STATISTICAL MODELING OF PRONUNCIATION AND PRODUCTION VARIATIONS FOR SPEECH RECOGNITION

F. Korkmazskiy and B.-H. Juang

Lucent Technologies Bell Laboratories, Murray Hill, NJ 07974, USA
yelena@research.bell-labs.com

ABSTRACT

In this paper, we propose a procedure for training a pronunciation network with criteria consistent with the optimality objectives for speech recognition systems. In particular, we describe a framework for using maximum likelihood(ML) and minimum classification error(MCE) criteria for pronunciation network optimization. The ML criterion is used to obtain an optimal structure for the pronunciation network based on statistically-derived phonological rules. Discrimination among different pronunciation networks is achieved by weighting of the pronunciation networks, optimized by applying the MCE criterion. Experinent results demonstrate improvements in speech recognition accuracy after applying statistically derived phonological rules. It is shown that the impact of the pronunciation network weighting on the recognition performance is determined by the size of the recognition vocabulary.

1. INTRODUCTION

Variations in pronunciation arise in natural speech due to phonological and productional reasons. Incorporating alternative pronunciations into a speech recognition decoding network can potentially improve the recognition accuracy. There has been extensive study in using decision trees to represent pronunciation variations [1]. For the Resource Management(RM) task(991 words with perplexity 60 grammar) this approach achieves 96.3% word accuracy versus 93.4% word accuracy obtained with conventional single baseform technique. A training procedure in this approach is based on splitting a pronunciation tree node to increase the degree of predictability, as measured by an entropy figure, of the subsequent phonemes. Although the approach improves the recognition performance, there exists a discrepancy between the entropy measure used during training and the likelihood measure used to select the best candidate during recognition. Another group of methods makes intensive use of the explicit phonological rules defined for the specific language [2]. Also in this case, we don't have common criteria used for phonological rule selection and recognition optimization. Yet another approach is to employ maximum likelihood (ML) in deriving multiple word pronunciations using speech data for the specified words [3]. For the RM task this approach achieves 18.4% word error rate reduction compared to mannualy generated single baseform lexicon. However, with this approach, knowledge about possible pronunciations can only be applied to the words that are presented in the training speech database. In order to extend this knowledge to new words, it is necessary to collect acoustic data for these words. In our paper we address the problem of word alternative pronunciation derivation using exclusively their baseform pronunciation and some set of rules. The set of rules is obtained by applying ML criterion to the acoustic data which include speech samples of words that can be different from the words we derive pronunciations for. We will refer to these rules as *statistically derived phonological rules.* These rules are defined as a set of transformations that can be applied to a string of phonemes producing some allophonic variations of the string.

In this study we focus on two issues: derivation of alternative pronunciation and methods to achieve word discrimination for the expanded representation. We also introduce the idea of segmental alternative which, unlike conventional approaches to alternative pronunciation using phonemes as unit, considers an acoustic realization to contain a legitimate alternative pronunciation only when it deviates from the baseform for a segment of more than a predetermined threshold.

2. PHONOLOGICAL RULES DERIVATION

We define a phonological rule $R(A \Rightarrow B)$ as a function that maps one string of phonemes $A = \{a_1, a_2, \ldots, a_m, \ldots, a_M\}$ to another string of phonemes $B = \{b_1, b_2, \ldots, b_m, \ldots, b_N\}$. Our goal is to find both A and $R(A \Rightarrow B)$ for a given lexicon through actual acoustic realizations of the words.

Assume we have a set of words, their pronunciation

baseforms(i.e. a sequence of phonems) and an associated set of acoustic data containing examples for these words. The acoustic data can be phonetically segmented and labeled (in terms of the baseform sequence) either manually or automatically (by using, for example, a forced alignment procedure). Our goal is to find a set of phonological rules that can provide a reasonable explanation for the differences between the baseform labels and the realized phoneme sequences. That is, we aim at finding a set $\Lambda = \{A_i\}$ containing the strings A_i representing parts of the word baseform pronunciations, and a set $\Gamma = \{R_i^{(k)}\}$ containing phonological rules $R_i^{(k)}$: $R(A_i \Rightarrow B_i^{(k)})$ for each of the string A_i. The number of the rules K_i that can be applied to the string A_i is restricted ($K_i \leq K_m$). Also, we account only for phoneme strings A_i and the corresponding mapping $R_i^{(k)}$ that can significantly increase the likelihood for the acoustic data representing the strings A_i.

For every (partial) baseform string label A_i, we collect from the labeled acoustic data all segments corresponding to the string A_i to form a set $X_i = x_i^{(j)}, 1 \leq j \leq J_i$. Assume $P(X|A_i)$ is the total likelihood evaluated for all the acoustic data from X_i by measuring the likelihood scores $P(x_j|A_i)$:

$$P(X|A_i) = \prod_{j=1}^{J_i} P(x_j|A_i) \tag{2.1}$$

If we supplement a string A_i representing part of the word baseform with a set $B_i = \{B_i^{(k)}, 1 \leq k \leq K_i\}$ of its alternative pronunciations $B_i^{(k)}$ obtained after applying the phonological rules $R_i^{(k)}$, the total likelihood $P(X|A_i)$ would change to $P(X|A_i, B_i)$,

$$P(X|A_i, B_i) = \prod_{j=1}^{J_i} P(x_j|A_i, B_i). \tag{2.2}$$

Here $P(x_j|A_i, B_i)$ is the maximum value for the likelihood scores $P(x_j|C_i)$ measured over all possible strings C_i that include a baseform string A_i and all alternative pronunciations $B_i^{(k)}$. It is clear that $P(X|A_i, B_i) \geq P(X|A_i)$. The more is the difference between $P(X|A_i, B_i)$ and $P(X|A_i)$ the better is the set of rules $R_i^{(k)}$ for the string A_i. Therefore, the optimal solution for phonological rule derivation can be defined in the following way: find a pair $(\Lambda_{opt}, \Gamma_{opt})$ of an optimal set of I strings in Λ_{opt} and an optimal set Γ_{opt} of the phonological rules applied to these strings that maximize the increase in the total likelihood over the strings A_i:

$$(\Lambda_{opt}, \Gamma_{opt}) = \underset{(\Lambda,\Gamma)}{\arg\max} \sum_{A_i} (\log P(X|A_i, B_i) - \log P(X|A_i)) \tag{2.3}$$

The task (2.3) can be solved by sequential optimization, first, over the set of rules Γ and then over the set of strings Λ. A set of optimal rules $B_i \in \Gamma$ can be found separately for each string A_i because there are no common rules for the different strings A_i. So, by applying an exhaustive search over all possible strings A_i and deriving an optimal set of rules B_i for each of them we complete optimization of (2.3) over the set of rules Γ. After selecting the I strings out of all possible strings A_i that give the largest increase in the total likelihood we complete optimization over the set of strings Λ. Optimization for each of the strings A_i is accomplished as follows:

$$B_{opt}(i) = \underset{B_i}{\arg\max} \ (\log P(X|A_i, B_i) - \log P(X|A_i)) \tag{2.4}$$

Here the set $B_{opt}(i)$ contains the optimal set of the phonological rules for the string A_i. In fact we only need to maximize $P(X|A_i, B_i)$ because for any given A_i, $P(X|A_i)$ is a constant; i.e.

$$B_{opt}(i) = \underset{B_i}{\arg\max} \sum_{j=1}^{J_i} \log P(x_j|A_i, B_i). \tag{2.5}$$

This task can be solved by using the K-means clustering procedure. The speech samples x_j are clustered according to their closeness (expressed in terms of the likelihood) to the string-templates represented by either a string A_i or by alternative strings $B_i^{(k)} \in B_i$. In this study the K phoneme strings representing the K-best string-centroids for each step of the K-means clustering procedure are selected from the list of M phoneme strings($K < M$). The baseform string A_i is always included as a template for the K-means clustering procedure unconditionally. In turn the list of the M phoneme strings - candidates for centroids is compiled as follows:

- Implement N-best decoding for all acoustical samples representing a string A_i. Pool decoded phoneme strings from the N-best lists for all samples into a common set C.
- Evaluate the total accumulated likelihood for each of the decoded phoneme strings from the set C over all samples, where this string appears in the N-best list.
- Among all decoded phoneme strings select the M ones that have maximum values of the total accumulated likelihood. These strings represent a list of the phoneme strings - candidates for templates.

The larger number K_i of the phonological rules(phoneme string clusters) the larger value of $P(X|A_i, B_i)$. But not any larger number K_i leads to the noticable increase in

$P(X|A_i, B_i)$. To avoid a high complexity of the derived phonological rules we select a value K_i as follows:

$$K_i = \arg\min_{0 \le K \le K_m} [\log P(X|A_i, B_i^{(K_m)}) - \log P(X|A_i, B_i^{(K)})] < \Delta \quad (2.6)$$

Here $P(X|A_i, B_i^{(K_m)})$ and $P(X|A_i, B_i^{(K)})$ correspond to the value of $P(X|A_i, B_i)$ for the number of phonological rules equal to K_m and K respectively and Δ is a small positive constant($\Delta > 0$). In particular, if $K = 0$ no phonological rule is accepted because of a nonsignificant increase in the likelihood they can provide.

All transformations constituting a rule are labelled with a probability of their use, evaluated by counting a number of speech samples assigned to a corresponding cluster.

In our experiments all possible strings A_i of different lengths, varying from 1 to 7, were examined to determine whether some phonological rules could be applied to them. If the number of speech examples representing a string is less than some marginal value (in our experiments 30) no rule is applied. We consider 2 ways of rule application in the vocabulary:

- The set of rules applied to a specified word should be determined prior to recognition. Such kind of rules we call *static rules*.
- The set of rules applied to a word is determined a posteriori during recognition allowing to select an optimal sequence of rules that maximizes the total likelihood for the word. Such kind of rules we call *dynamic rules*.

In our study static rules are applied to a word as follows. First, a higher priority is given to the rules applied to the longer strings A_i within a word. Second, rules are applied according to the string position within a word from left to right without overlapping. Application of dynamic rules implies use of a pronunciation network composed of all possible rules that can be applied to the given word. During recognition an optimal sequence of the phoneme string is selected according to the likelihood accounting for both the probability of the proper rules generation and the speech signal scores. By merging the pronunciation networks for separate words we create a general pronunciation network for decoding.

In our experiments, an isolated-word telephone channel speech PhoneBook database was used. Over 1,300 native speakers of American English reflecting different pronunciation styles and dialects were recorded in the database. The database consists of the utterances of almost 8000 different words. In our experiments 7500 words(10 utterances per a word) were used for the phonological rules derivation and another 500 words(5 utterances per a word) included in the test set for recognition.

In the benchmark experiment, only single-word baseform pronunciations were used for recognition of the vocabulary of 500 words. These pronunciations were obtained by Bell Labs text-to-speech system. 41 context independent HMMs(3 states, 8 mixtures per a state) representing 40 phonemes and a silence model were trained using a different from the PhoneBook database. The word error rate obtained in this experiment was 10.0%. Applying static rules allowed to improve the word error rate after introducing a constraint on the minimum length of the strings A_i (i.e. rules can be applied only to strings of 3 or more phonemes). The word error rate in this case was 9.5%. The word error rate of 9.3% was achieved by applying dynamic rules without introducing any constraints on the string minimum length. We found that a special constraint on the possible string transformation can further improve the recognition performance. This constrain allows application of only such rules that do not modify the first and the last phoneme of the string A_i. Using this constraint, we obtained 9.1% word error rate after applying static rules and 8.5% word error rate after applying dynamic rules. The last result represents a 15% word error rate reduction in comparison to the recognition experiment where only word baseforms are used.

3. PRONUNCIATION NETWORKS DISCRIMINATION

To enhance discrimination between competing words, weight coefficients can be assigned to the elements (arcs, HMM states) of the prinunciation networks. The optimality of the weights is defined in the framework of discriminative training via classification error minimization([4, 5]). In the current study we consider weighting of HMM state scores. Assume an output of the recognition system is represented by the N phoneme strings which are the best N decoded pronunciations for the R words:

$$N = \sum_{r=1}^{R} N_{v_r} \quad (3.1)$$

Here N_{v_r} is the total number of the decoded alternative pronunciations for the word v_r. A likelihood score $g(X|v_r^{(n)})$ for the n-th($1 \le n \le N_{v_r}$) decoded pronunciation $v_r^{(n)}$ of the word v_r can be expressed as follows:

$$g(X|v_r^{(n)}) = \sum_{l=1}^{L(v_r^{(n)})} \sum_{s=1}^{S(l)} \rho_s^{(l)}(X|v_r^{(n)}) \quad (3.2)$$

Here $L(v_r^{(n)}$ is a total number of the phonemes (subword HMMs) in the pronunciation $v_r^{(n)}$, $S(l)$ is the total number of states in the l-th HMM, $\rho_s^{(l)}(X|v_r^{(n)})$ is a likelihood score estimated at the s-th state of the l-th phoneme of the pronunciation $v_r^{(n)}$. A weighted version of the likelihood score $\tilde{g}(X|v_r^{(n)})$ takes on such a form:

$$\tilde{g}(X|v_r^{(n)}) = \sum_{l=1}^{L(v_r^{(n)})} \sum_{s=1}^{S(l)} [\rho_s^{(l)}(X|v_r^{(n)}) \cdot a_s(v_r^{(n)}) + b_s(v_r^{(n)})] \tag{3.3}$$

Here $a_s(v_r^{(n)})$ and $b_s(v_r^{(n)})$ represent multiplicative and additive weighting terms for the state score $\rho_s^{(l)}(X|v_r^{(n)})$. Weight coefficients for all states of the word pronunciation networks are included in the common set W. A likelihood score $g(X|v_r)$ for the word v_r is estimated as a maximum weighted likelihood score among the N_{v_r} decoded pronunciations for this word:

$$g(X|v_r) = \underset{1 \le n \le N_{v_r}}{\arg\max} \; \tilde{g}(X|v_r^{(n)}) \tag{3.4}$$

A classification error $L(X|v_r, W)$ for a speech sample X is conditioned by the set of weight coefficients W and can be evaluated as follows:

$$L(X|v_r, W) = -g(X|v_r) +$$

$$+ \quad 1/\eta \log \frac{1}{R-1} \sum_{\substack{f=1 \\ f \neq r}}^{R} \exp[-\eta \cdot g(X|v_f)], \eta > 0 \tag{3.5}$$

In our experiments we used likelihood weighting in the postprocessing stage of the recognition process. A state level word segmentation was evaluated prior to score weighting. Our attempts to include state weights during segmentation revealed that such a procedure caused degradation in recognition accuracy. The use of postprocessing score weighting for a 500 word vocabulary decreased the word error rate from 8.5% to 8.2%. For subsets of the 500 word vocabulary consisting of 100 words each, the average word error rate was reduced from 4.8% to 4.4% after the score weighting. For subsets of the 500 word vocabulary consisting only of 20 words, postprocessing score weghting reduced the word error rate from 1.8% to 1.2%. The last result corresponds to 33% of word error rate reduction.

4. CONCLUSIONS

In this paper we have shown a possibility of using ML criterion for the phonological rules derivation. Derived phonological rules can be used to create pronunciation networks for any new word. In the process of this research we have found that knowledge of word baseform pronunciation is a crucial factor in the construction of the high quality pronunciation networks. Also different constraints (such as a minimum allowable number of phonemes in the strings subjected to the rule conversion or a ban on the boundary phoneme deviation) can substantially improve the recognition rate. The principles of phonological rule construction described in this paper can be extended to continuous speech recognition task.

Pronunciation networks constructed by applying statistically derived phonological rules can be further optimized by using discriminative likelihood score weighting. Our studies have shown that such a technique provides different levels of the recognition rate improvement with most noticeable results occurring in a small vocabulary task (of a few dozen words). The use of discriminative weights is justified at the postprocessing phase of recognition.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. C.H. Lee and Dr. F.K. Soong for their suggestions and supports during the process of this research.

5. REFERENCES

[1] Riley, M., Ljolje, A., Hindle, D., and Pereira, F., "Automatic generation of detailed pronunciation lexicons" *In Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F.K. Soong, and K.K. Paliwal, Eds. Kluwer Academic, Boston, March 1996, ch. 12.

[2] E. Giachin, A. Rosenberg and C.-H. Lee, "Word juncture modeling using phonological rules for HMM-based continious speech recognition," *Computer Speech and Language*, **5**, 1991.

[3] T. Holter and T. Swendsen, "Maximum Likelihood Modeling of Pronunciation Variation," *Proc. Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc*, pp. 63-66, May, 1998.

[4] F. Korkmazskiy, B.-H. Juang and S. Katagiri, "Discriminative Training of the Pronunciation Networks," *Proc. Automatic Speech Recognition and Understanding, Santa Barbara*, pp. 223–229, December, 1997.

[5] K.-Y. Su and C.-H. Lee, "Speech Recognition Using Weighted HMM and Subspace Projection Approaches," *IEEE Trans. on Speech and Audio Processing* vol. 2, no. 1, pp. 69–79, January, 1994.