

CONTEXTUAL WORD AND SYLLABLE PRONUNCIATION MODELS

Eric Fosler-Lussier

International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704-1198
Tel: (510) 643-9153, Fax: (510) 643-7684, Email: fosler@icsi.berkeley.edu

ABSTRACT

This work focuses on the evaluation of models of syllable and word pronunciations constructed automatically using the Broadcast News corpus of radio and television news reports. Previous work [4] introduced the concept of extended-length decision tree models; here I report on ASR-independent assessment of these models. This study also discusses integration of static and dynamic pronunciation evaluation using the ROVER algorithm for combining hypotheses, and details the improvements of dynamic pronunciation evaluation on the 1998 DARPA Broadcast News test set. The new pronunciation models improve system robustness for speech that is not pre-planned and recorded under studio conditions; these models appear to represent both linguistic variation (as in spontaneous speech) and variation due to channel effects in telephone-bandwidth speech.

1. INTRODUCTION

Recent studies have shown that appropriate pronunciation models for speech recognition systems are critical for good performance in large-vocabulary tasks, particularly when the speaking style is spontaneous [5]. One popular approach to pronunciation modeling is to use decision trees to automatically learn patterns of variation within automatically- or hand-generated transcriptions.

Most decision tree based systems model pronunciations on a phone-by-phone basis. Each baseform phone is associated with a decision tree that predicts how the phone is realized in context. During recognition, the appropriate decision tree leaf for a given context determines a small piece of a finite state grammar (FSG), which is concatenated with other phone grammar fragments into an FSG for the entire utterance.

One problem with this technique of naïve concatenation is that the choice of pronunciations for each phone is independent of all other phones. For example, in the word *baseball*, the final vowel [ah] can be realized as [eɪ] if the final [l] is deleted; if phone realizations are considered independently, the unlikely pronunciation [b eɪ l] may result for the final syllable. One solution is to include a dependence on the previous decision tree output, as suggested by Riley [6], which improves the predictive power of the trees. Weintraub et al. [8] added phone *n*-gram constraints to the FSGs using a maximum entropy model; this extra information degraded recognizer performance significantly in initial experiments, although these results were not conclusive.

This work continues a strategy of modeling the distributions of phone pronunciations jointly at the syllable and word levels [4]. This longer-term modeling captures many of the coordinated phone pronunciation variations not handled by independent phone trees. Since phones at syllable boundaries still vary with context,

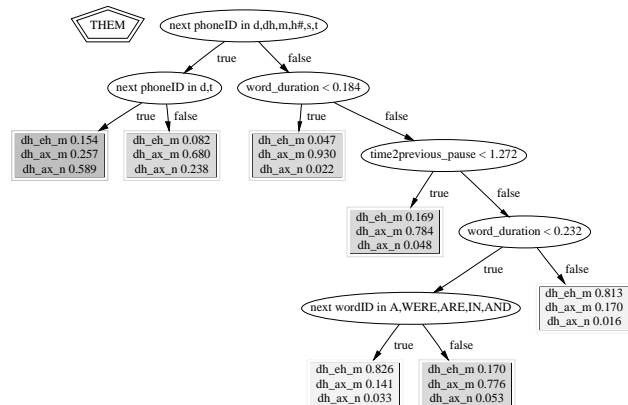


Figure 1: Word decision tree for *them*.

pronunciations in these models include dependencies on the neighboring baseform phone. Other forms of context, such as word identity, speaking rate, and word predictability, are also included in the model. These models are employed in the rescoring of *n*-best lists, dynamically choosing appropriate pronunciations based on hypothesis context. The work described here extends previous studies by including independent evaluation of word and syllable decision trees, as well as integrating hypotheses from static and dynamic dictionaries using ROVER [3].

2. DECISION TREE MODELS

To train decision trees to predict word pronunciations, a pronunciation training set is needed, consisting of a phone recognition transcript aligned to canonical dictionary pronunciation models [4]; each instance of a word in the corpus is therefore annotated with its pronunciation suggested by the phone recognizer. One tree is constructed for each word, using context cues such as neighboring word or phone identity to predict pronunciations. For example, the pronunciations [dh eh m], [dh ax m], and [dh ax n] were often suggested for the word *them* by the phone recognizer. Since the pronunciation [dh ax n] can be confused with the word *than*, perhaps ASR performance will improve if the dictionary dynamically restricts the contexts in which this pronunciation can represent *them*; as the top decision in Figure 1 shows, this pronunciation is favored when the next word starts with dental consonants.

Building separate decision tree models for each word has the drawback that only words with enough training data can be modeled, whereas with phone trees one can model every phone in the corpus. A way to increase coverage is to use syllable models, so that words like *baseball* and *football* can share pronunciation

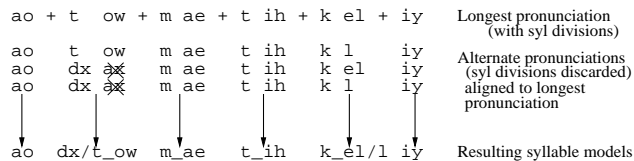


Figure 2: Selection of syllable models for the word *automatically*. Alternative pronunciations are aligned to the longest pronunciation. Reduced phones ($[ax]$) and phone deletions are eliminated if unreduced variants exist across variants; similar phones are clustered together.

models for their shared syllable. Determining appropriate syllable models for each word is non-trivial, however. Ideally, multiple pronunciations of a syllable should be incorporated into the syllable model. For instance, the word *some* has two pronunciations in the baseline dictionary, $[s ah m]$ and $[s ax m]$, but the variation between these alternatives should be provided by one syllable model $[s_ah_m]$. An algorithm to determine syllable models from a baseline dictionary is shown operating on the word *automatically* (which can be pronounced with five or six syllables) in Figure 2.

Initial experiments used the 1997 Broadcast News training set as the source of pronunciations for the word and syllable trees. The training set was phonetically transcribed automatically by means of smoothed phone recognition [4] using a combination of neural network acoustic models.

550 word models were constructed from the the 1997 training set (BN97 word trees). The word d-trees included phonetic, word identity, speaking rate, and predictability features to select appropriate pronunciation distributions. 800 d-trees based on syllable distributions were also trained (BN97 syllable trees). In addition to the features found in the word trees, syllabic tree context features included the lexical stress of the syllable, its position within the word, and the word’s identity.

3. EVALUATING D-TREES

In order to judge the quality of trees constructed with different subsets of context features, I extended the measurement paradigm of Riley et al. [7], in which the average log (base 2) probability of a held-out test set is calculated, giving a measurement related to the perplexity. This score can be obtained by filtering the test set down through the trees to the leaves; as each sample reaches a leaf, its probability according to the leaf distribution is recorded.

The average log probability is problematic as a metric for evaluating pronunciation models. Some test examples receive zero probability from the pronunciation model; this makes the measure unusable, as $\log_2(0) = \infty$. In pronunciation modeling, test transcriptions can occur that are not covered by the model due to both the nature of statistical modeling and pronunciation pruning at decision tree leaves; also, in syllable and word d-trees (unlike phone trees) not every syllable or word is modeled due to lack of training data. Disallowing zero probabilities by assigning a minimum probability does not match the way models are used within an ASR system, as each word has a finite number of baseforms. One can ignore “out-of-vocabulary” pronunciations and compute the log probability, but this metric does not penalize OOV test set errors made by systems that heavily prune pronunciation dictionaries.

Features	No pruning		Prune < 0.1	
	Avg \log_2 Prob.	Pron. Coverage	Avg \log_2 Prob.	Pron. Coverage
<i>Word trees (550 trees, 58.9% word coverage):</i>				
1. None (baseline)	-0.70	92.6%	-0.53	89.4%
2. Word context only	-0.65	92.6%	-0.47	89.6%
3. Word and phone context	-0.55	92.6%	-0.33	88.7%
4. All	-0.45	92.6%	-0.26	89.4%
<i>Syllable trees (800 trees, 78.5% syllable coverage):</i>				
5. None (baseline)	-0.70	96.0%	-0.46	91.7%
6. Word/syllable/phone context	-0.49	96.0%	-0.26	92.2%
7. All	-0.44	96.0%	-0.21	92.1%
<i>Phone trees (89 trees, 78.5% syllable coverage):</i>				
8. None	-1.45	95.0%	-0.96	84.5%
9. Phone context	-0.60	95.0%	-0.33	90.4%
10. All	-0.54	95.0%	-0.25	90.4%

Table 1: Word and syllable test set probabilities. Unmodeled segments are not included in totals. Phone tree probabilities were combined to form syllable pronunciations, and were scored on the same subset of syllables used to score syllable trees.

To address these issues, three statistics were compiled: the average log probability for baseform pronunciations receiving non-zero probabilities, the percentage of evaluated baseforms included in the scoring (labeled “pron. coverage”)¹, and the percentages of words or syllables in the test set that are actually modeled. This paradigm allowed testing of pronunciation models under the assumption of pruning within the ASR system. In unpruned models, pronunciation coverage remains the same no matter what features are used, but when pruning is invoked, the coverage varies depending on which pronunciations are eliminated at each tree leaf.

3.1. Word trees

Table 1 (lines 1–4) shows the \log_2 probability of a held-out part of the BN97 training set determined by word d-trees with different context features. In the baseline model (1), the pronunciation probabilities were set to the prior training set distributions. This model corresponds to a (simple) automatic baseform learning scheme. Comparing the unpruned to the pruned coverage numbers, roughly 3% of pronunciations in the test corpus had probabilities of 0.1 or less according to the prior model. Two metrics exist for calibrating improvement from this baseline model: increase in the log probability and in the pronunciation coverage for the pruned model.

Including just the word context (corresponding to a multi-word model, line 2) only increases average log-likelihood by 0.05. A bigger gain comes from adding in the surrounding phone context (3); using all of the features (4), including speaking rate, trigram probabilities, and durations, gives the best gain (35% improvement). The percentage gain is remarkably similar to that of Riley et al.[7]; they found relative gains of 20% to 32% depending on the training data. Yet one must be careful in comparing these results: Riley’s team was testing phone models on hand-transcribed data, whereas I am working with word models on automatically transcribed data.

When pruning is invoked, larger percentage gains result; the trees using all features show a 51% improvement. This means

¹Infrequent pronunciations are removed during d-tree construction.

that, on average, it is the pronunciations with higher probabilities ($p > 0.1$) in the baseline model that are increasing in likelihood due to the contextual modeling. The actual percentage of test pronunciations that have a probability above 0.1 does not change significantly with the increased context.

3.2. Syllable trees

The 800 BN97 syllable trees covered the test set more completely (79% compared to 59% word coverage for the word trees). The relative gains of the syllable models were a little higher than those for the word d-trees² (cf. line 5 and 7), reaching 37% for unpruned and 54% for pruned models. The real gain, however, was in pronunciation coverage: 9% of the pronunciations lost in pruning the baseline model were recovered under the d-tree models. The non-segmental features did not improve the model as much as in the word trees (cf. line 6 and 7). The increase in \log_2 probability is only about half of that seen when these features are included in word tree construction.

To compare syllable models with the more conventional phone-based methods, the syllable training set was broken into phones, from which phone models were trained (8–10). Since the syllable models contained variants (e.g., the syllable [k_l_ow_s/z] has encoded the fact that the final phone can alternate as [s] or [z]), this would give them an advantage over regular phone models. Therefore, separate trees were built for the phone variants listed in the syllable models, e.g., the final segment of [k_l_ow_s/z] was modeled by the phone [s/z]. The phone trees were then scored only on the syllable level, where pronunciations for the syllable were determined by concatenating the individual phone pronunciations from each tree; syllable pronunciation probabilities were obtained by multiplying together the phone probabilities. The subset of test syllables modeled by the syllable trees were used for scoring these models.

Without context (8), phone trees exhibit a large decrease in log likelihood compared to the syllable baseline (a relative difference of -107%). Adding contextual elements (9–10), the phone trees perform only a little bit worse than syllable trees, although the pronunciation coverage is significantly worse for both the unpruned and pruned cases. Syllable trees utilizing only segmental features outperform the phone d-trees with all features at their disposal. Thus, it seems that syllable models are as good, if not better, than phone models as an organizational structure for modeling the variation in pronunciations, although this is not a completely fair comparison because the previous output context is not used as a d-tree feature (cf. Riley [6]). Syllable models have the drawback of less coverage overall; one can model the entire corpus with phone models, but with syllable models, coverage will be incomplete.

4. IMPLEMENTATION IN THE ASR SYSTEM

In another set of experiments, I trained 920 word and 1300 syllable trees on the 1997 and 1998 Broadcast News training sets. For comparative purposes, the baseline static dictionary was taken from the 1996 ABBOT system [1]; a new static dictionary was also re-estimated from the same training set used for the word and syllable tree building (BN97+98 Static Dictionary) [4]. Several parameters for the dynamic trees were tuned on a 173-utterance subset of the 1997 Broadcast News Evaluation Set. A search was conducted

²It is important to compare relative increases in \log_2 probability, and not actual probabilities, as the test sets have different coverages.

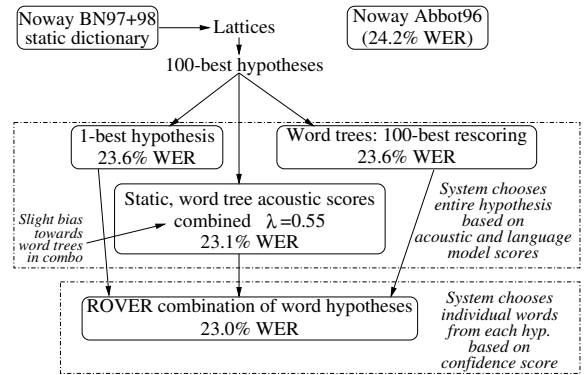


Figure 3: Results of tuning the combination of word-based dynamic pronunciation models with static dictionary components for 1997 Hub4E Subset.

for both syllable and word trees over the factors included in the tree learning algorithm, the pruning threshold at d-tree leaves, and the interpolation parameter λ between acoustic scores provided by evaluation of static and dynamic dictionaries on n -best lists.

Figure 3 shows the entire processing chain for word d-trees. Lattices were constructed by the NOWAY recognizer provided by Sheffield University, using the BN97+98 Static Dictionary. From these lattices, the 100 best hypotheses were derived with a 1-best hypothesis error rate of 23.6%. The best word d-trees from the optimization process proved to be the word and phone context trees, in contrast to the \log_2 metric results, where extra-segmental features such as speaking rate and duration improved modeling. Without interpolation, the dynamic pronunciation model had the same error rate as the static 1-best hypothesis. However, interpolating the two scores brought a 0.5% absolute error reduction.

Even though the word trees had the same baseline error rate as the 1-best hypothesis, an examination of the actual word sequences produced by each system found that the word hypotheses were often different. In these cases, it is often advantageous to combine hypotheses at the word level. The ROVER system from NIST [3] blends hypotheses annotated with confidence scores. I integrated the first-best hypothesis, the best hypothesis from the word d-trees without acoustic score interpolation, and the best post-interpolation hypothesis; each word in all three hypotheses was annotated with a posterior acoustic confidence score. The resulting word error rate was very similar to the interpolated acoustic score result (23.0%). Despite the small gain, I suspected that using ROVER in this way would provide robustness in recognition on independent test sets. Similar results were obtained using the syllable trees (23.1% WER).

4.1. Evaluation on 1998 Broadcast News test set

The best word and syllable d-tree system, as well as the ABBOT96 and BN97+98 static dictionaries, were evaluated on the 1998 Broadcast News (Hub4E-98) test set (Table 2). The static dictionary provides most of the improvement (0.6% overall, $\rho=0.031$); including the word d-trees increases the improvement to 0.9% ($\rho=0.014$ compared to ABBOT96). Compared to the BN97+98 dictionary, word d-trees give a small improvement, whereas syllable d-trees show no improvement. As hypothesized, ROVER does improve performance; when the word d-trees are evaluated independent of the ROVER combination, the word error rate is 21.7% in the unin-

Dictionary	Overall WER (%)	Focus condition WER (%)							Gender WER (%)	
		F0	F1	F2	F3	F4	F5	FX	Female	Male
[Word Count]	[32435]	[9944]	[6246]	[1095]	[1385]	[9142]	[235]	[4388]	[13160]	[19247]
ABBOT96	22.0	13.9	25.5	35.4	27.2	20.7	27.7	32.4	22.0	21.8
Static BN97+98	21.4	14.0	24.9	32.1	26.3	19.8	26.8	32.2	21.4	21.3
Dynamic word trees	21.1	14.1	24.5	31.1	25.8	19.7	24.3	31.2	21.0	21.1
Dynamic syllable trees	21.4	14.2	24.9	31.1	26.0	19.8	25.1	31.4	21.3	21.2

Focus conditions: 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)

Table 2: Categorical word error rate for Hub4E-98

terpolated case, and 21.4% in the interpolated case.

Table 2 also shows the word error rates for the focus conditions defined in the Broadcast News corpus, as well as separate error rates for female and male speech. The new static and dynamic pronunciation models never help in the planned speech condition (F0). For studio spontaneous speech (F1), word trees almost double the static dictionary’s performance increase over ABBOT96 (0.6% to 1.0%). For the other focus conditions, the dynamic word trees almost always seem to improve performance, the only exception being in the degraded acoustics condition (F4). The biggest absolute performance increases for the word trees were in the difficult F5 and FX conditions, although neither gain is significant.

The most impressive combined static/dynamic performance, though, is for the telephone speech condition (F2): the automatically derived dictionaries were 12% better (relative) in this condition. Even with the smaller test set size this is a significant difference ($p=0.016$). This may be due to an interaction of the automatically derived pronunciation models with the acoustic model: one of the three neural net models was trained on 8kHz (telephone) bandwidth speech using modulation-filtered spectrogram (MSG) features. The automatically-learned dictionaries may reflect the improved acoustic modeling for this focus condition. There are no significant patterns in word error rate due to gender. Dynamic rescoring with syllable trees was almost always worse than rescoring with word trees when compared across focus conditions; this is surprising since evaluation of earlier models with the \log_2 metric suggested the opposite. This may be due to the doubling of training data between the two experiments, but other experiments not reported here have suggested that word error rate and log probability improvements do not always go hand-in-hand.

5. CONCLUSIONS

This work describes advances in decision tree models of syllable and word pronunciations. Perplexity-like evaluations of d-tree models indicate that incorporation of context, both segmental and extra-segmental (e.g., speaking rate), does improve model quality; syllable models appear to have the best coverage and performance under this metric. However, not all of these gains transfer to word error rate improvements: for example, syllable d-tree models performed worse than word models in the ASR system, and speaking rate and word predictability measures were found to decrease recognition performance in these experiments. This latter fact contrasts with the work of Finke and Waibel [2], who found that speaking-mode related factors did improve their phonological rule based models. The relationship of these factors to top-down (phonological rule) versus bottom-up (automatic decision tree) systems is an interesting direction for future study.

The automatically learned pronunciation model presented here appears to be capturing some linguistic variation in spontaneous speech (as shown by improvements in the F1 focus condition) as well as non-linguistic variation in the acoustic models due to channel conditions (demonstrated by improvements in the telephone condition). Since the pronunciation model is the interface between the acoustic and language models, the best improvements may result from modeling both top-down linguistic variability and bottom-up acoustic variability.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

- [1] G.D. Cook *et al.* Transcription of broadcast television and radio news: The 1996 ABBOT system. In *DARPA Speech Recognition Workshop*, Chantilly, Virginia, February 1997.
- [2] M. Finke and A. Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Eurospeech-97*, Rhodes, 1997.
- [3] J. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, 1997.
- [4] E. Fosler-Lussier. Multi-level decision trees for static and dynamic pronunciation models. In *Eurospeech-99*, Budapest, 1999.
- [5] D. McAllaster, L. Gillick, F. Scatone, and M. Newman. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. In *ICSLP-98*, pages 1847–1850, Sydney, 1998.
- [6] M. Riley. A statistical model for generating pronunciation networks. In *IEEE ICASSP-91*, pages 737–740, 1991.
- [7] M. Riley *et al.* Stochastic pronunciation modelling from hand-labelled phonetic corpora. In *ETRW on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 109–116, Kerkrade, Netherlands, April 1998.
- [8] M. Weintraub *et al.* 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.