Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 26 (2012) 274–292

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Two-stage phone duration modelling with feature construction and feature vector extension for the needs of speech synthesis ☆

Alexandros Lazaridis *, Todor Ganchev, Iosif Mporas, Evaggelos Dermatas, Nikos Fakotakis

Wire Communications Laboratory, Department of Electrical and Computer Engineering, University of Patras, 26500 Rion-Patras, Greece

Received 2 December 2009; received in revised form 18 August 2011; accepted 31 January 2012
Available online 24 February 2012

Abstract

We propose a two-stage phone duration modelling scheme, which can be applied for the improvement of prosody modelling in speech synthesis systems. This scheme builds on a number of independent feature constructors (FCs) employed in the first stage, and a phone duration model (PDM) which operates on an extended feature vector in the second stage. The feature vector, which acts as input to the first stage, consists of numerical and non-numerical linguistic features extracted from text. The extended feature vector is obtained by appending the phone duration predictions estimated by the FCs to the initial feature vector. Experiments on the American-English KED TIMIT and on the Modern Greek WCL-1 databases validated the advantage of the proposed two-stage scheme, improving prediction accuracy over the best individual predictor, and over a two-stage scheme which just fuses the first-stage outputs. Specifically, when compared to the best individual predictor, a relative reduction in the mean absolute error and the root mean square error of 3.9% and 3.9% on the KED TIMIT, and of 4.8% and 4.6% on the WCL-1 database, respectively, is observed.

Keywords: Feature construction; Phone duration modelling; Statistical modelling; Text-to-speech synthesis

1. Introduction

In a text-to-speech (TTS) system the accurate modelling and control of prosody leads to high quality synthetic speech. Prosody can be regarded as the implicit channel of information in the speech signal that conveys information about the expression of emphasis, attitude, assumptions and the affected state of the speaker that provides the listener clues to support the recovery of the verbal message (Huang et al., 2001). Prosody is shaped by the relative level of the fundamental frequency, the intensity and, last but not least, by the duration of the pronounced phones (Dutoit, 1997; Furui, 2000). The duration of the phones controls the rhythm and the tempo of speech (Yamagishi et al., 2008). Flattening the prosody in a speech waveform would result in a monotonous, neutral and toneless speech, without rhythm, sounding unnatural and unpleasant to the listener, and sometimes even scarcely intelligible (Chen et al., 2003). Thus, the accurate modelling of the duration of the phones is essential in speech synthesis, contributing to the naturalness of

☆ This paper has been recommended for acceptance by 'Martin Russell'.

* Corresponding author. Tel.: +30 2610 996496; fax: +30 2610 997336.
E-mail addresses: alaza@upatras.gr (A. Lazaridis), tganchev@ieee.org (T. Ganchev), imporas@upatras.gr (I. Mporas), dermatas@wcl.ee.upatras.gr (E. Dermatas), fakotaki@upatras.gr (N. Fakotakis).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2012.01.009

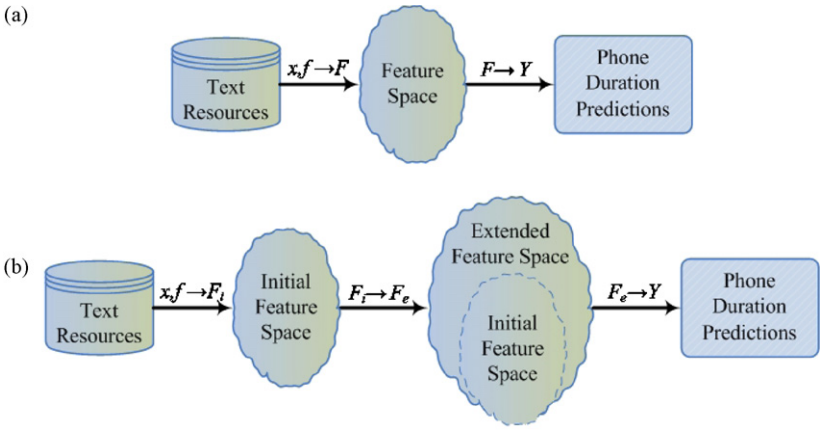

Fig. 1. Phone duration prediction: (a) the classical approach, (b) the two-stage approach involving a feature vector extension step.

synthetic speech and consequently to the quality of the speech (Huang et al., 2001; Yamagishi et al., 2008; Chen et al., 1998; Iwahashi and Sagisaka, 2000; Goubanova and King, 2008; Krishna and Murthy, 2004; Mobius and van Santen, 1996; Lee and Oh, 1999a; Klatt, 1976).

Various studies concerning phone duration modelling (Iwahashi and Sagisaka, 2000; Goubanova and King, 2008; Krishna and Murthy, 2004; Mobius and van Santen, 1996; Lee and Oh, 1999a,b; Klatt, 1976; Bartkova and Sorin, 1987; Simoes, 1990; Carison and Granstrom, 1986; Kohler, 1988; Epitropakis et al., 1993; Lazaridis et al., 2007, 2011; Teixeira and Freitas, 2003; Takeda et al., 1989; Riley, 1992; Chung, 2002; van Santen, 1994; Crystal and House, 1988; Gregory et al., 2001; Bell et al., 2003) have been made over the last few decades. The existing phone duration modelling methods are divided into two major categories: rule-based (Klatt, 1976; Bartkova and Sorin, 1987; Simoes, 1990; Carison and Granstrom, 1986; Kohler, 1988; Epitropakis et al., 1993) and data-driven methods (Iwahashi and Sagisaka, 2000; Goubanova and King, 2008; Krishna and Murthy, 2004; Mobius and van Santen, 1996; Lee and Oh, 1999a,b; Lazaridis et al., 2007, 2011; Teixeira and Freitas, 2003; Takeda et al., 1989; Riley, 1992; Chung, 2002; van Santen, 1994). The rule-based methods utilise manually produced rules which are extracted from experimental studies on large sets of utterances or are based on prior knowledge. The extraction of these rules requires linguistic expertise. One of the first and most well known attempts in the field of rule-based phone duration modelling is the one proposed in (Klatt, 1976) for the English language. In this method, rules based on linguistic and phonetic information, such as positional and prosodic factors, were used in order to predict the duration of the phones. These rules were derived by analysing a phonetically balanced set of sentences. Initially, a set of intrinsic values was assigned to each phone which was modified each time according to the extracted rules. Similar models were developed in other languages and dialects such as French (Bartkova and Sorin, 1987), Brazilian Portuguese (Simoes, 1990), Swedish (Carison and Granstrom, 1986), German (Kohler, 1988) and Greek (Epitropakis et al., 1993). The main disadvantage of the rule-based methods is the difficulty to represent and manually tune all the linguistic, phonetic and prosodic factors which influence the duration of the phones in speech. As a result, it is very difficult to collect all the appropriate (or even enough) rules without long-term commitment to this task (Klatt, 1987). Therefore, in order to be able to deduce the interaction among these factors and extract these rules, the rule-based duration models are restricted to controlled experiments, where only a limited number of contextual factors are involved (Rao and Yegnanarayana, 2007).

The creation of large databases made the development of data-driven methods for the task of phone duration modelling possible (Kominek and Black, 2003). Data-driven methods overcame the problem of manual rule extraction by employing machine learning techniques that automatically produce phonetic rules and construct duration models from large speech corpora. The classical approach, which can be summarised by the process shown in Fig. 1(a), relies on features extracted from a database which are then projected onto the phone duration space through a machine learning method. The main advantage of data-driven methods in comparison to rule-based methods is that this process significantly reduces the efforts (manual work) of linguists.

The present work was inspired by studies on processing, combining or transforming an initial feature set into a new set of features (Breiman et al., 1984; Matheus and Rendell, 1989; Pagallo and Haussler, 1990; Watanabe and Rendell, 1991; Flach and Lavrac, 2000; De Jong et al., 1992; Heath et al., 1993; Quinlan, 1993; Ragavan et al., 1993; Wenk and