Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition

Hua Yu and Tanja Schultz

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

10.21437/Eurospeech.2003-563

Abstract

Modeling pronunciation variation is key for recognizing conversational speech. Rather than being limited to dictionary modeling, we argue that triphone clustering is an integral part of pronunciation modeling. We propose a new approach called *enhanced tree clustering*. This approach, in contrast to traditional decision tree based state tying, allows parameter sharing across phonemes. We show that accurate pronunciation modeling can be achieved through efficient parameter sharing in the acoustic model. Combined with a *single pronunciation dictionary*, a 1.8% absolute word error rate improvement is achieved on Switchboard, a large vocabulary conversational speech recognition task.

1. Introduction

Modeling conversational speech is a major challenge for current speech recognition research. Conversational speech is characterized by rampant pronunciation variations, where accurate pronunciation modeling can lead to high recognition performance. Traditionally, people have tried to add alternative pronunciations to the recognition dictionary. Despite extensive investigation, this has yielded only marginal improvement.

What is pronunciation modeling? Is pronunciation modeling a synonym for dictionary modeling? As shown in Figure 1, the boundary between pronunciation modeling and acoustic modeling is not clear. For a given word, the lexicon is first looked up to convert the word into a phoneme sequence, which is subsequently translated into a state sequence using a phonetic decision tree. The state sequence is ultimately used to align with the acoustic observation. Hence the phonetic decision tree, a traditionally acoustic modeling concept, also plays a key role in the mapping from symbolic (phoneme) level to model (state) level.

Subtle pronunciation variations may actually be better modeled implicitly at the acoustic model level, rather than modeled explicitly at the lexical level. For example, phoneme AX can be alternately realized as phoneme IX in certain words:

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

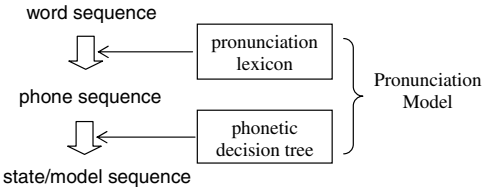

Figure 1: Pronunciation model as a mapping from symbolic level to model level

Instead of adding a variant in the dictionary, we can keep the dictionary unchanged, and either augment the mixture model of AX with Gaussians from the mixture model of IX (as in [1]), or simply tie these two models together.

This paper focuses on implicit pronunciation modeling using decision tree based tying. In Section 2, we will first examine the relative merits of different pronunciation modeling methods, namely, explicitly adding dictionary variants versus implicit modeling using decision tree based state tying. Section 3 introduces enhanced tree clustering that allows parameter sharing/tying across different phonemes. The new approach is evaluated on Switchboard, a large vocabulary speech recognition (LVCSR) task. Section 4 presents experiments and discussions. Related research work is reviewed in Section 5.

2. Explicit vs. Implicit Pronunciation Modeling

Early attempts at pronunciation modeling take the form of manually editing a lexicon. For example, to model the flapping of T in BETTER, an alternative lexical entry BETTER(2) is introduced:

```
BETTER      B EH T AXR
BETTER(2)   B EH DX AXR
```

Interestingly, for this kind of pronunciation variation, triphone modeling turns out to be better than dictionary editing. Triphone was originally introduced to model context dependency. As there is a large number of triphones in an LVCSR system, decision tree based state tying is widely used to cluster triphones [2]. This ensures sufficient training data for each model as well as better

generalization to unseen contexts. It turns out that many coarticulation rules, such as the flapping of T, can be well captured by triphone models in a purely data-driven fashion. This gives us another solution: leave the dictionary under-specified, and use automatic triphone clustering for pronunciation modeling[1]. The use of mixture model as the underlying distribution is also important, since we are using the same triphone model for both the flapped and unflapped version of T.

Compared to triphone clustering, manually editing a lexicon is both labor intensive and error prone. In the example of the flapping of T, one needs to be extremely careful to make sure that all relevant dictionary entries are modified, while nothing else is erroneously changed. This turns out to be quite difficult in reality. Whereas in the automatic solution, it is a lot easier to keep the dictionary simple and consistent.

One could also use automatic procedures to generate pronunciation variants, in order to avoid the pitfall of manual editing. There has been a lot of research in this area. So far, the improvement has been marginal. This could be attributed to several undesirable side effects of pronunciation variants:

- First, adding variants increases lexical confusability in a recognition dictionary;
- Second, if not done properly, adding variants increases model confusability during training. As illustrated in Figure 2, 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. In cases where the variant is spurious, model B will be contaminated with data belonging to A, making A and B more confusable.

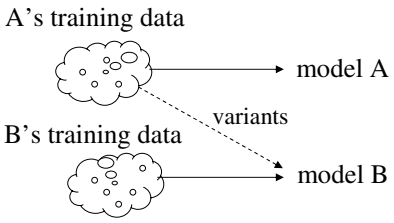

Figure 2: Model Contamination

In summary, due to the complex interaction between lexicon and acoustic modeling, adding pronunciation variants should be exercised with great care. In the next section, we introduce enhanced tree clustering for implicit pronunciation modeling.

3. Enhanced Tree Clustering

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 3(a)). Parameter sharing is not allowed across different phones or sub-states. With enhanced tree clustering, a single decision tree is grown for all sub-states of all the phones (Figure 3(b)). The clustering procedure starts with all polyphones at the root. Questions are asked regarding the identity of the center phone and its neighboring phones, plus the sub-state identity (begin/middle/end). At each node, the question that yields the highest information gain is chosen and the tree is split. This process is repeated until either the tree reaches a certain size or a minimum count threshold is crossed. Compared to the traditional multiple-tree approach, a single tree allows more flexible sharing of parameters. Any nodes can potentially be shared by multiple phones, as well as sub-states.

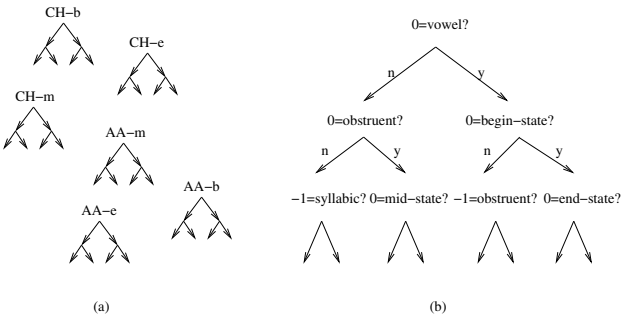

Figure 3: (a) shows the traditional clustering approach: one tree per phone and sub-state. (b) shows the concept of enhanced clustering using a single tree.

In sloppy speech, people don't differentiate phonemes as much as they do in read speech. Different phonemes tend to exhibit more similarity. Single tree clustering is well suited to capture these cross-phone parameter sharing, whereas the traditional approach does not allow such sharing.

Furthermore, sharing parameters across phones alleviates certain problems in a dictionary, namely, overspecification and inconsistencies.

Examples of these include the handling of T and DX, AX and IX as mentioned before. Some lexicons choose to differentiate them, while others do not. In lexicons that do, they are most often not marked consistently throughout. This is also referred to as the problem of choosing an optimal phoneme set. Unfortunately, such an optimal phoneme set does not exist. By allowing parameter sharing across phonemes, we no longer face this tough decision: if phonemes are indistinguishable under certain context, they will be allowed to share the same model; if they show sufficient differences under certain other context, they will be allowed to use different models.

Under the same argument, enhanced tree clustering is also a preferable choice for multilingual speech recogni-

[1]This actually resonates with single pronunciation dictionary [3]

tion or non-native speech recognition, where the phone set is not well defined and different phones show increased similarity.

4. Experiments and Discussions

4.1. Setup

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 hour subset of the 160 hours of SWB data for fast experimentation. The baseline system is developed using the Janus speech recognition toolkit [4]. The front-end uses vocal tract length normalization, cluster-based cepstral mean normalization, and an 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 [5], similar to soft-tying [6]): 24k distributions sharing 6k codebooks, with a total of 74k Gaussians. It has a word error rate (WER) of 34.4% [7]. Unless otherwise stated, all results reported here are based on first-pass decoding, i.e. no adaptation or multi-stage processing.

Computational cost is the main difficulty for growing a single big tree. As the number of unique quinphones on the Switchboard task is around 600k, direct 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. 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. The reason is: first, we expect little parameter sharing between vowels and consonants; furthermore, separating them reduces computation.

Initial experimentation 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 they are asked earlier 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.

4.3. Single Pronunciation Dictionary

Motivated by Hain's work on single pronunciation dictionaries (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 pronunciations, 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 using SPD with traditional clustering gives a 0.3% improvement, which is comparable to Hain's results. More interestingly, cross-phone clustering responds quite well with SPD. Overall, we achieve a 1.3% gain by cross-phone clustering on a single pronunciation dictionary (Table 1).

Dictionary	Clustering	WER(%)
multi-pronunciation	regular	34.4
	cross-phone	33.9
single pronunciation	regular	34.1
	cross-phone	33.1

Table 1: Cross-Phone Clustering Experiment

Note that experiments in Table 1 are based on the 66 hour training set and triphone clustering. The gain holds when we switch to the full 180 hour 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. WER improves from 33.4% to 31.6%, a 1.8% absolute gain.

4.4. Discussion

To explain why cross-phone clustering helps more with SPD (from 34.1% to 33.1%), than with a regular dictionary (from 34.4% to 33.9%), 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 (Figure 2), effectively allowing parameter sharing between phones. Hence, cross-phone parameter sharing exists even without explicit cross-phone clustering. But this sharing will only contaminate models and hurt

performance, if those variants are not carefully scrutinized. As discussed in Section 2, adding pronunciation variants is not as straightforward as it may seem. Changes to a dictionary should be coordinated closely with acoustic modeling.

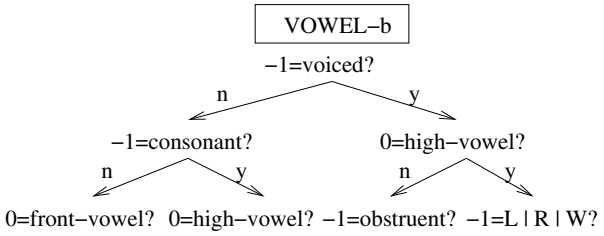

Figure 4: Top part of Vowel-b tree (beginning state of vowels). "−1=" questions ask about the immediately left phone, "0=" questions ask about the center phone.

The top portion of the tree for the beginning state of vowels is shown in Figure 4. 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.

4.5. 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. Related Work

Pronunciation modeling has received a lot of attention recently. To explain why simply modifying the lexicon does not work as expected, Jurafsky et al. argued that triphones can already capture many kinds of pronunciation variations [8], 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.

Recent studies focus on implicit pronunciation modeling by leveraging various acoustic modeling mechanisms. Saraclar et al. proposed a state level pronunciation model, which tries to add Gaussians from the surface form model to the baseform model. Pronunciation modeling at a deeper level allows greater expressive power and modeling resolution. For example, a pronunciation variant, as a phone sequence, can always be translated into a state/model sequence, but not vice versa. A majority of state/model sequences cannot be represented as valid phone sequences. To address this, Hain proposed the Hidden Model Sequence Model [9].

6. Conclusion

This paper shows that polyphone clustering is an integral part in pronunciation modeling. Enhanced tree clustering is proposed to allow efficient parameter sharing across phonemes. This effectively handles blurred phone identities commonly found in conversational speech, instead of making a hard decision in the lexicon. Combined with a single pronunciation dictionary, the new approach achieves a 1.8% WER reduction on the Switchboard task, over state of the art decision tree based state tying. We believe this approach also holds promise in other tasks, such as multilingual speech recognition and non-native speech recognition.

7. References

[1] 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.

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

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

[4] 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.

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

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

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

[8] 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.

[9] Thomas Hain, *Hidden Model Sequence Models for Automatic Speech Recognition*, Ph.D. thesis, Cambridge University, 2001.