PHONE TRANSITION ACOUSTIC MODELING: APPLICATION TO SPEAKER INDEPENDENT AND SPONTANEOUS SPEECH SYSTEMS

Jon P. Nedel, Rita Singh, and Richard M. Stern
Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{jnedel, rsingh, rms}@cs.cmu.edu, http://www.cs.cmu.edu/~robust/

ABSTRACT

HMM-based large vocabulary speech recognition systems usually have a very large number of statistical parameters. For better estimation, the number of parameters is reduced by sharing them across models. The parameter sharing is decided by regression trees which are built using phonetic classes designed either by a human expert or by data-driven methods. In situations where neither of these are reliable, it may be useful to have techniques for non-decision-tree based state tying which perform comparably to those based on traditional methods. In this paper we propose two methods for non-decision tree based parameter learning in HMM-based systems. In the first method (context-dependent state tying), we restructure acoustic models to explicitly capture the transitions between phones in continuous speech. In the second method (transition-based subword units), we redefine the basic sound units used to model speech to model transitions between sounds explicitly. Experiments show that context-dependent state tying is a viable option for large vocabulary systems. They also show that using transition-based subword units can improve performance on spontaneous speech.

1. INTRODUCTION

Large vocabulary continuous speech recognition systems use Hidden Markov models (HMMs) to characterize sound units or phones which are smaller than words. Phones change dramatically depending on the context in which they are found, and also depending on the speaking style and situation. The triphone, which links together a phone with its left and right contexts, is the generally accepted way to model the effects of phonetic context. However, the number of parameters in a complete triphone model is huge, and complete coverage of the triphonetic space is practically impossible, even with speech databases of 100 hours or more. Therefore, modern systems use linguistically-defined decision trees [1] to cluster the models so that the number of parameters is reduced to a manageable level and unseen triphone units are modeled effectively [2].

While decision trees provide an efficient way of distributing the parameters, they are at the same time dependent on the definition of phone classes which are used as linguistic questions to partition the trees at every stage. Regardless of whether these questions are manually or automatically designed, the process of using decision trees to tie parameters in these systems is quite time consuming and cumbersome. In systems that need to be rapidly deployed or trained, alternate effective and fast methods of state tying need to be devised. We address this problem in our paper and propose two methods of effective non-decision tree state tying: redefinition of HMM topologies to capture the effect of context on phones in continuous speech more effectively, and redefinition of subword units to explicitly focus on these contexts by capturing phone transitions rather than steady-state regions in speech.

Previously, other researchers have experimented with different subword units and state-tying techniques to focus acoustic modeling efforts on the speech signal as it transitions from one phone to the next. Tying HMM states based on context alone has been successful for speaker dependent systems [3-4]. Alternate subword units that explicitly model phone transitions, such as demiphones and diphones, have been shown to produce acoustic models that efficiently capture the coarticulation effects prevalent in natural speech [5-7].

In this paper we are further interested in the use of phone transition modeling for speaker independent recognition of broadcast news and spontaneous speech. In the following section we present our experimental framework in the context of which, in later sections, we explain our proposed methods of state tying.

2. EXPERIMENTAL FRAMEWORK

2.1 Databases

For the tests on large-vocabulary, speaker-independent systems, we selected data used from the 1997 and 1998 DARPA HUB4 Broadcast News Evaluation corpora [ref]. We used approximately 90 hours of training data and 3 hours of testing data from all the focus conditions in the broadcast news domain.

For the tests on spontaneous/conversational speech, we used the Multiple Register Speech Corpus (MULT_REG) which was collected at SRI and distributed by NIST. MULT_REG is a parallel corpus for comparison of spontaneous and read speech. Fifteen spontaneous conversations on assigned topics were recorded and carefully transcribed. The same speakers then returned to

re-read their conversations in dictation register. The speakers also read 40 Wall Street Journal sentences. Although the corpus was relatively small, we were able to use approximately 2 hours of read speech to train various acoustic models. These models were then tested on separate test sets (0.5 hours of speech each); one set contained read speech, and the other spontaneous speech.

2.2 Speech Recognizer and HMM Configuration

We used the CMU SPHINX-III speech recognizer for all experiments. The standard HMM topology to model each triphone unit was a 3-state left-to-right model with no state skipping. SPHINX-III supports fully-continuous, semi-continuous, and discrete HMMs. For the broadcast news corpus, there were sufficient data to train fully-continuous models. For the MULT_REG corpus, we used semi-continuous models (codebook size 256) due to the limited amount of data available.

3. USING GENERALIZED TRIPHONES WITH CONTEXT-TIED STATES

The use of decision trees to tie states, while still the most successful method for reducing the number of parameters and modeling unseen triphone units, is constrained by many factors. In most cases the questions used to segregate the data must be designed by a linguistic expert with extensive knowledge of the language to be modeled. The questions are also designed based on linguistic theories of phoneme "similarity", which may or may not reflect the behavior of real speech, especially when it is highly spontaneous and variable. In a few cases where the questions are automatically generated from the data, their generality depends on the amount and type of data available. In the following paragraphs we present an alternate and simple rule-based method of state tying which is based on phone-transitions.

3.1 HMM States that Model Transitions

For conceptual simplicity we present our method in the context of HMMs with three states and no allowable skips between the states. Two such models are depicted in Figure 1. In each model depicted,

1) The central state models the canonical or "steady-state" form of the phone, which should not be highly affected by the surrounding context.
2) The left state models part of the *transition* from the previous phone into the current phone. This is highly dependent on the left context.
3) The right state models part of the *transition* from the current phone into the next phone. This is highly dependent on the right context.

It is to be noted, however, that in the practical implementation of an HMM-based system there is nothing that explicitly forces the model to follow the simple conceptual partitioning of states that we have enumerated above.

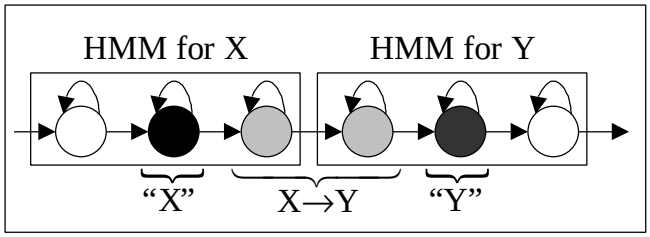

Figure 1. Motivation for 3-state HMM framework. Core states of phones model the canonical phone itself and are not highly affected by context. Remaining states model the context-dependent transition from phone to phone.

Thus, while it is reasonable to expect that the scheme presented above is generally true, it is likely that the representation of transitions and steady states is smeared over all the states of the HMM.

One way to impose this structure is through context-dependent state tying. For a given base phone, we tie together all of the central states. We also tie together the left states that have the same left context and the right states that have the same right context. Unlike decision tree approaches, this state tying does not rely on the data itself or on expert-designed linguistic questions. There are no inconsistencies in the resulting acoustic model structure. It has been previously shown that reasonably good acoustic models for speaker-dependent systems can be created using this approach [3-4].

The following experiments were designed to investigate the extension of this approach to speaker-independent systems and spontaneous-speech systems.

3.2 Context-Tied States and Broadcast News

We generated the tied model definition for the broadcast news corpus via the context-dependent state tying described above. Similar to [6], a minimum threshold N representing the number of observances was required to model a given state. In the case of broadcast news, we chose N=50. We used a decision tree built on context-independent models with one Gaussian per state to merge any transition states seen less than 50 times in the training corpus. (Note that because of the amount of data in the broadcast news corpus, only a few states that modeled rare phone transitions were replaced by this method. Removing the dependence on decision trees altogether by excluding the triphones that contain "undertrained" states resulted in only a slight degradation in performance.) Using this method, the resulting number of tied states was 5048, including the states used to model the CI phones. We built and trained a fully continuous acoustic model on the broadcast news training corpus.

For comparison, we also trained fully continuous models using decision trees to merge states in the standard manner.

We pruned the trees to ensure that our baseline model would have exactly the same number of tied states as the context-tied model (5048).

We decoded the test set using 4, 8, and 16 Gaussian per state models. The resulting word error rates (WER) are shown in Table 1. The standard decision tree-tied models slightly outperformed the context-tied models. However, these results confirm that context-tied modeling is a viable option for large-scale speaker independent systems.

	4gau/st	8gau/st	16gau/st
Standard	26.1%	23.8%	22.4%
Context-tied	28.5%	25.5%	23.7%

Table 1. Comparison of WER of standard-tied models with context-tied models on the Broadcast News corpus

3.3 Context-Tied States and MULT_REG

We generated the context-tied model definition for the read portion of the MULT_REG corpus. Due to the limited amount of training data, we merged all tied states seen less than N=10 times using a decision tree built from semi-continuous CI models. The resulting context-tied semi-continuous HMMs trained on the read portion of the MULT_REG corpus contained 2772 tied states. For comparison, we trained baseline semi-continuous models with states tied by decision trees. The baseline models contained 2500 distinct tied states. We tested these models on both the read and spontaneous test sets extracted from the MULT_REG corpus. The WER results are shown in Table 2. The resulting performance was consistent with the broadcast news results. Again, the context-tied models performed adequately and were slightly outperformed by the standard models.

	read test	spontaneous test
Standard	14.0%	38.5%
Context-tied	14.8%	41.8%

Table 2. WER comparison of standard-tied models with context-tied models on the MULT_REG corpus

Because the models were trained on read speech, the spontaneous results suffered from a mismatch in training and testing conditions. Standard models trained on the spontaneous corpus yielded a WER of 34.5% when evaluated on this particular spontaneous test set. In the next section, we discuss possible ways of doing transitional phone modeling to improve recognition when there is a mismatch in the level of spontaneity between the training and testing corpora.

4. PHONE TRANSITION SUBWORD UNITS AND SPONTANEOUS SPEECH

4.1 Transitional Subword Units

In 1997, Mariño *et al.* introduced the *demiphone* as an alternative subword unit for continuous speech recognition [5]. The phone is divided into two parts, the left part to handle the beginning of the phone and the left side coarticulation, and the right part to handle the end of the phone and the right side coarticulation. Using Mariño's labeling, the word "left" would be transcribed with demiphones as: F-l l+eh l-eh eh+f eh-f f+t f-t t+F. The symbol "F" is used to indicate a word boundary. The phone labels concatenated with "-" are left side demiphones, and those concatenated with "+" are right side demiphones. In [6], Mariño reported an improvement of recognition performance using demiphones as compared with standard triphone modeling.

In 1999, Dobrišek *et al.* directly used *diphone* units for recognition [7]. (The diphone unit is currently used in many speech synthesis systems.) A diphone is a representation which models the transitions that stretch from the "center" of one phone to the "center" of the next. Transcribed with diphones, the word "left" would be: l~eh eh~f f~t.

Similar to the above, we define *transition phone* units that explicitly model the transition from one phone to another. We use separate units to model the transition from one phone to the next, and we also use separate units to model the core of each phone. Using our notation, the word "left" is transcribed: <B>.l l l.eh eh eh.f f f.t t t.<E>. (<B> and <E> are used to mark begin and end word boundaries, respectively.) The phone pairs concatenated with "." represent the transition from one phone to the next. The isolated phone labels represent the canonical "center" of each phone. Having separate "steady" and transition states gives us the desired flexibility of model configuration required for our experimentation with spontaneous speech.

Table 3 summarizes the different proposed subword units that model coarticulation effects between phones. Again, the word "left" is transcribed using each representation.

demiphones:	F-l l+eh l-eh eh+f eh-f f+t f-t t+F
diphones:	l~eh eh~f f~t
transition phones:	<B>.l l l.eh eh eh.f f f.t t t.<E>

Table 3. Transcription of the word "left" using various proposed phone transition subword units.

4.2 Transition Phones and MULT_REG

For the experiments with transition phones, we expanded our base phone set to contain all of the transition phone units and trained context-independent semi-continuous HMMs on the read training data from the MULT_REG corpus. In this experiment, we used one state to model the

core of each phone and two states to model the transition from phone to phone. This setup provides the transition phone model equivalent to the standard 3-state triphone models used as a baseline for comparison. Again, due to limited training data size, we used decision trees built on standard CI triphone models to provide a model for unseen transition phone units. The results are shown in Table 4. The transition phone units performed better than standard CI models, but clearly they were unable to compete with standard CD models in recognizing either the read or spontaneous test sets.

	# tied states	read test	Spon test
Standard CI Triphones	165	23.3%	53.5%
Transition Phones (CI)	1357	17.8%	45.7%
Standard CD Triphones	2500	14.0%	38.5%

Table 4. WER results using phone transition units on MULT_REG. Comparison with standard CI and CD models. Number of tied states (proportional to the number of parameters) in each model is also shown.

4.3 Transition Phones and Train/Test Mismatch

The flexibility of the transition phone framework allowed us to experiment with the following hypothesis: Phones have a canonical "steady-state" or "target" feature value. When speech is carefully pronounced, the acoustic features transition gracefully from one target value to the next and almost always reach the steady state value. However, when speaking spontaneously, the canonical feature values are rarely attained. Often all that we see are transitions towards targets that are never quite achieved.

We used the read MULT_REG speech to train transition phone models with two states that model the "core" portion of each phone and two states that model the transitions from phone to phone. The core states are designed to capture the canonical feature values described above. When decoding, we used a modified decode dictionary that contained only the transition phone units and *not* the core units. For example, the decode dictionary transcription of "left" would be reduced to <B>.l l.eh eh.f f.t t.<E>. Dropping the core state from the decoding dictionary enabled us to test our hypothesis.

We decoded the read and spontaneous test sets using the transition phone models, both with and without the core states. The results are shown in Table 5. Dropping the core states from the model trained on read speech yielded a significant performance increase when decoding spontaneous speech, confirming our hypothesis. Also, dropping the core states resulted in a performance degradation on read speech, as expected.

	Read test	Spon test
Core included	19.8%	52.5%
Core dropped	23.4%	50.7%

Table 5. WER results using phone transition units on MULT_REG. Models were trained on read speech. Models were tested both with and without the core states that model "target" feature values observed in read speech.

5. DISCUSSION

We believe that the two methods described in Section 4 to be adequate for speaker-independent systems. These methods both simplify and speed up the state-tying process during system training. Since these methods do not explicitly require the knowledge of the sounds represented by the subword units, they are potentially of great use in rapid deployment of systems in foreign languages.

We also note that although standard triphone modeling outperformed the transition phone units in our experiments, there was an improvement in accuracy for spontaneous speech when the states modeling the core of each phone were not considered during decoding. This confirms the hypothesis that spontaneous speech can be thought of as a series of transitions towards canonical target values for each phone that are never completely reached.

6. ACKNOWLEDGEMENTS

The authors thank Dr. Bhiksha Raj for many fruitful discussions on the subject of this paper. This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

7. REFERENCES

[1] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Wadsworth Inc., 1984.

[2] M-Y Hwang et al., "Predicting Unseen Triphones with Senones", *IEEE Trans. on Speech Audio Processing, Vol.4 no 6*, 1996, pp. 412-419.

[3] L. C. Wood et al., "Improved vocabulary-independent sub-word modeling", *Proc. ICASSP91*, pp. 181-184.

[4] J. J-X Wu, L. Deng, J. Chan, "Modeling context-dependent phonetic units in a continuous speech recognition system for Mandarin Chinese", *Proc. ICSLP96*, pp. 2281-2284.

[5] J. B. Mariño, A. Nogueiras, A. Bonafante, "The demiphone: an efficient subword unit for continuous speech recognition", *Proc. EUROSPEECH97*, pp. 1215-1218.

[6] J. B. Mariño, P. Pachès-Leal, A. Nogueiras, "The demiphone versus the triphone in a decision-tree state-tying framework", *Proc. ICSLP98*, pp. 2463-2466.

[7] S. Dobrišek, F. Mihelic, N. Pavešic, "Acoustical modeling of phone transitions: biphones and diphones - what are the differences?", *Proc. EUROSPEECH99*, pp. 1307-1310.