Data-Driven Model Construction for Continuous Speech Recognition Using Overlapping Articulatory Features

Jiping Sun, Xing Jing, Li Deng,*
Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, Canada
(*Current address: Microsoft Research, One Microsoft Way, Redmond, WA.)

ABSTRACT

A new, data-driven approach to deriving overlapping articulatory-feature based HMMs for speech recognition is presented in this paper. This approach uses speech data from University of Wisconsin's Microbeam X-ray Speech Production Database. Regression tree models were created for constructing HMMs. Use of actual articulatory data improves upon our previous rule-based feature overlapping system. The regression trees allow construction of the HMM topology for an arbitrary utterance given its phonetic transcription and some prosodic information. Experimental results in ASR show preliminary success of this approach.

1. INTRODUCTION

Over the past several years, we have been developing a new, data-driven approach to deriving overlapping articulatory-feature based HMMs for speech recognition. This approach uses simultaneous articulatory and acoustic data from the University of Wisconsin Microbeam X-ray Speech Production Database [2,14]. It then builds statistical models using regression trees [12]. Use of the actual articulatory data improves upon our previous rule-based feature overlapping system [7,8,9,15]. The regression trees learned from the articulatory data allow direct construction of the HMM topology appropriate for any arbitrary utterance if given its phonetic transcription and high-level prosodic information such as stress value and syllabic function of each phone.

The basic framework of our approach is the five articulatory tiers or feature dimensions: the lips, the tongue tip, the tongue dorsum, the velum, and the larynx. In each of these articulatory dimensions, a phonetic unit is associated with one or more symbolic features. Based on this framework and on the findings from experimental phonetics and autosegmental phonology, we established a set of rules that describe the temporal overlapping of features between neighboring phones. Many of the pronunciation alternations are naturally accounted for by this feature overlapping process, for example, the assimilation of velum features (nasalization), lip features (lip rounding) and larynx features (voicing/unvoicing), etc.

In contrast to the conventional allophone-based approach to pronunciation modeling, this articulatory feature-based approach links itself to the physical process of speech production. This link makes it possible to use experimental data to enhance our earlier rule-based HMM topology construction method. The rule-based method now is expanded to include numerical parameters: the percentile temporal overlap between a pair of features. This allows us to incorporate in the new system a learning component using articulatory data. In our recent experiments, a Java-based graphical interface has been developed for hand-labeling of articulatory feature overlapping with the Microbeam X-ray data. The hand-labeled data is used for training regression trees. This labeling process is carried out by hands and eyes, aided by the Java-based graphical interface.

To test the effectiveness of this new, data-driven approach, the TIMIT speech corpus is used for training and testing the newly constructed, articulatory-feature based HMMs. The initial results have shown superior performance over the triphone-based approach in the phone recognition tasks. In the remaining sections of this paper, we introduce our new data-driven framework, the use of the X-ray microbeam data, the construction of the HMM topology, and some preliminary ASR experimental results.

2. THE ARTICULATORY FEATURE FRAMEWORK

We created a five-tier framework of articulatory features for use in our system development. These five tiers describe active articulators involved in the pronunciation of speech sounds. Each articulator is located at one of these five tiers. An articulator may take up a feature from each of a few feature dimensions. Each feature dimension has a set of possible features. The tier to articulator correspondence is shown in Table 1.

TIER	*ARTICULATORS*	*DIMENSIONS*
1	Upper Lip, Lower Lip	1: shape,2: manner
2	Tongue Tip, Tongue Blade	1: place,2: manner
3	Tongue dorsum, Tongue Root	1: place,2: manner
4	Velum	1: nasal opening
5	Glottis	1: phonation

Table 1. Articulators on five tiers.

At each tier, an articulator takes up one feature from each feature dimension. Each tier may be specified by one or more feature dimension. Each feature dimension contains a set of possible features. Which feature will be taken up depends on the phone that is pronounced. If we do not consider asynchrony of features at the five tires, which is a character of spontaneous

speech and will be explained later, the pronunciation of a phone can be described statically by a bundle of simultaneous features. Thus we say a pronunciation unit can be expressed by a feature bundle using features from five tiers.

A few examples of phones expressed by feature bundles are given below. (The TIMIT style phone names are used.)

- [dx] as in *ladder*. Lip = [flat, open], Tongue Tip = [alveolar, flap], Tongue Root = [low, open], Velum = [high], Glottis = [voicing]
- [nx] as in *ma**nn**er*. Lip = [flat, open], Tongue Tip = [alveolar, flap], Tongue Root = [low, open], Velum = [low], Glottis = [voicing]

We may call these static feature bundle descriptions of phones their lexical descriptions, which can be affected by overlapping features of neighboring phones in spontaneous speech. When this happens, features at each tier will have different temporal behaviors and may overlap with features of other phones.

In the following example, we show how such alternation phenomena as lip rounding and velum lowering (nasalization) can be accounted for by feature overlapping. Consider the word *strong* and its pronunciation [s t r ao ng]. The nasal consonant [ng] can overlap its velum feature with features of [r] and [ih], and [r] can overlap its lip feature with features of [s] and [t]. As a result, the phones [s t r ao] of this word can assimilate features from neighboring phones and their pronunciations undergo a process of alteration. This can be illustrated by the gestural score representation as shown in Fig 1.

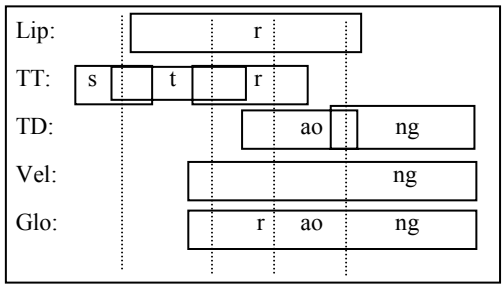

Figure 1. Feature bundles of *strong.*

Fig 1 uses the gestural score representation to show feature bundles of phones in their overlapping relations. In this figure we can see that the velum feature of [ng], i.e. the nasal lowering feature overlaps with several phones and so does the lip feature of [r], i.e. the lip rounding feature. In the feature overlapping situation, a phone is no longer represented by a single feature bundle of static nature, but by a number of feature bundles. This feature bundle series just form the basis for our construction of HMM topologies: each feature bundle corresponding to a HMM state. This is in comparison with the triphone-based models that use several states (normally 3) to represent a context-dependent phone, in which the boundary states represent the transition from phone to phone. In a triphone model, boundary states only reflect the influence of the immediate neighboring phones while in our model a state may reflect influence of a more distant neighboring phone.

3. USE OF THE X-RAY MICROBEAM SPEECH PRODUCTION DATABASE

In this section we describe the use of the Wisconsin X-ray speech production database. Based on the five-tier articulatory feature framework described in section 2, we wanted to collect information from real speech data on the duration and overlap of articulatory features. We used the University of Wisconsin's X-ray Microbeam Speech Production Database [2] for the intended work. Consequently, a feature overlapping database with regression-tree based prediction models has been created and used in our speech recognition research.

3.1 The X-ray Speech Production Corpus

The University of Wisconsin's Microbeam X-ray Speech Production database used in this study contains natural, continuous spoken utterances in both isolated sentences and short paragraphs. The speech data were recorded from 32 female speakers and 25 male speakers. Each speaker completed 118 tasks. Some of the tasks are unnatural speech, which were not used in our work. The data come in three forms: text data, which are the orthographic transcripts of the spoken utterances; digitized waveforms of the recorded speech; and X-ray trajectory data of articulator movements, simultaneously recorded with the waveform data.

The trajectory data are recorded for the individual articulators. The articulators are arranged as Upper Lip, Lower Lip, Tongue Tip, Tongue Blade, Tongue Dorsum, Tongue Root, Lower Front Tooth (Mandible Incisor), Lower Back Tooth (Mandible Molar). On each articulator of the speaker a pellet is attached to record its movement in the sagittal plane.

Based on this data set, we first carried out a number of necessary transformations. The orthographic transcripts are converted into phonetic transcripts. The conversion is based on the TIMIT dictionary. The phoneme set used by the dictionary is extended with allophones that are predictable by the phonetic context. The waveform data are transformed into wideband spectrograms that can be displayed in a window of the graphical labeling tool. The trajectory data is displayed as two-dimensional curves of time versus position for each of the eight articulators. The positions are factored into X-component and Y-component for forward-backward and up-down movements in the sagittal plane.

3.2. Labeling Articulatory Features

The feature labeling work is based on the theory of autosegmental phonology [3,11] and articulatory phonology [4]. These theories propose non-linear segmental features, especially articulatory features. This labeling work is also based on our previous work of feature overlapping models in speech recognition application [7,8,9,15].

we first performed segmentation and alignment. The spectrograms are aligned with the trajectories. The starting and end positions of both figures are aligned. Next, the

spectrograms are segmented according to the speech tasks and aligned with the phones of the utterance. The labeling is focused on the identification and tagging of articulatory features in the trajectories and aligning them with the phonetic symbols and appropriate sections of the spectrogram. Based on the five-tier articulatory feature model, both the trajectory and spectrogram data are used for locating features. For example, a lip opening feature can be identified on the Y position curve of the Upper or the Lower Lip, depending on the phone. A lip rounding feature can be identified on the Lips X position curve, and so on. Fig 2 shows some labeled features for the sentence *The other one is too big*, in which the articulators Upper Lip, Tongue Tip and Tongue Root are used for identifying tier 1, 2 and 3 features respectively, while other articulators are used only for reference. The tier 4 and 5 features are mainly identified from the spectrogram.

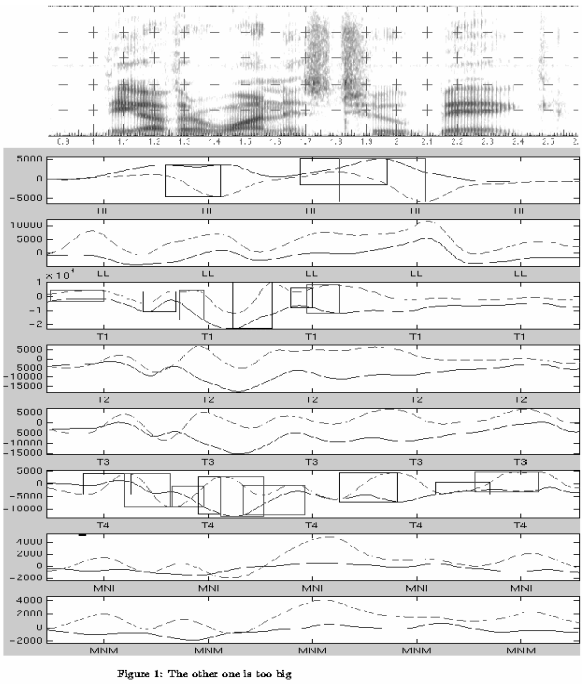

Figure 1: The other one is too big

Figure 2. The labeled sentence *The other one is too big*.

With a Java based labeling tool developed by our group, we are able to align spectrograms, phones and features graphically, save and reload labeled utterances and obtain the numerical data of feature duration, prominence and overlap. Currently we only use the duration and overlap information for deriving regression trees and gestural scores. The prominence (position) data is also retained, which can be used for estimating constriction degrees or build speech synthesis models.

The result of the labeling work is a feature overlapping database that provides numerical data of articulatory feature duration and overlap for natural English speech. Based on this database, we are able to derive predictive models for creating gestural scores if given an arbitrary phone string of an utterance.

3.3. Building a Predictive Model

The model for predicting overlaps of articulatory features is based on regression trees, which are automatically learned from the data of the labeled corpus. We expect feature overlapping to be context-dependent. Thus, since the labeled corpus only contains limited contexts for each phone, there is need to generalize the labeled corpus so that an arbitrary phone sequence of a speech task can be best dealt with.

A set of regression trees is trained for predicting feature duration and overlapping at for phones in context. The training data has numerical values as the **dependent variable** and symbolic features of left and right phones as the **predictors**. The University of Minnesota's Firm regression tree learning tool [12] is used. The predictors we used for training a regression tree include the features of its left and right two-phones. The predictors also include these phones' higher-level prosodic information: word stress, syllabic function (onset, coda or nucleus) and word boundary information. So a training example for a feature duration or overlap consists of 32 predictor values. Following is a training example of the tier-1 overlapping of stop consonants:

```
18, wi, 0, n, 0, 0, mmopn, n0, v1, wi, 0, m, labcls, 0, 0, n1, v1,
    wi, 1, n, 0, 0, lfopn,    n0, v1, wi, 1, n, 0, 0, hfcrt,    n0, v1
```

The number 18 is the dependent variable, meaning an overlapping of 18 units (one unit is 0.866 ms). This is followed by four neighboring phones' features each consisting of boundary, stress, syllabic information and tier-1 to tier-5 features. Altogether 60 regression trees were trained for 30 tiers of 10 phone types. The regression trees generalize for every possible five-phone context since only features are used as context information. One of the applications of this model is to predict Hidden Markov Model topologies in automatic speech recognition systems. Here is a HMM model toplogy for [s].

```
~o <VecSize> 39 <MFCC_0_Z_D_A>
~h "t_253"
<BeginHMM>
 <NumStates> 6
 <State> 2
 ~s "s296"
 <State> 3
 ~s "s37"
 <State> 4
 ~s "s393"
 <State> 5
 ~s "s1413"
<TransP> 6
0.0 1.0 0.0 0.0 0.0 0.0
0.0 0.230769 0.769231 0.0 0.0 0.0
0.0 0.0 0.692308 0.307692 0.0 0.0
0.0 0.0 0.0 0.230769 0.769231 0.0
0.0 0.0 0.0 0.0 0.115385 0.884615
0.0 0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

4. EXPERIMENTAL RESULTS

Using the data-driven predictive model we carried out experiments in speech recognition. The TIMIT phone recognition task is chosen for our experiments. Compared with the triphone-based approach, the feature-based approach predicts model states by considering larger-span context, up to two or three phones to each side of a central phone. This results in more discriminative training of the models.

Using the HTK toolkit [16], we have trained all the context-dependent phones as predicted by the overlapping model from the training section of TIMIT corpus. This resulted in 64230 context dependent phones based on 39 monophone set. Then we used the decision tree based state tying to overcome the data insufficiency problem. Our questions for decision tree based state tying are designed according to the predictions made by the feature overlapping model. Five-phone context is used in the question design. The contexts that are likely to affect the central phones through feature overlapping, as predicted by the model, form questions for separating a state pool. For example, the nasal release of stops in such context as [k aa t ax n], [l ao g ih ng] will give rise to questions as *+ax2n, *+ih2ng, etc, where the '2' is used to separate first right context phone from second right context phone. The experiment results for phone recognition are as follows.

SYSTEM	CORRECTION %	ACCURACY %
Triphone (Baseline)	73.99	70.86
Overlapping-feature	74.70	72.95

The test was done on the 1680 test files of the TIMIT corpus. There are a total number of 53484 phone tokens appearing in these files. The initial application of the feature overlapping model based on corpus data and machine learning has shown that this is a powerful model.

Currently we are continuously labeling the feature overlapping database. With more data available we expect better results will be achieved. We also plan to incorporate rule-based prediction models with the data-driven models for speech recognition experiments. In our future work, we plan to apply the overlapping model obtained from English data to other languages. It is our assumption that articulatory features and their overlapping patterns can be shared by all languages to a high degree.

5. REFERENCES

1. Abbs, J. H., "Invariance and Variability in Speech Production: a Distinction between Linguistic Intent and its Neuromotor Implementation:, in J. S. Perkell and D. H. Klatt (eds) Invariance and Variability in Speech Processes, pp. 202-218, Hilldale, NJ: Lawrence Erlbaum Associates, 1986.
2. Abbs, J. H., *Users' Manual for the University of Wisconsin X-ray Microbeam*. Madison, WI: University of Wisconsin Waisman Center, 1987.
3. Bird, S., *Computational Phonology: A Constraint-based Approach*. Cambridge University Press. 1995.
4. Browman, C.P., and L. Goldstein, "Articulatory Gestures as Phonological Units". *Phonology, 6:201-251*, 1989.
5. Church, K. W., *Phonological Parsing in Speech Recognition*. Kluwer Academic Publishers, 1987.
6. Coleman, J., *Phonological Representations*, Cambridge University Press, 1998.
7. Deng, L., "Autosegmental Representation of Phonological Units of Speech and Its Phonetic Interface", *Speech Communication*, 23(3):211-222, 1997.
8. Deng, L., "Finite-state Automata Derived from Overlapping Articulatory Features: A Novel Phonological Construct for Speech Recognition", *Proceedings of the Workshop on Computational Phonology in Speech Technology, (Association for Computational Linguistics), Santa Cruz, CA, pp. 37-45,* 1996.
9. Deng, L., "Integrated-multilingual Speech Recognition Using Universal Phonological Features in a Functional Speech Production Model", *Proceedings of the IEEE International Conference on Acoustics Speech, and Signal Processing, 2:1007-1010*, 1996.
10. Deng, L. and D. Sun, "A Statistical Approach to Automatic Speech Recognition Using the Atomic Units Constructed from Overlapping Articulatory Features", *J. Acoust. Soc. Am.,:2702-2719.*1995.
11. Goldsmith, J.A., *Autosegmental and Metrical Phonology*. Blackwell. 1990.
12. Hawkins, D. M., *Firm: Formal Inference-based Recursive Modeling, Release 2.2 User's Manual*, University of Minnesota, 1999.
13. Jensen, J.T., *Phonology*. John Benjamins Publishing Company, 1993.
14. Kiritani, S., "X-ray microbeam method for Measurement of Articulatory Dynamics: Techniques and Results", Speech Communication 5, pp. 119-140, 1986.
15. Sun, J. and L. Deng, "Use of High-level Linguistic Constraints for Constructing Feature based Phonological Model in Speech Recognition", *Australian Journal of Intelligent Information Processing Systems*, 5:4 PP. 269-76, 1998.
16. Young, S., "A Review of Large-Vocabulary Continuous-Speech Recognition", *IEEE Signal Processing Magazine*, Vol. 13, No. 5, pp. 45-57,1996.