STOCHASTIC PRONUNCIATION MODELLING FROM HAND-LABELLED PHONETIC CORPORA

M. Riley[1], *W. Byrne*[2], *M. Finke*[3], *S. Khudanpur*[2], *A. Ljolje*[1], *J. McDonough*[2], *H. Nock*[4], *M. Saraclar*[3], *C. Wooters*[5], *G. Zavaliagkos*[6]

AT&T Labs – Research, Florham Park, NJ, USA[1]
Johns Hopkins University, Baltimore, MD, USA[2]
Carnegie-Mellon University, Pittsburgh, PA, USA[3]
Cambridge University Engineering Department, Cambridge, UK[4]
U.S. Department of Defense, Fort Meade, MD, USA[5]
BBN, Cambridge, MA, USA[6]

ABSTRACT

In the early '90s, the availability of the TIMIT read-speech phonetically transcribed corpus led to work at AT&T on the automatic inference of pronunciation variation. This work, briefly summarized here, used stochastic decisions trees trained on phonetic and linguistic features, and was applied to the DARPA North American Business News read-speech ASR task.

More recently, the ICSI spontaneous-speech phonetically transcribed corpus was collected at the behest of the 1996 and 1997 LVCSR Summer Workshops held at Johns Hopkins University. A 1997 workshop (WS97) group focused on pronunciation inference from this corpus for application to the DoD Switchboard spontaneous telephone speech ASR task. We describe several approaches taken there. These include (1) one analogous to the AT&T approach, (2) one, inspired by work at WS96 and CMU, that involved adding pronunciation variants of a sequence of one or more words ('multiwords') in the corpus (with corpus-derived probabilities) into the ASR lexicon, and (1+2) a hybrid approach in which a decision-tree model was used to automatically phonetically transcribe a much larger speech corpus than ICSI and then the multiword approach was used to construct an ASR recognition pronunciation lexicon.

1. INTRODUCTION

Most speech recognition systems rely on pronouncing dictionaries that contain few alternate pronunciations for most words. In natural speech, however, words seldom adhere to their citation forms. The failure of ASR systems to capture this important source of variability is potentially a significant source for recognition errors, particularly in spontaneous, conversational speech. We report methods used to address this issue applied to read speech at AT&T [9] and to spontaneous speech at and after WS97, the Fifth LVCSR Summer Workshop, held at Johns Hopkins University, Baltimore, in July-August, 1997 [2].

As a first step towards alleviating this common limitation of pronouncing dictionaries, we identify a systematic way of generating alternate pronunciations of words by using phonetically labelled portions of the TIMIT [5] and Switchboard [6] corpora. One viewpoint we explore is that pronunciation variability may be modelled by a statistical mapping from canonical pronunciations (baseforms) to symbolic surface forms, and we use decision trees to capture this mapping. A second way we exploit the hand transcriptions is by enhancing the dictionary using frequently seen pronunciations. While the former has the potential to generalize to unseen words and pronunciations, the latter is more conservative and hence potentially more robust.

As many researchers have observed earlier, simply adding several alternate pronunciations to the dictionary increases the confusability of words to the extent that the gains from having them are often more than nullified. We address this problem in two ways. We assign costs to alternate pronunciations so that, *e.g.*, if a frequent pronunciation of "cause" and an infrequent pronunciation of "because" are identical, a penalty is incurred to attribute the pronunciation to "because" rather than "cause." In addition, we account for context effects so that, *e.g.*, "to" is allowed the pronunciation [`ax`], which is a frequent pronunciation of "a," only if "to" is preceded by "going," as in [`g aa n ax`].

Our pronunciation modelling efforts may be divided into two broad categories. In our *tree based dictionary expansion* experiments, we apply decision tree based pronunciation models to entries in our baseform dictionary to obtain alternate pronunciations, which are then used in testing. In our *explicit dictionary expansion* experiments, we apply the decision tree based pronunciation models first to the training corpus, and perform a forced alignment with the acoustic models to "choose" amongst the alternatives. The dictionary is then explicitly augmented with novel pronunciations which occur sufficiently often. The tree based expansion implicitly adds many more new pronunciations than the explicit expansion. However, it does not attempt to model any cross-word coarticulation. The explicit expansion does so by allowing as dictionary entries a select set (cf. [4]) of *multiwords* – word pairs and triples.

We demonstrate in Sections 2 and 3 that the tree-based method gives a reduction in the word error rate (WER) for the read-speech North American Business (NAB) News task while both methods give reductions for the conversational telephone speech Switchboard task over baseline systems using only a citation-form dictionary. Further, we show in Sections 4 and 5 that reductions persist when the baseline systems are improved by coarticulation sensitive acoustic modelling and improved language modelling.

2. TREE BASED DICTIONARY EXPANSION

Our tree based pronunciation models were inspired by phonological rules in acoustic phonetic studies (cf., *e.g.*, [7]) which characterize allophonic variations in certain phonemic contexts, and by the successful use of similar methods to model pronunciation variability and constraints by other researchers (*e.g.*, [3, 8, 4, 10, 12, 11]). Figure 1 illustrates the deletion or alteration of a phoneme in context which we modelled via decision trees.

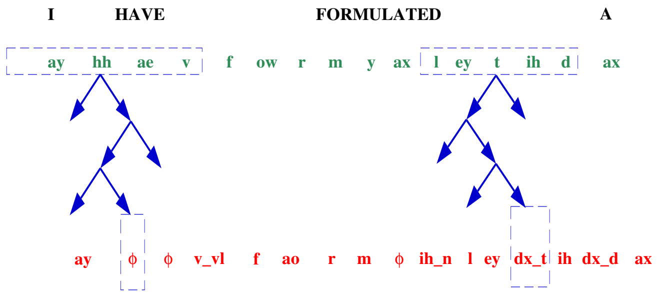

Figure 1: Decision Trees as Phone Predictors

PHONEME	PHONE	WORD
f	f	for
ao	ao	
r	-	
dh	dh	the
ax	iy	
f	f	for
ao	-	
r	-	
dh	dh	the
ax	ax	
d	jh	drug
r	r	
ah	ah	
g	g	

Table 1: Phoneme-to-phone Alignment.

2.1. Decision Trees from Hand Labelled Data

The decision trees built for these tasks drew from a substantial portion (134,000 phones) of the TIMIT data set and approximately 3.5 hours (96,000 phones) of the phonetically labelled transcriptions of the Switchboard (ICSI) data set.

The labels used by the annotators were slightly richer than the phonetic labels in the pronouncing dictionary (PronLex) used for these experiments. However, since the acoustic models for the baseline system were trained using the PronLex phone set, the hand annotations were mapped down to this phone set for reasons of consistency.

Next, based on the orthographic transcriptions and the pronouncing dictionary, a phonemic transcription of the utterances was obtained. Whenever the dictionary permitted more than one pronunciation for a word, a choice was made via a forced alignment of the acoustic signal with the alternatives using the baseline acoustic models.

The phonemic transcriptions were then lined up with the phonetic labels, using as the alignment criterion the minimization of the phonetic feature distance between the two symbol streams [9]. Table 1 gives an example alignment from the ICSI corpus. This gave us a corpus of *phoneme-to-phone* transformations together with the phonemic environment or *context* for each instance.

Decision tree models were then built to represent this phoneme to phone mapping. The context included the identity of the phoneme to be mapped as well as three neighbouring phonemes on either side (each encoded in terms of its phonetic features – see Table 2), the lexical stress on neighbouring vowels as obtained from the pronouncing dictionary, and the distance of the phoneme from the nearest segment boundary on either side, where the segment could be either a syllable, a word or a phrase. A separate tree was grown for each phoneme. The tree growing criterion was minimization of the empirical entropy of the surface phone, the stopping criterion was a minimum sample count at both parent and child nodes, and the trees were pruned via internal fivefold cross-validation.

- **Consonant-Manner:**
  voiced stop, unvoiced stop, voiced fricative, unvoiced fricative, voiced affricate, unvoiced affricate, nasal, rhotic, lateral, not–applicable.
- **Consonant-Place:**
  bilabial, labiodental, dental, alveolar, palatal, velar, pharyngeal, not–applicable.
- **Vowel-Manner:**
  monophthong, r–colored vowel, w–diphthong, y–diphthong, glide, not–applicable .
- **Vowel-Place :**
  front–low, front–mid–low, front–high, central–mid–low, central–mid–high, back–low, back–mid–low, back–mid–high, back–high, not–applicable.

Table 2: Phoneme Encoding Scheme. Each phoneme is represented as a four-element feature vector, `(consonant-manner, consonant-place, vowel-manner, vowel-place)`. For example, `/s/` is encoded as `(voiceless fricative, palatal, n/a, n/a)` and `/iy/` is encoded as `(n/a, n/a, y-diphthong, high-front)`.

Training Data	Average $\log_2$-prob (Efficiency)	
	ICSI-test	TIMIT-test
TIMIT	$0.76 \rightarrow 0.60$ (20%)	$0.34 \rightarrow 0.17$ (51%)
ICSI	$0.72 \rightarrow 0.50$ (30%)	
ICSI+TIMIT	$0.71 \rightarrow 0.48$ (32%)	

Table 3: Prediction Entropy for Pronunciation Trees

2.1.1. Predicting Surface Forms from Baseforms

Each leaf in a tree thus assigned probabilities in some context to more than one surface form realization of the phoneme it modelled. A way to judge the goodness of these trees, therefore, was to apply them as predictors on a held out portion of the hand labelled corpora. Test sets from TIMIT and ICSI corpora were held out for this purpose. The results in Table 3 summarize the predictive ability of the trees on these sets.[1] Relative to the context independent distribution of surface form realizations of a phoneme, decision trees built on the TIMIT portion of the trainig set reduce the entropy by about 50%, when tested on TIMIT. Those built on the Switchboard portion of the training set reduce the entropy by about 30%, when tested on Switchboard. Trees based on TIMIT alone are much less effective on the Switchboard test set (20%), but adding them to the Switchboard training data (ICSI+TIMIT) results in a small additional gain (32%). These results suggest there is more variability in pronunciations in Switchboard, relative to TIMIT, which is not captured by either the phonemic context cues or the modelling paradigm we considered.

2.1.2. The Effect of Leaving Out Features

In order to investigate the conditional utility of each of our contextual features given the others, trees were built at WS97 by leaving features out from the context one at a time. Table 4 summarizes the results of these experiments. The trees were trained on all of the Switchboard and TIMIT data mentioned above, and the test set was the same as the one used for the Switchboard results of Table 3. Note that, at least for this corpus size, there was little additional predictive power in neighbouring phonemes more than one position away, when the triphone context, word boundary, and lexical stress related information was specified.

2.1.3. The Impact of Some Additional Features

We also experimented at WS97 with adding new features to the decision trees.

- Based on the number of distinct pronunciations of a word that were seen in the ICSI-portion of the corpus, words were categorized into ten bins: from words having many pronunciations to words having few pronunciations. The bin number was then provided to the trees for each phoneme of the word. It was hoped that knowing how stable a word's pronunciation was would help predict the surface form better.
- It is varyingly conjectured that frequently used words, function words or low information bearing words often tend to be mispronounced. The frequency of occurrence of a word in the 60 hour acoustic training corpus was provided to the trees for each phoneme of the word.
- In the hope of capturing limited phonotactics, as well as to indirectly model deletion or reduction of units larger than phonemes, the trees were provided the surface form realization of the previous phoneme.

[1] So that test observations in contexts unseen in training do not make entropy figures infinite, the worst 10% of the test data (i.e., highest $\log_2$-prob) is removed from each entropy measurement in this paper.

Features Provided as Context	$\log_2$-prob
All Features	0.485
$\curvearrowleft$ 3rd Phoneme $\curvearrowright$ Excluded	0.484
$\curvearrowleft$ 2nd and 3rd Phonemes $\curvearrowright$ Excluded	0.485
Lexical Stress Excluded	0.487
Segment Boundary Cue Excluded	0.490
Vowels (manner and place) Excluded	0.497
Stress and Segment Boundary Cues Excluded	0.498
Consonants (manner, place) Excluded	0.527
Right Phonemic Context Excluded	0.537
Left Phonemic Context Excluded	0.547
Entire Phonemic Context Excluded	0.606
All Context Excluded (root trees)	0.714

Table 4: Leaving Out Features from the Context

Features Added to Context	$\log_2$-prob
None	0.485
Word level Pronunciation Variance (10 bins)	0.481
Word Frequency (from 60 hr training)	0.483
Pron. Variance and Word Frequency	0.483
Pron. Var., Word Freq., and Previous Phone	0.451

Table 5: Adding New Features to the Context

Table 5 summarizes our results. All the features used by the ICSI+ TIMIT trees of Table 3 are already present in the context. Note that while we were unable to successfully exploit the information about empirical pronunciation variability or frequency of a word, knowing the previous surface form seems to be of significant value in this modelling paradigm, perhaps because it compensates for some of our conditional independence assumptions in modelling the phoneme to phone mapping very locally.

2.2. Generating Automatic Phone Transcriptions

Using decision trees is a data intensive modelling technique. Large quantities of automatic phonetic transcriptions were generated to augment the hand-labelled corpora using the 37,000 training sentences (SI-284 training data set) for the NAB task at AT&T and using the 60-hour acoustic training corpus for the Switchboard task at WS97. Unlike [11], where unconstrained phone recognition was used to generate phone transcriptions, we constrained the words in our training utterances to assume only pronunciations generated by application of the decision trees to their phonemic baseforms: a forced alignment was performed on the resulting network of alternate pronunciations in an utterance and the most likely sequence of

Model	# Trn tokens	$\log_2$-prob
ICSI+TIMIT	96,040	0.525
Recount Weights	2.36 million	0.585
Rebuild Trees	2.36 million	0.542

Table 6: Rebuilding v/s Retuning the Pronunciation Trees

Dictionary	WER
TTS	12.7%
TIMIT	10.8%
Retrained2	10.0%

Table 7: NAB recognition results with Enhanced Dictionaries

pronunciations was chosen to be the phonetic transcription. Pronunciation probabilities derived from the trees were used as 'language model' weights during alignment, and since the word transcription was provided, word level language model weights are redundant and were not used.

Anecdotal evidence suggests that this method of obtaining automatic transcriptions is reasonable: it agrees more with human annotations than the phonemic baseforms do, though not by much. For the hand labelled portion of the ICSI corpus, for instance, we aligned the baseforms with the hand labels and found the *phone error rate* of the citation form transcription to be about 30%. The error rate for the automatic transcriptions for the same portion was 25%.

words	just	because	they're	grandparents ...
dict	jh ah s t	b ax k ah z	dh ey r	g r ae n p ey r ih n t s
icsi	jh ah s t	b ax k ah z	dh ey r	g r ae n p ey r ih n t s
auto	jh ax s	b ax k ao z	er	g r ae n p eh r s

It is also not clear if total agreement with the hand labels is desirable. Occasionally, as in the transcriptions shown above, a large number of human listeners preferred the automatic transcriptions to those of the annotators! Readers who would wish to listen to this particular utterance can find it on the 1997 LVCSR workshop pronunciation project web page at www.clsp.jhu.edu.

2.2.1. *Building Decision Trees from Automatic Transcriptions*

These transcriptions were used to build new decision trees. Two options were explored to use this large amount of training data – retain the topology (*i.e.*, the sequence of questions, or the equivalence classification of the contexts) of the original phonetically hand labelled corpus trees, and only update their leaf distribution by *pouring* this new training data down those trees; or rebuild the trees altogether. When applied to Switchboard, there is very little difference between the two methods as far as prediction entropy on a held out set goes, as illustrated in Table 6. It is also not surprising that the prediction entropy of these trees is higher than the ICSI+TIMIT trees trained on hand labels alone, because there is an obvious mismatch between the automatic derivation of the training transcriptions, and the hand labelling of the test set. The fully rebuilt trees were named *Retrained trees*.

Since we now had much more training data, we also built trees which additionally included in the context the previously realized surface form so as to capture some of the dependency in the surface string. Trees built this way were named *Retrained2 trees*.

2.3. Dictionary Expansion Using Pronunciation Trees

We applied the ICSI+TIMIT trees of Table 3 to successive phonemes of each baseform in the WS97 baseline dictionary to obtain a weighted pronunciation network as described in [9]. Figure 2 illustrates such a network for the word `pretty`. Applied statically, this

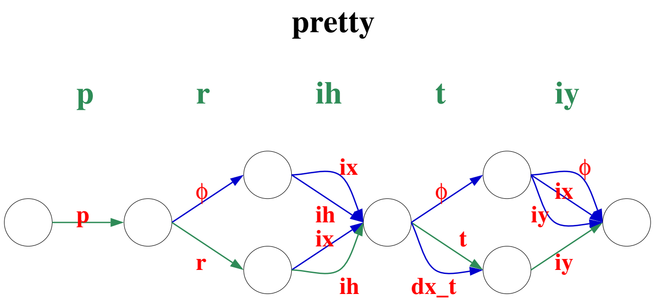

Figure 2: Pronunciation Network for `pretty`

resulted in an expanded dictionary which we call the *ICSI+TIMIT dictionary*.

We also applied the Retrained trees to baseforms in the baseline dictionary as before, to obtain a second enhanced dictionary, which we call the *Retrained dictionary*. Finally, expanding the baseforms in the baseline dictionary using the Retrained2 trees resulted in the *Retrained2 dictionary*.

2.4. Testing with Tree Based Dictionaries

At AT&T, both trees built on TIMIT and retrained trees built on the automatically transcribed SI-284 training corpus were used to constuct recognition dictionaries for the NAB Eval '95 test set. These were compared with a baseline system whose pronunciations came from the AT&T TTS text-to-speech system. Table 7 shows these recognition results. We see that the TIMIT-based trees gave a 1.9% WER reduction over the citation-form TTS dictionary, while trees retrained on the SI-284 training corpus gave an additional 0.8% reduction. In this earlier work, the full TIMIT phone set (minus the stop closures) was used, which contained 53 phones compared to the TTS inventory of 41 phonemes. Thus, new acoustic models had to be built for the larger phone set. In other words, the acoustic models used for the TIMIT and Retrained2 entries in Table 7 were different than the TTS-based test. At WS97, this was not required, since as mentioned before, the phone realizations had been forceably mapped to the PRONLEX set.

At WS97, bigram lattices for the WS97 development-test were rescored using the enhanced dictionaries described above using the WS97 baseline acoustic models[2]. Table 8 shows recognition performance using the three static but weighted enhancements to the

[2]The baseline acoustic models were state clustered cross-word triphones comprising about 7000 states, each with twelve-component Gaussian mixture output densities, trained on about sixty hours of Switchboard data.

Dictionary	WER	DEL	SUB	INS
Pronlex	44.66%	10.85%	29.47%	4.34%
ICSI+TIMIT	46.14%	11.65%	30.39%	4.10%
Retrained	43.99%	10.90%	29.08%	4.02%
Retrained2	43.75%	10.87%	28.85%	4.02%

Table 8: Switchboard Recognition Results with Enhanced Dictionaries

Dictionary	Weights	WER	DEL	SUB	INS
PronLex	–	44.66%	1987	5398	796
ICSI+TIMIT	$\sum = 1$	46.14%	2134	5568	751
ICSI+TIMIT	$\max = 1$	46.13%	1904	5653	893

Table 9: Scaling Pronunciation Probabilities

dictionary based on the ICSI+TIMIT trees, the Retrained trees and the Retrained2 trees.

The degradation in performance from the ICSI+TIMIT dictionary came as a surprise, especially since the AT&T NAB experiments showed an apparently opposite effect. There were, however, many differences between the two tests including (1) a read speech, low error task versus a spontaneous speech, high error rate task, (2) the TTS-dictionary versus the PRONLEX dictionary, (3) 53 phone set versus a 43 phone set, (4) potentially different phonetic transcription conventions between the TIMIT and ICSI labellers, and (5) acoustic model retraining in the AT&T experiments but not in the WS97 experiments. In fact, preliminary attempts at WS97 to retrain acoustic models using tree-based pronunciation lexicons lead to significantly worse results [1].

There were various conjectures made why the ICSI+TIMIT dictionary gave a worse result and we launched an series of experiments to investigate them. These are described in the next few paragraphs. From Table 8, we also see that the Retrained and Retrained2 trees yielded modest but statistically significant improvements in word error rate over the WS97 baseline system.

2.4.1. Are Words with Many Pronunciations Being Penalized?

It is conceivable that a word such as `and`, which admits many pronunciations, may be unnecessarily penalized relative to a word with few pronunciations such as `an`. *e.g.*, the phones `[ae n]` are the most likely pronunciation for both `an` and `and` in conversational speech. Since they have a much higher likelihood amongst pronunciations of `an` than amongst those of `and`, other things being equal, it costs less to map these phones to the word `an`. If Viterbi decoding is employed, many researchers have suggested that this problem is alleviated by scaling the pronunciation probabilities of every word so that the most likely pronunciation has unit weight[3].

We scaled our enhanced ICSI+TIMIT dictionary in this manner, and found an insignificant gain (see Table 9), ruling this out as the major cause of the degradation in performance.

[3] This would perhaps be unnecessary if the likelihoods were properly summed over all pronunciations of a word, but is a sensible adjustment for Viterbi decoding, as is the additional scaling of the pronunciation probabilities by the language model scale (12) to bring them on par with the acoustic scores.

Dictionary	Context	WER	DEL	SUB	INS
PronLex	–	44.66%	1987	5398	796
ICSI+TIMIT	None	46.14%	2134	5568	751
ICSI+TIMIT	1 Phone	46.09%	2112	5590	743

Table 10: Word-Internal v/s Cross-Word Pronunciation Modelling

Dictionary (# Expanded)	WER	DEL	SUB	INS
PronLex	44.66%	1987	5398	796
ICSI+TIMIT (All Words)	46.14%	2134	5568	751
ICSI+TIMIT (Top 100)	45.50%	2213	5456	666

Table 11: Expanding Only the Most Frequent Words Using Trees

2.4.2. Do the Trees Badly Need Crossword Context?

Recall that the enhanced dictionaries were obtained by applying the pronunciation trees to baseforms in isolation, and thus they could not utilize crossword context. We wrote additional software utilities so that the pronunciation model could be applied to the bigram lattices directly. However, looking at three neighbouring phonemes across word boundaries would have resulted in a drastic expansion of the lattice. We therefore implemented crossword pronunciation trees which looked at only one neighbouring phoneme in the context. This, we expected, would not be a severe limitation in light of the fact (from Table 4) that the deleted context is of little additional value in prediction. The results in Table 10 indicates that this too is not the entire reason for the poor performance of the ICSI+TIMIT dictionary. We conjecture that crossword pronunciation context is perhaps more important for some words than others (*e.g.*, `and I`, `want to`).

2.4.3. Are the Trees Generalizing Too Much?

The motivation for using local decision tree based models is to be able to observe phoneme to phone transformations which are universally applicable. However, it may be argued that since many words exhibit remarkably stable pronunciations in the hand labelled data set, the pronunciation model when applied to these words creates confusion without generating useful new pronunciations. We therefore expanded only the hundred most frequent words in the corpus using the ICSI+TIMIT trees, and tested using this instead of the ICSI+TIMIT dictionary.

As the results in Table 11 indicate, this is a significant improvement over expanding all dictionary entries, and should be investigated further. However, this is still not the sole reason for the poor performance of the ICSI+TIMIT dictionary, as the performance continues to be below that baseline system. It may be argued, for instance, that expanding only the 100 most frequent words simply brings the system closer to the baseline, and the recognition performance tracks this regression.

Pronunciation	Probability	
	ICSI+TIMIT Dictionary	Empirical
WANT TO: `[w aa n t t ax]`		
w aa n ax	0.04	0.34
w aa n t ax	0.20	0.28
w aa t t ax	0.05	–
WANT TO: `[w ah n t t ax]`		
w ah n ax	0.05	0.37
w ah n t ax	0.26	–
w ah n t ah	0.06	–

Table 12: Empirical vs. ICSI+TIMIT Dictionary Probabilities

Pronunciation	Probability	
	Retrained2 Dictionary	Empirical
WANT TO: `[w aa n t t ax]`		
w aa n ax	0.08	0.34
w aa n t ax	0.49	0.28
w aa n t uw	0.08	–
WANT TO: `[w ah n t t ax]`		
w ah n ax	0.10	0.37
w ah n t ax	0.61	–
w ah n t uw	0.10	–

Table 13: Empirical vs. Retrained Dictionary Probabilities

2.4.4. Can the Weights in Dictionary be Bettered?

Application of the decision tree model one phoneme at a time entails a conditional independence assumption between the surface forms given the baseforms, much as in a hidden Markov model (HMM). Thus the resultant probability of a pronunciation (obtained as a product of the conditional probabilities of the surface phones) is, at best, as reflective of the observed frequency of the pronunciation as the goodness of this independence assumption. To check this, we compared the probabilities of the pronunciations in the ICSI+TIMIT dictionary for a few hand-picked words with their relative frequency in our automatic transcriptions. Table 12 suggests that the tree probabilities, and perhaps the independence assumption as well, are very unsatisfactory. Much room for research and improvement remains here.

Since the Retrained trees were based on much more data (which also happened to be the same data from which the empirical probabilities of the pronunciations were inferred), we conducted a similar comparison for the Retrained2 dictionary. The example in Table 13 further reinforces our conclusion that it is the HMM-like independence assumption more than the leaf probability estimation which skews the tree based pronunciation probabilities away from their empirically observed values. Alternative probability assignments at the surface string level should be investigated in the future.

We also conducted an experiment, which clearly brings out the importance of correct pronunciation weight estimation even when the HMM-like independence assumption *is* made. Since we were not satisfied with the pronunciation probabilities of the ICSI+TIMIT trees, we poured the 60 hours of automatically transcribed data down

Dictionary	Weights	WER	DEL	SUB	INS
PronLex	–	44.66%	1987	5398	796
ICSI+TIMIT	ICSI+TIMIT	46.14%	2134	5568	751
ICSI+TIMIT	Retrained	44.05%	1982	5351	736

Table 14: Impact of Reestimating Pronunciation Tree Probabilities

the trees and reestimated the leaf distributions, as described in the context of Table 6. These trees continued to assign mismatched pronunciation probabilities to words, much as above, but they had considerably better recognition performance, as indicated in Table 14. We were unable to investigate due to time limitations during the workshop why the three retrained trees help in spite of not always being in tune with empirical pronunciation frequencies.

2.5. Summary of Tree Based Experiments

- Pronunciation probabilities based on TIMIT trees for NAB helped performance (+1.9%) and reestimated trees helped more (+0.8%).
- Pronunciation probabilities based on ICSI+TIMIT trees for Switchboard hurt performance (-1.5%), but those from reestimated trees help (+0.9%).
- Reestimated pronunciation probabilities still don't agree with empirical frequencies in training. Word level pronunciation probabilities should be examined.
- Words have variable tendencies to be mispronounced. All words in the dictionary should not be expanded equally.

3. EXPLICIT DICTIONARY EXPANSION

The degradation in performance due to the ICSI+TIMIT dictionary admits the possibility that the ICSI+TIMIT trees either generalize incorrectly or do a poor job of assigning costs to the alternate pronunciations. Both of these are crucial to the success of dictionary enhancement based methods. An alternate, more conservative approach to dictionary enhancement was therefore examined at WS97. As such, all experiments from here on apply to Switchboard.

3.1. ICSI Multiword Dictionary

The PronLex dictionary is first enhanced with all the pronunciations for words seen in the hand-labelled (ICSI) portion of the corpus. A candidate list of 172 multiwords (cf. [4]) is also appended to the dictionary to capture coarticulation, and pronunciations for these are similarly extended using the hand-labelled corpus. The word transcription of the training corpus is then expanded using these alternate pronunciations and aligned with the acoustics using our baseline models. New pronunciations which are chosen sufficiently often are deemed *bona fide* entries to the *ICSI Multiword dictionary*; the others are discarded. Pronunciations are assigned weights based on their relative frequency.

3.2. Auto Multiword Dictionary

Instead of the forced alignment among alternate pronunciations extracted from the hand-labelled portion of the corpus as described

Dictionary	WER	DEL	SUB	INS
PronLex	44.7%	10.9%	29.5%	4.3%
ICSI Multiword	44.6%	10.3%	29.7%	4.6%
Auto Multiword	43.8%	10.4%	29.1%	4.3%

Table 15: Lattice-Rescoring with Explicitly Expanded Dictionaries

above, new pronunciations for words and multiwords may be chosen from the large automatically transcribed corpus described in Section 2.2. This alternative approach yields the *Auto Multiword dictionary*. Qualitatively speaking, this dictionary invokes the decision tree pronunciation models to generate alternatives, but keeps only those which occur frequently enough in the automatic transcription. Again, weights are assigned to each pronunciation based on its relative frequency.

3.3. Recognition Results using Expanded Dictionaries

Bigram lattices for the WS97 dev-test, generated using the PronLex dictionary, are rescored using the enhanced dictionaries described above. Table 15 shows recognition performance using the two dictionaries. The 0.9% improvement due to the Auto Multiword dictionary is encouraging, particularly in contrast to the lack of improvement obtained from the ICSI Multiword dictionary. This comparison further reinforces the impression that the hand-labelled data is good for bootstrapping, but not reliable enough for directly estimating pronunciation models. At the least, incorporation of human expert knowledge into statistical information processing systems has been shown again to be a difficult problem in which naive approaches do not work as well as modelling techniques that match the supplied knowledge to the capabilties of the system.

4. COARTICULATION SENSITIVE CLUSTERING

Context dependent acoustic models such as triphone HMMs are capable of implicitly modelling some allophonic variation. However, the models in our baseline system do not distinguish between word-internal and cross-word triphones, and one may hypothesise that the gains above, especially those from the Multiword experiments, are due to better modelling of common cross-word effects. To investigate this possibility, the triphone clustering procedure in our (HTK) system is enhanced, as described next.

The major deviation from the baseline system is to mark the phones in the the PronLex dictionary to permit acoustic triphone state clustering routines to make explicit use of information about word boundary location. Another important modification is the use of a specific interjection phone set. This is not so much to model interjections better as to prevent the very frequent interjections from overwhelming the clustering and modelling of phones in noninterjections. Acoustic model training is carried out in the same manner as the baseline system, with the difference that the question set for triphone state clustering is augmented with questions regarding the word boundary tags and interjection phone set. A set of acoustic models, named the *INTWBD models*, comparable to the baseline in terms of the number of states and Gaussian components, is thus estimated.

Next, the training data is retranscribed using these models and the pronunciation networks of Section 2.2. The Retrained2 dictionary and the Auto Multiword dictionary of Sections 2.2 and 3.2 respectively are then regenerated from these transcriptions.

Dictionary	WER	DEL	SUB	INS
Baseline Acoustic Models				
PronLex	43.4%	9.8%	29.4%	4.1%
INTWBD Acoustic Models				
PronLex	41.8%	10.1%	27.8%	3.9%
Retrained2	41.3%	10.2%	27.5%	3.7%
Auto Multiword	41.1%	9.7%	27.5%	4.0%

Table 16: Lattice-Rescoring with New AMs

Dictionary	WER	DEL	SUB	INS
Baseline Acoustic Models				
PronLex	40.9%	8.9%	27.8%	4.2%
INTWBD Acoustic Models				
PronLex	39.4%	9.2%	26.2%	4.0%
Retrained2	38.9%	9.2%	25.9%	3.8%
Auto Multiword	38.5%	8.6%	25.8%	4.2%

Table 17: Lattice-Rescoring with new AMs and a Trigram LM

4.1. Recognition Results Using Improved Acoustic Models

Table 16 shows the results[4] of rescoring the WS97 dev-test set using the INTWBD acoustic models, and indicates that enabling the state clustering to take advantage of word boundary information and separate phones for interjections result in significant improvement in performance (1.6%). Observe that the two dictionary enhancement techniques continue to provide added improvements (0.7%), though to a slightly smaller extent now.

5. LANGUAGE MODEL IMPROVEMENTS

In the spirit of investigating whether pronunciation modelling via the two expanded dictionaries continues to be of benefit when other components of the system are improved, lattices generated by a bigram language model and the baseline PronLex dictionary are rescored using a trigram language model and the Retrained2 and Auto Multiword dictionaries. The results in Table 17 are therefore directly comparable with those in Table 16, which are based on bigram scores.

Observe that the improvement from the INTWBD models over the baseline models is 1.5%, which matches the 1.6% improvement with the bigram language model. The additional improvement of 0.5% from the Retrained2 dictionary also continues to hold, and the improvement from the Auto Multiword dictionary over the PronLex dictionary actually increases from 0.7% to 0.9%. All these results indicate that our straightforward pronunciation models and the coarticulation sensitive acoustic modelling provide gains which are additive to language model improvements.

[4] Though these results are for the same baseline system and test set, the baseline performance here differs slightly from the one shown in Tables 8 and 15. This is mostly due to a change in the acoustic segmentation of the test set between the two experiments, evidently for the better, and to a smaller extent due to a small change in the scoring software.

Models	WER	DEL	SUB	INS
Bigram LM				
INTWBD	41.8%	10.1%	27.8%	3.9%
MWINTWBD	41.3%	9.6%	27.5%	4.2%
Trigram LM				
INTWBD	39.4%	9.2%	26.2%	4.0%
MWINTWBD	39.0%	8.7%	26.1%	4.2%

Table 18: Lattice-Rescoring with Retrained Acoustic Models

6. ACOUSTIC MODEL RETRAINING

The baseline as well as the INTWBD acoustic models are trained on the PronLex dictionary, prompting the concern that these models are not appropriate for use with the new dictionaries. In particular, given the prevalence of reduced variants in the new dictionaries, the acoustic contexts upon which the triphone states are clustered in the baseline system are suspected to be poorly matched to the new dictionaries. This section describes a procedure used to retrain models better matched to the ICSI Multiword dictionary[5]. This work makes use of training techniques developed by the Hidden Pronunciation Mode group at the 1996 LVCSR Workshop.

First, the state clustered triphone INTWBD models and the regenerated ICSI Multiword dictionary of Section 4 are used to obtain a phonetic transcription of the corpus, which then remains fixed during training. Untied triphones for this transcription are then cloned from the monophone HMMs created during the training of the baseline system. Finally, the training procedure for the INTWBD models is mimicked starting with triphone HMM reestimation, followed by state clustering, *etc.*. The resulting HMMs, comparable in the number of states and Gaussian components to the baseline system, are called *MWINTWBD models*.

6.1. Recognition Results using Retrained Acoustic Models

Bigram lattices for the WS97 dev-test, generated using the baseline acoustic models and the PronLex dictionary, are rescored using the MWINTWBD acoustic models and the ICSI Multiword dictionary. Table 18 shows the results of the rescoring experiment.

Recall from Table 15 that the ICSI Multiword dictionary gives essentially no gain by itself, and thus the gain here (0.4%) may be attributed to the acoustic retraining. It is expected that substantially higher gains will be attained by acoustic retraining with better phonetic transcription such as those obtained using the Auto Multiword dictionary.

7. CONCLUSION

This research suggests that significant improvement in speech recognition can be made by suitably modelling systematic pronunciation variation. Further, our results indicate that while a hand-labelled corpus is very useful as a bootstrapping device, estimates of pronunciation probabilities, context effects, *etc.*, are best derived from larger amounts of automatic transcriptions, preferably done using the same set of acoustic models which will eventually be used for recognition.

On NAB, using pronunciation modelling with acoustic model retraining, we saw a 2.7% reduction in WER over a TTS baseline system. On Switchboard, without acoustic model retraining, we saw a 0.9% reduction in WER over a Pronlex baseline system, which is demonstrably additive to improvements in language (2.5%) and acoustic (1.5%) modelling, and to gains from adaptation (not reported here). Work is underway to develop effective acoustic model retraining methods for Switchboard when these statistical pronunciation lexicons are employed.

[5] The acoustic retraining was not on our best (Auto Multiword) dictionary for historical reasons: the ICSI Multiword dictionary was obtained first, and a retraining effort was started before the superiority of the Auto Multiword dictionary was established.

8. REFERENCES

[1] W. Byrne, *et al*, “Pronunciation Modelling for Conversational Speech Recognition: A Status Report from WS97,” presented at the 1997 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara, CA, Dec. 1997.

[2] W. Byrne, *et al*, “Pronunciation Modelling Using a Hand-labelled Corpus for Conversational Speech Recognition”, *Proc. ICASSP ’98* Seattle, WA.

[3] F. Chen, “Identification of Contextual Factors for Pronunciation Networks,” *Proc. ICASSP ‘90,* S14.9, 1990.

[4] M. Finke and A. Waibel, “Speaker Mode Dependent Pronunciation Modelling in Large Vocabulary Conversational Speech Recognition,” in *Proc. EUROSPEECH’97*, 1997.

[5] W. Fisher, V. Zue, J. Bernstein, and D. Pallet, “An Acoustic-Phonetic Data Base,” *J. Acoust. Soc. Am.* **81**, Suppl. 1,1987.

[6] S. Greenberg, “The Switchboard Transcription Project,” *1996 LVCSR Summer Workshop Technical Reports*, 1996, `http://www.icsi.berkeley.edu/real/stp/`

[7] P. Ladefoged, *A Course in Phonetics*, Harcourt Brace Jovanovich, Inc., New York, 1975.

[8] M. Randolph “A Data-Driven Method for Discovering and Predicting Allophonic Variation,” *Proc. ICASSP ‘90,* S14.10, 1990.

[9] M. Riley and A. Ljolje, “Automatic generation of detailed pronunciation lexicons.” *Automatic Speech and Speaker Recognition: Advanced Topics*. Kluwer. 1995.

[10] G. Tajchman, E. Fosler, and D. Jurafsky, “Building Multiple Pronunciation Models for Novel Words using Exploratory Computational Phonology”, *Proc. Eurospeech ’95*, 1995.

[11] M. Weintraub, E. Fosler, C. Galles, Y. Kao, S. Khudanpur, M. Saraclar, S. Wegmann, “Automatic Learning of Word Pronunciation from Data,” *1996 LVCSR Summer Workshop Technical Reports*, 1996.

[12] M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, C. Baldwin, D. Bell, “Linguistic Constraints in Hidden Markov Model Based Speech Recognition,” *Proc. ICASSP ’89,* S13.2, 1989.

[13] S. Young, J. Jansen, J. Odell, D. Ollasen, P. Woodland, *The HTK Book (Version 2.0)*, Entropic Cambridge Research Laboratory, 1995.