HIERARCHICAL PARTITION OF THE ARTICULATORY STATE SPACE FOR OVERLAPPING-FEATURE BASED SPEECH RECOGNITION

Li Deng[1] *and Jim Jian-Xiong Wu*[1,2]

[1]Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada N2L 3G1.
[2]Nortel Technology, 16 Place du Commerce, Nuns' Island, Verdun, Quebec, Canada H3E 1H6

ABSTRACT

We describe our recent work on improving an overlapping articulatory feature (sub-phonemic) based speech recognizer with robustness to the requirement of training data. A new decision-tree algorithm is developed and applied to the recognizer design which results in hierarchical partitioning of the articulatory state space. The articulatory states associated with common acoustic correlates, a phenomenon caused by the many-to-one articulation-to-acoustics mapping well known in speech production, are automatically clustered by the decision-tree algorithm. This enables effective prediction of the unseen articulatory states in the training, thereby increasing the recognizer's robustness. Some preliminary experimental results are provided.

1. INTRODUCTION

In the work described in this paper, we address the problem of how aspects of speech production related to coordinated articulators' movements can be effectively used to design the phonological component of a speech recognizer grounded on the principles from articulatory phonology [3]. Our previous efforts in the development of this overlapping articulatory feature based recognizer have been reported in [4, 5, 6, 7]. This paper reports our recent work aimed at improving the performance of the recognizer under the condition of limited amounts of training data where many articulatory states may not have their associated acoustic data in training.

One main characteristics of our recognizer has been its comprehensive utilization of the speech production knowledge and its systematic and consistent formulation of the computational framework in which statistical learning can be successfully applied to the recognizer design. By the objectives of the design, the recognizer is most effective for highly fluent utterances when phonological variation and articulatory dynamics become most prominent.

Theoretically, the articulatory-feature based recognizer has advantages over the conventional ones in that it is compact in the parameter size and yet it naturally covers the context-dependent behaviors spanning over several phonetic segments. However, the recognizer developed prior to this work encountered two practical difficulties. First, under the condition that only a limited amount of training speech data are available, the probability distributions associated with articulatory states estimated from the training data often do not cover all the possible states required to specify the test utterances. Second, the total number of articulatory states in the recognizer was fixed at a number independent of the amount of training data. To improve robustness of the recognizer, it is desirable to devise a scheme in which the total number of states can be adapted to the training data size at a minimal loss of accuracy in modeling co-articulation.

Both of the above practical difficulties are resolved in this work by applying the general methodology of the decision-tree based classification. In particular, we will describe how the articulatory state space is partitioned hierarchically by a decision-tree based algorithm so that articulatory states associated with similar acoustic realizations are automatically clustered, thus controlling the total number of states in the recognizer. We will also describe how the algorithm allows the articulatory states unseen in the training speech data to be predicted by their corresponding cluster representatives (i.e., upper level nodes in the articulatory-state partition tree).

2. OVERVIEW OF THE RECOGNIZER

The articulatory state space underlying the recognizer is defined over J dimensions; the dimensionality is determined by the number of largely independent articulatory tiers responsible for speech production. Each dimension in the state space is made explicitly associated with one distinct tier of the articulatory structure, which we call an articulatory "feature" due to its symbolic nature. The i^{th} dimension, Θ_i, in the articulatory state space is characterized by N_i distinct symbolic values: $\Theta_i \in \{\alpha_i^1, \alpha_i^2, \cdots, \alpha_i^{N_i}\}$, each indexed by a phoneme. While taking a particular symbolic feature value, the i^{th} articulatory feature can be regarded as being residing in one of the N_i states at any particular time point (or frame) during the speech utterance. The J features in separate dimensions, whose change of values over time forms the state evolution process in the articulatory space, are assumed to be largely independent of each other, allowing for asynchronous timing or overlapping across the J articulatory dimensions. A Markov chain $\Lambda_i = \{\pi_k^i, a_{kl}^i\}$ is employed to represent the state evolution process for the i^{th} articulatory dimension, where π_k^i and a_{kl}^i are initial state occupation probabilities and state transition probabilities of Λ_i, respectively.

Each individual one-dimensional Markov chain $\Lambda_i, i = 1, 2, ..., J$,

is only a subcomponent of the underlying speech generation process. To complete the specification of the entire generation process, we construct from these individual Markov chains a J-dimensional, composite Markov chain $\Lambda = \{\pi_k, a_{kl}\}$ spanning the space $\Theta = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_J$. The relationship between the composite articulatory state (which represents a fixed, complete articulatory configuration) and the expected acoustic correlates associated with the state can then be characterized by an "phonetic-interface" model.[1] A state in the composite Markov chain Λ is defined as a J-tuple vector: $s = (s(1), s(2), \cdots, s(J))$, with $s(i) \in \Theta_i$ (i is the feature dimension index).

In our current implementation of the speech recognizer, five articulatory features ($J = 5$) are employed: Lips, Tongue blade, Tongue dorsum, Velum, and Larynx. Each articulatory state is dynamically constructed from a phonemic transcription[2] of an arbitrary speech utterance without limits on the size of the vocabulary (American English).

3. SYSTEM TRAINING

3.1. A new decision-tree algorithm

Decision trees have been successfully applied recently in many speech recognition problems (e.g., [2]). The algorithm developed in this work, with specific applications to the articulatory feature based speech recognizer discussed in Section 2, differs from the previous decision-tree algorithms in several key aspects. First, our decision-tree based clustering algorithm is employed to build a hierarchical partition for the entire phonological/articulatory space of speech utterances, which is constructed via elaborate articulatory timing analysis according to the speech production theory. In contrast, in other conventional speech recognizers, the decision tree was used to cluster phonetic contexts only for each individual phones. Second, since each state in our system is associated explicitly with a five dimensional articulatory feature bundle, our decision-tree algorithm is able to systematically and exhaustively ask all questions at a very detailed level of component articulatory features for *individual states*[3]. The decision tree algorithms used in the conventional systems, in contrast, asked only isolated, non-systematic, and sparse questions (often compiled by linguists' intuition) for a fixed number of *nearby segments*. Third, since the articulatory state topology in our system for each phone-in-context is constructed by a fractional feature overlapping process operating asynchronously over five articulatory dimensions, the state clustering process can be and is made to start after the range of context dependency is already determined, thereby incorporating identifiable physical constraints in articulator motions responsible for co-articulation. In contrast, in other decision-tree based speech recognizers, the heuristic left-to-right state topology has to be employed and the range of context-dependency is determined during the tree growing process with no physical constraints built in.

3.2. System training and state clustering

Algorithm I: SystemTraining

1. Train an initial model using the method described in [7], except that the acoustic distribution associated with each articulatory state is represented by a uni-modal (for computational reasons) Gaussian with a common diagonal covariance matrix;
2. Build a partition tree for the entire articulatory state space according to the Gaussian parameters obtained in Step 1;[4]
3. Train the final speech model using the state-tying information obtained in Step 2 and represent the acoustic distribution of each state with mixture Gaussian densities (with a separate diagonal covariance matrix for each different state). The standard segmental k-means training algorithm is used.

Step 2 above (involving decision trees) is the heart of the system-training Algorithm I, and is detailed here. Let Θ_i be the collection of all distinct values taken by the i-th articulatory feature ($i < J$), and let $S = (S(1), S(2), \cdots, S(J))$ denote a partition or clustering of articulatory states each consisting of a J-tupled feature vector $s = (s(1), s(2), \cdots, s(J))$. Apparently, $s(i) \in \Theta_i$, $S(i) \subset \Theta_i$ and $s \in S$ if $s(i) \in S(i)$ for $i = 1, 2, \cdots, J$, and $S_0 \stackrel{\triangle}{=} (\Theta_0, \Theta_1, \cdots, \Theta_J)$ represents the entire articulatory state space (all allowable feature bundles with no constraints built in).

Now let x be an acoustic observation, $X(S)$ be the collection of all acoustic realizations of all articulatory states $s \in S$, and $n(S) = |X(S)|$ be the sample size of set $X(S)$. During the process of building the hierarchical partition of the articulatory state space, $X(S)$ is modeled by a single Gaussian density (for computational simplicity) with a mean vector $\mu(S)$ and a common diagonal covariance matrix Σ; i.e., $P(x|x \in X(S)) = \mathcal{N}(x; \mu(S), \Sigma)$.

Further, let $(S \rightarrow S_l, S_r)$ denote the operation of splitting a partition S into two sub-partitions S_l and S_r (left and right, respectively) with $S_l(i) \cup S_r(i) = S(i)$ for $i = 1, 2, \cdots J$. A split is conditional on dimension m if $S_l(m) \cup S_r(m) = S(m)$ with $S_l(m) \cap S_r(m) = \bigcirc$ ($\bigcirc$ is the empty set), $S_l(m) \neq \bigcirc$, $S_r(m) \neq \bigcirc$ and $S_l(i) = S_r(i) = S(i)$ for $\forall i \neq m$. We only consider such a conditional split (denoted as $S \stackrel{(m)}{\rightarrow} S_l, S_r$) in our current implementation.

The decision on whether a partition should be further split is made depending on the value of the likelihood ratio[8]:

$$\left(1 + \frac{n(S_l)n(S_r)}{n(S)^2}(\mu(S_l) - \mu(S_r))^t \Sigma^{-1} (\mu(S_l) - \mu(S_r))\right)^{-\frac{n(S)}{2}} \quad (1)$$

[1] A stationary version of this interface model was described in [5], and a non-stationary version in [6].

[2] Research on incorporating the prosodic information and syllabic structure in the state construction (especially useful for multi-lingual speech recognition including Asian languages) is currently underway.

[3] The computational complexity associated with such a detailed level is mitigated by a novel constrained K-means algorithm (See Algorithm III in Section 3.2). At a minimal loss of accuracy, this algorithm avoids exhaustive searching over all possible question sets from individual features in order to find a best node-splitting question.

[4] Many articulatory states may share the same acoustic distribution after the partition tree is constructed, with the underlying physical basis of many-to-one mapping from articulation to acoustics[1].

which leads to one of the two hypotheses:

- H_0: the observation set $X(S)$ is generated from one distribution $\mathcal{N}(x; \mu(S), \Sigma)$;

- H_1: the observation set $X(S)$ is generated from two distributions $\mathcal{N}(x; \mu(S_l), \Sigma)$ and $\mathcal{N}(x; \mu(S_r), \Sigma)$.

Use of the likelihood ratio in Eqn.(1) for deciding whether or not to further split a partition S is equivalent to maximization of the following distortion measure or decision function:

$$\delta(S \stackrel{(m)}{\rightarrow} S_l, S_r) \triangleq (\mu(S_l) - \mu(S_r))^t \Sigma^{-1} (\mu(S_l) - \mu(S_r)), \quad (2)$$

which we have implemented in building our recognizer.

Given the above notations and the decision function $\delta(S \stackrel{(m)}{\rightarrow} S_l, S_r)$, the hierarchical partition of the articulatory state space is built by the following tree-building algorithm:

Algorithm II: TreeBuilding

1. Put S_0 into a stack of nonterminal partitions;
2. Iterate until the nonterminal partition stack becomes empty:
   (a) Pop up a partition S from the stack;
   (b) Find the optimal split of S:

$$\delta^* = \max_{m=1,2,\cdots J} \max_{S_l, S_r} \delta(S \stackrel{(m)}{\rightarrow} S_l, S_r); \quad (3)$$

   (c) If either δ^* or $n(S)$ is below a preset threshold, label S as a terminal partition; otherwise push the sub-partitions S_l^* and S_r^* (obtained by applying the optimal conditional split in Eqn.(3)) back to the stack and continue with Step 2.

The optimal point of $\max_{S_l, S_r} \delta(S \stackrel{(m)}{\rightarrow} S_l, S_r)$ can be obtained by enumerating all possible ways of binary splitting $S(m)$, the set of distinct feature values in m-th dimension for node S. However, it is practically impossible because the number of alternatives is too high. For example, there are 20 variants of distinct tongue dorsum features in our system so the number of possible split at the root node for the tongue dorsum dimension would be 2^{20}. Defining a within-cluster distortion measure as

$$\begin{aligned}
&\delta_1(S \stackrel{(m)}{\rightarrow} S_l, S_r) \\
&\quad \triangleq E\{(x - \mu(S_l))^t \Sigma^{-1} (x - \mu(S_l)) | x \in X(S_l)\} \\
&\quad + E\{(x - \mu(S_r))^t \Sigma^{-1} (x - \mu(S_r)) | x \in X(S_r)\}. \quad (4)
\end{aligned}$$

Since

$$\begin{aligned}
&\delta(S \stackrel{(m)}{\rightarrow} S_l, S_r) + \delta_1(S \stackrel{(m)}{\rightarrow} S_l, S_r) \\
&\quad = E\{(x - \mu(S))^t \Sigma^{-1} (x - \mu(S)) | x \in X(S)\} \quad (5) \\
&\quad = const,
\end{aligned}$$

one can maximize $\delta(S \stackrel{(m)}{\rightarrow} S_l, S_r)$ by minimizing $\delta_1(S \stackrel{(m)}{\rightarrow} S_l, S_r)$, which can be achieved by applying the following constrained iterative k-means ($k = 2$) algorithm:

Algorithm III: NodeSplitting

1. Create temporary minimum partitions for the m-th feature dimension of S, $S_1, S_2, \cdots, S_{|S(m)|}$ with $S_i(j) = S(j)$ for $\forall j \neq m$ and $S_i(m) = s \ \ (s \in S(m))$;
2. Initialize $\mu(S_l)$ and $\mu(S_r)$;
3. Set S_l and S_r to empty sets;
4. For each minimum partition $S_i, i = 1, 2, \cdots, |S(m)|$, set $S_r = S_r \cup S_i$ and add $X(S_i)$ to $X(S_r)$ if

$$\begin{aligned}
&(\mu(S_i) - \mu(S_r))^t \Sigma^{-1} (\mu(S_i) - \mu(S_r)) < \\
&(\mu(S_i) - \mu(S_l))^t \Sigma^{-1} (\mu(S_i) - \mu(S_l)), \quad (6)
\end{aligned}$$

   otherwise set $S_l = S_l \cup S_i$ and add $X(S_i)$ to $X(S_l)$;
5. Updating $\mu(S_l)$ and $\mu(S_r)$ from $X(S_l)$ and $X(S_r)$;
6. Goto Step 4 until S_l and S_r are the same as that obtained from the previous iteration.

The above algorithm is just a two-means ($k = 2$) clustering algorithm except for the constraint that all $x \in X(S_i)$ should be clustered into the same descendent node.

4. EXPERIMENTS

Preliminary experiments have been conducted to evaluate the effectiveness of the decision tree algorithm for adaptive clustering of articulatory states and for predicting unseen articulatory states as described in Section 3. The task is the phonetic recognition of standard 39 folded phone classes in continuous TIMIT sentences. To reduce computation complexity in the recognition experiment, we adopt the strategy of re-evaluating N-best phonetic label hypotheses for each TIMIT sentence using the computation intensive feature-based, long-span context dependent models. Given the N-best phonetic label sequences, re-scoring each sequence using the feature-based model described in this paper is as follows. For each phone in the sentence, we take both of its left and right contexts, expressed in terms of each individual feature component (which is often spread from several phones away), into account to construct the articulatory HMM states. Given the resulting state topology for each contextual phone in the N-best sequences, we concatenate them into a sentence-level state topology according to the N-best hypotheses. Then the Viterbi-like algorithm is applied to re-score all the N phonetic label sequences and the new top sequence is regarded as the output of the recognizer.

The feature-based speech recognizer was implemented with and without use of the hierarchical partition of the articulatory state space. The testing set consists of 48 randomly selected SX sentences from 48 speakers (the selection process guarantees that each region has four male speakers and two female speakers).

Table 1 shows the phonetic recognition performances, in terms of percent correct, percent accurate, percent substitution error, percent

deletion and insertion errors, for the feature-based system with the decision tree algorithm for state state partition implemented (row A), in comparison with the benchmark system with no state partition implemented (row B). A total of 3,696 sentences from 462 TIMIT speakers were used in the training. In Table 2 are the performance figures with use of only 480 sentences from 60 speakers in the training.

	Corr.	Acc.	Sub.	Del.	Ins.
A	69.83%	55.33%	25.28%	4.89%	14.50%
B	69.39%	53.90%	26.46%	4.15%	15.49%

Table 1. Performance of the speech recognizer with (row A) and without (row B) use of decision-tree algorithm for state partition. 462 speakers in the training data.

	Corr.	Acc.	Sub.	Del.	Ins.
A	59.91%	49.81%	32.34%	7.74%	10.10%
B	59.95%	46.47%	34.10%	5.96%	13.47%

Table 2. Same as Table 1 except only 60 speakers used in training.

The results in Tables 1 and 2 show that the improvement of the recognizer performance via use of the decision tree based algorithm has been marginal or negligible. This has not been our expectation. [5] Due to the preliminary nature of the algorithm development, we have not been able to draw conclusions on the effectiveness of the idea of partitioning and clustering the articulatory state space. It is likely that several assumptions implicitly or explicitly made in the decision tree algorithm described in Section 3 will require serious examinations before the theoretical advantages of the ideas behind the algorithm can be realized.

5. SUMMARY AND DISCUSSIONS

Compared with conventional recognizers using phoneme-sized speech units, the overlapping articulatory feature based recognizer we developed over the past few years has theoretical advantages of compactness in the model parameterization and of the ability to cover the context-dependent behaviors of speech data. The improvement of the recognizer described in this paper is intended to push the above advantage of compactness further under the condition of unseen articulatory states (training and testing mismatch), thus increasing the robustness of the recognizer and making the recognizer potentially adaptive to the size of the training data.

The methodology we employed to achieve the robustness and to predict the unseen articulatory states is based on the decision tree algorithm which has already enjoyed a wide success in the conventional phonetic HMM based speech recognizers. In contrast to the conventional decision tree method which clusters HMM states only on the basis of the surface acoustic similarity in the speech signal, the new decision tree algorithm we developed which is made specific to our articulatory feature based recognizer is grounded on the physical phenomenon of many-to-one articulation-to-acoustics relations [1]. Although overlapping of the output distributions associated with separate articulatory states already allows the recognizer to embody the many-to-one relations, this does not resolve the problem of training and testing mismatch exhibited by the presence of abundant unseen articulatory states which we observed prior to this work. The strong tying and partitioning of the articulatory states determined by the decision tree algorithm eliminates the problem of unseen states by explicitly forcing the acoustic distribution *parameters* associated with many articulatory states to be identical (many-to-one mapping), rather than just making the possible outcomes from the acoustic distributions to coincide as in the previous version of our recognizer.

Given the physical basis of many-to-one articulatory-to-acoustic mapping which justifies the articulatory state partitioning, we developed a new decision tree algorithm that has relied upon the articulatory interpretation of the HMM states. Algorithmically, it also differs from the previously published decision tree algorithms in several aspects. For example, our algorithm theoretically allows to exhaustively ask all the relevant questions at the detailed level of articulatory features, needing no linguists' insights to design necessarily incomplete question sets. Also, the decision tree is employed to partition the entire articulatory state space instead of clustering phonetic contexts for individual phones in other systems.

Unfortunately, at the time of this writing, the many theoretical advantages of our decision tree algorithm offered by the above several theoretical reasonings have not been demonstrated in evaluation experiments. Some preliminary, discouraging experimental results have been provided in this paper while more comprehensive evaluations are underway.

[5] It has been expected that the results in Table 2 show a much greater performance improvement than those in Table 1 because of the robustness of the recognizer achieved, at least theoretically, by state clustering for use with a small amount of training data.

6. REFERENCES

1. B. Atal, J. Chang, M. Mathews, and J. Tukey. "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique," *JASA.*, Vol. 63, pp. 1535-1555, 1978.
2. L. Bahl, P. de Souza, P. Gopalakrishnan, D. Nahamoo and M. Picheny. "Decision trees for phonological rules in continuous speech," *Proc. ICASSP'91*, pp.185-188, 1991.
3. C. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, Vol.49, pp. 115-180, 1992.
4. L. Deng. "Design of a feature-based speech recognizer aiming at integration of auditory processing, signal modeling, and phonological structure of speech," *JASA*, Vol. 93, No.4, pp. S2318, April, 1993.
5. L. Deng and D. Sun. "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *JASA*, Vol. 95, No. 5, May 1994, pp. 2702–2719.
6. L. Deng and H. Sameti. "Transitional speech units and their representation by the regressive Markov states: Applications to speech recognition," *IEEE Trans. Speech Audio Proc.*, July, 1996.
7. L. Deng, J. Wu and H. Sameti. "Improved speech modeling and recognition using multi-dimensional articulatory states as primitive speech units," *Proc. ICASSP'95*, pp.385-388, 1995.
8. A. Kannan, M. Ostendorf and J.R. Rohlicek. "Maximum likelihood clustering of Gaussians for speech recognition ," *IEEE Trans. Speech Audio Proc.*, Vol.2, No.3, pp. 453-455, 1994