USING A SIGMOID TRANSFORMATION FOR IMPROVED MODELING OF PHONEME DURATION

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

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 model contextual influences on phoneme duration. The associated regression 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 sigmoid 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 more than 30%. Alternatively, for a given level of performance, it halves the number of parameters required by the model.

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].

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 more extreme 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 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.

As a result, alternative transformations may result in better models, as we first showed in [6]. Following this work, this paper presents empirical and theoretical evidence supporting a sigmoid-based formalism. The next section motivates a closer look at the assumptions underlying the SoP approach, and Section 3 examines the theoretical basis for an alternative solution. In Section 4, we describe the sigmoid transformation. Finally, Section 5 reports on a series of experiments illustrating the resulting benefits.

2. EMPIRICAL MOTIVATION

This work arose from the evaluation of 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. 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 [7]. 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. All instances of every syllable type systematically sample from all the phonemes in each class of each of the syllable onsets, nuclei, and codas. The corpus was spoken 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

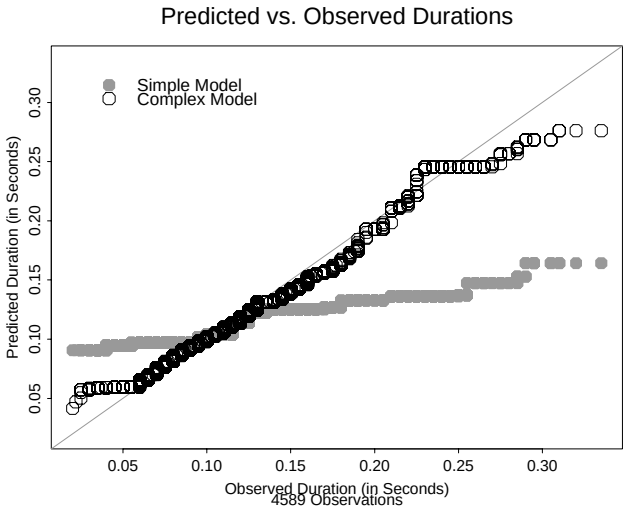

Fig. 1. Effects of Adding More Regression Parameters.

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. One distinct model was computed for each of 15 classes of phonemes, across which, for simplicity, we used a common set of factors. These included accent, preceding and following phoneme identity, and similarly well-known factors reported in the literature. In all cases, the standard log transformation was used.

Across the entire dataset, this approach left 15.2% of the standard deviation in the durations unexplained.[1] This overall fit is comparable to published results. 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. 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. If the predictions were perfect, all the points would lie on the dotted grey diagonal line. Instead, the overall shape of both sets of predictions 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 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 apply an appropriate transformation to the raw durations, to compensate as much as possible for 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 [8], 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 "certain conditions" mentioned above have to do with factor independence. Specifically, one may only construct a function F and a set of factor scales $a_{i,j}$ such that (1) holds *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*

[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.

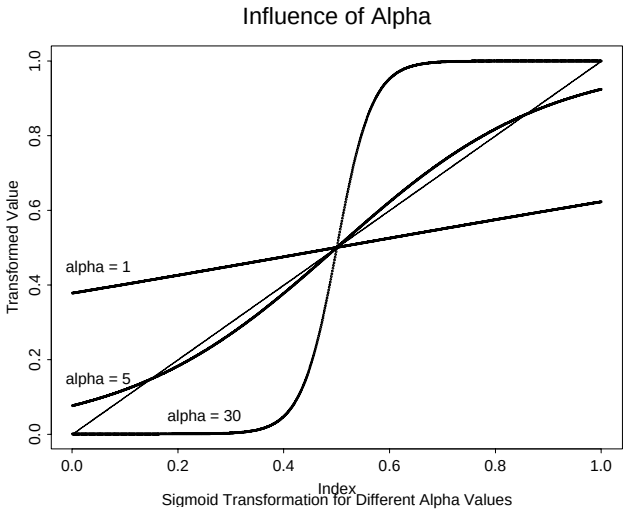

Fig. 2. Transformation Shape for Various α.

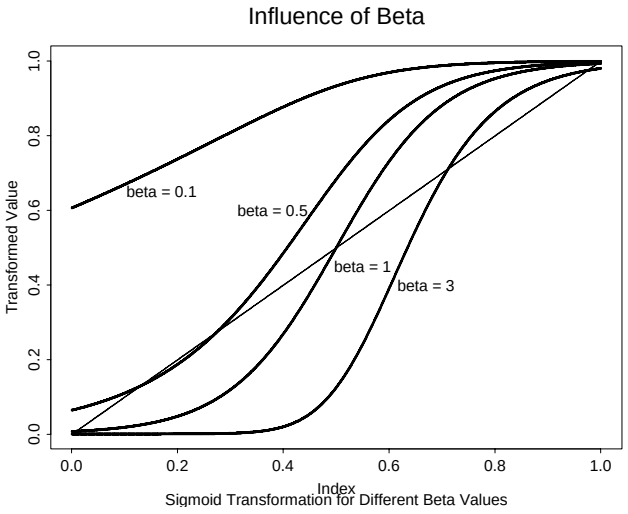

Fig. 3. Transformation Shape for Various β.

sums-of-products model will fit the [appropriately transformed] durations" [1] (emphasis ours).

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 what shape the transformation should have in the face of all interactions, amplificatory or otherwise.

4. NEW TRANSFORMATION

In fact, the data of Fig. 1 suggests that some interactions are only amplificatory for long durations (when durations are short, these interactions seem to exert the opposite influence). As a result, we opted to look for a transformation F with opposite behavior at the two ends of the range. In the first approximation, this entails at least one inflection point in F. This observation first led us to consider a sinusoidal function [6]. But the parameters in this function turned out to be somewhat non-intuitive, which called for an alternative formulation [9].

We then focused on a more conventional sigmoid function, of the type widely used in neural networks:

$$F(x) = \left[1 + \exp\left\{ -\alpha \left(\frac{x - A}{B - A} - \frac{1}{2}\right)\right\}\right]^{-\beta}, \qquad (2)$$

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, α controls the slope of the curve at the inflection point, and β controls the position of the inflection point within the range of durations observed.

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 increasing α makes the curve steeper, which means durations at the extreme of the range become comparatively more compressed than durations in the middle of the range. And, as Fig. 3 illustrates, larger values of β moves the curve moves to the right, which leads to a comparatively greater compression of the shorter durations, while smaller values of β moves the curve to the left, which affects the longer durations comparatively more. Not surprisingly, in the lower limit, the shape becomes somewhat logarithm-like.

Referring back to Fig. 1, it seems that, regardless of model complexity, the residuals for unvoiced fricatives are disproportionately greater in long durations than in short durations. Thus, we would expect the associated transformation to impact long durations more than short durations. From the above, this points to a value of β less than one, which was confirmed experimentally (we found an optimal value of $\beta = 0.7$ for this phoneme class). 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 way contextual factors influence the durations of particular phonemes.

In the experiments described below, we used a gradient descent algorithm to iteratively adjust α and β for each phoneme class, using the goodness of fit of the subsequent regression as the criterion. We have found that the values $4 \leq \alpha \leq 8$ and $0.7 \leq \beta \leq 1.5$ are adequate for a wide range of classes, which entails that most of the associated shapes (for 13 out of the 15 classes) are markedly different from a logarithm curve.

5. EXPERIMENTAL RESULTS

The baseline result (15.2% of the standard deviation unexplained) was obtained using the standard logarithmic transformation of the raw durations (multiplicative model), as described in Section 2. The same independent variables were then regressed against the sigmoid-transformed durations, using the same weighted least squares implementation.

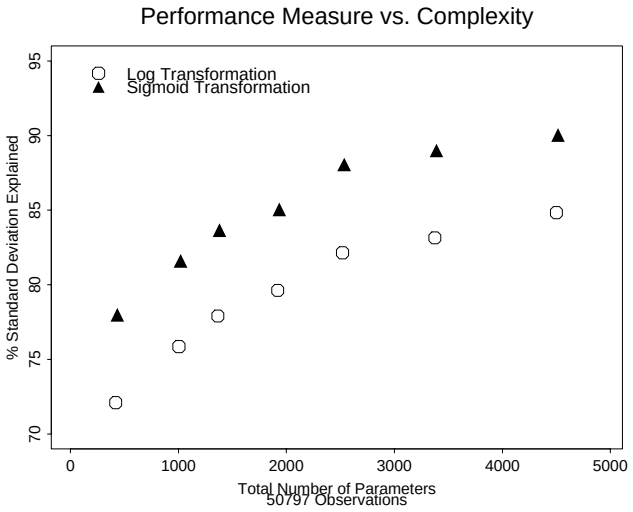

Fig. 4. Performance Comparison.

The sigmoid formalism left only 9.9% of the standard deviation unexplained, which corresponds to a reduction of 34.9% in the proportion not accounted for by the model. Note that this is slightly—albeit significantly—better than the reduction of 32.2% achieved with the root sinusoidal transformation described in [6].

The above experiments were then repeated with a range of different numbers of equation parameters, representing different choices of factors and interaction terms, to eliminate the possibility that the above result might somehow be 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 sigmoid 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 sigmoid transformation provides for a more parsimonious representation of the regular patterns in the observed data. Specifically, for a given level of performance, the sigmoid approach allows the underlying SoP expression to dispense with approximately half the number of parameters. For example, to explain 85% of the standard deviation in the durations would require slightly more than 4500 parameters with the log transformation, but only about 2000 parameters with the sigmoid transformation.

6. CONCLUSIONS

This paper has presented both theoretical and preliminary empirical evidence for the use of a sigmoid transformation in the well-known sums-of-products approach to duration modeling. Compared to the standard log transformation, the sigmoid function reduced the proportion of the standard deviation unexplained by about 35%, which improves slightly on the results of [6]. For a given level of performance, the sigmoid transformation more than halved the number of parameters required in the duration model.

This improved modeling 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.

7. ACKNOWLEDGEMENTS

We would like to thank Kevin Lenzo and Victoria Anderson for their care, skill, and infinite patience in the creation of the corpus used in the experiments. 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] J.R. Bellegarda and K.E.A. Silverman, "Improved Duration Modeling of English Phonemes Using a Root Sinusoidal Transformation," in *Proc.* 1998 *Int. Conf. Spoken Language Proc.*, Sydney, Australia, December 1998.

[7] 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 *Int. Conf. Spoken Language Proc.*, Banff, Canada, 1992.

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

[9] K. Lenzo, personal communication, August 1998.