Cross phone state clustering using lexical stress and context

Högberg, J. and Sjölander, K.

journal: TMH-QPSR
volume: 37
number: 3
year: 1996
pages: 007-012

http://www.speech.kth.se/qpsr
Cross phone state clustering using lexical stress and context*

Jesper Högberg and Kåre Sjölander
(Names in alphabetical order)

Abstract
This study deals with acoustic phonetic modelling in HMM based continuous speech recognition. Context dependent phone models were derived by a decision tree clustering algorithm. In particular, lexical stress was introduced as a clustering variable in addition to the phonetic context. The parameter sharing model was extended by tying HMM states across different target phones. For instance, one or more states of a tense vowel and the corresponding lax vowel were tied if they proved to be acoustically similar. The results indicate that the use of lexical stress information in acoustic modelling might be fruitful when large amounts of training data are available.

Introduction
State of the art speech recognition typically makes use of context dependent HMMs. These units are often triphones or even quinphones. However, to model all existing contexts for many speakers is an unfeasible task due to the enormous amounts of data such a task would require. Therefore, parameters are often shared between several context dependent models to achieve robust training for all possible contexts. Decision trees have been successfully used for creating robust acoustic models by clustering model parameters with respect to context, e.g. Hon & Lee (1991) and Bahl et al. (1991). Also, decision trees provide useful structures for predicting unseen triphones. The pursuit of finer context dependent models can be extended to word specific models. This has been done for function words which are very frequent and often unstressed (Lee, 1989). A related idea is to use lexical stress, that is the stress level assigned to syllables of words spoken in isolation.

This paper addresses the problem of narrow acoustic-phonetic modelling for speech recognition. Decision trees are used for analysis of phone context dependencies. Lexical stress is introduced as a clustering variable in addition to the immediate left and right phonetic contexts. Lexical stress information can be useful for acoustic modelling even though syllables with high lexical stress levels need not always be stressed in fluent speech. van Bergem (1993) found that lexical stress level and word class affiliation (function word vs. content word) significantly influenced the formant frequencies of Dutch vowels in fluent speech. These factors were more important than sentence accent. Lexical stress information has been used before in acoustic-phonetic modelling for speech recognition (Hon & Lee, 1991). The use of lexical stress was said to give no recognition improvement but the potential gain in using stress information is no doubt language dependent. The current system is designed for Swedish while the one reported in Hon & Lee (1991) was designed for English.

In this study, we investigate cross phone acoustic similarities by pooling all vowels in one cluster and all consonants in another. These data clusters are used as input to the decision tree analysis for further clustering. One vowel and one consonant tree are grown. This means that some parameters are shared by HMMs of different phones. For instance, one or more states of a tense vowel and the corresponding lax vowel are tied if they prove to be acoustically similar.

Speech material
The speech data used for training and testing in this investigation were collected as parts of the Waxholm spoken dialogue project. The Waxholm application is an information retrieval system providing data on boat traffic, restaurants, hotels, etc. in the Stockholm archipelago (Blomberg et al., 1993; Bertenstam et al., 1995a).

* This paper has also been presented at ICSLP 96, The fourth International Conference on Spoken Language Processing, Philadelphia, USA, 1996.
One part of the speech material consists of phonetically rich test sentences read by all 66 speakers. The other part pertains to man-machine dialogues recorded in sessions where the speech recognition module was substituted by a Wizard-of-Oz. The speech files have been transcribed with some 60 different labels. A broad transcription has been practised at a semi phonetic-phonematic level and the segmentation has been manually corrected. The consonant set amounts to 29 labels. Plosives were split into an occlusive segment and a release specified by one unique label each. Retroflex segments are also assigned special labels. The vowel system incorporates 23 labels including pre-R allophones. The Waxholm application database is described in Bertenstam et al. (1995b).

The training material comprises speech of 56 different speakers, 13 of which are females. The recorded speech material of the remaining 10 male and female speakers has been selected as test material. There is about two hours of speech altogether.

Method

Tree growing

Decision trees were used to iteratively partition the speech data into acoustically similar clusters. The phones were uniformly segmented into three parts prior to the tree analyses. Thus, each phone provided three tokens pertaining to the initial, middle and final parts respectively. Separate trees were grown for each part.

At each tree node, predefined binary questions were asked about the close phonetic context of the target phone. These questions could for instance be: "Is the left context a nasal?" or "Is the right context /s/?" The set used for all trees consisted of 140 such questions related to phones and phone classes. Only intra-word context was modelled. All partitions were evaluated by calculating a cost function \( R \),

\[
R = R_l + R_r
\]

\[
R_l = \frac{N_l}{N_l + N_r} \sum_j \sum_i (x_{i,j} - \bar{x}_j)^2
\]

\[
R_r = \frac{N_r}{N_l + N_r} \sum_j \sum_i (x_{i,j} - \bar{x}_j)^2
\]

where \( N_l \) and \( N_r \) are the numbers of tokens bound for the left and right nodes respectively. \( N_c \) is the number of coefficients in the spectral feature vector (normalised energy and 12 mel-frequency cepstral coefficients.) \( x_{i,j} \) is the arithmetic mean of three consecutive spectral feature vectors from the mid-part of the token. Spectral feature vectors aligned to neighbouring phones were never used in the calculation which means that some instances of \( x_{i,j} \) were derived from only one or two spectral feature vectors. \( \bar{x}_j \) is the mean of the \( j \)th spectral coefficient. Thus, the cost function, \( R \), is basically the sum of the node variances weighted by their relative node size.

When building the tree, the one question that corresponded to the lowest value of \( R \) was chosen to specify the partition of that node. The tree was grown until the final node contained less than a predefined number of tokens (165) or the context was fully specified. A threshold, \( L \), constraining the minimum node size was also defined. All questions that partitioned a node in two subnodes either of which contained less than \( L \) tokens were discarded.

All vowels were pooled in the root node to form a generic vowel prototype. Hence the label of the center phone was treated as a tree-growing variable. In practise, this means that the leaves of the final tree contain speech segments of different phone identities if they prove to be acoustically similar. The consonants were processed in the same way.

Lexical stress

Lexical stress was introduced as an additional cluster variable. Three stress levels were assumed: primary stress, secondary stress and no stress. Secondary stress is assigned to the stressed syllable in the second root morph of a compound. Consonants were assigned the same stress level as the vowel in the same syllable. The syllabification was carried out with the rule module of the KTH text-to-speech system (Carlson et al., 1991). The main rule used in the syllabification was to maintain maximal initial consonant clusters. This division was overruled by morph boundaries. Consonants following a lax stressed vowel were affiliated with the preceding syllable. The stress variable of all phone instances was consistently assigned the value corresponding to the lexical stress level. Thus, phones pertaining to function words retained their lexical stress, including primary and secondary stress levels.

Two sets of trees were constructed. One set included lexical stress as a clustering variable. These two sets contained six trees each: one consonant tree and one vowel tree for the initial, middle and final part of the speech segments respectively. This was done to obtain one tree
for each HMM state of the speech recogniser which is described in the next section. Thus, state tying between phones, for vowels and consonants, could be implemented. To ensure a fair comparison at recognition, the threshold $L$ was set to different values for the two tree sets to yield trees of the same size ($L=60$ and $L=70$).

### Top-down bottom-up clustering

Data fragmentation is a potential problem with the iterative partitioning algorithm. The contents of leaves in different branches can turn out to be quite similar. Therefore, an alternative tree growing approach was evaluated. In this scheme the trees were grown larger in a first stage. The trees were then pruned by merging similar leaves until the desired number of leaves was obtained. The merging procedure was carried out by combining nodes with minimal squared distances between the centroids. Thus, this approach included an initial top-down partitioning followed by bottom-up clustering. This technique has also been used to form compound questions from the original feature set (Hon & Lee, 1991). The stress information was only used in the vowel trees of this set.

### Speech recognition

The resulting vowel and consonant clusters were used to specify intra-word triphones which were trained with the HTK toolkit (Young et al., 1993). The triphone models were designed as three state, diagonal covariance, left-to-right HMMs without skips. The HTK facilities for state tying were employed in accordance with our decision tree analyses. The trees corresponding to the initial segments of all vowels and consonants were used to create clusters for the first state of the triphone models and the trees for the middle and final parts were used for state two and three respectively. All model sets were trained with identical sequences of Baum-Welch re-estimations and increases in the number of mixture Gaussian components. Phone bigrams were used with the recogniser.

### Results

#### Cluster analysis

For each of the three HMM states, the clustering algorithm produced about 350 clusters. The training set with stress markings contained some 2700 unique triphones whereas the set without stress comprised about 2000. After clustering these numbers decreased to approximately 1900 and 1400 respectively. That is, some model definitions became identical and indistinguishable. However, the total number of parameters of the model sets decreased by 87% and 83% respectively. Thus alleviating the data insufficiency problem and allowing more robust model estimation.

The degree of parameter sharing between models varied with the respective state number. For the vowel tree set including partitions on lexical stress, 45% of the state-one parameters were shared by different vowel models. The corresponding numbers were 32% and 52% for the middle and last state respectively. The consonants of this set shared relatively fewer parameters namely 28% for the first state, 24% for state two, and 30% for state three. Thus, the cross model state-tying was more extensive for vowels than for consonants. The fact that the middle states of different models shared fewer parameters than state one and three was anticipated. The unstressed set displayed a similar pattern with the exception for the consonants where as many parameters were shared for state two as for state one.

The state-related sharing is phone dependent. For instance, the short unstressed allophone of /e/ that occurs in post-stressed syllables, mostly in suffixes, shares state one with seven other phones, including stress levels. Only one of these vowels is stressed and tense. In state two the /e/ allophone does not share parameters with any stressed tense vowel. However, in the third and segment final state, the /e/ allophone shares parameters with no less than six different stressed tense vowel phones. This is very likely to be caused by the centralisation of stressed, tense vowels as pronounced in the mid Swedish dialects spoken by most subjects in the database.

About ten percent of the consonant partitions were made on questions concerning lexical stress for all three states of that tree set. The corresponding vowel tree for state one included five partitions on the stress variable. The number of questions on stress increased to eight for state two and twelve for the final state. These numbers correspond roughly to 4%, 6%, and 9% of the partitions respectively. It is hard to tell whether this left to right increase in stress dependency is significant or not. Figure 1 provides an example of a partition for which stress level was more important than phone identity in the same context. In order to clearly visualise the differences in the spectra, a filterbank representation is shown for state one for two different stress levels. The dotted and solid lines indicate spectra of /e/ and /e/ respectively. The thick lines represent primary
stressed phones and the thin lines represent secondary and unstressed phones. As can be seen in the figure, primary stress partitions the data better than phone identity into two spectrally similar clusters. Recognition results after each training iteration, the model sets were tested using phone recognition on the test set. All stress markings were removed prior to evaluation in order to make the comparison fair. The recognition results for the model sets with lexical stress information and without stress are shown in Figure 2 (the solid line depicts the stress models and the dotted the unstressed models.) As can be seen, the stress model set has a performance lead of approximately two percentage points for less than four mixture components. As the number of components increases they perform about the same. The results for the model set including stress information derived with the top-down algorithm (dotted line) compared to the combined top-down bottom-up method (solid line) can be seen in Figure 3. The combined top-down bottom-up method provides the best recognition results. After seven mixture components, the recognition score is about one percentage point higher using these models.

Discussion
Stress information adds a new dimension to acoustic modelling and improves the phone recognition accuracy when few mixture components are trained. The added complexity increases the need for training data. If larger amounts of training material were available an improvement would probably be seen for more mixture components as well.

The results shown in Figure 3 indicate that the combined top-down bottom-up clustering should be subjected to further investigations and refinements. Larger trees could be grown and the merging procedure should be elaborated. It can be argued that the primary stress markers should be removed from syllables pertaining to function words prior to clustering since these words often are reduced. However, some initial experiments did not yield higher recognition accuracy for that condition. On the contrary, the scores were lower. This might be a consequence

Figure 1. Filterbank representations of speech spectra. Dotted lines: /æ/. Solid lines: /e/. Thick lines: primary stress. Thin lines: secondary or no stress. The graph indicates that primary stressed /æ/ is spectrally more similar to primary stressed /e/ than to secondary or unstressed /æ/ and vice versa for /e/.

Figure 2. Phone accuracy for the model set using lexical stress (solid line) and for the reference set without stress (dotted line).
of the heterogeneity of the speech material which contains both read and spontaneous utterances. The function words in read speech might be relatively less reduced than those pertaining to spontaneous speech. Therefore, combinations of lexical stress and function word information should be used in clustering.

In this investigation results for phone recognition were reported. The continued work will include word recognition using the lexical stress dependent models.

Acknowledgements
This work has been supported by grants from The Swedish National Language Technology Program.

References


