IMPROVED DURATION MODELING OF ENGLISH PHONEMES USING A ROOT SINUSOIDAL TRANSFORMATION

Jerome R. Bellegarda and *Kim E. A. Silverman*

Spoken Language Group
Apple Computer, Inc.
Cupertino, California 95014, USA

ABSTRACT

Over the past few years, the "sums-of-products" approach has emerged as one of the most promising avenues to account for contextual influences on phoneme duration. This approach is generally applied after log-transforming the durations. This paper presents empirical and theoretical evidence which suggests that this transformation is not optimal. A promising alternative solution is proposed, based on a root sinusoidal function. Preliminary experimental results obtained on over 50,000 phonemes in varied prosodic contexts show that this transformation reduces the unexplained deviations in the data by 32.2%.

1. INTRODUCTION

In natural speech, durations of phonetic segments strongly depend on contextual factors such as the identities of surrounding segments, stress, accent, and phrase boundaries (cf., e.g., [1]). For synthetic speech to sound natural, these duration patterns must be closely reproduced. Two approaches have been followed for duration prediction: (i) general classification techniques, such as decision trees and neural networks [2], and (ii) "sums-of-products" (SoP) methods, based on multiple linear regression in either linear or log domain [3].

These two approaches differ in two key aspects: the amount of linguistic knowledge required, and the behavior of the model in situations not encountered during training. General classification techniques are largely data-driven and unsupervised, and therefore require a large amount of training data. Furthermore, they cope with never-seen circumstances by using coarser representations, thus sacrificing resolution. In contrast, SoP models are supervised on the basis of linguistic knowledge, which makes them more robust to missing data. In addition, they predict durations for unseen contexts through interpolation, by making use of the ordered structure uncovered during analysis of the data [1]. Given the typical size of training corpora currently available, the second approach tends to outperform the first one, particularly when cross-corpus evaluation is considered [4].

When SoP models are applied in the linear domain, they lead to various derivatives of the additive model originally proposed by Klatt [5]. When they are applied in the log domain, they lead to multiplicative models such as described in [1]. The evidence appears to indicate that the latter perform better than the former. Two reasons why this might be the case are: (i) the distributions tend to be less skewed after the log transformation; and (ii) the fractional approach underlying multiplicative models is better suited for small durations. There is, however, no evidence that the log transformation is optimal. Rather than eliminating skewness in the data, it tends to merely reduce it (and reverse its sign). And while it is true that contexts such as phrase-final position are likely to lengthen long phonemes more than short phonemes, there is no *a priori* reason for all factors to be strictly multiplicative across all durations.

This paper presents empirical and theoretical evidence supporting an alternative transformation which results in better models. The next section motivates a closer look at the assumptions underlying the SoP approach. Section 3 examines the theoretical basis for an alternative formalism. In Section 4, we propose a new transformation based on a sinusoidal function. Finally, Section 5 reports on a series of experimental results illustrating the benefits of the proposed framework.

2. EMPIRICAL MOTIVATION

This work arose from evaluating the SoP approach on a large corpus collected at Apple Computer in the summer of 1996. This corpus systematically represents the known contextual factors influencing prosodic phonetic structure for a canonical speaking style. It contains all possible syllable types as defined by a comprehensive grammar based on phoneme classes. There is at least one instance of each syllable with each of no pitch accent, $\mathbf{L}$+$\mathbf{H}^*$, and $\mathbf{H}^*$, in each of prenuclear, intermediate-phrase-nuclear, or phrase-final nuclear position [6]. Furthermore, there is at least one instance of each accented syllable separated from the end of its word, the following accent, and the end of the phrase, by each of 0, 1, 2, 3, and 4 intervening syllables. In addition, all of the instances of every syllable type systematically samples from all the phonemes in each class of each of the syllable component. The corpus was spoken

Fig. 1. Effects of Adding More Regression Parameters.

by a linguistically-trained speaker, with close monitoring of the intended intonation.

In the experiments, the phonemic alphabet had size 40, and the portion of the corpus considered comprised 50,797 observations. Thus, on the average, there were about 1270 observations per phoneme. Phoneme boundaries were automatically aligned using a speaker-dependent version of the Apple large vocabulary continuous speech recognition system. The SoP approach was implemented via weighted least-squares multiple regression, as implemented in the Splus v3.2 software package. The standard log transformation was used. Across the entire dataset, this model left 15.2% of the standard deviation in the durations unexplained.[1] This overall fit is comparable to prior results reported in the literature.

Close analysis of the residuals showed that they were not spread evenly throughout the data range. Specifically, long durations tended to be underestimated and short durations overestimated. This is of course a common modeling phenomenon, which typically becomes less and less severe as the models acquire more independent variables representing higher-order interactions between contexts.

Fig. 1 illustrates this error reduction for a subset of the above data (consisting of the four unvoiced fricatives). The predicted and observed values have each been sorted in ascending order, and the two distributions plotted against each other. If the predictions were perfect, all the points would lie on the grey "$y = x$" line. Instead, the grey filled circles represent the predictions from a simple SoP model with about 20 parameters, which accounts for 32.6% of the total standard deviation. The black hollow circles represent a more complex model with about 200 parameters, which accounts for 87.2% of the deviation. The additional parameters allow the model to more closely predict the more extreme observations in the data. However, the overall shape of the plot suggests that the overestimation of short durations and underestimation of long durations is a structural pattern over a wide range of regression equations. Moreover, this observation is consistent across the entire dataset.

There are two possible (non mutually exclusive) approaches to reducing these erroneous duration predictions. The traditional approach, as illustrated in Fig. 1, is to add more independent variables to the regression equation. However, each parameter added to the more complex equation represents only one particular higher-order interaction between factors, and thus only one specific subset of the data. As more interaction terms are added, they are trained on fewer and fewer points and account for smaller and smaller particular subsets of the outliers. At the extreme, this raises the issue of parameter reliability, as well as generalization to new combinations of context.

The other approach is to first make sure the raw durations are transformed appropriately, given the structural nature of the pattern observed in the residuals. This led us to re-examine the underlying assumptions of the SoP model.

3. THEORETICAL FRAMEWORK

The origin of the SoP approach can be traced to the "axiomatic measurement" theorem [7], as applied to duration data. This theorem states that under certain conditions the duration function D can be described by the generalized additive model, given by:

$$F\left[D(f_1, f_2, \ldots f_N)\right] = \sum_{i=1}^{N} \prod_{j=1}^{M_i} a_{i,j} f_i(j)\,, \qquad (1)$$

where f_i $(i = 1, \ldots, N)$ represents the ith contextual factor influencing D, M_i is the number of values that f_i can take, $a_{i,j}$ is the factor scale corresponding to the jth value of factor f_i, denoted by $f_i(j)$, and F is an *unknown* monotonically increasing transformation. Thus, $F(x) = x$ corresponds to the additive case and $F(x) = \log(x)$ corresponds to the multiplicative case. As mentioned before, $F(x) = \log(x)$ is normally used.

The conditions mentioned above have to do with factor independence. Specifically, one can construct a function F and a set of factor scales $a_{i,j}$ such that (1) holds *only if* the factors f_j, $j = 1, \ldots, N$, exhibit all possible forms of independence, i.e., *only if* joint independence holds for all subsets of $2, 3, \ldots, N$ factors. Clearly, this is not going to be the case for duration data. For example, accent and phrasal position interact in their influence on vowel duration, i.e., these factors are not independent. The justification for applying (1) anyway is, generally, that such interactions tend to be well-behaved, in that their effects are amplificatory, rather than reversed or otherwise permuted [1]. The "regular patterns of amplificatory interactions," in van Santen's words, make it "quite plausible that *some* sums-of-products model will fit the [appropriately transformed] durations" [1] (emphasis ours).

[1] In this paper we report the fit on the complete corpus, rather than setting aside a test subset. In our experiments we have found the same patterns as those reported here, when we evaluate the models with a train/test subdivision of the data.

Fig. 2. Transformation Shape for Various α.

Fig. 3. S

Violation of the joint independence assumption, however, may substantially complicate the search for the transformation F. In particular, the optimal transformation F may no longer be strictly increasing, opening up the possibility of inflection points, or even discontinuities. In other words, it is worth revisiting the likely behavior of the transformation in the face of amplificatory interactions.

4. NEW TRANSFORMATION

For simplicity, in the generalized additive model (1) we use a common set of factors across 15 classes of phonemes. This common set includes well known factors such as accent, preceding and following phoneme identity, and others reported in the literature. The data of Fig. 1 suggests that the interactions mentioned above are only amplificatory for long durations. When durations are short, the interactions seem to exert the opposite influence.

As a result, we opted to look for a transformation F with opposite properties at the two ends of the range. In the first approximation, this entails at least one inflection point in F. This observation led us to consider the sinusoidal function:

$$F(x) = \left\{ \sin\left[\frac{\pi}{2} \left(\frac{x-A}{B-A} \right)^{\alpha} \right] \right\}^{2+\beta}, \qquad (2)$$

where A and B denote the minimum and maximum durations observed in the training data, and the parameters α and β control the shape of the transformation. Specifically, these parameters control (i) the position of the inflection point within the range of durations observed, and (ii) the amount of shrinking/expansion which happens on either side.

Fig. 2 and 3 depict the shape of the function (2) for various values of α and β. It can be seen from Fig. 2 that with values $\alpha < 1$, the curve moves to the left, which leads to an expansion of the shorter durations and a compression of the longer durations. On the other hand, with values $\alpha > 1$ the curve moves to the right, which means the shorter durations shrink and the longer durations become more separated. Furthermore, the two parameters can be independently adjusted to also control the slope of the function at the inflection point. As Fig. 3 illustrates, this slope can be reduced by using a relatively large value of α and a relatively small value of β, or increased by using the opposite combination.

From our data, it also seemed that the residuals are disproportionately greater in long durations than in short durations (cf. Fig. 1). Thus, relatively speaking, the transformation should impact long durations more than short durations. It it important to note, however, that the optimal values of the parameters α and β depend on the phoneme (or class) identity, since the shape of the function is tied to the duration distributions observed in the training data.

In the experiments described below, the procedure we followed to generate these parameters was to iteratively adjust α and β for each phoneme class, using the goodness of fit of the subsequent regression as the criterion. It would be straightforward to automate this procedure using, e.g., standard gradient descent algorithms. As it turns out, we have found that the values $\alpha = 0.8$ and $\beta = 0$ are adequate for a wide range of phonemes/classes. For this reason, we call the resulting transformation the *root sinusoidal* transformation.

5. EXPERIMENTAL RESULTS

The baseline result (15.2% unexplained) was obtained using the standard multiplicative model, as described in Section 2. The same independent variables were then regressed against the root sinusoidal transformation of the raw durations. In both cases, the SoP coefficients (after the appropriate transformation) were estimated using weighted least squares as implemented in the Splus v3.2 software package.

Applying the root sinusoidal transformation left only 10.3% of the standard deviation unexplained, which corresponds

Fig. 4. Performance Comparison.

to a reduction of 32.2% in the proportion not accounted for by the model.

The above experiments were then repeated with a range of different numbers of equation parameters, representing different choices of factors and interaction terms, to see if the result was somehow linked to the particular regression model selected. Fig. 4 reports the outcome, in terms of the percentage of standard deviation explained as a function of the total number of parameters in the modeling (including the parameters required for the transformation). It can be seen that the root sinusoidal transformation (filled triangles) is consistently superior to the log transformation (hollow circles) across the entire range of parameters considered.

A consequence of Fig. 4 is that the root sinusoidal transformation provides for a more parsimonious representation of the regular patterns in the observed data. Specifically, for a given level of performance, the root sinusoidal approach allows the underlying SoP expression to comprise approximately half the number of parameters. For example, to explain 85% of the standard deviation in the durations would require less than 2500 parameters with a root sinusoidal transformation, but slightly more than 4500 parameters with a log transformation.

6. CONCLUSIONS

This paper has presented both theoretical and preliminary empirical evidence for the use of a root sinusoidal transformation in the well-known sums-of-products approach to duration modeling. Compared to the standard log transformation, this new transformation reduced the proportion of the standard deviation unexplained by more than 30%. Alternatively, for a given level of performance, the new transformation roughly halved the required number of equation parameters.

This improved duration model has implications for the voice generation in a speech synthesizer, because of the greater quantity of both shorter and longer phonemes that it is able to generate. Short phonemes are difficult to synthesize because they are typically associated with undershoot of articulatory targets. Mere warping (in the time domain) of units that sound appropriate with longer durations is likely to result in unnaturally sudden spectral transitions. Similarly, the longer durations produced by this model will require careful voice processing to avoid unnaturally salient steady states. Consequently, we believe that as duration models improve, there will be greater need for articulatory approaches to voice generation.

In future work, the parameters of the transformation will be automatically optimized, and the different transformations will be compared by calculating the unexplained deviations in the raw data rather than in the transformed domain.

7. ACKNOWLEDGEMENTS

We thank Kevin Lenzo and Victoria Anderson for their care, skill, and infinite patience in the creation of the corpus. We are also grateful to Devang Naik for generating and verifying the phoneme boundaries.

8. REFERENCES

[1] J.P.H. van Santen, "Assignment of Segmental Duration in Text-to-Speech Synthesis," *Computer Speech and Language*, 1994.

[2] M.D. Riley, "Tree-based Modeling for Speech Synthesis," in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoit, and T.R. Sawallis, Eds, Amsterdam: Elsevier, pp 265–273, 1992.

[3] J.P.H. van Santen, "Contextual Effects on Vowel Duration," *Speech Communication*, Vol. 11, pp. 513–546, 1992.

[4] A. Magbouleh, "An Empirical Comparison of Automatic Decision Tree and Linear Regression Models for Vowel Durations," in *Proc.* 1996 *Ann. Meet. ACM*, Santa Cruz, CA, 1996.

[5] D.H. Klatt, "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," *J. Acoust. Soc. Amer.*, Vol. 59, pp. 1209–1221, 1976.

[6] K.E.A. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J.B. Pierrehumbert, J. Hirschberg, "TOBI: A Standard for Labelling English Prosody," in *Proc.* 1992 *International Conference on Spoken Language Processing*, Banff, Canada, 1992.

[7] D.H. Krantz, R.D. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement, Vol. I*, New York: Wiley, 1971.