

Modeling Pronunciation Variation in Conversational Speech using Syntax and Discourse

Rebecca Bates, Mari Ostendorf

Electrical Engineering
University of Washington, Seattle, WA USA

[becky,mo]@ssl.i.ee.washington.edu

Abstract

A significant source of variation in spontaneous speech is due to intra-speaker pronunciation changes. Previous work in automatic speech recognition has identified several factors that affect pronunciation variability such as phonetic context and speaking rate. This work examines new higher level information sources: syntax and discourse structure, specifically the relationship between these factors and pronunciation variation as seen in reduction and hyper-articulation. Analyses of hand-labeled data are used to determine features for phone-independent variables characterizing pronunciation changes, which in turn are used in a decision-tree based dynamic pronunciation model. Pronunciation prediction experiments show a reduction in phone error rate of 10% over a baseline model using only phonetic context.

1. Introduction

A significant source of variation in spontaneous speech is due to the pronunciation changes made by an individual speaker, as illustrated by differences in word error rates for different speaking styles [1]. This work seeks to address pronunciation variation by incorporating new information sources in automatic speech recognition (ASR) systems, namely syntax and discourse structure. This work examines the relationship between these high-level factors and pronunciation variation as seen in reduction and, to a lesser extent, hyper-articulation.

The basis of the pronunciation model is a data analysis of a phonetically hand-labeled subsection of the Switchboard corpus developed at ICSI [2]. This analysis shows a relationship between word pronunciation quality and both syntactic features (part-of-speech labels) and features related to discourse structure (like dialog acts and location of word in the utterance). The ASR pronunciation model builds on previous work using decision trees [3, 4, 5], but introduces intermediate predictors of pronunciation distance and reduction to help address data sparsity issues in training with high-level features. The model is dynamic in the sense of depending on word and utterance context. Pronunciation prediction experiments are described, comparing a baseline model using only phonetic context with ones using different types of higher-level structure.

In the remainder of the paper, we begin by reviewing influential prior work in pronunciation modeling and motivation for investigating the particular conditioning factors used in this study. We then describe the corpus that this work is based on and the pronunciation distance measure used for analysis. Next, the analysis of prediction variables and pronunciation prediction results are provided, concluding with a summary of the main findings.

2. Background

2.1. Pronunciation Modeling

Particularly in the last five years, there has been a large body of work on pronunciation modeling in ASR, motivated by the challenges of modeling dialectal and spontaneous speech variation. In this section, we briefly summarize work that has influenced the approach described here.

The problem of pronunciation modeling is one of determining which pronunciations of a word to include and the relative probability of each. There are a wide variety of approaches, including selecting the most frequently observed variants in a corpus, automatic learning of word-dependent pronunciation networks for frequently observed words, and more general (but context-dependent) prediction of phone or sub-phone transformations. In this work, we follow the more general approach in order to allow multiple pronunciations for words that are unseen or rarely seen in training. There are two main ways to generalize models: define rules about phone changes and train probabilities of the rules, e.g. [6, 7, 8], or learn a probability distribution of unrestricted phone transformations and then prune to the most likely cases, e.g. [1, 5, 9, 10]. We follow this last approach, building on previous work using decision trees.

Decision tree pronunciation modeling of phone transformations involves training decision trees that take as input a phone from a pronunciation baseform and contextual features (e.g. neighboring phones, lexical stress) and provide as output a probability distribution for the possible surface realization. The trees are trained from pairs of base phones and surface realizations (which may include deletions and insertions) that are obtained by aligning baseforms with phonetically labeled data. The data may be labeled by hand or automatically labeled using a phone recognizer or forced alignment with another (presumably less refined) pronunciation model. There is typically one decision tree for each baseform phone with a probability distribution at each leaf node, so there are a large number of parameters to train. While low probability events are typically pruned, this approach is still sensitive to data sparsity, as evidenced by poor performance when only a small hand-labeled subset is used to train pronunciation models [5]. To address this problem, initial models are trained on a subset of hand-labeled data, and automatic phonetic alignments are used to make a larger data set available for training the decision trees.

Adding new pronunciations involves a trade-off between the increased chance of correctly recognizing a word uttered with that pronunciation and the added confusability with other words in the vocabulary. For example, an analysis by Jurafsky *et al.* [11] shows the many pronunciation variants of high frequency function words (where one might think pronuncia-

tion models would be most important) are often confusable with other short words that may or may not be function words. One way to reduce this problem is dynamic pronunciation modeling: condition the pronunciation probabilities on the surrounding lexical context, so that not all possible pronunciations are allowed at all times. This allows for more variability where it is pertinent, while reducing the confusability problem that comes from having too many choices. The choices can be dependent on local word context, capturing phenomena like “gonna” without the added complexity of multiwords. Dynamic modeling can also include longer range effects associated with speaking style. Fosler-Lussier and colleagues have successfully implemented dynamic modeling and shown improvements for spontaneous speech [4, 12]. Finke and Waibel performed dynamic modeling by including duration and speaking rate measures in pronunciation prediction trees [8]. In this work, we will look at introducing new factors into the decision tree feature set that would require such dynamic models.

2.2. Higher Level Information

In Lindblom’s theory of variation in speech [13], speakers adjust their articulatory effort to accommodate the listener and the importance of the information. Phonemes are hyperarticulated during points of emphasis or clarification and reduced at very predictable points. This can be seen in the reduction of such words as “to” and “of” to the point where it is difficult to associate a measurable segment duration with them. To some extent, this phenomenon can be captured by word predictability as quantified by local n-gram language model scores, which analyses show to be a useful predictor of pronunciation variability [11, 14]. In addition, we hypothesize that this variation in articulation quality can be associated with the related phenomena of information structure, including syntax and discourse structure. More broadly, factors that have been shown to improve language models may be candidates for predicting pronunciation variation.

A simplified representation of syntax that is easy to incorporate in a dynamic pronunciation (or language) model is the part-of-speech (POS) tag sequence. POS labels have been used successfully by Wakita *et al.* [15] to predict pronunciation changes in Japanese. However, they did not go beyond the POS of the particular word being recognized. The surrounding POS context has been used to improve language modeling, and anecdotal examples that we have observed (e.g. “to” in “gonna”) can be predicted by the neighboring POS labels. In addition, it is well known that POS labels are among the most important factors for predicting phrasal prominence, or pitch accent location, and prominent words are unlikely to be reduced (and may even be hyperarticulated). For purposes of this work, we use a grouping of POS tags into 8 classes that was also used in prominence prediction study for speech synthesis [16]. Classes include adjectives, nouns, cardinal numbers, adverbs, more accentable verbs (past participle verb, present participle verb), other verbs, more accentable function words (quantifiers and negatives) and other function words. We also included a separate POS classification feature that uses 5 classes aimed at conversational speech phenomena: backchannel, filled pause, content word, more accentable function word, and other function words. This division was motivated by observations of Jurafsky *et al.* [11] that planning problems – as represented by repetitions, pauses and filled pauses (“um” and “uh”) – are strongly correlated with reduction of vowels and function words.

Another simplified representation of syntax, proposed by

Meteor [17], divides an utterance into three regions: the pivot point (the main verb) and the words before and after that point. Pivot information was used in language modeling with some success [18]. Though it is based on syntax, it has been conjectured that relation to the pivot point provides some discourse information. The words before a pivot are typically “given” while following words are often “new”, in the sense of relating to something that has or has not already been introduced into the conversation.

Dialog act labels describe the intention of an utterance, an aspect of discourse. A general dialog act mark-up structure was devised by Allen and Core [19] and modified by Jurafsky *et al.* [20] for use with the Switchboard corpus of conversational speech. Conditioning on dialog acts has been shown to improve language models [18, 21], and for this reason we hypothesize that these will be useful for pronunciation modeling. Again, because of data sparsity problems, we will explore a smaller (clustered) set of dialog acts than those described in [20], specifically the frequent groups identified in [21].

3. Corpus

This work uses the Switchboard corpus, which includes spontaneous telephone conversations and is generally thought to have a great deal of pronunciation variability. Detailed descriptions of the collection methods and the corpus can be found in [22]. The corpus has been divided into training and test sets. There is a large amount of speech data available for training the acoustic models used in ASR including training data from the Callhome corpus. We use a subset of 254 hours that has some data eliminated (based on low forced alignment likelihoods) and other data eliminated to ensure independent test and training in future planned experiments. 2226 of the available 2497 Switchboard conversations have been labeled with dialog acts. A subset of 1131 have been labeled with discourse markers and disfluencies.

A four hour portion of the training set has been phonetically hand-transcribed by Greenberg *et al.* [2] at ICSI. The ICSI set is used to train the pronunciation models used in the baseline experiments, with half an hour of data held out for evaluation purposes. The held out utterances (about 10% of the phones) are from conversations used in the 1996 development test set as well as the 1997 JHU Workshop test set. The hand-transcribed corpus includes syllable times as well as phone labels.

A forced alignment between dictionary pronunciations and the ICSI phone labels using the finite state transducer tools developed at AT&T [23] was done using a phonetic feature distance (described next). The resulting alignment gives both the baseform and surface form phone sequences used in training the pronunciation model decision trees. In addition, the normalized word-level distance – as well as deletion, insertion and substitution statistics about particular phones – is used in the analyses of prediction variables.

In the ICSI subset (both training and held out sets), 26.6% of the phones do not match the baseform phone (although in the case of substitutions the surface form is most often an acceptable replacement). The relative proportion of phone errors is 59.0% substitutions, 36.9% deletions and 4.1% insertions. We hypothesize that the insertions might be used to automatically identify hyper-articulation (vs. reduction) phenomena. The most frequent insertion is the glottal stop, which is often used at word onsets of prosodically salient events [24].

4. Phonetic Distance Measure

We developed a phone distance matrix derived from articulatory features. In this section only, we use the term “features” here to describe symbolic linguistic characteristics that characterize or distinguish a specific sound or phoneme. Groups of phonemes may share features but, by definition, no two will have the same combination of features. Additionally, not all possible combinations of features are used in the set of phonemes defined for American English. This allows some flexibility in the production of a phoneme. If most, but not all, of a phoneme’s features are present, the sound is still likely to be identified as that phoneme.

The features are closely related to those described by Stevens in [25]; see the appendix for a complete list of the 22 features used here. The set of features can take on binary values described as “+” or “-”, or they can be unspecified. Some of the features that are unspecified can take on values depending on the surrounding speech context (e.g. vowels can become nasalized); others are not defined in combination with specific features (e.g. vowels cannot be + or - strident). The distance measure is a weighted sum of the distance between feature values, where more weight is given to features higher in the articulatory feature tree. A “+” feature has value 1; a “-” feature has value -1; an allowable but unspecified feature has value 0; and an undefined feature has value -2. The weights are determined heuristically, with a small amount of tuning to ensure that the biggest distances are between the phones in different manner classes. The cost of substituting one phone for another ranges from 2 (for a single low-level feature difference, such as voicing in “p” vs. “b”) to 72 for some vowel-consonant differences (e.g. “aa” vs. “t”). Deletion and insertion costs are proportional to the maximum substitution cost for a particular phone.

We use the distance measure to align the baseform pronunciations of the transcribed words with the hand-transcribed surface form and to provide a single word-level measure of pronunciation “quality”. The average total cost (or, pronunciation distance) of a word pronounced in the ICSI set is about 20, which is roughly 6.5 when normalized by the number of words in a phone. Excluding deletions, the average per phone cost is 2, so it appears that much of the distance is due to reduction phenomena (not surprisingly). Slightly less than half of the word tokens have a surface form that is identical to the baseform (zero cost).

5. Analyses of High-level Features

In this section, we look at the usefulness of different high-level factors for predicting the above length-normalized word pronunciation distance, as well as a simple classification of phone transformation in terms of hyperarticulation vs. reduction. The goal is to predict a low-dimensional pronunciation variation factor based on high-level features that can be used in combination with local phonetic and lexical stress context for pronunciation prediction.

5.1. Predicting Pronunciation Distance

In this section, we look at the relative importance of local words, syntax, and discourse cues to pronunciation variation, beginning with an illustration of the distribution of the pronunciation distance. Many have observed that function words have more pronunciation variation than content words, and indeed Figure 1 confirms this for the set of words that have non-zero cost. However, note that both word types have broad distributions, and the percentage of times that the two have non-zero cost is not so

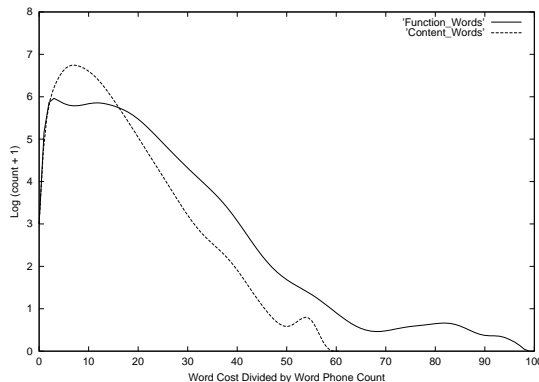


Figure 1: *Smoothed content word and function pronunciation distance histograms for tokens with non-zero cost, showing the different distributions of pronunciation quality (normalized average phone cost per word vs. log counts).*

different: 54% of function word tokens (21K words) and 51% of content word tokens (29K words).

In previous work, Fosler-Lussier showed a correlation between log trigram language model probabilities and word accuracy [14], providing support to the hypothesis that pronunciation variability is related to word predictability. Since word accuracy is only indirectly related to word variability (due to language model contributions in ASR decoding), we conducted a similar analysis using the pronunciation distance rather than word accuracy. The correlation between these variables is 0.1, and a scatter plot illustrating the relationship is given in Figure 2. While there is some information in the language model probability, especially for low probability words, this factor alone is not enough for useful prediction. Using the negative log trigram score in an analysis of variance gives a root mean-squared error (RMSE) of 9.5. For comparison, the RMSE using only the 5-class POS categorization is 9.4.

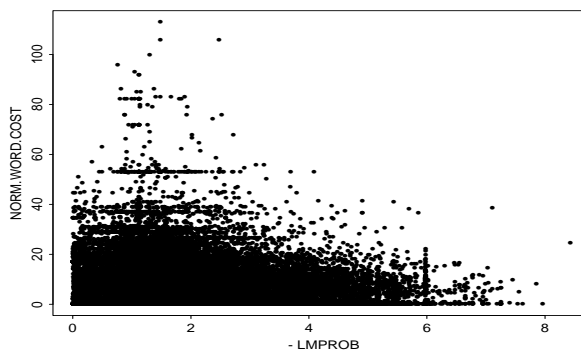


Figure 2: *Relationship between word predictability and average cost/phone ($-\log p(w_i|w_{i-1}, w_{i-2})$ vs. normalized pronunciation distance).*

Next, we investigated the relationship between different high-level factors and the pronunciation distance. The average word cost (normalized pronunciation distance) for the 9 POS groups used in this work differ as seen in Figure 3. The two groups defined by being more likely to be accented have the highest average pronunciation distance. While the pronun-

ation distance for dialog acts differs, as seen in Figure 4, the differences are smaller than that seen for the POS classes. Comparing the pronunciation distance for words before vs. after the pivot, we find an increased distance before the pivot (7.4 vs. 5.8, respectively), which is consistent with the notion that the words before the pivot tend to be “given” and thus are more likely to be reduced. The average pronunciation distance for words in utterances where there is no pivot is slightly lower: 5.3.

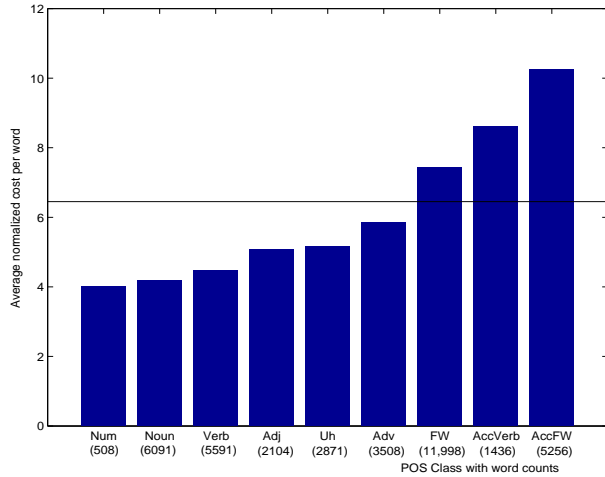


Figure 3: Relationship between part-of-speech and normalized pronunciation distance. The reference line is the average for all classes.

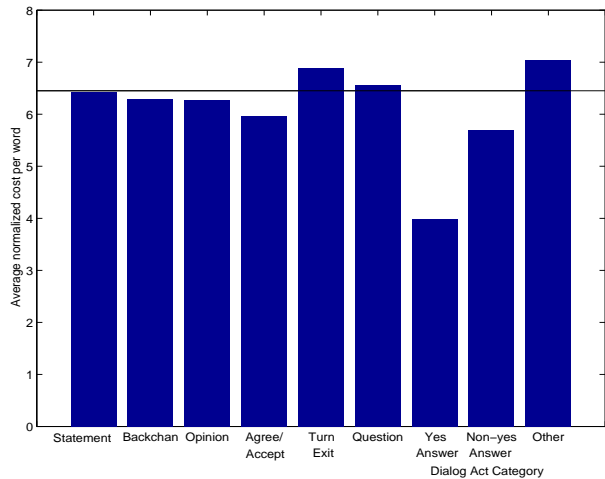


Figure 4: Relationship between dialog act label and average normalized pronunciation distance. The reference line is the average for all classes.

Finally, we looked at different syntactic and discourse cues for improving the prediction of pronunciation distance. The value predicted by the trigram GLM is used as a feature in combination with various categorical features in a regression tree. The results are summarized in Table 1. The standard deviation of the normalized pronunciation distance can be considered a baseline. For the training set, it is 9.54 and for the test set, it is 10.77. Part-of-speech is very useful in the tree. In particular, a three word window of POS tags brought down the RMSE for

both the training set and the held out set. Additional features including a three word window of content/function word (cw/fw) tags, dialog act labels, location of the word in the utterance (beginning, middle, end) and location of the word with respect to the pivot point further reduced the RMSE for the training set.

Table 1: Pronunciation distance prediction error using different features in a regression tree. RMSE = root mean squared error computed using the training set and a held out portion of the ICSI test set.

Expt	Factors	RMSE (Train)	RMSE (Held Out)
1a	trigram (glm)	9.2	11.1
1b	word POS	9.3	10.5
1c	POS window	9.2	10.5
2a	(1c) & trigram scores	9.0	10.7
2b	(1c) & trigram (glm)	8.9	10.6
2c	(1c) & dialog act	9.1	10.6
2d	(1c) & location in utt	9.1	10.6
2e	(1c) & cw/fw	9.1	10.5
2f	(1c) & cw/fw window	9.1	10.5
3	(2f) + trigram	8.9	10.6
4a	(3) + dialog acts	8.9	10.6
4b	(3) + location in utt	8.9	10.6
4c	(3) + pivot utt	8.9	10.6
5	all features	8.9	10.6

5.2. Analysis of Transformation Types

A limitation of using a single cost function is that it does not distinguish between hypo- and hyper-articulation. In an attempt to represent this difference, we characterized certain phone changes as associated with either hyper-articulation or reduction phenomena. Phone insertions and substitutions of full vowels where a reduced vowel is expected suggest hyper-articulation. Phone deletions, substitutions of flaps for /d/, /t/, or /n/, substitutions of reduced vowels for expected full vowels and feature changes such as devoicing suggest reduction. Table 2 shows the relative frequency of these different types of phone transformations, compared to the frequency that the baseform phone is used. In addition, there are other phone changes (e.g., from one full vowel to another) that could not be easily categorized as a reduction or hyper-articulation. Possibly these are associated with dialect differences. Not surprisingly, phone transformations that may be associated with reduction are the most frequent class. Hyper-articulation is relatively rare, which may reflect the speaking style and/or may be a consequence of the particular baseforms used, but in any case it appears the least important source of variability in spontaneous speech.

Table 2: The relative frequency of different types of phone transformations in the ICSI training set.

Baseform Phone	74.0%
Hyper-articulation	3.9%
Reduction	15.8%
Other	6.3%

We also looked at the hyper-articulation vs. reduction categories as a function of part-of-speech label, with the hypothesis that POS classes that are more likely to be accented would

tend to have higher percentages of hyper-articulation transformations and vice versa for POS classes that are not likely to be accented. We found that more accentable function words did indeed have a higher frequency of hyper-articulation phone transformations than other function words (9.1% vs. 5.1%, respectively), but we observed no such pattern for the more accentable verbs. (Of course, these categories were based on analyses of radio news speech, which may not hold for spontaneous speech.) The highest rate of insertions (3.1%) was for the “uh” POS class, which includes exclamations like “really,” “goodness,” and “man” as well as backchannels and filled pauses. Interestingly, the more accentable function words had a similar rate of reduction compared to the other function words (22.2% vs. 21.2%, respectively), so the more accentable function words seem to be exhibiting bimodal behavior. Both sets of function words had a substantially higher rate of reduction phenomena than any other POS type.

Using the same high-level features as in the previous section, as well as features associated with phone category and position in the word, we designed decision trees to predict 5 classes of phone transformations: deletion, reduction-related substitution, no change, hyper-articulation-related substitution or insertion, and other substitutions. The misclassification rate on a held out data set was 28.9%, which can be compared to 32.0% error for assigning all cases to the “no change” class. Even though the hyper-articulation class is infrequent, the decision tree does find contexts for predicting it.

6. Surface Form Prediction

In order to assess whether high-level features have a significant impact on the pronunciation model – and whether an intermediate predictor is useful – we conducted experiments predicting the surface form phones from the baseforms. The baseline pronunciation model was patterned after that in [5], including the following factors: position of phone in word, phone category (i.e., vowel, consonant, glide) and manner and placement categories for the left and right neighbors, dictionary stress and whether the previous phone was deleted. Several new trees were then grown incorporating higher-level features: log trigram score (or the same processed by a glm), POS tags for the word and its left and right neighbors, word location in the utterance, word location relative to the pivot, and dialog act. We compared the use of these factors directly to using an intermediate predictor (pronunciation distance or phone transformation type), and to the combination of these approaches. For the cases using the phone transformation type predictor, two features were used: the most likely transformation class and the probability difference between this and the next most likely class.

The results are summarized in Table 3. The best result represents roughly a 10% reduction in phone prediction error relative to the baseline using phonetic context and other word-internal features. The predicted pronunciation distance was more useful than the predicted phone transformation type, but neither intermediate predictor alone outperformed the case where all the features were used directly. Using both intermediate and direct features together gave the best result, but the main contribution was due to the predicted pronunciation distance.

7. Conclusions

This work shows that there are connections between high-level factors – specifically syntax and discourse – and pronunciation

Table 3: *Misclassification rates for phone transformation (base-form to surface form) prediction.*

Type of Features	Error Rate
Chance	25.2%
Baseline: phone context	21.5%
+ individual high-level features	19.7%
+ trigram glm	21.3%
+ predicted pronunciation distance	20.8%
+ phone xform type	21.3%
+ all of above	19.4%

variability, and that these factors can be used to improve the accuracy of pronunciation prediction. We have shown that the most important factors are POS tags of the word and its neighbors. In addition, various analyses support the hypothesis that words which are likely to be prominent (because of POS tag or discourse structure) tend to have fewer pronunciation changes than words which are less likely to be prominent. For this reason, we expect that improved prediction accuracy is possible by incorporating acoustic-prosodic cues into the prediction of pronunciation variation.

We hypothesized that the decision tree pronunciation model would be less sensitive to the problem of sparse data, if the various high-level factors were incorporated into a single (or low dimension) intermediate prediction value. However, it turned out that better performance was obtained by using the different prediction factors directly, although the intermediate variable was useful in combination with these factors. Of course, it may be that the intermediate variable becomes more useful as additional (prosodic) prediction factors are added to the pronunciation model.

Note that the experiments reported here only test the accuracy on the pronunciation model, and not its impact on ASR word accuracy. While we anticipate improvements, there is no guarantee of better word accuracy with an improved pronunciation model because of the interaction of many factors (acoustic model, language model, vocabulary, etc.) in determining speech recognition performance.

Acknowledgments

This work was supported in part by an Intel Ph.D. fellowship and by NSF, award number IIS-9618926. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendix: Phonetic Features

Following [25], the features are organized in a hierarchy, where those nearer the top of the tree are more important for distinguishing between larger classes of sounds and thus receive higher weight in our distance measure. At the highest level are the articulator-free, or ‘manner’, features that indicate whether the sound is a vowel, glide, consonant, or syllabic consonant – indicated using two binary-valued features. The scheme further divides consonants into types (liquids, fricatives, stops, affricates) using binary categories of sonorant, continuant and strident. Lower in the tree are the features that indicate which of the seven articulators is used in making a particular sound: the vocal folds, glottis, pharynx, soft palate, tongue body, tongue blade and the lips. At the lowest level of the tree are the articulator bound features. For the first four articulators, the interme-

diate articulator feature is redundant and not used in the distance measure. The vocal folds can be “stiff” or “slack” (combined as a single binary feature here), indicating consonant voicing. The glottis can be “spread” or “constricted” (again, a single feature); this feature is only used in English for the /h/ sound (spread) and for glottal stops (constricted). Pharynx position is described with the terms “advanced tongue root”, such as with non-back, tense vowels, and “constricted tongue root”, such as with back vowels. The soft-palate articulator is associated with the “nasal” feature. Tongue body position in the mouth is described as being either “high”, “low”, or “back”. The features associated with the tongue blade describe the position, shape and relative size of the constriction made by the tongue with the oral cavity. “Anterior” describes contact with the alveolar ridge while “distributed” shows how much of the tongue blade is involved (“+” denotes a broad portion). “Lateral” describes a tongue blade configured so that air can flow around the description and is canonically seen with /l/. If the tongue is rounded, as for /r/ sounds, it is described as “rhotic.” The position of the lips is described as +/- “round”, being + for /ow/ and - for /v/. Finally, there is a “delayed release” feature, which allows affricates to behave like both stops and fricatives. In total, we use 22 features in the distance.

8. References

- [1] M Weintraub, K Taussig, K Hunnicke-Smith, and A Snodgrass. Effect of speaking style on LVCSR performance. In *Proceedings of ICSLP*, pages S16–S19 (addendum), 1996.
- [2] S Greenberg. The Switchboard transcription project. Technical report, The Johns Hopkins University (Center for Language and Speech Processing) Summer Research Workshop, 1995. <http://www.icsi.berkeley.edu/real/stp>.
- [3] M Weintraub, S Wegmann, Y-H Kao, S Khudanpur, C Galles, E Fosler, and M Saraclar. Automatic learning of word pronunciation from data. Technical report, The Johns Hopkins University (Center for Language and Speech Processing) Summer Research Workshop, 1996.
- [4] JE Fosler-Lussier. *Dynamic Pronunciation Models For Automatic Speech Recognition*. PhD thesis, ICSI, UC Berkeley, CA, USA, 1999.
- [5] M Riley, W Byrne, M Finke, S Khudanpur, A Ljolje, J McDonough, H Nock, M Saraclar, C Wooters, and G Zavaliagos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29:209–224, 1999.
- [6] M Cohen. *Phonological Structures for Speech Recognition*. PhD thesis, Computer Science Division, Department of Electrical Engineering and Computer Science, University of California, CA, USA, 1989.
- [7] G Tajchman, E Fosler, and D Jurafsky. Building multiple pronunciation models for novel words using exploratory computational phonology. In *Proceedings of Eurospeech*, pages 2247–2250, 1995.
- [8] M Finke and A Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Proceedings of Eurospeech*, pages 2379–2382, 1997.
- [9] M Riley and A Ljolje. Automatic generation of detailed pronunciation lexicons. In Chin-Hui Lee, Frank K Soong, and K Paliwal, Kuldeep, editors, *Automatic Speech and Speaker Recognition*, chapter 1, pages 1–17. Kluwer Academic Press, 1996.
- [10] M Saraclar, H Nock, and S Khudanpur. Pronunciation modeling by sharing gaussian densities across phonetic models. *Computer Speech And Language*, 14(2):137–160, 2000.
- [11] D Jurafsky, A Bell, E Fosler-Lussier, C Girand, and W Raymond. Reduction of English function words in Switchboard. In *Proceedings of ICSLP*, pages VII–3111–3114, 1998.
- [12] E Fosler-Lussier. Multi-level decision trees for static and dynamic pronunciation models. In *Proceedings of Eurospeech*, pages 463–466, 1999.
- [13] B Lindblom. *Speech Production and Speech Modelling*, chapter Explaining Phonetic Variation: A Sketch of the H&H Theory. Kluwer Academic Publishers, 1990.
- [14] E Fosler-Lussier and N Morgan. Effects of speaking rate and word frequency on conversational pronunciations. In *Proceedings of ESCA Pronunciation Modelling Workshop*, pages 35–40, Kerkrade, The Netherlands?, 1998.
- [15] Y Wakita, H Singer, and Y Sagisaka. Multiple pronunciation dictionary using hmm-state confusion characteristics. *Computer Speech And Language*, 13:143–153, 1999.
- [16] K Ross and M Ostendorf. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language*, 10:155–185, 1996.
- [17] D Jurafsky, R Bates, N Coccaro, R Martin, M Meteer, K Ries, E Shriberg, A Stolcke, P Taylor, and C Van Ess-Dykema. SWBD discourse language modeling project final project report. Technical report, Center for Language and Speech Processing, Johns Hopkins University, 1997.
- [18] D Jurafsky, R Bates, N Coccaro, R Martin, M Meteer, K Ries, E Shriberg, A Stolcke, P Taylor, and C Van Ess-Dykema. Automatic detection of discourse structure for speech recognition In *Proceedings of IEEE Workshop on Speech Recognition and Understanding*, 1997.
- [19] J Allen and M Core. Coding dialogs with the DAMSL annotation scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, 1997.
- [20] D Jurafsky, E Shriberg, and D Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report 97-02, University of Colorado Institute of Cognitive Science, 1996.
- [21] Y Lobacheva. Discourse mixture language modeling. Master’s thesis, Boston University, 2000.
- [22] J Godfrey, E Holliman, and J McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, 1992.
- [23] M Mohri, FCN Pereira, and M Riley. Weighted finite-state transducers in speech recognition. to appear in *Computer Speech and Language*, 2002.
- [24] L Dilley, S Shattuck-Hufnagel, and M Ostendorf. Glottalization at word onsets in american english. *J. Phonetics*, 24:423–444, 1996.
- [25] K Stevens. *Acoustic Phonetics*. The MIT Press, 1998.