Automated Lexical Adaptation and Speaker Clustering based on Pronunciation Habits for Non-Native Speech Recognition

Antoine Raux

Language Technologies Institute
Carnegie Mellon University
antoine@cs.cmu.edu

Abstract

This paper describes a method to improve speech recognition for non-native speech in a spoken dialogue system. Based on very general rules about possible vocalic substitutions, the frequency of occurrence of each substitution in different phonetic contexts is estimated on a small set of recordings. The most frequently observed substitutions are applied to the lexicon of the recognizer. Speakers in the training set are automatically clustered according to their preferred phonetic variants, and a specific lexicon is built for each cluster. Acoustic adaptation is also performed on each cluster. Experiments show that lexical adaptation provides a 16.4% relative WER reduction over acoustic adaptation alone. Lexical clustering can further reduce WER if the system can reliably select the cluster best matching each input utterance.

1. Introduction

Poor speech recognition on accented speech, particularly for non-native speakers, is a major obstacle to the widespread use of speech interfaces outside the laboratory. Even for task-specific spoken dialogue systems, which only have to deal with a limited subset of natural language, recognition accuracy for non-native speech is often too low to make these systems of practical use. In the past decade, several attempts have been made to adapt speech recognition to non-native speakers, usually with some good results (e.g. [1]). Some approaches use linguistic knowledge of the speaker's native language (L1) to predict non-native phonetic and/or acoustic realization patterns of the target language (L2). Others are data-driven and extract such patterns from a corpus of non-native speech from a specific population. In order to avoid the difficulty of collecting large non-native databases, Goronzy[2] proposed a method to derive non-native pronunciations solely from native databases of L1 and L2. All these approaches target a specific non-native population. While this is reasonable for some applications whose users belong to a defined group (e.g. Computer-Assisted Language Learning), it is not applicable to publicly available speech-based systems, where there is no prior knowledge of the speaker's L1.

In order to study ways to improve the accessibility of speech applications to the general public, we built the CMU Let's Go bus information system, which provides schedule information for buses in the Pittsburgh area. Because of the large international student population in Pittsburgh, many users of the public transportation system are non-native speakers of English, coming from all over the world.

2. Automatic Lexicon Adaptation

2.1. Pronunciation Variant Analysis

Although our approach is generalizable to any type of lexicon adaptation, this work focuses on vocalic substitution. Because our goal is to deal with speakers from a wide variety of L1, we do not manually write accent-specific substitution rules. Rather, we manually define possible vocalic substitutions (see Table 1), without hypothesizing which substitutions actually occur in which context.

Table 1: *List of possible vocalic substitutions*

AA → AO	AA → OW
AE → AA	AE → EH
AH → AA	AH → UH
AH → UW	AO → OW
AO → AA	AW → OW
AW → UW	AY → EY
EH → AE	EH → EY
ER → AA	EY → EH
IH → IY	IY → IH
OW → AW	OW → AO
UH → UW	UW → UH

Following previous work on lexical adaptation (e.g. [1] and [3]), we learn the distribution of the substitutions from data. Note that our goal is not to define a set of rules that captures non-native accents in a general way but rather to improve recognition accuracy for one particular system and its user population. Therefore, our adaptation

assumes that a rather small amount of data (3280 utterances in these experiments) matching the task and user distribution of the target system. The occurrence of pronunciation variants is detected with a speech recognizer in forced-alignment using a lexicon expanded according to the possible substitutions. In our case, the original, hand-written, lexicon had 552 words, with an average of 1.3 pronunciations per word. The expanded lexicon, which contains all possible substitutions between confusable vowels, has 4.3 pronunciation variants per word on average. From the result of the forced alignement, we compute the frequency of each pronunciation variant.

2.2. Rule Selection and Application

The pronunciation variant analysis described in the previous section gives us the distribution of variants for words included in the forced-alignement lexicon. If the same lexicon was used for recognition, we could simply select the most likely variants to build an adapted lexicon. Unfortunately, even though they cover the same domain, forced-alignment lexicon and recognition lexicon are very different in our system for several reasons. First, the forced-alignment lexicon contains only words that actually appear in the adaptation data. The recognition lexicon on the other hand contains many more words that users are likely to say in the future. In particular, in our case, all bus stop names covered by the system (559 in total) are included, often with several variants. Also, in order to reduce word confusability, these stop names are included as single entries even when they contain several words (e.g. "Forbes and Grant"). This results in very long lexicon entries that would generate a very large number of pronunciation variants (since they contain many vowels), resulting in an "explosion" of the lexicon.

In order to generalize the variants observed in the adaptation data, we first infer from them a set of transformation rules that are then applied to the recognition lexicon. All our rules follow the format P [S] N $\rightarrow$ T, where S is the source phone, T is the target phone, and P and N are the previous and next phones. We computed the number of times each rule was "selected" by a speaker in the adaptation data based on the pronunciation variant analysis. Finally, after pruning rules that appeared less times than a given cutoff threshold, we apply the rule set to the recognition lexicon.

3. Lexicon-Based Speaker Clustering

One problem with the method described above is that it models the pronunciation habits of speakers with different accents together. This is suboptimal since it has often been observed that adding many possible variants into a single lexicon increases word confusability and can harm WER. We address this issue by automatically clustering speakers from the adaptation set based on their pronunciation habits. Each dialogue session is represented by a vector containing the number of times each pronunciation variant appears. We assume that this vector was generated by a certain pronunciation model, or probabilistic lexicon. Clustering is then done using model-based k-means, as follows:

1. Assign each session to one of two random clusters
2. For each word and each variant, compute the probability that the variant appeared, given that the word appeared $P(V|W,C)$, using the maximum likelihood estimate:

$$P(V|W,C) = \frac{n_{V,W,C}}{n_{W,C}}$$

   where $n_{V,W,C}$ is the number of times variant V of word W was observed in the sessions of cluster C, and $n_{W,C}$ is the number of times word W was observed in cluster C.
3. Reassign each session to the cluster whose lexicon gives the highest likelihood to the pronunciation variants observed in it, where likelihood is computed as follows:

$$L(S|C) = \prod_{(W_i,V_i)\in S} P(V_i|W_i,C)$$

   where S is a session and (W_i, V_i) represents the ith word of the session and its pronunciation variant.
4. If not converged, return to step 2.

Of course, the final result of this algorithm depends on the random starting point. One way to reduce this dependency is to repeat this algorithm a large number of time with different starting points and keep the result that yielded the maximum global likelihood over the whole data (i.e. the product of all the session likelihoods).

4. Experiments

4.1. Data and Baseline Performance

We conducted our experiments with the Sphinx 2 speech recognizer [4], which is also used by the Let's Go! system. The baseline acoustic models are gender-specific trained on a total of 52 hours of calls to the Communicator system. The two sets of acoustic models are used in parallel and, for each utterance, the system selects the hypothesis with the highest recognition score. Here, we report the WER both for the score-based selection method and for an oracle method where the model that gives the lowest error rate is selected for each utterance. The oracle gives a lower bound of the WER for the overall system given the set of models. All results reported here are based on a test set of 449 non-native utterances from past conversations with the Let's Go! system.

Although the channel condition (telephone speech) of the baseline models is roughly that of Let's Go!, they were trained on a different domain and, most importantly, mainly on native speech. As a result, whereas the baseline WER on a control set of 452 utterances from native users' calls to Let's Go! is 17%, on our non-native test set, the performance drops to a WER of 43.1% (resp. 15.1% and 39.8% for oracle selection).

In order to adapt the system to non-native speech, we used 3164 transcribed utterances (169 minutes) from other calls to Let's Go!. This data was separated into male (1676 utterances) and female (1488 utterances) speakers and used for lexical adaptation, clustering, and acoustic adaptation. The latter is performed using a new adaptation method for semi-continuous models[5]. We applied it separately to the male and female data, without any lexical adaptation or clustering. The WER of these adapted models on non-native data is 35.8% (resp. 33.0% for oracle selection), a 17% relative improvement over the baseline.

4.2. Lexicon Adaptation

We ran our lexicon adaptation algorithm on the male and female adaptation sets and created the two corresponding lexicons. Table 2 shows the top 10 substitution rules and their frequency counts in the adaptation data for male and female speakers. For both genders, the rule T UW # → UH [1] dominates, occurring more than 220 times whereas the others occur less than 150 times. Most rules correspond to common vocalic mistakes by non-native speakers of English such as tense/lax substitution (e.g. IH → IY, AO → OW).

Table 2: *The 10 most frequent substitutions on the male and female data*

Male		Female	
Rule	Count	Rule	Count
T [UW] # → UH	267	T [UW] # → UH	226
M [AH] R → AA	134	R [AH] M → AA	148
# [AE] V → AA	96	M [AH] R → AA	146
F [IY] L → IH	95	# [EH] M → AE	114
DH [IY] # → IH	92	# [EH] M → EY	114
N [UW] # → UH	91	M [AH] R → UH	91
# [AE] N → EH	89	# [EY] # → EH	91
# [AE] N → AA	89	B [IH] G → IY	89
# [IH] N → IY	87	L [OW] # → AW	81
B [IH] G → IY	79	F [AO] R → OW	69

Figure 1 shows the WER obtained by using lexicons generated with different thresholds for rule selection, resulting in different average number of variants per word. Results are shown for the case when the expanded lexicons are used only at recognition time as well as when they are used both during adaptation and recognition. It appears that a small amount of rules does improve WER by 16.4% from the adapted baseline. However, including more rules does not seem to provide any additional benefit. Moreover, acoustic adaptation did not prove helpful for these models. It even degraded the performance for small amounts of rules.

[1] We use ARPAbet for phonetic transcriptions throughout this paper

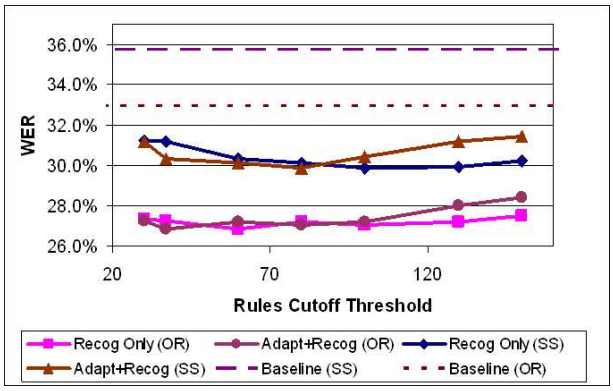

Figure 1: *Word Error Rate when using lexical adaptation for acoustic adaptation (Adapt) and recognition (Recog), with score-based hypothesis selection (SS) and oracle selection (OR)*

4.3. Lexicon-based Speaker Clustering

We tried to improve over the previous results by clustering both male and female adaptation sets using the approach described in section 3. The total number of clusters is now 4, 2 per gender, among which the system picks one for each utterance based on the recognition score or an oracle.

We inspected the clusters generated by this approach with respect to the 6 male and 5 female speakers about whom native language was known. Altogether, these speakers participated in 116 sessions, with an average of 15.1 utterances per session. In the vast majority of cases, the system was able to cluster different sessions from the same speaker together. Indeed only 5 sessions (2 male and 3 female) were not clustered with the majority of their speaker's sessions.

The system clustered dialogue sessions by language proficiency (as judged by the author). For example, one female cluster contained both Indian and Japanese speakers with a native-like pronunciation while the other grouped less proficient speakers.

In this case, we evaluated the two following approaches:

1. using the adapted lexicon and the baseline acoustic models (adapted on each gender's data),

2. using the adapted lexicon for acoustic adaptation and recognition.

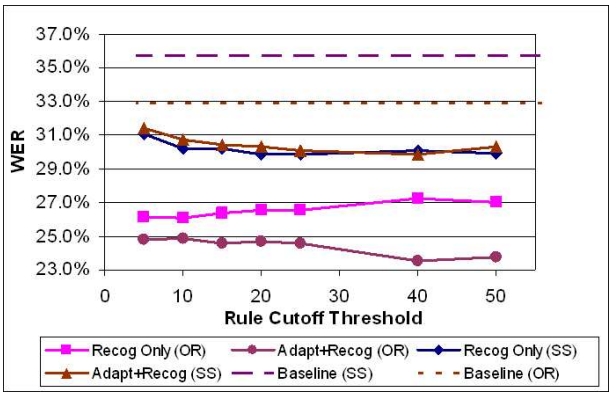

Figure 2: *Word Error Rate when using lexical clustering for acoustic adaptation (Adapt) and recognition (Recog), with score-based hypothesis selection (SS) and oracle selection (OR)*

In the first method, both clusters of each gender share the same acoustic models, whereas this is not the case in the latter. The overall recognition results are shown in figure 2. Again, it appears that while the first rules do bring a lot of improvement, adding more rules does not help recognition. Also, for score-based hypothesis selection, clustering barely improves WER. On the other hand, the oracle performance based on acoustic models adapted on each cluster did bring a 25.5% improvement over the adapted baseline, against 18.5% when no clustering is done.

5. Discussion and Future Works

The main findings of this study concern speaker clustering based solely on pronunciation variants. Inspection of the clusters indicated that the pronunciation variants from a single dialogue session contain enough information to identify the speaking habits of a given speaker. Also, performing acoustic adaptation on the corresponding clusters can significantly improve the accuracy of speech recognition, provided that we can design a reliable way to select among the hypotheses generated by different cluster models. As often observed, raw recognition scores are not a very good indicator of the confidence one can have in a recognition hypothesis. In this case, this is made worse by the fact that we are comparing scores from models adapted on different data. However, many features other than raw score are available in a spoken dialogue system like Let's Go! : how well the hypothesis can be parsed by the language understanding module, how well its semantics fit with the current dialogue state, etc. These features and others have been used in the past for confidence annotation[6]. We plan to follow a similar approach to improve hypothesis selection and take advantage of our clustering method.

Also, the fact that a few specific rules brought a large improvement in WER shows that this simple data-driven approach can be used to test the adequacy of the lexicon of a system given some (native or non-native) user data. Finding such discrepancies manually is hard and time-consuming. Of course, while we restricted ourselves to vowels in this study, more work needs to be done to test our method on other phonemes. In this context, we will study the relationship between the automatically generatd clusters and the set of considered substitutions.

6. Conclusion

We presented a method to analyze the pronunciation variants used by a given user population of a spoken dialogue system in order to improve speech recognition. A very small number of automatically detected vocalic substitution rules helped improve significantly recognition accuracy. We further proposed a new speaker clustering method that uses only pronunciation variant distributions to represent individual speakers. This resulted in further reduction of the WER when acoustic adaptation is performed on the generated clusters, provided that the system can reliably select which hypothesis to use for each utterance.

7. References

[1] L. Mayfield Tomokiyo, "Lexical and acoustic modeling of non-native speech for lvcsr," in *Proc. ICSLP '00*, Beijing, China, 2000.

[2] S. Goronzy, R. Kompe, and S. Rapp, "Generating non-native pronunciation variants for lexicon adaptation," in *Adaptation 2001*, Sophia Antipolis, France, 2001.

[3] I. Amdal, F. Korkmazskiy, and A. C. Suredan, "Data-driven pronunciation modelling for non-native speakers using association strength between phones," in *ASR2000*, Paris, France, 2000.

[4] X. Huang, F. Alleva, H.-W. Hon, K.-F. Hwang, M.-Y. Lee, and R. Rosenfeld, "The SPHINX-II speech recognition system: an overview," *Computer Speech and Language*, vol. 7, no. 2, pp. 137–148, 1992.

[5] A. Raux and R. Singh, "Maximum-likelihood adaptation of semi-continuous hmms by latent variable decomposition of state distributions," in *submitted to ICSLP 2004*, Jeju Island, Korea, 2004.

[6] D. Bohus and A. Rudnicky, "Integrating multiple knowledge sources for utterance-level confidence annotation in the cmu communicator spoken dialog system," Carnegie Mellon University, Tech. Rep. CS-190, 2002.