Phonetically Enriched Labeling in Unit Selection TTS Synthesis

Yeon-Jun Kim, Ann K. Syrdal, Alistair D. Conkie, Mark C. Beutnagel

AT&T Labs - Research, Florham Park, NJ, USA

{yjkim,syrdal,adc,mcb}@research.att.com

Abstract

Unit selection techniques have improved the quality of text-to-speech (TTS) synthesis. However, mistakes which had been less noticeable previously in poorer quality synthetic speech become very noticeable in more natural-sounding synthetic speech. Many problems appear to be caused by mismatches between phones requested by the TTS front-end and phones selected from the labeled speech inventory. Given the input text and the added information predicted by the TTS front-end, finding the optimal units from a speech inventory database still remains a challenge in unit selection TTS synthesis.

Consonants in American English affect intelligibility of speech synthesis and they are realized differently depending on their position in the syllable. Pre-vocalic plosives must have a release burst before the vowel begins while post-vocalic consonants may or may not be released. When a post-vocalic consonant is chosen to synthesize a pre-vocalic consonant, it may cause problems such as missing consonants, consonant confusion or word-boundary confusion.

In this paper, a new phone labeling method which differentiates pre-vocalic and post-vocalic consonants is proposed. The proposed phone labeling method leads unit selection to choose contextually accurate phone units and minimizes unit selection errors caused by lack of specification in TTS front-end transcriptions and phone labels in the speech inventory. In a listening test the TTS voices labeled with the pre-vocalic / post-vocalic distinction were rated significantly higher (+0.33) compared to reference voices that did not use this distinction.

Index Terms: speech synthesis, unit selection, phonetic variations.

1. Introduction

Unit selection based synthesis has brought huge improvement in text-to-speech (TTS) synthesis quality and is widely used in many applications [1]. To generate the desired utterance, previous synthesizers generally parameterized and regenerated speech with signal modification that reduces the quality of synthesized speech. On the other hand, unit selection based synthesizers choose suitable fragments from a database of speech recorded from a speaker and join them together with minimal signal modifications. Unit selection based synthesizers using minimal modification of the speech signal produce highly intelligible and natural sounding utterances instead of buzzy or robotic sounding speech.

Minimal modification in unit selection based synthesis does not only bring high synthesis quality, but also causes some problems. Some of the problems with unit selection synthesis weren't problems in the earlier TTS systems because they used signal modification. So, for example, plosive closure and burst durations were modified to suit the context. In addition, listeners who experience highly quality synthesis speech by the unit selection based systems are not forgiving. They perceive even minor mistakes and rate synthesis quality lower because of that.

Often problems are caused by the discrepancy between phones asked for by a TTS front-end and phones selected from a labeled voice database [2]. We usually label speech databases with phonemic symbols rather than phonetic ones. However, the same phoneme can be realized in different forms (*allophones*) depending on certain phone contexts. The phoneme /t/ in American English, for example, generates several allophones [3].

There two possible approaches to alleviate this problem: (1) specify greater allophonic detail in TTS front-end and database labels, or (2) identify contexts, such as pre-vocalic / post-vocalic positions within a syllable, that determine, in part, the allophonic variations. In our previous work [4], we tried to reduce such discrepancies by introducing allophones in the phone set. We differentiated one of the most variable phonemes, /t/, with three allophones: normal (with stop closure and burst) [t], flapped [dx], glottalized [q]. We updated letter-to-sound rules to predict such allophones in the certain phone context and re-labeled voice databases with the detailed phone set.

Synthesis quality was improved by that technique, however some other mismatches still remained unresolved. Selection of inappropriate consonant variants resulted in various phenomena. For example, unreleased /p/ chosen for /p/ in "PIN number" sometimes sounded like "bin number". In another case when the phone sequence /t ey t/ in "eight eight" is chosen for "Tate", the initial /t/ sound is missing, making it sound like "ate" instead of "Tate".

In this paper, a new phone labeling method that creates

better matches with phone realization in speech is proposed, which is a new technique to solve the phone variant problem in the current unit selection based TTS synthesis. The new phone set includes the distinction of consonant variants dependent on their position in the syllable structure, pre-vocalic and post-vocalic, which reduces missing consonants and consonant confusion.

2. Phonetic Variations

2.1. Allophone Mapping

Finding the optimal units from a speech inventory database is a key to synthesize high quality speech in a unit selection TTS system. However, it is not an easy problem because there are mismatches between the unit (phoneme) sequences called for by the TTS front-end and units (phone) labeled in the actual speech inventory. Those discrepancies started from the trivial fact that the TTS front-end is mainly written in *grapheme-to-phoneme* mapping rules rather than phone mapping [5] [6]. Before we discuss phonetic variations of a phoneme, we need to be reminded that "a phoneme is not a single sound, but a group of sounds. In fact, phonemes are abstract units that form the basis for writing down a language systematically and unambiguously." [7]

There are several approaches to bridge the gap between phoneme and phone: CART based methods [8], a method using a dictionary of alternate pronunciations [9]. In our previous work [4], we applied phoneme-to-phone mapping (allophone specification) rules to the /t/ sound which was frequently chosen inaccurately by unit selection.

- *flapping rule*:
  When an alveolar stop consonant like /t/ or /d/ is between two vowels, the second of which is unstressed, it becomes a voiced tap [dx]. For example, the /t/s in "pretty [p r ih dx iy]", "data [d ey dx ax]" may be replaced by a [dx].
- *glottalization rule*:
  When a voiceless alveolar stop locates before an alveolar nasal in the same syllable, it becomes a glottal stop. For example, the /t/ before syllabic [n] as in "button" may be replaced by a glottal stop [q].

Even though there are phenomena as shown above, it is still difficult to make a complete phoneme-to-phone mapping rule set because of uncertainty. For example, a word, "suit" in the TIMIT corpus [10] was found in four different phonetic realizations, [s uw tcl t], [s uw tcl], [s uw dx], [s uw q].

2.2. Phonetic Variations in Syllable

Phonetic variations of a consonant may be caused not only by surrounding phonetic context, but also by the position in the syllable [11]. A syllable in American English is generally composed of *onset* and *rhyme*. Any consonant or consonant cluster before the vowel forms the onset and the rhyme consists of a vowel and any consonant or cluster after the vowel.

The consonants before and after a vowel are often realized differently depending on their position in the syllable. For example, pre-vocalic stop consonants must have a burst part before the vowel begins while post-vocalic stop consonants may or may not have a burst part. For example, /d/ in "dark" has both the closure [dcl] and the burst [d] while /k/ after the vowel has only the closure [kcl]. Therefore, it may cause problems in speech synthesis, such as a drop-out, consonant confusion or word boundary confusion when a post-vocalic consonant segment is chosen to synthesize a pre-vocalic consonant.

3. Phonetic Enrichment Labeling

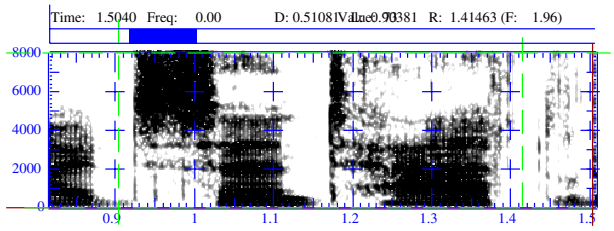
Selection of stop consonants is crucial in intelligibility of unit selection based TTS synthesis [4]. To avoid this problem, the penalties have been given to the units which violate syllable boundaries and word boundaries when the unit selection algorithm computes the target cost and the join cost of those units. However, it still occasionally chooses inappropriate units and makes conspicuous mistakes in synthesizing speech. In this paper, we introduce the pre-/post-vocalic distinction which prevents consonants in the rhyme from being used to synthesize onsets, and vice versa.

Table 1: Transcriptions using the pre-/post-vocalic distinction

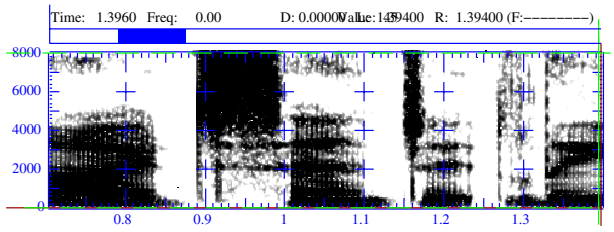
Word	Phonetic (TIMIT)	Proposed
clu**b**	kcl k l ah bcl b kcl l ah bcl	k l ah b_
grou**p**	gcl g r uw pcl p gcl g r uw pcl	g r uw p_
ha**nd**ba**g**	hh ae n dcl b ae gcl g hh ae n dcl b ae gcl	hh ae n_ d_ b ae g_
be**s**t	bcl b eh s tcl t bcl b eh s tcl bcl b eh s q	b eh s_ t_
da**rk**	dcl d aa r kcl k dcl d aa r kcl dcl d aa kcl k	d aa r_ k_
fu**ll**	f uh l f el	f uh l_
mo**re**	m ao r m ao ax m ao er m ao	m ao r_

The proposed phone labeling method distinguishes pre-vocalic and post-vocalic consonants. New phone symbols for the post-vocalic consonants are introduced while the phone symbols of pre-vocalic consonants are the same as the existing phone symbols. We label the post-vocalic consonant by adding an underscore ('_') like as /b_, d_, g_/. In addition to stop consonants, more distinctions are introduced to transcribe dark /l, r/s with /l_, r_/ and syllable final nasals with /m_, n_/. As shown in Table 1, each post-vocalic consonant covers various phonetic transcriptions by itself.

The voice database in the new TTS system is first labeled phonemically instead of allophonic variations. Then the pre-/post-vocalic distinction is applied to phonemic labels according to syllable boundary information given by the TTS front-end. The configuration of the TTS system is also changed according to the proposed phone set extension. In the new TTS system, the pre-/post-vocalic distinction module replaced the allophone mapping module used in the previous configuration. Instead of applying allophone mapping rules to the phoneme sequence predicted by the TTS front-end, the new TTS system assigns pre-/post-vocalic consonant symbols using the given syllable boundary information. The proposed distinctions embedded in the speech inventory also feed more suitable segments to the search algorithm of unit selection.



(a) by the reference TTS system



(b) by the proposed TTS system

Figure 1: Spectrogram of "sent at" in the prompt, "*A landslide sent at least a dozen homes crashing down a hill early Wednesday in Laguna Beach.*"

Figure 1(a) is a spectrogram that illustrates a type of common word-boundary confusion, for example in synthesis of "sent at" by the reference TTS system. The confusion is caused by selection of a word-initial (pre-vocalic) aspirated /t/ (taken from a recording of "...women **t**o ...") in the voice database and used instead in a word-final context. The resulting synthesized utterance sounds like "sen **t**at" instead of the intended "sen**t** at". In contrast, the spectrogram shown in Figure 1(b) illustrates the proper selection of an unaspirated syllable-final (post-vocalic) /t/ (taken from the context "...agreemen**t** at ..." in the recorded voice database). This version of "sent at", synthesized by the new phonetically enriched TTS system, causes no word boundary confusion to listeners.

4. Experiment

4.1. Listening Test

A listening test was conducted to evaluate whether the pre-/post-vocalic distinction leads to a measurable improvement in synthesis quality. The listening test was designed to compare two voices (female and male) and two TTS systems (the reference TTS version and the TTS version with phonetically enrichment), each used to synthesize 15 sentences (6 interactive prompts and 9 sentences from on-line news articles).

All 60 test stimuli were energy normalized to -20 dBov. Test files were renamed through symbolic links to prevent identification of test conditions. Listening tests were interactive and web-based. Listeners rated each test sentence on a 5-point scale from 1 (Bad) to 5 (Excellent). Listeners were 21 adults from the AT&T research community; 14 were native speakers of English, 7 were fluent non-native speakers of English.

4.2. Test Results

In the subjective rating test, the voices with the new phone set extension were rated significantly higher than the previous ones, 0.4 mean opinion score (MOS) improvement in the female voice and 0.26 MOS improvement in the male voice as shown in Figure 2. A repeated measures analysis of variance (ANOVA) was performed on the ratings data. ANOVA design consists of Voice + System + Sentence + Voice * System + Voice * Sentence + System * Sentence + Voice * System * Sentence.

All three main effects were statistically significant. The female voice (MOS = 3.505) was rated significantly ($p < 0.001$) higher than the male voice (MOS = 3.276). (Voice: $F(1,20) = 15.115$, $p < 0.001$) The phonetically enriched TTS version (MOS = 3.556) was rated 0.330 MOS higher than the existing version (MOS = 3.225), and that difference was highly significant ($p < 0.0001$). (System: $F(1,20) = 61.516$, $p < 0.0001$) There were also significant differences in ratings among test sentences. (Sentence: $F(14,280) = 20.381$, $p < 0.0001$)

Three of the four interactions were significant, but the most interesting interaction for our purposes, Voice*System, did not reach statistical significance (F(1,20)

= 3.454, $p < 0.078$). This indicates that the effect of improvements by the new phone set extension was statistically equivalent for both voices tested.

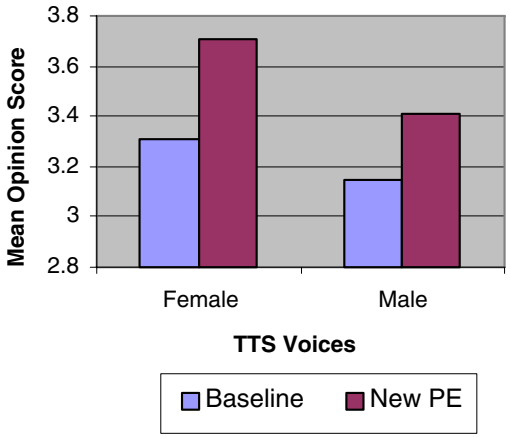


Figure 2: Comparison of the reference (baseline) TTS system versus the phonetically enriched (PE) TTS system with pre-/post-vocalic distinction

5. Discussion and Summary

Listening test result indicated that the proposed pre-/post-vocalic distinctive labeling improves synthesis quality of the test sentences. Several of the sentences synthesized by the reference TTS system have clear mistakes, but even in the other sentences which don't have evident mistakes it was observed that the proposed system is generally superior to the reference system.

Preserving the syllable structure by the pre-/post-vocalic distinction could lead to smoother joins in unit concatenation, not only avoiding selection of inappropriate synthesis units [12]. Even though the synthesis unit as used in our system is not limited to syllables or demi-syllables, the pre-/post-vocalic distinction eventually limited consonants in the rhyme (coda) not to be used for initial consonant (onset) synthesis. It could make it possible to have both flexibility and robustness in the unit selection based TTS synthesis.

In summary, a new phonetically enriched labeling method that differentiates pre-vocalic and post-vocalic consonants is proposed. The proposed method contributed significant improvement of synthesis quality in the unit selection based TTS system.

The proposed phone labeling method led unit selection to choose contextually accurate phone segments and minimized unit selection errors caused either by discrepancies between TTS front-end transcriptions and phone labels in the speech inventory or by lack of specificity in phoneme labels.

6. 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] Matthias Jilka and Ann Syrdal, "The AT&T German Text-to-Speech System: Realistic Linguistic Description," in *Proceedings of ICSLP*. 2002, Denver.

[3] Victor W. Zue and Martha Laferriere, "Acoustic study of medial /t,d/ in American English," *J. Acoust. Soc. Am.*, vol. 66, no. 4, pp. 1039–1050, 1979.

[4] Yeon-Jun Kim, Ann K. Syrdal, and Matthias Jilka, "Improving TTS by Higher Agreement between Predicted versus Observed Pronunciations," in *Proceeding of The 5th ISCA ITRW on Spech Synthesis*, 2004.

[5] Cecil Coker, "A Dictionary-intensive Letter-to-Sound Program," *J. Acoust. Soc. Am*, vol. 78, no. 1, pp. 78–87, 1985.

[6] Walter M. P. Daelemans and Antal P.J. van den Bosch, "Language-Independent Data-Oriented Graphme-to-Phoneme Conversion," in *Progress in Speech Synthesis*, pp. 77–89. Springer-Verlag, 1996.

[7] P. Ladefoged, *A Course in Phonetics*, New York: Harcourt, Brace, and Jovanovich, 1993.

[8] M.D. Riley and A. Ljojle, "Automatic generation of detailed pronunciation lexicons," in *Automatic Speech and Speaker Recognition*, chapter 12. Kluwer Academic Publishers, 1995.

[9] Wael Hamza, Ellen Eide, and Raimo Bakis, "Reconciling pronunciation differences between the front-end and the back-end in the IBM speech synthesis system," in *INTERSPEECH 2004*, 2004.

[10] W. Fisher, V. Zue, D. Bernstein, and D. Pallet, "An Acoustic-Phonetic Database," *J. Acoust. Soc. Am.*, vol. 81, 1986.

[11] Cecil Coker, Kenneth Church, and Mark Liberman, "Morphology and Rhyming: Two Powerful Alternatives to Letter-to-Sound Rules for Speech Synthesis," in *The ESCA Workshop on Spech Synthesis*, 1990, pp. 83–86.

[12] Ann K. Syrdal and Alistair D. Conkie, "Perceptually-based Data-driven Join Costs: Comparing Join Types," in *Proceedings of Interspeech '05*, 2005.