PRONUNCIATION MODELLING FOR CONVERSATIONAL SPEECH RECOGNITION: A STATUS REPORT FROM WS97

B. Byrne, M. Finke, S. Khudanpur,
J. McDonough, H. Nock, M. Riley,
M. Saraclar, C. Wooters, G. Zavaliagkos
(The WS97 Pronunciation Modelling Group)
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/

Accurately modelling pronunciation variability in conversational speech is an important component for automatic speech recognition. We describe some of the projects undertaken in this direction at 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 a recognition experiment as well as in the acoustic training of an automatic speech recognition system. Our results show a reduction of word error rate in both cases – 0.9% without acoustic retraining, and 2.2% with acoustic retraining.

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 at WS97, the Fifth LVCSR Summer Workshop, held at Johns Hopkins University, Baltimore, in July-August, 1997.

As a first step towards alleviating this problem, we identified a systematic way of generating alternate pronunciations of words by using a phonetically labelled portion of the Switchboard corpus [1]. One viewpoint we explored was that pronunciation variability may be modelled by a statistical mapping from canonical pronunciations (baseforms) to symbolic surface forms, and we used decision trees to capture this mapping. A second way we exploited the hand transcriptions was 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 addressed this problem in two ways. We assigned costs to alternate pronunciations so that, *e.g.*, if a frequent pronunciation of "cause" and an infrequent pronunciation of "because" happened to be identical, a penalty was incurred to attribute the pronunciation to "because" rather than "cause." More importantly, we accounted for context effects so that, *e.g.*, "to" was allowed the pronunciation `ax`, which is a frequent pronunciation of "a," only when "to" was 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 applied decision tree based pronunciation models to baseforms in the PronLex dictionary to obtain alternate pronunciations, which were then used in testing. In what we have termed *explicit dictionary expansion experiments*, we applied the decision tree based pronunciation models first to the training corpus, and performed a forced alignment with the acoustic models to "choose" amongst the alternatives. The dictionary was then explicitly augmented with novel pronunciations which occurred sufficiently often.

The tree based expansion implicitly added many more new pronunciations than the explicit expansion. However, it did not exploit any crossword coarticulation while the explicit expansion did so by allowing as dictionary entries a select set of word paris and triples – we call them *multiwords*. We obtained a reduction of 0.9% in the word error rate (WER) over a baseline system which used a PronLex dictionary by both expansion methods.

We also retrained acoustic models on a phonetic transcription of the training data obtained using an explicitly expanded dictionary (based on the hand transcriptions alone). Recognition using these models and the expanded dictionary they were trained on resulted in a 2.2% reduction in WER, which is partly attributable to some changes in the acoustic training procedure, and partly to improved training transcriptions resulting from the pronunciation modelling.

TREE BASED DICTIONARY EXPANSION EXPERIMENTS

Our tree based pronunciation models were inspired by phonological rules in acoustic phonetic studies (cf., *e.g.*, [2]) 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, 4, 5, 6, 7, 8, 9]). Figure 1 illustrates the deletion or alteration of a phoneme in context which we modelled via decision trees.

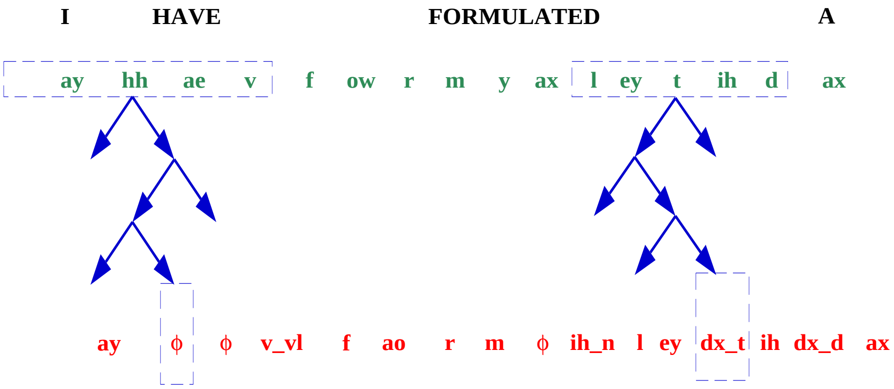
I
HAVE
FORMULATED
A
ay hh ae v f ow r m y ax l ey t ih d ax
ay ϕ ϕ v_vl f ao r m ϕ ih_n l ey dx_t ih dx_d ax

Figure 1: Decision Trees as Phone Predictors

Features Provided as Context	$\log_2$-prob*
All Features	0.485
↶ 2nd and 3rd Phonemes ↷ Excluded	0.485
Stress and Segment Boundary Cues Excluded	0.498
All Context Excluded (root trees)	0.714

Table 1: Prediction Entropy for the ICSI+TIMIT Trees

Decision Trees from Hand Labelled Data

The first set of decision trees built during WS97, named *ICSI+TIMIT trees*, were based on approximately 3.5 hours of the phonetically labelled transcriptions of Switchboard (ICSI) augmented with about 5 hours of the TIMIT data set. The context included 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 on either side. 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 cross-validation. As indicated in Table 1, the trees reduce the entropy of the surface form by 32%, as tested on a held out set. The dictionary obtained by applying these trees to the baseforms was named the *ICSI+TIMIT dictionary*.

Decision Trees from Automatic Phone Transcriptions

We also applied these trees to the training transcriptions to obtain a pronunciation network, and used the baseline acoustic models to obtain an automatic retranscription of the corpus. We then built decision trees from these transcriptions as

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: Rescoring Bigram Lattices with Tree Based Expanded Dictionaries

described earlier and named them *Retrained trees*, and the dictionary obtained by applying them to baseforms the *Retrained dictionary*. We also built a third set of decision trees from this transcription, named *Retrained2 trees*, which included in the context the surface form realized at the previous phonemic position. We then obtained the corresponding *Retrained2 dictionary*

Recognition Results using Tree Based Dictionaries

Bigram lattices for the WS97 development-test were rescored using the enhanced dictionaries described above and the WS97 baseline acoustic models[1]. Table 2 shows recognition performance using the three dictionaries. The best result here is a 0.9% reduction in WER. We also conducted several experiments to ascertain reasons for the failure of the ICSI+TIMIT based dictionary, but no single cause was found and the most likely suspect was the mismatch between the human-perceptual nature of the hand transcriptions and the signal as “perceived” by the acoustic phonetic models. Both the Retrained trees by virtue of being trained on automatic transcriptions do not suffer from this mismatch, and this may explain their superior performance. Details of our investigations may be found on our web site.

EXPLICIT DICTIONARY EXPANSION EXPERIMENTS

The degradation in performance due to the ICSI+TIMIT dictionary, if the ICSI transcriptions are to be trusted for our purposes, opens the door to the possibility that either the ICSI+TIMIT trees generalize incorrectly or do a poor job of assigning costs to the alternate pronunciations, which is crucial to the success of dictionary enhancement based methods. We therefore examined a more conservative approach to dictionary enhancement.

ICSI Multiword Dictionary

In particular, we first enhanced the dictionary with all the pronunciations for any particular word seen in the hand labelled portion of the corpus. A candidate list of 172

[1] 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. The acoustic features were MEL-frequency PLP cpestral coefficients. The test data was ML-VTL normalized. No speaker adaptation was used.

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: Rescoring Bigram Lattices with Explicitly Expanded Dictionaries

multiwords was provided by Michael Finke [5]. Pronunciations for these in the hand labelled corpus were also added to the dictionary. We then offered these alternate pronunciations to the training corpus and aligned using our baseline acoustic models. New pronunciations which were chosen sufficiently often were deemed *bona fide* entries to the dictionary; the others were discarded. The resulting dictionary was called the *ICSI Multiword dictionary*.

Auto Multiword Dictionary

Instead of choosing new pronunciations for words and multiwords from the hand labelled portion of the corpus, we had the alternative of choosing them from the large automatically transcribed corpus described in Section 2.2. This alternative approach yielded what we call the *Auto Multiword dictionary*. Qualitatively speaking, we relied on the decision-tree pronunciation models at transcription time when confusability was lower but allowed only the frequent pronunciations, including a few multiwords, at test time when confusability was higher.

Recognition Results using Explicitly Expanded Dictionaries

Bigram lattices for the WS97 development-test were rescored using the enhanced dictionaries described above and the WS97 baseline acoustic models[2]. 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 without retraining. This comparison also suggests that acoustic model retraining based on the Auto Multiword pronunciations is worth pursuing (instead of the retraining based on the ICSI Multiword dictionary which we did at the workshop and which we described next).

ACOUSTIC MODEL RETRAINING

The WS97 baseline acoustic models were trained from the original (unmodified version of the) PronLex dictionary which prompted the concern that these models were

[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. The acoustic features were MEL-frequency PLP cpestral coefficients. The test data was ML-VTL normalized. No speaker adaptation was used.

Word	Tagged Pronunciation
ABBA	ae:s b ax:e
A	ax:m
A	ey:m
HUH-UH	hh:i ah:i ah:i
HUM	hh:i ah:i m:i
HUMAN	hh:s y uw m ax n:e

Table 4: Word Boundary Phone Tags

not appropriate for use with the new dictionaries. In particular, given the prevalence of reduced variants in the new dictionary, the acoustic contexts upon which the triphone states were clustered in the baseline system were suspected to be poorly matched to the new dictionary. This section describes the procedures used to retrain models better matched to the new dictionary. This work made use of training techniques developed by the Hidden Pronunciation Mode group at the 1996 LVCSR Workshop.

A major deviation from the WS97 baseline system was to mark the phones in the the multiword dictionary to permit acoustic triphone state clustering routines to make explicit use of information about word boundary location. We took the view that since coarticulation at word boundaries is a major effect, permitting triphone clustering to be sensitive to this information is a form of pronunciation modelling. Another important modification was the use of a specific interjection phone set. The motivation for this was not to model interjections better, but rather to prevent interjections from contributing to, and thus overwhelming by their frequency of occurrence, the clustering and modeling of phones in non-interjections. The dictionary entries were enhanced with tags that distinguished word-initial and word-final phones. Phones in monophone words were tagged separately, as were phones in interjections. Examples of dictionary entries using this tagged phone set are given in Table 4.

The phone transcriptions found from forced alignment through the training set pronunciation networks in the process of constructing the ICSI Multiword dictionary were stripped to monophones. These were tagged to be consistent with the dictionary tagging and transformed to triphones based on their context. The monophone HMMs created in training the baseline system were then cloned to provide a model for each triphone in the training set. Acoustic model training was then carried out in the same manner as the baseline system, with the difference that the question set for triphone state clustering was augmented with questions regarding the word boundary tags and interjection phone set. A system comparable to the baseline in terms of the number of states and Gaussian components was built.

Dictionary	WER	DEL	SUB	INS
No Acoustic Retraining				
PronLex	44.7%	10.9%	29.5%	4.3%
Acoustic Retraining				
ICSI Multiword	42.5%	10.1%	28.1%	4.3%

Table 5: Rescoring Bigram Lattices with Retrained Acoustic Models

Recognition Results using Retrained Acoustic Models

Results of rescoring lattices generated using the WS97 baseline system are shown in Table 5. As mentioned earlier, bigram language model scores were available in the lattice and the acoustic features were MEL-frequency PLP cepstral coefficients. The test set data was ML-VTN adjusted based on the workshop baseline models; no adjustments of the VTN warp were made for the new models and no speaker adaptation was used in either system.

A substantial improvement of 2.2% percent over the baseline system was seen with the retrained acoustic models in this experiment. Many factors were incorporated simultaneously and therefore it is difficult to gauge the contribution of each individual change to the improvement over the baseline. Work has been undertaken since the conclusion of the workshop to determine the beneficial effect, if any, of each change.

CONCLUSIONS

The research conducted at the workshop, while clearly preliminary in nature, indicates 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 saw a 0.9% reduction in word error both with the decision tree method and the explicit multiword dictionary expansion. With enhanced acoustic training, the ICSI-multiword approach showed a word error rate reduction of 2.2%.

While we were heartened by the improvements seen and the knowledge gained, there were nevertheless many avenues and details left unexplored at the conclusion of the workshop. These include accessing the relative contribution of variations to the baseline to our overall WER improvements, trying acoustic retraining using the auto-multiword approach, other effective methods for acoustic retraining, and discovering an effective unsupervised learning procedure for modelling pronunciations.

Some of these issues are being addressed in the student project spawned from this team. Others are being be explored by team members at their respective sites and through collaborations formed in the workshop.

References

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

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

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

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

[5] M. Finke and A. Waibel, “Speaker Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition,” to appear in *EUROSPEECH’97*, 1997.

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

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

[8] 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.

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