DICTIONARY REFINEMENTS BASED ON PHONETIC CONSENSUS AND NON-UNIFORM PRONUNCIATION REDUCTION

Gustavo Hernández-Ábrego, Lex Olorenshaw

Spoken Language Technology
Sony Electronics, Inc. PTCA
3300 Zanker Rd, San Jose CA 95134
gustavo@slt.sel.sony.com

Raquel Tato, Thomas Schaaf

Speech and Sound Group
Sony Corporate Labs Europe
Hedelfinger Str. 61
70327 Stuttgart, Germany

ABSTRACT

In this paper we present a procedure to refine the recognition dictionary based on a composite approach to prune the unneeded pronunciations. First, pruning is applied in a non-uniform manner according to the characteristics of each word. Even though this straightforward operation may produce high-quality dictionaries, it makes the refined dictionary heavily dependent on the data used in this process. For the words not observed in the data, we propose, in second place, to use multiple sequence alignment techniques in order to find phonetic consensus among the pronunciation variants and select the worthy pronunciations that will represent the unobserved words. Experimental results show that our dictionary refining method helps to improve the recognition performance in two relevant aspects: it increases the recognition accuracy by reducing the cross-word confusibility and it improves the recognition speed by reducing the complexity of the search space.

1. INTRODUCTION

Due to the variability inherent to speech communication, the use of pronunciation variants in speech recognition is a natural and often used solution to increment recognition accuracy. Pronunciation variants allow coverage of the particular ways in which one word is said across different speakers, speaking styles, speech stress, dialectal variations, speaking rates, etc. In state-of-the-art recognition systems, it is common to produce pronunciation variants automatically, based on phonological rules that are applied to all the words included in the recognition dictionary. Rules like these tend to over-generate the pronunciation variants, no matter how accurately the set of rules was designed. On the other hand, these rules rarely consider the pronunciations across different words. They tend to introduce similar pronunciations to represent different words. Cross-word confusibility hinders recognition accuracy [1] and, therefore, it is highly desirable to eliminate it. Moreover, to over-populate the recognition dictionary with alternative pronunciations results in a considerable increment of the search space and represents a considerable drawback for its practical applications.

A natural solution to the aforementioned problems is to avoid adding unnecessary pronunciation variations. This would require a word-by-word approach to pronunciation design which, by definition, results in a tedious process that is hard to generalize to new words or languages. Another possible solution is to remove the excess of pronunciations. Here, it is assumed that the original dictionary contains sufficient pronunciations in every word to accurately represent the diverse spoken instances of it. It is hard to demonstrate the correctness of this assumption. Nonetheless, pronunciation pruning remains popular. Pruning has been tried from different approaches with remarkable success [2]. Many of the data-driven pruning methods are based on the premise that the pronunciations to be kept are the ones that were observed when forced aligning a set of speech samples against their word transcriptions. The forced alignment is usually done with large speech databases like the ones used for acoustic model training. This solution has been effective when there is enough speech material in the training database for a large portion of words in the dictionary to be treated in this way. However, suitable speech material is not always available in sufficient amounts for this purpose [3].

To cope with data limitation in a data-driven method for dictionary pruning, we propose the introduction of phonetic string consensus to deal with the pronunciations of unobserved words. Phonetic consensus can be seen as a condition primarily independent of the training data. On the other hand, when dealing with words that do have pronunciations observed in the training database, we use the nature of the word to alter the pruning process. This contrasts with previous pruning and pronunciation modeling approaches, where the same rules and the same pruning strategies are applied to all the words in the dictionary regardless of the particularities of each word.

In the following section, the concepts of phonetic consensus and non-uniform pruning are explained in detail. In section 3, the method to combine the data-driven method with phonetic consensus to prune the dictionary is described. Experimental results are presented and discussed in section 4 and our concluding remarks are given in section 5.

2. CONSENSUS AND NON-UNIFORM PRUNING

When dictionary pruning is based on the pronunciations observed on the speech training database, two problems arise: 1) words that were not included in the data do not have information to be treated with and 2) some words tend to keep pronunciations that were rarely observed.

To solve the unobserved words problem, we propose to use "central" or "summary" pronunciations in the pruned dictionary. The aim of this sort of pronunciation is to capture the phonetic contents included in the set of pronunciation variants of each word and to consolidate them in a reduced or summarized pronunciation set. If every pronunciation is to be understood as a sequence of phones, the central pronunciation in a word can be formulated after a sequence analysis procedure. Sequence and string analysis has been extensively addressed by computational biology researchers. Multiple sequence alignment (MSA) is used to find central or consensus sequences that represent a common pattern within a sequence of proteins or genes [4]. It is not difficult to modify the methods of computational biology to handle phone and word sequence multiple alignment. The consensus pronunciation should not be confused with the "baseform" pronunciation. Generally speaking, the baseform pronunciation is a "canonical" representation that does not include pronunciation variants. The pronunciation variants are typically generated from such canonical pronunciation.

The consensus phone sequence is here defined as the sequence resulting from collecting the phones which are always present in the imaginary columns resulting from MSA. Figure 1 shows an example of multiple phone sequence alignment. From MSA, we formulate three output sequences. The sequence marked as *consensus* in figure 1 is rather restrictive. According to the previous definition, the minimum lack of homogeneity in the alignment of the original sequences affects consensus. The *majority* sequence is defined as the sequence resulting from selecting the phone that appears in the majority (50% plus 1) of the rows in every column of the aligned sequences. The *plurality* sequence is defined as the sequence of phones resulting from collecting the phone that appears the most at every MSA column.

In the remaining of this paper, we will refer to this part of the procedure as the *consensus step* in general. However, this does not mean that the strictly defined *consensus sequence* will always be utilized to drive the pruning. The majority or plurality sequences might be used for this purpose as well.

```
              l ah cl k sh ax r iy
              l ah cl k sh er - iy
              l ah cl g zh ax r iy
              l ah cl g zh  - r iy

consensus   : l ah cl - -  - - iy
majority    : l ah cl - -  - r iy
plurality   : l ah cl - - ax r iy
```

Fig. 1. Alignment for the pronunciations of word "luxury"

In addition, we propose the use of a "non-uniform" approach to determine which of the pronunciations that were observed in the forced-alignment step will be kept in the dictionary. It is natural to expect that all the observed pronunciations are somewhat relevant for recognition purposes. This is not necessarily the case when the observation frequency is very much concentrated in a subset of the pronunciation variants. We consider that pronunciations that are rarely observed in the training data might bring only small benefit to the overall recognition performance at the high expense of adding complexity to the recognition operation. The problem with this consideration is that the statistics for the observed events depend on the amount of the data used in forced alignment. To reduce the quantitative effects introduced by the amount of speech material, we propose to use criteria relative to the number of observations of each word in order to determine which of the pronunciations observed are considered spurious and should be pruned off. In such a way, highly-observed pronunciations are always kept while the not-so-frequent ones are removed. We go further with this concept of non-uniform pronunciation pruning and apply different refinement strategies depending on the nature of each word. It is not difficult to relate certain word features to the need to prune them more or less aggressively. In many of the multi-pronunciation dictionaries used for speech recognition, it is natural to have a larger number of pronunciation variants in longer words. This is because the phonological rules tend to produce the alternative pronunciations in a combinatorial manner. In contrast, the number of pronunciations in short words tends to be small. The irony of this condition is that, in English as in several other languages, short words are more frequent and contribute considerably more to the overall recognition error than long and infrequent words. We consider that many more pronunciations should be removed from the long words while the short ones should keep all those pronunciations that help to model them accurately. In such a way, pruning is not only non-uniformly applied to the pronunciations of a word but it is also non-uniformly applied to the words of a dictionary.

3. PRUNING METHOD

Pronunciation reduction is handled differently depending on whether a word was observed or not during forced alignment:

1. Observed: non-uniform pruning is applied
2. Unobserved: pruning based on consensus is applied.

For the words observed in the forced alignment, two different kinds of pruning strategies are used depending on their length: a) cumulative observation thresholding for the short words, and b) thresholding over the standardized distance to the most observed pronunciation for the long words. Words with more than 7 characters are considered long. It is also possible to define word length in terms of the number of phones but we did not find evidence that such consideration has an impact on the pruning results.

For the short words, the observed pronunciations are sorted in descending order based on the frequency of occurrence. The set of pronunciations that account for a cumulative 95% of the observations is kept in the dictionary. If one of the kept pronunciations was observed only once, it is also removed from the dictionary. If there is a word with pronunciations that were all observed only once, one of those pronunciations is randomly chosen. This straightforward criterion helps to keep most of the observed pronunciations while it removes the infrequent ones.

The long words are pruned more aggressively. For every observed pronunciation, we find the standardized distance between such pronunciation and the pronunciation with the highest number of observations. The standardized distance, similar to the one--dimensional case of the Mahalanobis distance, can be computed according to:

$$d_j(w_i) = \frac{\left|\max\left(\vec{Q}(w_i)\right) - q_j(w_i)\right|}{\sigma_{\vec{Q}(w_i)}} \tag{1}$$

where q_j is the observation count for the j-th pronunciation for word w_i and $\vec{Q}(w_i)$ is the vector of observation counts for all the pronunciations for that word. This distance is normalized by the standard deviation of vector $\vec{Q}(w_i)$. A pronunciation is pruned off the dictionary when its distance $d_j(w_i)$ is bigger than an arbitrary threshold that in this case is set to be 2.

The key issue about non-uniform pruning is to be able to define when to use each pruning strategy and which thresholds to use for that purpose. Our definition of "long" and "short" are arbitrary, and so are our pruning thresholds. Nevertheless, these settings have the clear purpose to eliminate many unworthy pronunciations while keeping the ones that would reliably model each word.

It seems apparent that the quality of a dictionary treated this way will very much depend on the amount and quality of the data used to select the pronunciations. As mentioned in section 2, it is possible to use the consensus criterion to eliminate unneeded pronunciations from any word in the dictionary. It is irrelevant if it was observed in the data or not because consensus is computed from within the dictionary only.

The example in figure 1 is useful to illustrate how to summarize the phonetic contents of the alternative pronunciations through consensus. Out of the three phonetic sequences resulting from the MSA step, the plurality sequence seems the less restrictive. Pruning is carried out by eliminating all the pronunciation variants but the one that has exactly the same phonetic contents as the plurality sequence. However, as seen in figure 1, the plurality sequence may not represent one of the original pronunciations for the word in question. It seems inaccurate to use the plurality sequence as the unique pronunciation in this case. We do not want to generate new pronunciations that might include only partial information. Instead, the pronunciation variant (or set of variants) closest to the plurality sequence will be used. In this case, we use yet another phonetic alignment step to determine the distance from the plurality sequence to each one of the pronunciations in a pair-wise alignment. From this alignment, the phone-level Levenstein distance is computed. The pronunciation or pronunciations with shortest distance will be used to represent a valid summary for the original set of pronunciation variants. In the example in figure 1, the pronunciations `l ah cl k sh ax r iy` and `l ah cl g zh ax r iy` have the same (and the shortest) distance to the plurality sequence. Therefore, these pronunciations are the only ones to be kept to represent the word "luxury" in the refined dictionary.

It is worthwhile noticing that no explicit symbol-to-symbol weights were used in any of the alignments. A single set of substitution and gap weights was used in the MSA and in the pair-wise alignment. The inclusion of a distance matrix for the MSA, possibly based on phonetic similarity, could be used as a way to find a more representative set of consensus pronunciations.

4. EXPERIMENTAL DEVELOPMENT

After refining the dictionary with this method, recognition performance is evaluated over one American English spontaneous speech database. This Sony-recorded testing database is regarded as "counseling domain". It includes spontaneous conversations held in an informal chatting mode on a number of different topics primarily related to a counseling session. In the testing database there are 16 different speakers (8 male, 8 female). The total number of utterances in the database is 469 for a total of 0.84 hours of acoustic material (speech and silence included). In this testing, only the "patient" side (which is assumed to be the most spontaneous party) of the counseling session is used. Speech was collected under clean studio conditions.

The acoustic model used is a Gaussian Mixture HMM with 2000 states. After state-tying, there are 24 Gaussian mixtures per state. The front-end parameterization is a 26 LDA-transformed one from an original 38 MFCC that includes static, first and second derivatives but excludes C_0. This model was primarily trained with spontaneous and conversational speech directly collected by Sony.

The language model (LM) was generated by interpolating different LMs trained with texts from different sources. The perplexity of the interpolated LM evaluated on the "counseling" test described above is 90.88 with an OOV rate of 0.23 %.

Arthur [5], the Sony-built recognition engine, is used for recognition experimentation. Its several parameters are automatically set up in an auto-tuning fashion [6]. Auto-tuning is run over a development testing data set in a cross-validation way.

Regarding the dictionary, there is one initial (baseline) dictionary that includes entries for 64,002 words. The pronunciation variants for the baseline dictionary were automatically generated by allophonic rules. The average number of pronunciations per word in this dictionary is 7.62.

In our experimental development, our objective was manifold: 1) to verify that dictionary refinements can bring consistent improvements to the recognition system, 2) to measure the impact of the nature of the data used to prune the dictionary and 3) to evaluate the effects of using pronunciations refined through phonetic consensus.

The recognition results for these experiments are presented in table 1. Although the recognition system was tuned to perform at maximum accuracy, two recognition attributes are considered for recognition evaluation: word error rate and computational time. Time is measured in real-time factors (RTF) for the recognizer running on a Sun Sparc workstation at 400MHz. We also include in table 1 the number of pronunciations in each dictionary.

Exp	Refinement type	# Pron	WER	RTF
1.1	Baseline	488035	23.59	12.8
1.2	Sony data	14741	23.22	7.8
1.3	Sony & BN data	30440	22.89	9.8
1.4	Sony data + consen.	115996	22.23	9.1
1.5	Sony & BN data + consen.	110580	22.67	10.0
1.6	only consensus	118883	36.40	8.2

Table 1. Recognition evaluation

To answer the first two questions, we first tried refining the dictionary using our own Sony training speech database which includes 66 hours of speech and is similar in nature to the testing database. Recognition results (experiment 1.2 in table 1) show that indeed the dictionary refinements have an impact over the recognition accuracy when compared against the baseline (exp. 1.1). Although similar in nature to the testing database, the Sony database was not able to deal with a large number of words and their pronunciations. Only 10,566 of the original dictionary words were observed in this data. The unobserved words that could not be treated this way had to be removed from the refined dictionary. This resulted in an OOV rate of 0.83 % over the testing database. To try to overcome this situation, a larger acoustic database (independent of our testing data) was added to the data used to prune the dictionary. We chose to use the 1996 Broadcast News acoustic training material [7] for this purpose. In such a way, we were able to prune the pronunciations of 25,302 words in our dictionary using the 50 hours of speech in this database. The OOV rate of the refined dictionary dropped to 0.64%. Recognition results also showed improvement (exp. 1.3) making evident that the amount of data used for pruning directly affects recognition results. The next question was related to the impact of using words pruned through consensus to complement the words pruned with forced-aligned data. When we used 10,566 words pruned with the Sony data and complemented them with the consensus pronunciations for the words that were not observed in such database the OOV rate is brought back to the baseline level (0.23 %), and we noticed a considerable performance improvement (experiment 1.4). Following this lead, we also tried complementing the 25,302 words treated with the Sony+Broadcast News data with the consensus pronunciations for the unobserved words (exp 1.5). Interestingly, we found that, in spite of keeping the same low level of 0.23 % OOV, the resulting dictionary produced lower accuracy. We think that Broadcast News is somehow adapting many words to its speaking style and therefore removing some pronunciations that are valuable to the counseling test. We believe that consensus does not produce this damaging effect. In order to try to gain more insight into this question, we tried doing recognition with a dictionary that was refined based only on consensus. No data was used this time for the pruning (exp. 1.6).

Results in table 1 show that the best way to refine the recognition dictionary is by using data close in nature to the data used in testing and by using consensus to deal with the words that were not included in the data used for pruning. It is worthwhile noticing that this approach may be superior to using more but possibly unrelated speech data for this purpose. However, consensus should be treated carefully. Table 1 also shows that the dictionary treated with consensus only does not retain adequate phonetic content, and thus accuracy drops considerably.

It seems fair to say that recognition accuracy can be improved through these dictionary refinements because the cross-word confusibility has been reduced while the effective modeling capabilities of the dictionary have not been seriously affected. One other motivation of this work was to evaluate how these refinements can help to reduce the complexity of the search space. We consider that this condition can be measured through the recognition computational time. In our first round of experiments, the effects of search space reduction are not very noticeable. This might be due to the fact of having a recognition system tuned to reach maximum accuracy. We ran once again the automatic tuning process in order to reach maximum accuracy but at faster operation. The target of the tuning can be easily changed by modifying the recognition merit computation [6]. With these new parameter settings, we ran again some of the recognition experiments of table 1.

Exp	Refinement type	# Pron	WER	RTF
2.1	Baseline	488035	24.82	6.1
2.2	Sony data	14741	24.03	3.0
2.3	Sony & BN data	30440	23.24	3.5
2.4	Sony data + consen.	115996	23.29	3.2
2.5	Sony & BN data + consen.	110580	23.27	3.6

Table 2. Search space evaluation

Results in table 2 show that the baseline recognition time can be reduced by 50% (exp. 2.1) after changing the tuning. In general, the accuracy levels are not as high as they were in the previous experiments but they have not dropped drastically. Accuracy improvements from the refined dictionaries look similar to what was observed in table 1. Search space reduction is more noticeable when the recognition system is adjusted to perform at higher speeds. While the new baseline has been able to see its computational time halved, the systems with the refined dictionaries have reduced their computational time by a bigger proportion. The dictionary that shows highest accuracy and speed combined is again the one pruned with data close in nature to the testing data and complemented with the consensus pronunciations for the unobserved words (exp. 2.4). These experiments show that, with a properly refined dictionary, accuracy can be improved with only a quarter of the computational effort spent in the non-refined dictionary (exp. 1.1).

The results of tables 1 and 2 show that this method is advantageous because there is no need to perform dictionary refinements using extensive and generic databases. To use data related to the testing database seems more rewarding. The consensus approach allows using non-generic databases because it is able to effectively deal with the words not included in the databases used for pruning. This opens the possibility to adapt the recognition dictionary to specific recognition tasks.

5. CONCLUDING REMARKS

Our experiments with the composite method for dictionary refining showed that: 1) refined dictionaries can consistently improve the speech recognition results, 2) non-uniform pruning can effectively be used to refine the dictionary, but it strongly depends on the data used, and 3) phonetic consensus is a valuable means to prune the pronunciations of unobserved words. After pruning, the dictionaries show reduced confusibility, which brings higher accuracy, and reduced complexity, which results in a reduced search space and faster recognition. Our results also show that it is more advantageous to use pronunciations refined through consensus than using data for pruning that is not related to the testing target. With our method, it is possible to effectively prune dictionaries with non-generic databases with limited lexicon and to adapt them to particular applications or speech modalities.

6. REFERENCES

[1] J. M. Kessens, C. Cucchiarini, and H. Strik, "A data-driven method for modeling pronunciation variation", *Speech Communication*, vol. 40, no. 4, pp. 517–534, June 2003.

[2] T. Hain, "Implicit pronunciation modelling in ASR", in *ITRW PMLA 2002*, Estes Park, CO, 2002, Invited Short Lecture.

[3] L. ten Bosch and N. Cremelie, "Pronunciation modeling and lexical adaptation using small training sets", in *ITRW PMLA 2002*, Estes Park, CO, 2002.

[4] D. Gusfield, *Algorithms on strings, trees and sequences*, Cambridge University Press, 1997.

[5] H. Lucke, H. Honda, and K. Minamino et al, "Development of a spont. speech recognition engine for an entertainment robot", in *Proceedings of SSPR 2003*, Tokyo, 2003.

[6] G. Hernández Ábrego, X. Menéndez-Pidal, T. Kemp, K. Minamino, and H. Lucke, "Automatic set-up for speech recognition engines based on merit optimization", in *Proceedings of ICASSP 2003*, Hong-Kong, April 2003, vol. I, pp. 196–199.

[7] D. S. Pallet, "The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program", *Speech Communication*, vol. 37, pp. 3–14, 2002.