ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2000)
Fragrant Hill Hotel, Beijing
October 13-15, 2000

Annotation and Use of Speech Production Corpus for Building Language-Universal Speech Recognizers

*Jiping SUN, Xing JING, and Li DENG**

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo

{Jsun,xjing,deng}@crg3.uwaterloo.ca,http://sip.uwaterloo.ca/{jsun,xjing,deng}

(*Current address: Microsoft Research, One Microsoft Way, Redmond, WA.)

ABSTRACT

A corpus linguistic study is reported in this paper, guided by articulatory phonology and by general phonetic principles of speech production. A direct application of this study is the construction of Hidden Markov Model topologies for automatic speech recognition, taking into account integrated multilingualism with the consideration of the common physiological organs and processes involved in the production of speech sounds from the world's languages. We demonstrate in this study that incorporation of speech production principles can provide effective constraints on pronunciation modeling for the purpose of building language-universal speech recognizers.

1. INTRODUCTION

Speech sounds of different languages are produced by many common elements in the articulation process and within the same vocal tract of the speaker. Therefore, the speech sounds necessarily share common acoustic properties despite the differences at the higher linguistic levels. As one example of the phonological universal, several broad categories of speech sounds are widely shared by all languages in the world, such as stops, fricatives, glides, nasals, vowels, etc. Further, the articulatory features either distinguishing or shared by sets of speech sounds, e.g., those that are concerned with the place of articulation or constriction manners, are of physiological nature and universal to all languages.

The purpose of this study is to establish a speech production model at the level of articulator movements as they are related to phonological entities, the phonemes and articulatory features, base on autosegmental and articulatory phonology [10, 2, 3]. We aim at the prediction of articulatory characteristics given an input phoneme sequence of an utterance. The phoneme sequence will carry higher-level, prosodic information, such as word and phrase boundary, syllabic functions of a phoneme, etc. The articulatory characteristics are described as feature spreading at several independent articulatory tiers. Each tier stands for an active articulator that participates in uttering speech sounds. Features of the same sound may start and end at different times; features of neighboring sounds may exist at the same time. Thus this model is known as the overlapping of articulatory features [7,14].

We started by annotating a large speech production corpus that contains both sound wave data and X-ray articulator trajectory data of continuous speech, using a graphical annotation tool we have developed. A substantial amount of annotation work has been carried out, resulting in a numerical articulatory feature database, tagging the natural, continuous utterances of the speech production database. Based on the annotated data, an articulatory feature-based speech production model has been constructed using regression trees [11]. Applying regression trees to any arbitrary phoneme sequence automatically produces gestural scores or feature-overlapping patterns.

The regression tree model serves as a mapping from phoneme sequence input (with higher-level prosodic information) to low-level, articulator movements in terms of duration and overlapping of features. The sequence of such feature bundles corresponds to acoustic realizations of speech. That is, each feature bundle corresponds to a relatively homogeneous stretch of acoustic signal and the transition from one feature bundle to another corresponds to a transition in acoustic signal properties. So the speech production model is used to construct context-dependent phone models for HMM-based ASR systems. The difference between our approach toward speech recognition model building and the triphone-based models will be discussed later.

We have used this speech production model to create context-sensitive phone model topologies and used these in training speech recognition systems. We have experimented with the TIMIT speech corpus data and compared this approach to triphone-based approach. In the phone recognition experiment and comparative study we have achieved better performance than a triphone baseline system.

2. A FEATURE SPECIFICATION SYSTEM

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

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

Table 1. Articulators on five tiers.

At each tier, an articulator may take up one feature from each feature dimension. Each dimension has a set of possible features. Which feature is taken up depends on the phone being pronounced. If we did not consider asynchrony of features at all the tires, the pronunciation

of a phone would be described statically by a bundle of features simultaneously at the five tiers. A few examples of such feature bundles are given below[1]:

- [dx] as in *le**tt**er*. 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]
- [p] as in *s**p**eak*. Lip = [flat, closed], Tongue Tip = [neutral, open], Tongue Root = [low, open], Velum = [high], Glottis = [-voicing]

We call these static feature bundle descriptions of phones the lexical descriptions, which are to be affected by spreading features of neighboring sounds in continuous speech. When this happens, features at each tier will have different temporal ranges and may overlap with features of other phones in time. One example is the above-mentioned word *"speak"*. In real speech, the phone [p] will become unaspirated.

The authors have previously summarized a set of articulatory phonological rules accounting for pronunciation alterations [14]. These rules account for phenomena such as assimilation, co-articulation, etc. in terms of the overlapping of articulatory features. Our present work is a data-driven approach to deriving a predictive system.

In the following example, we show how such alteration phenomena as lip rounding and velum lowering (nasalization) can be accounted for by feature overlapping. Consider the word *string* and its pronunciation [s t r ih ng]. The nasal consonant [ng] can overlap its velum feature with [r] and [ih], and [r] can overlap its lip feature with [s] and [t]. As a result, the phones [s t r ih] in this word can assimilate features and alter their pronunciations. The gestural scores can represent this as shown in Fig 1.

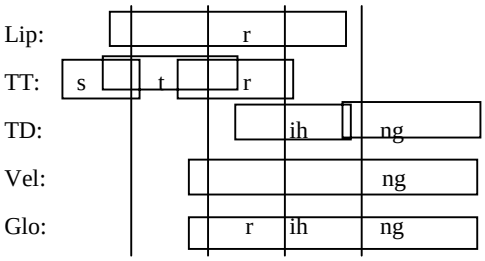

Figure 1. gestural scores and feature bundles of the word "*string*".

Fig 1 uses the gestural score format to show feature bundles of phones in their overlapping relations. In this figure 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.

3. LABELING THE SPEECH CORPUS

In this section we describe the labeling of the Wisconsin X-ray speech production corpus [1]. Based on the five-tier articulatory feature framework described in section 2, we want to collect information from real speech data of the duration and overlap of such features. This corpus provides the possibility for such work.

3.1 The X-ray Speech Production Corpus

The University of Wisconsin's Microbeam X-ray Speech Production database as used in this study contains natural, continuous spoken utterances in both isolated sentences and short paragraphs. 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 from 57 speakers each performing 118 speech tasks.

The trajectory data are recorded for individual articulators. The articulators are arranged as Upper Lip, Lower Lip, Tongue Tip, Tongue Blade, Tongue Dorsum, Tongue Root, Lower Front Tooth (Mandible Incisor) and 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 phone set 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 set to the form of two-dimensional curves in time and position for each of the eight articulators. The positions are factored into X-component and Y-component for forward-backward and up-down movements (see Fig 2 for an example).

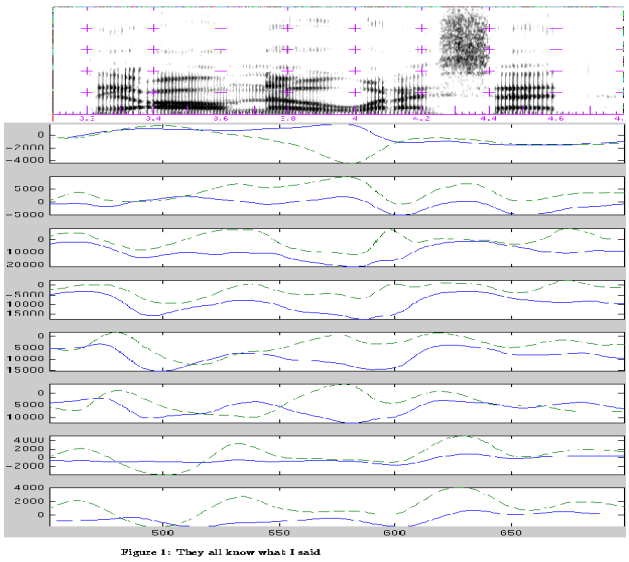

Figure 2. The Microbeam X-ray Speech Data.

3.2. The Labeling Process

The feature labeling work is based on the theory of autosegmental phonology that proposes non-linear segmental features and on our previous work of feature overlapping models in speech recognition

[1] Throughout this paper, phone names are written in the TIMIT style.

application.

After transforming the data into appropriate forms, we performed segmentation and alignment. First, the spectrograms are aligned with the trajectories. The starting and ending positions of both figures are aligned. Next, the spectrograms are segmented and aligned with the phonetic symbols of its corresponding utterance.

The labeling work 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, the trajectory and spectrogram data are used for locating features on each of the five tiers. 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 X position curve. Fig 3 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, while other articulators are also referred to. The tier 4 and 5 features are mainly identified from the spectrogram.

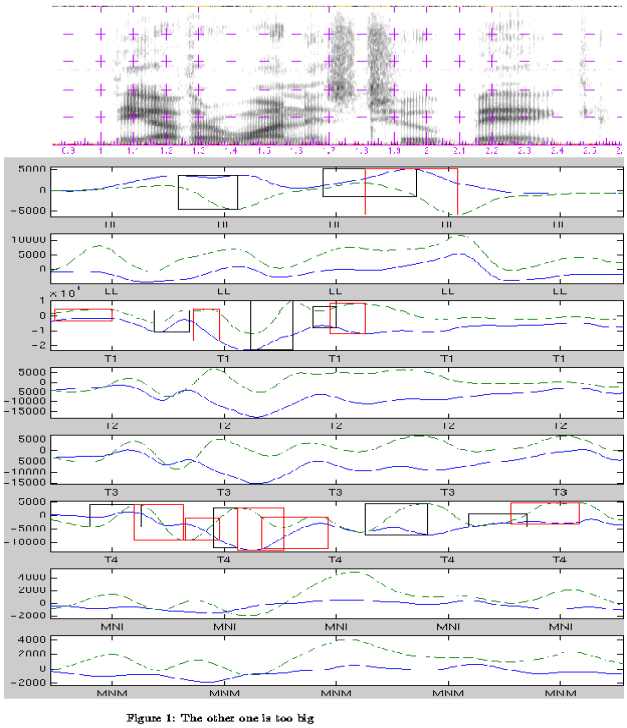

Figure 3. The labeled sentence "*The other one is too big*".

With a Java based labeling tool developed by our group, we are able to align spectrogram sections, phonetic symbols and features, save and reload labeled utterances and obtain the numerical data. Currently we only use the duration and overlap information for deriving regression trees and gestural scores. The position data are also saved, 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 when given an arbitrary phone string of an utterance. Fig 4 shows the interface of the labeling tool.

4. THE 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 estimated.

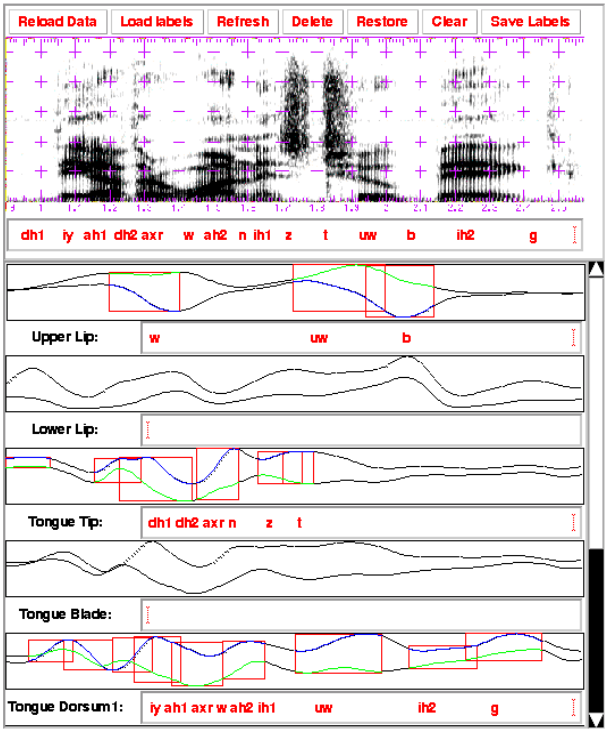

Figure 4. The feature labeling tool interface.

A set of regression trees is trained for predicting feature duration and overlapping at each tier for phones in context. The training data for regression trees have numerical values of duration and overlapping as the **dependent variable** and phonological features of left and right phones as the **predictors**. University of Minnesota's Firm regression tree learning system is used [11]. The predictors include the five-tier features of its left and right two-phone context and these phones' higher-level prosodic information: word stress, syllabic function (onset, coda or nucleus) and boundary information (word beginning, word internal or word end and utterance boundaries). So a training example for a feature (either for duration or for overlap) consists of 32 predictor values. This is a training example of the tier-1 overlap for 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 (wi: word internal, 0: unstressed or neutral feature, n: nucleus, mmopn: vowel tongue dorsum middle and open, m: word internal consonant, n0: velum high, v1: voicing, labcls: labial-closure lip feature, etc.). 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, and the regression trees put features with maximal information gain at the top of the tree. As a result, even though a context has not been seen in the corpus, the best possible estimate based on the similarity between the feature matrices is calculated by the regression trees.

This resulted in a five-dimensional speech production model with outputs in the form of articulator gestural scores. One of the applications of this model is to predict Hidden Markov Model topologies used in constructing automatic speech recognition systems. Our model has been used in this way and the initial results have shown better performance in terms of recognition rate to the currently prevailing triphone model.

5. SPEECH RECOGNITION RESULTS

Experiments have been carried out using the tool for predicting feature overlapping as described so far in this paper. The TIMIT phonetic recognition task is chosen for our experiments. Compared with the triphone-based approach, the feature-based approach predicts model states by considering larger-scope context, up to two or three phones on each side of a central phone. This results in more discriminative training of the models.

Using the HTK toolkit [15], we have trained all the context-dependent phones as predicted by the overlapping models from the training set of TIMIT corpus. This resulted in 64230 context dependent phones based on 39-phone 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. A five-phone context is used in the question design. The contexts that are likely to affect the central phone by 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 '2' is used to separate first right context from second right context. The experimental results for phonetic recognition are as follows.

SYSTEM	**ACCURACY%**
Triphone HMM (Baseline)	70.86
Overlapping-feature HMM	72.95

The test was carried out on the 1680 test files of the TIMIT corpus. There are a total 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. In our future work, we plan to apply the overlapping model obtained from English data to a set of languages including both French and Mandarin Chinese.

6. REFERENCES

[1] Abbs, J. H., *Users' Manual for the University of Wisconsin X-ray Microbeam. Madison*, WI: University of Wisconsin Waisman Center, 1987.

[2] Bird, S., *Computational Phonology: A Constraint-based Approach.* Cambridge University Press. 1995.

[3] Browman, C.P., and L. Goldstein, "Articulatory Gestures as Phonological Units". *Phonology, 6:201-251*, 1989.

[4] Church, K. W., *Phonological Parsing in Speech Recognition.* Kluwer Academic Publishers, 1987.

[5] Coleman, J., *Phonological Representations*, Cambridge University Press, 1998.

[6] Deng, L., "Autosegmental Representation of Phonological Units of Speech and Its Phonetic Interface", *Speech Communication*, 1997, 23(3):211-222.

[7] 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, 1996, pp. 37-45.

[8] 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*, 1996, 2:1007-1010.

[9] Deng, L. and H. Sameti., "Transitional Speech Units and Their Representation by the Regressive Markov States: Applications to Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, 1996, 4(4):301--306.

[10] Goldsmith, J.A., *Autosegmental and Metrical Phonology.* Blackwell. 1990.

[11] Hawkins, D. M., *Firm: Formal Inference-based Recursive Modeling, Release 2.2 User's Manual*, University of Minnesota, 1999.

[12] Jensen, J.T., *Phonology.* John Benjamins Publishing Company, 1993.

[13] Lee, C.H., F. Soong, and K. Paliwal. (eds.), *Automatic Speech and Speaker Recognition -- Advanced Topics.* Kluwer Academic, 1996.

[14] 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.

[15] Young, S., "A Review of Large-Vocabulary Continuous-Speech Recognition", *IEEE Signal Processing Magazine,* Vol. 13, No. 5, 1996, pp. 45-57.