Predicting Consonant Duration with Bayesian Belief Networks

Predicting Consonant Duration with Bayesian Belief Networks

Olga Goubanova, Simon King[*]

Centre for Speech Technology Research
University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK
s9808417@inf.ed.ac.uk

Abstract

Consonant duration is influenced by a number of linguistic factors such as the consonant's identity, within-word position, stress level of the previous and following vowels, phrasal position of the word containing the target consonant, its syllabic position, identity of the previous and following segments. In our work, consonant duration is predicted from a Bayesian belief network (BN) consisting of discrete nodes for the linguistic factors and a single continuous node for the consonant's duration. Interactions between factors are represented as conditional dependency arcs in this graphical model. Given the parameters of the belief network, the duration of each consonant in the test set is then predicted as the value with the maximum probability. We compare the results of the belief network model with those of sums-of-products (SoP) and classification and regression tree (CART) models using the same data. In terms of RMS error, our BN model performs better than both CART and SoP models. In terms of the correlation coefficient, our BN model performs better than SoP model, and no worse than CART model. In addition, the Bayesian model reliably predicts consonant duration in cases of missing or hidden linguistic factors.

1. Introduction

In a concatenative text-to-speech (TTS) system, the duration of a phone is usually predicted from a database of feature vectors that each consist of a set of linguistic factors' values describing a phone in a particular context. Databases used to train phone duration models are usually sparse and un-balanced: they cover only a fraction of all linguistically possible combinations of feature vectors; different factor combinations occur with unequal frequencies. However, it has been shown [1], [2] that the probability of a rare feature vector occurring even in a small sample of text is quite high. Furthermore, factors affecting phones' duration interact: a set of two or more factors may amplify or attenuate the effect of other factors. A robust model for predicting phone duration must generalise well in order to successfully predict the duration of phones with these rare feature vectors. Since linguistic factors affecting segment duration interact, we expect that modelling these factor interactions will give a better model.

There have been a number of models developed for predicting a phone's duration, ranging from rule-based [3] to classification and regression tree (CART) [4] to sums-of-products (SoP) models [1], [2]. In the CART model, a phone's duration (absolute or z-score) is predicted by finding the data cluster in the decision tree that matches as many of the feature vector attributes as possible (in the order specified by the tree). The CART model is easy to build, robust to errors in data, but performs poorly when the percent of missing data is too high. In the SoP model, the log of a phone's duration is predicted as a sum of factors' product terms. The SoP model predicts phone duration with high accuracy, even in cases of hidden or missing data. However, this is done at the cost of substantial data pre-processing. In addition, the number of different sums-of-products models grows hyper-exponentially with the number of factors. Therefore, one must use some heuristic search techniques to find the model that fits the data the best.

We model a phone's duration using probabilistic Bayesian belief networks (BN) [5], whereby linguistic factors that influence a phone's duration are represented as the nodes in a directed acyclic (DAG) graph, and factors' interactions are modelled by causal relationships among the nodes in the DAG. The BN model makes robust predictions in cases of missing or incomplete data, therefore thoroughly addressing data sparsity and data imbalance problems. We successfully applied a Bayesian model for predicting phone duration in our previous work [6], [7], [8]. We discuss the Bayesian models for predicting phone duration in more detail in [9].

The structure of the paper is as follows. We briefly introduce the theory of Bayesian belief networks in Section 2. We describe the database used for predicting consonant duration in Section 3 and define a BN for predicting consonant duration in Section 4. We describe the training procedure in Section 5 and discuss the results in Section 6. We draw conclusions and discuss future work in Section 7.

2. Bayesian belief network basics

A *Bayesian network* for a set of variables $\mathbf{U} = \{X_1, X_2, ..., X_n\}$ is a pair (G, P), where $G = (\mathbf{V}, \mathbf{E})$ is a directed acyclic graph (DAG) with the vertex set $\mathbf{V}$ and the set of directed edges $\mathbf{E}$. The vertex set $\mathbf{V}$ encodes the information about the problem domain variables $\mathbf{U}$ and the edges set $\mathbf{E}$ encodes the relations between the variables of the domain set $\mathbf{U}$. $P(\mathbf{U})$ is a joint probability distribution (JPD) over the variables of $\mathbf{U}$ that factors according to the graph structure G:

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

where n is the size of the network, $\mathbf{Pa}(X_j)$ is the set of parents of variable X_j.

For consonant duration prediction we use a special kind of hybrid (containing discrete and continuous variables) Bayesian network, namely a *Conditional Gaussian* (CG) network. We say that the variables $X_j \in \mathbf{U}; j = 1, \cdots, n$ of a hybrid BN have a *conditional Gaussian* (CG) distribution, if the BN's continuous variables follow a multivariate Gaussian distribution

[*] Supported by EPSRC Advanced Research Fellowship GR/T04649/01.

given the values of the discrete variables. For a CG network with a single continuous variable for a consonant's duration, the distribution of that durational variable, *D*, has a 1-dimensional CG distribution, with its PDF being:

$$p(D = d|\mathbf{i}) = \frac{1}{\sqrt{(2\pi)}\sigma(\mathbf{i})}\ exp\{-\frac{(d-\mu(\mathbf{i}))^2}{2\sigma^2(\mathbf{i})}\}, \quad (2)$$

where for each configuration (i.e. instantiation with certain values) of the discrete parents $\mathbf{i} \in \mathbf{I}$ of the duration variable *D*, $\mu(\mathbf{i})$ and $\sigma^2(\mathbf{i})$ are the conditional mean and variance of *D*.

3. Database

Number of consonant feature vectors			
Voice	**Train**	**Test**	**Total**
lja	54,489	6,015	60,504
rjs	138,635	14,998	153,633
erm	85,048	9,039	94,087

Table 1: The number of consonant tokens in the train, test sets, and the total number for the 3 voices: **lja**, **rjs**, and **erm**.

The data were derived from 3 Rhetorical PLC databases: 2 RP English voices *rjs* (male) and *lja* (female), and 1 GA English voice *erm* (male). The databases consist of a set of utterances, one set for each voice. The set of utterances was divided into train (90%) and test (10%) sets. The train and test data were dumped as a vector of categorical features for each consonant token, along with the consonant's duration, using Rhetorical internal tools. The amount of consonant data for the 3 voices is shown in Table 1.

4. Bayesian model for consonants

4.1. Linguistic variables chosen

A consonant's duration is influenced by a number of linguistic factors such as the consonant's identity, frontness of the syllabic vowel, identity of the previous and following segments, a consonant's within-word and syllabic positions, stress level of the previous and following vowels, phrasal position of the word containing the target consonant [1]-[3], [10].

Variable	# Values	Example
manner-voice MV	9	voiced fricative
within-word position Wpos	3	initial
stress S	2	stressed
within-utterance position Utt	3	utterance medial
syllabic position Syl	3	coda
previous segment identity Cpre	3	consonant
following segment identity Cpos	3	silence
frontness of syllabic vowel Front	3	front

Table 2: Linguistic variables chosen for the Bayesian prediction of consonant duration.

Hence, for our Bayesian model we selected the 8 linguistic factors shown in Table 2. Consonant identity was encoded as a compound variable *MV* that represents manner of production and voicing distinctive features; it takes on values: *voiceless stops*, *voiceless affricates*, *approximants*, *voiceless fricatives*, *nasals*, *voiced stops*, *voiced affricates*, *voiced fricatives* and *liquids*. The within-word position variable *Wpos* represents the position of a consonant within a word; it takes on *initial*, *medial*, and *final* values. The stress variable *S* represents the stress level of a syllabic vowel, and takes on *stressed* and *unstressed* values. The utterance position variable *Utt* describes phrasal position of a word with a target consonant; it takes on *initial*, *medial*, and *final* values. The syllabic position variable *Syl* represents the position of a consonant within a syllable; it takes on the values *onset*, *coda*, and *syllabic*. The identity of the previous (following) segment variable(s) *Cpre* (*Cpos*) represents the information about the previous (following) segment in a broad sense; it takes on 3 values: *consonant*, *vowel*, and *silence*. The frontness of a syllabic vowel variable *Front* takes on 3 values: *front*, *medial*, and *back*. Hence, the "universe" (i.e. the nodes) of the BN consisted of 9 (including the duration variable *D*) variables: $\mathbf{U} = \{MV, Wpos, S, Utt, Syl, Cpre, Cpos, Front, D\}$.

4.2. Learning belief network structure

To learn the belief network structure we applied the K2 structure learning algorithm (see [11] for details). The K2 algorithm uses a greedy heuristic approach whereby, given a fixed ordering of the variables (with parents preceding children), a parent variable is successively added to the parent set of each variable in such a way that maximally improves the joint probability of the training data given the model. Since there are no network structure learning algorithms for hybrid BNs, we applied the K2 algorithm to a version of the data in which the continuous values of durations were uniformly discretized. We chose several levels of discretisation ranging from 2 to 10 bins (i.e. 9 different versions of the data sets for each of the three voices). We then applied the K2 algorithm to each set. As a result, we identified 8 different network topologies with which to perform further experiments using the original, continuously-valued, duration data. Each network is a representative of a class of networks: the networks within a class have the same duration variable *D* parent set $\mathbf{Pa}(D)$, being different otherwise (the parent sets of the linguistic variables within the same class may be different). If all linguistic variables are observed, all networks within a class will give the same conditional PDF for *D*.

Name	Pa(D)	# params
CBN1	*MV, Cpos*	27
CBN2	*MV, Syl, Front*	81
CBN3	*MV, Wpos, S, Syl, Cpre, Cpos, Front*	4,374
CBN4	*MV, Wpos, S, Utt, Syl, Cpre, Cpos*	4,374
CBN5	*MV, Wpos, S, Utt, Syl, Cpre, Cpos, Front*	13,122
CBN6	*MV, Wpos, Syl, Cpre, Cpos*	729
CBN7	*MV, Wpos, Syl, Cpre, Cpos, Front*	2,187
CBN8	*MV, Wpos, Utt, Syl, Cpre, Cpos, Front*	6561

Table 3: BNs learnt by the K2 algorithm, with consonant durations being uniformly discretized. The number of the CG pdf parameters of the *D* variable is shown in the third column of the table.

The duration variable *D* parent sets $\mathbf{Pa}(D)$ for 8 networks are shown in Table 3. An example BN with the parent set $\mathbf{Pa}(D) = \{MV, Wpos, S, Utt, Syl, Cpre, Cpos, Front\}$ is shown in Figure 1.

5. Model training

The goal of the training experiments was to study the performance of the networks learnt from the data and to compare this to baseline CART and SoP models. In addition, we wanted to

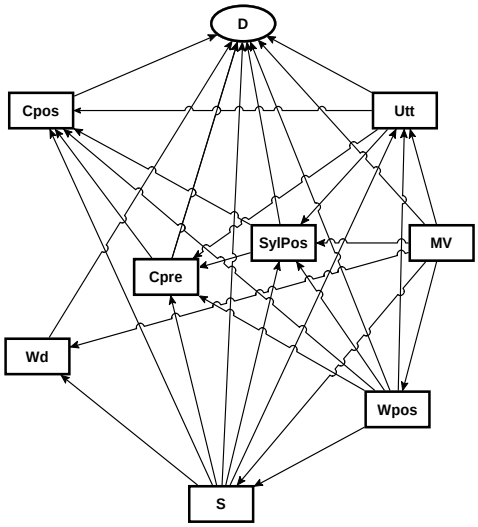

Figure 1: Bayesian network learnt by the K2 algorithm; duration node D parent set $\mathbf{Pa}(D) = \{$*MV, Wpos, S, Utt, Syl, Cpre, Cpos, Front*$\}$.

find the best (among the 8 belief networks learnt) BN model for each type of consonant. By best model we mean a network that predicts consonant duration with the maximum correlation and the minimum RMS error. We trained all 8 models on each of the three voices: *lja*, *rjs*, *erm*.

We trained each model by estimating the networks' parameters. We assumed that the discrete variables follow multinomial distribution. For the discrete variables we calculated their parameters as the MAP estimates. The prior values of the discrete (linguistic) variables were estimated as Dirichlet priors with equivalent sample size of 2. To calculate the MAP estimates, we used the EM algorithm, with the duration variable D being hidden and the discrete variables being observed.

We assumed the continuous variable for consonant's duration D follows a 1-dimensional CG distribution, with probability density function defined in Equation 2. We estimated its parameters $\Theta(\mathbf{i}) = (\mu(\mathbf{i}), \sigma^2(\mathbf{i}))$ as ML estimates: for each instantiation $\mathbf{i}$ of the discrete parents $\mathbf{Pa}(D) = \mathbf{i}$ in the train set we calculated the mean and standard deviation of a consonant's duration.

6. Results

6.1. Overall behaviour

To compare the performance of our BN models to CART and SoP models, we used 2 metrics: *test sample correlation coefficient* and *Root Mean Squared Error* (RMS error) in milliseconds (ms). After we trained our Bayesian models, we predicted the duration of each consonant in the test set via Bayesian inference: we calculate the PDF of D given the observed values of the linguistic variables. The most likely value of D (i.e. the conditional mean) is chosen as the consonant's duration. Table 4 shows the correlation and RMS error results for the 8 Bayesian models as well as SoP and CART. As can be seen from the table, for the *lja* and *rjs* voices the *CBN3* network predicts consonant duration with the maximum test sample correlation (0.84 and 0.80), beating the CART (0.78 and 0.80) and SoP (0.73 and 0.79) models. For the *erm* voice, the *CBN4* network predicts consonant duration with a maximum correlation of 0.80 which

Model	**Voice**					
	lja	**rjs**	**erm**	**lja**	**rjs**	**erm**
CBN1	0.80	0.77	0.69	3.8	4.4	3.8
CBN2	0.73	0.76	0.67	5.1	5.6	5.1
CBN3	**0.84**	**0.80**	0.69	**3.5**	**4.1**	**3.6**
CBN4	0.72	0.74	**0.80**	4.6	5.1	4.5
CBN5	0.71	0.73	0.74	3.7	4.3	4.5
CBN6	0.80	0.74	0.75	4.6	5.2	4.6
CBN7	0.76	0.73	0.73	4.7	5.3	4.7
CBN8	0.56	0.49	0.75	3.5	4.1	3.7
CART	0.78	0.80	0.82	21	20	24
SoP	0.73	0.79	0.76	25	26	33
SoP-German	0.896					
SoP-Dutch	0.77			23.4		

Table 4: The correlation and RMS error results by model type and voice. **SoP-German** – SoP model for German [13]; **SoP-Dutch** – SoP model for Dutch [12].

is higher than that of the SoP (0.76), but smaller than that of the CART (0.82) models. The best BN models better than the SoP model for Dutch (0.77) [12] and no worse than the SoP model for German (0.896) [13]. In terms of the RMS error, all BN models beat both the SoP and CART models.

6.2. Best network for each consonant type

For each consonant type we chose the best network in terms of maximum correlation and minimum RMS error. Figure 2

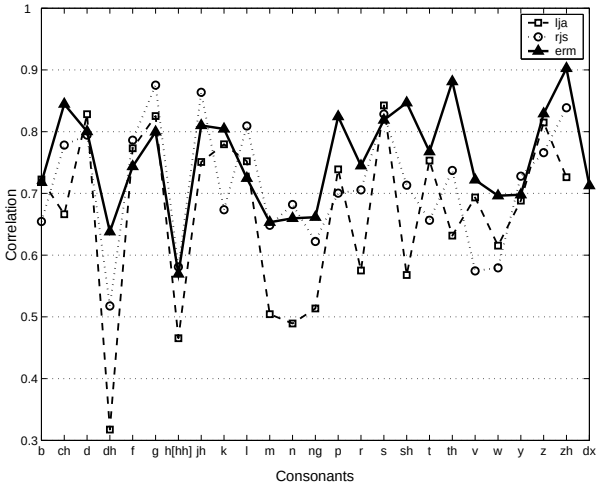

Figure 2: The correlation between predicted and actual durations by consonant type and voice. The best (maximum correlation) network for each consonant type is chosen.

shows the correlation results for each consonant type for the 3 voices (*lja*, *rjs* and *erm*). As can be seen from the figure, for the *lja* voice, the correlation ranges from 0.31 ($/dh/$) to 0.84 (for $/s/$). There are 5 consonants: $/dh, h, m, n, ng/$ for which the correlations are around or below a value of 0.5. For 37% of the consonants, the best BN models predict duration with a correlation greater or equal to 0.75, which is better than the test set correlation of the the SoP (0.73) model, and no worse than that of the CART (0.78) model.

For the *rjs* voice, the correlation ranges from 0.51 ($/dh/$) to 0.88 ($/g/$). There is one consonant $/dh/$ for which the correlation is 0.51. There are also 3 consonants: $/h, v, w/$ for which the correlations are around or slightly below 0.6. For 37% of the consonants the best models predict consonant duration with

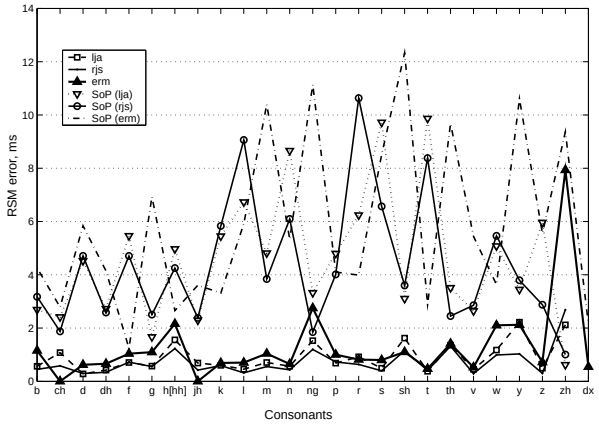

Figure 3: The RMS error by consonant type and voice. The best (minimum RMS error) network for each consonant type is chosen.

a correlation greater than 0.75 which is no worse than the test set correlation of the SoP (0.79) and the CART (0.80) models.

For the *erm* voice, the correlation ranges from 0.59 ($/hh/$) to 0.90 ($/zh/$). There is just one consonant $/hh/$ for which the correlation is below 0.6. There are also 6 consonants: $/dh, m, n, ng, w, y/$ for which the correlation is around or slightly below 0.7. For 44% of the consonants, the best BN models predict consonant duration with a correlation around or above 0.8 value which is better than the SoP (0.79) and no worse than the CART (0.82) models.

Figure 3 shows the RMS error results for the Bayesian and SoP models, for each consonant type, for the 3 voices (*lja, rjs* and *erm*). (Since the results for the CART model are of the same level of magnitude as those for the SoP model, they are not shown in the figure.) For all voices overall, the belief models predict consonant duration with RMS errors that are significantly (e.g. $t_{23} = 4.43, p < 0.001$ for *rjs* voice) smaller than these of the CART and SoP models.

7. Conclusions and future work

We have used a Bayesian model for predicting consonant duration using 8 linguistic variables (factors) that are known to influence consonant duration. We applied the K2 structure learning algorithm to the discretized duration data and found 8 belief networks that describe the data the best. We then trained each model by calculating the networks' parameters as ML estimators for continuous duration variable *D*, and MAP estimators for the discrete variables. To calculate the MAP estimates we used the EM algorithm. We analysed the performance of each BN model on the 3 voices: *lja, rjs*, and *erm* and compared it to the SoP and CART models.

On average, the Bayesian models predict consonant duration with a correlation that is better (0.72-0.84) than that of the SoP model (0.73-0.76), and no worse than the that of the CART model (0.78-0.82). In terms of RMS error, our belief networks (1.5-3.5ms) are better than either the SoP (25-33ms) or CART (20-24ms) models.

We chose the best model for each consonant type. In terms of the RMS error, for each of the consonant types, the corresponding best model predicts consonant duration with a RMS error smaller than that of the SoP and CART models. In terms of correlation, for at least 37% of the consonants, the best BN model performs better than either SoP or CART. However, there are 4 consonants: $/dh, h, m, n, ng/$ for which these best models give a correlation that is below 0.5. This can not be explained by low frequencies of these consonants in the data sets since the counts are high (e.g. over $18,000$ $/n/$ segments in the train set for the *rjs* voice). Still, our best model for $/n/$ predicts duration with a correlation of 0.4. Hence, it may well be that the chosen *best* model is not really the best one in terms of the linguistic variables chosen for the analysis. One possible explanation of such an unsatisfactory performance of the model for $/n$ is that consonant identity was represented by voice and manner distinctive features which may not be the best descriptor for this consonant. In future, we may try representing consonants such as $/n/$ with a place of articulation feature instead. In addition, we should search for a better model by analysing which linguistic factors are the *strongest* predictors of consonant duration for each consonant type separately.

8. References

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

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

[3] Klatt, D.H., "Linguistic uses of segmental duration of English: Acoustic and perceptual evidence", J. Acoust. Soc. Amer., 59, 1976, 1209-1211

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

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

[6] Goubanova, O., and Taylor, P. "Using Bayesian belief networks to model duration in text-to-speech systems", CD-ROM Proc. ICSLP 2000, 2000, Beijing, China

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

[8] Goubanova, O., "Bayesian modelling of vowel segment duration for text-to-speech synthesis using distinctive features", CD-ROM Proc. 15th Int. Conf. Phonetic Sciences, 2003, Spain

[9] Goubanova, O., "Bayesian networks for predicting durations of phones", Ph.D. Thesis, University of Edinburgh, Submitted 2005

[10] Van Son, R.J.J.H. and Van Santen, J.P.H., "Strong interaction between factors influencing consonant duration", CD-ROM Proc. Eurospeech 97, 1997, Rhodes, Greece

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

[12] Klabbers, E. and Van Santen, J., "Predicting segmental duration for Dutch using sums-of-products approach", CD-ROM Proc. of the ICSLP 2000, 2000, Beijing, China

[13] Moebius, B. and Van Santen, J., "Modelling segmental duration in German text-to-speech synthesis", CD-ROM Proc. ICSLP'96, 1996, Pennsylvania, USA