

Schwa Variants in American English

H. Timothy Bunnell¹, Jason Lilley²

¹Nemours Biomedical Research, A. I. duPont Hospital for Children, Wilmington, DE, USA

²Department of Linguistics, University of Delaware, Newark, DE, USA

bunnell@asel.udel.edu, lilley@asel.udel.edu

Abstract

In this study, we examine the acoustic structure of schwa variants in a speech corpus intended for concatenative TTS. Our goals are two-fold. First, as a matter of academic interest, we seek to characterize schwa acoustics in a speech corpus designed to provide broad coverage of di- and tri-phone contexts. This characterization extends recent work on the distinction between the barred-i variant of schwa /ɨ/ and the more central /ə/ form of schwa [1]. Our second goal is to improve the concordance between our transcriptions as generated by the front-end of our TTS system, and the phonetic behavior of talkers who record corpora for concatenative TTS. Based on analysis of a single talker’s corpus, our results support the claim in [1] that /ɨ/ is more common than generally assumed. However, the claim that /ɨ/ characterizes all stem-medial schwas is not well-supported by our data.

Index Terms: acoustic phonetics, schwa, speech synthesis, unit selection, parallel-state HMM

1. Introduction

The acoustic structure of English schwa is believed to vary greatly with context. For example [2], [3], [4], and others have provided evidence that schwa pronunciation is greatly influenced by the neighboring context, by adjacent consonants as well as by neighboring vowels in neighboring syllables. Making matters more complex, descriptions of English generally seem to agree that English has two reduced vowels, but disagree on the proper transcription and distribution of the pair. In [1] it is argued that the contrast in minimal pairs such as *Rosa’s/rozes* in some dialects demonstrates that morphological boundaries can play a role in reduced vowel distribution. In particular, they argue that stem-final reduced vowels such as in *Rosa’s* are pronounced with a non-high, somewhat back vowel properly transcribed [ə], while most stem-medial reduced vowels are high and more front (but still generally central), and so should be transcribed with the “barred-i” [ɨ]. However, it’s possible that the two vowels are merely the extrema of a continuous range of possible schwa pronunciations.

While these questions are of obvious interest to phoneticians and phonologists, they also have practical implications for the design of concatenative TTS systems, particularly those that aim to generate high quality speech based on minimal recorded speech corpora. Such systems must choose material for a speech corpus carefully to include all perceptually distinct subtypes of segments. However, including unnecessary segmental distinctions—if they are coded as different segments—leads to geometric increases in corpus size. Moreover, as a practical matter, discordances between the formal transcriptional system used by a TTS system and the

actual productions of speakers who are recording speech for a unit selection TTS system lead to errors in constructing and searching the speech synthesis database.

To address both of these issues, we have examined a corpus of speech recorded for the ModelTalker TTS system. The 55-symbol transcription set for ModelTalker has two symbols, “AX” and “IX”, for schwa subtypes. “AX” is intended to stand for the low back subtype, and “IX” for the high front subtype. The ModelTalker transcription routines look at a number of factors to determine which schwa symbol to use. The goal of this research is to test and improve the performance of these routines.

Traditional instrumental studies of acoustic phonetics generally analyze formant frequencies of segments to determine acoustic structure. But the process of measuring formants is time-consuming, and typically requires careful human judgments to correct formant tracking errors. This problem is exacerbated for segments like schwa, which may have very brief durations and less well-defined formant patterns. Because of the time involved, it is impractical to examine large amounts of data.

As an alternative to traditional formant-based analyses, in this paper we explore the application of HMM training procedures to automatically classify phonetic types and subtypes. The HMM training uses Mel-frequency cepstral coefficients (MFCCs) as the input vectors. Unlike formant frequencies, MFCCs can be calculated quickly and automatically by machine.

Although the traditional HMM design (Figure 1), in which the states are arranged in series and no back-transitions are allowed, provides an accurate enough model to be used in many speech synthesis and recognition applications, it is not so useful as a tool to investigate subphonemic variation, because

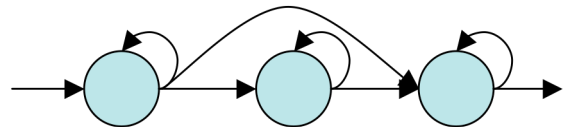


Figure 1: Typical 3-state HMM.

there is little natural allowance for variation built into the model. In its most basic form, each model state is described by a single vector of mean cepstral coefficients which characterizes a single spectral shape, and although the number of analysis frames associated with each state may vary from example to example, there is no variation in which states are used.

Many HMM systems improve their performance by explicitly modeling subphonemic variation, in various ways. Some do this by building a set of models for each phoneme, each model built for an explicit context, e.g. triphone models.

An investigator studying contextual variation can then analyze these model sets using statistical clustering techniques. One problem with this approach is that a lot more data is needed to train a large set of models, as compared to a single one. Another problem is that this approach requires one to guess in advance how the variation is distributed. Using triphone models, for example, presupposes that subphonemic variation is determined chiefly by the immediately adjacent phonemes, but minimal pairs such as *Rosa's/rozes* show that much different schwa subtypes can occur in the same triphone context.

Other systems accommodate variation through the use of mixture models, in which each state is described by a set of Gaussian distributions, rather than a single one. The probability of a state generating a particular observation vector is then defined as the weighted sum of the probabilities computed from each Gaussian. Gaussian mixture models can capture much more complicated variability in segment structure. However, because each mixture is associated with a single state, there is no possibility for different state-to-state transition probabilities to be associated with individual Gaussians.

We avoid this problem by using what we call *parallel-state* HMMs (psHMMs). In a psHMM, each of the states of a serial HMM is replaced with a *set* of parallel alternative states (see Figure 2). The number of alternative states in each set need not be identical. Transitions are allowed from a state in one state-set to any of the states in the following adjacent state-set (but not to previous state-sets). Self-transitions (transitions back to

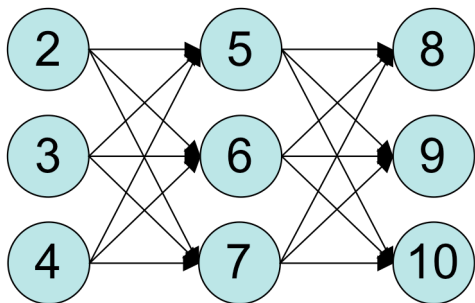


Figure 2: *Parallel-state HMM used in the present study. Not shown here, skips from each of the first set of states (2-4) to each of the third set (8-10), were also allowed, as were self-transitions within each state.*

the same state) are also allowed, but transitions between different states in the same set are forbidden. Skip transitions to non-adjacent following state-sets can also be allowed.

After such a model is trained on all the examples of a phoneme in a corpus, the model can be used to label each of the examples according to the sequence of states which has the highest probability of generating that example. Such labeling provides a natural clustering of the examples according to their acoustic similarity in successive regions of the segment. In our example in Figure 2, there are $3 \times 3 \times 3$ possible paths through the model (not counting self-transitions), resulting in 27 possible paths or clusters. If skip transitions between states in the first and third columns are allowed, the number of possible paths increases to $3 \times 4 \times 3$. It is this 36-path configuration we use in the study described below. In addition to the path data, the log likelihood of the example associated with each state and the number of frames associated with each state can be used as additional factors in a traditional statistical clustering analysis.

Another advantage of this method is that the clusters produced do not require a prior hypothesis about the expected nature or distribution of subtypes. The data will self-organize into the acoustically best set of clusters in each state-set as a consequence of the HMM training procedure.

2. Method

2.1. Stimuli

The materials for this investigation were the recordings of 1837 utterances (from single words to complete sentences) of a single adult female speaker. The utterances were designed to be phonetically diverse and cover a wide range of the most common biphones of English. Recording was done in a sound-attenuated recording booth at a sampling rate of 16 KHz with 16-bit sample resolution.

2.2. Acoustic analysis

All acoustic analysis, as well as the HMM training and recognition described below, was done with tools provided by HTK [5]. The raw speech signal was converted into input vectors using a 25 msec window and 10.0 msec frame step. The speech signal was pre-emphasized by a factor of 0.97, Hamming windowed, and converted to Mel-frequency cepstral coefficients. As is common practice, we used 13 basic MFCCs (C_0 to C_{12}) as well as the corresponding delta and acceleration coefficients (for a total input vector size of 39).

2.3. HMMs

A single parallel-state HMM was constructed for both the “AX” and “IX” schwa labels in the ModelTalker transcription set. This psHMM had 3 serially-arranged state-sets, with 3 parallel single-Gaussian states in each set as per Figure 2. Initially, no skip transitions were allowed. However, skip transitions from the first state-set to the third were added after initial model training, as described below.

Traditional 3-state monophone HMMs were constructed for each of the other 53 “phones” (including 3 silence “phones”) in the ModelTalker transcription set for English. Each model consisted of 3 single-Gaussian states, arranged in series. No skip-transitions were allowed in these models.

2.4. Training

In each training step described below, the parallel-state schwa model was trained on all instances of both schwa subtypes (AX and IX) in the corpus.

Transcriptions of the 1837 utterances in the training corpus were automatically aligned to the utterances using an independently developed HMM set. These alignments were used in the first two steps of the training process. In the first step, each of the 54 models was trained in isolation on the corresponding segments of speech in the corpus using the Viterbi algorithm. For each model, the algorithm was repeated 20 times or until convergence was achieved. Model parameter variances were not allowed to fall below 1% of the global corpus variance. In the second step, the same procedure was repeated, but this time using the Baum-Welch (BW) algorithm.

Next, all models were simultaneously given two rounds of embedded-unit BW re-estimation. The pruning threshold ranged from 250.0 to 1000.0 in increments of 150.0. As before, variances were kept above 1% of the global variance. Next,

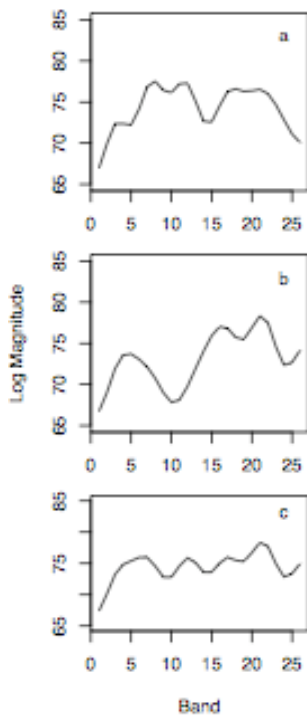


Figure 3: Example magnitude spectra calculated from the mean cepstra of the schwa model states. One log magnitude is calculated for each of 26 filter bands spaced evenly over the Mel-frequency spectrum from 0 to the Nyquist frequency.

skip transitions were added to the schwa psHMM from every state in the first state-set to every state in the third state-set, making a total of nine added skip transitions. Finally, all models were given three more rounds of embedded BW re-estimation.

After training, the model set was used to re-segment the corpus utterances. The end result was a set of state alignments for every utterance in the corpus.

3. Results

3.1. Alignment

To check the accuracy of the alignments, an expert manually segmented 160 of the utterances, and the differences between the hand aligned segment boundaries and those determined by HMM training were calculated. There were a total of 2071 segment boundaries in the 160-sentence comparison set. Over all segment boundaries, the median absolute difference between the hand-aligned boundaries and HMM-aligned boundaries was 9.2 msec. For the 362 boundaries involving schwa, the median absolute error was 8.6 msec. For the sake of comparison, we built an identically trained HMM set in which the single schwa psHMM was replaced with separate standard 3-state models for AX and IX. The median absolute error for boundaries involving either variant of schwa in this set was 9.4 msec.

3.2. Spectral properties

By inverting the cepstrum analysis associated with the first 13 feature vector elements of the HMM states, it is possible to recover the Mel frequency spectrum associated with each state. We generated spectra for all nine states of the psHMM, and

among these observed three distinct shapes, as illustrated in Figure 3. The top panel (a) of Figure 3 shows what we can refer to as a “low-back” form due to the high frequency of F1 and low frequency of F2. The second shape (panel b), termed “high-front” here, more closely resembles spectra typically associated with high front vowels: a comparatively low-frequency F1 and high-frequency F2. The third form we observed (panel c) is here termed a “central” form due to the more nearly even spacing of formant peaks (in Mel unit spacing). Strikingly, we found a clear example of each spectral shape associated with a state in each of the state-sets in our psHMM.

3.3. HMM path analysis

In the process of aligning the corpus transcriptions, we recorded the state path through the schwa psHMM for each of the 1904 instances of schwa in the corpus. Of the 36 possible unique paths through the psHMM, 25 were actually used. These are shown in Table 1, along with the spectra associated with each path and a count of the number of AX and IX vowels (per our normal transcription) that took each path.

Overall, a relatively small number of schwa tokens (318 out of the 1904) were associated with a low-back form in any state, compared to the central (1251) and high-front (1495) state forms. As one might expect, spectral shape seems to play a large role in determining possible state paths: nine of the 11 unused state paths involve direct transitions from one of the extreme spectral shapes (low-back or high-front) to the other. For the majority (1072) of schwa tokens, the broad spectral shape of the schwa remained unchanged between the initial and final state-sets. Of the rest, 360 tokens transitioned from the central form to one of the other two, and 435 transitioned from one of the extreme forms to the central one. Surprisingly, however, while only one schwa took a path from the initial high-front state to the final low-back one, 118 schwas transitioned from low-back to high-front states (including 17 which do not pass through a central shape along the way).

ModelTalker transcribes schwas associated with low-back forms in any state as “AX” 90% of the time (including 86% of the 118 tokens which transition to a final high-front state). In comparison, ModelTalker transcribes only 49% of the schwas associated with any high-front states as “IX” (even schwas which pass through 2 high-front states are only labeled “IX” 55% of the time). As for central-form states, ModelTalker transcribes 66% of the schwas associated with these as “AX.”

3.4. Contextual factors

One claim for schwa is that phonetic (and other) contextual factors strongly influence schwa acoustic structure. If that is the case, it should be possible to identify specific context factors that are strongly associated with each path and spectrum of schwa. To investigate this, we examined the preceding and following segmental and prosodic contexts for each path and its associated spectrum type. The pattern that emerged from our preliminary analysis suggested that high-front forms are most frequently seen adjacent to obstruents, and central forms adjacent to sonorants. On the other hand, low-back forms frequently occurred in word-initial or word-final positions, particularly utterance-initial or -final. These patterns were more consistent for the final state-set than the initial one.

We looked more closely at the 118 schwa tokens which took a path from an initial low-back state to a final high-front one. These followed the same pattern: all occurred before an

Table 1: Paths used by both nominal schwa segments after psHMM training. States are referenced to Figure 2. “LB,” “HF”, and “C” indicate low-back, high-front, and central spectra. An “s” indicates the middle state was skipped.

Path	Path Spectra	AX	IX	Total
2-5-8	LB-LB-LB	25	0	25
2-5-9	LB-LB-HF	4	0	4
2-5-10	LB-LB-C	12	0	12
2-7-8	LB-C-LB	1	0	1
2-7-9	LB-C-HF	88	14	102
2-7-10	LB-C-C	35	7	42
2-s-8	LB-s-LB	21	0	21
2-s-9	LB-s-HF	10	2	12
2-s-10	LB-s-C	26	8	34
3-6-9	HF-HF-HF	37	206	243
3-6-10	HF-HF-C	91	69	160
3-7-8	HF-C-LB	1	0	1
3-7-9	HF-C-HF	105	22	127
3-7-10	HF-C-C	39	13	52
3-s-9	HF-s-HF	218	212	430
3-s-10	HF-s-C	45	90	135
4-5-8	C-LB-LB	51	0	51
4-5-9	C-LB-HF	1	0	1
4-5-10	C-LB-C	5	0	5
4-6-9	C-HF-HF	8	26	34
4-6-10	C-HF-C	1	26	27
4-7-8	C-C-LB	7	0	7
4-7-9	C-C-HF	116	51	167
4-7-10	C-C-C	59	8	67
4-s-10	C-s-C	69	57	126

obstruent, and most were word-initial (64 were utterance-initial). Why only one of the several schwa tokens in the mirror-image context (word-final after an obstruent) took the mirror-image path (from high-front to low-back states) is still unclear; all the other such schwa tokens began in a central or low-back state.

4. Discussion

These results support previous descriptions of schwa that suggest there are distinct forms that differ in both height and fronting (e.g., [1]) and arise in different contexts ([2]-[4]): a relatively low, back variant, commonly designated [ə], and a higher, fronter vowel [ɨ]. Our work also supports the claim in [1] that the low-back variant is relatively uncommon, while [ɨ] is more prevalent than often assumed. Our study, however, reveals a third, centralized variant, which complicates transcription in a system which would allow only two symbols

for schwa. In [1], it is claimed that the relatively uncommon [ə] has a fairly constant pronunciation, while the common [ɨ] is highly variable, particularly in F2. If this is the case, then the “central” form might best be thought of as an extreme pronunciation of [ɨ] and be uniformly labeled IX. However, over 100 schwa tokens were associated with both “central” and “low-back” states, suggesting that at least some of these schwas are better labeled AX. Obviously, a third option is to use a separate symbol for centralized schwa subvariants. In addition, we found instances of schwa which change over time from a low-back shape to a high-front one, defying easy transcription. The distribution of this subtype, and the question of why the reverse trajectory is very rare, merit further study.

5. Conclusions

While many of the patterns determined by HMM training align fairly well with our TTS symbol set and usage of AX and IX, the analysis revealed several hundred cases where segments labeled AX are better fit to the HMM paths for IX-like schwa. To a lesser extent, there are also examples of segments we label IX aligning to paths associated more commonly with AX. Many of these instances represent cases where we expect to be able to make improvements to our present transcription system, and hopefully to the resulting TTS quality.

An interesting matter for future study is the question of whether adding a third more neutral schwa symbol to code the third class of schwa from our analyses might result in improvements in synthetic speech quality.

6. Acknowledgements

This work was supported by grant number R42-DC006193-03 from NIH and by Nemours Biomedical Research.

7. References

1. Flemming, E. and S. Johnson, *Rosa's roses: reduced vowels in American English*. Journal of the International Phonetic Association, 2007. **37**(1): p. 83-96.
2. Browman, C. and L. Goldstein, *Targetless schwa: An articulatory analysis*, in *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, G. Docherty and R. Ladd, Editors. 1992, Cambridge University Press: Cambridge.
3. Koopmans-van Beinum, F.J., *What's in a schwa? Durational and spectral analysis of natural continuous speech and diphones in Dutch*. Phonetica, 1994. **51**: p. 68-79.
4. Lindblom, B., *Spectrographic study of vowel reduction*. J Acoust Soc Am, 1963. **35**(143-162).
5. Young, S.J., *The HTK Hidden Markov Toolkit: Design and Philosophy*, in *Technical Report*. 1993, Department of Engineering, Cambridge University.