JOINT PRONUNCIATION MODELLING OF NON-NATIVE SPEAKERS USING DATA-DRIVEN METHODS

Ingunn Amdal, Filipp Korkmazskiy and Arun C. Surendran

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974, USA
amdal@tele.ntnu.no, {yelena,acs}@research.bell-labs.com

ABSTRACT

Modelling non-native speakers with different mother tongues is a difficult task for automatic speech recognition due to the large variation among speakers. One possibility for jointly modelling all speakers is to use the same speaker independent acoustic models and a joint lexicon to capture the variation.

We have modified the reference lexicon using pronunciation rules that are derived in a totally data-driven manner from a set of adaptation data using the reference recognizer and the reference lexicon. Deriving common rules for such diverse sources simultaneously is difficult. The challenge is to combine these rules to a common set without increasing the confusability.

In this paper we compare several methods of combining the individual rules to form a common lexicon for all speakers. Using a new log likelihood rule pruning measure presented in this paper, we achieved improved performance compared with more traditional rule pruning methods based on rule probability, and with much fewer rules. With a confusability reduction scheme we reduced the number of rules even further.

1. INTRODUCTION

Due to the large variation among non-native speakers, constructing a single lexicon for speakers with different mother tongues is a difficult task. In this paper we have addressed this tough problem using a rule based method that is totally data-driven. The motivation for this is: 1) Linguistic information about different accents may not be available. 2) A data-driven approach allows the use of the same optimization metrics as used for acoustic and language modelling. Data-driven techniques have been used before [1, 2, 3, 4], and they usually follow these four steps:

1. automatically generate alternative transcriptions
2. align the reference and alternative transcriptions
3. derive initial rules from the alignment
4. prune the initial rules

The issue we would like to focus on is rule pruning. We propose to use a new log likelihood measure for rule pruning, as this is consistent with the training part of the recognizer. We compute an improvement measure for each rule by adding the improvements in log likelihood seen when assessing the rules on the adaptation set. This will give a higher score to rules that result in pronunciations that are more likely given the adaptation data and the acoustic models. We have compared this to the use of estimated rule probability in the rule pruning.

Many pronunciation variants in the lexicon may increase the confusability and thereby the error rate. Adding variants also slows down recognition. We have considered several confusability reduction schemes to remove confusable rules.

The paper is organized as follows: Different rule pruning measures, including the new log likelihood based improvement measure, are described in Section 2. Section 3 describes our approach to pronunciation modelling, and Section 4 gives a brief overview of the experiments. The results are presented in Section 5, and a summary is given in Section 6.

Ingunn Amdal is a research fellow at the Norwegian University of Science and Technology. This work was done while visiting Bell Labs.

2. JOINT RULE BASED LEXICON MODIFICATION

In this paper we focus on part 4 in the rule pronunciation modelling procedure, the rest of the system will be described in Section 3. In our pronunciation rule formulation a *rule source segment* consists of the affected phone `A` as well as the two neighbouring phones `x1` and `x2`. A rule for mapping `A` to `B` can be written as: `x1-A+x2` $\rightarrow$ `B`, where `B` is the *target* of this rule. `B` can be a single phone (substitution of `A` with `B`), several phones (insertion/substitution), or `DELETED` (deletion of `A`).

2.1. Rule Probability Based Rule Pruning

Rule pruning by using estimated rule probability is frequently used, e.g. [1, 3]. The rules are sorted according to the estimated rule probability and the rules below a threshold are discarded. As in [1] we have used frequency counts in the estimation. All mappings from each rule source segment to each of the target phone(s) or deletions are counted. From the alignment the occurrence of all rule source segments are also counted. An estimate of the rule probability is the ratio between these counts:

$$\hat{P}(\texttt{x1-A+x2} \rightarrow \texttt{B}) = \frac{\text{count}(\texttt{x1-A+x2} \rightarrow \texttt{B})}{\text{count}(\texttt{x1-A+x2})} \quad (1)$$

To reduce the number of rules without reducing performance, we have modified this rule probability pruning measure. This is done by retaining the most useful rules, i.e. the rules with highest rule source segment probability. This probability was estimated from frequency counts on the task lexicon (the 5k WSJ lexicon, see Section 4).

2.2. Log Likelihood Based Rule Pruning

We propose to use log likelihood improvements as a rule pruning measure, because this is more consistent with training of the rest of the recognizer. The acoustic models are trained using a maximum likelihood formulation and we have therefore chosen to use the same metric for rule pruning. We compare the log likelihood of the pronunciations affected by each rule with the log likelihood of the corresponding reference transcription. The measure will thus be a log likelihood ratio, giving a normalization that makes rule comparison easier. The rule pruning measure $\mathcal{I}$ for an acoustic segment x_i affected by a rule using log likelihood improvement is:

$$\begin{aligned}\mathcal{I}(x_i) &= \log\left[\frac{p(x_i|B_i^{\text{alt}},\theta)}{p(x_i|B_i^{\text{ref}},\theta)}\right] \\ &= \log[p(x_i|B_i^{\text{alt}},\theta)] - \log[p(x_i|B_i^{\text{ref}},\theta)]\end{aligned} \tag{2}$$

Here, B_i^{ref} is the reference pronunciation for the word belonging to x_i, and B_i^{alt} is the alternative pronunciation for the same word after modification according to the rule we want to assess. θ is the set of parameters of the recognizer used in computing the log likelihoods.

For each rule, we combine the positive contributions from the words affected to compute an improvement measure $\mathcal{M}$:

$$\mathcal{M}(\text{rule } k) = \sum_{k' \in \mathcal{K}} \mathcal{I}(x_{k'}) \tag{3}$$

Here, $x_{k'}$ is an acoustic segment for a word affected by rule k, and we sum over all such segments $\mathcal{K}$ where $\mathcal{I}(x_{k'}) > 0$. We do not include negative log likelihood contributions because we always keep the reference transcription in the lexicon, and $\mathcal{I}(x_{k'}) < 0$ will mean that the reference pronunciation will be chosen. All positive contributions are added so that rules that are applied more frequently are favoured. This is deliberate, because we assume the rule source segments which are most frequent in the adaptation also are most useful in testing. If this is not the case, a weighting factor should be applied.

2.3. Confusability Reduction

Because rule source segments that occur often will get a relatively higher score in our improvement measure $\mathcal{M}$, we may get a lot of confusable rules, i.e. rules with the same rule source segment, but different target phone(s). These rules may be modelling the same variation and add superfluous complexity and confusability. An example is the rules `sh-ax+n` $\rightarrow$ `ah`[1] and `sh-ax+n` $\rightarrow$ `eh`, where the actual pronunciation often seems to be somewhere in-between `/ah/` and `/eh/`. We have applied a confusability reduction approach by restricting the rule set to consist of at most one rule per rule source segment. We retained the rule with highest log likelihood improvement $\mathcal{M}$. (In the example this is `/ah/`.)

The improvement measure $\mathcal{M}$ only assesses the performance of a rule on the correct word. To achieve a discriminative effect, the performance on the other words in the vocabulary will have to be assessed. A misclassification measure for an acoustic segment x_i belonging to the word i, compared with a new pronunciation for another word B_j^{alt}, can be defined as:

$$d_1(B_j^{\text{alt}}, x_i) = \log[p(x_i|B_j^{\text{alt}},\theta)] - \log[p(x_i|B_i^{\text{max}},\theta)] \tag{4}$$

[1] We have used the ARPABET phonetic alphabet for the transcriptions.

B_i^{max} is the pronunciation for word i with the highest score. If this measure is positive, we have a classification error. One possible loss function is therefore to count any errors introduced by the new pronunciation, implicitly making no model assumptions:

$$l_1(B_j^{\text{alt}}, x_i) = \begin{cases} 0, & d_1(B_j^{\text{alt}}, x_i) \leqslant 0 \\ 1, & d_1(B_j^{\text{alt}}, x_i) > 0 \end{cases} \tag{5}$$

We sum over all acoustic segments x_i belonging to a word i to get the loss $l_2(B_j^{\text{alt}}, \text{word } i)$. To assess the performance of a new pronunciation B_j^{alt}, we combine the loss computed for all other words to find the total error count. Each rule will probably be applicable to several words, and we sum over all these words. A possible evaluation function for a rule k is to count the number of errors it introduces for all words $\mathcal{K}$ where this rule is applicable:

$$\mathcal{L}(\text{rule } k) = \sum_{k' \in \mathcal{K}} \sum_{\text{all words } i} l_2(B_{k'}^{\text{alt}}, \text{word } i) \tag{6}$$

The measure $\mathcal{L}(\text{rule } k)$ should be compared to the number of errors introduced by the corresponding reference pronunciations. A rule that increases the number of errors will add confusability and should be discarded.

One problem with both these confusability reduction schemes is that we only consider one rule at a time. To truly reduce confusability we have to look at the interaction between the rules, i.e. sets of rules. The error counting scheme can be extended to consider sets of rules by combining errors for several rules.

For a measure consistent with minimum classification error training we consider a metric similar to the one used in the training of the acoustic models, e.g. [5]:

$$\begin{aligned}d_2(x_i) = &-\log p(x_i|B_i^{\text{max}},\theta) \\ &+ \log\left[\frac{1}{N-1}\sum_{j,j\neq i} e^{\log p(x_i|B_j^{\text{max}},\theta)\cdot\eta}\right]^{1/\eta}\end{aligned} \tag{7}$$

B_i^{max} is the best pronunciation for word i, and B_j^{max} is the best pronunciation for a competing word j. This will be a smoothed misclassification measure where the most confusable pronunciations are given higher weight. η is a positive constant governing the weighting.

3. DATA-DRIVEN PRONUNCIATION RULE MODELLING

To achieve a purely data-driven approach the recognizer is used to generate the alternative transcriptions [1, 2, 3]. Instead of modelling the pronunciations for each word directly, we have modelled pronunciation rules, because the rule source segments will occur more frequently than words and thereby give more reliable estimates. Besides, for our task the vocabulary of the adaptation set is quite different from the test set.

3.1. Alternative Transcriptions

All systematic differences between the reference and the alternative transcriptions should be considered as possible pronunciation rules. We have therefore chosen to use a phone loop grammar as in [3]. The alternative transcriptions can contain two types of errors. Either they can be too similar to the reference transcription

and hide the differences that actually exists in the data, or they can contain transcription errors. We assume that the transcription errors will not be systematic, and since the rule derivation methods we use rely on revealing systematic differences between the two transcriptions, these errors will be discarded.

The phone loop transcription was performed with 5-best recognition to get more transcriptions. As we have a limited amount of data this is favourable. In order to maintain time information we have performed the phone loop recognition on isolated words. This ensures that we compare transcriptions belonging to the same acoustic data. The drawback is that it restricts our experiments to finding word internal rules.

3.2. Alignment of Reference and Alternative Transcriptions

A reference transcription of the data was obtained using the reference lexicon. The alternative transcription obtained by phone loop recognition was aligned to this reference transcription by dynamic programming. The cost of a substitution was set inversely proportional to a measure based on statistical co-occurrence of phones. This measure is called *association strength* and estimates probabilities for phone-to-phone mappings from the acoustic data, making the alignment totally data-driven. Algorithm details on the association strength can be found in [6]. As we have more phones in the reference than the alternative transcription, the cost in the dynamic programming is set higher for an insertion than for a deletion.

3.3. Rule Extraction

The pronunciation rules were derived from the alignment of the reference and alternative transcriptions; this is an approach similar to e.g. [1]. For our experiments we have small amounts of data, and we have therefore chosen to use only one preceding and one succeeding phone as context for the phone-to-phone rules. In [4] the immediate phone neighbours are shown to be the most important context.

Only rules that were applicable, i.e. where the rule source segment appeared in one or more of the words in the task lexicon, were maintained. We merged the rule-generated pronunciations with the reference lexicon as we have little data [2, 3]. We used no pronunciation probabilities in the lexica. A rule was not counted if it appeared less than 6 times. This threshold was chosen because we wanted to prohibit rules made from just one word uttered by a single speaker, as we used a 5-best phone loop. We derived only word internal rules.

4. EXPERIMENTS

We have applied our method to the Wall Street Journal (WSJ) adaptation and test set for non-native speakers of American English (Spoke3, November 93). This part of the test consists of 10 speakers reading 40 sentences for adaptation and 40–43 sentences for testing.

The vocabulary of the test sentences is the 5k WSJ vocabulary. The reference 5k lexicon was generated with the Bell Labs Text-to-Speech system. For the test set a closed trigram language model for the 5k vocabulary was used. The adaptation set consisted of other words, and we used the lexicon generated for the 64k WSJ vocabulary. A reference recognizer with 12 Mel frequency cepstral coefficients plus log-energy term and their first and second derivatives was trained on 84 native speakers (WSJ0 SI-84). Phonetic decision tree state tying was used to build triphone HMMs. This recognizer gave the baseline result of 29.2% WER. For tests on all 10 speakers the 95% confidence interval for this WER is [28.2 – 30.3]%.

The alternative transcriptions were generated using the speaker independent recognizer trained on native speakers. Speaker adapted acoustic models may help produce a better set of pronunciation rules. On the other hand these rules will be more difficult to merge into a common lexicon, and the mismatch between the models used for recognition (common models) and the adapted ones used for rule derivation may be a problem. Preliminary experiments did not give promising results.

5. RESULTS

5.1. Rule Probability Based Rule Pruning

We found individual rules by computing individual association strengths and alignments for each speaker. The variation between speakers is large, and it seems reasonable to find rules for each speaker separately and later merge them to make a common lexicon. The individual rules were selected using different thresholds for the estimated individual rule probability (IRPR).

A joint lexicon based on rule probability was made by merging the counts for the individual rules. Using IRPR $\geqslant 50\%$ for each speaker, merging the counts, and using a threshold for the estimated joint rule probability (JRPR)$> 50\%$, gave a WER of 28.8% compared to the baseline WER of 29.2%. This lexicon had an average number of pronunciations per word (PPW) of 1.34 using 75 rules, and will be used for later comparison to log likelihood based lexica with the same number of rules. Results using different thresholds for both IRPR and JRPR are shown in Table 1. The best result was a WER of 28.4% obtained with IRPR $> 50\%$ and JRPR $> 40\%$, using 140 rules giving a PPW of 1.50. As we can see, lower thresholds on IRPR give less improvement.

IRPR	**JRPR**$> 30\%$	**JRPR**$> 40\%$	**JRPR**$> 50\%$
$> 40\%$	28.6%	28.9%	29.0%
$\geqslant 50\%$	28.8%	28.6%	28.8%
$> 50\%$	28.4%	28.4%	28.6%

Table 1: Results in WER for merged rules using different thresholds on the individual (IRPR) and joint rule probability (JRPR).

The modified rule probability scheme was tried by using the top 30 rules for each speaker after sorting the rules by rule source segment probability. These $40 \cdot 30 = 1200$ rules were then merged by adding the counts. Using a JRPR$> 50\%$ we got 82 rules, a PPW of 1.41 and WER 28.4%. We achieved equal performance with a smaller lexicon, showing that sorting by rule source segment probability retained the most useful rules.

5.2. Log Likelihood Based Rule Pruning

Using the log likelihood pruning on rules sorted by merged rule counts did not give further improvement, in some cases less improvement over baseline. The rules restricted by the joint rule probability threshold may be too pruned. Instead, we used only the log likelihood improvement measure to find the best combined

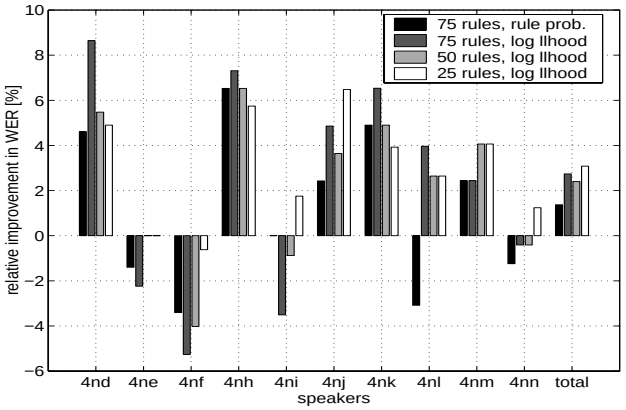

Figure 1: Relative improvement in WER for different number of rules and rule pruning methods, IRPR $\geqslant 50\%$.

rules from the complete set of individual rules. Using different IRPR thresholds we used the top 75, 50, and 25 rules according to the improvement measure $\mathcal{M}$. Equivalent or better results were achieved using fewer rules compared to the rule probability approach. In Figure 1, the lexica based on IRPR $\geqslant 50\%$ are compared to the merged lexicon with 75 rules. We have used relative improvement in WER in this comparison because the variation between the speakers is large.

A test on *native* speakers (WSJ H2 adaptation set) using the 25 rule lexicon gave the same result as baseline for this task: 7.8% WER. For the modified merged lexicon with 82 rules, we got a deterioration: 8.2% WER. As we can see there is a large gap in the performance between the native and the non-native speakers. This is due to the diversity among the speakers, which affect both the acoustic models and the lexicon.

5.3. Confusability Reduction

Including low probability rules, IRPR $\geqslant 20\%$, gave more rules to choose from (1401 compared to 294 for IRPR $> 50\%$) and more confusable rules among the top rules as sorted by $\mathcal{M}$. These lexica gave little improvement over baseline because of the increased confusability.

To reduce the confusability we restricted the rule set to contain at most one rule per rule source segment selected by the log likelihood improvement measure $\mathcal{M}$. Results for confusability reduction are shown in Table 2. The best result we achieved was a WER of 28.3% using only 19 rules (1.11 PPW), and a WER of 28.2% using 39 rules (1.16 PPW). As we can see, low probability rules now perform equally well as higher probability rules, showing that for rule pruning the log likelihood improvement measure is more important than rule probability.

We performed preliminary experiments on the metric in equation (6) to reduce the number of rules even further. We used the 19 rules from the log likelihood pruned lexicon, but the results were counterintuitive, i.e. tests on adaptation and test set were not consistent. Our test and adaptation set contain quite different vocabularies and this may be one of the reasons, because an error counting based measure will not take into account the unseen errors. More study is warranted for a deeper understanding.

Treating the rules individually will not give optimal performance, because the rules will interact. The confusability measure should therefore apply to a set of rules. To this end, we are now experimenting with the smoothed misclassification measure in equation (7).

IRPR	before		after	
	#rules	WER	# rules	WER
$\geqslant 20\%$	25	28.9%	11	28.8%
$\geqslant 20\%$	50	28.8%	22	28.4%
$\geqslant 50\%$	25	28.3%	19	28.3%
$\geqslant 50\%$	50	28.5%	36	28.4%
$> 50\%$	25	28.4%	21	28.4%
$> 50\%$	50	28.3%	39	28.2%

Table 2: Results for top rules sorted by log likelihood improvement measure before and after confusable rules are removed.

6. SUMMARY

In this paper we have compared a new log likelihood based rule pruning measure with more traditional rule probability based measures. The results show that log likelihood works better as a rule pruning measure. We can achieve equivalent or better performance with fewer rules and also avoid deterioration for native speakers using the same lexicon, i.e. less confusability. Lexica with fewer rules are also favourable, as more rules give more pronunciations in the lexicon and slows down recognition. Thus log likelihood based rule pruning gives a better way of combining individual rules to generate a joint lexicon. Adding confusability reduction, we achieved the same result with even fewer rules. We achieved the best result using a 19 rule lexicon which gave a WER of 28.3% compared to the baseline WER of 29.2%.

To achieve better results, we believe a confusability measure for sets of rules is necessary.

7. REFERENCES

[1] N. Cremelie and J.-P. Martens, "In search of better pronunciation models for speech recognition," *Speech Communication*, vol. 29, pp. 115–136, 1999.

[2] E. Fosler-Lussier and G. Williams, "Not just what, but also when: Guided automatic modeling of Broadcast News," in *Proc. DARPA Broadcast News Workshop*, (Herndon (VA), USA), 1999.

[3] J. J. Humphries and P. C. Woodland, "Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition," in *Proc. EUROSPEECH-97*, (Rhodes, Greece), pp. 2367–2370, 1997.

[4] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Communication*, vol. 29, pp. 209–224, 1999.

[5] F. Korkmazskiy and B.-H. Juang, "Discriminative training of the pronunciation networks," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, (Santa Barbara (CA), USA), pp. 137–144, 1997.

[6] I. Amdal, F. Korkmazskiy, and A. C. Surendran, "Data-driven pronunciation modelling for non-native speakers using association strength between phones," in *Proc. ISCA ITRW ASR2000*, (Paris, France), 2000.