



# A comparison of pronunciation modeling approaches for HMM-TTS

Gabriel Webster, Sacha Krstulović, Kate Knill

Toshiba Research Europe, Cambridge Research Laboratory, Cambridge, United Kingdom

{gabriel.webster, sachak, kate.knill}@crl.toshiba.co.uk

## Abstract

Hidden Markov model-based text-to-speech (HMM-TTS) systems are often trained on manual voice corpus phonetic transcriptions, despite the fact that because these manual pronunciations cannot be predicted with complete accuracy at synthesis time, the result is training/synthesis mismatch. In this paper, an alternate approach is proposed in which a set of manually written post-lexical effects (PLE) rules modeling a range of continuous speech effects are applied to canonical lexicon pronunciations, and the resulting *matched PLE* phone sequences are used both in the voice corpus markup and at synthesis time. For a US English system, a subjective evaluation showed that a system trained on matched PLE markup and a system trained on manual phone markup were equally preferred, suggesting that it may be possible to replace manual pronunciations with matched PLE pronunciations, dramatically decreasing the time and cost required to produce an HMM-TTS voice.

**Index Terms:** speech synthesis, HMM-based, pronunciation modeling, post-lexical effects, continuous speech effects

## 1. Introduction

An important element of any text to speech (TTS) system is pronunciation modeling. This modeling consists not only of determining the pronunciation of each word in an input utterance independently, but also of modeling the continuous speech effects that result from pronouncing a sequence of words in quick succession. Continuous speech effects are typically dependent on word context and thus cannot be directly encoded in the lexicon. Thus during synthesis, a typical TTS system first generates the pronunciations for individual words, and then applies a set of post-lexical effects (PLE) rules to model the continuous speech effects.

For research in hidden Markov model-based TTS (HMM-TTS) [1][2], training and synthesis are often both conducted using a voice corpus with manually corrected phonetic transcriptions [3]. However, in a full TTS system where pronunciations are predicted from text, training on manual transcriptions has two potential drawbacks. Firstly, manually labeling a corpus is time consuming and costly. Secondly, manually corrected phone sequences are not fully predictable at synthesis time due to intra-speaker variation, where a single speaker realizes the same linguistic context differently at different times. This unpredictability introduces an element of training/testing mismatch into the system, which is known to hurt performance for machine learning tasks, including unit selection TTS [4][5][6][7][8].

An alternative to manual markup is to automatically mark up the speech corpus with the same kind of lexicon-plus-PLE-rule label sequences that are predicted at synthesis time. This approach has the benefit of using fully automatically derived phone labels (apart from any manual writing of PLE rules), and it results in voice corpus phonetic transcriptions that precisely match the synthesis-time pronunciations.

Furthermore, since the transcriptions model continuous speech effects, they are more accurate than simply using canonical pronunciations from the lexicon. Finally, in this scheme only a single pronunciation for each word token is input to the voice corpus markup process (see section 3.5.1), which in at least one study has been shown to give the best results for HMM-based automatic speech recognition (ASR) [9]. Because in this approach, the same set of PLE rules is applied in both the voice corpus transcriptions and the pronunciations generated at synthesis time, it is referred to here as *matched PLE*.

The two issues of accurate transcriptions and training/synthesis pronunciation matching thus form a classic trade-off question. On the one hand, manually corrected phone sequences guarantee an accurate labeling of the acoustic training data, which suggests that the acoustic models will be more accurately trained, but it results in training/synthesis mismatch which may degrade performance. On the other hand, matched PLE rules eliminate the training/testing mismatch but do not guarantee a fully accurate labeling of the training data. Given this trade-off, which approach yields the best synthesis quality? There is currently no literature that addresses this question for HMM-TTS. Therefore, for this research, two different approaches to voice corpus phonetic transcription were used to train HMM-TTS systems:

- (1) Manual phone sequences system: During training, manually corrected phone label sequences are used.
- (2) Matched PLE system: Canonical pronunciations are generated from a TTS front-end, and a set of manually written PLE rules modeling a range of continuous speech effects are applied to these pronunciations to form the input to forced alignment.

At synthesis time, both systems use the same front-end, consisting of the same canonical pronunciations and PLE rules that were used to generate the pronunciations for the matched PLE system.

## 2. HMM modeling of continuous speech effects

An HMM-TTS synthesis system implicitly models many phonetic context effects. During training, HMM-TTS uses a decision tree to cluster together similar training instances into shared HMM states. The questions available to the decision tree building process are based on the linguistic context features associated with the training instances, with the result that the training instances associated with each decision tree leaf are acoustically similar and share a matching subset of context feature values. This allows the synthesizer to generate acoustically different variants of the same phone based on its phonetic (and higher-level) context.

Systems based on both manual and automatic transcriptions make use of this implicit context modeling. In a manual phone sequence system, the effects that are modeled

are largely sub-phonetic. For example, a voiceless stop in English such as /t/ may be aspirated or unaspirated, depending on its phonetic context, and yet most phone sets only contain the single symbol /t/. In contrast, a manual system models continuous speech effects explicitly. For example, the vowel in the word *the* is canonically /i/, but changes to [ə] before words beginning with consonants and pauses.<sup>1</sup> A manual system transcribes both vowels accurately in the voice corpus transcriptions, and then uses a PLE rule at synthesis time to ensure that a correct and matching phone sequence is synthesized.

Because continuous speech effects are fundamentally just another kind of phonetic variation, an HMM synthesis system can in principle use its ability to implicitly model phonetic variation to model continuous speech effects as well. Consider again the case of vowel reduction in *the*. Even if the vowel is always transcribed canonically /i/ in a voice corpus, the acoustically distinct [i] and [ə] variants can be clustered into two different leaf nodes, because the acoustic realization is determined by the identity of the following phone, which is a context feature available to the decision tree. The correct variant will be chosen automatically at synthesis time, even though in this case the explicit phone is still /i/ in both the voice corpus markup and during synthesis. Crucially, this ability to implicitly model continuous speech effects depends on the phonetic context input to training and synthesis being the same.

In practice, however, there are two situations in which this implicit modeling of continuous speech effects may not work well. The first is when not enough training instances exist to create separate decision tree leaf nodes representing each acoustic realization. The second situation is for continuous speech effects that insert phones, such as glottal stop insertion between similar vowels (as in the word sequence *many evils*), because the number of HMM states that is normally used to model a single phone (typically 5) must now be used to model two phones. In these cases, explicit modeling of continuous speech effects through the application of PLE rules may give better results. When a PLE rule changes the phone sequence to explicitly model the relevant context, then it may change the context of that training instance to one which is accurate enough and for which enough training instances exist to robustly model.

In summary, what this means is that HMM-TTS may not need pronunciations that are perfectly accurate, but rather just accurate enough for the HMMs to implicitly model any remaining variation. This is the principle behind matched PLE pronunciation modeling: it provides training and synthesis pronunciations that are perfectly matched, while at the same time mostly accurate.

### 3. Method

#### 3.1. Data

All experiments were conducted with a voice corpus of 2,422 sentences spoken by a female speaker of American English. Pauses were included in the phone sequences. The corpus was manually transcribed using a proprietary but mostly conventional phone set. The phone set used separate symbols

---

<sup>1</sup> A symbol in forward slashes, such as /i/, is used to indicate an explicitly modeled phone, while a symbol in square brackets, such as [ə], indicates an implicitly modeled phonetic variant.

for released and unreleased stops, resulting in a total of 48 phones.

#### 3.2. PLE rules

The PLE rules that were used for the experiments were written by hand. The rules were written with the three goals of being linguistically well formed, being speaker-independent, and having high precision, in the sense that every PLE rule represented a continuous speech effect that is likely to be spoken by all speakers of American English in all contexts.

A different possibility for deriving PLE rules would have been to use manual PLE rules for the matched PLE system, and to automatically derive PLE rules from the manual phonetic transcriptions for the manual system [6][8]. This approach was not used for two reasons. Firstly, it would have confounded the effects of the phonetic transcription method with the accuracy of the PLE rule learner. Secondly, trying to learn PLE rules from relatively small voice corpora faces significant data sparsity problems. Some of the automatic rules generated in [8] overfit the training data, despite steps taken to minimize such overfitting, and many continuous speech effects that are desirable to model did not appear often enough in the training data for an automatic rule to be induced.

Space limitations prevent all of the PLE rules from being listed here. There were 188 rules in total. The rules modeled a variety of well known continuous speech effects, consisting principally of the following:

1. Changing the final vowel of the word *the* from /i/ to /ə/ before a consonant-initial word, pause, or end of utterance
2. Flapping a post-vocalic or post-rhotic word-final /t/ when the next word is vowel-initial
3. Changing the final vowel of *to/into/onto* from /u/ to /ə/ before a consonant-initial word, pause, or end of utterance
4. Inserting glottal stops between a word-final vowel and word-initial vowel, when these vowels are too similar
5. Reducing the vowel of *and/from/was* to /ə/ when the word is followed by something other than pause or end of utterance
6. Changing released stops to unreleased stops in appropriate contexts, such as when directly followed by another stop or affricate

Objective analysis of the PLE rules shows that they achieve a phone-level accuracy of 90.1% relative to the manual phone sequences, compared to 83.0% for the unmodified canonical pronunciations.

#### 3.3. Canonical versus matched PLE phone sequences

Given the ability of HMMs to implicitly model phonetic context, how well do they model continuous speech effects when only canonical lexical pronunciations are used for the voice corpus markup and at synthesis time? As a pilot experiment to determine whether such a system might be competitive, two HMM-TTS systems were built using a female American English voice corpus, one using only canonical pronunciations in the markup and for synthesis, and one using only matched PLE pronunciations.<sup>2</sup> Phone boundaries were

---

<sup>2</sup> In the canonical system, some reduced vowels were used in function words when they were more frequent. For example, the vowel of *the*,

automatically derived in both sets of voice corpus markup. Utterances containing a wide range of continuous speech effects were synthesized with both systems and analyzed by a native speaker of American English with phonetic training.

The results showed that the canonical pronunciations system was able to correctly model some continuous speech effects; however, the analysis also revealed several cases in which the canonical system was unable to fully model the continuous speech effect, while the matched PLE system correctly modeled the effect. These effects included unnatural glottal stops inserted between *the* and certain vowel-initial words and incorrect vowel reduction (or lack thereof) in many function words. Thus, a system using only canonical pronunciations was not used in the main evaluation.

### 3.4. Manual versus automatic time alignments

The manual markup for the voice corpus contains both manual phone label sequences and manual time alignments for those labels. The matched PLE markup, in contrast, can contain manual time alignments for the phone sub-sequences that match the manual phone sequences, but the phone sub-sequences that differ must contain automatically placed time alignments. This is because these non-matching matched PLE phone sequences are phonetically wrong (precisely because they differ from the manually marked phone sequences), and it would be often be difficult for a human to determine where boundaries between nonexistent (because incorrect) phones pairs should be placed.

The fact that the manual phone label sequences contain manual time alignments, whereas the matched PLE sequences must contain some automatic time alignments is a potentially confounding factor in an experiment comparing the two. Furthermore, because training the HMMs involves re-estimating the phone boundary locations, there is reason to believe that manual phone boundaries might not be a significant factor in synthesis quality. For these reasons, a second pilot experiment was conducted comparing a system trained on manual phone label sequences and manual time alignments to a system trained on manual phone sequences and automatic time alignments. Forced alignment using the Hidden Markov Model Toolkit (HTK) [10] was used to generate the latter markup, with the use of a token-level lexicon to guarantee that each word token in the automatic time alignment markup received the same label sequence as in the manual markup.

An informal subjective evaluation of 11 listeners each listening to 30 utterance pairs was conducted, as was a linguistic evaluation by a native speaker of American English with phonetic training. Both analyses found no overall preference for manual time alignments over automatic time alignments. Thus in the main evaluation, automatic time alignments were used for both systems.

### 3.5. Main evaluation system training

#### 3.5.1. Voice corpus phonetic markup

HTK was used to generate the time alignments for the voice corpus phonetic markup for both systems of the main evaluation. For both systems, the manually corrected pause locations were used, in order to avoid any confounding effects

of automatic pause prediction. To generate the time alignments, HTK was run in forced alignment mode with a token-specific pronunciation lexicon (i.e., a lexicon containing a specific pronunciation for each word token in the voice corpus); this was necessary for both systems, since manually corrected pronunciations are obviously token-specific, and many PLE rules apply across word boundaries, resulting in context-specific pronunciations that are easily handled with token-specific pronunciations.

#### 3.5.2. HMM synthesizer training

The HMM-based Speech Synthesis System (HTS) [1][2] was used to train the systems. The model setup was the following: 5 states per model; observation vector comprising 39 Line Spectrum Pair coefficients, log gain and 5 band-aperiodicity coefficients from STRAIGHT analysis; log-F0 from a proprietary automatic F0 extractor; addition of first-order derivatives only; and MDL factor defaulted to 1.0 for all streams. The context features and set of questions were computed by proprietary software. Features and questions generated from part of speech (POS) and parser output were turned off for this experiment. The training setup reproduces the mono-speaker training procedure illustrated in the HTS tutorial demonstrations, with some parallelization to speed up the training. The speech samples were synthesized with `hts_engine`, with Global Variance turned off, and a proprietary external vocoder.

### 3.6. Main evaluation: Manual versus matched PLE phone sequences

For the main experimental condition of comparing manual phone sequences with matched PLE phone sequences in the voice corpus markup, a subjective evaluation was carried out. The first system was trained on manual phone sequences, while the second system was trained on matched PLE phone sequences. Because the second pilot experiment (see section 3.4) found no difference between automatic and manual phone boundaries, automatic boundaries were used for both systems. At synthesis time, the same front-end, including the same set of manual PLE rules, was used to generate the phone sequences. This means that the manual phone sequence system was the same as the manual phone sequence/automatic time alignment system from the second pilot experiment.

The matched PLE phone sequences were generated by running the text corresponding to the voice corpus utterances through the front-end processing of the Toshiba ToSpeak TTS system. This means that homographs were disambiguated using the part of speech (POS) tagger of the ToSpeak front-end. While this approach is not error-free, it does precisely match the pronunciations that are generated at synthesis time, thus potentially allowing the HMM modeling to compensate for any errors.

A total of 15 native speakers of North American English took part in the evaluation. Forty-two utterances were synthesized for the evaluation. The utterance texts were chosen such that about half contained clear continuous speech effects handled by the PLE rules. Each listener wore headphones and listened to 30 utterances chosen at random. For each utterance, the listener was presented with the utterance as synthesized by the two systems, in random order, and then was asked to indicate which of the sentences sounded more natural: the first sentence, the second sentence, or neither sentence. Subjects were allowed to listen to each sentence pair multiple times if they wanted.

---

though canonically [i], is most often pronounced [ə], so it was used as the vowel in the canonical pronunciation.

## 4. Results and discussion

The overall results of the subjective evaluation are given in Table 1. The manual phone sequence system was preferred in 43% of all responses, while the matched PLE system was preferred 41% of the time. To test the statistical significance of this difference, an ANOVA was performed on the results. The difference was not significant ( $p=0.42$ ).

	Manual	Matched PLE	Neither
Count	195	184	71
Percentage	43%	41%	16%

Table 1: Overall preference of manual and matched PLE systems

This result suggests that it may be possible to use automatic matched PLE phonetic transcriptions to generate an HMM-TTS system that is equal in quality to a system trained on manually corrected phonetic transcriptions (bearing in mind that for this experiment, the matched PLE system included manual pause locations, so the phonetic markup was not completely automatic).

In addition to the subjective evaluation, a phonetic analysis focusing on how well continuous speech effects were modeled was carried out. This analysis showed that two main kinds of continuous speech effects appeared to be clearly less well modeled in the matched PLE system. The first effect was flapping of /t/ when the next word began with a vowel: instead of flapping the /t/, the matched PLE system (implicitly) inserted a glottal stop at the beginning of the second word. The second effect was vowel changing in words like *the* and *to* when the next word began with a vowel: the matched PLE system changed the vowel appropriately, but in addition, it inserted a glottal stop at the beginning of the second word. Preliminary investigation revealed that this behavior was the result of optional variation within the voice corpus, where the speaker would, for example, sometimes flap a /t/ as expected, and sometimes insert a glottal stop instead. In contrast, the manual system did not insert any glottal stops in these contexts, and the result was clearly smoother sounding modeling of these effects.

To analyze the perceptual effect of these differences, a post-hoc analysis of the subjective evaluation data was conducted. In this analysis, the utterances were divided into two categories based on whether the last eight words of the utterance contained a continuous speech effect whose acoustic realization was clearly different in the matched PLE and manual systems (the maximum distance from utterance end was chosen to minimize the possibility of a single effect being perceptually “lost” within a long utterance). Fourteen of the 42 utterances contained such differences. The results are shown in Table 2.

	Manual	Matched PLE	Neither
No difference	44%	40%	16%
Difference	41%	44%	15%

Table 2: Preference of manual and matched PLE systems for sentences containing clear continuous speech effect modeling differences, or no such differences

Interestingly, the presence of clearly differently modeled continuous speech effects appears to make no significant difference in system preference. In fact, the preference for the matched PLE system actually trends towards *increasing*

slightly in cases where the matched PLE system models a continuous speech effect differently. Given that many of the differences involved glottal stop insertion in the matched PLE system, it may be the case that listeners prefer the more deliberate enunciation implied by the additional glottal stops over the manual system, which pronounces these effects in a smoother, but quicker fashion. If this is true, a preference for more clearly enunciated speech may be due to the increased amount of attention required to understand synthesized speech relative to natural speech.

## 5. Conclusion

The experiments described in this paper show that an HMM-TTS system trained on a voice corpus marked up with matched PLE phonetic transcriptions, which consist of canonical pronunciations with PLE rules applied and which precisely match the phone sequences that are predicted at synthesis time, can achieve synthesis quality that is no worse than that of a system trained on manual phonetic transcriptions. This result suggests that it may not be necessary to spend the time and money needed to manually correct phonetic transcriptions in order to achieve the best possible quality synthesis.

There are several following steps for this research. Firstly, the manual pauses used in the experiments for this paper should be replaced by automatically detected pauses, in order to generate an HMM-TTS system based on fully automatic phonetic markup. Secondly, similar experiments should be run on different voices, languages, and corpus sizes, in order to determine whether the results obtained here hold for other types of voice corpora. And thirdly, experiments should be performed in which the voice corpus forced alignment process is allowed to make very limited choices, in terms of phone sequence, in cases when the forced alignment is likely to make the right choice and it is likely to help overall system quality, with the goal of further improving the already promising quality of the matched PLE approach.

## 6. References

- [1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis”, in *Proc. ICASSP*, pp.1315-1318, 2000.
- [2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, “The HMM-based speech synthesis system version 2.0”, in *Proc. Sixth ISCA Workshop on Speech Synthesis*, Bonn, Germany, 2007.
- [3] H. Zen, personal communication.
- [4] M. Jilka and A. Syrdal, “The AT&T German Text-to-Speech System: Realistic Linguistic Description,” in *Proc. ICSLP*, Denver, 2002.
- [5] J. Fackrell, W. Skut, and K. Hammervold, “Improving the accuracy of pronunciation prediction for unit selection TTS”, in *Proc. Eurospeech*, Geneva, 2003.
- [6] Y.J. Kim, A. Syrdal, and A. Conkie, “Pronunciation Lexicon Adaptation for TTS Voice Building”, in *Proc. ICSLP*, Jeju, Korea, 2004.
- [7] Y.J. Kim, A. Syrdal, and M. Jilka, “Improving TTS by higher agreement between predicted versus observed pronunciations”, in *Proc. Fifth ISCA Workshop on Speech Synthesis*, 2004.
- [8] G. Webster, T. Burrows, and K. Knill, “A Comparison of Methods for Speaker-Dependent Pronunciation Tuning for Text-to-Speech Synthesis”, in *Proc. Eurospeech*, Lisbon, 2005.
- [9] T. Hain, “Implicit modelling of pronunciation variation in automatic speech recognition”, *Speech Communication* 46:171–188, 2005.
- [10] S. Young et al., *The HTK Book (for HTK Version 3.4)*, Cambridge: Cambridge University Engineering Department, 2009.