IRRELEVANT VARIABILITY NORMALIZATION IN LEARNING HMM STATE TYING FROM DATA BASED ON PHONETIC DECISION-TREE

Qiang Huo and Bin Ma

Department of Computer Science and Information Systems,
The University of Hong Kong, Pokfulam Road, Hong Kong (e-mail: qhuo@csis.hku.hk)

ABSTRACT

We propose to apply the concept of *irrelevant variability normalization* to the general problem of *learning structure from data*. Because of the problems of a *diversified training data set* and/or possible *acoustic mismatches* between training and testing conditions, the *structure* learned from the training data by using a maximum likelihood training method will not necessarily generalize well on mismatched tasks. We apply the above concept to the structural learning problem of phonetic decision-tree based hidden Markov model (HMM) state tying. We present a new method that integrates a linear-transformation based normalization mechanism into the decision-tree construction process to make the learned structure have a better modeling capability and generalizability. The viability and efficacy of the proposed method are confirmed in a series of experiments for continuous speech recognition of Mandarin Chinese.

1. INTRODUCTION

Currently, in building an HMM-based automatic speech recognition (ASR) system, a common practice is

- to collect a large amount of speech data in the *target application domain* for the case of building a task-dependent (TDEP) system, or in several possible application domains for the case of task-independent (TIND) system; then
- to determine empirically and/or learn in a partially automatic way the *structure* of the HMMs for the adopted basic speech units from the collected training speech data; and
- finally with the so learned *structure*, to estimate the HMM parameters from training data.

Here *structure* refers to the model topology, parameter sharing schemes, the complexity of models, etc. Under the currently adopted statistical ASR framework, the training data have to be representative enough to achieve the required generalizability thus supporting somehow the performance robustness. In many real applications, this will lead to a training set with *diversified variabilities* which might be caused by different speakers, speaking styles, accents, dialects, transducers, transmission channels, environments, etc. People noticed that the common practice of maximum likelihood (ML) training of HMMs on this *pooled* training data set will usually result in a set of *diffused* models which might not work optimally for any particular testing conditions, even on the same task.

There are many ways to obtain a set of *sharp* and thus hopefully more *discriminative* models. One way is to adopt the *discriminative training* with the hope of maximizing the separation between models of speech units so that the robustness of a recognizer can be improved. Another way is to embed some *normalization mechanisms* of the *irrelevant variabilities* into the conventional ML HMM training to make better use of the diversified training data with the hope of obtaining a set of sharper and more appropriate *generic* speech models. These generic models are expected to work reasonably well only for those speech data similar to the generic training speech data, but not for others. The so-called *speaker adaptive training* (SAT) originally proposed by the BBN researchers is such an example [1]. The efficacy of the above two strategies is highly dependent on the nature and size of the training data as well as the task itself. It is now well-known that the performance of an ASR system often degrades drastically whenever there exist some acoustic mismatches between the training and testing conditions. So, another strategy, namely, developing efficient feature/model compensation and adaptation techniques, has been one of the most active research areas to address the above problems [8]. For example, if the application scenario allows, by using the above set of generic models as seed models and performing a fast adaptation on demand for individual application and/or condition, a better performance or robustness can be achieved [1].

For applications in which the target vocabulary is either not specified *a priori* or changing frequently from one task to another, a training procedure aiming at task-independent (TIND) subword modeling becomes necessary. The goal of TIND training is to create a set of subword models that is capable of handling new tasks without the need of collecting new training materials, capable of generating a context rich set of subword units and models to handle new vocabularies and capable of producing a reasonable performance even for unseen tasks [7]. Among many issues, *unit selection and modeling* is a very important subject. Research in TIND training was pioneered in [4] under the notion of *vocabulary learning*. Instead of fixing a set of context-dependent (CD) phone models at training time, one can use the context information of the target vocabulary and task grammar and select a new set of phone models to train for each new task. This was shown to produce a good performance by incorporating such task-specific context information. Authors in [7] suggest to use the *complete sets* of right and left CD units as the basic phone sets. If required, the triphone model can be *composed* from the existing sets of single-context and context-independent (CI) phone models. Good performance was also obtained on several tasks. In order to achieve high performance in large vocabulary speech recognition, detailed double-context dependent speech units such as triphones are usually needed for modeling both intraword and interword linguistic phenomena. To flexibly control the required model complexity in terms of the total number of states for the intended recognition task and the available amount of training data, phonetic decision-tree based state-tying technique is usually adopted (e.g., [2, 3, 5, 6, 10]). Decision-tree also provides a convenient way for the synthesis of models for contexts which do not occur in the training data. It thus has the potential of being a good TIND modeling tool. However, because of the abovementioned problems of *diversified training data set* and the possible *acoustic mismatch* between training and testing conditions, the *structure* learned from the training data by using the current ML training method will not necessarily generalize well on the mismatched task.

In this paper, we propose to apply the concept of *irrelevant variability normalization* to the general problem of *learning structure from data*. As a first step, we choose to apply this concept to the structural learning problem of phonetic decision-tree based HMM state tying which is adopted in many large vocabulary ASR systems.

2. METHODOLOGY AND ALGORITHM

In the general problem of modeling and learning, two concepts are very important:

- to model what we intend to model, thus
- to learn what we intend to learn.

Let's take the phonetic decision-tree based state tying in [10] as an example. In this case, the phonetic decision tree is used to recursively partition a set of states into subsets by answering some linguistically-motivated questions about phonetic (here triphone) context in which each state occurs. States reaching the same leaf node are judged to be similar and thus tied. So, the variability caused by co-articulation in different contexts is the primary source we intend to model by using the decision-tree. Other variabilities are irrelevant in this regard. However, the approximate ML learning of the *state-tying* in (e.g., [10] and other decision-tree based approaches) lacks a mechanism of normalizing the effects of those irrelevant variabilities in decision-tree construction. Consequently, if the training data are very diversified, the learned *state-tying* might reflect insufficiently the co-articulation effects, and/or worse, reflect more the effects of the other variabilities. This might lead to a poor generalization ability.

Based on the above considerations, we integrate the linear-transformation based normalization technique in [1] into the decision-tree construction process in [10] to derive a new procedure outlined as follows:

Step 1: Partition the training data set $\mathcal{X}$ into R different subsets denoted as $\{\mathcal{X}^{(r)}\}_{r=1,\cdots,R}$ with each being "homogeneous" according to some criterions. We refer each set to one "condition".

Step 2: Train an initial set of *untied* triphone HMMs with a single Gaussian distribution per state as in [10].

Step 3: Using the above untied models as seed model, perform a *condition-normalized* ML training as in [1] to obtain

- A set of *generic* triphone HMMs denoted as Λ, of which $\{\mu_s, \Sigma_s\}$ denote the mean vector and covariance matrix respectively of the Gaussian distribution of the *untied state* s;
- A set of linear transformations $\mathcal{G} = \{G^{(r)}\}_{r=1,\cdots,R}$ with $G^{(r)} = (A^{(r)}, b^{(r)})$ being the linear transformation(s) of the mean vectors $\{\mu_s\}$ for r-th condition. Note that in the following experiments, a single linear transformation is used for each "condition". It is possible however to use multiple linear transformations. The extension of the related formulation is straightforward;
- The necessary *condition-normalized* statistics $\gamma_s^{(r)} = \sum_t \gamma_s^{(r)}(t)$ for decision-tree construction, where $\gamma_s^{(r)}(t) = \Pr(\mathbf{X}_t^{(r)} \in s | \mathcal{X}^{(r)}, G^{(r)}(\Lambda))$ with $\mathbf{X}_t^{(r)}$ being an observation feature vector from training set $\mathcal{X}^{(r)}$.

Step 4: Construct the decision tree by using the new method in which the *goodness-of-split* evaluation function is computed by using the *generic* HMM parameters and the *condition-normalized* statistics.

More specifically, let S denote a set of *untied states* for a node to be split and Q denote a set of binary questions about the context. A question $q \in Q$ will split S into two subsets denoted as $S_y(q)$ and $S_n(q)$ based on the outcome of question being "yes" or "no" respectively. Then the new *goodness-of-split* evaluation function is defined as

$$\begin{aligned} m(q,S) &= \frac{1}{2}\gamma_S \ln|\Sigma_S| - \frac{1}{2}\gamma_{S_y(q)} \ln|\Sigma_{S_y(q)}| \\ & \quad -\frac{1}{2}\gamma_{S_n(q)} \ln|\Sigma_{S_n(q)}| \qquad (1) \end{aligned}$$

where

$$\Sigma_S = \frac{\sum_{s\in S}\sum_{r=1}^{R}\gamma_s^{(r)}(\Sigma_s + \mu_s^{(r)} \cdot \mu_s^{(r)'})}{\sum_{s\in S}\sum_{r=1}^{R}\gamma_s^{(r)}} - $$

$$(\frac{\sum_{s\in S}\sum_{r=1}^{R}\gamma_s^{(r)} \cdot \mu_s^{(r)}}{\sum_{s\in S}\sum_{r=1}^{R}\gamma_s^{(r)}}) \cdot (\frac{\sum_{s\in S}\sum_{r=1}^{R}\gamma_s^{(r)} \cdot \mu_s^{(r)}}{\sum_{s\in S}\sum_{r=1}^{R}\gamma_s^{(r)}})'$$

$$\begin{aligned} \mu_s^{(r)} &= A^{(r)} \cdot \mu_s + b^{(r)} \\ \gamma_S &= \sum_{s\in S}\sum_{r=1}^{R}\gamma_s^{(r)} \ . \end{aligned}$$

Other terms $\Sigma_{S_y(q)}$, $\Sigma_{S_n(q)}$, $\gamma_{S_y(q)}$, $\gamma_{S_n(q)}$ can be calculated in the same way. Note that $\gamma_S = \gamma_{S_y(q)} + \gamma_{S_n(q)}$. Based on the above-defined *goodness-of-split* evaluation function $m(q, S)$, a node $S*$ will be chosen to split by using a question $q*$ if

$$(S*, q*) = \arg\max_{S,q \in Q} m(q, S)$$

and $m(S*, q*)$, $\gamma_{S*_y(q*)}$, $\gamma_{S*_n(q*)}$ exceed their associated thresholds which are selected empirically.

After the construction of the above decision-tree, by pooling all the training data together, we can train a tied-state HMM system with an increased number of mixture components per state as in [10]. Furthermore, we can also train a condition-normalized *generic* tied-state HMM system as in [1]. Using above different model sets as seed models, model adaptation can be performed for new conditions and/or applications.

3. EXPERIMENTS

3.1. Experimental Setup

To examine the viability and the efficacy of the proposed method, a series of experiments for continuous speech recognition of Putonghua (Mandarin Chinese) are performed. The database we used is the HKU96 Putonghua Corpus developed in our laboratory [11]. The HKU96 corpus consists of a total of 20 native Putonghua speakers, 10 females and 10 males, each speaking: (1) all Putonghua syllables in all tones at least once, (2) 11 words of 2 to 4 syllables, (3) 16 digit strings of 4 to 7 digits, (4) 3 sentences of 7 rhymed syllables with /a/, /i/ and /u/ endings respectively, and (5) hundreds of sentences with verbalized punctuation from newspaper text. All speech recording were made in a quite room with a single National Cardioid Dynamic Microphone. Speech was digitized using a Sound Blaster 16 ASP A/D card plugged into a 486 PC at 16-bit accuracy and with a sampling rate of 16KHz. We used 18224 sentences (about 23 hours of raw speech) from 18 speakers (9 females and 9 males) for training; 200 sentences from 2 speakers (1 female and 1 male, 100 sentences randomly chosen from each speaker) for testing; and the remaining sentences from those testing speakers for adaptation.

Input speech was initially pre-emphasized $(1 - 0.97z^{-1})$ and grouped into frames of 25ms with a frame shift of 10ms. For each frame, a Hamming window was applied followed by the computation of 12 MFCC's. The 39-dimensional feature vector used in this study consists of 12 MFCC's and log-scaled energy normalized by the peak of the individual sentence, plus their first and second order derivatives. Sentence-based cepstral mean subtraction (CMS) is applied for acoustic normalization both in training and testing.

The adopted context-independent (CI) phone set consists of 37 phones plus silence. With this phone set definition, there are 8358 triphones in Putonghua. Among them, 5633 triphones are observed in our training data set, with only 4796 triphones each appearing at least 3 times. Each phone is modeled by a left-to-right three-emitting-state Gaussian-mixture continuous density HMM (CDHMM) without state skipping. Each state has 3 Gaussian mixture components with each component having a diagonal covariance matrix. A special three-state CDHMM is also used for silence modeling.

The recognition task is the recognition of 410 Putonghua *base syllables* disregarding tones. The recognition network enforces silence at the start and end of sentences. As for syllable language model, a uniform grammar with a syllable perplexity of 410 (i.e., each syllable can be followed by any of the 410 base syllables) is used. All the recognition experiments are performed with the search engine provided by HTK2.1 toolkit.

3.2. Effects of Normalization in Decision-Tree Construction

The baseline system is a speaker independent, cross-syllable-triphone, decision-tree-based tied-state system and is trained by using the HTK2.1 toolkit. 152 linguistic questions are used in decision-tree construction and the relevant thresholds for stopping criterion are adjusted to generate 3450 tied states. For this system, an averaged syllable accuracy of 73.7% over 2 testing speakers is achieved.

Considering the nature of HKU96 corpus, speaker difference is the main source of the irrelevant variabilities in decision-tree construction. So, we partition the training set into 18 subsets according to speaker identity (condition). For each condition, we use one affine transformation for normalization purpose of the mean vectors of the CDHMMs. We build a new decision-tree using the procedure described in Section 2. Then, a new recognition system is built which achieves a syllable accuracy of 74.8%. In comparison with the HTK baseline system, 4.2% error reduction is achieved. This is a quite encouraging result, because if we view the contextual variability caused by co-articulation as the main source we intend to model with the triphones, apart from speaker variability, there is no much other irrelevant variabilities existed in the speech data of HKU96 corpus. The benefit of normalization is expected to be bigger in a more realistic situation with a diversified training data set.

3.3. Effects of Structure-Normalization on Adaptive Modeling

By using the above two sets of models as seed models, we performed supervised speaker adaptation on two testing speakers by using the so-called batch-mode MLLR adaptation method in [9]. Two regression trees are first built for all of the Gaussian mixture components of the above two systems by using a divisive Gaussian distribution clustering method with a distortion measure being the symmetric divergence measure between two Gaussian distributions. In adaptation, different number of affine transformations are adaptively chosen based on the amount of available adaptation data. Figure 1 (a) shows the performance (syllable accuracy in %) comparison averaged over 2 testing speakers as a function of number of available adaptation sentences among two systems with above two different sets of seed models. The performance of new method is consistently better than that of HTK system. The benefit of the irrelevant variability normalization in decision-tree construction is also confirmed in adaptive modeling. We do have learned a better structure!

As we mentioned before, starting from above two sets of seed models, we further perform condition-normalized training (CNT) as described in [1], to obtain two sets of new models. Then the above MLLR-based adaptation experiments are repeated. Figure 1 (b) shows a similar performance comparison of two systems with two different sets of seed models. Once again, normalization in decision-tree construction proves to be helpful. The results show that the benefits of the normalization in structure learning and the normalization in generic model parameters learning can be combined to generate a set of the best generic models. For example, in the case of 200 adaptation sentences, the three systems with different, namely, HTK-based, new decision-tree based, and new decision-tree plus CNT based, seed models, achieve respectively the syllable accuracies of 80.4%, 81.4%, and 82.6%. This represents error reductions of 5.1% and 11.2% respectively of the later two systems from the first system.

4. CONCLUSION

In this paper, we propose a new concept of irrelevant variability normalization in learning structure from data. As a first step, we develop the technique to apply such a concept to the structure learning problem of decision-tree based HMM state tying. In a series of preliminary experiments, we show the benefits of the new method. The same concept can also be applied to other structure learning problem. As future works, we will perform experiments on larger scale and more diversified corpus. We will examine the effects of the proposed method on task-independent training and testing scenarios. More intelligent stopping criterion is another topic of future research. Finally, as a language specific issue, we will refine our question set for Putonghua (Mandarin Chinese) recognition.

REFERENCES

[1] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," *Proc. ICSLP-96*, 1996, pp.1137-1140.

[2] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo and M.A.Picheny, "Decision trees for phonological rules in continuous speech," *Proc. ICASSP-91*, 1991, pp.185-188.

[3] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan and M.A.Picheny, "Context dependent vector quantization for continuous speech recognition," *Proc. ICASSP-93*, 1993, pp.II-632-635.

[4] H.-W. Hon and K.-F. Lee, "Vocabulary learning and environmental normalization in vocabulary-independent speech recognition," *Proc. ICASSP-92*, 1992, pp.I-485-488.

[5] M.-Y. Hwang, X.-D. Huang and F.A. Alleva, "Predicting unseen triphones with senones," *IEEE Trans. SAP*, Vol.4, No.6, pp.412-419, 1996.

[6] A. Kannan, M. Ostendorf and J.R. Rohlicek, "Maximum likelihood clustering of Gaussians for speech recognition," *IEEE Trans. SAP*, Vol.2, No.3, pp.453-455, 1994.

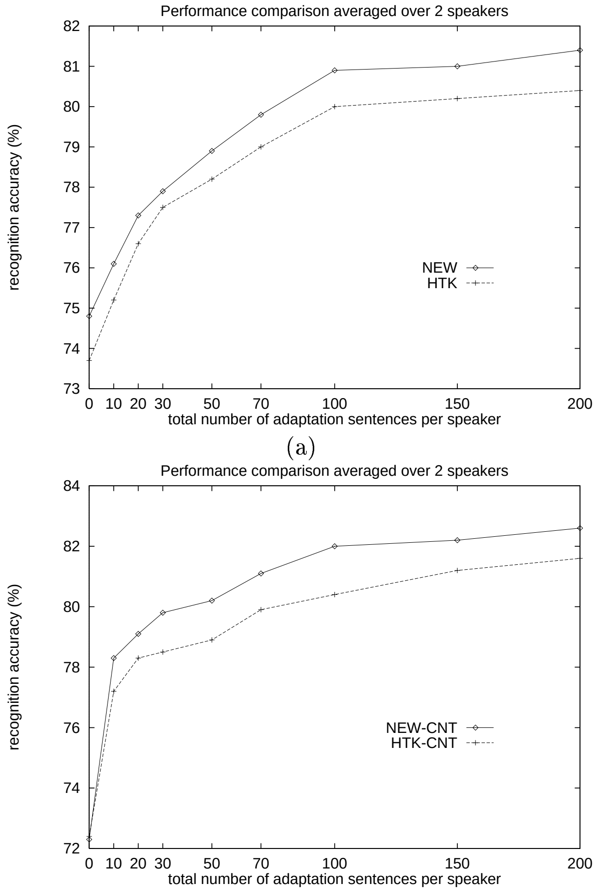
[7] C.-H. Lee, B.-H. Juang, W. Chou and J.J. Molina-Perez, "A study on task-independent subword selection and modeling for speech recognition," *Proc. ICSLP-96*, 1996, pp.1820-1823.

[8] C.-H. Lee, "On feature and model compensation approach to robust speech recognition," *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition For Unknown Communication Channels*, 1997, pp.45-54.

[9] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, Vol. 9, pp.171-185, 1995.

[10] S.J. Young, J.J. Odell and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," *Proc. ARPA Human Language Technology Workshop*, 1994, pp.307-312.

[11] Y.-Q. Zu, W.-X. Li, M.-C. Ho and C. Chan, "HKU96 – a Putonghua corpus (CD-ROM version)", Speech Lab., Dept. of Computer Science, Univ. of Hong Kong, 1996.



(a)

(b)

Figure 1: Performance (syllable accuracy in %) comparison averaged over 2 testing speakers as a function of number of available adaptation sentences per speaker: (a) HTK vs. New decision-tree construction with normalization, (b) HTK plus condition-normalized training (CNT) vs. New decision-tree plus CNT.