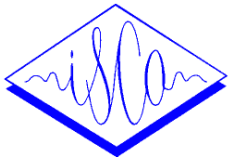
ISCA Archive
http://www.isca-speech.org/archive



ITRW on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA2002)
Aspen Lodge, Estes Park, Colorado, USA
September 14–15, 2002

# MODELING PRONUNCIATION VARIATION IN CONVERSATIONAL SPEECH USING PROSODY

*Rebecca Bates and Mari Ostendorf*

University of Washington
Electrical Engineering
Seattle, WA USA

## 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, as well as syntactic structure. This work examines prosody as a cue to pronunciation variability, as represented by attributes derived from F0, energy and duration values. Analyses of hand-labeled data are used to determine useful instances of prosodic variables for characterizing pronunciation changes, which in turn are used in a decision-tree-based dynamic pronunciation model. Experiments predicting phone changes show an improvement over chance when prosodic attributes are used. Including prosodic variables in a model using phonetic context and word-based information shows a 14% reduction in entropy and a slight improvement in phone error rate over the baseline model.

## 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 another information source in automatic speech recognition (ASR) systems, namely prosody, and examines the relationship between prosodic factors and pronunciation variation as seen in reduction and, to a lesser extent, hyper-articulation. In addition, we look at the interaction between prosodic and syntactic variables.

The basis of the pronunciation model is an analysis of a phonetically hand-labeled subsection of the Switchboard corpus developed at ICSI [2, 3]. The ASR pronunciation model builds on previous work using decision trees [4, 5, 6], but introduces intermediate predictors of pronunciation distance and phone transformation categories in an attempt to address data sparsity issues in training with combinations of prosodic and word-based attributes. We focus on the use of decision trees for analysis because of the ease of incorporation into ASR systems. The model is dynamic in the sense of depending on local word and utterance context. Pronunciation prediction experiments are described, comparing a baseline model using only phonetic and syllabic context with ones using different types of higher-level structure.

In the remainder of the paper, we review influential prior work and motivate the exploration of prosody in this study. We describe the corpus that this work is based on and the generation of the prosodic and word-based values used here. An analysis of prediction variables and pronunciation prediction results are provided. We conclude with a summary of the main findings and future directions.

## 2. BACKGROUND

### 2.1. Pronunciation Modeling

Recently, there has been a large body of work on pronunciation modeling in ASR, motivated by the challenges of modeling dialectal and spontaneous speech variation. 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. [7, 8], or learn a probability distribution of unrestricted phone transformations and then prune to the most likely cases, e.g. [4, 6, 9]. We follow this last approach, building on previous work using decision trees.

### 2.2. Prosody and Pronunciation

It is widely believed that speakers adjust their articulatory effort to accommodate the listener and the importance of the information. Phonemes are hyperarticulated during points of emphasis and reduced at very predictable points [10]. For example, such words as "to" and "of" can be reduced 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, 12]. We hypothesize that, while pronunciation is affected by many factors, it may be more directly related to prosody than word sequence characteristics.

Prosody is often represented by symbolic prosodic events consisting of prosodic phrase boundaries, prominences, and tones. While symbolic prosodic events require hand labeling, their acoustic correlates can be found automatically and incorporated into ASR systems. These include fundamental frequency (F0), both local pitch movements as well as the pitch range over a phrase or utterance, duration at the segmental level, phrase level speaking rate, energy of the speech, and articulation quality.

It should be noted that there are conflicting cues in prosody. A high value of F0 suggests emphasis while low F0 values can suggest either emphasis or mumbling. Wightman and Ostendorf [13] show that duration reflects both speaking rate as well as stress and phrase boundaries. Short durations can suggest either a fast speaking rate or reduction while long durations can suggest emphasis, phrase final lengthening or simply a slow speaking rate. For example, with phrase final lengthening, deleted phones are still possible. Any one cue is not reliable.We need to examine both local and longer term measures, as well as the interaction between F0, duration and energy in order to disambiguate their causes.

Only a few studies have used prosody in pronunciation modeling. Fosler-Lussier and collaborators have done a great deal of work developing automatic methods for measuring speaking rate and have shown that using speaking rate in conjunction with dynamic pronunciation modeling can lead to improvements in recognition accuracy [11, 5, 14]. Fundamental frequency, signal-to-noise ratio (SNR) and duration have been explored in other work with duration being the most important cue [8, 15]. While previous work has not found F0 to be useful, it may be a normalization or measurement issue. Since F0 is a cue to other structures known to influence pronunciation, e.g., syntax [16] and disfluencies [11], we hypothesize that F0 is directly useful for pronunciation modeling.

## 3. CORPUS

This work uses the Switchboard corpus, which is composed of spontaneous telephone conversations and is generally considered to have a great deal of pronunciation variability. Detailed descriptions of the collection methods and the corpus can be found in [17].

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. 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 [18] was done using a phonetic feature distance [16] as the cost function. The resulting alignment gives both the baseform and surface form phone sequences used in training the pronunciation model decision trees.

In the ICSI subset, 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 transformations is 59.0% substitutions, 36.9% deletions and 4.1% insertions. The most frequent insertion is the glottal stop, which in [19] is found to be used at word onsets of prosodically salient events

Of the 888 unique words in the held out set, 346 are seen less than three times in the ICSI training set and 211 are not seen at all. There is obviously a need for a generalizable pronunciation modeling approach. It is possible to generate optional pronunciation strings with decision trees trained on the ICSI training set for these words. For example, the word "above" is not in the training set but for use in recognition, six reasonable[1] pronunciation strings are generated:

- /ax/ /b / /ah/ /v/ (canonical)
- /ax/ /b / /ah/
- /b / /ah/
- /b / /ah/ /v/
- /ih/ /b / /ah/ /v/
- /ih/ /b / /ah/.

[1] "Reasonable" for conversational speech.

Because we have no fronted schwa in our recognition phone set, the acoustic model for /ih/ is used. Hence, /ih/ appears frequently as an option in our pronunciation model.

## 4. PREDICTION VARIABLES

The pronunciation prediction uses both acoustic and word-based factors, referred to here as "attributes". In the work presented here, we use three types of **acoustic measures:** duration, energy and F0. Duration values include length of the utterance, word and phone as well as the word duration normalized by (divided by) the utterance length and the phone duration normalized by the word duration. Energy measures include mean, minimum and maximum energy values over: the utterance, word, and windows of 15 and 30 frames prior to the end of the word. These values were also normalized by the energy at the beginning of the conversation for an SNR-type measure by dividing the values by the average energy of the first ten frames of the conversation side. In some SNR cases, log values were also used.

F0 values were generated using `get_f0`, the ESPS pitch tracker [20, 21] but were modified to be more useful in decision tree prediction using a piecewise liner model to generate a stylized curve [22]. The resulting F0 contour included values very similar to the original F0 values but with a reduction of doubling and halving artifacts. The curves fitted to the F0 values were useful in that falling and rising trends are more easily seen. The slopes of the piecewise lines and the line endpoints are used as attributes. Along with mean, minimum and maximum F0 values for the utterance, word and 15 and 30 frame windows, the slope values and the number of slope changes in both the utterance and word are used in our predictive trees. F0 values are normalized by the speaker baseline F0 in two ways; the baseline value can either be subtracted or used as a divisor. Normalizing F0 values is important to reduce speaker dependent variation.

Along with the prosodic variables, we use **word cues** explored in [16]. Word level attributes included three-word windows of part-of-speech (POS) labels (where the labels are clustered into nine groups) and content/function word tags, location of the word in the utterance (beginning, middle, end), and trigram language model probabilities. Phone level attributes include manner and placement information about the dictionary phone and stress information and location of the phone in the syllable and word.

## 5. ANALYSES OF PROSODIC VARIABLES

In this section, we look at the usefulness of prosody for predicting two intermediate variables: a pronunciation distance and 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 variables that can be used in combination with local phonetic and lexical stress context for pronunciation prediction.

**Table 1**. *Word-level attributes most correlated with pronunciation distance. Correlation of remaining attributes were between* $\pm$ *0.03.*

| Attribute | Correlation |
|---|---|
| Trigram LM Probability | 0.10 |
| Max. word energy | -0.04 |
| Min. word energy | -0.08 |
| Word duration | -0.16 |
| Word dur/Utt. dur | -0.10 |
| F0 num. changes in word | -0.11 |
| F0 num. changes in word/dur | 0.14 |

### 5.1. Predicting Pronunciation Distance

We developed a phone distance matrix derived from articulatory features which is described in detail in [16]. Along with values describing articulatory features, we take into account vowel stress and the syllable location of consonants. We use this distance measure to provide a single word-level measure of closeness to the baseform, which will be referred to as *pronunciation distance*. We define pronunciation distance as the sum of the phone distances between the baseform and surface-form phones in an instance of a word divided by the number of phones in the baseform pronunciation. The pronunciation distance – as well as deletion, insertion and substitution statistics about particular phones – is used in the analyses of prediction variables. The average pronunciation distance of a word in the ICSI set is roughly 6.5. Excluding deletions, the average per phone distance is 2, so it appears that much of the pronunciation 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 distance). In [16], we use our phonetic distance measure to show the connection between pronunciation variation and syntactic cues.

Because many of the prosodic attributes are numerical rather than categorical, we included them in generalized linear models (GLMs) based on a Gamma distribution. The attributes that stood out as being most correlated with the pronunciation distance are shown in Table 1. The pronunciation distance predicted by the GLM can be used as an attribute in combination with various categorical and numerical attributes in a regression tree. The standard deviation of the pronunciation distance can be considered a baseline. It is 9.5 for the training set and 10.8 for the held out por-

**Table 2**. *Pronunciation distance prediction error using prosody-based attributes in a regression tree. RMSE = root mean squared error, computed for the training set and a held out portion of the ICSI set.*

| Expt | Factors | RMSE (Train) | RMSE (Held Out) |
|---|---|---|---|
| 0 | Baseline | 9.5 | 10.8 |
| 1 | Duration | 9.2 | 10.9 |
| 2 | Best of F0 | 9.2 | 10.8 |
| 3 | Best of Energy | 9.1 | 11.0 |
| 4 | Energy + F0 + Duration | 8.9 | 10.5 |
| 5 | (4) + word-based | 8.7 | 10.5 |

tion. The results for intermediate trees built using prosodic attributes directly are summarized in Table 2. Intermediate trees built with the individual sets of prosody variables (Expts. 1-3) do not give improvements over the baseline (Expt. 0), but when combined (Expt. 4), they are useful. The addition of word-based variables in the tree reduces the error further on the training set but does not change the held out set result. When using the output of the GLM built with the variables from Table 1 in a decision tree, the training set error is reduced over the decision tree using the GLM built with the trigram scores alone. However, the use of the GLM-based predictors had a negative impact on the held out set results.

### 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 [10], both of which might yield phone changes that give similar distances but are associated with different prediction factors and different surface forms. 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 (3.6% of phones). Phone deletions, substitutions of flaps, substitutions of reduced vowels for expected full vowels suggest reduction (12.7%). 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 (9.5%).[2] 74.2% of phones match their dictionary baseform. Not surprisingly, phone transformations that may be associated with reduction are the most frequent class. Hyper-articulation is relatively rare, by this definition, which may reflect the nature of conversations between strangers and/or may be a consequence of the particular baseforms used, but in any case it appears to be the least important source of variability in spontaneous speech.

[2]Regional dialect, age or sociolect differences may be the source for many of these cases.

Our guiding motive is to improve speech recognition. While reducing the error rate of prediction is a goal, we also use entropy as a measure of improved output distributions of the pronunciation model. Improving the overall pronunciation model will be most useful to a recognition system. We calculate entropy over the test set by summing the entropy for each phone transformation decision. The entropy is calculated as

$$H = -\frac{1}{T}\sum_{j=1}^{T}\left(\sum_{i=1}^{N} p(\phi_i|\phi_j)\log p(\phi_i|\phi_j)\right) \qquad (1)$$

where $\phi$ is a phone, $N$ is the total number of phones that can be predicted and $T$ is the number of phones in the held out set.

Since the goal of the intermediate trees is to improve the prediction of lower level information, we present in Table 3 the results of phone transformation type prediction using the pronunciation distance, both known (as an oracle experiment) and predicted. Note that pronunciation distance itself performs as well as the use of phonetic context and significantly better in combination. Hence, improved prediction of pronunciation distance, and by extension other intermediate predictors, should help in this case as well as in surface form prediction.

**Table 3**. *Phone transformation type prediction error using predicted and known pronunciation distance with baseline phone-based attributes in a classification tree for the held out portions of the ICSI set.*

| Factor | % Error | Entropy |
|---|---|---|
| Baseline: | 31.9 | NA |
| Phone information | 30.7 | 0.70 |
| (1) + predicted pron. distance | 30.8 | 0.68 |
| (1) + oracle pron. distance | 24.5 | 0.53 |
| oracle pronunciation distance | 29.7 | 0.72 |

The best starting point for predicting transformation categories was using word-based phone and word-level variables. This combination has an 8% reduction in error on the held out set over the chance baseline. Results are presented in Table 4. While the addition of various F0 values did not further reduce the error rate, it did reduce the entropy on the held out set. Adding duration and energy did not improve the result although they were chosen as attributes in the tree. Decision trees based on prosody alone were not better than chance. The F0 values used in the tree include the number of F0 slope changes in the word (normalized by the word duration and unnormalized), the maximum and minimum F0 values of the word, the difference between the F0 slope

at the end of the word and the slope at the beginning of the next word, and the count of the frames affected by pitch halving in the word. Predicted transformation categories, along with a confidence score based on the difference between the most likely and second most likely categories, are used in the surface form prediction.

**Table 4**. *Phone transformation type prediction error using different attributes in a classification tree for the held out portion of the ICSI set.*

| Expt | Factor | % Error | Entropy |
|---|---|---|---|
| 0 | Baseline: | 31.9 | NA |
| 1 | Word variables | 29.4 | 0.67 |
| 2 | (1) + F0 | 29.5 | 0.66 |
| 3 | (1) + F0, energy | 29.6 | 0.67 |
| 4 | (1) + F0, duration, energy | 29.4 | 0.68 |

The prediction results for pronunciation distance and the transformation categories do not seem to support each other, in that the prosodic variables are useful for pronunciation distance but not transformation type. However, it should be noted that they are predicting different levels of information. One is a word-level value and the other is at the phone level. The prosodic attributes used in prediction are generally at the word-level or higher so may be better suited to predicting pronunciation distance, which can then be used as a useful predictor for transformation category and surface form changes.

## 6. SURFACE FORM PREDICTION

In order to assess whether prosodic variables have a significant impact on the pronunciation model we conducted experiments predicting the surface form phones from the baseforms. Decision trees were built using individual sets of prosodic attributes and pruned using a cost-complexity parameter of k=5 in cross-validation on the training data. The results on the held out set are given in terms of percentage error as well as entropy as summarized in Table 5. Again, because the goal is to include the probability distributions in a speech recognition system, we look for a reduction in entropy to show improved output distributions. While adding the word-based attributes actually raises the error rate, there is an 18% reduction in entropy. Word-based information is more powerful than prosody alone, but Expt. 4 shows that both may help as entropy is low and the error rate is reduced. Including the intermediate predictors had an entropy value 0.56 but resulted in an error rate of 31.0%. When using the true values of the intermediate predictors for pronunciation distance and the transformation category with the word and phone-based attributes (no prosody), the error rate was significantly lower (a 56% reduction in error) and the entropy was reduced by 72%. This strongly suggests the use and further improvement of the intermediate predictors for baseform to surface form prediction.

**Table 5**. *Misclassification rates for phone transformation (baseform to surface form) prediction on the held out set.*

| Expt | Factors | % Error | Entropy |
|---|---|---|---|
| 0 | Chance | 33.0 | NA |
| 1 | Phone context | 30.8 | 0.66 |
| 2 | (1) + word-based | 30.9 | 0.54 |
| 3 | (1) + prosody | 30.8 | 0.58 |
| 4 | (2) + prosody | 30.7 | 0.57 |
| 5 | (2) + intermediates | 31.0 | 0.56 |
| 6 | (2) + oracle intermediates | 13.7 | 0.15 |

The individual attributes used most often in the baseform to surface form decision trees in Expt. 3 are duration and energy features. These are chosen more often than attributes describing phonetic context. As a group, F0 values are used most often in tree questions. Trees for some baseform phones are built solely of questions based on prosodic attributes. When word information is included as in Expt. 5, prosodic attributes are still chosen often but questions about the part-of-speech window are amongst the most common.

## 7. SUMMARY AND FUTURE WORK

While word-based variables are most useful in predicting pronunciation distance and surface form changes, prosody does improve the intermediate predictor models and has an effect on surface-form prediction. In particular, word-level duration and energy values improve the language-model-based GLM reported in [16]. All prosodic variables contributed to surface-form prediction. Though oracle experiments with intermediate prediction values gave significant gains, the specific realization here did not realize this potential. The next step with this work will be to incorporate the prosody-based pronunciation models into an ASR system.

**Acknowledgments**

This work was supported in part by an Intel Ph.D. fellowship and by the NSF, award number IIS-9618926. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The software for generating fitted F0 values was created at SRI by Harry Bratt, Kemal Sönmez and Andreas Stolcke who were supported by NSF, award STIMULATE IRI-9619921, and NASA, award NCC 2-1256.

## 8. REFERENCES


[1] M Weintraub, K Taussig, K Hunicke-Smith, and A Snodgrass. Effect of speaking style on LVCSR performance. In *Proc. of ICSLP*, pages S16–S19 (addendum), 1996.

[2] S Greenberg. The Switchboard transcription project. Technical report, The Johns Hopkins University (CLSP) Summer Research Workshop, 1995. `http://www.icsi.berkeley.edu/real/stp`.

[3] S Greenberg. Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29:159–176, 1999.

[4] M Weintraub and *et al.* Automatic learning of word pronunciation from data. Technical report, The Johns Hopkins University (CLSP) Summer Research Workshop, 1996.

[5] JE Fosler-Lussier. *Dynamic Pronunciation Models for Automatic Speech Recognition*. PhD thesis, University of California,Berkeley, CA, USA, 1999.

[6] M Riley and *et al.* Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29:209–224, 1999.

[7] G Tajchman, E Fosler, and D Jurafsky. Building multiple pronunciation models for novel words using exploratory computational phonology. In *Proc. of Eurospeech*, pages 2247–2250, 1995.

[8] M Finke and A Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Proc. of Eurospeech*, pages 2379–2382, 1997.

[9] 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.

[10] B Lindblom. *Speech Production and Speech Modelling*, chapter Explaining Phonetic Variation: A Sketch of the H&H Theory, pages 403–439. Kluwer Academic Puflishers, 1990.

[11] D Jurafsky, A Bell, E Fosler-Lussier, C Girand, and W Raymond. Reduction of English function words in Switchboard. In *Proc. of ICSLP*, pages VII–3111–3114, 1998.

[12] E Fosler-Lussier and N Morgan. Effects of speaking rate and word frequency on conversational pronunciations. In *Proc. of ESCA Pronunciation Modelling Workshop*, pages 35–40, Kerkrade, The Netherlands, 1998.

[13] C Wightman and M Ostendorf. Automatic labeling of prosodic patterns. *IEEE Trans. on Speech and Audio Processing*, 2(4):469–481, 1994.

[14] N Morgan and E Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proc. of ICASSP*, pages II–729–732, 1998.

[15] M Ostendorf and *et al.* Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. Technical Report ECE-97-0002, Boston University, 1997.

[16] R Bates and M Ostendorf. Modeling pronunciation variation in conversational speech using syntax and discourse. In *Proc. of the Workshop on Prosody in Speech Recognition and Understanding*, pages 17–22, 2001.

[17] J Godfrey, E Holliman, and J McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proc. of ICASSP*, pages 517–520, 1992.

[18] M Mohri, FCN Pereira, and M Riley. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16(1):69–88, 2002.

[19] L Dilley, S Shattuck-Hufnagel, and M Ostendorf. Glottalization at word onsets in american english. *J. Phonetics*, 24:423–444, 1996.

[20] Entropic Research Laboratory. *ESPS Version 5.0 Programs Manual*. StatSci, 1993.

[21] D Talkin. Pitch tracking. In W Kleijn and K Paliwal, editors, *Speech Coding and Synthesis*. Elsevier Science B.V., 1995.

[22] K Sonmez, E Shriberg, L Heck, and M Weintraub. Modeling dynamic prosodic variation for speaker verification. In *Proc. of ICSLP*, Sydney, Australia, 1998.