IMPROVED PRONUNCIATION MODELLING BY INVERSE WORD FREQUENCY AND PRONUNCIATION ENTROPY

Ming-yi Tsai, Fu-chiang Chou, Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
Email: pancho@speech.ee.ntu.edu.tw

ABSTRACT

In this paper, we propose a new approach to rank the potential pronunciations for each word by their pronunciation frequency and inverse word frequency (***pf-iwf***) weights. The pronunciation set obtained in this way can then be pruned with different criteria. This approach not only considers the frequencies of occurrence of the pronunciations, but tries to minimize the extra confusion which may be introduced by the pronunciation variations, such that the best overall performance can be achieved. A new entropy-based approach for pruning the pronunciation variations is also proposed. Experimental results showed that the proposed approach can not only improve the recognition performance, but make the performance more stable and less sensitive to various parameters, factors and options including the different pruning criteria. All the experiments were performed with the LDC Mandarin Call Home corpus, although the approaches and principles are definitely not limited to Mandarin Chinese.

1. INTRODUCTION

It has been well known that the pronunciation variation in spontaneous speech may very often seriously deteriorate the performance of ASR systems. Pronunciation variation is usually modeled by enumerating appropriate pronunciations for each word in the vocabulary using a pronunciation lexicon, with a prior probability for each pronunciation. As have been observed earlier, simply adding several alternative pronunciations to the pronunciation lexicon naturally increases the homophone rate and hence may not be helpful to the recognition performance. In spontaneous speech, for example, some function words are often pronounced similar to other function words which may not be easily distinguished by the language model. To combat such potential confusion, one approach is to assign costs to alternative pronunciations [1]. For instance, if a frequent pronunciation of one word and an infrequent pronunciation of a different word are identical, a penalty is incurred when infrequent pronunciation is used rather than frequent one. However, how much penalty should be incurred to those infrequent pronunciations respectively is still a good question, and if doing this those words which admit more pronunciations may be unnecessarily penalized as compared to those with fewer pronunciations. Although this problem may be somehow alleviated by rescaling the pronunciation probabilities of each word, the introduced penalty has been a possible cause of the detrimental performance [1]. Besides, different approaches were also developed to determine which pronunciation variants are more appropriate to be augmented. Good examples for such approaches include the maximum likelihood criterion [2], the confidence measures [3], the degree of confusability between the variants [4], etc.. In this paper we proposed to use pronunciation frequency and inverse word frequency (***pf-iwf***) weights to rank the pronunciations, and then pruning criteria can be applied. In this different way the extra confusion which may be introduced by the pronunciation variations can be minimized.

In the following, section 2 summarizes the baseform generation procedures and section 3 presents three different baseform pruning criteria. The new baseform weighting and ranking approach is then discussed in section 4. The recognition experiments are given in section 5. The conclusion is finally made in section 6.

2. BASEFORM GENERATION

The main steps to acquire automatically the pronunciation confusion table are as follows:

1. Acquiring the canonic transcriptions for the training data.
2. Using the unconstrained phone recognizer to obtain the surface forms for the training corpus.
3. Aligning surface forms with the canonic froms using the dynamic programming algorithm
4. Generating the confusion table and obtaining the statistics at the word level.

3. BASEFORM PRUNING

Although adding variants to the existing pronunciation lexicon is straightforward, the added variants may cause extra confusion with other words and introduce new errors. Therefore, to select the best set of pronunciation variants from the confusion table to be added to the lexicon so as to minimize the total errors is crucial. This can be done with the traditional probability-based, or the count-based [5] pruning methods, and in this paper a new entropy-based pruning method is also proposed.

3.1. Probability-based Pruning

For each word, the pronunciations with priori probabilities less than a parameter α (which is empirically tuned) times of P_{max}, the probability for the most probable pronunciation, are not to be added into the lexicon.

3.2. Count-based Pruning

For each word, the number of the pronunciations to be added into the lexicon is a parameter β (which is empirically tuned) times of the log of the total count of the word in the corpus, truncated to an integer.

3.3. Entropy-based Pruning

Entropy has been found to be a good measure for the spread of pronunciations in a training set. For the pronunciation set of a word w_j with probability distribution estimates $p_{i,j}$ for different associated pronunciations v_i, the entropy H_j is defined as

0-7803-7343-X/02/$17.00 © 2002 IEEE

$$H_j = -\sum_i p_{i,j} \log_{10} p_{i,j} \quad \text{...(1)}$$

In the entropy-based pruning criterion proposed in this paper, the total number of pronunciations for a word w_j to be added into the lexicon is a parameter γ (which is empirically tuned) times of the entropy H_j, truncated to an integer.

4. BASEFORM RANKING

When handling the pronunciation variation, usually a list of potential baseforms was compiled under each eligible word in the vocabulary ordered by their frequency of occurrence in the corpus. When some pruning method is applied, several preceding pronunciation variants with higher frequencies are selected, while those with lower frequencies deleted. The problem to be addressed here by the proposed approach is the selected pronunciation variants with high enough frequencies may tend to be confused with pronunciations of other words, thus causing extra errors which won't occur without the pronunciation variation. Therefore the scheme proposed in this paper is to re-rank the pronunciation variants not only based on their frequencies of occurrence, but considering the possible confusion with other undesired words which may be introduced by the extra pronunciation variants. This possible confusion is represented by a measure called inverse word frequency (***iwf***) in this paper. This is analogous to the inverse document frequency (***idf***) used in information retrieval [6]. In information retrieval, an indexing term frequently appearing in many different documents usually implies low discriminating functions in identifying relevant documents. Therefore the importance of the indexing terms is re-ranked by the inverse of its frequency of appearance in different documents. Similarly, if a pronunciation variant also occurs in many other different words, this pronunciation variant may cause more confusion and should be ranked lower. After re-ranking the existing potential baseforms by the inverse word frequency (***iwf***), the pruning criteria mentioned above can follow just as usual.

In a formal formulation, assuming D is the size of the vocabulary and w_j (j=1,2,...D) represents a word in this vocabulary. Let T be the total number of different pronunciations for all the words in D, and v_i(i=1,2,...,T) stands for one pronunciation among them. A good example of the word confusion table as obtained by the steps summarized in section 2 is draw in **Table.1**, in which c_{ij} is the count for the pronunciation v_i which is realized for the word w_j. Note that many of c_{ij}'s in **Table.1** are zero, and c_{ij}'s in **Table.1** are not ranked by their values.

4.1. Pronunciation Frequency (*pf*)

As was done conventionally, the importance of a pronunciation v_i for a word w_j is assumed to be related to the number of times the word w_j is pronounced as v_i, normalized to all pronunciations of the word w_j

$$pf_{ij} = \frac{c_{ij}}{\sum_{i=1}^{T} c_{ij}} = P(v_i \mid w_j) \quad \text{...(2)}$$

This is called pronunciation frequency (***pf***) here in this paper, also analogous with the term-frequency (***tf***) in information retrieval. In information retrieval, an indexing term appearing more frequently in a document very often implies its higher relationship with the content of the document. Therefore the importance of an indexing term is related to the term frequency, very similar to the situation here. Conventionally, a list of pronunciations under each word ordered by this factor is to be pruned by a certain criterion as mentioned above.

	w_1	w_2	...	w_j	...	w_D
v_1	c_{11}	c_{12}	...	c_{1j}	...	c_{1D}
...		...				
v_i	c_{i1}	c_{i2}	...	...	...	...
...		...				
v_T	c_{T1}	c_{T2}	...	c_{Tj}	...	c_{TD}

Table.1 Word confusion table.

4.2. Inverse Word Frequency (*iwf*)

While the pronunciation frequency discussed above concerns the pronunciation within a word, the inverse word frequency (***iwf***) discussed here concerns the pronunciation occurrences across a group of confusing words. The original concept of ***iwf*** is that a pronunciation frequently occurring in many other words may introduce extra confusion and hence its importance should be repressed. This is analogous to the inverse document frequency (***idf***) in information retrieval. Therefore, the inverse word frequency for a pronunciation v_i may be initially defined as

$$iwf_i = \log \frac{D}{d_i} \quad \text{...(3)}$$

where D is the vocabulary size and d_i is the number of different words that may include the pronunciation v_i. This definition in equation (3) is almost identical to the definition of inverse document frequency in information retrieval. However, in this formulation all different words that include the pronunciation v_i are treated equally regardless of the different word frequencies as well as the different probabilities of the pronunciation v_i for these words. In order to incorporate these considerations, a better definition of the inverse word frequency for the pronunciation v_i under the word w_j is given by

$$iwf_{ij} = \frac{1}{\sum_{\substack{all\ k \\ k \neq j}} P(v_i \mid w_k) P(w_k)} \quad \text{...(4)}$$

4.3. Pronunciation Frequency (*pf*) and Inverse Word Frequency (*iwf*)

Combining ***pf*** and ***iwf*** formulated above, the importance of a pronunciation v_i under a word w_j can be given by a ***pf-iwf*** weight,

$$h_{ij} = (pf_{ij}) * (iwf_{ij})^a \quad \text{...(5)}$$

This expression is again analogous to the term frequency and inverse document frequency in information retrieval. The rationale is that, for each word w_j, a pronunciation v_i of it is assigned a weight h_{ij}, which should be higher if the pronunciation occurs more frequently for the word w_j, but should be lower if it appears to be a frequent pronunciation of other different frequently used words w_k, k≠j. if the parameter a is set equal to zero, this weight h_{ij} is reduced to the traditional concept of pronunciation frequency (pf_{ij}). But when the inverse word frequency (iwf_{ij}) is included in h_{ij} (i.e. a>0), the confusion for the pronunciation across a group of different words can be considered. Therefore, the existing pronunciations for each word w_j can be first ranked according to the ***pf-iwf*** weights h_{ij} and then those pronunciations with relatively higher weights h_{ij} will be more qualified to be selected by a pruning criterion. Any pruning

那個 (that)	
Ranked with *pf* only	Ranked with *pf-iwf*
/n Ei g e/	/n Ei g e/
/n Ei/	/n Ei g uo/
/n Ei g uo/	/n Ei g ai/
/n Ei g Ei/	/n Ei g Ei/
/n Ei g ai/	/n Ei g E/
...	...

Table.2 Pronunciations of an example word "那個(that, with canonic pronunciation /n Ei g e/)" ranked by *pf* only and *pf-iwf* weights respectively.

criterion mentioned above (e.g. probability-based, count-based or entropy-based) can then be applied.

An example is shown in **Table.2**, for the word "那個(that, with canonic pronunciation /n Ei g e/)", a more frequent alternative pronunciation /n Ei/ suffering the deletion of the second syllable /g e/ is seriously confused with the pronunciations of a few other commonly used words, including "內(inside, with canonic pronunciation /n Ei/)", and "哪 (where, with canonic pronunciation /n Ei/)". As a result, even if the pronunciation /n Ei/ was ranked higher when only the pronunciation frequency (*pf*) was used, the ranking became very low and even pruned when the *pf-iwf* weight h_{ij} was used so as to avoid any possible confusion.

5. EXPERIMENTS AND RESULTS

The experiments were performed with the HTK tools on a part of the Mandarin Call Home corpus. After removing the laughters, filled pauses, corruptive background and channel noise, and those words in other languages, about 5.57 hours of speech in Putonghua accent was used in training, including 2.91 hours for male and 2.66 hours for female. It was used to train the gender-dependent acoustic models consisting of 58 three-state Initials and 22 four-state Finals, with 24 Gaussian mixtures per state, regardless of the tone. Here all Initial/Final models were context independent. The acoustic features were 13 MFCCs, 13 delta MFCCs and 13 acceleration MFCCs. The trained acoustic models were used both to acquire the surface forms and to perform the recognition experiments as well. Another set of 44 minutes of data in the same corpus outside the training set annotated with reasonable level of quality (channel noise, background noise, crosstalk, difficulty and distortion, as annotated on the Call Home Corpus) and the same accent was taken as the evaluation set. We also used another set of 30 minutes of data with the same accent from the remaining of the corpus to be the fast turnaround development set. A lexicon of 10321 words was used, most of which were present in the 5.57 hours of training data. A bigram language model trained using the text transcription of about 128K words from the training set regardless of accent was used.

In order to contend with the inevitably high rate of errors occurring in the free-Initial/Final decoding process, only those words occurred sufficiently frequently in the training set were considered to provide reliable enough pronunciation variation statistics. A threshold of 200 samples was set as the reference for such reliable statistics in the experiments, and as a result a total of 57 words may have alternative pronunciations. These 57 words actually covered about 53% of the evaluation data. The prior probabilities for all baseforms v_i under a word w_j are

		Acc.	Del.	Sub.	Ins.
Baseline*(canonic)*		27.07	13.95	56.47	2.52
Probability-based	*pf*	27.48	11.62	56.71	4.19
	iwf	28.58	12.29	55.77	3.36
Count-based	*pf*	27.09	11.32	57.01	4.49
	iwf	28.28	12.28	56.11	3.33
Entropy-based	*pf*	26.43	11.11	57.01	4.49
	iwf	28.69	12.29	56.11	3.33

Table.3 Recognition performance of the baseline and the three pruning criteria with *pf* ranking and *pf-iwf* ranking respectively.

proportional to their frequencies c_{ij} within the word, but rescaled so as to be summed up to unity after all ranking and pruning processes.

Some typical experimental results are shown in **Table.3**, including the character accuracy, deletion, insertion and substitution rates for the three pruning approaches applied with *pf* ranking only and *pf-iwf* ranking. The results in **Table.3** are for the case that even with different ranking and pruning approaches the average number of pronunciation per word was always tuned to be 1.008. As can be seen in this table, ranking the pronunciation variants by the proposed *pf-iwf* weights h_{ij} and then pruning by any of the three criteria always resulted in better performance than those of ranking by *pf* only. Besides, it was found from **Table.3** that the count-based pruning did not perform as well as the probability-based pruning with either ranking scheme. This is in good agreement with the previous investigation for Mandarin Chinese [7]. It was found previously [7] that in Chinese language quite many most frequently used words are mono-character words with relative less pronunciation variations. This may be the reason why count-based pruning approach works worse than the probability-based pruning here. This is why a new entropy-based pruning approach was proposed in section 3.3. Since the entropy is a good measure for the spread of the pronunciations of a word, adding the pronunciation variants into the dictionary based on the entropy of each word may make better sense than simply based on the word frequencies. As shown in **Table.3**, the proposed entropy-based pruning performed slightly better than the probability-based pruning if the *pf-iwf* ranking was used, although it was worse if the ranking was base on *pf* alone. Note that the pruning approaches are to determine the appropriate number of pronunciations to be included in the dictionary, its performance therefore has to be conditioned on the existence of a good set of pronunciations with correct ranking. Without the correct ranking, a good pruning criterion may perform worse. Another nice observation on **Table.3** is that, with the *pf-iwf* ranking, the character accuracy became much less sensitive to the different criteria. In other words, as long as a correct ranking is given, the selection of pruning criterion may become less important.

More complete results for average number of pronunciations per word ranging from 1.000 to 1.012 with *pf* ranking and *pf-iwf* ranking for the three different pruning approaches are plotter in **Fig.1**, **2** and **3** respectively. First consider the case of probability-based pruning with *pf* ranking (the lower curve) in **Fig.1**, where the point of 1.000 pronunciation per word is the

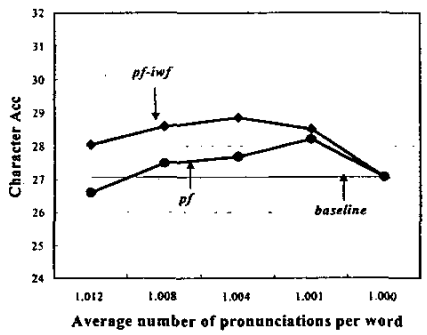

Fig.1 Character accuracy for probability-based pruning with *pf* and *pf-iwf* ranking respectively.

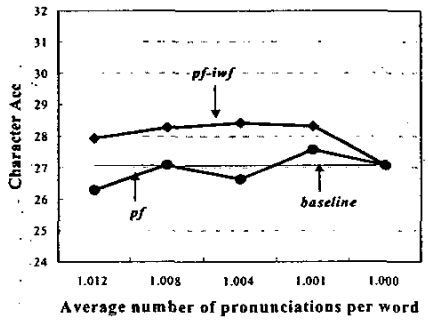

Fig.2 Character accuracy for count-based pruning with *pf* and *pf-iwf* ranking respectively.

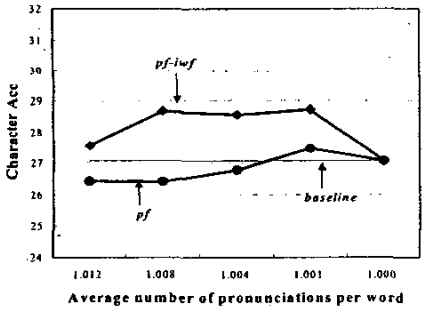

Fig.3 Character accuracy for entropy-based pruning with *pf* and *pf-iwf* ranking respectively.

baseline with canonic forms. It can be found that the performance was improved when the pronunciations were increased from 1.000 to 1.001 per word, but degraded when the pronunciations are further increased. Apparently some more words can be corrected by limited number of alternative pronunciations, but too many alternative pronunciations naturally brought more confusion. The performance even became worse than that of the baseline when the average number of pronunciations per word is 1.012. However, for probability-based pruning with *pf-iwf* ranking (the upper curve) in **Fig.1**, the performance not only was significantly higher, but remained relatively high when the average number of pronunciations was increased all the way up to 1.012. We can also find from **Fig.1** that when the average number of pronunciations per word is only 1.001, both ranking techniques perform very similarly since not too much confusion was introduced. However, when the number of pronunciations were added increased, ranking by *pf* suffered more confusion but ranking by *pf-iwf* didn't. In other words, pruning with *pf-iwf* ranking performs not only better, but more stably over a considerable range for the size of pronunciation variations. Similar trends can be observed for the count-based and entropy-based pruning cases in **Fig.2** and **3**. By comparing the upper curves in **Fig.1**, **2** and **3**, it is also observable that with *pf-iwf* ranking the performance is more robust to not only the size of pronunciation variations, but the pruning approaches used.

6. CONCLUSION

In the pronunciation modeling problem, it is not only the number, but also which variations to be added. The good performance of a pruning criterion also relies on a good pronunciation set with correct ranking. The approach proposed in this paper to rank the potential pronunciations for each word by its *pf-iwf* weights was shown to provide improved performance with different pruning criteria, since it determines an appropriate pronunciation set to be selected. It is also found that with the proposed *pf-iwf* weighting, the recognition performance becomes much less sensitive to the many empirical factors, used in the experiments.

7. REFERENCES

[1] M. Riley et al., "Stochastic Pronunciation modeling from hand-labelled phonetic corpora," Speech Communication, pp. 209-224, 1999.

[2] T. Holter et al., "Maximum Likelihood Modelling of Pronunciation Variation," Speech Communication, pp. 177-191, 1999.

[3] D. A. G. Williams, "Knowing What You Don't Know: Roles for Confidence Measures in Automatic speech recognition," Ph.D. thesis, University of Sheffield, Sheffield, England, 1999.

[4] M. Wester et al., "A Comparison of Data-driven and Knowledge-based Modeling of Pronunciation variation," Proc ICSLP., 2000.

[5] E. Fosler et al., "Not just what, but also when: Guided Automatic Pronunciation Modeling for Broadcast News," DARPA Broadcast News Workshop, 1999.

[6] R. Baeza, B. Ribeiro, "Modern Information Retrieval," ACM Press, New York, 1999.

[7] Ming-Yi Tsai et al., "Pronunciation Variation Analysis with respect to Various Linguistic Levels and Contextual Conditions for Mandarin Chinese," Eurospeech, 2001.

[8] F. Korkmazskiy and B.-H. Juang, "Statistical Modeling of Pronunciation and Production variations for Speech Recognition," Proc. International Conference on Spoken Language Processing, Sydney, pp. 149-152, December, 1998.

[9] H. Strick et al., "Modeling pronunciation variation for ASR: A survey of the literature", Speech Communication, pp. 225-246, 1999.

[10] E. Fosler, "Dynamic Pronunciation Models for Automatic Speech Recognition," Ph.D. thesis, University of California, Berkeley, 1999.