Predicting segment duration using sums of products model

Olga Goubanova
olga@ling.ed.ac.uk

Centre for Speech Technology Research
Edinburgh University

ABSTRACT

Database imbalance and factor interaction are one of the majour factors that make segment duration modelling in text-to-speech systems a challenging task. We present the results of segment duration modelling in a text-to-speech system based on sums of products approach. We contrasted the results of sums of products model with those of the classification and regression tree (CART) model. We trained and tested the two models on the same data. The results have shown that sums of products model better predicts vowel duration compared to the CART approach (median correlation of 0.94 vs. 0.82 and RMSE of 7 ms vs. 20 ms).

1. Introduction

Segment duration is influenced by a number of contextual factors such as segment identity, stress, accent, identity of preceding and following segments, position of a target segment within a syllable, word, and utterance. If a machine learning approach is taken, a database is used to infer the parameters of the algorithm making the duration prediction. This presents a number of problems for current text-to-speech systems. In general, databases that are used to model segment duration suffer from data imbalance problem. On the one hand, only a small and uneven fraction of linguistically allowed factor combinations is present in a training database; different factor combinations occur with unequal frequencies. Yet as was shown by [4] rare combinations of factor levels occur quite often in any given text. Furthermore, factors affecting segment duration interact; a set of two or more factors may amplify or attenuate the affect of other factors. Nevertheless, the properties of duration data are such that "the interactions are often quite regular" [4].

Previous researchers applied various computational techniques to segment duration modelling from rule-based [1], to statistical (classification and regression trees [7]), to supervised data-driven approaches (the sums-of-products, or SoP duration model by [4]). In the rule-based model by [1] a segment duration was modified by applying a set of rules that described contextual effects influencing a segment's inherent duration. These rules were tailored to fit data the best; data sparsity problem was not considered in the model. Likewise, the CART approach by [8] applied to modelling of segment duration, underperformed when the percent of missing data was too high [4]. It also responded badly to noise in the data.

Among the above mentioned models, the SoP model of segment duration accounts for data imbalance and factor interaction problems the best. It is an example of a general linear model whereby segment duration is represented as a sum of factors' product terms that effect segment duration. In the SoP model by [2] segment duration was modeled as a log-transformation of factor terms. Other researches reported to have successfully applied root sinusoidal transformation [6] to modelling durational data. Furthermore, SoP model had also been applied to model segment duration of languages other than English, e.g., Japanese [5].

2. Sums of products model

2.1. Definitions

Suppose $\mathcal{F}$ is a set of factors $\{F_1, F_2, \dots F_n\}$ that affect segment duration. The feature space is represented by a cartesian product $F = F_1 \times F_2 \dots \times F_n$ of the factors. The sums of products model ϕ is a mapping from a space of all possible combinations of factor levels represented by a cartesian product to the space of real niumbers $\Re$ such that:

$$\phi \to \Re : \phi(F) = \sum_{i \in K} \prod_{i \in I_i} S_{ij}(f_j) \qquad (1)$$

where K is a set of intergers; I_i are subsets of K (index sets), and $S_{ij}(f_j)$ is a factor scale for a factor F_j having a level f_j.

Within sums of products framework, a number of assumptions are made.

- Interaction terms are products of single factor scales.
- Main effects and interaction terms do not necesseraly have zero sums.
- Index sets, corresponding to a set of factors in ith term are all unique.
- For every factor F_j having at least two levels a_j and b_j scale values $S_{ij}(a_j) \neq S_{ij}(b_j)$ for all index sets I_i that involve factor F_j (joint "non-constancy property" [3]).

2.2. Factorial Scheme

In our analysis we adopted van Santen's factorial scheme [4].
The factor set $\mathcal{F}$ consists of eight factors $\{S, A, V, Wpost, Utt, Cpre, Cpost, Wpre\}$. We describe below the meaning of each factor. Syllabic stress factor S. It has 2 levels, stressed and unstressed. Accent factor A; it represents accent status of a word with a target vowel. It has 2 levels, accented and un-accented. Vowel identity factor V. It has 15 levels according to the number of vowel types chosen for our analysis. Preceding consonant $Cpre$ identity. Based on van Santen's analysis [2], we limited number of the factor levels to two, voiced stop and other. Following consonant $Cpost$ indentity factor. The values $Cpost$ factor takes on were based on voicing and manner of production features for a consonant; voiceless stops, voiceless affricate, liquids, voiceless fricatives, nasals, voiced stops, voiced affricate, and voiced fricatives. $Wpost$ is a within word position factor; it takes on values corresponding to initial, medial, and final position of a syllable with a target vowel in a word. $Wpre$ node corresponds to the number of consonants that precede a target vowel; zero, one, and more than one. Phrasal position of a vowel Utt. It take on the values initial , medial, and final.

Given these factors, the duration of a vowel segment can be modelled as follows.

$$\begin{aligned} log(DUR) = {} & S_{1,1}(Stress) \times S_{1,2}(A) + S_{2,3}(V) + \\ & S_{3,4}(Cpre) + S_{5,6}(Wpre) + S_{6,7}(Wpost) + S_{4,5}(Cpost) + S_{7,5}(Cpost) \times S_{7,8}(Utt) \end{aligned} \tag{2}$$

where $S_{i,j}$ are scales corresponding to van Santen's factors.

3. Implementation details

3.1. Speech database

Speech databases consisted of phonetically rich sentences recorded from an American male speaker of English. We used two sets of data corresponding to two types of speaking styles, read speech and news commentary. The former consisted of 452 (22 minutes) isolated sentences (TIMIT); for our analysis, we selected 3900 vowels. The latter consisted of over 200 excerpts (18 minutes) from broadcast news; we chose 6102 vowels for our analysis. Each database was divided into training (90%) and test sets (10%). Each segment in the database was marked with segment and syllable-level phonetic information. The databases were also labeled with word boundaries, lexical stress, and word-level accent information. We conducted our experiments on each database separately, so as to exclude from the analysis the effects of speaking rate and speaking style variation.

3.2. Model's parameters estimation

We estimated the parameters of the SoP model in z-scores domain. The absolute values of duration were transformed to z-scores, the model's parameters were estimated, then backward transform was performed, with predicting duration values of a test set based on a probability distribution function $\mathcal{N}(\mu, \Sigma)$ of each vowel class. The SoP's parameters were estimated using ordinary least squares regression with dummy variable coding for the factor levels.

4. Results and their discussion

In our experiments, we compared the results of the SoP model prediction with those of the CART model. We estimated the performance of the models based on the values of a correlation coefficient and a root mean square (RMS) error. Figures 1 and 2 show the correlation values predicted from the SoP and CART models for TIMIT and sole data bases. As can be seen from the figures, the SoP model outperforms the CART model. For sole database, the median correlation values are 0.91 and 0.71 for the SoP and CART models respectively. For TIMIT database, the median correlation of 0.94 compares against 0.82 for the CART model.

It can be concluded that the SoP model predicts durations better that the CART model. For TIMIT database both models give quite uniform correlation values; for sole database correlation values are more spread. One possible explanation would be that TIMIT is phonetically-balanced database created to fully cover all possible phonetic contexts; sole database comprises the excerpts from news commentary and is thought to be less balanced phonetically.

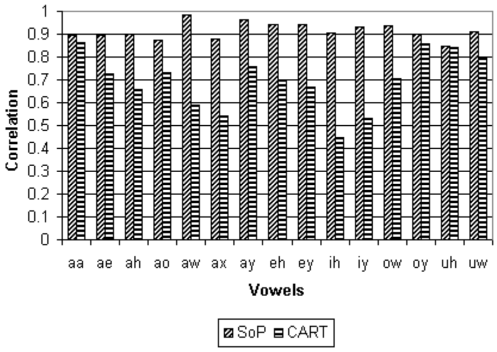

Figure 1: Correlation values predicted from the SoP and CART models. TIMIT database; test data set of 390 vowels.

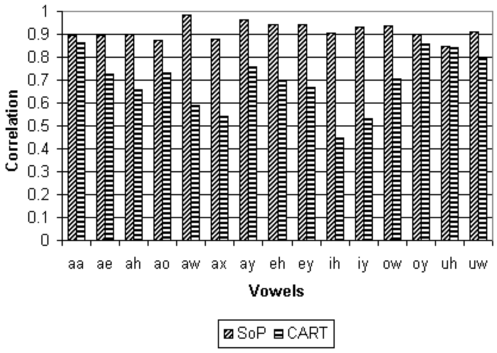

Figure 2: Correlation values predicted from the SoP and CART models. sole database; test data set of 608 vowels.

Figures 3 and 4 show the RMS error values (in milliseconds) predicted from the SoP and CART models for TIMIT and sole data bases. As can be seen from the figures, the SoP model predicts vowel durations with much lower RMS error compared to the CART model. From analising RMS error values, it can be concluded that the SoP model performs significantly better than the CART model. For example, for TIMIT database the median RMS errors are 9 ms and 20 ms for the SoP and CART models respectively.

5. Conclusions

We showed that sums-of-products approach to modelling segment duration produces promising results in terms of RMS error and correlation values. The results are better than those produced by the CART model. Across the vowel classes, the SoP model gives the median RMS error value of 9 msec and the correlation value of 0.94; the corresponding values are 20 msec and 0.82 for the CART models (TIMIT database). One of the advantages of the SoP model is its stable behaviour when there is a noise in the data or a substantial amount of data are missing. The reason for the SoP model's good performance is that it uses interpolation for predicting noisy or never seen feature vectors. The CART model instead uses pooling therefore making rather crude estimations. As a continuation of the current research, we plan to apply the SoP model to predicting consonants' durations.

6. REFERENCES

1. Klatt, D. H. Linguistic uses of segmental duration of English: Acoustic and perceptual evidence. Journal of the Acoustic Society of America, 59, 1209-1211
2. Van Santen, J.P.H. Contextual effects on vowel durations. Speech Communication, 11, 513-546, 1992
3. Van Santen, J.P.H. Analising N-way tables with sums-ofproducts models. Jounal of Mathematical Psychology, 37, 327-371, 1993
4. Van Santen, J.P.H. (1994). Assignment of segmental duration in text-to-speech synthesis. Computer Speech and Language, 8, 95-128
5. J. J. Venditti and Jan P. H. van Santen. "Modeling vowel durations for Japanese text-to-speech synthesis", in CD-ROM Proc. ICSLP 98, 1998

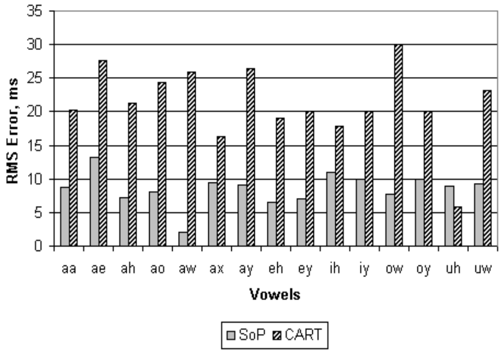

Figure 3: The RMS error values of predicted durations estimated from the SoP and CART models. TIMIT test data set of 390 vowels.

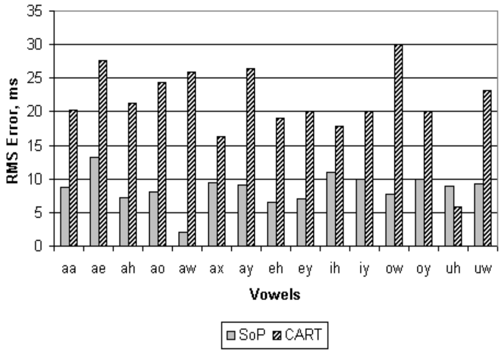

Figure 4: Te RMS error values of predicted durations estimated from the SoP and CART models. Sole test data set of 608 vowels.

6. Jerome R. Bellagarda and Kim E.A. Silverman. "Improved duration modeling of English phonemes using root sinusoidal transformation", in CD-ROM Proc. ICSLP 98, 1998
7. A. W. Black, P. Taylor, and R. Caley. The Festival Speech Synthesis System: system documentation. The Centre for Speech Technology Research, University of Edinburgh, 1.4.0 edition, 2000. http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html
8. L. Breiman, J. Friedman, and R. Olshen. Classification and Regression Trees. Wadsworth and Brooks, Pacific Grove, CA, 1984.