Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 301–311
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Bayesian networks for phone duration prediction

Olga Goubanova, Simon King *

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

Received 15 November 2006; received in revised form 9 July 2007; accepted 23 October 2007

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.

Keywords: Text-to-speech; Bayesian networks; Duration modelling; Sums of products; Classification and regression trees

1. Introduction

1.1. Duration modelling for text-to-speech synthesis

We present comparative experimental results for three 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

* Corresponding author. Tel.: +44 131 651 1725; fax: +44 131 650 6626.
E-mail addresses: ogoubanova@netscape.net (O. Goubanova), Simon.King@ed.ac.uk (S. King).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.10.002

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 unstressed 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 effect 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; Klatt, 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: e.g. /*oi*/is more than twice as long as /*i*/ (van Santen, 1992) (Throughout the paper we use Machine Readable Phonemic Alphabet (MRPA) to represent phones as described in (Hiller and Laver, 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 two 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*. 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 the *Length* factor for reasons of data sparsity and to reduce the computational complexity when estimating BN model parameters.

Table 1
Linguistic factors chosen for predicting vowel 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

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*.