PRONUNCIATION MODELLING USING A HAND-LABELLED CORPUS FOR CONVERSATIONAL SPEECH RECOGNITION

W. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock,
M. Riley, M. Saraclar, C. Wooters and G. Zavaliagkos

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218-2686
ws97_pron@mail.clsp.jhu.edu
http://www.clsp.jhu.edu/ws97/pronunciation

ABSTRACT

Accurately modelling pronunciation variability in conversational speech is an important component of an automatic speech recognition system. We describe some of the projects undertaken in this direction during and after WS97, the Fifth LVCSR Summer Workshop, held at Johns Hopkins University, Baltimore, in July-August, 1997. We first illustrate a use of hand-labelled phonetic transcriptions of a portion of the Switchboard corpus, in conjunction with statistical techniques, to learn alternatives to canonical pronunciations of words. We then describe the use of these alternate pronunciations in an automatic speech recognition system. We demonstrate that the improvement in recognition performance from pronunciation modelling persists as the system is enhanced with better acoustic and language models.

1. INTRODUCTION

Pronunciations in spontaneous, conversational speech tend to be much more variable than in careful read speech where pronunciations of words are more likely to adhere to their citation forms. Most speech recognition systems, however, rely on pronouncing dictionaries which contain few alternate pronunciations for most words. This limitation in capturing an important source of variability is potentially a significant cause for the relatively poor performance of recognition systems on large vocabulary conversational speech recognition (LVCSR) tasks. We report some of the methods investigated to address this issue both during [1] and after WS97, the Fifth LVCSR Summer Workshop, held at Johns Hopkins University, Baltimore, in July-August, 1997.

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 a phonetically labelled portion of the Switchboard corpus [4]. 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." More importantly, 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 baseforms in the PronLex 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. [3]) of *multiwords* – word pairs and triples.

We demonstrate in Sections 2 and 3 that both expansion methods lead to a modest reduction[1] in the word error rate (WER) over a baseline system which uses a PronLex dictionary. More importantly, we show in Sections 4 and 5 that this reduction persists when the baseline system is improved by coarticulation sensitive acoustic modelling and improved language modelling. In other experiments not reported here, we have seen this improvement persist after adaptation as well.

Though our pronunciation modelling effort is preliminary due to the six week duration of the workshop, we have been able to

[1] A lattice rescoring framework is used throughout this article for obtaining meaningful results in a reasonable time-frame. Lattices are generated for the WS97 dev-test set (2427 utterances comprising about 18,000 words) using a set of baseline acoustic models, the PronLex dictionary and a bigram language model. The acoustic models are state clustered cross-word triphone HMMs comprising about 7000 states, each with twelve-component Gaussian mixture output densities, trained on about sixty hours of Switchboard data. The acoustic features are MEL-frequency PLP cepstral coefficients. The test data is ML-VTL normalized on these models, but the warps are not readjusted for any of the new acoustic models. No speaker adaptation is used. The back-off bigram language model is trained on about 2 million words of the Switchboard corpus. New results reported here are compared with this baseline system.

Features Provided as Context	$\log_2$-prob*
None (root trees)	0.714
Stress and WB Cues Only	0.606
Stress, WB and 3 Phonemes on the Left	0.537
3 Phonemes on Either Side	0.498
Stress, WB and 1 Phoneme on Either Side	0.485
Stress, WB and 3 Phonemes on Either Side	0.485

Table 1: Prediction Entropy for the ICSI+TIMIT Trees

demonstrate that hand-labelled corpora are best used as a bootstrap device and not directly as sources of pronunciation modelling, and that the pronunciation modelling gain is persistent across other system improvements.

2. TREE BASED DICTIONARY EXPANSION

Our tree based pronunciation models are inspired by phonological rules in acoustic phonetic studies (cf., *e.g.*, [5]) 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.*, [2, 3, 6, 7, 8, 9, 10]). Figure 1 illustrates the deletion or alteration of a phoneme in context which we model via decision trees.

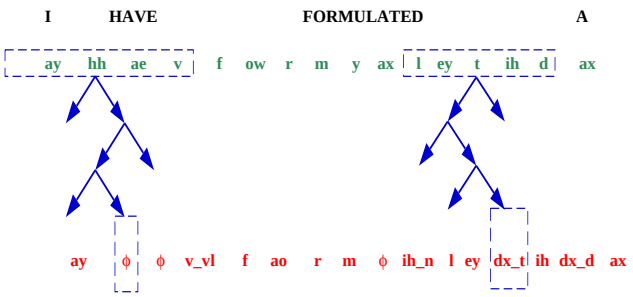

Figure 1: Decision Trees as Phone Predictors

2.1. Decision Trees from Hand-Labelled Data

Our starting point is a set of decision trees, named *ICSI+TIMIT trees*, based on nearly 3.5 hours of phonetically labelled transcriptions (ICSI) of Switchboard augmented with about 5 hours of the TIMIT data set. The tree context includes three neighbouring phonemes on either side (each encoded in terms of its phonetic features [7]), the lexical stress on neighbouring vowels as obtained from the pronouncing dictionary, and the distance of the phoneme from the nearest word boundary (WB) on either side. These trees reduce the prediction entropy of the surface form on a held out set by 32% as indicated in Table 1. The *ICSI+TIMIT dictionary* is obtained by expanding each baseforms in the PronLex dictionary into a small network using these trees. Note that the ICSI+TIMIT dictionary utilizes the pronunciation model in a word-internal manner.

2.2. Decision Trees from Automatic Phone Transcriptions

As a means of constraining the automatic phonetic transcription of a larger corpus, cross-word ICSI+TIMIT trees are applied to the training word transcriptions to obtain pronunciation networks.

Dictionary	WER	DEL	SUB	INS
PronLex	44.7%	10.9%	29.5%	4.3%
ICSI+TIMIT	46.1%	11.6%	30.4%	4.1%
Retrained	44.0%	10.9%	29.1%	4.0%
Retrained2	43.8%	10.9%	28.9%	4.0%

Table 2: Lattice-Rescoring with Tree Based Dictionaries

Next, the baseline acoustic models are used to obtain a phonetic retranscription of the corpus. Decision trees, named *Retrained trees*, are then built from these transcriptions, and applied to baseforms to obtain the *Retrained dictionary*. A third set of decision trees, named *Retrained2 trees*, is built from the automatic transcription which improves upon the Retrained trees by including in the context the surface form realized at the previous phonemic position. The corresponding *Retrained2 dictionary* is similarly obtained.

2.3. Recognition Results using Tree Based Dictionaries

Bigram lattices for the WS97 dev-test, generated using the PronLex dictionary, are rescored using the enhanced dictionaries described above. Table 2 shows recognition performance using the three dictionaries. Observe that a direct application of the hand-label based models degrades performance[2], while models based on the automatic transcriptions reduce WER (0.9%). The Retrained2 dictionary outperforms the Retrained dictionary, as expected.

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 is therefore examined.

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. [3]) 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.

[2] Several experiments were conducted to ascertain reasons for the failure of the ICSI+TIMIT dictionary, but no single cause was found. A likely suspect is the mismatch between hand transcriptions based on human perception and recognition based on machine perception (by the acoustic phonetic models). The Retrained dictionaries, in addition to being trained on a larger corpus, do not suffer from this mismatch by virtue of being trained on automatic transcriptions. This may explain their superior performance. Details of our investigations are on our web site.

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 3: Lattice-Rescoring with Explicitly Expanded Dictionaries

3.2. Auto Multiword Dictionary

Instead of the forced alignment among alternate pronunciations extracted from the hand-labelled portion of the corpus as described 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 3 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.

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 4: 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 5: Lattice-Rescoring with new AMs and a Trigram LM

4.1. Recognition Results Using Improved Acoustic Models

Table 4 shows the results[3] 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. APPOSING 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 5 are therefore directly comparable with those in Table 4, 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.

[3] Though these results are for the same baseline system and test set, the baseline performance here differs slightly from the one shown in Tables 2 and 3. 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 6: 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[4]. 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 6 shows the results of the rescoring experiment.

Recall from Table 3 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 conversational 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.

Using pronunciation modelling without any acoustic retraining, we see a 0.9% reduction in WER 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 better incorporate a pronunciation model in acoustic model training than reported here, *e.g.* by using the Auto Multiword or the Retrained2 dictionary, and larger gains are expected from this. Discovery of an effective unsupervised learning procedure for modelling pronunciations (to obviate need of a hand-labelled corpus) is an open research issue at this point.

[4]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. ACKNOWLEDGMENTS

We are grateful to AT&T Labs-Research, Florham Park, NJ, for the FSM Library used for pronunciation network processing, to Andrej Ljolje of AT&T Labs-Research, for informative preworkshop discussions, and to Entropic Cambridge Research Laboratory, Cambridge, UK, for providing the software in support of the workshop for lattice generation and rescoring.

9. REFERENCES

[1] B. 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] F. Chen, "Identification of Contextual Factors for Pronunciation Networks," *Proc. ICASSP '90*, S14.9, 1990.

[3] M. Finke and A. Waibel, "Speaker Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition," in *Proc. EUROSPEECH'97*, 1997.

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

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

[6] M. Randolph "A Data-Driven Method for Discovering and Predicting Allophonic Variation," *Proc. ICASSP '90*, S14.10, 1990.

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

[8] G. Tajchman, E. Fosler, and D. Jurafsky, "Building Multiple Pronunciation Models for Novel Words using Exploratory Computational Phonology", *Proc. Eurospeech '95*, 1995.

[9] 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.

[10] 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.

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