Bayesian Modelling Of Vowel Segment Duration For Text-to-Speech Synthesis Using Distinctive Features

Olga V. Goubanova

Centre for Speech Technology Research
University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK
olga@cstr.ed.ac.uk

Abstract

We apply a Bayesian belief network (BN) approach to vowel duration modelling, whereby vowel segment duration is modelled as a hybrid Bayesian network consisting of discrete and continuous nodes, with the nodes in the network representing linguistic factors that affect segment duration. Factor interaction is modelled in a concise way by causal relationships among the nodes in a directed acyclic (DAG) graph. New to the present research, we model segment identity as a set of distinctive features. The features chosen were frontness, height, length, and roundness. In addition, the BNs were augmented with the word class feature (content vs. function). We experimented with different BNs, and contrasted the results of the belief network model with those of Sums-of-Products (SoP) and classification and regression trees (CART) models. We trained and tested all three models on the same data. In terms of the RMS error and correlation coefficient, our BN model performs better than CART and SoP model.

1. Introduction

Segment duration is known to be affected by a number of linguistic factors such as segment identity, stress level of the syllable containing the segment, accent of the word the syllable is a part of, identity of preceding and following segments, and position of a target segment within a syllable, word, and utterance. When modelling segment duration for a text-to-speech system (TTS), large databases are used to estimate the parameters of the duration model. Databases used for duration modelling usually do not cover all the possible combinations of linguistic factors; data are *sparse*. In addition, databases are *not balanced*: different factor combinations occur with unequal frequencies. Nevertheless, the probability of rare factor combinations occurance is quite large even for a small sample of text [1]. Therefore, durational model should generalise well to successfully predict durations of these rare feature vectors. Since linguistic factors affecting segment duration interact, it should also model these factor interactions well.

Past approaches to segment duration modelling for TTS include rule-based [2], statistical (classification and regression trees [3]), and supervised data-driven methods (the Sums-of-Products, or SoP duration model [1],[4]). In general, CART models predict segment duration well, though they perform badly when data are noisy or the amount of missing data is large. In the SoP model the problems of data imbalance, data sparsity and factor interaction are treated satisfactorily by using general statistical techniques. However, this requires substantial data preprocessing, and consequently a large number of the model's parameters have to be estimated.

As an alternative to the conventional techniques of data modelling, we model segment duration using probabilistic Bayesian belief networks (BN) [5]. Our previous work on Bayesian modelling of segment duration proved to be promissing in overcoming unbalanced data and data sparisty problems [6], [7]. Factor interaction is modelled in a concise way by causal relationships among the nodes in a directed acyclic (DAG) graph. The BN model makes robust predictions in cases of missing or incomplete data. Compared to sums-of-products model, BN model also requires fewer parameters to be estimated.

The structure of the paper is as follows. We give a brief overview of Bayesian belief approach in section 2. We give the details of applying BN approach to modelling segment duration in section 4. We give the details of the databases used for segment duration modelling in section 3. We describe the experiments and discuss the results in section 5. We make the conclusions and discuss future work in section 6.

2. Bayesian Belief Networks

When using Bayesian networks for modelling segment duration, we represent linguistic factors that affect segment duration as nodes in a graph. Throughout the paper we use the terms *node*, *variable*, and *factor* interchangebaly. A Bayesian belief network is defined by a triple (G, Ω, P), where $G = (U, E)$ is a directed acyclic graph (DAG) with a node set U representing 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 $P(U)$ is a joint probability distribution (JPD) for all of the nodes in the graph G. Learning the whole JPD $P(U)$ requires an exponential number of BN parameters to be calculated. By using the so-called *Markov property* of BNs (each variable in a network is independent of its non-descendants given its parents), the joint probability $P(U)$ factorises into local conditional probabilities for each variable in the network. The $P(U)$ factorisation is:

$$P(U) = P(X_1, X_2, ..., X_n) = \prod_{j=1}^{n} P(X_j|Pa(X_j)) \quad (1)$$

where $Pa(X_j)$ is the set of parents of node X_j. We modelled vowel segment duration as a *hybrid* Bayesian network; consisting of discrete and continuous nodes. The problem domain set U of a hybrid BN is divided into a set of discrete variables Δ and a set of consinuous variables Γ, i.e. $U = \Delta \cup \Gamma$. The variables $U = (X_1, X_2, \ldots, X_n)$ in a hybrid BN are said to have a *conditional Gaussian (CG)* distribution; given a particular instantiation of discrete nodes $\mathbf{i} \in \Delta$, the continuous variables

$\mathbf{Y} = \{Y_1, Y_2, Y_3, \cdots, Y_k\} \in \Gamma$ follow a multivariate Gaussian distribution, i.e., the probability distribution function (pdf) over the continuous nodes has the form:

$$P(\mathbf{y}|\mathbf{i}) = \frac{1}{\sqrt{(2\pi)^d det\Sigma(\mathbf{i})}}$$
$$exp\{-\tfrac{1}{2}(\mathbf{y}(\mathbf{i}) - \vec{\mu}(\mathbf{i}))^T \ \Sigma(\mathbf{i})^{-1}(\mathbf{y}(\mathbf{i}) - \vec{\mu}(\mathbf{i})) \quad (2)$$

where d is the cardinality of the set Γ, $\mathbf{y}(\mathbf{i}) = (y_1, y_2, \cdots, y_k)$ are the instantiations of the continuous variables $\mathbf{Y} \in \Gamma$, $\vec{\mu}(\mathbf{i})$ and $\Sigma(\mathbf{i})$ are the mean vector and covariance matrix of the multivariate Gaussian distribution given the values of the discrete nodes $\mathbf{i} \in \Delta$; here the covariance matrix $\Sigma(\mathbf{i})$ is assumed to be positive definite.

3. Durational database

The databases used for this research were derived from Rhetorical Systems speech data. We used three databases; one database of General American (GA) English male speaker 'erm'; and two databases of Received Pronunciation (RP) English speakers, a female database 'lja' and a male database 'rjs'. Each database was divided into train (90%) and test (10%) sets. 'rjs' database of $98,763$ vowels was divided into $88,997$-segment train and $9,766$-segment test sets. 'lja' database of $39,224$ vowels was divided into $35,348$-segment train and $3,876$-segment test sets. 'erm' database of $63,188$ vowels was divided into $57,104$-segment train and $6,084$-segment test sets. Each segment in the data was labeled with segment, syllable, word, and utterance level phonetic and phonological information.

4. Bayesian analysis of segment duration

4.1. Defining linguistic factors of durational BN

In the case of durational BN, the set Γ consists of just one scalar node D that corresponds to the duration value of a vowel segment. The set Δ varies according to what causal factors are selected for analyis. For the present analysis we selected 9 linguistic causal factors that affect vowel duration shown in Table 1. Within word position factor $Wpost$ has 3 possible values corresponding to initial, medial, and final position of a syllable with a target vowel in a word. Stress factor S can take 2 values; stressed and unstressed. Within utterance position factor Utt describes phrasal position of a word with a target vowel taking on 3 values; initial, medial, and final. The identity of the following segment factor $Cpost$ takes on 10 values. When the following segment is a consonant, the values of $Cpost$ node are based on voicing and manner of production features for consonant; voiceless stops, voiceless affricates, liquids, voiceless fricatives, nasals, voiced stops, voiced affricates, and voiced fricatives. In addition, $Cpost$ node takes on values 'vowel' and 'silence'.

Factor	Wpost	S	Utt	Cpost	
# Levels	3	2	3	10	
Example	initial	stressed	final	voiced stop	
Factor	Front	Height	Length	Round	WdCl
# Levels	3	3	4	2	2
Example	back	high	long	round	content

Table 1: Linguistic factors selected for the Bayesian modelling of vowel duration.

We also introduced a word class factor represented by a binary discrete node $WdCl$, describing whether a word with a target vowel is content (open class) or function (closed class). Word class factor is meant to implicitly represent word frequency information. From the studies of the effect of word frequency on duration of content [8] and function [9] words, it is known that the duration of a more frequent word tends to be shorter than the one of a less frequent word. Therefore, we assumed that word frequency should have an effect on word duration and consequently on a word's segment (vowel) durations. In the future, we plan to use continuous word frequency factor directly.

4.2. Modelling vowel identity

We modelled vowel segment identity as a combination of four factors corresponding to the following phonological (distinctive) features. The frontness of a target vowel is represented by the factor $Front$ that can have 3 values; front, medial, and back. The height of a vowel segment is represented by the factor $Height$ that can have 3 values; high, medial, and low. The factor $Length$ can take on 4 values; short, long, diphtong, and shwa. The factor $Round$ can have 2 values, rounded and unrounded.

4.3. Learning durational BN

The process of BN learning consists of BN structure learning and BN parameter learning. Once the BN structure is known, the parameters of the BN, i.e. the parameters of the conditional probability distributions (CPDs) of the nodes are estimated. The CPD parameters of the discrete nodes are just the entries in the Conditional Probability Table (CPT). The parameters of the continuous nodes are the mean vector ($\vec{\mu}$ and covariance matrix Σ) of the Gaussian pdf. First, we performed BN structure learning. We used the K2 structure learning algorithm (see [10] for details). In brief, the K2 algorithm uses a greedy heuristic approach whereby, given the fixed ordering of the nodes (with parents preceding children), a parent node is succesively added to a parent set of each node in such a way that maximally improves the joint probability of a network structure and data. Since there are no network structure learning algorithms developed for hybrid BNs, we applied the K2 algorithm to the durational data that were uniformly discretised. We chose several levels of discretisation ranging from 2 to 7 bins. We applied the K2 algorithm to 3 discretised data sets; 'erm', 'rjs', and 'lja'. The learning resulted in 7 different network structures; the BNs differed in the connections between the causal nodes and the durational node D. After removing some linguistically superficial connections (between the causal nodes) learned by the K2 algorithm, we then estimated the nodes CPDs. An example BN with $Pa(D) = \{Cpost, Front, Length, Round\}$ is shown in Figure 1. The number of BN parameters as well as the linguistic nodes connected to the D node for 7 BNs learned are shown in

BN #	**Pa(D)**	**# params**
BN1	Cpost Length Round	80
BN2	Cpost Front Length Round	240
BN3	Cpost Front Height Length Round	720
BN4	Cpost Front Height Length WdCl	720
BN5	Wpost S Cpost Round	120
BN6	Wpost Cpost Length Round WdCl	480
BN7	Wpost Utt Cpost Front Height Length WdCl	6480

Table 2: Connections to the durational node D learned by the K2 algorithm applied to the discretised data.

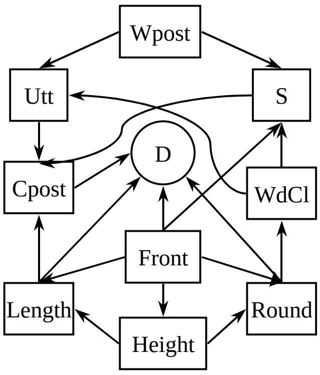

Figure 1: Example durational Bayesian network of size 10; boxes represent discrete nodes, oval represents a continuous node.

Table 2. The connections among the causal nodes themselves are fixed for all the learned BNs; they are the same as those shown in Figure 1. The prior CPD parameters of the discrete linguistic nodes were estimated as Dirichlet priors. Since D is a scalar node with all the parents being discrete, for each instantiation of its discrete parents $\mathbf{i} \in Pa(D)$ the conditional probability distribution (CPD) is given by an univariate Gaussian distribution with mean $\mu(\mathbf{i})$ and standard deviation $\sigma(\mathbf{i})$:

$$P(y|\mathbf{i} \in Pa(D)) = \frac{1}{\sqrt{(2\pi)\sigma^2(\mathbf{i})}} exp\{-\frac{(y(\mathbf{i}) - \mu(\mathbf{i}))^2}{2\sigma^2(\mathbf{i})}\} \quad (3)$$

The prior parameters of this univariate Gaussian distribution $N(y; \mu(\mathbf{i}), \sigma^2(\mathbf{i}))$ were estimated from the training set as sample means. All calculations were done in the z-score domain. The learning of the parameters of the BNs was done via the EM algorithm, with the causal nodes observed and the durational node D hidden. Following the BN parameter learning, the inference was performed on the test set. The learning and inference were done for 7 different BNs, for each database separately.

5. Experimental Results and Discussion

Given 7 different BNs learned by the K2 algorithm, we set out to find the model that would be optimal in terms of RMS error (minimal) and correlation coefficient (maximal). We call this Maximum Correlation – Minimum RMS Error (MAXC-MINEr) criterion. In Figure 5 the results of the mean (across the database) RMS error values of the predicted vowel durations by model type are shown. In Figure 5 the results of the mean (across the database) correlation coefficient values of the predicted vowel durations by model type are shown. In general, in terms of RMS error all the BNs selected for the analysis perform better that both SoP and CART models. For 'rjs' database $BN4$ model produces the mean RMS error of 1.5 msec compared to 8 msec and 32.5 msec for SoP and CART models respectively. In terms of the correlation values, there are some BNs (e.g. $BN3$ and $BN4$) that perform better than CART model, and no worse than SoP model. For 'lja' database $BN1$ model produces the mean correlation value of 0.76 compared to 0.69 and 0.94 for CART and SoP models respectively. Based on MINC-MINEr optimisation criterion, we selected 3 optimal BNs: $BN1$, $BN3$, and $BN4$.

Since our optimal BN model selection criterion is based on the RMS error and correlation values averaged across a paricular database, we also looked at the performance of the optimal BNs for each vowel class separately. We assumed that for each vowel class there may exist a different optimal network. The analysis of the RMS error values for $BN4$ model for 'lja' database revealed that the model makes robust predictions of the vowel segment durations, with the RMS error values ranging $1 - 2$ msec. The results of the correlation values of the predicted vowel durations by vowel class for $BN4$ model for 'lja' database are shown in Figure 5. As can be seen from the figure, for the majority of the vowel classes the correlation values range $0.47 - 0.85$. The obvious outlier was vowel $/u@/$; with the RMS error being 18 msec and the correlation being 0.07. Comparing the correlation values for the segment $/u@/$ across all the BNs had shown that $BN4$ being on average an optimal choice, is not an optimal BN for this vowel. In fact, $BN6$ is a better choice with the correlation value of 0.91. Likewise, for the vowel $/@/$ it is the network $BN1$ that is optimal with the RMS error and the correlation values being 2 msec and 0.89 respectively. In Figure 5 the correlation values by vowel class by model type are shown for 'lja' database'. The search for an

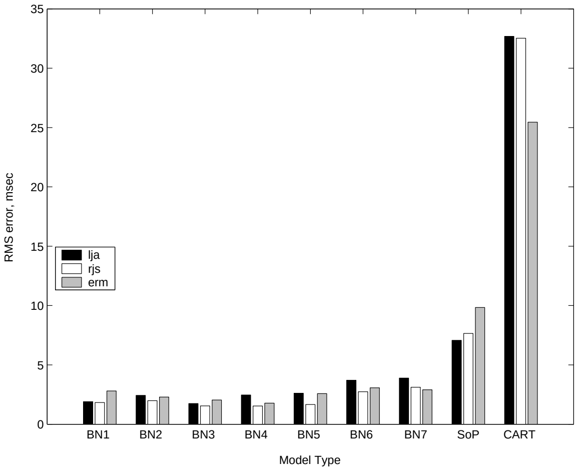

Figure 2: The mean RMS error values of the predicted vowel durations by model type (Bayesian, CART and SoP) by database ('lja', 'rjs', and 'erm').

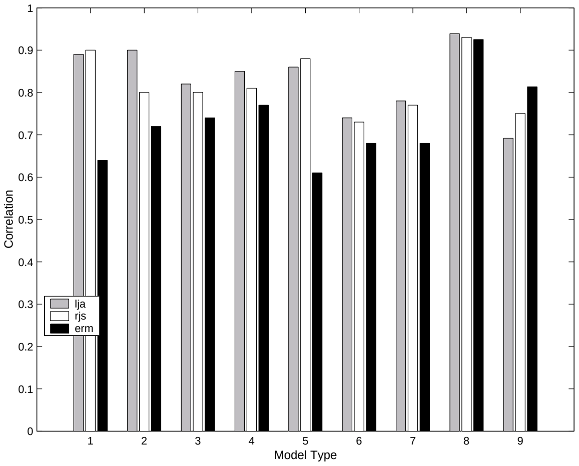

Figure 3: The mean correlation coefficient values of the predicted vowel durations by model type (Bayesian, CART, and SoP) by database ('lja', 'rjs', and 'erm').

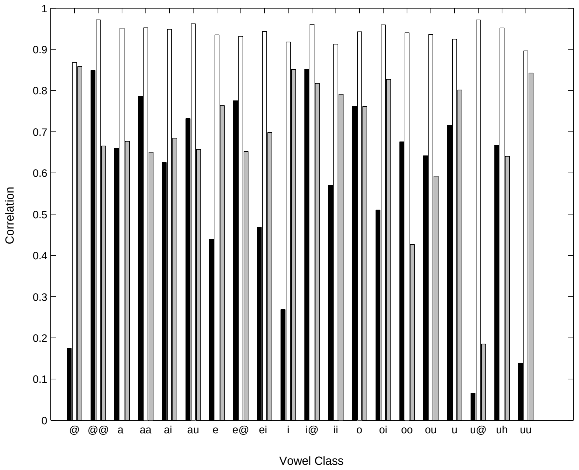

Figure 4: The correlation values of the predicted vowel durations by vowel class by model type (Bayesian, SoP, and CART); 'lja' database; BN4 Bayesian model. Black - 'lja', white - 'rjs', gray - 'erm'.

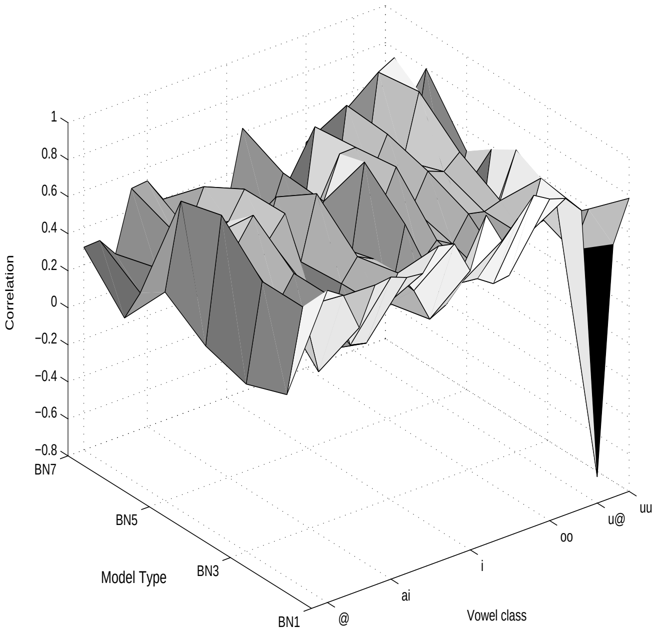

Figure 5: The correlation values of the predicted vowel durations by vowel class by BN model type for 'lja' database.

optimal BN model for each vowel class can be thought of as a search for the maximum peaks on this correlation surface. For each vowel class for 'lja' database the optimal model selected produces the correlation values ranging $0.66 - 0.99$; these values are better than the results for CART $0.18 - 0.86$ and no worse than those for SoP $(0.87 - 0.97)$ models.

6. Conclusions and Future Work

First, we implemented the BN structure learning procedure for discretised durational data using the K2 structure learning algorithm. Second, we analysed 7 BNs learned by the K2 algorithm and chose the maximum correlation – minimum RMS error optimal candidate network for each database. Third, for each vowel class we selected the optimal BN separately. For each vowel class the optimal BN model produces promissing results in terms of RMS error values; our BN model significantly outperforms both CART and SoP models. In terms of the correlation coeffcient, the BN model results are better than CART model and comparable to the SoP model results. Therefore, Bayesian belief network model can be sucessfully used for vowel duration modelling for text-to-speech systems. In the future, we will consider other linguistic factors such as word frequency and boundary type for our BN analysis. We will also implement the BN durational model in the Festival [11] speech synthesis system.

7. Acknowledgments

We are thankful to Rhetorical Systems PLC, Edinburgh, UK, for using their speech data. We would like to acknowledge the financial support from the CSTR, University of Edinburgh, that made this work possible. We would like to thank Kevin Murphy for using his BayesNet Toolkit.

8. References

[1] J.P.H. van Santen, "Assignment of segmental duration in text-to-speech synthesis", Computer Speech and Language, Vol. 8, 1994, 95-128,

[2] D.H. Klatt, "Linguistic uses of segmental duration of English: Acoustic and perceptual evidence", Journal of the Acoustic Society of America, 59, 1976, 1209-1211

[3] L. Breiman, J. Friedman, and R. Olshen, Classification and Regression Trees, Wadsworth and Brooks, Pacific Grove, CA, 1984.

[4] J.P.H. van Santen, "Contextual effects on vowel durations", Speech Communication, 11, 1992, 513-546

[5] R. Cowell, R., A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter, "Probabilistic networks and expert systems", Springer, 1999

[6] O. Goubanova, and P. Taylor, "Using Bayesian Belief Networks for model duration in text-to-speech systems", CD-ROM Proceedings ICSLP2000, Beijing, 2000.

[7] O. Goubanova, "Predicting segmental duration using Bayesian belief networks", CD-ROM Proceedings 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Scotland, 2001.

[8] M. Gregory, A. Bell, D. Jurafsky, and W. Raymond. "Frequency and predictability effects on the duration of content words in conversation, Journal of the Acoustic Society of America, 110, 2738.

[9] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, D. Gildea. "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation", Journal of the Acoustic Society of America, 113, 2003, 1001-1024.

[10] G. F. Cooper, E. Herskovits. "A Bayesian method for the induction of probabilistic networks from data", Machine Learning, 1992, 309-347.

[11] 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.
URL http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html