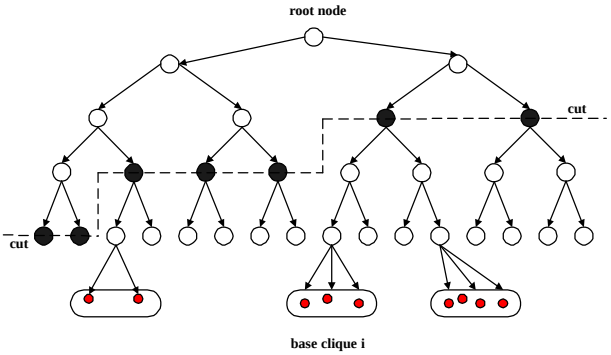

Figure 2: Construction of model tree

A. Construction of Model Tree

For convenience, we use binary tree in our approach. The tree is constructed bottom-to-top by K-mean algorithm. To alleviate the

constrain imposed by binary tree, we first assign those distributions which "look" very similar to each other in one clique, then construct a binary tree upon those cliques as illustrated in fig. 2. We choose the absolute value of correlation coefficient $|\rho|$ as distance measurement instead of divergence or Euclidean distance between Gaussian mean. Distance between Gaussian mean of ith and jth model d_{ij} is defined as:

$$d_{ij} = |\rho| = \frac{1}{N} \sum_{s=1}^{N} (\mu_{s,i} - \mu_{SI,i})^T (\mu_{s,j} - \mu_{SI,j})$$

N is the total number of training speakers, $\mu_{s,i}$ is the mean of jth vector Gaussian distribution of speaker S' SD model. $\mu_{SI,i}$ is the mean vector of ith Gaussian distribution of SI model.

We choose this distance measurement because, for different models to share the same transform, their trend of variation across speaker plays more important role than the model's spatial distance[5]. The model tree is constructed as follows,

1. Calculate distance d_{ij} between any two Gaussian distribution i and j as defined above.
2. Generate cliques in which all the distance between any two node is small than threshold *th_0*
3. Construct binary tree by k-mean algorithm.

The distance of two classes is defined on pairwise basis by averaging distances between all member of either class.

B. Control Strategy

Control strategy determines the current scale of transformation, in other word, degree of sharing. A good control strategy must be both effective and simple. By "effective" we mean that it can reflect the relationship between data amount and model complexity. By "simple" we mean it should be easy to implement. In [6], a control strategy based on information theory is put forward. Though with sound basis, it is a bit costly in computation.

In our proposed method, we used a simpler recursive control strategy based on sufficient stochastic, to be specific, count of state occurrence. Cut_finding goes in the following steps.

Cut_Finding(k)
0. k = Root Node
1. if k is a NULL node
2. return
3. if k is leave node and statistic(k) > Model_Complexity
4. append k
5. return
6. if(statistic(LChild(k)) < Model_Complexity or statistic(RChild(k)) < Model_Complexity)
7. Cut_Finding(LChild(k))
8. Cut_Finding(RChild(k))
9. else
10. append k
11. return

C. Implementation of Incremental Transformation

The input speech is first analyzed and represented by a set of sequential feature vector O $\{o_1,...,o_t,...,o_T\}$. O is segmented using viterbi algorithm with current SI model. Next, sufficient statistics for base cliques are collected. After cut-finding, current transformation blocks are determined. Gaussian components of the same block are modified accordingly by the block-wised transformation we proposed before [3].

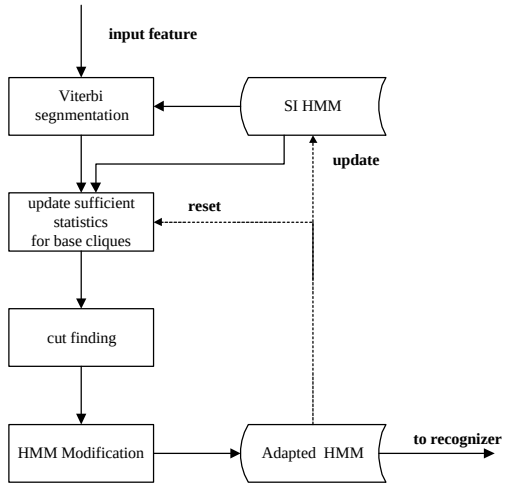

Figure 3: Implementation of Incremental Adaptation

When a lot of data from the new speaker is processed, to alleviate the effect of old data and enable the system to continuously track the variations of current speech, the system should be updated at certain interval. In figure 3, the data line represents such "forgetting" mechanism, that is, the old SI HMM is replaced by the adapted HMM and sufficient statistics is reset to zero.

4. EXPERIMENT

Experiments were carried out to evaluate the validity and effectiveness of the proposed methods on our Chinese large vocabulary continuous speech recognition system (Chinese LVCSR system). In our system, the speech was first digitized at 16kHz. The sampled speech was then partitioned into 20ms frames with 10ms-overlap in-between. The feature used was a vector of 45 component with mel-scaled cepstrals and frame energy, their first and second order derivatives.

4.1. Baseline

For convenience, we use a scaled down version of our Chinese CSR system as baseline, which is trained with 10 hours of continuous speech. It had 856 context-dependent states with one Gaussian distribution per state. In recognition stage, tests are carried out in ***pure acoustic level with no grammatical constrain*** in order that we can tell the effect of the proposed method and of course this is easier to implement compared to the integrated system. We use character right rate (CRR) as the performance measurement of our system (In Chinese, one character corresponds to one syllable). In adaptation test, we use 50 sentences from each out-set speaker.

Character right rate is defined as follows,

$$\mathrm{CRR} = \frac{\text{number of correctly recognized characters}}{\text{total number of characters}} \times 100\%$$

4.2. Test for vocal tract length normalization

VTLN is implemented at both training and recognition stage. To get the warping factor of each speaker, we use 3 sentences from each speaker in unsupervised mode. Table I shows that VTLN gives a 5.4% of error reduction of character error rate.

Speaker / System	Baseline(GD)	VTLN
M25	60.31	61.87
M33	65.22	67.18
M45	65.45	65.84
F00	63.66	67.24
F28	58.44	60.81
F34	65.39	67.60
Avg.	**63.08**	**65.09**

Table 1: Test for VTLN

4.3. Result for tree-based transformation

To test the effectiveness of tree-based transformation, the base system was adapted by 50 random chosen sentences in

supervised and unsupervised mode. In unsupervised mode, final recognition results are used as label. From table II, we can see incremental adaptation is superior to static adaptation, which give approximately 25% reduction of error rate. Table II also manifested that the unsupervised incremental adaptation method can still improve the recognition rate, which indicated that our proposed method is robust to error in label.

Speaker / System	Baseline (GD)	Static Adap.	Incremental Adap.	
			Sup.	Unsup.
M25	60.31 (86.17)*	66.24	69.97 (93.37)	68.90 (92.31)
M33	65.22 (86.26)	71.49	71.50 (91.60)	69.38 (89.31)
M45	65.45 (88.56)	71.33	67.08 (91.05)	67.24 (90.27)
F00	63.66 (88.33)	78.68	79.14 (95.25)	76.65 (94.71)
F28	58.44 (85.50)	71.67	71.50 (91.35)	68.53 (89.82)
F34	65.39 (88.30)	74.22	75.06 (93.30)	72.43 (91.43)
Avg.	**63.08 (87.19)**	**72.27**	**72.38 (92.65)**	**70.52 (91.31)**

Table 2: Test for VTLN

* ( ) stands for the CRR of first five candidates

4.4. Asymptotic property of tree-based transformation

Figure 4 compares the recognition results obtained by tree-based transformation with those obtained in SI, MAP and global transformation recognition experiments. In global transformation, all the distributions share the same transformation. When adaptation data are small, it can improve the recognition rate sharply, but the performance satiates rapidly when more data became available. On the other hand, MAP has good asymptotic property but relatively slow adaptation rate. From figure 4, we can see that the proposed tree-based method has the advantage of both techniques.

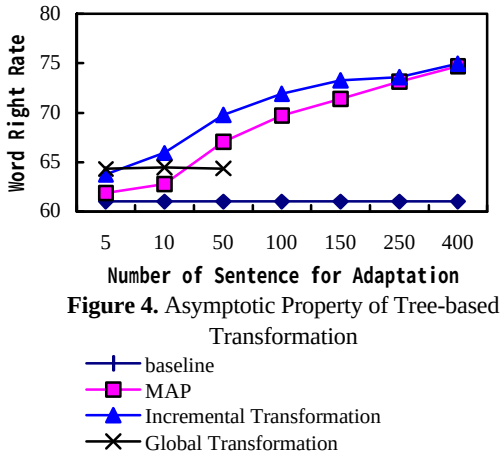

Figure 4. Asymptotic Property of Tree-based Transformation

4.5. Combined incremental learning

Lastly, we test the effectiveness of our combined scheme. For comparison, we also list the performance of VTLN and incremental adaptation when used separately. All the tests are in unsupervised mode.

Speaker / System	Baseline (GD)	VTLN	Incremental Adaptation	Combined Scheme
M25	60.31 (86.17)*	61.87	68.90 (92.31)	67.84 (91.90)
M33	65.22 (86.26)	68.36	69.38 (89.31)	70.23 (90.25)
M45	65.45 (88.56)	65.21	67.24 (90.27)	67.86 (90.97)
F00	63.66 (88.33)	65.84	76.65 (94.71)	77.82 (94.86)
F28	58.44 (85.50)	60.39	68.53 (89.82)	69.47 (90.67)
F34	65.39 (88.30)	67.09	72.43 (91.43)	74.81 (92.45)
Avg.	**63.08 (87.19)**	**64.90**	**70.52 (91.31)**	**71.34 (91.85)**

Table 3: Test for Incremental Learning

5. SUMMARY

In this paper, we presented a framework for incremental adaptation. Two crucial techniques are investigated: VTLN and incremental transformation. Incremental adaptation is proved to be a flexible and effective way to remove the difference of an outlier speaker.

6. REFERENCES

1. Li Lee and Richard Rose, "A Frequency Warping Approach to Speaker Normalization", *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 1, pp. 49-60.
2. Rong Zheng and Zuoying Wang, "Speaker Adaptation: an Overview", *Chinese Journal of Electronics*, April 1998, pp. 122.
3. Fan Zhang and Zuoying Wang, "Speaker adaptation using hierachical transformation and Bayesian approach", *Proc. CJSLP'97*, pp.181~186.
4. Puming Zhang and Martin Westphal, "Speaker Normalization Based on Frequency Warping", *Proc. Int'l Conf. on Acoust., Speech and Signal Processing*, 1997, pp.I-704~I-707.
5. Ben M. Shahshahani, " A Markov random field approach to Bayesian speaker adaptation", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 2, Mar. 1997.
6. Koichi Shinoda and Takao Watanabe , "Speaker adaptation with autonomous model complexity control by MDL principle", *Proc. Int'l Conf. on Acoust., Speech and Signal Processing*,1996 , pp. II-717-720.