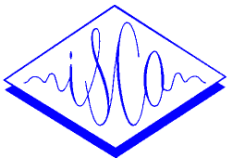
ISCA Archive
http://www.isca-speech.org/archive



ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

FLEXIBLE PARAMETER TYING FOR CONVERSATIONAL SPEECH RECOGNITION

Hua Yu and Alex Waibel

Interactive Systems Lab, Carnegie Mellon University, Pittsburgh, PA 15213
Email: hyu@cs.cmu.edu

ABSTRACT

Modeling pronunciation variation is key for recognizing conversational speech. Previous efforts on pronunciation modeling by modifying dictionaries only yielded marginal improvement. Due to complex interaction between dictionaries and acoustic models, we believe a pronunciation modeling scheme is plausible only when closely coupled with the underlying acoustic model. This paper explores the use of flexible parameter tying for pronunciation modeling. In particular, two new techniques are investigated: Gaussian tying and flexible tree clustering. We report a 1.3% absolute WER improvement over the traditional modeling framework on the Switchboard task.

1. INTRODUCTION

Modeling conversational speech is an important, yet difficult task for speech recognition. While conventional techniques work well on read, prepared speech, real life situations, such as meetings and telephone conversations, still pose great challenges for current technology.

One way to model rampant pronunciation variation in conversational speech is to add alternative pronunciations to a dictionary [1]. The improvement, however, has been quite limited. Recently, there has been a trend to model pronunciation implicitly. The reasoning is that the mapping from a phonetic string to the acoustic model sequence is a complex one, influenced by context clustering and state tying. Therefore, simply modifying the lexical representation of a word may not achieve the desired effect. Jurafsky et al. have argued that triphones can capture many kinds of pronunciation variations [2], including phone substitution and reduction. Hain questioned the use of pronunciation variants in a recent work called "single pronunciation dictionary" [3]. By systematically removing variants, he showed a slight gain over a state-of-the-art Switchboard system.

The studies above suggest that pronunciation modeling needs to be closely coupled with the underlying acoustic modeling. The reason is multi-fold. Firstly, the boundary between an acoustic model and a pronunciation model is not clearly cut. Both seek to model variations in speech. For certain weak forms of variation, it may be better to model implicitly at the acoustic model level, rather than introducing a variant in the lexicon. Secondly, a joint approach allows leveraging various acoustic modeling mechanisms for pronunciation modeling, which translates to greater expressiveness and modeling power. For example, a pronunciation variant, as a phone sequence, can always be translated into a state/model sequence, but the reverse is not true. A majority of state/model sequences cannot be represented as valid phone sequences. Recently, there has been several works in this direction, including state-level pronunciation model (SLPM) [4] and Hidden Model Sequence Model [3].

Parameter tying is an important aspect in acoustic modeling. In particular, decision tree based state tying [5] is key to context dependent systems. Tying is originally designed to achieve statistical robustness, but it also plays a crucial role in the mapping from symbolic (phoneme) level to state/model level. Hence it can be used for pronunciation modeling purposes, too. As an example, phone substitutions can be modeled by the tying of appropriate states.

In this paper, we explore two novel flexible parameter tying methods for pronunciation modeling: Gaussian tying and flexible tree clustering. While they seem to be purely acoustic modeling techniques, we emphasize that our primary interest is in how they may improve modeling conversational speech. We will introduce the two ideas first, then give detailed discussion together with experiment results on the Switchboard task.

2. GAUSSIAN TYING

Gaussian tying is motivated by state level pronunciation model (SLPM) [4]. If a baseform phone, say, AX, is alternately realized as the surface form IX in certain contexts, such as

```
AFFECTIONATE     AX F EH K SH AX N AX T
AFFECTIONATE(2)  AX F EH K SH AX N IX T
```

SLPM augments the mixture model of AX with all Gaussians from the mixture model of IX. Gaussian tying generalizes from SLPM in that it allows sharing of a selected subset of Gaussians from the IX model (rather than all of them), hence providing a finer level of control.

Gaussian tying is actually a general tying framework, covering all different families of tied mixture models. Let

$\mathcal{G}$, the Gaussian pool, be the complete set of Gaussians in a system. A model, m_i, is then defined by mixture weights over a subset of $\mathcal{G}$.

$$p(\cdot|m_i) = \sum_{j \in S_i} \pi_{ij} g_j(\cdot)$$

where $S_i \subset \mathcal{G}$ is the set of Gaussians used by m_i. Any acoustic modeling scheme can be completely specified by the weights matrix (Π_{ij}), where the ith row represents model m_i, the jth column the jth Gaussian in $\mathcal{G}$. Hence Gaussian tying covers: tied mixture, senones , genones, soft tying, etc.

Note most of the models allow only tying within the same phone and the same sub-state (begin/middle/end), corresponding to a very constrained form of the tying matrix. But this need not be the case. For example, neighboring states, such as the end state of one phone and the begin state of the next phone, tend to share many similarities.

In Section 4.1, we apply Gaussian tying on top of an existing state-tying framework to fix this deficiency. While gaussian tying allows more flexibility, it is not trivial to determine the exact form of tying. Here are some potential approaches:

- *similarity based tying*: Gaussians that are close in the model space are tied;
- *mapping baseform to surface form*: as with SLPM, we can augment the baseform model selectively with Gaussians from the surface form model;
- *error corrective tying*: the idea is similar to [6]. We monitor competition between models on the training data. In a perfect world, the correct model, as specified by transcripts, should achieve the best likelihood. If there is a strong competing model, we can augment the reference model with Gaussians from the competing model.

3. FLEXIBLE TREE CLUSTERING

Conventional decision tree based state tying allows parameter sharing at leaf nodes of a tree. Typically, one decision tree is grown for each sub-state (begin/middle/end) of each phone. With 50 phonemes in the phone set, 150 separate trees are built (Figure 1(a)). Parameter sharing is not possible across different phones or sub-states. Gaussian tying is one way to introduce more sharing. A potentially better approach is to build a single decision tree, from the very beginning, for all sub-states of all the phones (Figure 1(b)). In such a tree, any nodes can potentially be shared by multiple phones/sub-states (hence the name *flexible* tree clustering).

Other than improving parameter tying, flexible tree clustering may be beneficial in two important aspects:

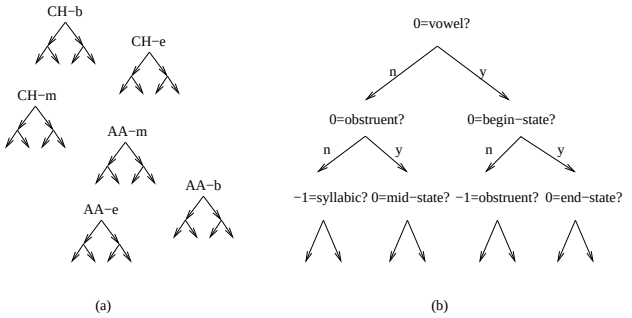


Fig. 1. (a) shows the traditional clustering approach: one tree per phone and sub-state. (b) shows the concept of flexible clustering using a single tree.

- dictionary over-specification and inconsistency: one can find all kinds of artifacts in a dictionary. Certain phones, such as DX (as in MATTER), may just be an allophonic variation of another phoneme. As it may not be marked consistently throughout the lexicon, data belonging to the same linguistic category might be splitted among several phone models. Cross-phone parameter sharing can help in this case.
- reduced phoneme differentiation in sloppy speech: people don't differentiate phonemes as much in sloppy speech as they do in read speech. This calls for an even greater amount of cross-phone parameter sharing in order to model them robustly.

4. EXPERIMENTS & ANALYSIS

Experiments are performed on the Switchboard (SWB) task. The test set is a 1 hour subset of the 2001 Hub5e evaluation set. The full training set includes 160 hours of SWB data and 17 hours of CallHome data. We typically use a 66 hours subset of the 160 hours SWB data for fast experimentation. The baseline system is developed using the Janus speech recognition toolkit [7]. The front-end uses vocal tract length normalization, cluster-based cepstral mean normalization, and a 11-frame context window for delta and double-delta. Linear discriminant analysis is applied to reduce feature dimensionality to 42, followed by maximum likelihood linear transform. We use a 15k vocabulary and a trigram language model trained on SWB and CallHome.

The baseline acoustic model uses a quinphone tree based, two level state tying scheme (described in [8], similar to soft-tying [9]): 24k distributions sharing 6k codebooks, with a total of 74k Gaussians. It has a WER of 34.4% [10]. All results reported in this paper are based on first-pass decoding, i.e. no adaptation or multi-stage processing.

4.1. Gaussian Tying

For similarity based tying, we tried three Gaussian distance measures:

- Euclidean distance between Gaussian means (variances are ignored)
- KL2 (symmetric Kullback-Leibler) distance:

$$KL2(A;B) = \frac{\sigma_A^2}{\sigma_B^2} + \frac{\sigma_B^2}{\sigma_A^2} + (\mu_A - \mu_B)^2(\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2})$$

- Likelihood loss: this measures how much we lose in likelihood if we choose to model samples as a single Gaussian rather than two.

With a chosen distance measure, greedy iterative bottom-up clustering is performed until a desired number of Gaussians are tied. One design issue concerns the size and quality of a cluster. Big clusters are likely to be problematic, as it causes too much smoothing and reduces discrimination between models. A simple solution is to impose a hard limit on the cluster size. We also tried complete linkage clustering rather than single linkage clustering to create "tight" clusters. In general, the improvement is quite small, despite extensive experimentation.

For error corrective tying, we consider two model sequences side by side: the "correct" model sequence obtained by viterbi alignment, and the most likely model/Gaussian sequence. If a particular Gaussian g of model m' "outvotes" a correct model m frequent enough, g is added to the mixture of m.

We conducted a cheating experiment to verify the concept. Model competition statistics is collected on the test set itself, using reference text for the correct model sequence. On our Broadcast News system, we are able to reduce WER from 19.1% to 15.1% (on 1998 Hub4e test set) by error-corrective tying. However, the gain doesn't hold when we repeat the same procedure on the training data.

The ineffectiveness of these methods might be blamed on their post-processing nature. It can be difficult to fix an existing sub-optimal tying scheme (as determined by decision trees). This leads us to flexible tree clustering.

4.2. Flexible Tree Clustering

Computation cost is the main difficulty for growing a single big tree. As the number of unique quinphones on the Switchboard task is around 600k, directly clustering on all of them is quite daunting. The traditional approach doesn't have this problem, since polyphones are divided naturally according to center phone and sub-state identities. For this reason, we conducted two experiments to investigate the effects of cross-phone tying and cross-substate tying separately.

4.2.1. Cross-Phone Clustering

We grow six triphone trees for cross-phone clustering: one for each of the begin/middle/end state of vowels and consonants. We could have built three big trees, without differentiating between vowels and consonants. But we expect little parameter sharing between vowels and consonants. Furthermore, separating them reduces computation.

Initial experiment gives a small, albeit significant improvement (from 34.4% to 33.9%). As the tree is grown in a purely data-driven fashion, one may wonder how much cross-phone sharing there actually is. It is possible that questions regarding center phones are highly important, therefore get asked early in the tree, resulting in a system which is no different from a phonetically tied system. We examined the six triphone trees, and found that 20% to 38% of the leaf nodes (out of a total of 24k) are indeed shared by multiple phones.

Motivated by Hain's work on single pronunciation dictionary (SPD) [3], we tried to reduce the number of pronunciation variants in the dictionary. The procedure to derive a new lexicon is even simpler than Hain's. First we count the frequency of pronunciation variants in the training data. Variants with a relative frequency of less than 20% are removed. For unobserved words, we keep only the baseform (which is more or less a random decision). Using this procedure, we reduced the dictionary from an average 2.2 variants per word to 1.1 variants per word. We are not using strictly single pronunciation, so that we can keep the most popular variants, while pruning away spurious ones. For example, the word A has two variants in the resulted dictionary:

```
A       AX
A(2)    EY
```

Simply retraining the baseline system using SPD gives a 0.3% improvement, which is comparable with Hain's results. More interestingly, cross-phone clustering responds quite well with SPD. As shown in Table 1, we achieve a 1.3% gain by cross-phone clustering on single pronunciation dictionary.

Dictionary	Cross-Phone Clustering	WER(%)
multi-pronunciation	no	34.4
	yes	33.9
single pronunciation	no	34.1
	yes	33.1

Table 1. Cross-Phone Clustering Experiment [1]

Note experiments in Table 1 are based on the 66 hours training set and triphone clustering. The gain holds when we switch to the full 180 hours training data and quinphone clustering. Due to high computation, we only compared two systems: one with multi-pronunciation lexicon and no cross-phone clustering, and the other with single-pronunciation lexicon and cross-phone clustering. WERs are 33.4% and 31.6%, respectively.

How to explain the interaction between cross-phone clustering and SPD? Why does cross-phone clustering help more with SPD (from 34.1% to 33.1%), comparing to from 34.4% to 33.9% with a multi-pronunciation lexicon? Let us consider the (unintended) side effects of pronunciation variants. When a variant replaces phone A by phone B, we are distributing to model B the data that was originally used to train model A, effectively allowing parameter sharing between phones. Therefore, even with no explicit cross-phone clustering, cross-phone parameter sharing already exists. However, those unintended sharing may be undesirable. This explains both why SPD works, and why cross-phone clustering doesn't help as much with the multi-pronunciation system as it does with SPD. Adding pronunciation variants also increases confusability in the lexicon, while SPD does not. In short, adding pronunciation variants is not as straightforward as it may seem. Changes to a dictionary should be coordinated closely with acoustic modeling.

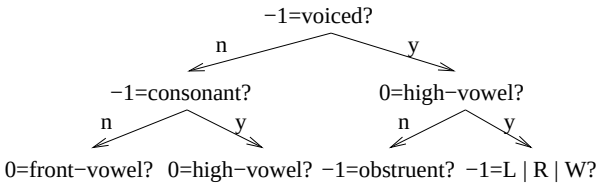


Fig. 2. Top part of Vowel-b tree (beginning state of vowels). "−1=" questions ask about the immediately left phone, "0=" questions ask about the center phone.

The top portion of the tree Vowel-b is shown in Figure 2. It is clear that questions about center phone identities are not necessarily preferred over contextual questions. Again, 20% to 40% of the leaf nodes are found to be shared by multiple phones. Consonants that are most frequently tied together are: DX and HH, L and W, N and NG. Vowels that are most frequently tied together are: AXR and ER, AE and EH, AH and AX.

4.2.2. Cross-Substate Clustering

In this experiment, we build one tree for each phone, which covers all three sub-states. Three new questions are added regarding sub-state identities. Contrary to our experience with cross-phone clustering, we find those three questions to be highly important. They are chosen in most cases as the top two questions. Hence the resulted tree is not any different from three separate trees as in traditional clustering.

5. CONCLUSION

This paper introduces two new methods: Gaussian tying and flexible tree clustering. While they are seemingly pure acoustic modeling techniques, we have shown how they relate to, and how they interact with modeling pronunciations in conversational speech. Flexible tree clustering gives a significant improvement over state of the art decision tree based tying. The fact that it works even better with a single pronunciation dictionary demonstrates the importance of tightly coupling pronunciation modeling and acoustic modeling.

6. REFERENCES

[1] W. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Pronunciation modelling using a hand-labelled corpus for conversational speech recognition," in *Proc. ICASSP*, Seattle, WA, USA, 1998, pp. 313–316.

[2] D. Jurafsky, W. Ward, J. Zhang, K. Herold, X. Yu, and S. Zhang, "What kind of pronunciation variation is hard for triphones to model?," in *Proc. ICASSP*, 2001.

[3] T. Hain, "Implicit pronunciation modelling in ASR," in *ISCA Pronunciation Modeling Workshop*, 2002.

[4] M. Saraclar, H. Nock, and S. Khudanpur, "Pronunciation modeling by sharing gaussian densities across phonetic models," *Computer Speech and Language*, vol. 14, no. 2, pp. 137–160, April 2000.

[5] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. ARPA HLT Workshop*, 1994.

[6] A. Nakamura, "Restructuring gaussian mixture density functions in speaker-independent acoustic models," in *Proc. ICASSP*, 1998.

[7] Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries, and Torsten Zeppenfeld, "The JanusRTk Switchboard/Callhome 1997 evaluation system," in *Proceedings of LVCSR Hub5-e Workshop*, 1997.

[8] Michael Finke and Ivica Rogina, "Wide context acoustic modeling in read vs. spontaneous speech," in *Proc. ICASSP*, 1997, pp. 1743–1746.

[9] X. Luo and F. Jelinek, "Probablistic classification of hmm states for large vocabulary continuous speech recognition," in *Proc. ICASSP*, 1999.

[10] H. Soltau, H. Yu, F. Metze, C. Fügen, Y. Pan, and S. Jou, "ISL meeting recognition," in *Rich Transcription Workshop*, Vienna, VA, 2002.