Modeling Linguistic Features in Speech Recognition[1]

Min Tang, Stephanie Seneff and Victor W. Zue

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139 USA
{mtang, seneff, zue}@sls.lcs.mit.edu

Abstract

This paper explores a new approach to speech recognition in which sub-word units are modeled in terms of linguistic features. Specifically, we have adopted a scheme of modeling separately the *manner* and *place* of articulation for these units. A novelty of our work is the use of a generalized definition of place of articulation that enables us to map both vowels and consonants into a common linguistic space. Modeling manner and place separately also allows us to explore a multi-stage recognition architecture, in which the search space is successively reduced as more detailed models are brought in. In the 8,000 word PhoneBook isolated word telephone speech recognition task, we show that such an approach can achieve a recognition WER that is 10% better than that achieved in the best results reported in the literature. This performance gain comes with improvements in search space and computation time as well.

1. Introduction

For nearly two decades [1], speech recognition researchers have recognized the need to model sub-word units by taking into account the context in which they appear. Results have shown that significant performance improvements could be realized by utilizing context-dependent models to capture the acoustic variabilities of these sub-word units. But such performance gains are achieved with a significant cost. The more elaborate these units are, the more severe the computation and storage demands are during training/testing, and, perhaps more importantly, the more data are needed to train these models, in order to avoid sparse data problems.

To alleviate such problems, researchers have often adopted a data-driven approach of first spawning a large number of highly-specific context-dependent units (e.g., tri-phones), and then combining units with similar context using automatic clustering techniques [2, 3]. This has resulted in more robust models, especially for rare combinations, and subsequently better overall performance.

When one examines the outputs of the automatic clustering algorithms, it is often the case that members of a cluster fall along natural linguistic dimensions such as manner or place of articulation. For instance, the following diphone cluster was generated by a decision-tree-based clustering process [4]:

```
ch|tcl sh|tcl jh|tcl zh|tcl
```

This cluster contains four di-phones between consonants and the voiceless closure (/tcl/). The left contexts in this cluster, namely /ch/, /sh/, /jh/, and /zh/, all share the same place of articulation *palatal*.

These observations inspire us to consider a knowledge-driven approach, in which linguistic features are employed as the basic units for acoustic modeling. Since the natural distributions of sound units are found to maximize the contrasts induced by linguistic features [5], we expect to encounter fewer data-sparseness problems when constructing models along the organizational lines of linguistic features. Furthermore, any given phoneme can be grouped along the complementary dimensions of its distinct manner and place of articulation classes, in order to form two distinct models that capture different aspects of its acoustic manifestations.

The choice of linguistic features is more than a remedy for data insufficiency. It provides an alternative solution to some fundamental problems in speech recognition [6, 7, 8, 9]. Most notably, the phonological rules that govern the context-dependent allophonic variations can now be expressed by the underlying movement of linguistic features and can be accounted for directly in acoustic models.

Feature-based models are potentially more compact and robust and they may require less computation. This prompts us to consider a multi-stage configuration for speech recognition, such as proposed originally in [10]. Such an approach may be an attractive alternative for incorporating speech recognition capabilities on networked devices that are computationally impoverished. Thus, a feature-based first stage can be exploited as a "fast-match" in situations where computational and memory resources are limited.

2. Linguistic Features

According to non-linear phonology, the speech stream can be viewed as a sequence of "feature bundles" organized along auto-segmental tiers [11, 12]. Manner and place of articulation are two classes of the auto-segmental features, grouped together in part based on their roles in phonological rules [13]. They are attractive as classes because members of the same manner/place class usually share common acoustic properties. In the remainder of this paper, they are used interchangeably with "linguistic features" unless otherwise stated.

Manner of articulation describes primarily the nature of the speech production source. Table 1 lists the eight manner classes we have adopted for this research. These manner classes are somewhat unconventional; they are empirically chosen based on their relative acoustic differences. Thus, for example, the vowel class is divided into three distinct subclasses based on energy, duration, and dynamic movements. Similarly, we distinguish between the closure and release portions of a stop consonant.

One of the barriers to using manner/place features for speech recognition lies in the complexity surrounding place of articulation, which has traditionally been defined differently for consonants and vowels. For consonants, it is defined to be the location of the main constriction in the vocal tract during pronunciation, such as *palatal* in the previous example. The place dimension has traditionally been defined for vowels based on *tongue position* and *lip rounding*. This makes it difficult to define a set of organizational classes that can be used across the full set of phonetic units.

As a working hypothesis for simplifying the modeling requirements, we decided to group all sounds into the *same* set of place-based features. Intuitively, /iy/ and /y/ are so similar that a "palatal" place for /iy/ is well-motivated. Similar relationships hold between /er/ and /r/, /uw/ and /w/, and, arguably, between /ao/ and /l/. A place assignment for other vowels is less clear, but in the interest of simplicity, we have coerced all vowels and diphthongs to be organized into the *same* place of articulation classes as the consonants. We are using seven distinct place class assignments, as listed in Table 1. We realize that our choices cannot be fully justified on the basis of linguistic theory, but we have nonetheless adopted the position that this is a reasonable first step, and that empirical results will ultimately guide us to further refinement.

Manner:	vowel, schwa, diphthong, fric, affr, stop, plosive, nasal
Place:	labial, dental, alveolar, retroflex, palatal, glottal, velar

Table 1: *Eight manner classes and seven place classes used in this Work. Note: "fric" stands for "fricative" and "affr" is an abbreviation of "affricate."*

In this framework, each phone is considered as a bundle of two features, and acoustic models can be trained along the two parallel feature dimensions: manner and place. A small number of context-dependent acoustic models are induced from the general features, along the manner and place dimensions respectively. Data are sorted into two distinct organizational groupings to separately train manner and place models. Figure 1 gives an example of how this framework can provide more natural description of co-articulation effects in speech.
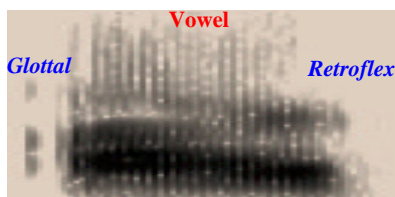

Figure 1: The strong co-articulation effect in an /aa/ + /r/ sequence in a stressed syllable is characterized by the "*vowel*" manner spreading from /aa/ to /r/. The place of articulation for this sequence transits from "*glottal*" to "*retroflex*."

3. SUMMIT Recognizer

The SUMMIT recognizer provides a probabilistic framework to decode graph-based observations [14]. We use the landmark-based framework of SUMMIT throughout our experiments. In landmark-based modeling, the acoustic observations are represented in terms of two types of landmarks: those corresponding to segment transitions and those that are segment internal. The segments are traditionally phonetic units.

In SUMMIT, various knowledge sources, e.g. phonetics, phonology, language models, etc. are pre-compiled into a single finite state transducer (FST). A set of phonological rules maps idealized phonemic forms to alternate phonetic realizations.

4. Modeling Features in SUMMIT

In this section, we report on several different system configurations that make use of manner and place models. Our interest is in designing the best method for combining the information contained in the manner class models as well as in the place class models, and exploring the most effective ways to optimize search and performance.

4.1. Feature-Based Landmark Modeling

The lack of a feature-transcribed corpus forces us to seek ways to model feature-based landmarks using the available phonetically transcribed data. We dictate manner and place values for each phone in the corpus. Feature-based landmarks can be obtained in this way through a mapping from the phone-based landmarks, as exemplified in Table 2. In the feature-based modeling, we enforce two different views toward the underlying acoustic observations, along the parallel manner and place dimensions. In each view, feature-based models are fully trained with the entire database.

Phone Landmark	Manner Landmark	Place Landmark
ch|tcl	affricate|closure	palatal|alveolar
jh|tcl	affricate|closure	palatal|alveolar
sh|tcl	fricative|closure	palatal|alveolar
zh|tcl	fricative|closure	palatal|alveolar

Table 2: *Phone and Feature-based Landmark Examples.*

The SUMMIT framework allows us to decide what feature-based landmarks to model. We can build complete feature-based models for both segment-internal boundaries and segment transitions. Or we can model feature-based landmarks only at segment boundaries, where sparse data problems become severe. The latter proves to be an efficient approach for some small vocabulary tasks, as will be discussed later.

4.2. Feature Integration

During the analysis (modeling) phase, we decompose acoustic signals into features for robust modeling. The information from different feature channels needs to be combined during recognition. We explore three different strategies to combine information from manner and place models: early, intermediate, and late integration.

Early Integration: In early integration, manner and place features are coupled into *one* recognizer. During the search, each hypothesized landmark is scored along both the manner and place dimensions, as illustrated in Figure 2. The optimal path is the one that maximizes the combined score for the two classes.
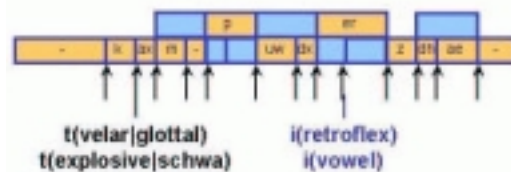

Figure 2: In early integration, each landmark is scored along both the manner and place dimensions.

Intermediate Integration: An interesting integration scheme is to model segment-internal landmarks along the manner dimension while modeling segment-transition landmarks along the place dimension, as illustrated in Figure 3. Manner and place features enforce constraints to the graph decoding algorithm at distinct landmarks. This integration scheme requires the least computation and yields significant improvement over either of the separate manner or place recognizers.
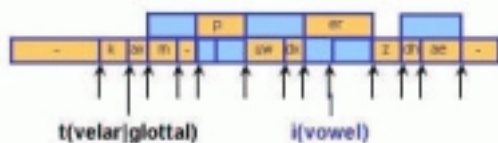

Figure 3: In the intermediate integration model, hypothesized landmarks are scored in only one feature dimension, which further reduces the computation cost.

Late Integration: Late integration, or hypothesis fusion, is a widely-used technique to reduce WER's using multiple recognizers [15]. In our experiments, we build recognizers along the manner and place dimensions. The N-best lists generated individually by manner- and place-based recognizers are integrated using a simple voting scheme.

4.3. Two-Stage Recognition

The feature-based acoustic models are more robust and compact, albeit less discriminative when the vocabulary is large. In this situation, it is ideal to use these feature-based models as a "filter" to limit the search space to a high-quality cohort so that context-dependent language modeling and/or acoustic-phonetic analysis techniques can be effectively applied [10]. We explore this venue in a two-stage configuration schematized in Figure 4. In the second stage, a detailed phonetic analysis system [16] searches within a small cohort generated by the feature-based models. This configuration is utilized in the large vocabulary task discussed in Section 5.2

Figure 4: A two-stage speech recognition system. Feature-based models in the first stage generate a high-quality cohort for detailed analysis in the second stage.

5. Results

We experiment with feature-based modeling within the context of the PhoneBook corpus [17], which contains over 80,000 read words of telephone-quality, collected over 1,000 speakers and on an 8,000-word lexicon. It also contains an independent test set of 6,598 utterances and a development set of similar size. There has been extensive prior research on this corpus reported in the literature [7, 16, 18, 19]

5.1. Results on Phonebook Small Vocabulary Task

The Phonebook small vocabulary task uses 20,000 utterances from the training set, and decodes only on the vocabulary of the test set, which contains about 600 words [18, 19, 7]. In this experiment, we choose to model feature-based landmarks only at segment transitions. We keep the phone-based segment-internal landmarks intact, since there are only a small number of them and data sparseness tends not to occur. In cases 5 and 6 in Table 3, we show the results of replacing detailed phone-based landmark models with manner- and place-based landmark models. On a small vocabulary, the loss of discriminant power due to feature-based landmarks is not serious. Best results are achieved when we combine the outputs of these two recognizers, using the weighted sum of the N-best list scores [15]. In case 7 of Table 3, the recognition error is 30% better than the nearest competitor reported in the literature (4.2% vs 3.0%)[1].

	WER
1. Hybrid HMM/ANN [18]	5.3
2. DBN [19]	5.6
3. HMM+HAMM [7]	4.2
4. Phone-based Landmark Models	3.6
5. *Transitional Manner Landmarks*	4.5
6. *Transitional Place Landmarks*	4.5
7. *5 and 6 N-best Fusion*	3.0

Table 3: *Phonebook Small Vocabulary Results for various systems. System 4 is our baseline system.*

5.2. Cohort Analysis on Phonebook Large Vocabulary Task

The Phonebook large vocabulary task [16] makes use of the entire 80,000 training utterances and decodes on the 8000-word vocabulary. Our goal on this task is to build feature-based models and apply them to prune the search space at an initial recognition stage. With a reduced search space, we can afford computation-intensive algorithms, such as context-dependent language understanding and/or acoustic-phonetic analysis, in later recognition/understanding stages. We are interested in the cohort quality, i.e. the "missing rates" – the percentage of words that fall outside the cohort – for a given cohort size, of different feature-based models and different fusion schemes. As shown in Figure 5, the cohort missing rate drops dramatically when proper information fusion techniques are applied. In particular, the early fusion scheme performs best when the cohort size is small (<50). The late fusion scheme performs extremely well when the cohort size is medium or large.
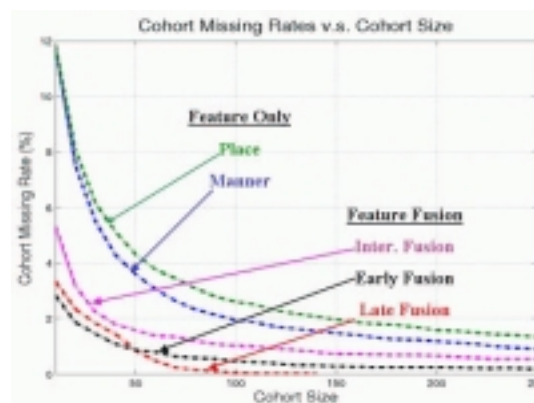

Figure 5: The cohort missing rates as a function of cohort size, for various conditions.

[1]Although only 17% better than the result we achieve using our baseline system (3.6% vs 3.0%).

5.3. Multi-stage Recognition Results

A fixed cohort size of 50 is chosen in the second stage of a two-stage recognition system, schematized in Figure 4. This corresponds to a reduction by a factor of 160 in terms of vocabulary size. The cohort is rescored using a phonetic-based system in [16]. The WER of the second stage is reduced to 8.4 as compared to 8.7 reported in [16], which is the best prior result in the literature to our knowledge.

	WER
Context-dependent Duration Modeling [16]	8.7
Two Stage System	8.4

Table 4: *Phonebook Large Vocabulary Results.*

To further improve upon this result, the second stage can be combined with the early, feature-based stages. Further reduction in WER is observed for all three different information fusion schemes employed at the first stages, as shown in Table 5.

We also explored a three-stage system: a 300-word cohort generated by the manner-based models is re-scored by the place-based models in the second stage, and a reduced 50-word cohort from the second stage is scored by the phone-based models in the final stage. Although the final result of this three-stage system is similar to that of the other two-stage systems, it is computationally more efficient, mainly because the first stage, by virtue of modeling *only* the manner class dimension, has a significantly reduced search space.

	WER
Early Integration	7.9
Intermediate Integration	8.0
Late Integration	8.0
Three Stage System	7.9

Table 5: *Recognition results for the 8,000 vocabulary experiments under different integration schemes.*

6. Summary and Future Research

In this paper, we explore the possibility to use linguistic features as basic units for acoustic modeling. We study in particular the parallel features of manner and place of articulation, and different strategies to combine them to optimize search and final performance. We have achieved improved performance using feature-based models on an isolated word task, along with a reduction in computational and memory requirements.

In the future, we will integrate this work with sub-lexical modeling [20] to build a domain-independent first-stage recognizer that supports open vocabulary continuous speech recognition. We will also study how language modeling [20] and analysis-by-synthesis techniques [12] can be applied to the cohort. We are also interested in feature-dependent front-end techniques that may further improve the performance.

7. Acknowledgment

8. References

[1] Y. Chow, M. Dunham, O. Kimball, M. Krasner, G. F. Kubula, J. Markhoul, P. Price, S. Roucos, and R. Schwartz, "BYBLOS: The BBN continuous speech recognition system," in *Proc. IEEE Int. Conf. ASSP*, 1987, pp. 89–92.

[2] K.-F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System.* Boston: Kluwer Academic Publishers, 1989.

[3] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state typing for high accuracy acoustic modelling," in *Proceedings ARPA Workshop on Human Language Technology*, 1994, pp. 307–312.

[4] C. Wang, S. Cyphers, X. Mou, J. Ponifroni, S. Seneff, J. Yi, and V. Zue, "Muxing: A telephone-access mandarin conversational system," in *Proc. ICSLP*, Beijing, P.R.China, October 2000.

[5] K. N. Stevens, *Acoustic Phonetics.* MIT Press, 1998.

[6] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, Der Technischen Fakultät der Universität Bielefeld, 1999.

[7] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models: Performance improvements and robustness to noice," in *Proc. ICSLP*, 2000.

[8] E. Eide, "Distinctive features for use in an automatic speech recognition system," in *Proc. Eurospeech*, 2001.

[9] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: Application to speech recognition," *Journal of Acoustic Society American*, vol. 111, no. 2, Feburary 2002.

[10] D. P. Huttenlocher and V. W. Zue, "A model of lexical access from partial phonetic information," in *Proc. IEEE Int. Conf. ASSP*, San Diego, CA, March 1984.

[11] J. A. Goldsmith, *Autosegmental and Metrical Phonology.* Basil Blackwell, 1989.

[12] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *Journal of Acoustic Society American*, vol. 111, no. 4, pp. 1872–1891, April 2002.

[13] E. C. Sagey, "The representation of features and relations in nonlinear phonolgy," Ph.D. dissertation, Massachusetts Institute of Technology, 1982.

[14] J. Glass, "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, p. To Appear, 2003.

[15] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, California, 1997.

[16] K. Livescu and J. R. Glass, "Segment-based recognition on the Phonebook task: Initial results and observations on duration modeling," in *Proc. Eurospeech*, 2001.

[17] J. Pitrelli, C. Fong, S. Wong, J. Splitz, and H. Leung, "Phonebook: A phonetically-rich isolated-word telephone-speech database," in *Proc. IEEE Int. Conf. ASSP*, 1995, pp. 101–104.

[18] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J. M. Boite, "Hybrid HMM/ANN systems for training independent tasks: Experiments on Phonebook and related improvements," in *Proc. IEEE Int. Conf. ASSP*, 1997.

[19] J.A.Bilmes, "Dynamic Bayesian multinets," in *Proc. 16 Conf. on Uncertainty in Artificial Intelligence*, 2000.

[20] S. Seneff, "The use of linguistic hierarchies in speech understanding," in *Proc. ICSLP*, 1998.