Using Bayesian Belief Networks for model duration in text-to-speech systems

Olga Goubanova
olga@ling.ed.ac.uk

Centre for Speech Technology Research,
University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK

[ ]

ABSTRACT

We present a new Bayesian-based probabilistic approach to modelling segmental duration in a text-to-speech system. Segment duration is influenced by a number of contextual factors such as segment identity, stress, accent, local context within a syllable, position of a target segment within a syllable, word, and utterance. The factors that affect segmental duration interact with each other in a complex way. Databases of speech data are often imbalanced with respect to frequencies of different factors' combinations. A model of segment duration should account for these problems. We propose a probabilistic Bayesian belief network (BN) approach to tackle data sparsity and factor interaction problems. In our work, we model segment duration as a hybrid Bayesian network consisting of discrete and continuous nodes; each node in the network represents a linguistic factor that affects segmental duration.

The interaction between the factors is represented as conditional dependence relations in the graphical model. For the purposes of the current research we used a database of over 1000 prosodically rich sentences from the speakers of American and RP English. We contrasted the results of the BN model with those of the sums of products (SoP) model by van Santen and the CART model implemented in the Festival text-to-speech system. We trained and tested all three models on the same data. The results have shown that our new model outperforms the CART model; it compares in performance with the SoP model. However, we think our model has many other advantages compared to SoP, for instance it is much easier to configure and experiment with new features. This should make it easier to adapt to new languages.

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 [1] rare combinations of factor levels occur quite often in any given text. Furthermore, factors affecting segmental duration interact; a set of two or more factors may amplify or attenuate the affect of other factors.

Previous researchers applied various computational techniques to segmental duration modelling from rule-based [2], to statistical (classification and regression trees [3]), to supervised data-driven approaches (the Sums-of-Products, or SoP duration model by [1]). In the rule-based model by [2] 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 part of the model. Likewise, the CART approach [4] applied to modelling of segment duration, underperformed when the percent of missing data was too high [1]. 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 [5] 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 [7].

One of a few drawbacks of SoP approach is that the number of different sums-of- products models grows hyper-exponentially with the number of factors. Therefore, one has to apply some clever techniques to finding a particular SoP model that describes data the best; brute force approach of exhaustive enumeration is infeasible. In addition, when modelling segment duration with a SoP model a substantial amount of data preprocessing is required to correct for factor interaction and data imbalance. Consequently, Bayesian belief

network (BN) approach seems like a good alternative to conventional deterministic techniques of data modelling.

The structure of the paper is the following. We give a theoretical motivation behind a BN approach in section 2. We explain the details of applying BN analysis to segment duration modelling in section 3. We proceed with describing the databases used for the present research in section 4. We describe the experiments and discuss the results in section 5. We conclude with discussing future work in section 6

2. THEORETICAL MOTIVATION

These considerations lead us to try and develop an general statistical framework for data prediction in which we could take principled approaches to tackling these problems. The approach was to use Bayesian belief networks. These networks are ideal for duration modelling because the basic topology of the model is flexible, which allows the model designer to use knowledge to control which factors can be considered independent. The consequence of this is that factor interactions can be captured by indicating the causal relationships of the factors in the connectivity of the nodes in a directed acyclic (DAG) graph. This in turn allows a significant reduction in the number of parameters to be estimated.

Formally a Bayesian network is defined by a triple (G, Ω, P), where $G = (\Phi, E)$ is a directed acyclic graph with a node set Φ representing a problem domain information; E is a set of edges that describes conditional dependency relations among domain variables; Ω is a space of possible instantiations of domain variables and P is a joint probability distribution for all of the nodes of the graph G.

The most important property of a Bayesian network, called Markov property, states that each variable in a network is independent of its non-descendants given its parents. This allows to factorise the joint probability distribution P into a set of univariate conditional distributions over variables of a network. Given a set of problem domain variables $P(X_1, X_2, ..., X_N)$ the joint probability distribution P factorises like so:

$$P(X_1, X_2, ..., X_N) = \prod_{i=1}^{N} P(X_i|pa(i)) \tag{1}$$

where N is a size of a BN, $pa(i)$ is a set of parents of a node X_i.

3. DURATIONAL BAYESIAN BELIEF NETWORK MODEL

In our work, we model duration of a vowel segment as a hybrid Bayesian network consisting of discrete and continuous nodes; each node in the network represents a linguistic factor that affects segmental duration. Interactions between factors are represented as conditional dependency relations in a graphical model. Duration estimation is accomplished via learning the parameters of the Bayesian network in a "from cause to effect" fashion; given a set of causal factors that affect segment duration, we find the most probable value of duration.

For the convenience of probabilistic analysis, the node set Φ of a hybrid BN is partitioned into a set of discrete variables Δ and a set of continuous variables Γ. In case of durational BN, the set Γ consists of just one scalar node D that corresponds to the duration values of a segment. The set Δ consists of discrete variables corresponding to contextual factors that affect vowel duration, $\Delta = (V, Wpost, S, A, Utt, Cpre, Cpost, Wpre)$.

The vowel BN of size 9 is shown in Figure 1. V is a vowel identity node (it takes on 15 values according to the number of the vowel phones chosen for analysis). $Wpost$ is a within word position node; it takes on values corresponding to initial, medial, and final position of a syllable with a target vowel in a word. S is a stress node, taking on stressed and unstressed values. A is a node describing an accent status of a word; it takes on accented and unaccented values. Utt node describes phrasal position of a word with a target vowel, taking on values initial , medial, and final. $Cpre$ describes the class of preceding consonant. We limited possible values for $Cpre$ to two, voiced stop and other. $Cpost$ variable corresponds to the class of the following consonant. Values for $Cpost$ node were based on voicing and manner of production features for consonant; voiceless stops, voiceless affricate, liquids, voiceless fricatives, nasals, voiced stops, voiced affricate, and voiced fricatives. $Wpre$ node corresponds to the number of consonants that precede a target vowel; zero, one, and more than one.

According to Markov property, the joint probability distribution P over the varibles $V, Wpost, S, A, Utt, Cpre, Cpost, Wpre, D$ factorises like so:

$$\begin{aligned} P(V, Wpost, S, A, Utt, Cpre, Cpost, Wpre, D) &= P(D|V, Wpost, S, A, Utt, Cpre, Cpost, Wpre)\times \\ P(V) \times P(Wpost) \times P(S|V, Wpost) \times P(A|S) \times P(Utt) \times P(Cpre) \times P(Cpost) \times P(Wpre) & \end{aligned} \tag{2}$$

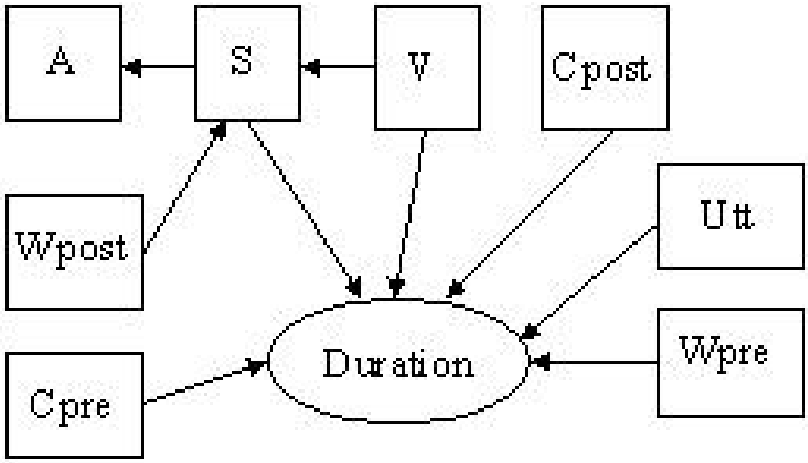

Figure 1: Duration Bayesian network of size 9; boxes represent discrete nodes, oval represents a continuous node.

The joint distribution P for a hybrid BN can be expressed as a conditional (CG) Gaussian (see [8] for details). In particular, we are interested in estimating the parameters of the conditional probability of a continuous duration node D given its parents $P(D|V, Wpost, S, A, Utt, Cpre, Cpost, Wpre)$. For every instantiation of discrete nodes $\delta \in \Delta$ the distribution over the duration node D is given:

$$p(D(\delta)|\delta \in \Delta) = \mathcal{N}(d, \mu(\delta), \Sigma(\delta)) \tag{3}$$

where $D(\delta)$ is a value of a vowel duration, $\mathcal{N}(\cdot)$ is a Gaussian pdf of the duration node D.

We estimated duration values in the following fashion. We initialised parameters of the Gaussian pdf $\mathcal{N}(\cdot)$ to prior values calculated as marginal means for every instantiation of the values of Δ in the training set. We applied EM algorithm to estimate the parameters of a junction tree, a secondary structure obtained from a BN [8]), based on the train set. Finally, we calculated the predicted values of duration for the test set.

4. SPEECH DATABASES

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.

5. EXPERIMENTAL RESULTS

One of the advantages of a BN approach is its flexibility in selecting problem domain variables and defining independence relations among these. Therefore, we can experiment with the networks of different sizes and varying connectivity. We tested the models on databases of different sizes, with sole database being twice as large as TIMIT database. BN parameter estimation was done in z-scores domain; values of duration node D were transformed to z-scores, model's parameters were estimated, then backward transform was performed, with predicting duration values of a test set based on a vowel's (μ, Σ) class. In our experiments, we compared the results of the BN model prediction with those of the SoP and CART models.

5.1. Single vowel analysis

We studied the quality of duration prediction on the networks of different sizes. We selected 5 subsets of discrete "causal" variables from the node set Δ. Table 1 shows the subsets of nodes selected for the analysis. The problem of hybrid BN structure learning is $\mathcal{NP}$- hard, therefore, we can not claim that our heuristic selection approach exhaustively selects all optimal subsets of "causal" nodes. We based our choice of factors selection upon the results reported by other researches (see for example, [1]). Initially, we performed

#	Subset	Nodes	BN Space Size
1	D V S A	4	60
2	D V Wpost S A	5	180
3	D V Wpost S A Utt	6	540
4	D V Wpost S A Utt Cpost Wpre	8	12,960
5	D V Wpost S A Utt Cpre Cpost Wpre	9	25,920

Table 1: BN's of different sizes selected for duration analysis.

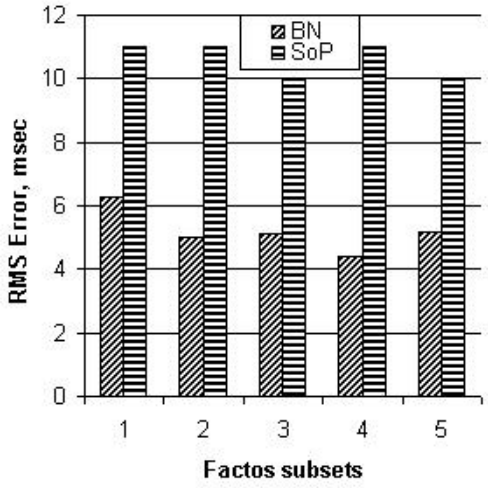

Figure 2: RMSE values of predicted durations for BN of different sizes; 378 train and 41 test /iy/ vowels, sole database. For subset of nodes see Table 1.

our analysis for a single vowel $/iy/$, extending the approach to all vowels afterwards. Figures 2 and 3 show the results of duration prediction for vowel $/iy/$. Figure 2 shows the values of RMSE for predicted duration depending on the BN complexity for sole database (378 train and 41 test segments). As can be seen from Figure 2, the RMS error ranges from 4.4 to 6.1 msec, with median RMSE value of 5 msec (compared to 11 ms as estimated from the SoP model). For TIMIT database, the RMS error of 5 msec for the BN model compares against 7 ms for the SoP model.

Figure 3 shows the values of correlation coefficient with respect to different BN structures. As can be seen from Figure 3, the correlation coefficient values change slightly with increasing BN complexity. For sole database, the median correlation of 0.94 compares against 0.91 for the SoP model); for TIMIT, the median correlation values were 0.95 and 0.94 for the BN and SoP models respectively.

5.2. All vowels analysis

One of the advantages of the BN approach is that it can be easily extended to tailor particular network architecture. In particular, we performed a BN analysis for all vowels for networks of different sizes, similar to the one we just described, by changing the number of possible values of the node V from 1 to 15. The results of duration estimation using the network of 9 nodes (see Figure 1) for sole database are shown in Figures 4 and 5. As can be seen from Figure 4, the RMSE values of duration predicted by the BN model are lower than those for the SoP and CART models, with median RMSE value of 5 msec compared to 9 msec for the SoP and 20 msec for the CART model. Figure 4 shows that for vowels $/aw/$ and $/ow/$ the SoP model gives better RMSE values (2 msec and 8 msec compared to BN's value of 12 and 13 msec for $/aw/$ and $/ow/$ vowels respectively). This effect may had to do with the fact that some local computations in the network may be non-optimal due to the uneven probability mass distribution for particular vowels.

In Figure 5 the values of correlation coefficient for sole database are shown; the median correlation of 0.94 for the BN model compares to the values of 0.9 and 0.7 for the SoP and CART models respectively. It can be concluded that on average the BN model gives a better prediction of vowel durations than the SoP and CART models.

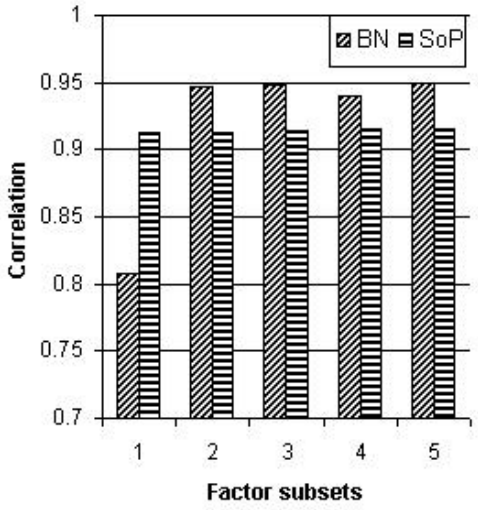

Figure 3: Correlation values of predicted durations for BN of different sizes; /iy/ vowel, sole database.

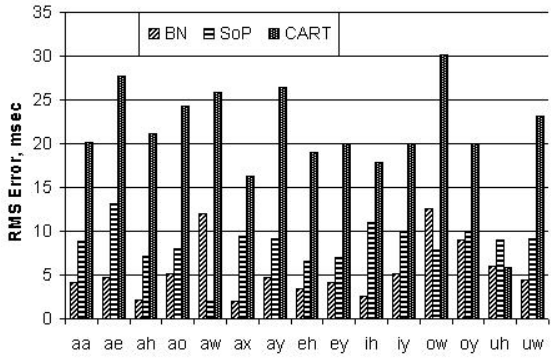

Figure 4: RMSE values of predicted durations; all vowels, sole database, BN of size 9.

6. CONCLUSIONS

We showed that Bayesian approach to modelling segment duration produces promising results in terms of RMSE and correlation values. The results are better or comparable to those produced by the SoP and CART models. Across the vowel classes, the BN model gives the median RMSE value of 5 msec and correlation value of 0.94; corresponding values are 9 msec and 0.9 for the SoP and 20 msec and 0.7 for CART models. One of the advantages of the BN approach is its flexibility; in the future, we plan to apply the BN approach to modelling duration of consonants. Given the abundance of existing speech databases, we also plan to do BN structure learning for durational networks.

7. ACKNOWLEDGEMENTS

We would like to thank Kevin Murphy ([10]) for using his Bayes Net Toolkit in our work.

8. REFERENCES

1. J. van Santen. Assignment of segmental duration in text-to-speech synthesis. Computer Speech and Language, 8, 95-128, 1994
2. D. H. Klatt. Linguistic uses of segmental duration of English: Acoustic and perceptual evidence. Journal of the Acoustic Society of America, 59, 1209-1211, 1976
3. A. W. Black, P. Taylor, and R. Caley. The Festival Speech Synthesis System: system documentation. The Centre for Speech

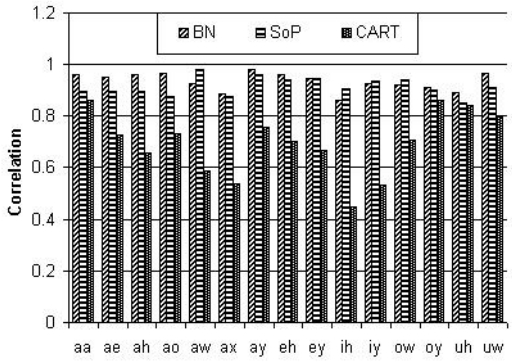

Figure 5: Correlation values of predicted durations; all vowels, sole database, BN of size 9.

Technology Research, University of Edinburgh, 1.4.0 edition, 2000.
http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html

4. L. Breiman, J. Friedman, and R. Olshen. Classification and Regression Trees. Wadsworth and Brooks, Pacific Grove, CA, 1984.
5. J. van Santen. Contextual effects on vowel durations. Speech Communication, 11, 513-546, 1992
6. J. R. Bellagarda and K. E. Silverman. Improved duration modeling of English phonemes using root sinusoidal transformationi. In CD-ROM Proc. ICSLP 98, 1998
7. J. J. Venditti and Jan van Santen. Modeling vowel durations for Japanese text-to-speech synthesis, in CD-ROM Proc. ICSLP 98, 1998
8. S. Lauritzen. Graphicl models. Oxford University Press. 1996.
8. C. Huang and A. Darwiche. Inference in Belief Networks: A procedural guide.
10. K. Murphy. Bayes Net Toolkit.
http://www.cs.berkeley.edu/m̃urphyk/Bayes/bnt.html