

Pronunciation Lexicon Adaptation for TTS Voice Building

Yeon-Jun Kim, Ann Syrdal, Alistair Conkie

AT&T Labs-Research
180 Park Ave. Florham Park, NJ 07932, USA
{yjkim,syrdal,adc}@research.att.com

Abstract

This paper describes reducing phone label errors in TTS voice building by means of modeling of speaker pronunciation variants.

Each speaker has his or her own unique pronunciations (and context-dependent variations), so that no one standard lexicon is able to cover all of the speaker's variations. Creating *speaker-dependent* pronunciation lexicons for automatic speech labeling of our TTS voice databases helped to eliminate many pronunciation errors that resulted from mismatches between *lexical* pronunciations and how the speaker (voice talent) *actually* pronounced a word. We also found that it contributed other synthesis quality improvement as well.

A perceptual test showed that our work contributed to MOS improvement for American English male and female voices.

1. Introduction

Currently, many commercial TTS systems employ unit selection synthesis techniques and deliver highly intelligible synthetic speech [1]. A unit selection TTS system works well most of the time and can synthesize very natural speech. A drawback of unit selection TTS is, however, that errors are often conspicuous when TTS synthesizes phonetically incorrect sounds. Often these errors are caused by phone label mistakes in the voice inventories though other causes are possible, e.g. a TTS front-end will transcribe some proper names wrongly.

Voice inventories of unit selection TTS are generally created by an automatic labeling method for which the accuracy has a direct influence on the segmental quality of the TTS system. Automatic labeling techniques save a large amount of human effort and time, and their error rates are generally low enough to ensure reasonable accuracy of phone labels. To improve the accuracy of phone labels in voice inventories, we have been working on fine-tuning acoustic models, reducing timing errors of phone labels, as well as the segmental quality improvement of synthesis speech [2] [3].

Through careful analysis of synthesized speech, we have also found that mismatches between speaker specific productions and standard transcriptions are another source of labeling errors in voice inventories. Automatic phone labeling for

TTS voice building is commonly done by *forced alignment*, where each phone label is assigned by a transcription of the corresponding word even if there is no transcription in the lexicon that matches precisely with a speaker's pronunciation.

Even though we might know the word sequence to be uttered, the phone sequence of utterance can vary depending a speaker's performance. Speaker performance is of course influenced by various factors such as discourse situation, speaking style or even level of concentration. Therefore, a lexicon used in forced alignment should have not only general word transcriptions but also speaker specific pronunciation variants to help us label units accurately.

This paper describes our efforts to reduce labeling errors in two *American English* voice inventories by means of adopting speaker pronunciation variants in the labeling lexicon.

In section 2, the requirements of a pronunciation lexicon used in labeling TTS voice inventories are described. In order to detect speaker pronunciation variants in large-scale speech corpora, the results suggested by a phone recognizer are analyzed in section 3. Section 4 deals with acoustic model adaptation incorporating speaker pronunciation variants into the lexicon. This paper concludes with a perceptual evaluation of synthetic quality based on the voice databases created using the methods described.

2. Lexicon for Automatic Labeling

A pronunciation lexicon is a core component in automatic phone labeling which maps the orthographic representation of a word to its pronunciation. The requirements of a pronunciation lexicon for automatic labeling are as follows:

- Transcriptions in the lexicon for automatic labeling needs to be as close as possible to the one from the TTS front-end so that TTS (unit-selection) can choose the best units for synthesizing a word.
- Transcriptions should also accurately describe speaker pronunciations for higher agreement between front-end and speech database.

Based on the requirements described above, our initial pronunciation lexicon consisted of transcriptions generated

by the TTS front-end together with an automatic speech recognition (ASR) lexicon having pronunciation variants. The ASR part of the lexicon contains numerous pronunciation variations generated with decision tree based phoneme-to-phone mappings [4].

However, it may not be advisable to label a TTS voice database and train the speaker-dependent acoustic models with a large multiple speaker transcription set such as the one from the TIMIT corpus [5] since it makes it difficult to focus on single speaker-specific pronunciation variants. In the previous voice building, it turned out that overly generated variants reduced the labeling accuracy.

To alleviate over-generation, we built a new pronunciation lexicon for TTS voice labeling consisting of 1) transcriptions from TTS letter-to-sound rules, 2) the PRONLEX [6] to cover well-known pronunciation variants, and 3) the target speaker’s own pronunciation variants. To obtain the target speaker specific pronunciation variants in recorded speech, a phone recognition technique was used in our work. The speaker specific variants suggested by a phone recognizer are classified and explained with phonological phenomena in section 3.

3. Analysis of Pronunciation Variants

For a unit selection TTS system having several hundred thousand labeled units in each speaker’s voice inventory, it is impractical to check transcription variants without ASR techniques. To detect speaker-specific pronunciation variants, we use a hidden Markov model (HMM)-based phone recognizer as described in [7].

Even though we train HMM models with speaker dependent data, the accuracy of the phone recognizer is limited. The outputs of the phone recognizer are not reliable enough to use without manual checking. Therefore, we identified frequent mistakes of the phone recognizer and applied them as a filter to reduce the number of transcriptions to check.

The transcriptions that our phone recognizer routinely mis-transcribes, but the speaker does not actually produce, are mostly caused by misinterpretations of co-articulation phenomena. The following examples show the frequent mis-transcriptions categorized into *insertion*, *deletion*, and *substitution* errors.

Insertion 1) *insertion of an intrusive plosive* between a nasal and an fricative (especially /s/¹ or /th/), e.g., “chance” as /ch aa n t s/, “means” as /m iy n d z/, “amongst” as /ax m aa ng k s t/,

2) *doubled intervocalic consonant* when the following syllable has primary stress, e.g., “initial” as /ih n n ih sh ax l/, “immense” as /ih m m eh n s/, “amount” as /ax m m aw n t/

Deletion loss of an *unreleased final plosive* before an initial stop (plosive or nasal) or a dental fricative, e.g., “as-

signed task” as “assign task”, “forced to go” as “force to go”, “returned there” as “return there”, “send mail” as “sen mail”

Substitution 1) plosive and fricative separated by a *morpheme boundary* interpreted as an affricate, e.g., “friendship” as /f r eh n ch ih p/, “roadshow” as /r ow ch ow/

2) an aspirated voiceless plosive following /s/ in a syllable onset is *interpreted as voiced*, e.g., “scales” as /s g ey l s/, “space” as /s b ey s/, “stood” as /s d uh d/

3) a strongly aspirated /t/ is transcribed as /th/, e.g., “tan” as /th ae n/, “target” as /th aa r g ih t/, “too” as /th uw/

4) stressed “dr” misinterpreted as “tr”, due to retroflex /r/: e.g., “draw” as /t r ao/, “drink” as /t r ih ng k/, “drive” as /t r ay v/

Many of these errors were automatically identified by a dynamic programming (DP) matching between forced alignment result and phone recognizer transcription, filtering these patterns leads to a 30-40% reduction of the manual labeling correction. These kinds of pronunciation variants suggested by the phone recognizer are disregarded in the continuing recognition process.

3.1. Identifying valid pronunciation variants

Some of the variants suggested by a phone recognizer can be considered unpredictable in the sense that they are determined by the speaker’s performance. Speaker performance is of course influenced by such various factors as discourse situation, speaking style or even level of concentration. Therefore, the analysis of variants produced by the phone recognizer is more experimental rather than based on theory-oriented approaches such as the explicit distinction of lexical and post-lexical variation.

The examples below are cases where the phone recognizer has correctly detected a pronunciation variant produced by the speaker that is not covered by the pronunciation lexicon.

- *vowel reduction* generally creates a lot of variation, e.g. “political” (/p aa 0 l ih 1 t ih 0 k ax l 0/, /p ax 0 l ih 1 t ih 0 k ax l 0/, /p ax 0 l ih 1 t ax 0 k ax l 0/), “prudential” (/p r uh 0 d eh n 1 sh ax l 0/, /p r ax 0 d eh n 1 sh ax l 0/), especially in the case of alternations of schwa /ax/ and unstressed /ih/, e.g., “biggest” as /b ih 1 g ih s t 0/ or /b ih 1 g ax s t 0/, “deliver” as /d ax 0 l ih 1 v er 0/ or /d ih 0 l ih 1 v er 0/
- *schwa-elisions before /r/*, e.g., “century” as /s eh n 1 ch r iy 0/, “commercially” as /k ax 0 m er sh 1 l iy 0/, “dangerous” as /d ey n jh 1 r ax s 0/, “delivery” as /d ax 0 l ih v 1 r iy 0/, “summary” as /s ah m 1 r iy 0/, “conference” as /k aa n 1 f r ih n s 0/

¹In this paper, we use the DARPABET symbols for specifying phones.

- *cluster simplifications*, e.g., “impacts” as /ih m 1 p æ k s 2/, “government” as /g ah 1 v er 0 m ih n t 0/, “products” as /p r aa 1 d ax k s 0/, especially in the case of two alveolar sounds at the end of a word, “analysts” as /æ 1 n ax 0 l ix s 0/, “journalists” as /jh er 1 n ax 0 l ix s 0/, “around” as /ax 0 r aw n 1/, “first” as /f er s 1/

3.2. Identifying speaker’s mistakes

These kinds of errors are of course not predictable, but it can be stated that they are more likely in unfamiliar words, like proper names, foreign words and technical vocabulary in general. In such cases the speaker can even produce more than one variation due to his/her uncertainty about the correct pronunciation.

- “Alvarez” as /aa l 1 v ax 0 r eh z 0/, /æ l 0 v æ 1 r eh z 0/, /ao l 1 v ax 0 r eh z 0/, “Bantu” as /b aa n 1 t ih 0/, /b æ n 1 t uh 0/, “Gendarme” as /zh aa n 1 d aa r m 2/, /zh eh n 2 d aa r m 1/, “Uppsala” as /uh p 1 s ax 0 l ax 0/, /ah p 2 s ey 1 l ax 0/

There is also the possibility of careless reading, a type of error that is difficult, if not impossible, to predict, e.g.

- “Fibreboard” as “Fireboard”, “veterinarian” as “vegetarian”

4. Acoustic Model Adaptation with Speaker-Specific Variants

Speaker-specific pronunciation variants aren’t only added to the lexicon for labeling, but also help to tune speaker-dependent acoustic models. Though the number of wrong labels in initial model training is small with respect to the total number of correct phone labels, mislabels in training contaminate acoustic models. We have seen improvement in phone labeling accuracy after acoustic training with the speaker-specific pronunciation lexicon. The procedure for adapting an acoustic model is described below.

As mentioned in section 3, we use an HMM-based phone recognizer to detect speaker-specific pronunciation variants. The phone recognition procedure consists of two phases: 1) Acoustic (HMM) phone modeling and 2) Phone recognition.

Initially, only existing speaker-independent (SI) HMMs and a basic pronunciation lexicon derived from a letter-to-sound module are used to produce seed phone labels by means of Viterbi alignment for speaker-dependent (SD) HMM training. The resulting speaker-dependent HMMs are trained to provide the segmentation for building an inventory of synthesis units. During the initial training a false mapping between a word and the phonetic units may contaminate the HMMs. But, the SD HMMs are eventually fine-tuned to produce optimal results by using the phone labels from the previous iteration as the input for HMM initialization and re-estimation [3].

In a second phase, phone recognition is performed with the trained SD HMMs and a phone bi-gram language model (LM). A phone recognizer with a phone bi-gram LM as opposed to *forced* alignment with a given lexicon can increase the degrees of freedom in the pronunciation transcription. The pronunciation variants which pass through filtering rules and are accepted by human listening serve as feedback to the pronunciation lexicon used in the next HMM training phase illustrated in Figure 1.

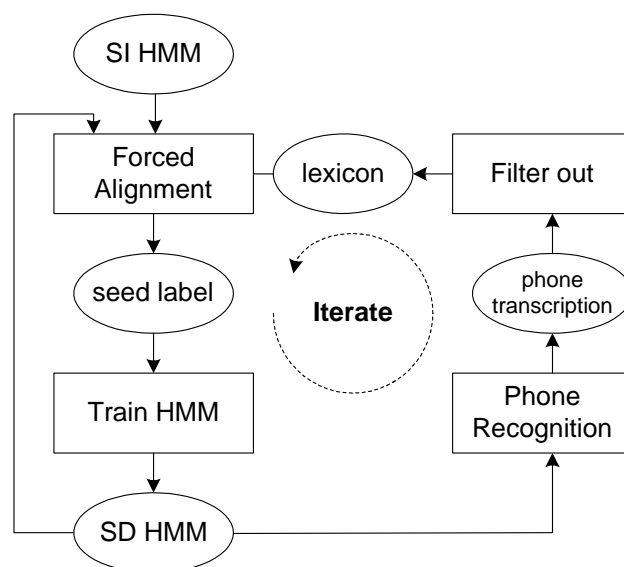


Figure 1: Iterative acoustic model adaptation

The process is repeated and the pronunciation lexicon is improved, yielding a more accurate acoustic model. Iterations are continued until there is no more improvement in the synthetic speech or no more acceptable variations are produced.

In summary, our objective is best achieved by an iterative process involving three main elements described above: 1) acoustic model training (using HMMs) with a high-quality dictionary and letter-to-sound rules that cover all phonologically predictable variants of a word, 2) phone recognition with the speaker-dependent HMMs, and then 3) analysis of the recognized pronunciation variants not covered by the given pronunciation lexicon.

5. System Evaluation

5.1. Resulting speaker-dependent lexicon

As a result of the procedure described above, the numbers of the transcriptions from different sources making up our speaker-dependent lexicons are shown in Table 1.

In the case of speaker A, 6,752 entries produced by our TTS letter-to-sound include both canonical transcriptions and context-dependent variants. We found 197 speaker-specific variants not shown in either, our TTS letter-to-sound or the PRONLEX.

Table 1: Number of transcriptions covered by TTS lts, PRONLEX, and manual transcription (The number in parenthesis means the number of all possible (well-known) variants for words in the lexicon. We add only PRONLEX variants chosen by the alignment procedure into the lexicon.)

Source of transcription	Number of transcription labeled	
	Speaker A	Speaker B
0) unique words	6432	6986
1) TTS letter-to-sound	6752	7354
2) PRONLEX	169 (1365)	154 (1388)
3) Phone recognizer & Human listening	197	246
Total	7118	7754

5.2. Perceptual evaluation

A perceptual experiment was performed to confirm that our work contributes a measurable improvement in synthesis quality. The experiment elicited listener ratings for three versions each of male and female American English synthesized speech. The three versions are organized as follows:

- 1) the reference front-end + the reference voice database
- 2) the reference front-end + the modified voice database
- 3) the modified front-end + the modified voice database

where the modified front-end has an improved post-lexical process which covers vowel reduction, schwa-elision, and cluster simplification described in section 3, and the modified voice database was re-labeled with the speaker-specific lexicon and the acoustic models tuned by the iterative training process described in section 4.

Twelve test sentences were synthesized per voice(2) and TTS(3) version, resulting in a total of 72 trials. The test sentences consisted of 17- to 32- word sentences selected from recent news articles, Aesop's fables in the IPA manual, and interactive service announcements.

Eighteen native and non-native speakers of English participated in the test. All were fluent in English, and most were unfamiliar with the TTS system and the voices tested. A total of 1296 judgments were collected from the eighteen subjects. The interactive and self-paced listening test was web-based, and generally lasted from 15 to 20 minutes. Subjects were instructed to rate the speech quality of an utterance on a 5-point rating scale by clicking on an associated radio button.

5.3. Perceptual test results

The results presented reflect the entire group of eighteen listeners. Listeners' ratings were analyzed with a repeated measures ANOVA, in which Voice (2), TTS Version (3), and Sentences (12) were repeated factors. Each of the three main effects, Voice ($F(1,17) = 0.865$, $p < 0.0001$), TTS Version ($F(2,16) = 19.700$, $p < 0.0001$), and Sentences ($F(11,7) = 7.632$, $p < 0.001$), was statistically significant.

The overall mean rating of each version indicated the TTS system differed significantly: version 1): 3.329, version

2): 3.204, and version 3): 3.576. The Sentences main effect (Sig. = 0.006 (< 0.05)) indicated that rated quality of the 12 test sentences differed insignificantly among each other.

There was a significant Voice by TTS Version interaction ($F(2,16) = 15.814$, $p < 0.007$), reflecting the fact that the male voice improved relatively more than the female voice.

6. Summary and Conclusions

This paper presents a procedure for recognizing speaker specific pronunciation variations in order to build more accurate voice databases for our unit selection TTS system. The main principle described is correction of pronunciation variants and an iterative training process of acoustic models considering variants. The approach not only provides higher-quality synthesis, but also constitutes a method that does not demand extensive amounts of time and effort, thus facilitating the relatively swift production of new voices.

Special emphasis was put on phonological interpretations for particularly frequent mismatches between recognition dictionary and actual speech. This may allow more time to search for less predictable deviations.

In our perceptual test experiment, labeling the speaker's actual pronunciations contributed to synthetic quality improvement.

7. References

- [1] J. Schroeter, A. Conkie, A. Syrdal, M. Beutnagel, M. Jilka, V. Strom, Y. Kim, H. Kang, and D. Kapilow, "A Perspective on the Next Challenges for TTS Research," in *IEEE 2002 Workshop on Speech Synthesis*, 2002.
- [2] Matthew J. Makashay, Colin W. Wightman, Ann K. Syrdal, and Alistair Conkie, "Perceptual Evaluation of Automatic Segmentation in Text-to-Speech Synthesis," in *Proc. ICSLP 2000, Beijing*, 2000, pp. 431–434.
- [3] Yeon-Jun Kim and Alistair Conkie, "Automatic Segmentation combining an HMM-based Approach and Spectral Boundary Correction," in *Proceedings of ICSLP. 2002, Denver*.
- [4] M.D. Riley and A. Ljole, *Automatic generation of detailed pronunciation lexicons*, chapter 12, Kluwer Academic Publishers, 1995.
- [5] W. Fisher, V. Zue, D. Bernstein, and D. Pallet, "An Acoustic-Phonetic Database," *J. Acoust. Soc. Am.*, , no. 81, 1987.
- [6] Linguistic Data Consortium, *COMLEX English Pronouncing Lexicon*, Trustees of the University of Pennsylvania, version 0.3, 1997.
- [7] M. Ravishankar and M. Eskenazi, "Automatic Generation of Context-dependent Pronunciation," in *Proceedings ESCA Eurospeech'97*, 1997.