PREDICTION OF PRONUNCIATION VARIATIONS FOR SPEECH SYNTHESIS: A DATA-DRIVEN APPROACH

Christina L. Bennett and Alan W Black
Language Technologies Institute, Carnegie Mellon University
{cbennett,awb}@cs.cmu.edu

ABSTRACT

The fact that speakers vary pronunciations of the same word within their own speech is well known, but little has been done to automatically categorize and predict a speaker's pronunciation distribution for unit selection speech synthesis. Recent work demonstrated how to automatically identify a speaker's choice between full and reduced pronunciations using acoustic modeling techniques from speech recognition. Here, we extend this approach and show how its results can be used to predict a speaker's choice of pronunciations for synthesis. We apply machine learning techniques to the automatically categorized data to produce a pronunciation variation prediction model given *only* the utterance text – allowing the system to synthesize novel phrases with variations like those the speaker would make. Empirical studies emphasize that we can improve automatic pronunciation labels and successfully utilize the results for prediction of future synthesized examples. The prediction results based on these automatic labels are very similar to those trained from human labeled data – allowing us to reduce manual effort while still achieving comparable results.

1. INTRODUCTION

A goal of data-driven unit selection speech synthesis is to produce a voice that sounds as similar as possible to the donor speech. Thus, much work has been done to model the speaker's acoustics, but the pronunciation habits of the speaker have not previously been a major area of research in the synthesis community. Pronunciation has been treated as a secondary issue, using knowledge-based resources to adapt them to a dialect or perhaps a speaker.

As a result, a large amount of human effort is required to study each dialect or carefully examine what the particular speaker does. Some work has been done to reduce this effort. For example, Fitt and Isard [5] have demonstrated ways to create dialect-independent lexica with encoded dialectal variation. However, these techniques do not account for the variation in pronunciations within the individual's speech, nor do they allow for changes in speaking style. Miller [7] proposed using a neural network to learn certain pronunciation habits from a speaker, such as realization of an underlying reduced vowel. Results were best when a speaker was consistent, but the technique was less able to differentiate within-speaker choices when both variants were frequently used, as well as when more than two variants were possible.

More recently, Bennett and Black [1] introduced a method using acoustic modeling and forced alignment to automatically label words expected to exhibit in-speaker variation. The method worked well for some words but had little impact for others. These methods were used only to identify what pronunciation choices had been made in a particular database but did not show how to use this knowledge to make appropriate variant choices when synthesizing a new utterance.

2. PROBLEM DISCUSSION

The purpose of this work is to automatically learn which of several possible pronunciations an individual would use if he/she were speaking the text. To that end, we require a corpus of examples of the words with accurate pronunciation labels. Thus, we propose an adaptation of the technique introduced by Bennett and Black [1] to improve automatic categorization of a speaker's pronunciation choice in a given speech corpus so that accurate prediction models can be trained. We then utilize a portion of these automatic categorizations to learn what the speaker would say in unseen utterances, allowing us to predict the speaker's pronunciation preference in a new utterance.

As in [1], the words analyzed are "for", "to", "the", and "a", which were chosen because of their pronunciation variability and frequency of occurrence. Initially, each word was assumed to have a single *full form* and *reduced form* pronunciation, but we have adapted this to include an additional reduced form when motivated by the data.

3. FRAMEWORK

In order to compare to the results in [1], we use the same experimental setup for automatic categorization. In addition to the *f2b* voice from the Boston University Radio News Corpus [8] used previously, we have applied our techniques to the *bdl_arctic* voice from the CMU ARCTIC database [6]. Both corpora were created for use in speech synthesis. The *f2b* voice is roughly fifty minutes of female American English in newsreader style and not designed to be phonetically balanced, whereas the *bdl_arctic* voice, is phonetically balanced and contains 1,132 utterances collected from storybook texts. The speaker is an American male.

Obtaining phonetic labels for the acoustic data is a necessary step in any corpus-based synthesis endeavor. The SphinxTrain acoustic modeling toolkit [4] was used to help automate this process, allowing us to build models directly from the database and then use them to perform forced alignment of the text to the spoken data. To determine the feasibility of using automatically obtained pronunciation categorizations to predict future pronunciation choices, we have chosen to build classification and regression trees (CART) [3]. The Wagon tree builder, part of the Edinburgh Speech Tools [9] was used. Experiments were run within the FestVox voice-building environment [2].

4. TECHNIQUES

4.1. Data distribution

In order to measure performance of the method, it was necessary to obtain a human categorization of pronunciations for the four words studied. Overall distributions for the f2b and bdl_arctic databases are shown in Table 1. Also shown is the number of occurrences of the words studied in the two chosen datasets. The style, as well as the method in which sentences were chosen when creating the database, has an effect on these counts.

		full form	reduced form	further reduced	undet	word count
f2b	**"for"**	51.13%	48.12%	---	0.75%	133
	"to"	13.97%	76.42%	---	9.61%	229
	"the"	12.36%	86.53%	---	1.10%	453
	"a"	0.54%	99.46%	---	0%	185
bdl_arctic	**"for"**	27.94%	54.41%	16.18%	1.47%	68
	"to"	18.62%	69.68%	n/a	11.7%	188
	"the"	12.50%	85.98%	n/a	1.33%	527
	"a"	0%	99.15%	n/a	0.85%	234

Table 1. Distribution of pronunciations, as determined by a human evaluator, for each of the words analyzed.

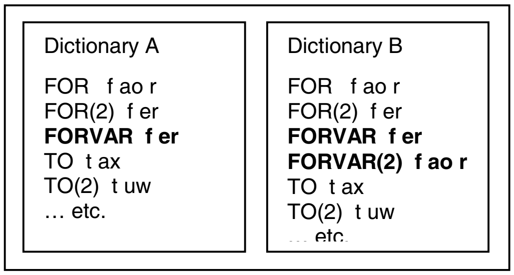
As in Bennett and Black [1], the human evaluator was given strict guidelines for determining which instance of each of the words belonged in the ***full*** or ***reduced form*** categories, and a complete description of each is given there. The human evaluator determined that the *bdl_arctic* speaker actually had three pronunciations for the word "for", thus a third category was added. Since the third pronunciation [F AX] is even more reduced than the predefined reduced form [F ER], we call this category ***further reduced***. Though a similar, but less clear-cut case was found in *f2b* for the word "to", the large number of cases in the undetermined category was attributed to the pronunciation [T UH], which was considered to be in between the other defined pronunciations ([T UW], full form and [T AX], reduced) and thus did not warrant its own category.

4.2. Experimental Setup for Automatic Categorization

Bennett and Black's method using acoustic modeling techniques to automatically categorize pronunciations was used for words with known in-speaker variation, and training was done solely on the database in question [1]. The method concentrated on only two pronunciations per word, a *full form* and a *reduced form*; one of which was the more common, default pronunciation, and the other the variant form. The method focused on iteratively finding instances of the variant form in the given database, where the first iteration consisted of: 1) training initial acoustic models on the corpus itself; 2) using forced-alignment to choose between possible pronunciations; 3) modifying the transcript to mark predicted variant pronunciations. After this, the procedure repeated by retraining acoustic models, this time taking into account the pronunciations chosen by forced alignment; these words were marked as different tokens, having only the variant pronunciation.

The differences in the approach presented here are twofold. Firstly, we forego the first training iteration for obtaining acoustic models in favor of using established models, which include correct labels for the variants being investigated. This prevents us from relying on predictably problematic models trained from the data upfront. We used the open source general speaker independent 6k Sphinx2 models, which were trained on wide band read data from the Wall Street Journal, for the first forced alignment pass. The variant choices from this forced alignment are then added to the transcript, and we are ready to begin a full iteration as outlined above.

In the previous work, whenever a variant was predicted, the transcript was modified for the next iteration by replacing the word with a token such as "forVAR" to indicate that the variant form of the word had been chosen. In the dictionary, that token had *only* the variant pronunciation as an option during forced alignment. Our second modification to this technique was to continue to give a choice of pronunciations for each instance of the word, regardless of what was decided in previous iterations. Thus, instead of simply accumulating instances where a variant form has been marked, we allow forced alignment to "change its mind" about the instances labeled as variants in previous iterations. This difference is shown in Figure 1.


Dictionary A

FOR f ao r
FOR(2) f er
FORVAR f er
TO t ax
TO(2) t uw
... etc.

Dictionary B

FOR f ao r
FOR(2) f er
FORVAR f er
FORVAR(2) f ao r
TO t ax
TO(2) t uw
... etc.

Figure 1. Dictionary A represents the technique of Bennett and Black [1]; Dictionary B represents our modification.

4.3. Experimental Setup for Prediction

To predict future pronunciation choice, the automatically categorized pronunciations must be utilized. A CART model of the speaker's pronunciation habits can be used as postlexical rules at synthesis time. Here, we chose to predict whether an instance of a word in a given context is expected to have a reduced pronunciation. In a real world scenario where human labels are not available, we would of course be forced to perform the training on the automatic categorizations described above. For this reason, trees have also been trained on the automatic categorization results achieved here and in [1].

We first used the human pronunciation judgments in order to establish a feature list and other experimental settings. Doing so also gives us a sense of whether or not the task is feasible for this learning method. Models were trained in three ways, for the purpose of comparison – using all four words together to create a single tree, using pairs of words (a preposition tree ("for"/"to") and a determiner tree "a"/"the"), and using each word separately to create four distinct trees. All trees were trained using 90% of the instances of the corresponding dataset (within which we used 90% to train and 10% to validate). The remaining 10% was held out to test performance of the resulting trees.

The list of features used to train the trees was accumulated over time based on preliminary experiments with the *f2b* database. Linguistic knowledge was used in constructing this list. For example, since people often say [DH IY] before a word beginning with a vowel, a binary phone-level feature was included to indicate whether the next phone was a vowel or a consonant. A total of 32 features were given in training.

5. AUTOMATIC CATEGORIZATION RESULTS

Below we refer to the method in [1] and its modification presented here, as Method 1 and Method 2, respectively.

5.1. Results for *f2b*

Table 2 shows choices made by Method 2 for the *f2b* dataset, after seven iterations. In this case, more iterations were required to reach convergence compared to Method 1, whose results can be found in [1]. At first glance, these numbers appear to be fairly close to those in Table 1, as chosen by the human evaluator; however, since there are both false negatives and false positives, the performance requires closer investigation.

f2b	full form	reduced form
"for"	48.12%	**51.88%**
"to"	**22.71%**	77.29%
"the"	**14.57%**	85.43%
"a"	**1.63%**	98.37%

Table 2. Automatic categorization distributions for the *f2b* database after seven iterations by Method 2. In bold are variant pronunciations (as opposed to default).

Note that an automatic method must always make a choice between full and reduced forms and can never choose the undetermined category, but either choice could be considered acceptable. We note that there seems to be some logical consistency in terms of the choices made by the method even when these choices conflict with those made by the human.

Because of the differences in the two techniques, the types of errors can be quite different. As a result, we cannot compute accuracy in exactly the same way as was done by Bennett and Black [1]. Thus, we will compare the results with different metrics in order to determine which gives a more accurate categorization. We have calculated false positives and false negatives predicted by each of the methods for the *f2b* voice, but due to space limitations, cannot list them here. Although Method 1 never produced false positives for this database, the number of false negatives (*i.e.* variants missed) is quite high. In terms of overall error (false positives plus false negatives), Method 2 had significantly less (30) than Method 1 (101), with most improvement in the problematic words "to" and "for".

Another valid comparison is that of correctness. Table 3 shows the percentage of the total number of instances of each word that were categorized correctly. While Method 1 is respectable, Method 2 clearly outperforms it for three of the four words under investigation. The fourth word, "a", is of course an unusual case since there was only one instance of the full form pronunciation in the entire dataset.

f2b	"for"	"to"	"the"	"a"	Overall
M1	88.72%	86.90%	87.64%	100%	89.88%
M2	96.99%	93.01%	98.23%	98.91%	96.53%

Table 3. Comparison of the two methods in terms of percentage correct, for the *f2b* database.

5.2. Results for *bdl_arctic*

In order to insure that the techniques discussed above are not database-specific or style-specific, we must consider other data. Thus, Methods 1 and 2 have both been applied to the *bdl_arctic* database, which is of a different gender and style from *f2b*.

Despite the fact that there were three feasible categories of pronunciations for the word "for", we first performed the experiments with only the two pre-existing categories. Because of space limitations, we will not include a detailed analysis of the results. The same experimental comparison was performed with the three pronunciations of "for" included. Results for each of the techniques are presented together in Table 4.

bdl_arctic		full form	reduced form	further reduced
M1	**"for"**	36.76%	**48.53%**	**14.71%**
	"to"	**13.3%**	86.7%	n/a
	"the"	**12.69%**	87.31%	n/a
	"a"	**0%**	100%	n/a
M2	**"for"**	32.35%	**44.12%**	**23.53%**
	"to"	**18.62%**	81.38%	n/a
	"the"	**16.1%**	83.9%	n/a
	"a"	**1.28%**	98.72%	n/a

Table 4. Automatic categorization distributions for the *bdl_arctic* database with three pronunciation choices for the word "for". In bold are variant pronunciations (vs. default).

Again, performance cannot be judged solely based on distribution, thus we calculated the numbers of false positives and false negatives for each of the techniques. For *bdl_arctic*, Method 1 performed much better than for *f2b*, and Method 2 also improved. Table 5 shows percentage correct for each. As shown, both methods performed remarkably well despite the added pronunciation possibility for one of the words. In fact, both perform nearly as well or better on this data compared to the previous database. It is worth noting, however, that Method 2 required only three iterations to converge (*i.e.* no further instances labeled with the variant forms), as opposed to 13 full iterations required by Method 1 to reach its convergence point.

bdl_arctic	"for"	"to"	"the"	"a"	Overall
M1	89.71%	95.21%	99.24%	100%	98.03%
M2	94.12%	95.74%	97.53%	99.57%	97.44%

Table 5. Comparison of the two methods in terms of percentage correct, for the *bdl_arctic* database.

6. PREDICTION RESULTS

Producing good automatic labels of pronunciation choice is a step in the right direction toward the goal described in Section 1; however, it is only a first step. We next need to use the labeled data to learn predictions for future occurrences of each word. We have done many experiments using CART trees with various configurations on both the human-labeled and automatically labeled data for both corpora; due to space limitations, only a small selection of results are given below. Note that human-labeled *undetermined* instances were removed before testing.

As described in Section 4.3, these experiments were actually done many times with various portions of the training data used to build the trees (*i.e.* using the words altogether, paired, and individually). For *f2b*, only the following features were selected when building the trees: identities of word and following word; break level after next and previous words; part of speech of word two words earlier; lexical stress of following two syllables; and following phonetic context. Many of the same features were selected for the *bdl_arctic* data, including identity of the word, part of speech of the word two words earlier, and following phonetic context. Additional features used were part of speech of next word and vowel in the next syllable.

Results are shown in Table 6, where, **Human-train**, **M1-train**, and **M2-train** refer to the trees trained on human-labeled data, automatically labeled data from Bennett and Black [1] (M1), and our automatically labeled data (M2), respectively. Likewise, **human-test** and **auto-test** refer to whether human-labeled data or automatically labeled data was used in testing.

f2b		**human - test**		**auto - test**	
single tree (all words)	**Human - train**	95.9%	94/98	--	--
	M1 - train	**81.6%**	**80/98**	94.9%	94/99
	M2 - train	**94.9%**	**93/98**	92.9%	92/99
bdl_arctic		**human - test**		**auto - test**	
separate tree (each word)	**Human - train**	97.9%	95/97	--	--
	M1 - train	**96.9%**	**94/97**	96.0%	95/99
	M2 - train	**97.9%**	**95/97**	96.0%	95/99

Table 6. Prediction results for *f2b*, using single trees trained on all four words together, and prediction results for *bdl_arctic*, using trees trained separately on each of the words.

For the *f2b* database, we report results when training a single tree from all four words together because it had the best results for the *human*-trained and *human*-tested combination. The word pair trees were equally good; however, using all words together gives more data, which is advantageous for smaller databases. In the case of *bdl_arctic*, trees trained on each word separately were most accurate. These trees were trained to predict from three possible pronunciations for the word "for".

As a sanity check, we have established a baseline prediction result as the percent correct if only the most commonly predicted pronunciation is chosen. That is, if the pronunciation form (full or reduced) that was chosen most often for each word by the automatic categorization methods was always assumed, how many examples in the test set would it predict correctly? For *f2b*, the most common label occurred 84.7% of the time in the test set of all four words together. With this baseline, we see that the results from Method 2's automatic labels give an improvement of 10 percentage points, thereby reducing error by a full third. Method 1 on the other hand, does worse than the baseline in this case. For the baseline in *bdl_arctic*, we use the aggregate percent correct in each of the individual words' test sets since training a tree for each word separately was best. Here the proportion of instances with the most commonly chosen pronunciation was 86.6%. We see that our predicted result for Method 2 surpasses the baseline by 11 percentage points, *i.e.* 84% error reduction. Furthermore, Method 2's performance is overall very similar to that of the trees trained on human labels.

7. DISCUSSION

In this paper, we have improved automatic pronunciation categorization methods using acoustic modeling, and utilized these results to predict variation in unseen data. Determining a speaker's pronunciation choice is difficult because of data sparseness and ambiguous pronunciation. In speech synthesis we require carefully recorded speech from single speakers, thus large quantities of data are not available, making it difficult to obtain substantial samples of the same word for study of pronunciation variation. Application of a strict label can be challenging even for a human, thus how to deal with ambiguous cases when also arises. Despite this, automatic pronunciation categorization techniques have performed well for multiple speech synthesis datasets, differing in terms of style and gender. We have also shown that these automatic labels can be used to predict future pronunciation choices and in fact perform similarly to human labeled data. Results also show improvement over a baseline in which only the most common form is used.

Given these positive results, we hope to extend this work to test whether resulting synthesized speech exhibits improvement from the ability to choose between pronunciations. We also wish to apply these techniques to more difficult words, both less common and with poorly understood variation (*e.g.* "sure"). Ultimately, we wish to discover, as well as categorize and predict, variant forms in languages not known by the researchers.

This work was funded in part by NSF grant (0219687) "ITR/CIS Evaluation and Personalization of Synthetic Voices". Opinions expressed in this paper do not necessarily reflect those of NSF.

8. REFERENCES

[1] C.L. Bennett and A.W. Black, "Using Acoustic Models to Choose Pronunciation Variations for Synthetic Voices," In *Eurospeech '03*, pp. 2937-2940, 2003.

[2] A.W. Black and K.A. Lenzo, "Building Voices in the Festival Speech Synthesis System," http://festvox.org/bsv/, 2000.

[3] L. Breiman, J.H. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth & Brooks, Pacific Grove, CA, 1984.

[4] Carnegie Mellon, "SphinxTrain: Building Acoustic Models for CMU Sphinx," http://www.speech.cs.cmu.edu/SphinxTrain/, 2001.

[5] S. Fitt and S. Isard, "Representing the Environments for Phonological Processes in an Accent-Independent Lexicon for Synthesis of English," In *ICSLP98*, pp. 847-850, 1998.

[6] J. Kominek and A.W. Black, "CMU ARCTIC speech databases for speech synthesis research," CMU-LTI-03-177, Carnegie Mellon, http://festvox.org/cmu_arctic, 2003.

[7] C. Miller, "Individuation of Postlexical Phonology for Speech Synthesis," In *Third ESCA/COCOSDA Speech Synthesis Workshop*, pp. 133-136, 1998.

[8] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," ECS-95-001, Boston University, 1995.

[9] P. Taylor, R. Caley, and A.W. Black, "The Edinburgh Speech Tools Library," http://www.cstr.ed.ac.uk/projects/speechtools.html, 1998.