Pronunciation Modeling for Large Vocabulary Conversational Speech Recognition

Kristine Ma *George Zavaliagkos* *Rukmini Iyer*

GTE/BBN Technologies
70 Fawcett Street
Cambridge, MA 02138
kma@bbn.com

ABSTRACT

In this paper, we address the issue of deriving and using more realistic pronunciations to represent words spoken in natural conversational speech. Previous approaches include using automatic phoneme-based rule-learning techniques [1, 2, 7], linguistic transformation rules [4, 8], and phonetically hand-labelled corpus [3] to expand the number of pronunciation variants per word. While rule-based approaches have the advantage of being easily extensible to infrequent or unobserved words, they suffer from the problem of over generalization. Using hand-transcribed data, one can obtain a more concise set of new pronunciations but it cannot be extended to unobserved or infrequently occuring words. In this paper, we adopt the hand-labelled corpus scheme to improve pronunciations for frequent multi and single words occurring in the training data, while using the rule-based techniques to learn pronunciation variants and their weights for the infrequent words. Furthermore, we experiment with a new approach for speaker-dependent pronunciation modeling. The newly expanded dictionaries are evaluated on the Switchboard and Callhome corpora, giving a slight reduction in word recognition error rate.

1. INTRODUCTION

Spontaneous speech tends to alter the canonical pronunciations of words, therefore in pronunciation modeling, we try to derive and use more realistic pronunciations to represent words spoken in conversational speech. For example, the word "because" is often observed in conversation speech with a full or a reduced vowel in the initial syllable, or sometimes the entire initial syllable is dropped to result in "'cause". Another frequently observed phenomena relates to multi-words, where contractions and reductions at word boundaries result in a word pair sounding like a single word; examples include "sort of" being reduced to "sorta", or "going to" being compressed to "gonna". Current approaches to modeling such pronunciation variations include using automatic phoneme-based rule-learning techniques [1, 2, 7], linguistic transformation rules [4, 8], and phonetically hand-transcribed corpus [3]. While rule driven approaches have the advantage (over using hand-transcribed data) of being easily extensible to infrequent or unobserved words, they typically suffer from the problem of over generalization.

The work reported here starts off with a conservative approach. Adopting from the experience gained in [3], we obtain the initial pool of pronunciation variants for the frequent multi- and single words from a phonetically hand-transcribed sample of Switchboard data [5]. Each pronunciation variant from the pool is then assigned a weight based on the frequency of occurrence in the complete training set available. Low frequency pronunciation variants are pruned to provide an expanded dictionary. We further improve this initial dictionary to include (i) pronunciation variants for infrequent words using existing linguistic transformation rules [4, 8], and (ii) speaker adapted pronunciation weights.

The paper is organized as follows. In Section 2.1, we describe our data-driven technique for enhancing the lexicon with multi-words, and new pronunciation variants and pronunciation weights for the frequent single words. In Section 2.2, we expand this baseline system with pronunciation variants for infrequent words in our dictionary. Section 2.3 describes our approach for speaker-dependent pronunciation modeling. Experimental results on the Switchboard corpus are presented in Section 3. Section 4 concludes with a discussion of the experimental results and future work.

2. PRONUNCIATION MODELING

2.1. Baseline System

The BBN canonical Switchboard dictionary contains approximately 25,000 words, and on the average, each word has about 1.1 pronunciations. Spontaneous conversational speech tends to have a lot more pronunciation variability than can be captured by the canonical pronunciations. We therefore focused first on improving the baseline dictionary by adding pronunciation variants for the frequent words, word pairs and word triplets. As mentioned earlier, word pairs and triplets are particularly useful for capturing contraction and reduction effects across word boundaries. All words and multi-words observed at least 40 times in training are considered to be "frequent".

Typically, pronunciation variants are derived from the output of automatic phone recognition systems [1, 2]. However, given the high phone recognition error rates for the Switchboard corpus, we preferred to extract pronunciation variants for the frequent multi- and single words from a small phonetically transcribed sample of the Switchboard data [5, 3]. As was found by other researchers, over-generating multiple pronunciations in the dictionary increases word confusability during recognition, often nullifying the advantages of modeling pronunciation variability. To avoid problems from over-generation, we performed an iterative forced Viterbi alignment procedure on the training data using the expanded dictionary. All pronunciations selected less than 5% of the time were pruned out to result in a smaller set of robust pronunciation variants.

There are three issues critical to using such an expanded dictionary in state-of-the-art recognition systems: (i) estimation of pronunciation weights (or costs for the different pronunciation variants), (ii) treatment of multi-words in the n-gram language model, and (iii) acoustic retraining of the pronunciations using phonetic transcriptions of the training data obtained with the expanded dictionary.

First, for the frequent words we estimated pronunciation weights using the relative frequency of the observed variants in the Viterbi-aligned training data. Adding gender-dependence to the pronunciation weight estimation did not improve recognition performance; on the other hand, using all the training data, independent of gender, resulted in more accurate pronunciation weights giving improved recognition performance. Second, we investigated two alternative approaches to multi-word language modeling: one where multi-words are treated as a single unit, and another in which where each token in the multi-word is modeled independently. Contrary to the observations made in [4], our studies indicated that each multi-word should be treated as a single unit to capture a wider language model context at the expense of having fewer n-gram training samples for the individual tokens. Third, we retrained our acoustic models based on the Viterbi-aligned training data using the expanded dictionary. Again, contrary to results reported in [4, 3], acoustic retraining did not result in performance improvements.

2.2. Rule-Based Dictionary Expansion

The pronunciation modeling approach described in Section 2.1 does not generalize to infrequent words, an advantage often found with using rule-based expansion techniques. However, the phonetically transcribed Switchboard sample does provide a reasonably tight set of pronunciation variants for the frequent words. To exploit the advantages from both the data-driven and the rule-based expansion techniques, we apply pronunciation transformation rules only to infrequent words whose pronunciation variants and weights cannot be accurately derived from the phonetic transcriptions.

The pronunciation variants of infrequent words are first derived using a subset of the rules described in [4, 8]. The pronunciation weights for the new variants are then estimated using decision trees via SPLUS. The probability of a rule being applied, $P(r_k)$, is estimated using a set of questions $Q = q_1, q_2, \ldots, q_T$ based on:

- the number of times a rule-generated pronunciation occur in the training data,
- 37 questions on each of the left and right context, center on the transformed phoneme (e.g. “is the left phone a liquid?”, “is the right phone a word boundary?”),
- the number of times the word is seen in the training data, and
- the length of the word in terms of the number of phonemes.

The trees are grown to minimize the error rate of predicting whether a rule should be applied or not. We grow one tree per pronunciation rule. Let r^+ be the rules which match with the base-form of word, w_j, and are used to derive pronunciation variant i. Let r^- be the rules which match with the base-form of word, w_j, but are not used to derive pronunciation variant i. The weight for each pronunciation variant, p_i^j, of each word w_j is estimated as given in [4, 8]

$$p_i^j = \frac{\prod_{r^+} P(r^+) \prod_{r^-} (1 - P(r^-))}{Z}$$

where Z is the normalizing factor so that for each word w_j,

$$\sum_i p_i^j = 1.$$

2.3. Lexical Adaptation

In our work we have found that weighting the pronunciation variants by their probability of occurrence is crucial. Our baseline approach described in the previous section estimates the pronunciation weights in a speaker independent (SI) fashion, that is it assumes that the pronunciation weights are the same for all speakers. Since speakers *are* different (for example, different speaking rates can result in different pronunciation weights), it is natural to look for techniques that adapt these weights to the patterns of a new speaker.

Similar to speaker adaptation, the main challenge here is that we have too many parameters (equal to the number of alternative pronunciations in the dictionary, which is of the order of thousands) and too few observations (a 5 minute conversation contains roughly 1000 word occurrences, and the subset of these words that have pronunciation variants is small). Hence, in order to successfully adapt the pronunciation weights we have to first cluster the variants together, then modify their probabilities in a constrained manner.

Tying pronunciation variants The tying of the pronunciation variants can be achieved with a rule based approach. For example, if we want just a few clusters we can group the variants based on whether the pronunciation change happens on the boundary or in the middle of a word, based on whether the change is a deletion, insertion or substitution of a phoneme or combinations of the above. To get more detailed clusters we can look at what happens at phoneme classes.

Automatic ways of clustering can be introduced by looking at the co-occurrence of pronunciation for each speaker: if two pronunciation variants change in tandem for all speakers, we will assume that their probabilities (pronunciation weights) will, too.

Adapting the pronunciation weights We start by parameterizing the pronunciation probability for variant i, p_i as

$$p_i(\alpha) = \frac{\exp(\alpha + \gamma_{\rm i})}{1 + \exp(\alpha + \gamma_{\rm i})}$$

where α is a parameter shared by all variants that belong to the same cluster and γ_i are variant specific. Let $p_i(0)$ be the SI pronunciation weight. Solving for γ_i, we get

$$\gamma_i = \ln \frac{\rm p_i}{1 - \rm p_i}.$$

With data from a new speaker, we will re-estimate α so as to increase the likelihood of the observed variants. Assume that for each speaker, we observe n_i occurrences of pronunciation variant i out of N_i possible ones. Then the likelihood of the pronunciation cluster is

$$L(\alpha) = \prod_i p_i(\alpha)^{n_i} (1 - p_i(\alpha))^{N_i - n_i}$$

and we select the value for α such that the above quantity or its log is maximized. The derivative of the log likelihood with respect to α has a very intuitive interpretation:

$$\frac{\partial \rm lnL(\alpha)}{\partial \alpha} = \sum_i (n_i - N_i p_i(\alpha)).$$

That is, when the actual number of observations for variant i, n_i, is bigger than the expected number of observations, $N_i p_i$, the gradient is positive and α increases, thus increasing the probability of occurrence, p_i. When the number of observations is smaller than expected, the gradient is negative and the probability of the variants is decreased.

The pronunciation adaptation process Similar to speaker adaptation, the adaptation of pronunciation weights for the Switchboard system is unsupervised: first we use the SI system with SI pronunciation variant weights to obtain putative transcriptions of the speech. These errorful transcription are then used to supervise both acoustic and pronunciation weight adaptation, and the adapted models and weights are used in a second recognition pass.

Adaptive training of pronunciation weights Once adaptation is used, the original estimate of γ_i is no longer maximum likelihood. One can introduce an adaptive training procedure (similar to speaker adaptive training [6]) where the γ's are initialized as said, then a speaker dependent estimate of α is obtained, and new values of the γ's are estimated given the adapted pronunciation weights.

3. EXPERIMENTS

Recognition results are reported on the Switchboard and Callhome corpora using the BBN Byblos System [10], a state-of-the-art speaker independent HMM system. The test set comprises of 7 Switchboard and 7 Callhome conversations drawn from the NIST 1997 Large Vocabulary Speech Recognition evaluation data set. The baseline decoding dictionary is a 25,000 word dictionary with no multi-words.

Acoustic training for the pronunciation experiments use an in-house 18 hour subset of the Switchboard data. However, pronunciation variants for all the multi-words and the 166 frequent single words are extracted from a sample of the Switchboard data set that has been manually aligned at the phonemic level [5]. Pronunciation weights are estimated from Viterbi alignments of the same 18 hours of Switchboard training data.

Table 1: *Word error rate on BBN development test set: 7 Switchboard and 7 Callhome conversations. Performance reported in terms of absolute WER decrements (increments) from the previous line.*

Experimental Conditions	WER (%)
Baseline canonical dictionary	54.6
+ 193 multi-words, pronunciation weights	-1.0
+ 1273 additional multi-words	-0.3
+ rule-based pronunciation variants	+0.2

The results are shown in Table 1. All experiments use the expanded dictionaries only during decoding. Multi-words are treated as single tokens during n-gram language model training. Adding a) 383 new pronunciation variants and weights for 166 frequent single words, and b) 193 multi-words with with 3.7 pronunciations each on the average resulted in a 1% improvement in performance. With this preliminary pronunciation enhanced dictionary, we observed that most of our system improvement is attributed to the new pronunciations introduced for multi-words rather than for single word. Therefore, we expanded the number of multi-words to 1466, whose pronunciation weights are estimated based on the frequency count on 60 hours of training data. This resulted in an additional 0.3% improvement in performance. Adding rule-based pronunciation variants and weights for the infrequent words resulted in a small degradation in performance. We hypothesize that this may be a result of over-generation of pronunciation variants which increases word confusability during recognition.

In another series of experiments, we measured the impact of pronunciation modeling with increasing the complexity of the baseline system, both in terms of the number of parameters in the system and increased acoustic training availability. As shown in Table 2, gains from pronunciation modeling reduces, both as the number of parameters as well as the training in the baseline system increases.

Finally, many of the pronunciation modeling gains reported in earlier work [3, 4] appear system-dependent. Pronunciation retraining as well as treating multi-words as multiple tokens in the n-gram model do not result in performance improvements.

4. CONCLUSION

In this paper, we have presented pronunciation modeling improvements for large vocabulary conversational speech

Table 2: *Word error rate on BBN development test set: 7 Switchboard and 7 Callhome conversations. Performance reported in terms of absolute WER decrements from the baseline.*

System description	No pron modeling	With pron modeling
triphone, 18 hours training	54.55	53.49
triphone, 160 hours training	47.94	47.19
quinphone, 160 hours training	44.40	44.11

recognition systems, using an approach that combines the advantages of data-driven and rule-based techniques. The data-driven approach provides robust pronunciation variants and weights for the frequent single and multi-words. On the other hand, the rule-based approach is used to derive pronunciation variants for infrequent words. While modeling pronunciation variability improves the recognition performance by as much as 1%, the gains reduce as the complexity of the system and the amount of acoustic training increase. Experiments with speaker dependent lexical adaptation were inconclusive due to suboptimal pronunciation clusters. However, we feel that with improved pronunciation clustering schemes, speaker dependent lexical adaptation will outperform simple SI pronunciation modeling.

5. REFERENCES

1. T. Fukada and Y. Sagisaka, "Automatic Generation of a Pronunciation Dictionary Based on a Pronunciation Network," in Proc. EUROSPEECH, 1997.

2. N. Cremelie and J. Martens, "Automatic Rule-based Generation of Word Pronunciation Networks," in Proc. EUROSPEECH, 1997.

3. W. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters and G. Zavaliagkos, "Pronunciation Modelling Using a Hand-Labelled Corpus for Conversational Speech Recognition," in Proc. ICASSP, 1998.

4. M. Finke, A. Waibel, "Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition," in Proc. Eurospeech, 1997.

5. S. Greenberg, "The Switchboard Transcription Project," 1996 LVCSR Summer Workshop Technical Reports, 1996.

6. J. McDonough, T. Anastasakos, G. Zavaliagkos, H. Gish, "Speaker-Adapted Training on the Switchboard Corpus", Proc. ICASSP, 1997.

7. H. J. Nock and S. J. Young, "Detecting and Correcting Poor Pronunciations for Multiword Units," Workshop Modeling Pronunciation Variation, Rolduc, May 1998.

8. M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld, "Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode," in Proc. ICSLP, 1996.

9. M. Ravishankar and M. Eskenazi, "Automatic Generation of Context-dependent Pronunciations," in Proc. EUROSPEECH, 1997.

10. G. Zavaliagkos, J. McDonough, D. Miller, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Siu, H. Gish, "The BBN Byblos 1997 Large Vocabulary Conversational Speech Recognition System," in Proc. ICASSP, 1998.