Acoustic Model Clustering Based on Syllable Structure

Izhak Shafran & Mari Ostendorf

Department of Electrical Engineering
University of Washington, Seattle, USA

Abstract

Current speech recognition systems perform poorly on conversational speech as compared to read speech, arguably due to the large acoustic variability inherent in conversational speech. Our hypothesis is that there are systematic effects in local context, associated with syllabic structure, that are not being captured in the current acoustic models. Such variation may be modeled using a broader definition of context than in traditional systems which restrict context to be the neighboring phonemes. In this paper, we study the use of word- and syllable-level context conditioning in recognizing conversational speech. We describe a method to extend standard tree-based clustering to incorporate a large number of features, and we report results on the Switchboard task which indicate that syllable structure outperforms pentaphones and incurs less computational cost. It has been hypothesized that previous work in using syllable models for recognition of English was limited because of ignoring the phenomenon of re-syllabification (change of syllable structure at word boundaries), but our analysis shows that accounting for re-syllabification does not impact recognition performance.

1. Introduction

Recognizing conversational speech has proved to be more challenging than read speech for automatic speech recognition (ASR) systems. For the best systems reporting results on the 1999 DARPA Broadcast News benchmark tests, word error rates on the spontaneous speech portion of the test set (14-16%) were nearly double those on the baseline condition of planned recordings* (8-9%) (Pallett et al., 1999). Those sites that also participated in 2000 DARPA Conversational Speech benchmark tests, performed at error rates of roughly 30%. The degradation in performance may be due to many factors such as channel effects, variability in speaking rate and dialect of speakers, less careful pronunciation, loosely structured language,

*The baseline or "F0" condition is defined as planned speech from a native speaker, over a high bandwidth channel, with no background noise.

and the presence of disfluencies. (Weintraub et al., 1996) demonstrated that a large part of the degradation is related to acoustic variation associated with speaking style. In this study, spontaneous speech was recorded and then the transcript of the speech was both read and acted by the same speakers to control for effects related to speaker, channel and language. While spontaneous speech was recognized with a word error rate of 52.6%, the acted and read versions were recognized at 37.4% and 28.8% error rates, respectively. The degradation with increasingly casual speaking style was observed across telephone-band and wide band speech and under matched training and test conditions (Saraclar et al., 2000).

In many ASR systems, the acoustic variation of words are modeled at two levels - the pronunciation model which maps word sequences to phonemes, and the acoustic model which maps phoneme sequences to multivariate acoustic models. Work with simulated data which was produced using the acoustic models of speech, have pointed to pronunciation variability as a key problem in recognizing conversational speech (McAllister et al., 1998). However, the work on pronunciation modeling in terms of phoneme-level substitutions, deletions and insertions has so far only yielded small performance gains (Byrne et al., 1997; Riley et al., 1999). In a recent work, (Saraclar et al., 2000) showed that modeling pronunciation at state level and allowing the components of a Gaussian mixture model to be shared across alternate pronunciations is more beneficial than modeling pronunciation at phoneme level. Experiments by Hain (Hain and Woodland, 2000) demonstrate an advantage in using phonetic context to directly influence model sequence. Both these studies support the notion that there is a need to represent variation of a more gradient nature where higher-level context (beyond triphones) influences the acoustic models of the phonemes as well as the pronunciation of a word.

Conventionally, phone-level acoustic variation has been captured by conditioning the acoustic models for a phoneme on the context of neighboring phonemes in the hypothesized sequence. Typically, in large vocabulary ASR, phonemes with immediate neighbors (triphones) and possibly two neighbors (pentaphones) are used. Conditioning only on phonemic context does not capture the acoustic variation of conversational speech fully. Already, position of phoneme in word has been found to be useful in acoustic modeling. This conforms to observations about word-position effects in linguistic studies of different consonants with electropalatography (EPG) (Keating et al., 1999). The linguopalatal (tongue-palate) contact, which affect the strength and duration of sound produced, for word-initial consonant is significantly different from word-final.

Our hypothesis is that the syllable structure is also useful in modeling the variation not accounted by phoneme context. Consider the phoneme "t" (in the context "iy t er") in "beater", "beat Ernest" and "baby turned". Even though it is the same triphone, the articulation of phone "t" in the three contexts is distinctly different - in the first it is flapped, in the second it is an unreleased closure and in the third it is a closure plus a release. These differences are closely related to syllable structure, and correspond to ambisyllabic, syllable-final, and syllable-initial contexts, respectively. In a statistical study, Greenberg and Fosler found systematic variation with respect to constituent of syllable, namely onset, nucleus and coda (Greenberg, 1998; Fosler et al., 1999; Greenberg and Fosler-Lussier, 2000). They used a subset of Switchboard corpus for conversational speech and a standard corpus called TIMIT for read speech, annotated by linguists at word, phone and syllable level. The perceived phones were compared with the lexical expansion of the spoken words. Both corpora show that the onset of a syllable maintains its canonical identity at most times (85-91%) regardless of the speaking style, and more so in the presence of consonant clusters. In general, the nucleus is prone to substitution by a wide range of vowels. The coda is less often realized in canonical form in conversational speech (63%) than in read speech (81%). The coda is prone to deletion, but the absence of a coda does not impact canonical realization of nucleus. These results together support the notion of syllable-initial strengthening, which has been observed as a more gradient phenomena in EPG studies that also suggest that the amount of strengthening may be equal to that in word-initial position (Keating et al., 1999). An analysis of errors made by state-of-the-art systems on recognizing conversational speech (Greenberg and Chang, 2000) suggests that accurate recognition of syllable onsets is more important for word recognition than syllable codas. While categorical phonetic changes can be accommodated by a larger phone inventory and a good pronunciation model, as in the TIMIT labeling conventions, phone substitutions and deletions fail to capture more gradient aspects of variation such as strength of a stop release. Thus, it is not surprising that state-based pronunciation models outperform phone-based models. In this work, we look at a complementary approach to state-based pronunciation modeling, which is acoustic model context conditioning on high-level contexts, specifically syllable and word structure.

One way to model syllable structure is to use syllable-sized units rather than phones. For small vocabulary tasks, a few researchers have successfully used the syllable as a unit for acoustic modeling (Jones et al., 1997; Hamaker et al., 1998). Marino *et al.* (Marino et al., 1997), and Lleida *et al* (Lleida et al., 1991) split the syllable into demi-syllable units, and used them for acoustic modeling. However, these two approaches lack the ability to effectively model syllables that are rarely or not seen in the training data. To overcome this deficiency, triphones were used in addition to frequent syllables in (Doddington et al., 1997) at the 1996 workshop on speech recognition at JHU. This approach in conjunction with improvement in temporal structure of the acoustic model gave a 2% reduction in error rate on a test set that was defined at the workshop. A major part of this improvement came from modeling syllables in monosyllabic words separately from other instances of the same syllable. A disproportionate number of errors were found to be in words recognized by the triphone, and may be due to poor sharing of parameters between the triphones and syllables.

The focus of our work is on learning contextual variation directly in the acoustic model using both word and syllable level information, since they seemed promising in both pronunciation models and previous acoustic studies mentioned above. In contrast to modeling the syllable explicitly as a unit, we use a tree-based clustering mechanism to allow sharing of parameters across all contexts for robust estimation. To tackle the problems that arise in using a large number of contextual features, we have extended the decision tree based clustering to use multiple stages of clustering.

The paper is organized as follows. In section 2, we present a brief review of tree-based acoustic model clustering, followed by our use of syllable features in them, and then outline issues related to re-syllabification. The details of the multiple stage clustering approach is presented in section 3. Experimental results on the use of syllable structure, a re-syllabification model, and multistage clustering are reported in section 4, using the Switchboard corpus of conversational speech. Finally, section 5 concludes and discusses future work.

2. Clustering with Syllable Features

Since we use decision trees extensively in this work, a brief review of tree based clustering is provided, which is followed by a discussion of related work and description of our use of syllable features.

2.1. Tree-based clustering

For large vocabulary ASR systems, decision tree distribution clustering is used to map the large number of possible triphone (or pentaphone) contexts into a smaller set of distributions that can be robustly estimated (Young et al., 1994). This technique is particularly attractive for parameter tying, as it allows mapping of any sub-word unit that is not seen in the training data to a cluster made up of acoustically similar units. In training, all the context specific phonemes observed in the training data are pooled at the root node of the tree. A set of predefined questions, typically about phonetic context (e.g. "Is the left phoneme a vowel?"), is used to define candidate binary partitions of a node in the tree. Assuming that all the data in a partition share a common Gaussian, the question corresponding to the partition that maximizes the likelihood of the data in a node is chosen as a candidate for the next split. From amongst these candidates, the node with the best likelihood gain from using 2 versus 1 Gaussian is split. The best partitions of the new clusters resulting from this split are added to the list of the candidate splits, and thus the tree is grown until some stopping criterion is met (*e.g.*, limit on the number of leaves or terminal nodes). After the tree is designed, more complex Gaussian mixture distributions can be estimated to model the

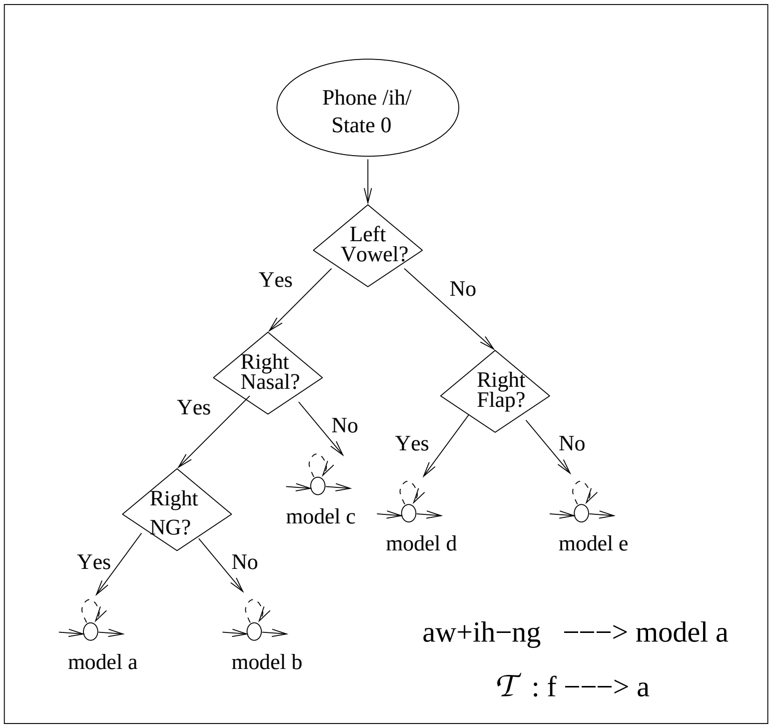

Figure 1: Decision tree maps context-specific phonemes to acoustic models

data in the leaves using Estimation-Maximization algorithm (Rabiner and Juang, 1993). In building a word model for decoding, a particular context specific phoneme is dropped down the tree and is guided by the linguistic questions at the branches. As illustrated in Figure 1, the acoustic model of the leaf that it lands in, is associated with the phoneme. Typically, a fixed topology of 3-5 five states is associated with each phoneme, and separate trees are used for each state of a phoneme.

In most ASR systems, decision tree questions are based on hand-specified phonetic classes (*e.g.* grouped by manner and/or place of articulation) on the neighboring phonemes. By incorporating a symbolic description of phonemes in the lexicon such as stress, position of phone in word, and syllable structure, it is possible to capture new phenomena with decision tree clustering, such as a tendency to reduce unstressed vowels and to more strongly release a stop consonant in word onset position. The phonemes in the training data are marked with these symbols. Then, the tagged models are estimated and clustered, just as for triphones, except that the decision tree must choose among questions about these tags as well as those defined in terms of phonetic context. Clustering with information other than phonetic neighbors is sometimes referred to as *tagged clustering*.

2.2. Related work on tagged clustering

The idea of using word- and syllable-level features in a decision tree framework for conversational speech is supported by a study conducted at a summer workshop in JHU (Ostendorf et al., 1996). Clustering a subset of standard training data for conversational speech with triphones that were coded with these features, it was found that questions regarding them were asked early, i.e., near the top of the tree. This may lead to finding better equivalence classes, and thus improve acoustic models. To make this study possible, the computational cost was reduced by ignoring the triphones that span word boundaries, i.e. evaluating performance with a word-internal triphone system. Thus, the usefulness of these features was not demonstrated conclusively.

The use of word position (initial, medial, final) as a context-conditioning feature has been shown to be useful in several studies, for both conversational speech (Finke, 1997; Gunawardana, 1998) and read speech (Reichl and Chou, 1999), and is used in many research systems. The use of syllable position alone has not so far proved to be useful (Paul, 1997; Gunawardana, 1998), though Paul reports a small gain when syllable position is used in combination with lexical stress tags. Paul's results for lexical stress are also mixed, with gains depending on the dictionary used. The mixed results on read speech could possibly be due to a variety of reasons, including inadequate levels for coding features or the sensitivity to the alignment of the training data used. Hence, we chose to re-evaluate the use of syllable vs. word position in clustering for conversational speech.

Tagged clustering studies have also looked at other features. Word type (function vs. content) (Lee and O'Shaughnessy, 1997) was found to be useful in experiments recognizing read speech, and a preliminary study with prosody show potentials of improving acoustics models for conversational speech (Shafran and Ostendorf, 2000). In our work, we will restrict our experimental study to word and syllable features, but the development of the multi-stage clustering approach makes possible the use of a greater number of features in general, which might include these and other features.

2.3. Use of richer syllable features

In this work, we used a rich set of symbols to represent syllable structure, which includes consonant cluster and ambisyllabicity. The lexicon and syllable coding system used in this work was developed at 1996 JHU workshop, and later extended for new words. The lexical expansion of words are coded at phoneme level as illustrated in table 1 (e.g. "arc aa:0:3:2:1 r:1:3:4:1 k:2:3:5:1"). Note that the position of the phone in the syllable distinguishes between onset consonants which are and are not in clusters, and marks consonants as onset even if they are not syllable initial, unlike previous work on syllable position. Also, unlike previous work, stress is marked using three levels (as in Pronlex and most dictionaries): primary, secondary and unstressed. Interpreting primary stress as the default position of the strong syllable of a word and secondary stress as a potential position for the strong syllable of a word, we labeled monosyllabic content and function words as having primary and secondary stress, respectively, so that all syllables in the dictionary were marked with one of the three levels.

Table 1: Coding of word- and syllable-features in the dictionary.

Digit	Phone position in word	Syllable position in word	Phone position in syllable	Stress
0	first	first	onset initial	stress-less
1	middle	middle	onset other	primary
2	last	last	nucleus	secondary
3	only	only	coda only	
4			coda initial	
5			coda other	
6			ambisyllabic	

The state-level segmentation of the training data (from a triphone system) for each phoneme was encoded with the word- and syllable-level features from the corresponding lexical expansion of the word in the lexicon. Initial acoustic models for each context-specific phoneme were estimated from these alignments. These models were then clustered using decision tree and questions about the word- and syllable-level features were used in addition to the standard contextual questions about neighboring phonemes. Subsequently, the estimate of the distribution associated with each cluster was refined using EM iterations.

Syllabification may vary systematically at word boundaries, depending on the neighboring word. For example, "just a" may be syllabified as "[jh ah s][t ax]" instead of "[jh ah s t][ax]", "choirs and" as "[k w ay r][z ax n]" instead of "[k w ay r z][ax n]" and "it's an" as "[ih t][s ax]" instead of "[ih t s][ax]", where the latter forms are obtained by stringing together syllables of each word. A complete description of the process of re-syllabification in English is relatively complex. However, a majority of cases occur at words that begin with vowels, and the process of re-syllabification moves the syllable boundary amongst the consonants to maximize the slope of sonority in onset-nucleus demi-syllable, and minimizes it in nucleus-coda demi-syllable (Sonority Dispersion Rule) (Clements, 1990; Kenstowicz, 1994). Since a small number of rules captures a large fraction of the cases of re-syllabification, it is possible to incorporate re-syllabification into the acoustic model of a word by allowing alternate pronunciations. The specific method used in our work is as follows:

- Candidates for re-syllabification include only open-vowel syllables that are word initial and do not follow a pause or a vowel (based on forced alignments in training and N-best hypotheses in testing).

- If the syllable is preceded by a single consonant, mark that consonant as optionally ambisyllabic.
- If the syllable is preceded by more than one consonant, apply the Sonority Dispersion Rule to obtain the alternate syllable boundary.

To study the effect of re-syllabification, we performed a series of experiments, using the rules mentioned above, as further described in section 4.

3. Multi-stage Clustering

There are a few limitations in using standard decision tree design techniques for clustering phonemes when they are coded with a large number of features. The number of elementary coded-phonemes increases drastically with the number of features, increasing the memory requirements for storing sufficient statistics of all possible coded phones. In addition, the large number of partitions that need to be tested to use these features raises the computational cost of clustering. Furthermore, phonemes in infrequent contexts, which constitute a large fraction of the phonemes, are estimated poorly and the partitions learned may not represent general trends in speech. These problems have restricted previous work on tagged clustering. For example, in (Ostendorf et al., 1996), experimental costs were reduced by restricting the use of syllable boundary and stress only to word-internal triphones, discarding triphones that spanned word boundaries. In other work, cross-word context is used but only with simple tag sets such as word position (begin, middle, end). We address these problems by introducing a new tree design technique based on multi-stage clustering.

Our approach to reduce the storage and computational costs for clustering is based on dividing the task into multiple stages. The decision tree can be viewed as a function, $\mathcal{T}$, that maps a feature vector, $\mathbf{f}$, consisting of contextual information to an index, a, of an acoustic model, thus $\mathcal{T} : \mathbf{f} \rightarrow a$. As illustrated in Figure 2, for two-stage clustering, we group the contextual information into two feature vectors $\mathbf{f}_1$ and $\mathbf{f}_2$, optionally allowing some common components between them. In the first stage, the training data is annotated only with the values of vector $\mathbf{f}_1$. Using the annotated data, we grow a decision tree, $\mathcal{T}_1$, which maps the different values of $\mathbf{f}_1$ to the index of its leaves b, thus $\mathcal{T}_1 : \mathbf{f}_1 \rightarrow b$. In the second stage, the training data is annotated with $\mathbf{f}_2$ along with the value of b which is obtained by dropping its context $\mathbf{f}_1$ down the tree $\mathcal{T}_1$. Using the newly annotated training data, a new decision tree, $\mathcal{T}_2$ is grown that maps $[b\ \mathbf{f}_2]$ to the index of acoustic models as represented by the leaves of $\mathcal{T}_2$, thus $\mathcal{T}_2 : [b\ \mathbf{f}_2] \rightarrow a$.

In current decision tree clustering for speech recognition, questions about features are defined by hand and are linguistically motivated. This is straightforward for the features in $\mathbf{f}_1$ and $\mathbf{f}_2$, but not for the index b. Allowing all possible partitions of b is impractical since there are $2^{|b|}$ binary partitions, and to use the features in the first stage adequately, the

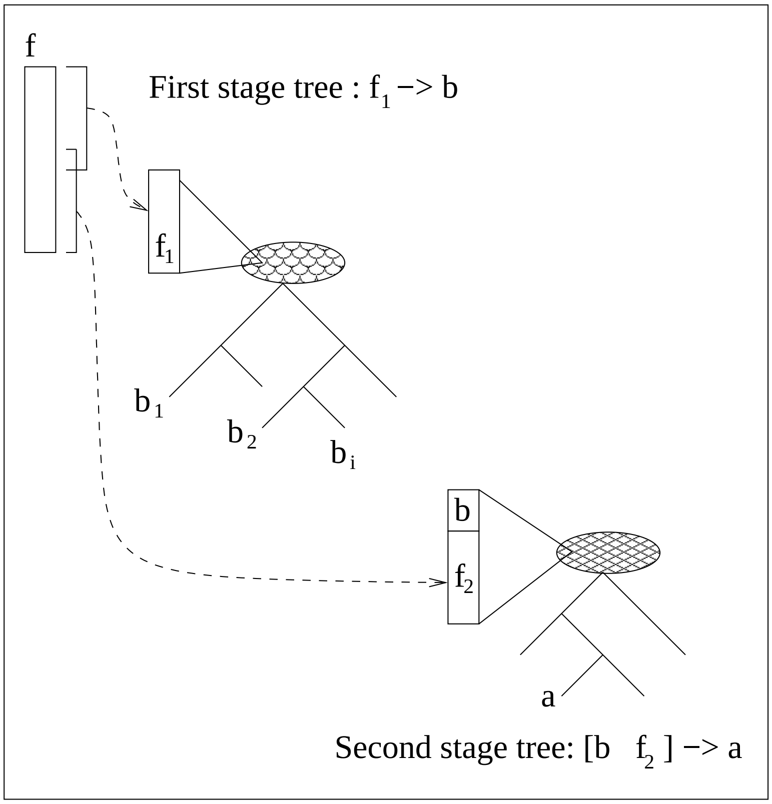

Figure 2: Multi-stage clustering illustrated here with two stages.

size of $\mathcal{T}_1$ needs to be large. To solve this problem we define questions that test whether a node b belongs to a subtree of the first tree $\mathcal{T}_1$ or not. Such questions are equivalent to compound questions which are obtained by performing an "and" operation on a set of binary questions about the features in $\mathbf{f}_1$. Defining questions on subtrees permits the decision tree to test a large number of partitions, and is more efficient than allowing all partitions.

Once the second stage tree $\mathcal{T}_2$ is grown, the questions on subtrees in $\mathcal{T}_1$ are replaced with the equivalent compound questions to obtain a single tree. Note that, in principle, there is no limit to the number of stages, but this work considers only two.

The multi-stage clustering techniques helps ameliorate the problem of sparse data by reducing the number of coded units for which sufficient statistics need to be estimated, since only a subset of the features are used at each clustering stage. The root node at every stage has all the data available to it or, in this case, all the data associated with the particular state and the phone. The number of elementary units that need to be clustered in stage i depends on the features $\mathbf{f}_i$ used in that stage and, if $i > 1$, the number of leaves of the preceding tree $\mathcal{T}_{i-1}$. Both of these factors can be controlled to reduce the effects of data fragmentation, essentially by trading off the potential for more directly modeling interaction between features (with a large dimension $\mathbf{f}_i$) with the robustness (and computational) advantages of a low dimension feature set $\mathbf{f}_i$. Note that robust estimation of statistics of elementary units also benefits from the general principle of increasing system complexity incrementally. In particular, we use phone alignments from our best triphone system, rather than bootstrapping from monophone models, as shown to be important in (Gunawardana, 1998).

The storage and computational cost of the multi-stage clustering depends on various factors. The number of sufficient statistics that need to be clustered in the two stages is determined by the number of components used in $\mathbf{f}_1$ and $\mathbf{f}_2$, and the size of $\mathcal{T}_1$, as mentioned above. The number of sufficient statistics is limited by the diversity of the data, and also depends on how uniformly the training data is divided into the clusters (how balanced the tree is). If a maximal tree is grown for $\mathcal{T}_1$, then the multi-stage clustering will be computationally more expensive than clustering a single tree. The size of $\mathcal{T}_1$ may be set using a constraint on total number of leaves, or saturation of likelihood increase. To reduce the number of partitions that are tested in the second stage, we selected only a subset of subtrees (the largest) from the first stage during clustering at the top of the second stage, where the largest proportion of the computation occurs in training decision trees.

4. Experiments, Results & Observations

4.1. Experimental framework

4.1.1. Speech corpus

We used the Switchboard and Callhome corpora, which together provide a collection of about 140 hours of spontaneous telephone conversations between pairs of callers in American English (Godfrey et al., 1992). The 1998 NIST Hub-5 development test set, consisting of about 12.5k words in approximately 30 minutes of speech from 28 speakers, is used for testing the performance of the recognizer using the standard criterion of word error rate ($= (C - I)/R$ where the number of words correct is C, inserted is I and the total number of words in the reference transcript is R).

4.1.2. Recognition system

The speech data is preprocessed to produce a 14 dimensional vocal tract length normalized MFCC vector sequence augmented with its first order derivatives, at a rate of 100 vectors per second. This serves as acoustic input to the recognition system (Zavaliagkos et al., 1998). A standard HMM topology with 5 states and skips is used to model the acoustic units, with a single full covariance Gaussian as the observation distribution for each state. Allophones of each phoneme and state are clustered to produce 10000 means and 2000 covariances which may be shared in a subtree.

To reduce experimentation time, we restricted our experiments to a lattice re-scoring decoding paradigm and did not adapt the models to the speaker being tested. The lattice was derived from the 100 best hypotheses provided by BBN†; the error rate of randomly selected hypotheses in the lattice is 55.4%. A 35k vocabulary was used to generate these hypotheses. To further reduce the search time, hypothesized word times were used to constrain the lattice search space in re-scoring, i.e. the re-scored times were forced to occur within a window of $\pm M$ frames ($M = 40$) of the hypotheses. The language model used in all the experiments was a part-of-speech smoothed trigram trained with Broadcast news data as well as the Switchboard and Callhome data (Iyer and Ostendorf, 1997). The acoustic models for all the recognition experiments described below were trained from clustered triphone state alignments with same fixed number of parameters.

4.1.3. Lexicon

For this study, we use the syllabified dictionary and the coding system that was developed at the 1996 summer workshop at John Hopkins University (JHU-WS-Lexicon, 1996). A brief review of the lexicon is given here; further details can be found in (Ostendorf et al., 1996). The stress markings for the multisyllabic words in this lexicon are taken from Pronlex dictionary; monosyllabic words were marked with either primary or secondary stress depending on whether or not the word was a content word, as described in Section 2.3. Syllabification for this lexicon was performed automatically using Fisher's implementation of Kahn's principles for English syllabification (Fisher, 1996). The syllabification is performed by assigning as many consonants to syllable onsets as possible (maximal onset rule) where permitted onsets were predefined. Among the possible syllabifications of a word, the most casual variant was selected to represent the nature of conversational speech. In this process, the morpheme boundaries were not taken into consideration. However, the use of casual variants of syllabification mitigates the associated syllabification errors, since many of the consonants at the boundary were labeled as ambisyllabic, rather than with the wrong syllable. To syllabify foreign words, an augmented list of permitted onsets was applied on those words that initially failed to parse.

4.2. Testing syllable features

We developed gender-dependent systems using information about word and syllable structure, as mentioned in section 2.3. During acoustic model training, the decision trees were allowed to ask questions about syllable and word information of the center and the immediate neighboring

†It is often useful in rescoring to combine scores from different acoustic models, but since our focus was on understanding the behavior of the syllable features, the BBN acoustic model scores were not used in the results reported here.

phonemes, in addition to the questions about triphone context. The recognition performance of these systems were compared with triphone and pentaphone systems with same number of model parameters. In all the systems described below, the clustered triphones were trained using single stage of clustering from the same base triphone alignment and then reestimated with a few passes of EM training. The results are summarized in the table below.

Table 2: Word error rates of systems with different features used in clustering.

System	
a) Triphone	44.56 %
b) Pentaphone	44.37 %
c) Triphone+Word-posn.	44.31 %
d) Triphone+Word-posn.+Syllable-feats.	44.05 %

While the difference in performance of the triphone and pentaphone systems is not statistically significant[‡], the system based on word- and syllable-features is significantly better than triphone system, according to NIST statistical significance tests (.05 Wilcoxon Signed Rank Test on speaker word error Rate (%); .01 McNemar Test on sentence error). This result has also been confirmed in subsequent experiments with a Gaussian mixture system on the same task. Contrary to other reported results, this gain is not simply due to use of word position features, which accounts for roughly half of the improvement from the triphone system to the system using syllable features.

The computational cost for decoding and training were also significantly lower for the system with word- and syllable-features. In training, the pentaphone system required testing 350 potential partitions for clustering up to 2.5M acoustic units, while the system with syllable features required testing only 200 potential partitions for 700k acoustic units. In decoding, the system with syllable features was 20% faster. In addition, unlike pentaphones, which incur extra computational expense and software flexibility to span the two forward contexts, the coded-triphones only looks ahead as much as a single phone and could be incorporated in a standard first pass triphone decoder.

4.3. Observations on use of syllable features

To study how the syllable features were used, we analyzed the questions that were chosen in the decision tree for clustering acoustic units. The number of questions asked about a feature in the decision trees and the percentage of total data affected by a feature can be used to understand

[‡]Other systems showing improved performance with pentaphones appear to have increased numbers of parameters in the pentaphone system, whereas here the number is constrained to be roughly the same.

how useful a feature has been in training acoustic models. To make a fair comparison across the columns, the number of partitions tested (or degrees of freedom) for each feature also needs to be taken into account.

Table 3: Word- and syllable-feature usage in decision trees trained on the Switchboard and Callhome corpora (about 140 hours of speech).

Feature	Questions in the tree	Data affected	Degrees of freedom
Triphone	66 %	76 %	180
Phone in Word	10 %	9 %	6
Syllable in Word	11 %	4 %	6
Phone in Syllable	3 %	6 %	21
Stress	11 %	5 %	4

Even though questions about phone identity allows 5 times as many degrees of freedom in partitioning the training data as all other features combined, the decision trees chose the latter about one third of the time. Among all the features, the fewest questions are asked about position of the phone in the syllable. However, it affected disproportionately a larger amount of data. This feature is likely to be used higher up in the tree, suggesting that questions about it generalizes over a large fraction of the data. The lexical stress and position of the syllable in the word is least useful, as it affects the least amount of data. It was also observed that the questions about the position of the center and the right phone in syllable is significantly more important (4-6 times the fraction of data affected) than that of left phone. Whether the center phone was in a monosyllabic word was among the top questions about the position of syllable in word, as one might expect from the gains observed in modeling monosyllabic word (Doddington et al., 1997), but the amount of data affected was not high so it did not stand out as a particularly important feature. In general, the pattern of usage of the features across gender is similar.

Interestingly, even though the pentaphone models had a higher likelihood on the training data, the syllable system had a better likelihood and word error rate on the test data, indicating its ability to generalize better.

4.4. Impact of re-syllabification

First, we studied the effect of re-syllabification on the training data. The method described in Section 2.3 was applied on a single phoneme path to generate alternate syllabifications across word boundaries for all vowel-initial words in the training set. This generated an additional (potential) 13% coded phonemes. Assuming that coded phonemes in re-syllabified paths exhibit the same acoustic characteristic as their word internal counterpart, we used the acoustic models generated in section 4.2 and let the decoder choose the best path. Of the alternate coded phonemes, the decoder chose only about 7% of them, which is less than 1% of the total labels in the training data. Thus, we do not expect re-syllabification to greatly impact the acoustic models.

The impact on test data was evaluated using N-best re-scoring. After expanding each of the 100 hypotheses into a lattice with alternate re-syllabification paths, we let the decoder choose the best path using the acoustic models developed in section 4.2. The word error rate did not show any improvements. This may be explained simply by the small number of tokens impacted by the change, but it may also be due to the fact that re-syllabification tends to occur in high frequency word pairs, and often the language model score adequately compensates for any loss in acoustic match.

4.5. Verification of multi-stage clustering

To evaluate the effectiveness of multi-stage clustering we trained gender-specific pentaphone systems using standard single stage clustering and two stage clustering. The systems were trained from a base triphone alignment with one pass of Viterbi training and a few passes of EM. In the first stage of the second system, we clustered the data into 1000 clusters for each of the five states, using the second phone neighbors as features. In the second stage, we clustered the data using the leaf indices of $\mathcal{T}_1$ along with the triphone context to obtain the final models.

Table 4: Word error rates of systems trained with one vs. two stages of clustering.

System	
a) Pentaphone: 1 stage	44.37 %
b) Pentaphone: 2 stage	44.39 %

The two-stage pentaphone system for both genders performed as well as the one-stage systems. Thus the result shows that incorporating features in multiple stages is a viable method for using a large number of features in acoustic modeling.

The memory used in clustering is directly proportional to the number of unique contexts to be clustered. In the single-stage pentaphone system, we had about 2M unique contexts, whereas in two stage system, we had only about 78K in the first stage and less than 0.5M in the second stage, thus reducing the memory requirement at any time by a factor of 4. This could be reduced further by shrinking the size of the first tree. The computational cost of two-stage clustering in this case is half that of single-stage clustering.

5. Conclusions

We have shown a small but consistent improvement in using syllable structure in addition to word position in a large vocabulary recognition task.

This, is in contrast to other reported results, and may be due to our use of high quality state alignments and a more detailed syllable coding system. The results also suggest that the syllable features generalize better than long span (pentaphone) phonetic context. Perhaps more importantly, the system using syllable features has lower training and decoding computational costs than a pentaphone system of equivalent size. In addition, our studies show that alternate paths of re-syllabification predicted by general linguistic rules do not provide any improvement in recognition performance. To take full advantage of syllable features, we conjecture that temporal variation must also be modeled. For example, the fixed-state topology could be replaced with a context-specific topology. This may be carried out within a decision tree framework such as (Eide, 1999).

We have developed a multi-stage clustering system that enables the use of a large number of features in clustering. Multi-stage clustering addresses the general issue of unreliable estimates of infrequent context as well as the higher computational cost incurred in clustering them. Speech recognition experiments show that the approach performs as well as a single stage of clustering with significantly reduced computational costs. The features explored in this work included phonetic context and syllable structure, but they could easily be expanded to include other features such as speaking rate (quantized into finite levels), word type (function vs. content word), and prosodic constituent structure. In addition, the work could be extended to incorporate data-driven techniques for organizing features into hierarchies for use in the different stages. In the work described here, the feature groups were chosen heuristically based on linguistic intuitions. Alternatively, statistics from their previous use, such as the total likelihood gained from the use of a feature or the fraction of data affected by it, could be used to determine the groups.

References

L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique in the statistical analysis of probabilistic functions of finite state Markov chain," in *Ann. Math. Stat.*, vol. 41, pp. 164–171, 1970.

B. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters and G. Zavaliagkos, "Pronunciation modeling for conversational speech recognition: A status report from WS97", in *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings (ASRU)*, Santa Barbara, CA, pp. 26–33, 1997.

G. N. Clements, "The role of the sonority cycle in core syllabification," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech, John Kingston and Mary E. Beckman (eds)*, pp. 283–333, Cambridge University Press, 1990.

L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of vowel-initial syllables as a function of prosodic structure," in *Journal of phonetics*, vol. 24, pp. 423–444, 1996.

G. Doddington, A. Corrada, and B. Wheatley et al., "Syllable based speech processing," in *Proc. 1997 summer workshop on speech recognition*, http://www.clsp.jhu.edu/ws97/syllable, 1997.

E. Eide, "Automatic modeling of pronunciation variation," in *Proc. European Conference on Speech Communication and Technology*, vol. 1, pp. 451–454, 1999.

E. Fosler-Lussier, S. Greenberg, and N. Morgan, "Incorporating contextual phonetics into automatic speech recognition," in *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 611–614, 1999.

W. Fisher, "A C implementation of Daniel Kahn's theory of English syllable structure," in *ftp://jaguar.ncsl.nist.gov/pub/tsylb2-1.1.tar.Z*, 1996.

M. Finke, and I. Rogina, "Wide context acoustic modeling in read vs. spontaneous speech", in *Proc. Int'l Conf. on Acoustic, Speech and Signal Processing*, Munich, 1997.

S. Greenberg, "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation," in *Proc. Workshop Modeling Pronunciation Variation in Automatic Speech Recognition*, pp 47–56, 1998.

S. Greenberg and E. Fosler-Lussier, "The uninvited guest: Information's role in guiding the production of spontaneous speech," in *Proceedings of the Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modeling, Kloster Seeon, Germany*, 2000.

S. Greenberg and S. Chang, "Linguistic dissection of Switchboard-corpus automatic speech recognition systems," in *Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millenium*, 2000.

J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. Int'l Conf. on Acoustic, Speech and Signal Processing*, vol. 1, pp. 517–520, 1992.

A. Gunawardana, "Using stress and syllabification features in state clustering for switchboard acoustic models," in *Unpublished Technical Report*, 1998.

T. Hain and P. C. Woodland, "Modelling sub-phone insertions and deletions in continuous speech recognition," in *Proc. Int'l Conf. on Speech and Language Processing*, vol. 4, pp. 172–176, 2000.

J. Hamaker, A. Ganapathiraju, J. Picone, and J. Godfrey, "Advances in alphadigit recognition using syllables," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp 421–424, 1998.

R. Iyer and M. Ostendorf, "Transforming out-of-domain estimates to improve in-domain language models," in *Proc. European Conference on Speech Communication and Technology*, vol. 4, pp. 1975–1978, 1997.

I. Shafran and M. Ostendorf, "Use of higher level linguistic structure in acoustic modeling for speech recognition", in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing*, vol. 2, pp.1021–1024, 2000.

M. Kenstowicz, "The syllable and syllabification," in *Phonology in Generative Grammar*, pp. 250–309, Blackwell Publishers, 1994.

JHU-WS-Lexicon *http://www.clsp.jhu.edu/ws96/ws96_workshop.html*, 1996.

R. J. Jones, S. Downey, and J. J. Mason, "Continuous speech recognition using syllables," in *Proc. European Conference on Speech Communication and Technology*, vol. 3, pp. 1171–1174, 1997.

P. Keating, R. Wright, and J. Zhang, "Word-level asymmetries in consonant articulation," in *UCLA Working papers in phonetics No. 97*, 1999.

C. Z. Lee and D. O'Shaughnessy, "Clustering beyond phoneme contexts for speech recognition," in *Proc. European Conference on Speech Communication and Technology*, vol. 1, pp. 19–22, 1997.

E. Lleida, J. B. Marino, C. Nadeu, and J. Salavedra, "Demisyllable-based HMM spotting for continuous speech recognition," in *Proc. Int't Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp. 709–712, 1991.

J. B. Marino, A. Nogueiras, and A. Bonafonte, "The demiphone : An efficient sub-word unit for continuous speech recognition," in *Proc. European Conference on Speech Communication and Technology*, vol. 3, pp. 1215–1218, 1997.

D. McAllister, L. Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," in *Int'l Conf. on Spoken Language Processing*, vol. 5, pp. 1847–1850, 1998.

M. Ostendorf, B. Byrne, M. Bacchian, M. Finke, et al., "Modeling systematic variation in pronunciation via language-dependent hidden speaking mode," in *Proc. 1996 summer workshop on speech recognition*, http://www.clsp.jhu.edu/ws96/zilla/hidden-mode/group.html, 1996.

M. Ostendorf, "Incorporating linguistic theories of pronunciation variation into speech recognition models," in *Royal Society Phil. Trans.*, vol. 358, pp. 1325–1338, 1999.

D. Pallett, J. Fiscuss, J. Garofolo, A. Martin, et al., "1998 broadcast news benchmark test results: English and non-English word error rate performance measures," in *Proc. DARPA broadcast news workshop*, pp. 5–12, 1999.

D. B. Paul, "Extensions to phone-state decision tree clustering: single tree and tagged clustering," in *Int'l Conf. on Acoustics Speech and Signal Processing*, vol. 2, pp. 1487–1490, 1997.

J. Pitrelli, M. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the TOBI framework," in *Int'l Conf. on Speech and Language Processing*, vol. 2, pp. 123–126, 1994.

L. Rabiner and B. H. Juang, in *Fundamentals of speech recognition*, Prentice Hall (Signal processing series), 1993.

W. Reichl and W. Chou, "A unified approach of incorporating general features in decision tree based acoustic modeling," in *Int'l Conf. on Acoustics Speech and Signal Processing*, vol. 2, pp. 573-576, 1999.

M. Riley, B. Byrne, M. Finke, S. Khudanpur, A. Lolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters and G. Zavaliagkos, "Stochastic pronunciation modeling from hand-labeled phonetic corpora," in *Speech Communication*, vol. 29, pp. 209–224, 1999.

M. Saraclar, H. Nock and S. Khudanpur, "Pronunciation Modeling By Sharing Gaussian Densities Across Phonetic Models," in *European Conference on Speech Communication and Technology*, vol. 1, pp. 515–518, 1999.

M. Saraclar, H. Nock, and S. Khudanpur, "Pronunciation modeling by sharing Gaussian densities across phonetic models," in *Computer Speech and Language*, vol. 14(2), pp. 137–160, 2000.

M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, "Effect of speaking style on LVCSR performance," in *Int'l Conf. on Speech and Language Processing supplement*, pp. 16–19, 1996.

Young, S. J and Woodland, P. C, "State clustering in HMM-based continuous speech recognition," in *Computer Speech and Language*, vol. 8(4), pp. 369-384, 1994.

G. Zavaliagkos, M. McDonough, D. Miller, A. El-Jaroudi, et al., "The BBN Byblos 1997 large vocabulary conversational speech recognition system," in *Proc. Int'l Conf. on Acoustic, Speech and Signal Processing*, vol. 2, pp. 905–908, 1998.

6. Acknowledgments

We thank Prof. Richard Wright for advice on re-syllabification rules, Prof. Jeff Bilmes for useful discussions on efficient partitioning, Dr. Michiel Bacchiani for making his software tools available, and BBN for Byblos first stage recognition output.

This work was supported by the National Science Foundation, grant no. ISI-9618926. The views and conclusions contained in this document are those of the authors and should not be interpreted as reflecting the official policies of the funding agency.