Individuation of postlexical phonology for speech synthesis

Corey Miller

Motorola Speech Processing Research Laboratory, 1301 E. Algonquin Road, IL02-982, Schaumburg, Illinois 60196, USA

ABSTRACT

Postlexical phonology exhibits a great deal of interspeaker variation, even within widely construed dialects such as American English. Contemporary speech synthesizers have concentrated on modeling the acoustic properties of individual speakers, but not necessarily their phonologies. We discuss a method for inferring individual postlexical phonologies from labeled corpora. Using this method, individualized phonologies can be combined with individualized acoustic models, thereby enabling a speech synthesizer to achieve a closer facsimile of an original voice.

1. INTRODUCTION

A predominant modern paradigm for speech synthesizer development can be characterized as learning as much as possible about the speech of an individual speaker in an effort to reproduce that speaker's voice as faithfully as possible in synthetic productions. This concentration on individual speech characteristics has usually been concerned with acoustic or "back-end" properties, while "front-end" linguistic issues have usually been handled in a general way for all speakers of a particular language or dialect supported by a given system.

We believe that coupling efforts to achieve individualized acoustic characteristics with efforts to achieve individualized higher-level linguistic characteristics can contribute to a more faithful representation of an individual's speech, and consequently, a more plausible and acceptable synthetic speech quality.

In this paper, we discuss a speech synthesizer architecture that supports this model. We then present a method for learning postlexical variation and apply it to the speech of three speakers of American English. Finally, we examine the differences in the postlexical phonologies of these speakers and show how the method described allows these differences to surface in their synthesized speech.

2. ARCHITECTURE

Let us assume a synthesizer that has two principal modules: a linguistic module for transforming text into a linguistic representation and an acoustic module for transforming linguistic representations into speech. In the Motorola speech synthesizer [1], the acoustic module is trained on linguistic representations derived from a hand-labeled speech database. This labeling is performed in conformity with a protocol derived from both the TIMIT [2] and CSLU [3] labeling guidelines. This means, among other things, that some allophony is represented, and that stop closures and releases are labeled separately.

The linguistic representation can be derived from text by consulting a dictionary and a letter-to-sound conversion procedure for word pronunciations. Most common electronic and printed dictionaries (and the letter-to-sound conversion procedures that train on them) do not supply transcriptions with the allophonic and phonetic detail employed in the speech labeling described above. As a result, there arises a necessity to map between "lexical" pronunciations in the dictionary and "postlexical" pronunciations in the speech labeling [4].

While it has been suggested that dialect-specific lexicons can be useful for speech synthesis [5], we believe that a single "generalized" or pandialectal dictionary [4] can be used in conjunction with the postlexical learning procedure to be described below.

3. VARIATION

Sociolinguistic studies have demonstrated that there is great variation in postlexical phonology, even among speakers in the same speech community [6]. We therefore believe that attempts to use general postlexical rules for all speakers of a given language or dialect would fail to capture the subtle variations that exist among speakers. Research in dialectology has shown that speakers from different regions or groups may differ in their lexical representation of word pronunciations [7]. In addition, since the choice of what symbol to label can be somewhat arbitrary, especially in the case of subphonemic labels, it is important to identify a mechanism to learn labeler idiosyncrasies.

There are two principal reasons to aim for fidelity between the kinds of linguistic representations that were labeled by hand and those that are generated by the linguistic module of a synthesizer. First, it is important to provide the acoustic module with linguistic representations that are as similar as possible to those upon which it was trained, in order to get the most faithful output. For example, if a dictionary provides the pronunciation /kæri/ for *carry*, and the labeled database contains /kɛri/, the quality of an /æ/ vowel before /r/ may prove difficult to synthesize plausibly, since such a sequence may not have been seen in the training data. Second, even if a plausible /æ/ could be synthesized in this environment, it might prove jarring to listeners, who might find the pronunciation unnatural, given either their familiarity with the speaker or the dialect being synthesized.

4. METHOD

We put lexical pronunciations from our dictionary and postlexical pronunciations from our labeled database into one-

to-one alignment using dynamic programming and a specialized cost function that takes the similarity of lexical and postlexical phones into account [4]. In the case of stops, which are often represented as two phones in the postlexical transcription, we collapsed the postlexical phones into pseudophones that could be aligned with a single lexical phone. We then encoded the lexical phones, along with prosodic information, in a numeric representation that was provided to a recurrent neural network. In order to incorporate preceding and following lexical phone context, the network was constructed with a running window of three phones. The neural network outputs a postlexical phone and diacritic, or silence in the case of deletion, for each input phone. Details about the neural network simulator used are provided elsewhere [1,4].

Figure 1 illustrates the block-oriented architecture of the network used. The output postlexical phone information is contained in block 11, the diacritic information in block 12; and the combined information is fed to block 1. The recurrent buffer, block 13, allows the network to take into account the previously emitted postlexical phone when deciding which postlexical phone to emit for a given lexical phone. This was found to improve performance; for example, without the recurrent buffer, the sequence [ʔ n] was frequently predicted for the final phones of words like "button". With the recurrent buffer, [ʔ ṇ] began to surface. Riley and Ljolje incorporate a similar constraint by using a Markov model [8].

Blocks 2-5 contain the input information as follows: block 2 contains lexical phone information, block 3 contains feature information for each lexical phone, block 4 encodes stress and accent information, and block 5 encodes information about adjacency to various syntactic and prosodic boundaries. Blocks 6-10 are hidden layers. The input encoding is discussed more thoroughly in [4].

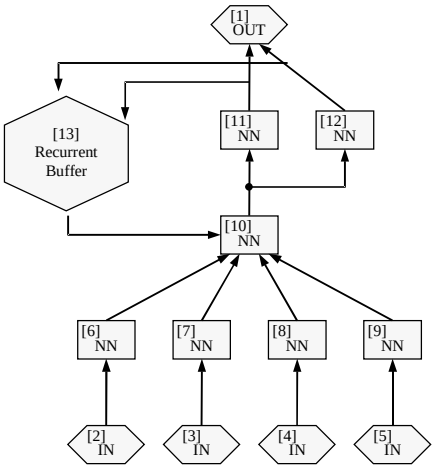

Figure 1: Postlexical neural network architecture. Blocks labeled "IN" are input blocks, block labeled "OUT" is an output block, and blocks labeled "NN" are hidden layer neural network blocks.

We trained this network on the hand-labeled speech of three individuals whose demographic information is shown in Table 1. The speech data for all speakers was read speech including the Harvard sentences [9]. Speaker mhb1 was labeled by an engineer at Motorola with some training in linguistics, speaker mps1 was labeled by two linguists at Motorola, and speaker fdb1 was labeled by a group of four linguists at Ohio State University. Speaker mhb1 was labeled predominantly according to the TIMIT labeling scheme, while speakers mps1 and fdb1 were labeled predominantly according to the CSLU labeling scheme. The major difference between these two schemes as far as this paper is concerned is that the CSLU scheme employs the system of Worldbet phones and diacritics [10]. In that system, phonemic information is symbolized through phone labels, while allophonic information is primarily symbolized through diacritic labels, separated from phone labels by an underscore, "_".

Speaker	Hometown	Age at recording (years)	Profession
mhb1	Chicago	36, 38	engineer
mps1	suburban Boston	40	professor
fdb1	suburban Chicago	38	trained as a nurse

Table 1: Speaker demographic information

Table 2 shows the amount of training data used for each speaker. The amount of training data varies across speakers due to different availability of data. The same test set, based on the Harvard sentences, was withheld from training for each speaker. Table 2 also shows phone and diacritic accuracy on the test set for each speaker. The higher performance for speaker mhb1 may be attributed to the larger amount of training data, as well as the fact that it was labeled by only one person, resulting in more consistency than might be possible with multiple labelers, despite our use of a rigorous cross-checking procedure [11].

Speaker	Training phones	Phone accuracy (%)	Diacritic accuracy (%)
mhb1	23076	89.6	NA
mps1	19897	86.8	93.7
fdb1	10489	82.1	90.6

Table 2: Size of training data and test set results for each speaker. Note that no diacritics were used in the labeling of speaker mhb1.

5. RESULTS

In an effort to assess the qualitative differences among the networks trained on each speaker, beyond the simple issue of accuracy addressed in Table 1, we will examine performance on the following phonological variables: schwa, /u/, /t/ and vowel onset glottalization. These variables were chosen because they promised to demonstrate interspeaker variability due to dialect or sociolinguistic differences. It should be reiterated that some differences among speakers may be attributable to labeler idiosyncrasies.

Figure 2 presents original labeled performance in the test set of the three speakers with respect to glottalization of word-initial vowels. Speakers mps1 and mhb1 are similar in a low rate of glottalization, in contrast to speaker fdb1. It appears that dialect area does not affect realization of this variable in the data presented here. This is a good example of a case where individualized phonological analysis is important; a dialect-based rule for glottalization would not be sufficient.

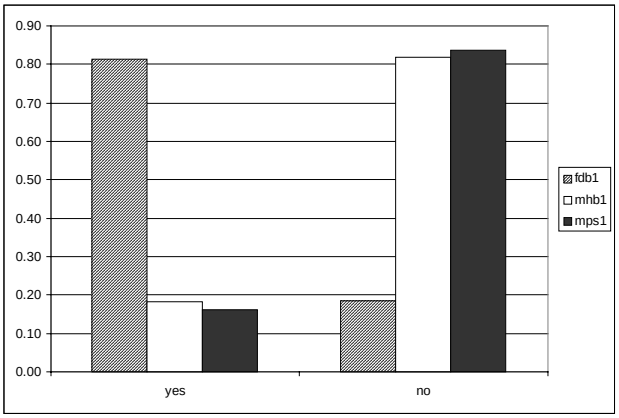

Figure 2: Word-initial vowel glottalization by speaker. "yes" indicates word-initial vowel glottalized, "no" indicates lack of glottalization.

Figure 3 presents original labeled performance in the test set of the three speakers with respect to realization of an underlying reduced vowel. In many dialects of English, reduced vowels vary in quality between [ə] and [ɪ]. In the figure, we have grouped [ɪ] and [ɨ] together, as they represent a higher/fronter reduced vowel variant than [ə]. It is not clear whether the speakers pattern by dialect area; mps1 prefers [ə], while mhb1 prefers [ɪ]/[ɨ], and fdb1 uses the two variants with equal frequency. It has been shown that realization of reduced vowels may correlate with the [coronal] specification of the surrounding consonants [12], but that is not investigated here.

Figure 4 presents original labeled performance in the test set of the three speakers with respect to realization of underlying /u/ as [u] or fronter [ʉ]. The speakers prefer the fronter variant of /u/ in the same relationship to each other as they prefer the fronter reduced vowel variant, as shown in Figure 3.

Figure 5 presents original labeled performance in the test set of the three speakers with respect to realization of underlying syllable-final /t/. The realizations were divided in an attempt to indicate increasing degrees of lenition. The most strongly articulated /t/'s included an aspirated release, the middle category included lack of release (closure only) or an unaspirated release, while the most lenited category includes flapped, glottalized, and deleted /t/'s. Given this organization of the data, it appears that mps1's /t/'s are the most strongly articulated, while mhb1's /t/'s tend much more to be lenited.

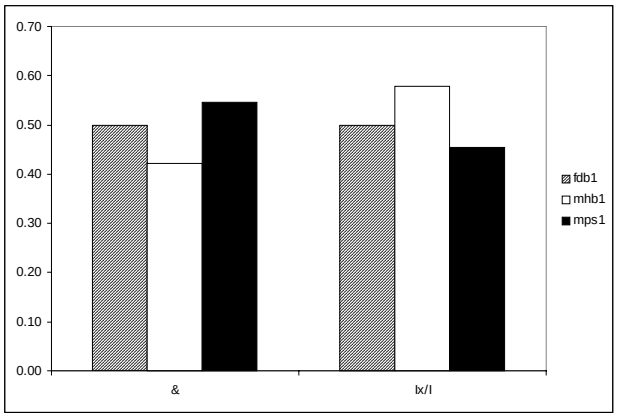

Figure 3: Reduced vowel realization by speaker. Figure labels are in Worldbet: & = [ə], Ix = [ɨ], and I = [ɪ].

Table 3 examines neural network accuracy in learning the phonological variables that have just been discussed with respect to the original speech. Accuracy on these phonological variables is lower than overall phone accuracy. This is not surprising, since there are many phones that rarely differ between lexical and postlexical levels (e.g. /ʒ/) that are included in the general accuracy score. It appears that word-initial vowel glottalization was the easiest for the network to learn, although this may be affected by the fact that there were only two postlexical options (glottalized or not), whereas the variable /t/ has several different realizations.

Speaker	Reduced vowel accuracy (%)	/u/ accuracy (%)	/t/ accuracy (%)	Glot-talization accuracy (%)
mhb1	82.8	81.5	72.7	86.4
mps1	68.8	63	56.3	83.7
fdb1	62.5	55.6	43.8	79.1

Table 3: Neural network accuracy by phonological variable.

6. CONCLUSION

We have demonstrated a practical method for learning postlexical variation on an individual basis for synthetic voices based on real people. The benefits for the text-to-speech

architecture described are twofold: a closer approximation to the material upon which the acoustic model was trained yields higher quality acoustic output, and a proper modeling of natural connected speech processes leads to speech that is likely to be considered more natural by listeners. Speech synthesis research has concentrated on using read speech as a model, yet this view may be misguided [13]. We note that much postlexical variation occurs even in read speech, as the results here suggest, however, perhaps modeling spontaneous speech will lead to further naturalness gains. Despite the fact that clear speech has been demonstrated to be more intelligible than normal speech [14], we believe that use of spontaneous speech as a model will draw attention away from *how* things are uttered to the content of *what* is being uttered, which is perhaps the most useful orientation applications employing synthetic speech can have.

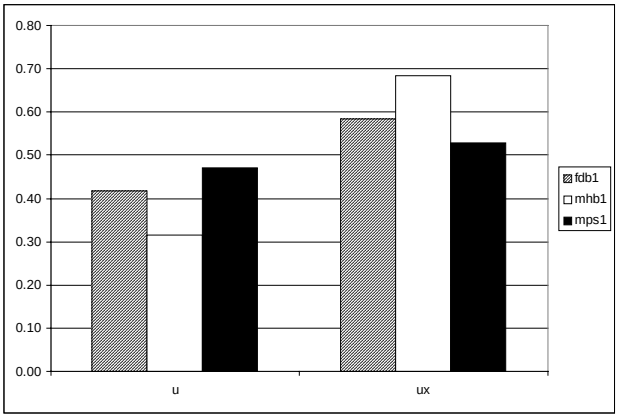

Figure 4: Realization of underlying /u/ by speaker. Figure labels are in Worldbet: ux = [ʉ].

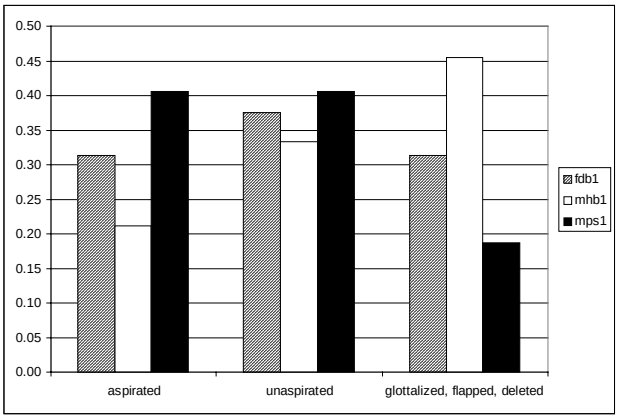

Figure 5: Realization of syllable-final underlying /t/ by speaker.

7. ACKNOWLEDGMENTS

I would like to thank our three speakers for their voices, Noel Massey and Otto Schnurr for advice on neural network architecture, Jerry Corrigan, Erica Zeinfeld, Joseph Goldberg, and Lynette Melnar for linguistic advice and labeling, and Jay Williams for useful discussions. I would also like to thank the labelers at Ohio State University for their efforts.

8. REFERENCES

1. Karaali, O., Corrigan, G., Massey, N., Miller, C., Schnurr, O., and Mackie, A. "Application of multiple neural networks for high quality text-to-speech synthesis," ICASSP Vol. 2: 1237-1240, 1998.

2. Seneff, S., and Zue, V. "Transcription and alignment of the TIMIT database," manuscript, 1988.

3. Lander, T. "CSLU labeling guide," Center for Spoken Language Understanding, Oregon Graduate Institute, 1997.

4. Miller, C. "Pronunciation modeling in speech synthesis," Institute for Research in Cognitive Science Technical Report 98-09, University of Pennsylvania, 1998.

5. Fitt, S. "The generation of regional pronunciations of English for speech synthesis," Eurospeech 97, 2447-2450, 1997.

6. Guy, G. "Variation in the group and the individual: the case of final stop deletion," in *Locating language in time and space*, ed. W. Labov, 1-36, Academic Press, New York, 1980.

7. Wells, J.C., *Accents of English* I, Cambridge University Press, Cambridge, 1982.

8. Riley, M., and Ljolje A. "Automatic generation of detailed pronunciation lexicons," in *Automatic speech and speaker recognition: advanced topics*, ed. C.-H. Lee, F.K. Soong, and K.K. Paliwal, 285-301, Kluwer, Boston, 1996.

9. Egan, J.P. "Articulation testing methods, II," OSRD Report No. 3802, 1944.

10. Hieronymus, J.L. "ASCII Phonetic Symbols for the World's Languages: Worldbet," Bell Laboratories manuscript, 1993.

11. Lander, T., Oshika, B., Cole, R.A., and Fanty, M. "Multi-language speech database: creation and phonetic labeling agreement," ICPhS 95, 166-169, 1995.

12. Clements, G. "Palatalization: linking or assimilation?," Chicago Linguistic Society 12, 96-109, 1976.

13. Laan, G.P.M. "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and read speaking style," *Speech Communication* 22: 43-65, 1997.

14. Picheny, M. A., Durlach N.I., and Braida, L.D., "Speaking clearly for the hard of hearing I: intelligibility differences between clear and conversational speech," *Journal of Speech and Hearing Research* 28: 96-103, 1985.