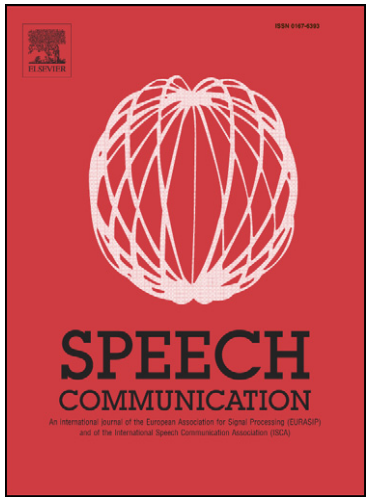
Bayesian networks for phone duration prediction

Olga Goubanova, Simon King



Bayesian networks for phone duration prediction

Olga Goubanova, Simon King

PII: S0167-6393(07)00175-6
DOI: 10.1016/j.specom.2007.10.002
Reference: SPECOM 1672

To appear in: *Speech Communication*

Received Date: 15 November 2006
Revised Date: 9 July 2007
Accepted Date: 23 October 2007

Please cite this article as: Goubanova, O., King, S., Bayesian networks for phone duration prediction, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.10.002

Bayesian networks for phone duration prediction

Olga Goubanova and Simon King

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

Abstract

In a text-to-speech system, the duration of each phone may be predicted by a duration model. This model is usually trained using a database of phones with known durations; each phone (and the context it appears in) is characterised by a *feature vector* that is composed of a set of linguistic factor values. We describe the use of a graphical model – a Bayesian Network – for predicting the duration of a phone, given the values for these factors. The network has one discrete *variable* for each of the linguistic *factors* and a single continuous variable for the phone's duration. Dependencies between variables (or the lack of them) are represented in the BN structure by arcs (or missing arcs) between pairs of nodes. During training, both the topology of the network and its parameters are learned from labelled data. We compare the results of the BN model with results for Sums of Products and CART models on the same data. In terms of the root mean square error, the BN model performs much better than both CART and SoP models. In terms of correlation coefficient, the BN model performs better than the SoP model, and as well as the CART model. A BN model has certain advantages over CART and SoP models. Training SoP models requires a high degree of expertise. CART models do not deal with interactions between factors in any explicit way. As we demonstrate, a BN model can also make accurate predictions of a phone's duration, even when the values for some of the linguistic factors are unknown.

Key words:
Text-to-speech; Bayesian Networks; Duration Modelling; Sums of Products; Classification and Regression Trees

Email addresses: ogoubanova@netscape.net (Olga Goubanova), Simon.King@ed.ac.uk (Simon King).

1 Introduction

1.1 Duration modelling for text-to-speech synthesis

We present comparative experimental results for 3 classes of phone duration prediction model: Bayesian Networks (BNs), Sums-of-Product (SoP) models van Santen (1992), and Classification and Regression Trees (CARTs). The principal application for these models is in text-to-speech synthesis.

In text-to-speech systems, it is often necessary to predict the prosody of the output speech; segment durations are an important aspect of prosody. Although in some unit-selection systems, such as Festival 2 (Clark *et al.*, 2004), no prediction of duration is required, this can lead to unpredictable prosody in the output speech. Even if the predicted durations are not imposed on the selected units via signal processing, the prediction of phone duration can still be used to compute a duration component of the target cost. In other cases, such as non-concatenative systems (e.g. Hidden Markov Model approaches, Tokuda *et al.*, 2002) or expressive/emotional speech synthesis (e.g. Strom *et al.*, 2006), explicit prediction of phone durations are necessary. Since duration is a factor affecting listener's perception of naturalness of synthetic speech (e.g. Mayo *et al.*, 2005), there is still a need for accurate duration predictions.

In common with many other areas of speech and language processing, the databases used to train phone duration models are unbalanced. In the space of all possible combinations of linguistic factor values, only some are linguistically plausible and, of those, only a small fraction will actually be observed in any corpus. Of the observed feature vectors (these are vectors of linguistic factor values), many will be very rare – i.e. low in frequency. However, as was shown by van Santen (1994), the joint probability mass of all these rare vectors taken together is sufficiently large to mean that they cannot simply be neglected. In other words, in any individual sentence, it is very likely that we will encounter one or more of these rare vectors. Therefore, models of phone duration must be robust: they must predict appropriate durations for rare (and indeed previously unseen) vectors.

In addition, there exists a problem of factor confounding: different factors occur with unequal frequencies in the training database. As a result, *raw* durations calculated from the database can be deceptive. van Santen (1994) gives an example of within-word position and stress factor confounding. Durations of vowels turn out to be shorter in word-final syllables than in non-word-final syllables, if stressed and unstressed vowels are analysed together. But, unstressed vowels are shorter than stressed vowels and word-final syllables are five times more likely to be unstressed than stressed. So, if stressed and un-

stressed vowels are analysed separately, the vowel duration in final syllables (all other factors being equal) is longer than in non-final syllables, as we would expect.

The linguistic factors affecting a phone's duration interact with one another; the value of one or more factors may amplify or attenuate the affect of another factor. van Santen (1994) showed that these effects are easily predicted.

A robust model for predicting phone duration must address all of these issues. It should generalise well in order to successfully predict the duration of phones with rare (or previously unseen) feature vectors. It may be desirable to allow some factors to be unspecified or have ambiguous values; this would be the case if these factors' values are predicted by some other model which is not 100% accurate – for example, part of speech or features relating to the position of syllable boundaries.

We expect a duration model that properly accounts for factor interactions and confounding to be more accurate than a model that does not.

1.2 Linguistic factors influencing segment duration

1.2.1 Vowels

Umeda (1975b), Klatt (1975) and Crystal and House (1988a) cited in van Santen (1992) report that vowels in stressed syllables have longer durations than in unstressed syllables. Nooteboom (1972), Sluijter and van Heuven (1995), Turk and White (1999), Turk and Shattuck-Hufnagel (2000) report that syllables (and their vowels) in accented words are longer than in de-accented words. van Santen (1992) found interaction between stress and pitch accent: stressed vowels in accented words were significantly longer than non-stressed vowels; in de-accented words the difference was smaller but still noticeable. Word-initial stressed syllables get shorter as the number of syllables in the word increases (Lehiste, 1972; Klatt, 1973; Port, 1981). Stressed vowels in word-final syllables are longer than those in non-word-final syllables(Nooteboom, 1972; Oller, 1973). The last vowel in an utterance is longer than other vowels (Oller, 1973; Lehiste, 1973; Klatt, 1975, 1976; Wightman *et al.*, 1992).

A vowel's duration depends on voicing and manner of production of the following consonant (Peterson and Lehiste, 1960; Crystal and House, 1988b; van Santen, 1992). van Santen (1992) defined the "standard order" of postvocalic consonant classes arranged in order of increasing vowel duration: *voiceless stops, voiceless affricate, liquids, voiceless fricatives, nasals, voiced stops, voiced affricate*, and *voiced fricatives*. Given the same linguistic context (e.g. stress and accent status, phrasal position), different vowels vary in duration:

Factor	**# Values**	**Possible values**
frontness *Front*	3	front, mid, back
height *Height*	3	high, mid, low
length *Length*	4	short, long, diphthong, shwa
frontness-height *FH*	9	see Table 2
roundness *Rnd*	2	rounded, unrounded
stress *S*	2	stressed, unstressed
within-word position of syllable *Wpos*	3	initial, medial, final
within-utterance position of word *Utt*	3	initial, medial, final
following segment identity *Cpos*	10	voiceless stop, voiceless affricate, liquid, voiceless fricative, nasal, voiced stop, voiced affricate, voiced fricative, vowel, silence
word class *Wd*	2	function, content

Table 1
Linguistic factors chosen for predicting vowel duration. The encoding of *FH* is given in Table 2. Either *Front*, *Height* and *Length* OR *FH* are used, plus the other factors; these two systems are referred to as *F+H+L* or *FH-compound* .

e.g. /*oi*/ is more than twice as long as /*i*/ (van Santen, 1992) (Throughout the paper we used Machine Readable Phonemic Alphabet (MRPA) to represent phones as reported in Hiller *et al.* (1990).)

Durations of more frequent words tends to be shorter than those of less frequent words (Gregory *et al.*, 2001). Function words tend to be shorter than content words (Bell *et al.*, 2003).

Based on these findings, we selected linguistic (causal) factors for predicting vowel duration, shown in Table 1. We represent vowel identity as 2 factors (a compound frontness-height factor, roundness). This set of factors will be referred to as *FH-compound* .

In Goubanova (2005), we reported a second way of representing vowel identity, in which separate *Front*, *Height* and *Length* factors were used instead of *FH*.

	Frontness		
Height	front	mid	back
high	1	2	3
mid	4	5	6
low	7	8	9

Table 2
The encoding of the frontness-height compound factor *FH*.

This set of factors will be referred to as *F+H+L* . We did not use the previous segment identity because, in preliminary experiments, we found this had an insignificant effect. We also did without *Length* factor for reasons of data sparsity and to reduce the computational complexity when estimating BN model parameters.

1.2.2 Consonants

van Santen (1994) found that duration of intervocalic consonants (VCV) depends on the consonant's manner of production and voicing, with voiceless stops being the shortest and voiced fricatives being the longest. van Son and van Santen (1997) reports an interaction of manner of production and voicing: voiced fricatives were found to be the longest and voiced stops the shortest. Consonant constriction duration is longer in word-initial than in word-medial position (Oller, 1973; Cooper, 1991; Fougeron and Keating, 1997).

van Son and van Santen (1997) found interaction between within-word position, stress and consonant identity (represented as the prime articulator – labial, coronal, post-coronal). Primary stressed syllable duration is affected by word boundary position (Turk and Shattuck-Hufnagel, 2000) with consonants in word-initial primary stressed syllables being longer than other consonants. Previous and following vowel stress affects stop consonant durations (Umeda, 1977). Pre-stressed consonants are longer than others (Oller, 1973; Klatt, 1974; Umeda, 1975a). van Santen (1994) and van Son and van Santen (1997) also found a significant effect of stress and within-word position on intervocalic consonant duration. Haggard (1973) cited in Klatt (1976) reported consonants being shorter in clusters than in a CV environment. van Santen (1994) also found that duration of consonants in clusters is affected by the preceding and following segment identity, position relative to syllable boundary, stress of the previous and following vowels, and accent status of the word. Consonants in phrase-final syllables are longer than those in phrase-medial positions (Oller, 1973; Lehiste, 1973; Klatt, 1975, 1976; Wightman *et al.*, 1992).

From these findings, we selected 9 linguistic factors for predicting consonant duration, shown in Table 3. Consonant identity is represented by the com-

Factor	**# Values**	**Possible values**
manner-voice *MV*	9	see Table 4
within word position *Wpos*	3	initial, medial, final
stress *S*	2	stressed, unstressed
within utterance position *Utt*	3	initial, medial, final
syllabic position *Syl*	3	onset, coda, syllabic
previous segment identity *Cpre*	3	consonant, vowel, silence
following segment identity *Cpos*	3	consonant, vowel, silence
frontness of syllabic vowel *Front*	3	front, mid, back
number of syllables in word *NSyls*	5	1, 2, 3, 4, >4

Table 3
Linguistic factors chosen for predicting consonant duration. The encoding of *MV* is given in Table 4.

	Voicing	
Manner	unvoiced	voiced
stop	1	6
affricate	2	7
approximant	3	
fricative	4	8
nasal	5	
tap		6
lateral	9	

Table 4
The factor encoding of the manner-voice compound factor *MV*.

pound manner-voice factor *MV* shown in Table 4, following from the results in van Santen (1994) and van Son and van Santen (1997)

1.3 Phone vs Syllable modelling

There is an on-going debate of whether to use the phone, the syllable or a perception-based unit for duration modelling. Despite the importance of the syllable in the prosodic organisation of speech, we chose to use phones in our BN model because it has many advantages over syllables (Campbell and Isard, 1991) or interperceptual centre groups (Barbosa and Bailly, 1994).

First, there is a data sparsity issue. There are only 40 or so phonemes of English, but there are around 2,000 possible syllables that can be generated from them. In models that use syllables, it is therefore common to collapse the syllable inventory down to far fewer types (e.g. Campbell (1992)) but this is unsatisfactory. These syllable-based models assume that syllables occurring within the same prosodic and positional context (within-word, within-phrase position) have equal durations (*syllable mediation* hypothesis), regardless of their segmental content. They assume that syllable duration does not depend on the identity of the syllable's constituent phones. This is the so-called *segmental independence hypothesis* Shih and van Santen (2000), but this hypothesis is not supported by experimental evidence presented in, for example, Shih and van Santen (2000).

In TTS, durations predicted at the syllable level must usually be mapped down to the phone level. Because phones are not equally "elastic", this requires a model of how phone durations change to fit their parent syllable's duration Campbell and Isard (1991). On the other hand, models which predict phone durations directly can take into account the effects of various linguistic factors on phone durations, without a mapping via syllables.

1.4 Types of models used for duration prediction

The earliest models developed for predicting phone duration were systems of rules (e.g. Klatt (1976)). Despite being successfully implemented in the MITalk synthesiser described in Allen *et al.* (1987), rule-based models have a few problems. Rule-based models do not account for factor interaction. Instead, they rely on manual adjustment of the model's parameters, which eventually makes the analytical representation of the model very complex. The models do not explicitly account for various speaking styles, speaking rates, and dialect differences.

Corpus-based methods, in which a model is automatically learned from labelled speech data, include both non-parametric models such as CART (Breiman *et al.*, 1984; Riley, 1992; Dusterhoff *et al.*, 1999) and parametric models such as Sums-of-Products (SoP) (van Santen, 1992, 1994). We will use CART and

SoP baselines (Sections 2 and 3 respectively).

CARTs are easy to build and robust to labelling errors in the training data. However, when faced with rare or unseen feature vectors, CARTs must "back off" and cannot interpolate. "Backing off" means relaxing the requirement to match all elements of the feature vector and allowing one or more elements to be essentially ignored. The performance of CARTs degrades when the percentage of such vectors is too high, as was shown in van Santen (1994).

SoP models use methods such as "ordinal data analysis" (Coombs, 1964) and "axiomatic measurement" (Krantz *et al.* (1964) cited in van Santen (1994)), that allow the discovery of regular patterns in data. By discovering and modelling regularities in duration data, SoP models can interpolate in cases of rare or unseen feature vectors. SoP models are also robust to noise in the data. However, a major problem of SoP models is that the search for the best model is a tedious and time consuming process. The number of possible SoP models is hyper-exponential in the number of linguistic factors (i.e. dimension of the feature vector). That is, the number of possible SoP models is of the order of 2^{2^n}, where n is the number factors. This means that finding the best model requires expert intuition and heuristic search techniques, which makes SoP modelling both an art and a science.

Bayesian models allow for an intuitive, straightforward representation of the problem domain information. The model's structure as well as parameters are estimated from the data. Building and training a model may be a time consuming process, but once the model is built and trained it could be easily implemented in any real TTS system.

1.5 Databases

The data used in all experiments were derived from 3 Rhetorical[1] TTS voice databases: 2 RP English voices – *rjs* (male) and *lja* (female) – and 1 GA English voice – *erm* (male). Each database consists of a set of utterances which we divided into train (90%) and test (10%) sets by taking out every 10th utterance. These databases were designed and recorded specifically for unit-selection TTS and can be regarded as having good coverage (measured in terms of diphones in different linguistic contexts). The commercial TTS system, *rVoice*, using these voices was widely regarded as very high quality. Tables 5 and 6 give the sizes of the databases.

[1] Rhetorical Systems Ltd, now part of Nuance. Many thanks to Paul Taylor for access to this data.

Voice	Number of vowel tokens		
	Train	Test	Total
lja	35,348	3,876	39,224
rjs	88,997	9,766	98,763
erm	57,104	6,084	63,188

Table 5
The number of vowel tokens.

Voice	Number of consonant tokens		
	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 6
The number of consonant tokens.

1.6 Performance metrics

In order to test the performance of our baseline (CART and SoP) as well as Bayesian models we used 2 metrics: *sample correlation coefficient* between observed and predicted values of duration and *Root Mean Squared Error* (RMSE) in milliseconds (ms). The sample correlation coefficient is defined as (see for example, Lee (1997)):

$$r = \frac{\sum_{m=1}^{M}(d_m^{obs} - \bar{d}^{obs})(d_i^{pred} - \bar{d}^{pred})}{\sqrt{\{\sum_{m=1}^{M}(d_m^{obs} - \bar{d}^{obs})^2\}\{\sum_{i=m}^{M}(d_m^{pred} - \bar{d}^{pred})^2\}}} \tag{1}$$

where M is the size of the test set, d_m^{obs} is an observed duration of a phone in the m-th feature vector, $\bar{d}^{obs}$ is the mean observed duration across the test set; d_m^{pred} is a predicted duration of a phone in the m-th feature vector, $\bar{d}^{pred}$ is the mean predicted duration across the test set. We will refer to the sample correlation coefficient as *correlation coefficient*, or just *correlation*. For the RMSE we adopt a definition used in Bishop (1998):

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(d_m^{obs} - d_m^{pred})^2} \tag{2}$$

where M is the size of the test set, d_m^{obs} is the observed duration, and d_m^{pred} is the predicted duration in the m-th feature vector, respectively.

In order to compare different experimental conditions we used a paired t-test. The independent variables tested were the sample correlation coefficient r defined in Equation 1 and the *Root Mean Squared Error* RMSE defined in Equation 2. The null hypothesis H_0 is that there will be no difference between the models. H_0 is tested against the two-sided non-null hypotheses H_1 ($\bar{r} > 0$) or $\bar{r} < 0$) by calculating the t statistic (a significance level of $p < 0.01$, one-tailed, is assumed, unless otherwise stated).

2 Classification and Regression Trees

CART models are very well known, particularly for tasks in TTS. The reader is referred to Breiman *et al.* (1984) for an introduction and to Dusterhoff *et al.* (1999), Krishna *et al.* (2004), among others, for examples of applications in TTS.

2.1 Experiments

We used *Wagon*, which is part of the *Edinburgh Speech Tools Library* (Black *et al.*, 2003) to build a regression tree. *Wagon* uses the data variance times the number of feature vectors (the "impurity") to determine the best question to add at each node of the tree, using a standard greedy algorithm to grow the tree. The algorithm stops growing a branch when one of the following is true: all questions about all elements of the feature vector have been asked; all the feature vectors at the current leaf are identical; the number of data points after the next split would fall bellow a threshold (the "stop value"); the improvement in impurity after the next split would fall below a threshold.

In addition, in order to prune an over-trained tree with a small stop value, one can use *held-out* data: some of the data is taken off to be used for testing a tree. The tree is built using the training data; it is then pruned back to where it best matches the held-out data. The advantage of this approach is that it allows the stop value to vary through different parts of the tree depending on how good the prediction is when compared against the held-out data.

Because the tree-building algorithm automatically selects the best questions to ask about the feature vector, it is safe to use a large number of linguistic factors to describe phones in context, shown in Tables 7-8, even if some are not predictive of duration, or are redundant. This is in contrast to the CART and SoP models.

Two CART models (one for vowels, one for consonants) were trained on each

Name	**Example**
segment's duration (s)	0.056
segment's name	*/ax/*
type of a segment	*vowel*
syllabic feature	+
length	*shwa*
height	*low*
frontness	*back*
rounded	+
manner of production	*fricative*
place of articulation	*labial*
voicing	+
previous segment name	*/b/*
previous segment type	*consonant*
previous segment syllabic	−
previous segment length	*short*
previous segment height	*low*
previous segment frontness	*mid*
previous segment rounded	−
previous segment manner	*affricate*
previous segment place	*dental*
previous segment voicing	−
next segment name	*/l/*
next segment type	*consonant*
next segment syllabic	+
next segment length	*long*
next segment height	*mid*
next segment frontness	−
next segment rounded	0
next segment manner	*stop*
next segment place	*palatal*
next segment voicing	0

Table 7
Linguistic factors used to build CART duration model. Part 1.

of the three databases (*lja, rjs*, *erm*) separately. We used the z-score of the duration (i.e., duration values are shifted and scaled so that their distribution in the training data has zero mean and unit variance). Held-out data was used to choose the stop value and balance factor. We also experimented with different amounts of held-out data: 5%, 10%, 15%. The local maxima of the correlation – or minima of the root mean square error (RMSE) – on the held-out set were used to select the best stop value, balance factor and amount of held-out data. We then re-trained the CART using these values. Results on

Name	**Example**
segment in stressed syllable	*true*
previous segment in stressed syllable	*false*
next segment	*false*
number of segments in syllable	5
number of segments in previous syllable	3
number of segments in next syllable	2
segment phrase initial	*true*
segment phrase medial	*false*
segment phrase final	*false*
previous segment phrase initial	*true*
previous segment phrase medial	*false*
previous segment phrase final	*false*
next segment phrase initial	*false*
next segment phrase medial	*true*
next segment phrase final	*true*
segment word initial	*true*
segment word medial	*true*
segment word final	*false*
previous segment word initial	*false*
previous segment word medial	*false*
previous segment word final	*false*
next segment word initial	*false*
next segment word medial	*false*
next segment word final	*true*
segment's syllable position	*true*
previous segment's syllable position	3
next segment's syllable position	2
onset/nucleous/coda type	*O1*
frontness of syllabic vowel	*back*
number of syllables in word	3
word class	*content*

Table 8
Linguistic factors used to build CART duration model. Part 2.

the test data are shown in Table 9.

We found that a stop value of 10, between 5% and 15% held-out data and a balance factor of 10%-15% worked best for vowels. For consonants, these values were 10, 10% and 5%-10% respectively.

Phone Type	Correlation			RMSE (ms)		
	lja	rjs	erm	lja	rjs	erm
Vowels	0.86	0.88	0.89	26	23	27
Consonants	0.73	0.79	0.82	21	20	24

Table 9
The test sample correlation and RMSE results for the the baseline CART models for vowels and consonants.

3 Sums of Products

Sums-of-Products (SoP) models (van Santen, 1992) are general linear models which compute the duration of a phone using a sum of product terms. Each product term involves various linguistic factors. In contrast to CARTs, the linguistic factors must be carefully chosen. It is usual to predict the log of duration. A full explanation of SoP is beyond the scope of this article. In brief, a SoP model computes the duration of a segment thus:

$$\log(\text{duration}) = \sum_{l \in \mathcal{A}} \prod_{j \in \mathcal{B}_l} S_{lj}(x_j) \tag{3}$$

where set $\mathcal{A}$ is a set of summation terms, $\mathcal{B}_l$ is the set of product terms for the l-th summation term, S_{lj}, are "factor scales" which correspond to the weight on the contribution of the value of the j-th linguistic factor x_j.

$S_{lj}(\cdot)$ is a "factor scale" and is a function of the value of its argument (its argument being a linguistic factor). When the linguistic factor is discrete (which is the case throughout this article), then this function is implemented as a table, which has one entry for each possible value of the corresponding linguistic factor. The table entries are learnt from data.

As van Santen (1994) demonstrates, many of the duration models in the literature can be represented as sums of products: Lindblom and Rapp (1973), Coker *et al.* (1973), Klatt (1976) and Kaiki *et al.* (1990), to name a few.

3.1 Experiments

3.1.1 Vowels

For his original SoP model, van Santen (1992) selected the following linguistic factors: the vowel identity *V*, syllabic stress *S*, accent status of the word *A*, number of syllables to the word end *Wpos*, number of consonants preceding

Model	Correlation			RMSE		
	lja	rjs	erm	lja	rjs	erm
Vowels	0.71	0.72	0.70	25	28	32
Consonants	0.74	0.79	0.76	25	26	33

Table 10
The test sample correlation and RMSE results for the the baseline SoP models for vowels and consonants.

the vowel in the word *Wpre*, preceding and following segment identity *Cpre* and *Cpos*, and utterance position *Utt*. The model predicted vowel duration (zero-mean and unit-variance normalised before taking the log) as:

$$\begin{aligned}&\log(\text{duration}) = \\ &K_{1,1}(A) \times K_{1,2}(S) + K_{2,3}(V) + K_{3,4}(Cpre) \\ &+K_{4,6}(Wpre) + K_{5,7}(Wpos) + K_{6,5}(Cpos) \\ &\quad +K_{7,5}(Cpos) \times K_{7,8}(Utt)\end{aligned} \tag{4}$$

For our baseline SoP model, we used Equation 4 but without the accent status term, because our data were not labelled with phrasal accent information. Our stress factor S has only two possible 2 values, *stressed* and *unstressed*, because in our data, no distinction is made between primary and secondary stressed vowels.

In Goubanova (2005), instead of the factor V representing vowel identity we used factors based on 4 phonological features: frontness, height, roundness, length, but here we describe only the model with vowel identity based on 2 factors: compound frontness-height FH and roundness Rnd because this model performs just as well and is slightly simpler. The complete list of the factors that we used for our baseline SoP model for vowels is shown in Table 1 and the SoP formulation is:

$$\begin{aligned}&\log(\text{duration}) = \\ &\quad K_{1,1}(S) + K_{2,2}(FH) + K_{3,3}(Rnd) \\ &+K_{4,4}(Wpos) + K_{5,5}(Cpos) + K_{6,6}(Utt) + K_{7,7}(Wd)\end{aligned} \tag{5}$$

The model was trained using a standard least-squares method on the data introduced in Section 1.5. The test sample correlation and RMSE (ms) results are shown in Table 10.

3.1.2 Consonants

van Santen (1994) performed an analysis of consonants in different segment-level contexts and suggested using separate models for intervocalic consonants and for consonants in clusters, with different linguistic factors for each. For consonants in clusters, he made even finer distinctions based on the syllabic and phrasal position of the target consonant. Consequently, 4 different models were defined for intervocalic consonants, consonants in syllable onsets, phrase-medial codas, and phrase-final codas. Correlation values reported by van Santen (1994) range from 0.824 for consonants in phrase-medial codas to 0.907 for intervocalic consonants.

For simplicity, we used the same model for all types of consonant. As in van Santen (1994), this was a simple multiplicative model (which is additive in the log domain), with the factors defined in Table 3 and the SoP formulation:

$$\begin{aligned}&\log(\text{duration}) = \\ &\quad K_{1,1}(MV) + K_{2,2}(S) + K_{3,3}(Wpos) \\ &+K_{4,4}(Utt) + K_{5,5}(Syl) + K_{6,6}(Cpre) \\ &\qquad +K_{7,7}(Cpos) + K_{8,8}(Front)\end{aligned} \tag{6}$$

Again, the model was trained using a standard least-squares method, and used the same data sets as the vowel models. In a pilot study described in Goubanova and King (2005) and Goubanova (2005) we also built a model in which consonant identity was represented directly with a factor that had 24 possible values but found that the compound MV encoding performed better. Results are given in Table 10.

Although direct comparison with van Santen's results is not possible, not least due to differing data sets, our results are broadly in line with those from van Santen (1992) and van Santen (1994).

4 Bayesian Networks

In many statistical modelling problems, we wish to do computations with the joint probability distribution (JPD) of a number of variables. Doing such computations directly would involve a potentially very large joint probability table (if all variables are discrete). This is often infeasible either computationally or because insufficient data is available to reliably estimate all the entries in this table. One solution to this challenge is to factor the joint distribution into the product of a number of simpler conditional distributions (each with a much

smaller probability table). For example, factoring the JPD of three discrete variables $P(A, B, C)$ into $P(A|B)P(B|C)P(C)$ is making the assumption that A is conditionally independent of C, given B. The tables needed to represent $P(A|B)$, $P(B|C)$ and $P(C)$ will, in total, have fewer entries than the table needed for $P(A, B, C)$. Similar arguments apply to the case of continuous variables, or combinations of discrete and continuous variables.

Bayesian Networks are a graphical representation of such a factorisation. They are directed graphical models with one node for each variable; the nodes are connected into a directed acyclic graph. Each variable thus has incoming arcs from zero or more other variables: these are called the parents of that variable. Because they are graphical, BNs are intuitive to work with and give an easy-to-interpret representation of this factorisation in which the conditional independence assumptions inherent in the factorisation can be simply read off the graph structure: the JPD is $\prod_i P(V_i|Pa(V_i))$ where i indexes the variables in the graph, and $Pa(V_i)$ is the set of variables that are parents of V. Importantly, given its parents, V_i is independent of all other variables. This structure (i.e. the factorisation of the JPD) can be devised by hand, or learnt from data.

Given a BN, and settings for some of the variables, *inference* can be used to determine the most probable value(s) for the remaining variable(s). This is how we will use BNs for duration prediction: During learning, all variable values will be known (from the training data) and the model's parameters (conditional probability tables and conditional Gaussians) will be learnt; To make a prediction of duration, the values of the discrete linguistic variables will be set to known values and the most probable value of the continuous duration variable will be inferred. For background material on Bayesian Networks and the algorithms available for parameter learning and inference, the reader is referred to Heckerman (1995).

The differences between the models presented in the following sections are both in the way that the linguistic factors are encoded as variables in the graph, and in the structures of the graphs.

4.1 Learning

Although we will use an automatic procedure for learning the network topology, we still need to specify by hand what the variables in the BN will be. For vowels, we experimented with the two versions of the set of linguistic factors, already described in Section 1.2.1 and referred to as *FH-compound* and *F+H+L* . For consonants, the factor set described in Section 1.2.2 will be used.

For a given network structure (topology), standard algorithms are used for

Factor	**# Values**	**Example**
consonant identity *C*	24	*/ch/*
within-word position *Wpos*	3	initial
stress *S*	2	stressed
frontness *Front*	3	back
number of the syllables in word *NSyls*	5	2

Table 11
Linguistic factors chosen for the hand crafted 6-factor model.

learning the parameters (Dempster *et al.*, 1977). A more difficult problem is that of learning the network topology: that is, what are the parents of each variable? We used an algorithm called K2 (Cooper and Herskovits, 1992).

4.1.1 Learning Bayesian Network structure

The most common approach to network structure learning is to apply some heuristic search techniques to search through the hypothesis space of possible network structures and evaluate a scoring metric (function) for each candidate network. The network with the highest score is selected. To this end, there are two common scoring functions used in network structure learning algorithms: *Minimum Description Length* (e.g. Lam and Bachus, 1994) and the *Bayesian measure* used in the present research and described below.

The number of possible networks of n nodes is hyperexponential in n. Exhaustive search is infeasible, so we must resort to heuristic approaches. Based on the Bayesian measure, Cooper and Herskovits (1992) developed an algorithm called K2, for learning the structure of networks with *discrete* variables from data.

The K2 algorithm starts with an ordered list of the variables and an empty parent set for each variable. It successively adds parents to each variable (parents are chosen from earlier in the list, thus ensuring the graph remains acyclic) to maximally improve the Bayesian measure. It must be pointed out that the variable ordering greatly affects the quality of the network structure learnt: it is essential to provide a good node ordering, and this must be done using expert intuition. We always placed the duration variable last in the ordering: any variable could be a parent of duration, but duration is no other variable's parent.

Since K2 only works for discrete variables, the duration variable D must be discretized, which we do using the z-scores of the duration values by assigning them to evenly spaced bins. We experimented with 9 different levels of dis-

cretization (from 2 to 10 bins) and, with the three voices we are using, this results in 27 distinct data sets.

For example, 9 linguistic factors (*F+H+L*) have $9! = 40,320$ different orderings. Trying all of these for all 27 datasets is computationally infeasible. Instead, for each of the three voices in turn, we tried all possible levels of discretization for just one variable ordering (*Wpos, Front, Height, Length, Rnd, Wd, Utt, S, Cpos, D*), and all possible variable orderings for just one discretization level (5 bins). For the *FH-compound* model, the single ordering was (*Wpost, S, Utt, Cpost, FH, Round, Wd*) and for the consonant model it was (*MV, Wpos, S, Utt, Syl, Cpre, Cpos, Front*).

4.1.2 Uniqueness of Bayesian Network structures

K2 algorithm assigns the same score (Bayesian measure) to two or more DAGs with different node orderings given the same discretization level. For each discretization level, the DAGs found have the same duration node parent set and differ only in a few (equal to the number of node permutations) of the CPTs. After we found all the DAGs for all discretization levels, we then get rid of the identical DAGs (up to the arc reversal), then group the rest into equivalence classes based on the identity of the duration node parent set. This reduced the number of networks to 947 (*vowels, FHLR*), 590 (vowels, *FH-compound*) and 915 (consonants, *MV-compound*). All networks had the same value for the Bayesian measure so, in order to further reduce the number of networks to a feasible number for experimentation, we considered the set of variables that are parents of the duration variable. Sets of networks with the same parent set for duration were considered equivalent, and only one of the set was retained. Note that, in the case of all discrete variables being observed, all networks within such classes will predict the same distribution for the duration variable. As a result, there were 7 different vowel *F+H+L* networks, 4 vowel *FH-compound* networks and 8 consonant networks.

The duration variable D parent sets $\mathbf{Pa}(D)$ for the various resulting networks are shown in Tables 4.1.2 and 12. An example network of each type is shown in Figures 1, 2 and 3. The rest are shown in Goubanova (2005).

4.2 Results

4.2.1 Vowels

Table 13 presents the results for vowels. All models perform no worse than the CART models in terms of correlation ($p > 0.01$, insignificant). In addition, all *F+H+L* models significantly ($p < 0.001$) outperform both the SoP and

Name	**Pa(D)**	**# params**
Vowel-F+H+L-1	Cpos Length Round	80
Vowel-F+H+L-2	Cpos Front Length Rnd	240
Vowel-F+H+L-3	Cpos Front Height Length Rnd	720
Vowel-F+H+L-4	Cpos Front Height Length Wd	720
Vowel-F+H+L-5	Wpos S Cpos Rnd	120
Vowel-F+H+L-6	Wpos Cpos Length Rnd Wd	480
Vowel-F+H+L-7	Wpos Utt Cpos Front Height Length Wd	6480
Vowel-FH-compound-1	Cpos FH	90
Vowel-FH-compound-2	Cpos FH Rnd	180
Vowel-FH-compound-3	Wpos S Cpos Rnd	120
Vowel-FH-compound-4	Wpos S Utt Cpos FH Rnd Wd	6480
Consonant-1	*MV, Cpos*	27
Consonant-2	*MV, Syl, Front*	81
Consonant-3	*MV, Wpos, S, Syl, Cpre, Cpos, Front*	4374
Consonant-4	*MV, Wpos, S, Utt, Syl, Cpre, Cpos*	4374
Consonant-5	*MV, Wpos, S, Utt, Syl, Cpre, Cpos, Front*	13122
Consonant-6	*MV, Wpos, Syl, Cpre, Cpos*	729
Consonant-7	*MV, Wpos, Syl, Cpre, Cpos, Front*	2187
Consonant-8	*MV, Wpos, Utt, Syl, Cpre, Cpos, Front*	6561

Table 12
BNs learnt by the K2 algorithm. The number of parameters in the conditional Gaussians of the duration variable is shown in the third column.

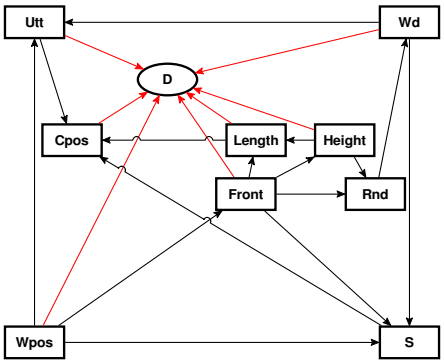


Fig. 1. A Bayesian network of size 10 learnt by the K2 algorithm, with vowel durations being uniformly discretized. Duration parent set $\mathbf{Pa}(D) = \{$ *Wpos, Utt, Cpos, Front, Height, Length, Wd* $\}$.

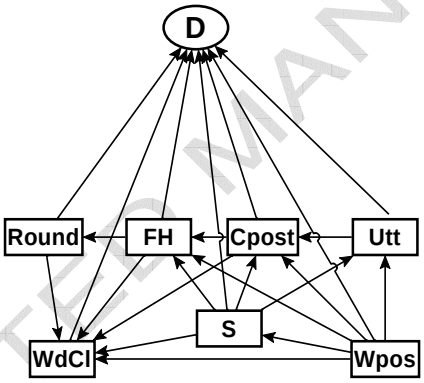


Fig. 2. A Bayesian network learnt by the K2 algorithm, with vowel durations being uniformly discretized. The duration D parent set $\mathbf{Pa}(D) = \{$ *Wpos, S, Utt, Cpos, FH, Rnd, Wd* $\}$.

CART baselines in terms of RMSE. In particular, the *Vowel-F+H+L-5* model significantly outperforms the SoP model in terms of correlation ($p < 0.01$) and RMSE ($p < 0.005$); the *Vowel-F+H+L-3* and *Vowel-F+H+L-4* models perform significantly better ($p < 0.005$) than both the SoP and CART models in terms of RMSE.

Hence, we chose the *Vowel-F+H+L-3* and *Vowel-F+H+L-4* models as the best models for the RP (*lja, rjs*) voices and the *Vowel-F+H+L-5* model as the best model for the GA (*erm*) voice.

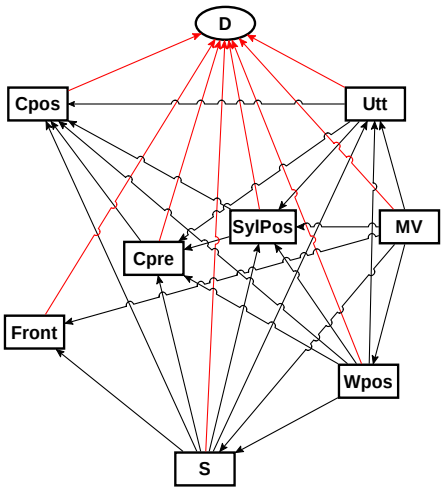


Fig. 3. Bayesian network learnt by the K2 algorithm, with consonant durations being uniformly discretized: *Consonant*-5 model for consonants.

Model	**Correlation**			**RMSE (ms)**		
	Voice					
	lja	**rjs**	**erm**	**lja**	**rjs**	**erm**
Vowel-F+H+L-1	0.83	0.80	0.67	2.3	2.2	2.6
Vowel-F+H+L-2	0.86	0.85	0.68	1.7	1.5	1.8
Vowel-F+H+L-3	**0.88**	**0.90**	0.68	2.5	2.4	2.7
Vowel-F+H+L-4	0.86	0.82	0.75	**1.5**	**1.5**	**1.7**
Vowel-F+H+L-5	0.83	0.84	**0.78**	1.9	1.6	1.8
Vowel-F+H+L-6	0.84	0.82	0.72	2.14	2.0	2.3
Vowel-F+H+L-1	0.82	0.84	0.60	2.54	2.4	2.7
Vowel-FH-compound-1	**0.84**	0.84	**0.81**	2.82	1.6	**2.8**
Vowel-FH-compound-2	0.84	0.84	0.81	**2.8**	**1.6**	**2.8**
Vowel-FH-compound-3	0.83	0.82	0.76	4.0	2.5	4.2
Vowel-FH-compound-4	0.84	**0.84**	0.80	4.1	2.4	4.7
SoP	0.71	0.72	0.70	25	28	32
CART	0.86	0.88	0.89	26	23	27

Table 13
Results for vowels. The maximum correlation / minimum RMSE values within each of the *F+H+L* and *FH-compound* groups of models, per voice, are shown in bold.

Model	Correlation			RMSE (ms)		
	Voice					
	lja	**rjs**	**erm**	**lja**	**rjs**	**erm**
Consonant-1	0.80	0.77	0.69	3.8	4.4	3.8
Consonant-2	0.73	0.76	0.67	5.1	5.6	5.1
Consonant-3	**0.84**	**0.80**	0.69	**3.5**	**4.1**	**3.6**
Consonant-4	0.72	0.74	**0.80**	4.6	5.1	4.5
Consonant-5	0.71	0.73	0.74	3.7	4.3	4.5
Consonant-6	0.80	0.74	0.75	4.6	5.2	4.6
Consonant-7	0.76	0.73	0.73	4.7	5.3	4.7
Consonant-8	0.56	0.49	0.75	3.5	4.1	3.7
SoP	0.74	0.79	0.76	25	26	33
CART	0.73	0.79	0.82	21	20	24

Table 14
Results for consonants. The maximum correlation / minimum RMSE values for the BN models, per voice, are shown in bold.

All *FH-compound* models perform significantly ($p < 0.01$) better than the SoP model and no worse ($p > 0.1$; insignificant) than the CART model in terms of correlation. In particular, the *Vowel-FH-compound*-1 and *Vowel-FH-compound*-2 models outperform the SoP model at a higher significance level of $p < 0.001$ in terms of correlation and RMSE. We therefore selected these two models as the best *FH-compound* models for all voices.

4.2.2 Consonants

Table 14 summarises results for consonants.

All consonant models perform no worse than the CART models in terms of correlation ($p > 0.05$, insignificant). They also significantly outperform both the SoP and CART baselines in terms of the RMSE ($p < 0.01$). In particular, the *Consonant*-3 model significantly outperforms both SoP and CART models in terms of RMSE ($p < 0.005$). The *Consonant*-4 model performs significantly better than the CART model in terms

For the two RP voices (*lja, rjs*), the *Consonant*-3 model is best; the parents of duration are all variables except for within-utterance position. For the GA voice (*erm*) the *Consonant*-4 model is best; the parents of the duration variable are all variables except the frontness of the syllabic vowel.

5 Conclusions and Future Work

We have demonstrated that Bayesian Network models can be successfully used to predict phone duration and that they outperform CART and SoP models. Building and training these models can be be time consuming but, once the model is trained, it is computationally very cheap to use for duration prediction, since it is essentially a look-up table. The BN structures found here could probably be used directly on other voice databases (i.e. relearn only the parameters, not the network structure), particularly for consonants.

5.1 Parents of the duration variable

As can be seen from Table 4.1.2, the *F+H+L* vowel models that performed best on RP voices (*Vowel-F+H+L*-3 and *Vowel-F+H+L*-4) had *Front*, *Height*, and *Length* as parents of duration *D*, whereas the best model for the GA *erm* voice, the *Vowel-F+H+L*-5 model, had the roundness *Rnd* variables as the parent of duration *D*. It appears that the best choice of parent variables is dialect dependent. It is also obvious that the *Length* variable is important, since 6 out of 7 models in Table 4.1.2 have this variable as a parent of the duration variable.

5.2 Effectiveness of the K2 algorithm

In an earlier study, we devised a model structure by hand (Goubanova and King, 2005; Goubanova, 2005). Overall, this model performed worse than the SoP and CART models. Almost all the models learnt by the K2 algorithm reported above perform significantly better than this hand-crafted model.

5.3 Limitations of the approach

5.3.1 Problem domain specification

Despite being successful overall, there were some limitations to predicting phone duration using Bayesian models. It turns out that for the GA *erm* voice, the models learnt from data demonstrated slightly lower performance in comparison to the results for the 2 RP voices: *lja, rjs*. This difference can not be explained by the training data size, since the training set for the *lja* voice contains $54,489$ consonant tokens, and for the *erm* voice $85,048$ tokens. The tendency of poorer performance on the *erm* voice occurs across different

models for both vowels and consonants. It may be that the linguistic factors we chose are appropriate for RP but not for GA. For GA vowels we may need extra factors, such as *Tense* (amount of articulatory effort). Since GA is rhotic, a factor accounting for this may have proved useful.

5.3.2 Number of model parameters

The number of parameters in the BN models is highly variable (Table 12) and can be quite large. The *Consonant*-5 model for consonants has 13,122 parameters in the conditional Gaussians alone, which is 70 times larger than the 196 parameters reported in van Santen (1994) for a SoP models for consonants. Parameter estimation is the most time-consuming part of the model training process and typically takes tens of hours of CPU time. Techniques exist for reducing the number of parameters: for example, replacing the full table of conditional Gaussians with a noisy-OR (Pearl, 1988), decision tree (Boutilier *et al.*, 1996), or default table (Friedman and Goldszmidt, 1996).

5.4 Future work

As mentioned above, van Santen used a number of different SoP models for different classes of segments. This would also be possible with BNs. A more sophisticated approach would be to automatically cluster phones into classes such that within a class, a single BN topology was the optimal one.

With BNs, duration predictions can still be made if the values of some variables are not specified (i.e. the variables are unobserved, or "hidden"); this is impossible with CART and SoP models. BNs can even be trained with some variables hidden. We have made a preliminary exploration of this possibility Goubanova (2005) and have demonstrated that good predictions can still be made.

References

Allen, J., Hunnicut, S., and Klatt, D. (1987). *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge.

Barbosa, P. and Bailly, G. (1994). Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, **15**, 127–137.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustic Society of America*, **113**(2), 1001–1024.

Bishop, C. (1998). *Neural Networks for Pattern Recognition*. Clarendon Press, Cambridge.

Black, A., Caley, R., King, S., and Taylor, P. (2003). Edinburgh Speech Tools Library: system documentation. Technical Report 1.2.0 edition, The Centre for Speech Technology Research, University of Edinburgh, UK.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). Context specific independence in baysian networks. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, Portland, Oregon, USA.

Breiman, L., Friedman, J., and Olshen, R. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove.

Campbell, N. (1992). Prosodic encoding of English speech. In *Proceedings of the 2nd International Conference on Spoken Language Processing*, Banff, Canada.

Campbell, W. and Isard, S. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, **19**, 37–47.

Clark, R., Richmond, K., and King, S. (2004). Festival 2 - build your own general purpose unit selection speech synthesiser. In *Proceedings 5th ISCA workshop on speech synthesis*, Pittsburgh, USA.

Coker, C., Umeda, N., and Browman, C. (1973). Automatic synthesis from ordinary English text. *IEEE Transactions on Audio and Electroacoustics*, **AU-21**, 293–298.

Coombs, C. (1964). *A theory of data*. Wiley, New York.

Cooper, A. (1991). Laryngeal and oral gestures in Englush /p, t, k/. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, volume 2, pages 50–53, Aix-en-Provance, France.

Cooper, G. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.

Crystal, T. and House, A. (1988a). Segmental durations in connected-speech signals: Syllabic stress. *The Journal of the Acoustical Society of America*, **83**(4), 1574–1585.

Crystal, T. and House, A. (1988b). Segmental durations in connected-speech sygnals: Current results. *The Journal of the Acoustical Society of America*, **83**(4), 1553–1573.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete date via the em algorithm. *Journal of the Royal Statistical Society*, **B 39**, 1–38.

Dusterhoff, K. E., Black, A. W., and Taylor, P. A. (1999). Using decision trees within the Tilt intonation model to predict F0 contours. In *CD-ROM Proceedings of Eurospeech 99*, Budapest, Hungary.

Fougeron, C. and Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, **101**(6), 3728–3740.

Friedman, N. and Goldszmidt, M. (1996). Learning Baysian networks with local structure. In *Proceedings of the 12th Conference on Uncertainty in*

Artificial Intelligence (UAI), Portland, Oregon, USA.

Goubanova, O. (2005). *Bayesian networks for predicting duration of phones*. Ph.D. thesis, University of Edinburgh.

Goubanova, O. and King, S. (2005). Predicting consonant duration with Bayesian belief networks. In *Proceedings of the Interspeech 2005*, volume 4, pages 1941–1944, Lisbon, Portugal.

Gregory, M., Bell, A., Jurafsky, D., and Raymond, W. (2001). Frequency and predictability effects on the duration of content words in conversation. *The Journal of the Acoustic Society of America*, **110**(5), 2738.

Haggard, D. (1973). Abbreviation of consonants in English pre- and post-vocalic clusters. *Journal of Phonetics*, **1**(1), 9–24.

Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Reserach, Microsoft Corporation, Redmond, USA.

Hiller, S., E.J., R., and Laver, J. (1990). Spell project speech stimuli:. Technical Report Staus report 1.1, The Centre for Speech Technology Research, University of Edinburgh, UK.

Kaiki, N., Takeda, K., and Sagisaka, Y. (1990). Statistical analysis for segmental duration rules in japanese. In *Proceedings of the 1st International Conference on Spoken Language Processing*, pages 17–20. Kobe, Japan.

Klatt, D. (1973). Interaction between two factors that influence vowel duration. *The Journal of the Acoustic Society of America*, **54**(4), 1102–1104.

Klatt, D. (1974). The duration of [s] in English words. *Journal of Speech and Hearing Research*, **17**, 51–63.

Klatt, D. (1975). Vowel lengthening is syntactically determined in connnected speech. *Journal of Phonetics*, **59**(3), 129–140.

Klatt, D. (1976). Linguistic uses of segmental duration of English: Acoustic and perceptual evidence. *The Journal of the Acoustic Society of America*, **59**(5), 1209–1211.

Krantz, D., Luce, R., Suppes, P., and Tversky, A. (1964). *Foundations of measurement*, volume 1. Wiley, New York.

Krishna, N., Tulukdar, P., Bali, K., and Ramakrishnan, A. (2004). Duration modelling for Hindi text-to-speech synthesis system. In *CD-ROM Proceedings of International Conference on Spoken Language Processing 2004*. Denver, USA.

Lam, W. and Bachus, F. (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence*, **10**, 269–293.

Lee, P. M. (1997). *Bayesian Satistics*. Arnold, Cambridge.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America*, **51**(6B), 2018–2024.

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *The Journal of the Acoustical Society of America*, **54**(5), 1228–1234.

Lindblom, D. and Rapp, K. (1973). Some temporal regularities of spoken

Swedish. In *PILUS*, volume 21, pages 1–59. Sweden.

Mayo, C., Clark, R., and King, S. (2005). Multidimensional scaling of listener responses to synthetic speech. In *Proceedings Interspeech 2005*, volume 4, pages 1725–1728. Lisbon, Portugal.

Nooteboom, S. (1972). *Production and perception of vowel duration*. Ph.D. thesis, University of Utrecht.

Oller, O. (1973). The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America*, **54**(5), 1235–1247.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Peterson, G. and Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, **32**, 693–703.

Port, R. (1981). Linguistic timing factors in combination. *The Journal of the Acoustical Society of America*, **69**(1), 262–273.

Riley, M. (1992). Tree-based modelling for speech synthesis. In G. Bailly, C. Benoit, and T. Sawallis, editors, *Talking Machines: theories, models and designs*, pages 265–273. Elsevier, Amsterdam, Netherlands.

Shih, C. and van Santen, J. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *The Journal of the Acoustical Society of America*, **107**(2), 1012–1026.

Sluijter, A. and van Heuven, V. (1995). Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in dutch. *Phonetica*, **52**, 71–89.

Strom, V., Clark, R., and King, S. (2006). Expressive prosody for unit-selection speech synthesis. In *Proceedings Intespeech 2006*, Pittsburgh, USA.

Tokuda, K., Zen, H., and Black, A. (2002). An hmm-based speech synthesis system applied to english. In *Proceedings 2002 IEEE Speech Synthesis Workshop*. Santa Monica, USA.

Turk, A. and Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, **28**(4), 397–440.

Turk, A. and White, L. (1999). Structural influenceson accentual lengthening in English. *Journal of Phonetics*, **27**(2), 171–206.

Umeda, N. (1975a). Another consistency in phoneme duration. *The Journal of the Acoustical Society of America*, **58**(S1), 62.

Umeda, N. (1975b). Vowel duration in American English. *The Journal of the Acoustical Society of America*, **58**(2), 435–445.

Umeda, N. (1977). Consonant duration in American English. *The Journal of the Acoustical Society of America*, **61**(3), 847–858.

van Santen, J. H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, **8**, 95–128.

van Santen, J. P. H. (1992). Contextual effects on vowel durations. *Speech Communication*, **11**, 513–546.

van Son, R. and van Santen, J. (1997). Strong interaction between factors influencing consonant duration. In *Proceedings of the Interspeech'97*, pages

319–322. Rhodes, Greece.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. (1992). Segmental durations in the vicinity of procodic phrase boundaries. *The Journal of the Acoustical Society of America*, **91(3)**, 1707–1717.