# MULTI-LEVEL DECISION TREES FOR STATIC AND DYNAMIC PRONUNCIATION MODELS

*Eric Fosler-Lussier*

University of California, Berkeley, and International Computer Science Institute
1947 Center Street, Berkeley, CA, 94704-1198, USA
fosler@icsi.berkeley.edu

## ABSTRACT

We have been focusing on improving pronunciation models for automatic transcription of television and radio news reports by modeling phone, syllable, and word pronunciation distributions with decision trees. These models were employed in two separate sets of experiments. First, decision trees facilitated selection of word pronunciations derived automatically from data for use in a standard speech recognizer dictionary. We have seen a small but significant improvement with these automatically constructed dictionaries in our one-pass decoding system. In a second set of experiments, we allowed decision tree models to determine the probability of word pronunciations dynamically, dependent on the linguistic context of the word during recognition. Dynamic models provided an additional insignificant decrease in error, but improvements were focused within the spontaneous speech portion of the test set.

## 1. INTRODUCTION

One goal of recent research within the ASR community has been to provide systems with better pronunciation models, particularly in hopes of improving performance on spontaneous speech tasks such as Switchboard and Call Home. We are working with the Broadcast News (BN) database, a collection of news reports and interviews in both planned and spontaneous focus conditions. In this study, we examine the effects of pronunciation modeling across the focus conditions in this corpus.

Our approach, like that of many others, is to automatically derive new pronunciation models using the acoustic models of our existing recognizer. In essence, we allow the acoustic models to suggest new candidate baseforms via phone recognition. It is debatable, however, whether such a source of pronunciations is linguistically defensible. Phonetic transcriptions from these systems often do not match linguistic expectations; thus, a system that uses hand-transcriptions (*e.g.*, [10]) as a seed may perform better. On the other hand, from an engineering standpoint, pronunciation models are the interface between acoustic models and word sequences: if acoustic conditions (*e.g.*, telephone versus studio recording environments) induce regular variations in the output of acoustic models, then it may be useful to capture these in the pronunciation model. This implies the dependence of an automatically-derived dictionary on a particular acoustic model.

Since pronunciations suggested by a phone recognizer are often linguistically "noisy," one can use the overall statistics of the variations in phone recognition to filter these automatic transcriptions, discarding anomalous events. *Smoothed phone recognition* [13] attempts to the generalize the generation of new pronunciations, using classifiers (*e.g.*, decision trees [1] or neural networks [8]) to constrain alternative pronunciations proposed by the phone recognizer.

In a *static* dictionary, pronunciation probabilities are fixed before recognition. However, a derived pronunciation may not be applicable in all contexts. In fast American English, for example, instead of the canonical pronunciation for *interesting*, `[ih n t er eh s t ih ng]`, the reduced pronunciation `[ih n axr s t en]` may be observed more frequently; in slower speaking rates the latter pronunciation may not be observed at all. Recognizers should *dynamically* change pronunciation models based on the linguistic context [11]. Previous examples of dynamic pronunciation models [13, 10] have determined pronunciations by modeling phone variations; here we study models of syllable and word pronunciations, incorporating an expanded set of context factors that influence pronunciation transcriptions [5].

This paper focuses on several dimensions of automatic pronunciation learning. First (and foremost), we examine differences between the static and dynamic evaluation of induced dictionaries. We also investigate model performance with changes in the amount of training data and associated acoustic models. Finally, for dynamic dictionary rescoring, we study effects of modeling syllables versus words, as well as the effects of employing different sets of contextual features.

## 2. STATIC DICTIONARIES

The first goal of our study was to build a good static dictionary, both for use in the SPRACH Broadcast News system [2], and also to ensure a fair comparison against any dynamic techniques. In addition, dynamic dictionaries are used to rescore lattices or $n$-best lists of hypotheses; constructing lattices or $n$-best lists requires the best possible static dictionary in a first decoding pass.

In this section, we describe experiments comparing the 1996 ABBOT system 65K vocabulary dictionary [3] to an automatically augmented dictionary used in the SPRACH system. Since automatic pronunciation techniques depend on the amount of data available and the acoustic models used, we also provide comparisons to a new static dictionary (BN97+98) induced from twice the training data and improved acoustic models.

### 2.1. Dictionary construction

The algorithm for smoothed phone recognition has been thoroughly documented elsewhere ([13, 6], *inter alia*); an outline of the algorithm is shown in Figure 1. To learn the variation between canonical phones (represented by black circles) and alternatives provided by the phone recognizer (mixed black and grey circles) we use decision trees (d-trees). In the experiments here, the context provided to the d-trees consisted of the identity, articulatory manner and place, and syllabic position of each baseform phone and its immediate neighbors. The d-trees were then used to generate pronunciation networks to be aligned with acoustic models, producing a smoothed transcription.

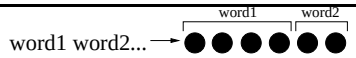
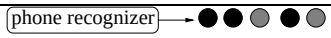
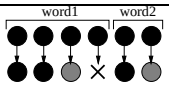
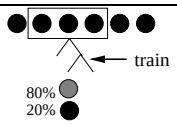
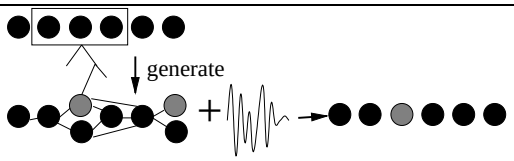
| | | |
|---|---|---|
| Step 1: | Generate canonical transcription with baseline dictionary |   |
| Step 2: | Generate alternative transcription with phone recognition |   |
| Step 3: | Align canonical and alternative transcriptions |   |
| Step 4: | Train decision trees to predict $P(\text{alternative}\|\text{canonical}, \text{context})$<br>Smooth by pruning low probability events from leaves |   |
| Step 5: | For each utterance in training set:<br>5a. Create network of pronunciation variations using trees<br>5b. Align with acoustic models to get smoothed transcription |   |
| Step 6: | Collect smoothed pronunciation variants into new dictionary; interpolate with baseline dictionary for smoothing. | |

Figure 1: Algorithm for generation of new dictionaries using smoothed phone recognition

In our initial experiments [6], we used an acoustic model from an intermediate stage in the SPRACH system's development, the best model available at the time. We combined the 1997 ABBOT PLP-based recurrent neural network (RNN) context-independent phone classifier with a 4,000 hidden unit multi-layer perceptron (MLP) using modulation-filtered spectrogram (MSG) features. Both networks were trained only on the 1997 BN training data. We refer to this combined acoustic model as A-Model I. To build the dictionary for the SPRACH system, we used A-Model I acoustics to perform smoothed phone recognition on the 100-hour 1997 BN training set.

Since the 1998 evaluation, we have retrained the pronunciation models using an improved acoustic model (A-Model II) that combines a PLP-based RNN, and two 8,000 hidden unit MLPs, trained on PLP and MSG features, respectively. All 200 hours of the 1997 and 1998 BN training sets were transcribed using the smooth phone recognition procedure; the resulting dictionary is labeled "BN97+98 training" in these studies.

### 2.2. Results I: Changing the acoustic model

To encourage fast testing, most of our experiments use a half-hour subset of the 1997 Broadcast News Evaluation test set (labeled Hub-4E-97-Subset). In Table 1, the left-hand column of results describe the experiment reported in [6]: the augmented SPRACH98 dictionary outperforms the baseline ABBOT96 dictionary. While this difference is not large, we have found that it is consistent across different test conditions.

One such change in test conditions was the inclusion of the improved acoustic model (A-Model II), as well as a change in decoder. The experiments with A-Model I also utilized the NOWAY time-synchronous stack decoder from our colleagues in Sheffield [9]; the results in the right column use the CHRONOS "time-first" decoder from our partners at Softsound [12][1]. The improvement from the SPRACH dictionary (0.6%) is unchanged with the new acoustic model and decoder. The automatically derived pronunciation model is at least *somewhat* independent of the acoustic models from which they were derived; therefore, one does not have to retrain the pronunciation models every time the acoustic models are changed. Nonetheless, the dictionaries are probably still dependent on the corpus and overall recognition system.

| | Acoustic Model/Decoder | |
|---|---|---|
| | A-Model I | A-Model II |
| Dictionary | NOWAY | CHRONOS |
| ABBOT96 (baseline) | 27.5 | 24.0 |
| SPRACH98 (BN97 training) | 26.9 | 23.4 |
| BN97+98 training | - | 23.2 |

Table 1: Word Error Rates for Hub4E-97-Subset

### 2.3. Results II: Adding pronunciation training data

Doubling the amount of pronunciation training data appears to have a very small effect. Comparing the last row of Table 1 (BN97+98 training) to the SPRACH98 results shows only a 0.2% absolute gain. We used these three dictionaries to decode the full 1997 Hub4E evaluation set using CHRONOS and A-Model II (Table 2), and found that the improvement pattern for the full set is very similar to that of the subset, with BN97+98 only just edging out SPRACH98. The improvement over the ABBOT96 dictionary is significant at $p < 0.05$.

It is interesting to note the focus conditions for which improvements are shown in the full evaluation set. Comparing ABBOT96 to BN97+98, we see decreases in word error rate in almost every category (except F3, speech with background music), leading to a significant improvement overall. Most of this gain was obtained from training with the first half of the data; between SPRACH98 and BN97+98, no difference was seen in the planned and spontaneous studio focus conditions (F0 and F1), from which the majority of the test data is drawn. The BN97+98 dictionary was marginally better in most of the non-studio conditions (F2-FX), although these differences are not statistically significant. In summary, both derived dictionaries helped in all focus conditions; the additional training data may have just captured variation in the acoustic model caused by more difficult acoustic environments.

[1] We have used two different decoders in our experiments because the CHRONOS decoder is an order of magnitude faster than NOWAY , but has the shortcoming of only producing a single best hypothesis, not a lattice of hypotheses. The CHRONOS decoder tends to outperform NOWAY by a few tenths of a percent with the particular parameter settings we are using.

| | | Focus Conditions | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dictionary | Overall | F0 | F1 | F2 | F3 | F4 | F5 | FX |
| ABBOT96 (baseline) | 23.0 | 14.6 | 24.4 | 31.8 | 31.3 | 27.0 | 22.9 | 35.4 |
| SPRACH98 (BN97 training) | 22.4 | 14.2 | 23.6 | 31.4 | 30.7 | 25.3 | 23.2 | 35.1 |
| BN97+98 training | 22.3 | 14.2 | 23.6 | 30.9 | 31.3 | 25.1 | 22.4 | 35.1 |

Table 2: Word Error Rates for Hub4E-97, using A-Model II. The focus conditions for Broadcast News include Planned Studio Speech (F0), Spontaneous Studio Speech (F1), Speech Over Telephone Channels (F2), Speech in the Presence of Background Music (F3), Speech Under Degraded Acoustic Conditions (F4), Speech from Non-Native Speakers (F5), All Other Speech (FX).

## 3. DYNAMIC PRONUNCIATION RESCORING

We have shown previously [5] that pronunciations in the Switchboard corpus depend heavily on other factors in addition to phonetic context. In particular, the frequency of a word influences the extent to which reduction processes are correlated with speaking rate: more frequent words have more variation at high rates of speech. Syllabic structure also plays an important part in determining which phones are more likely to vary; coda consonants are much more likely to be non-canonical than onset consonants. Others ([4], *inter alia*) have found that modeling tuples of words (*multiwords*) has a beneficial effect.

These studies suggest that an orientation towards larger linguistic units (*i.e.*, syllables or words) may prove beneficial in pronunciation modeling. This is easy to implement in our paradigm; instead of d-trees modeling one phone each, they now model one syllable or one word each. One can integrate more forms of context in syllable or word trees; for instance, in a phone tree, the identity of a neighboring word probably has little meaning.

### 3.1. Building syllable and word trees

In our initial experiments, we built 550 word models (BN97 word trees) from the smoothed transcriptions obtained by aligning A-Model I to the 1997 training set, as in the training of the SPRACH98 dictionary (Figure 1, step 5b). The word d-trees used the phonetic features from Section 2.1 and the surrounding word identities as a set of primary *segmental context* features. Additional context features included word length, several estimates of speaking rate, and the trigram probability of the word. Slightly less than half of the trees in each case used a distribution other than the prior (*i.e.*, were grown to more than one leaf).

We also trained roughly 800 d-trees based on syllable distributions (BN97 syllable trees). Each word was given a single canonical syllable transcription, so that words with similar syllabic-internal pronunciation variations in the ABBOT96 dictionary shared the same syllable model. In addition to the features found in the word trees, syllabic tree context features included the lexical stress of the syllable, position within the word, and the word's identity.

As in our static dictionary experiments, when A-Model II became available we regenerated both sets of trees using the 1997 and 1998 training sets (BN97+98 trees), providing 1300 syllable and 920 word classifiers. We also trained a separate set of trees on the segmental context features alone, to determine the influence of secondary features such as speaking rate.

| Dictionary | 100-best | lattice |
|---|---|---|
| SPRACH98 (baseline) | 26.7% | 27.0% |
| BN97 Word trees | 26.5% | 26.6% |
| BN97 Syllable trees | 26.3% | 26.4% |

Table 3: Hub4E-97-Subset Word Error Rates for dynamic rescoring of tree models using A-Model I.

### 3.2. Results III: Lattice versus N-best rescoring

Starting from A-Model I and the SPRACH98 dictionary, we generated lattices for the Hub4E-97-Subset test set, and also decoded the lattices into a list of the 100 best hypotheses for each segment. $N$-best lists were particularly easy to rescore in our paradigm: each hypothesis was expanded into a pronunciation graph, and then aligned to the acoustics (as in Figure 1, step 5), resulting in a new acoustic score for the hypothesis. The hypotheses were then re-ranked after adding in the language model scores.

We also integrated the dynamic pronunciation model earlier in the search by rescoring lattices [7]. To do this, a decoder must determine the pronunciations of words on-the-fly. We implemented an acoustic-rescoring lattice decoder (JOSÉ[2]); the search algorithm is a typical stack-based lattice decoder, with one minor difference. In a regular time-synchronous stack decoder, hypotheses (word sequences) are extended by a word at every time step; each new hypothesis is inserted into a stack corresponding to its particular end time. In JOSÉ, before inserting an hypothesis extension, we rescore the penultimate word with the dynamic pronunciation model, using the word sequence (and associated phonetic and syllabic information) as the context for the d-trees.

In both paradigms, we found that averaging the rescoring model with the original lattice acoustic score in a multistream-like approach improved results. We did not tune the combination parameter, but instead weighted each acoustic score evenly.

Since different decoders are used in each paradigm, we tested the influence of the decoding process by recomputing the baseline scores. The $n$-best decoder and lattice decoder were run with the SPRACH98 static dictionary as a calibration (Table 3, line 1). The results in both cases were similar to those of the first-pass decoding (26.9% word error rate (WER)).

As Table 3 shows, the dynamic decoding of trees gave us an insignificant increase in accuracy over our improved static dictionary, with syllable trees performing the best. The difference between lattice decoding and $n$-best rescoring seems to be minimal in this test. As opposed to the across-the-board improvements seen with the SPRACH98 static dictionary in most focus conditions, we found that the 0.4% difference between $n$-best decoding of the baseline and the syllable trees was accounted for almost completely by a 1.4% improvement in WER in the spontaneous broadcast speech focus condition, and a 0.9% improvement for speech with background music.

### 3.3. Results IV: Changing the acoustic model

We regenerated $n$-best lists for Hub4E-97-Subset using A-Model II and the BN97+98 static dictionary, and rescored the lists using both the BN97 and BN97+98 trees. Table 4 shows that none of the trees made a significant difference in performance. However, we can see some general trends across the experiments: first, BN97 trees performed worse than both the baseline and

[2] So named because NOWAY produces lattices for it beforehand.

| Dictionary | | Overall WER | F0 | F1 | F2 | F3 | F4 | F5 | FX |
|---|---|---|---|---|---|---|---|---|---|
| Static: | BN97+98 (baseline) | 23.6 | 13.5 | 23.3 | 34.5 | 29.2 | 26.6 | 16.8 | **44.4** |
| Word trees: | BN97 | 24.0 | 13.5 | 25.2 | 34.7 | 26.9 | 27.1 | 17.6 | 45.0 |
| | BN97+98 All Features | 23.4 | **13.2** | 23.0 | 34.6 | 27.8 | 26.8 | 17.6 | 44.6 |
| | BN97+98 Segmental Context | **23.3** | 13.5 | **22.4** | 34.4 | **27.2** | **26.2** | 17.6 | 44.8 |
| Syllable trees: | BN97 | 24.1 | 13.4 | 23.5 | 36.7 | 29.5 | 27.1 | **16.0** | 45.7 |
| | BN97+98 All Features | 24.0 | 13.5 | 24.2 | 34.6 | 28.9 | 27.2 | 19.3 | 46.1 |
| | BN97+98 Segmental Context | 23.5 | 13.5 | 22.8 | **33.9** | 28.1 | 27.1 | **16.0** | 45.4 |

Table 4: Hub4E-97-Subset WER for dynamic evaluation of tree models using A-Model II.

BN97+98 trees. This suggests that dynamic models may be more susceptible to changes in the acoustic model, since the BN97 trees were trained using A-Model I. Also, in a reversal of our earlier experiment, the BN97+98 word trees outperformed the syllable trees. The increase in training data, which allowed for greater coverage of the corpus, may have contributed to this result.

When non-segmental features like speaking rate and trigram probability were removed from the trees, performance improved. We have highlighted the lowest error rate in each focus condition across all seven experiments; the lowest error in five of seven focus conditions occured with a segmental context model. Our measures of speaking rate and word predictability may not be robust enough for use in a dynamic model.

Finally, as in our initial experiments, the modest improvements of the dynamic models (*e.g.*, BN97+98 segmental word trees) were concentrated in the non-F0 portions of the corpus, although, with the lack of statistical significance, we do not wish to make any strong claims.

## 4. CONCLUSIONS

We have experimented with static and dynamic pronunciation models that use decision trees at various representational levels. Automatically learned static dictionaries from decision tree smoothed phone recognition gave roughly 3% relative improvement on the Broadcast News task; this result was used in the SPRACH Broadcast News system for the 1998 Hub4E evaluation. The learned static dictionaries seem to be somewhat robust to changes in decoding conditions. Experiments using additional training data showed diminished returns.

In our experiments with dynamic evaluation of syllable and word tree models, small gains were seen over static dictionaries; these models may be capturing some of the pronunciation variation in the spontaneous portion of the BN corpus. Unlike static dictionaries, performance of these models are more dependent on the acoustic model used. Questions remain about the robustness of speaking rate and word predictability features, as trees only using segmental context outperformed trees using all features. In future work, we plan to investigate the effectiveness of individual features in hopes of improving the dynamic model.

## 5. ACKNOWLEDGMENTS

The author gratefully thanks Nelson Morgan, Gary Cook, Dan Ellis, Dan Gildea, Adam Janin, Brian Kingsbury, and Gethin Williams for advice and system support. This work was supported by NSF grant IRI-9712579.

## 6. REFERENCES

[1] F. Chen. Identification of contextual factors for pronounciation networks. In *IEEE ICASSP-90*, pages 753–756, 1990.

[2] G. Cook, J. Christie, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson, and G. Williams. The SPRACH system for the transcription of broadcast news. In *DARPA Broadcast News Workshop*, Herndon, Virginia, February 1999.

[3] G. Cook, D. Kershaw, J. Christie, and A. Robinson. Transcription of broadcast television and radio news: The 1996 ABBOT system. In *DARPA Speech Recognition Workshop*, Chantilly, Virginia, February 1997.

[4] M. Finke and A. Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Eurospeech-97*, 1997.

[5] E. Fosler-Lussier and N. Morgan. Effects of speaking rate and word frequency on conversational pronunciations. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 35–40, Kerkrade, Netherlands, April 1998.

[6] E. Fosler-Lussier and G. Williams. Not just what, but also when: Guided automatic pronunciation modeling for broadcast news. In *DARPA Broadcast News Workshop*, Herndon, Virginia, March 1999.

[7] J. E. Fosler-Lussier. *Dynamic Pronunciation Models for Automatic Speech Recognition*. Ph.D. thesis, University of California, Berkeley, 1999. Forthcoming.

[8] T. Fukada, T. Yoshimura, and Y. Sagisaka. Automatic generation of multiple pronunciations based on neural networks. *Speech Communication* 27:63–73, 1999.

[9] S. Renals and M. Hochberg. Efficient search using posterior phone probability estimators. In *IEEE ICASSP-95*, pages 596–599, Detroit, MI, 1995.

[10] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 109–116, Kerkrade, Netherlands, April 1998.

[11] M. Riley, F. Pereira, and E. Chung. Lazy transducer composition: a flexible method for on-the-fly expansion of context dependent grammar networks. In *IEEE Automatic Speech Recognition Workshop*, pages 139–140, Snowbird, UT, December 1995.

[12] T. Robinson and J. Christie. Time-first search for large vocabulary speech recognition. In *IEEE ICASSP-98*, pages 829–832, Seattle, WA, May 1998.

[13] M. Weintraub, E. Fosler, C. Galles, Y.-H. Kao, S. Khudanpur, M. Saraclar, and S. Wegmann. WS96 project report: Automatic learning of word pronunciation from data. In F. Jelinek, editor, *1996 LVCSR Summer Research Workshop Technical Reports*, chapter 3. Center for Language and Speech Processing, Johns Hopkins University, April 1997.