

Predicting segmental duration using Bayesian belief networks

Olga Goubanova

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

olga@cstr.ed.ac.uk

Abstract

Modelling segment duration in text-to-speech systems is hindered by the database imbalance and factor interaction problems. We propose a probabilistic Bayesian belief network (BN) approach to overcome data sparsity and factor interaction problems. The belief network approach makes good estimations in cases of missed or incomplete data. Also, it captures factor interaction in a concise way of causal relationships among the nodes in a directed acyclic (DAG) graph. Furthermore, a belief network approach allows a significant reduction of the number of parameters to be estimated. 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. We contrasted the results of belief network model with those of sums of products model and classification and regression tree (CART) model. We trained and tested all three models on the same data. Our BN model of vowels performs better than the SoP model: the belief network achieves a RMS error of 3 milliseconds compared with 7 ms from SoP. The CART model also produces an error of 3 ms, and hence our new model isn't any worse in terms of final performance. The BN model for consonants also produces promising RMS error values; the BN gives a value of 2 milliseconds versus 4 ms for SoP and 1 ms for the CART. The consonant BN architecture is not optimal in terms of correlation values; a search for better model will be done in the future. 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

This work is a continuation of our previous efforts to model segment duration using Bayesian belief network approach [1]. Segment duration is influenced by a number of linguistic factors such as segment identity, stress level of a syllable with a target segment, accent of a word the syllable is a part of, identity of preceding and following segments, position of a target segment within a syllable, word, and utterance. In machine learning approach to segment modelling, databases are used to calculate the parameters of a durational model.

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. Furthermore, factors affecting segmental duration interact; a set of two or more factors may amplify or attenuate the affect of other factors.

Previous researchers tried to overcome the above problems by applying different techniques ranging from rule-based [2], to statistical (classification and regression trees [3]), to supervised data-driven approaches (the Sums-of-Products, or SoP duration model [4]-[5]).

The CART durational model was the one successfully implemented in many research text-to-speech systems [6]. Such a popularity was due to the ease of building a model and its satisfactory prediction power. Still the CART model underperformed when the percent of missing data was too high. It also responded badly to noise in the data.

Another approach, namely sums-of-product model [5], received a lot of attention over the past decade. 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 [4] segment duration was modeled as a log-transformation of linguistic factors. In the SoP model the problems of data imbalance and factor interaction are thoroughly addressed. However, this is done at the expense of substantial data preprocessing. In addition, the number of different sums-of-products models grows hyper-exponentially with the number of factors. Therefore, one should apply some heuristics search techniques to find a model that fits data the best.

Therefore, we put forward a Bayesian belief network (BN) approach as an 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. Bayesian belief networks theory

We apply Bayesian belief network approach to modelling segment duration as an general statistical framework for data prediction in which we could take principled approaches to tackling data imbalance and factor interaction problems. We believe a BN approach is well suited for duration modelling because the basic topology of the model is flexible, which allows a model designer to use a problem domain 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 $P(X_1, X_2, \dots, X_N)$ the joint probability distribution P factorises like so:

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

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 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 varies according to whether the segment is a vowel or a consonant; it also varies depending on what "causal" factors are selected for analysis.

3.1. BN model of vowels

In case of the vowel duration model, the set of discrete nodes Δ consists of discrete variables corresponding to contextual factors that affect vowel duration, $\Delta = (V, W_{post}, S, Utt, C_{pre}, C_{post}, W_{pre})$. The vowel BN of size 8 is shown in Figure 1. V is a vowel identity node; it takes on 20 values according to the number of the vowel phones chosen for analysis. W_{post} 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. Utt node describes phrasal position of a word with a target vowel, taking on values initial, medial, and final. C_{pre} describes the class of preceding segment. We limited possible values for C_{pre} to two, voiced stop and other. C_{post} variable corresponds to the class of the following segment. When the following segment is a consonant, the values for C_{post} node are based on voicing and manner of production features for consonant; it takes on values voiceless stops, voiceless affricate, liquids, voiceless fricatives, nasals, voiced stops, voiced affricate, and voiced fricatives. In addition, C_{post} node can take on values vowel and silence. W_{pre} node corresponds to the number of syllables that precede a target vowel; zero, one, and more than one. According to Markov property, the joint probability distribution P over the variables

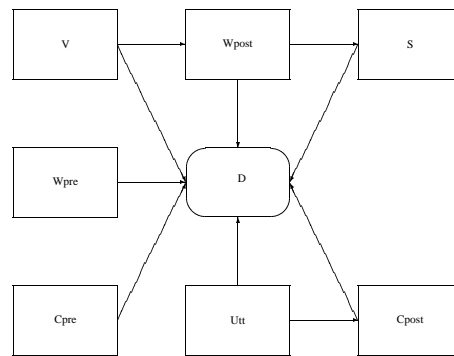


Figure 1: Vowel duration Bayesian network of size 8; boxes represent discrete nodes, oval represents a continuous node.

$V, W_{post}, S, Utt, C_{pre}, C_{post}, W_{pre}, D$ factorises like so:

$$P(V, W_{post}, S, Utt, C_{pre}, C_{post}, W_{pre}, D) = P(D|V, W_{post}, S, Utt, C_{pre}, C_{post}, W_{pre}) \times P(V) \times P(W_{post}) \times P(S|V, W_{post}) \times P(Utt) \times P(C_{pre}) \times P(C_{post}) \times P(W_{pre}) \quad (2)$$

The joint distribution P for a hybrid BN can be expressed as a conditional (CG) Gaussian (see [7] for details). In particular, we are interested in estimating the parameters of the conditional probability of a continuous duration node D given its parents $P(D|V, W_{post}, S, Utt, C_{pre}, C_{post}, W_{pre})$. For every instantiation of discrete nodes $\delta \in \Delta$ the distribution over the duration node D is given:

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

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 BN based on the train set. Finally, we calculated the predicted values of duration for the test set.

3.2. BN model of consonants

For the consonant model, the variables set G consists of a continuous node D corresponding to consonant duration values, and a set of discrete "causal" nodes Δ corresponding to the contextual factors that affect consonant duration, $\Delta = (C, W_{post}, S, NSyls, VFront)$. Our choice of discrete nodes was based on the work done by [8] on the factor interaction affecting consonant duration. It should be stressed again, this choice of factors is not necessarily the optimal solution in our Bayesian analysis. The BN selected for the analysis is shown in Figure 2. A node C corresponds to a consonant identity factor; it takes on 24 values. W_{post} node corresponds to a consonant position within a word, taking on initial, medial, and final values. S node represents a stress value (stressed or unstressed) for a syllabic vowel. $NSyls$ is a node describing a number of syllables in a word with a target consonant. It can take on values

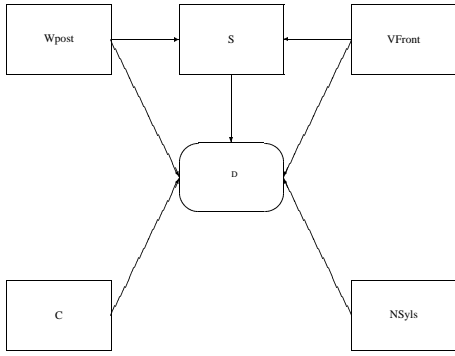


Figure 2: Consonant duration Bayesian network of size 6; boxes represent discrete nodes, oval represents a continuous node.

corresponding to a word with one, two, three, four or more than four syllables. $VFront$ is a variable describing the frontedness of a syllabic vowel, taking on front, middle, and back values.

The joint distribution for the consonant BN factorises like so:

$$\begin{aligned}
 P(C, Wpost, S, NSyls, VFront, D) = & \\
 & P(D|C, Wpost, S, NSyls, VFront) \times \\
 & P(C) \times P(Wpost) \times P(S|VFront, Wpost) \quad (4) \\
 & \times P(NSyls) \times P(VFront) \quad (5)
 \end{aligned}$$

We are interested in estimating the parameters of the conditional probability of the continuous duration node D given its parents $P(D|C, Wpost, S, NSyls, VFront)$ which is given by Equation 3. The parameter estimation for the consonant BN was done similarly to the parameter estimation for the vowel BN.

4. Databases

For our research we used a database of a RP male speaker of English (rjs). We used only a subset of rjs database that contained 1268 utterances (over 115,000 segments). The database was divided into a train (90%) and test sets (10%). The vowel set was divided into 41,066 train segments and 4,447 test segments. The consonant set comprised 64,618 train and 7,110 test segments. Each segment in the databases was marked with segment and syllable-level phonetic information. The databases also contained word boundaries and lexical stress information.

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.

5.1. Vowels

For the vowel analysis, we used the network of size 8 with discrete node set $\Delta = (V, Wpost, S, Utt, Cpre, Cpost, Wpre)$. In Figure 3 the RMS error values of the predicted vowel durations for rjs database for the BN of size 8 are shown. As can be seen from the figure, the BN model performs better than the SoP

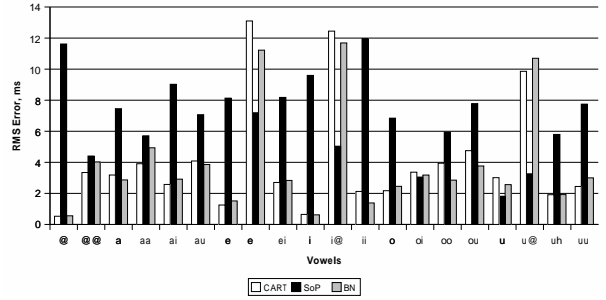


Figure 3: RMSE values of predicted vowel durations for rjs database.

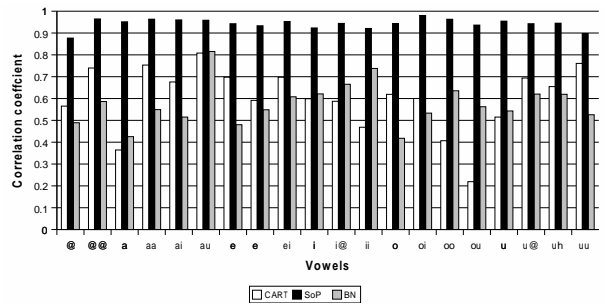


Figure 4: Correlation values of predicted vowel durations for rjs database.

model across different vowel classes, with the median RMSE values being 3 ms and 7 ms correspondingly. In that respect the BN model compares in performance with the CART model that produces the same median RMSE value of 3 ms. In Figure 4 the correlation coefficient values of the predicted vowel durations for rjs database are shown. As one can notice from the figure the BN model produces slightly lower correlation values, with the median correlation coefficient values across different vowel classes being 0.56, 0.61, and 0.94 for the BN, CART, and SoP models respectively. However, when calculated across the whole test set the correlation values are 0.84 for the BN model, and 0.82 and 0.72 for the CART and SoP model respectively. We can speculate that the BN model is too sensitive to the amount of data available for prediction, making good estimates for larger test sets.

We also investigated the problem of the network size on the accuracy of the BN parameter estimation. We experimented with the networks of sizes 5 and 6, with the discrete nodes sets being subsets of the discrete node set of the network of size 8.

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, [5]). Table 1 shows the

#	Subset	Nodes	BN Space Size
1	D V Wpost S Utt	5	360
2	D V Wpost S Utt Cpost	6	3600
3	D V Wpost S Utt Cpre Cpost Wpre	8	21600

Table 1: Vowel BNs of different sizes selected for duration analysis.

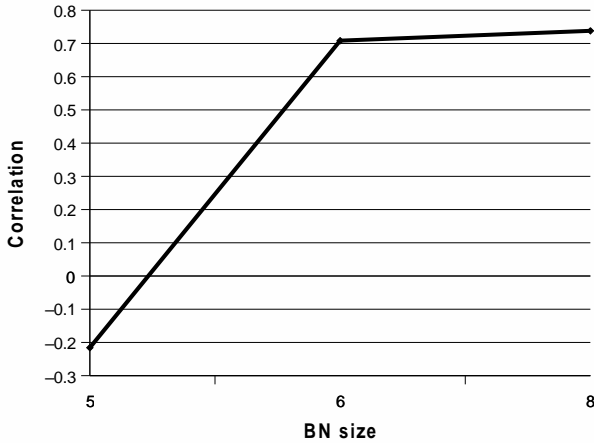


Figure 5: Correlation values for the vowel *ii* (424 segments) for the BNs of different sizes. Test set (4,447 segments).

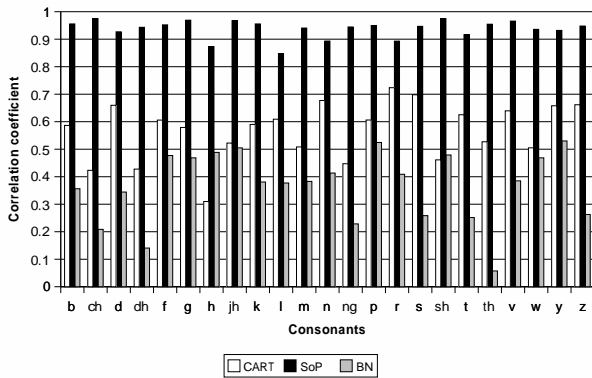


Figure 6: Correlation values of predicted consonant durations for *rjs* database.

node sets selected for our analysis. For the BN of size 6 we selected the discrete set $\Delta = (V, W_{post}, S, Utt, C_{pre})$. For the BN of size 5 the discrete set consisted of the nodes $\Delta = (V, W_{post}, S, Utt)$.

The correlation values for the networks of different sizes for the vowel *ii* are shown in Figure 5. As one can see from the figure, the correlation changes only slightly, from 0.73 to 0.70, when the node set does not contain previous context information at the segment and word level (nodes C_{pre} and W_{pre}). If the following context information at the segment level is disregarded (node C_{post}), the prediction power of the network degrades substantially. Such a behaviour of the BN may indicate the importance of the following context information for the BN parameter estimation.

5.2. Consonants

For our consonant duration analysis we selected just one network of size 6, with the node set $G = (D, C, W_{post}, S, NSyls, V_{Front})$.

In Figure 6 correlation coefficient values of the predicted consonant durations for *rjs* database are shown. As can be seen from the figure, the BN model slightly underperforms in terms of the correlation values compared to the CART and SoP models, with the median correlation coefficient values across differ-

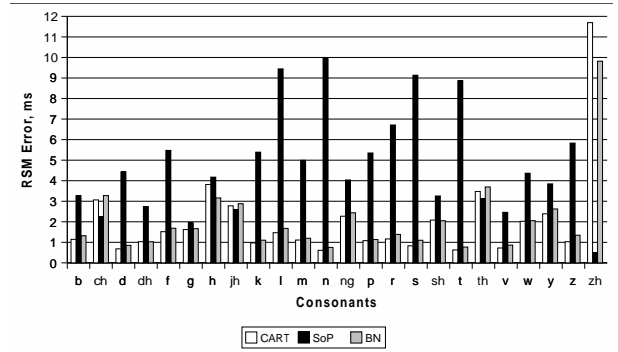


Figure 7: RMSE values of predicted consonant durations for *rjs* database.

ent vowel classes being 0.38, 0.59, and 0.95 for the BN, CART, and SoP models respectively. However, the calculations carried out for the whole test set produce the values 0.6 for the BN model, and 0.72 for the CART, and 0.51 for the SoP models.

In Figure 7 the values of RMS error values of the predicted consonant durations for *rjs* database are shown. As one can see from the figure, the BN model of size 6 gives the median RMS error value of 2 ms, performing better than the SoP model (the RMS error value of 4 ms) and worse than the CART model (the RMS error value of 1 ms).

From the above results, it follows that the consonant BN model performs no worse than the SoP model. We may also conclude that the network selected for our consonant analysis is not an optimal solution for the consonant duration prediction. In the future, we plan to experiment with networks of bigger sizes that would take into account previous and following context information.

6. Conclusions and future work

Our Bayesian analysis of vowel duration produces some promising results in terms of RMSE 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 3 ms; the corresponding values are 7 ms for the SoP and 3 ms for the CART models. However, it should be pointed out that the BN model is quite sensitive to the amount of data used for prediction, producing lower correlation values for smaller sets.

Based on the the consonant analysis results, we may conclude that the network selected for our consonant analysis is not an optimal solution for the consonant duration prediction. In the future, we plan to experiment with networks of bigger sizes that would take into account previous and following context information. In addition, we plan to explore the possibilities of learning network structure from data. This would aid tremendously in selecting optimal network configurations and maximizing the predicting power of the Bayesian analysis.

7. References

- [1] Goubanova, O., and Taylor, P., "Using Bayesian Belief Networks for model duration in text-to-speech systems", CD-ROM Proceedings ICSLP2000, Beijing, 2000.
- [2] Klatt, D. H., "Linguistic uses of segmental duration of En-

glish: Acoustic and perceptual evidence", Journal of the Acoustic Society of America, 59, 1976, p 1209-1211

- [3] Breiman, L. , Friedman, J., and Olshen, R., Classification and Regression Trees, Wadsworth and Brooks, Pacific Grove, CA, 1984.
- [4] van Santen, J. P. H., "Contextual effects on vowel durations", Speech Communication, 11, 1992, 513-546
- [5] van Santen, J.P. H., "Assignment of segmental duration in text-to-speech synthesis", Computer Speech and Language, Vol. 8, 1994, p 95-128,
- [6] Black, A.W., Taylor, P., and Caley, R. "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
- [7] Lauritzen, S., Graphical models, Oxford University Press, 1996.
- [8] van Son, R. J. J. H., and van Santen, J. P. H., "Strong interaction between factors influencing consonant duration", ESCA Eurospeech97, 1997, p 319-322