

Using Acoustic Models to Choose Pronunciation Variations for Synthetic Voices

Christina L. Bennett and Alan W. Black

Language Technologies Institute
Carnegie Mellon University
{cbennett, awb}@cs.cmu.edu

Abstract

Within-speaker pronunciation variation is a well-known phenomenon; however, attempting to capture and predict a speaker's choice of pronunciations has been mostly overlooked in the field of speech synthesis. We propose a method to utilize acoustic modeling techniques from speech recognition in order to detect a speaker's choice between full and reduced pronunciations.

1. Introduction

In the field of speech synthesis, much attention has been focused on modeling the acoustic characteristics of an individual speaker, but the pronunciation habits of the particular speaker have been virtually ignored. Pronunciation rules have instead been derived from generalized lexicons modeled after the average, or most common, pronunciation. This is often done on a very large, *language*-level scale (which is often actually a highly generalized dialectal scale, such as "American English" or "British English"), despite the well-known variations in dialect within them.

To create voices in these different dialects, large linguistic studies of the dialect as a whole must be undertaken to determine what characteristics can again be *generalized* across speakers of the dialect. Generally, a new pronunciation dictionary or set of transformation rules must then be created or obtained from an existing source in order to build a voice. Fitt and Isard [1] have examined ways to create dialect-independent lexica while encoding dialectal variation. These techniques again, however, ignore the actual individualized pronunciations of the donor speaker.

There are currently no well-defined methods for automatically learning the idiosyncrasies of pronunciation for an individual. Miller [2] has proposed a method to learn a speaker's behavior regarding certain observed characteristics; however, this technique worked best when a speaker consistently used only one of the variants but did not predict well the speaker's choice between variants. It also struggled when more than two variations were possible.

2. Problem discussion

The focus of this work is to find methods to automatically determine the variations in pronunciation on the level of the individual. That is to say, given a database for synthesis, when does the speaker pronounce word X with one pronunciation versus another? In order to eventually learn which pronunciation to produce during synthesis, we must determine which pronunciation variant was produced by the speaker and in what context. In future work, we may find that the contextual decision basis may be different for each speaker

(or even each of the speaker's styles), but we expect the technique for discovering it will be transferable to multiple speakers.

The method we describe should be applied when variation is known to occur in a database, but a human determination of which instances are examples of which pronunciation has not been made. Specifically, we must know which language we are working with; we must know that a word existing in the language can have multiple pronunciations; and, we must know what those pronunciations are. In order to evaluate the effectiveness of a method, we must of course also know when the user has actually produced each pronunciation.

We have focused on words which are known to have varying pronunciations yet are common (i.e. frequently occurring) enough to be studied without the need to amass extreme quantities of data. The following words are known to have multiple pronunciations for most speakers of American English, yet the distribution of such remains unclear: *the*, *a*, *to*, and *for*. The actual pronunciations that exist for these words may vary by speaker, but in most cases we can expect to minimally find [DH AX] versus [DH IY], [AX] versus [EY], [T AX] versus [T UW], and [F ER] versus [F AO R].

Most people believe the word *the* to have a fairly clear rule for pronunciation variation: [DH AX] is the default, but [DH IY] occurs before vowels. Yet even this seemingly clear rule has exceptions. When adding emphasis or referring to a specific example (*the* car), a speaker may use [DH IY]. On the other hand, when speech is sloppy, or in a more casual style, a user may use [DH AX] despite the following vowel.

Despite what may have been *prescribed*, we are interested in *describing* what actually happens – which may or may not reflect this "rule". Indeed, the purpose of this work is to find a procedure to automatically determine the distribution of pronunciations for a particular speaker.

3. Framework

3.1. Data

In the work described here, we have used the *f2b* voice from the Boston University Radio News Corpus [3]. This corpus was collected specifically for use in speech synthesis and contains roughly forty-nine minutes of speech. For this voice, an American female spoke the utterances in newscaster style. This corpus was not designed to be phonetically balanced, but it is not believed to be notably unbalanced.

All phonetic representations throughout the paper are given in the DARPA phoneset, which is used by Festival, described below.

3.2. SphinxTrain acoustic modeling

In the use of corpus-based synthesis, labeling of data is both necessary and time consuming. Thus various techniques have been proposed to automate the phonetic labeling of data. In speech synthesis, there are specific advantages over acoustic modeling for speech recognition. Here we have a single speaker and the recordings are very high quality. Except where there are errors (which exist but are rare), the prompt list given to the speaker will be very nearly a correct transcription of what was spoken. Also, if a database is suitable for unit selection synthesis it will be phonetically balanced. In this work we use SphinxTrain [4] to build new acoustic models from our database, which are then used in forced alignment of our prompts to the data.

3.3. Synthesis framework

We are working within the FestVox [5] voice-building environment, which offers tools, scripts and documentation for building unit selection voices within the Festival Speech Synthesis System [6].

Within FestVox, the unit selection technique we are using is that described in [7]. This method clusters units of the same labeled type. This requires reliable phonetic labeling. Although the Boston University (BU) Radio corpus is distributed with phonetic labeling, this unit selection technique also depends on a strong correlation between the lexical pronunciations and the data labeling. As our lexicon, CMUDICT [8], does not correspond to BU original labeling, and that we wish to investigate automatic speaker specific labeling, we generate our own labels.

4. Techniques

For this work we have run a number of experiments using the dataset and tools described above. The following is a description of the distribution of data and our experimental techniques.

4.1. Data distribution in f2b

In order to determine when and to what extent the procedure was able to identify the correct pronunciation, we undertook a human evaluation of all the occurrences of the words under investigation. In total, there were 133 occurrences of the word *for*, 229 occurrences of *to*, 453 occurrences of *the*, and 185 occurrences of the word *a* in the f2b corpus, as it was used for the experiments described herein. (Some portions of the original corpus were excluded based on recording conditions.)

Given the nature of the words chosen for this work, each could be categorized as being **full form** or **reduced form**. The category of **undetermined** was added for those cases that were deemed too difficult to categorize. The results of this analysis are shown in Table 1 and the discussion that follows.

For the word *for*, **full form** refers to the [F AO R] pronunciation, whereas **reduced form** refers to a [F ER] pronunciation. We should point out that many of the utterances in the database end with the same phrase (because of the nature of radio journalism). Since this phrase contains the word in question, it is certainly expected that its regularity may have influenced the way in which it was pronounced, particularly giving it a more careful pronunciation.

	full form	reduced form	undetermined
<i>for</i>	51.13%	48.12%	0.75%
<i>to</i>	13.97%	76.42%	9.61%
<i>the</i>	12.36%	86.53%	1.10%
<i>a</i>	0.54%	99.46%	0.00%

Table 1: Distribution of pronunciations, as determined by a knowledgeable human evaluator.

The assessment of the word *to* was more subjective since the word *to* may be commonly reduced in a variety of ways. In this case, the **full form** refers to the [T UW] pronunciation, but [T AX] or just [T] were both considered to be **reduced form**. Several cases were deemed to be **undetermined** for one of two reasons. Some seemed to be neither fully formed, nor completely reduced, instead resulting in a pronunciation somewhat like [T UH]. Others had been affected by, and indeed could not be separated from, the following context.

Full and reduced forms of *the* correspond to [DH IY] and [DH AX], respectively. Despite the fact that the reduced form may be considered to be the primary pronunciation, whereas [DH IY] would be the less common variant, we have chosen to designate them as such based on the feeling that the [DH AX] pronunciation is the more relaxed, less careful pronunciation of the two. This categorization is parallel to the choice made for *to* above.

In general, most full form instances of *to* and *the* preceded a vowel, as could be expected, whereas others were most frequently reduced. However, this was not a hard rule, as exceptions were found in both cases.

As above, we have chosen to designate [EY] as the **full form** and [AX] as the **reduced form** pronunciation of the word *a*. The sampling of occurrences of the word *a* shows a highly consistent preference for the reduced form. These examples, however, only reflect pronunciations of the determiner *a*, as opposed to its pronunciation as a letter, in the case of spelling or abbreviations. If we were to include these instances as well, we would expect an equally regular distribution of only the full form pronunciation, as the reduced form is widely accepted to be an illegal pronunciation in this usage.

4.2. Experimental setup

The basic procedure for our experiments was as follows:

- Setup the database as described in the FestVox manual;
- Train using SphinxTrain;
- Perform forced alignment where a choice in pronunciations is given;
- Evaluate the predictions and add them to the text for the next iteration;
- Repeat.

In more detail, we follow the procedure outlined in section 13.2 of the FestVox documentation [5], using SphinxTrain to perform automatic labeling by way of forced alignment. Before the forced alignment step, we add the predicted pronunciation variations to the automatically generated dictionary. The procedure up to this point is depicted in Figure 1. This allows SphinxTrain to choose between the

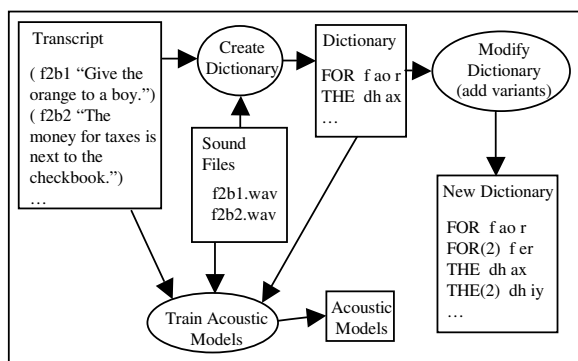


Figure 1: Diagram depicting the various components involved in the acoustic training process and the procedure leading up to the forced alignment (automatic labeling) step.

variants when assigning labels to the utterances. Pronunciation predictions are made in the form of label files (phonetic representations), shown in Figure 2. Once these choices have been made, we can take these predictions and "hard code" them into the transcript text by, for instance, replacing *for* with *for2* in the transcript wherever [F ER] was predicted instead of [F AO R]. Thus a new transcript is created based on these predictions as in Figure 2. By repeating the procedure (back to Figure 1), we allow more choices to be made each time, while getting more accurate models with each iteration.

In the experiment described above, the default pronunciation (which was automatically added to the lexicon by SphinxTrain, i.e. the most common pronunciation) was used for training. Note that this is sometimes the full form pronunciation (as is the case for the word *for*), but other times it is the reduced form that is the more common, or most frequently expected, pronunciation (as is the case for *to*, *the*, and *a*).

In another experiment, we sought to alleviate the potential "overloading" of the /AX/ phone (in both *to* and *the*) caused by the fact that the default pronunciation contained this phone (i.e. [T AX] and [DH AX]). For this experiment, we instead used the full forms of these words in training. Results of both experiments are described in the next section.

5. Results

As can be seen in Table 1, the actual distribution of the pronunciations of the words under investigation varies greatly depending on the word. *For* has roughly equal distribution of its common full form and variant reduced form; however, the word *a* had only one occurrence of its variant (in this case, the full form) in the entire database. *To* and *the* had similar distributions of their variant full forms. Since the procedure we've discussed is automatic, there is no option to choose undetermined. For this reason, we have chosen to consider these cases as pronounced in the default manner, for the purpose of comparison to the automatic labels.

Table 2 shows the choices made by our method after five iterations. At each iteration, the method identified more instances of the secondary, less common, pronunciation for one or more of the words investigated. A sixth iteration was performed, but we found it had converged to a stable point (i.e. no more secondary variant pronunciations were chosen).

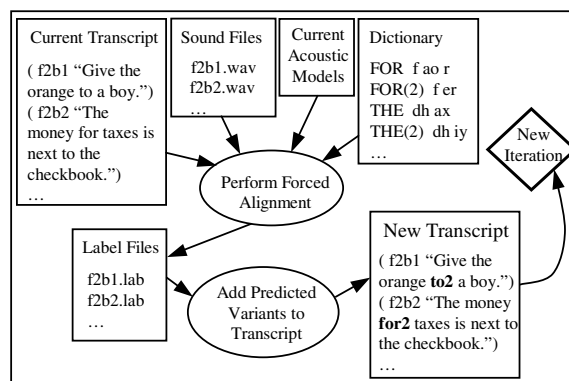


Figure 2: Diagram depicting the various components involved in the forced alignment (automatic labeling) step.

Thus, there was no motivation to run further iterations. We refer specifically to the secondary variants because these are the points of interest. Remember that normally only the most common pronunciation is available, and likewise, it is this pronunciation that is used during training. Therefore, it is the choice of the secondary pronunciation that is of interest to us.

	<i>full form</i>	<i>reduced form</i>
<i>for</i>	63.16%	36.84%
<i>to</i>	0.87%	99.13%
<i>the</i>	0.00%	100.00%
<i>a</i>	0.54%	99.46%

Table 2: Distribution of pronunciations, as chosen automatically after five iterations. Numbers in **bold** are the secondary variants.

To put these results in perspective, we have calculated the accuracy of our method for comparison with the expected baseline accuracy. Normally when automatically assigning pronunciations there would only be one pronunciation for each word; therefore, the default accuracy represents the percentage of times that the most common pronunciation actually occurred in the database. This default accuracy, the accuracy of our method, and the resultant percentage of relative error reduction are all shown in Table 3 below.

	Default Predicted Accuracy	Accuracy of Method	Relative Error Reduction
<i>for</i>	51.13%	87.97%	75.38%
<i>to</i>	76.42%	77.29%	3.70%
<i>the</i>	86.53%	86.53%	0.00%
<i>a</i>	99.46%	100.00%	100.00%

Table 3: Comparative accuracy of pronunciation prediction for the proposed method versus the baseline.

It is worth noting that the method never chooses a secondary variant incorrectly; that is to say, there were zero false positives. This can be seen by the fact that the relative error reduction is never negative. Thus we can clearly state

that the method can only provide improvement in identifying pronunciation choice since it will never falsely predict a secondary variant. That said, it does not provide improvement for all of the words; the reasons for this discrepancy are discussed in Section 6.

5.1. Secondary experiments

In the secondary experiment mentioned in Section 4.2, we replaced the more common pronunciations for the words *to* and *the* with their respective variants during training. By doing this, we avoided the overloading problem for the /AX/ phone; however, since the distribution shown in Table 1 shows that the majority of examples did contain the /AX/ phone, we of course must expect confusability to now exist in the models for the /UW/ and /IY/ phones.

After four iterations of this process, although it was more likely to make a choice between pronunciations for the words *to* and *the*, we determined that this ordering of pronunciations still clearly caused an over-prediction of the *trained* phone, just as it had done previously. This leads us to conclude that there is more to this problem. The most likely explanation is that the following phone causes confusability, since the secondary pronunciation for each almost always immediately precedes another vowel.

Another experiment was performed wherein one known example of the variant pronunciation for each of the words *to* and *the* was "hard-coded" in the transcript for purposes of having an example during training. We found however that this did not significantly impact the results.

6. Discussion

As we can see from Tables 2 and 3 above, the technique works well for the words *for* and *a* in this dataset. However, the alternate pronunciation for the word *to* was not selected nearly as much as by the human evaluator, and it was never chosen for *the*. As we alluded in Section 4.2, we believe there may be a problem here in overloading the /AX/ phone during training. Since we know that there are several instances (for both words) in which a full vowel exists, training with the reduced form, despite its prevalence, will lead to very poor models of /AX/. This confusability could be further enforced by full pronunciations of other words in the database, in which the lexicon expected reduction.

Furthermore, we note that nearly all of the full form examples of *the* and *to* ([DH IY] and [T UW]) occur before a vowel. Since automatic labeling of vowel type is known to be problematic, we hypothesize that this may contribute to the difficulty in these cases. On the other hand, since the goal is to find where the variations occur, we can be satisfied to rely on this rule for mostly accurate predictions of these words.

The ultimate evaluation criteria for this technique is how it effects a unit selection speech synthesizer built from the predicted labeled data. Evaluating speech synthesizers is hard, but not impossible, and we have yet to set up the experiments to test our labeling, though informally we have noted selection of mislabeled vowels is a distinct problem.

We also note that detecting these variants automatically is only one half of the ultimate problem. Once the variations are detected we must also build predictors, to be used at synthesis time, to choose between variants.

In the future we intend to further investigate the differences in performance described, by including other

contexts and/or other words that have different circumstances, as well as continuing the approach described here with modifications. One such modification would be to use established models, which include correct labels for the variants being investigated, rather than training predictably problematic models from the data itself.

We also plan to extend this work to other datasets, including corpora of other languages in which the variation is unknown. Additionally, there are many extensions of the problem, which we intend to investigate. In particular, we will examine the effects of different speaking styles on pronunciation choice, as well as lexically more difficult variations, such as the English word *sure*, which can be pronounced in a number of ways, and for which there is no clear notion of easily predictable distribution.

7. Conclusions

In conclusion, we have proposed a method to automatically determine which pronunciation to assign to words that commonly vary in known ways. The method was very successful for those words without a contextually predictable variability, i.e. those words that do not follow a specific contextual rule for variation. For the words with contextually predictable pronunciations, further investigation is required as they were difficult to predict acoustically. These words are known to have several other factors that may contribute to confusability.

8. Acknowledgements

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

9. References

- [1] Fitt, S. and Isard, S., "Representing the Environments for Phonological Processes in an Accent-Independent Lexicon for Synthesis of English", in *Proceedings of ICSLP98*, pp. 847-850, Sydney, Australia, 1998.
- [2] Miller, C., "Individuation of Postlexical Phonology for Speech Synthesis", in *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 133-136, Jenolan Caves, Australia, 1998.
- [3] Ostendorf, M., Price, P., and Shattuck-Hufnagel, S., "The Boston University Radio News Corpus", ECS-95-001, Electrical, Computer and Systems Engineering Department, Boston University, 1995.
- [4] Carnegie Mellon University, "SphinxTrain: Building Acoustic Models for CMU Sphinx", <http://www.speech.cs.cmu.edu/SphinxTrain/>, 2001.
- [5] Black, A. and Lenzo, K., "Building Voices in the Festival Speech Synthesis System", <http://festvox.org/bsv/>, 2000.
- [6] Black, A., Taylor, P., and Caley, R., "The Festival Speech Synthesis System", <http://festvox.org/festival/>, 1998.
- [7] Black, A. and Taylor, P., "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis", in *Proceedings of Eurospeech97*, V. 2, pp. 601-604, Rhodes, Greece, 1997.
- [8] Carnegie Mellon University, "The Carnegie Mellon University Pronouncing Dictionary", <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>, 1998.