Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 85–97
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Improving phone duration modelling using support vector regression fusion

Alexandros Lazaridis *, Iosif Mporas, Todor Ganchev, George Kokkinakis, Nikos Fakotakis

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

Received 30 October 2009; received in revised form 7 July 2010; accepted 21 July 2010

Abstract

In the present work, we propose a scheme for the fusion of different phone duration models, operating in parallel. Specifically, the predictions from a group of dissimilar and independent to each other individual duration models are fed to a machine learning algorithm, which reconciles and fuses the outputs of the individual models, yielding more precise phone duration predictions. The performance of the individual duration models and of the proposed fusion scheme is evaluated on the American-English KED TIMIT and on the Greek WCL-1 databases. On both databases, the SVR-based individual model demonstrates the lowest error rate. When compared to the second-best individual algorithm, a relative reduction of the mean absolute error (MAE) and the root mean square error (RMSE) by 5.5% and 3.7% on KED TIMIT, and 6.8% and 3.7% on WCL-1 is achieved. At the fusion stage, we evaluate the performance of 12 fusion techniques. The proposed fusion scheme, when implemented with SVR-based fusion, contributes to the improvement of the phone duration prediction accuracy over the one of the best individual model, by 1.9% and 2.0% in terms of relative reduction of the MAE and RMSE on KED TIMIT, and by 2.6% and 1.8% on the WCL-1 database.

Keywords: Duration modelling; Parallel fusion scheme; Phone duration prediction; Support vector regression; Text-to-speech synthesis

1. Introduction

In text-to-speech synthesis (TTS) there are two major issues concerning the quality of the synthetic speech, namely the intelligibility and the naturalness (Dutoit, 1997; Klatt, 1987). The former refers to the capability of a synthesized word or phrase to be comprehended by the average listener. The latter represents how close to the human natural speech, the synthetic speech is perceived. One of the most important factors for achieving intelligibility and naturalness in synthetic speech is the accurate modelling of prosody.

Prosody can be regarded as the implicit channel of information in the speech signal that conveys linguistic, paralinguistic and extralinguistic information related to communicative functions. Such functions are the linguistic functions of prominence (stress and accent), the phrasing, the discourse segmentation, the information about expression of emphasis, attitude, assumptions, the emotional state of the speaker, the information about the identify of the speaker (particular with respect to habitual factors). These functions provide to the listener clues supporting the recovery of the verbal message (Clark and Yallop, 1995; Laver, 1980, 1994). The accurate modelling and control of prosody in a text-to-speech system leads to synthetic speech of higher quality.

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

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

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.07.005

and the tempo of speech (Yamagishi et al., 2008) and the flattening of the prosody in a speech waveform would result in a monotonous, neutral, toneless and without rhythm synthetic speech, sounding unnatural, unpleasant to the listener or sometimes even scarcely intelligible (Chen et al., 2003). Thus, the accurate modelling of phones' duration is essential in speech processing.

Several areas of speech technology, among which TTS, automatic speech recognition (ASR) and speaker recognition benefit from duration modelling. In TTS, the correct segmental duration contributes to the naturalness of synthetic speech (Chen et al., 1998; Klatt, 1976). In hidden Markov model (HMM)-based ASR, state duration models improve the speech recognition performance (Bourlard et al., 1996; Jennequin and Gauvain, 2007; Levinson, 1986; Mitchell et al., 1995; Pols et al., 1996). Finally, significant improvement of the performance in the speaker recognition task was achieved by Ferrer et al. (2003), when duration-based speech parameters were used for the characterization of the speaker's voice.

Various approaches for segment duration modelling and many factors influencing the segmental duration have been studied in the literature (Bellegarda et al., 2001; Crystal and House, 1988; Edwards and Beckman, 1988; Riley, 1992; Shih and Ao, 1997; van Santen, 1994). The features related to these factors can be extracted from several levels of linguistic information, such as the phonetic, the morphological and the syntactic level. With respect to the way duration models are built, the duration prediction approaches can be divided in two major categories: the rule-based (Klatt, 1976) and the data-driven methods (Campbell, 1992; Chen et al., 1998; Lazaridis et al., 2007; Monkowski et al., 1995; Rao and Yegnanarayana, 2005; Riley, 1992; Takeda et al., 1989; van Santen, 1992).

The rule-based methods use manually produced rules, extracted from experimental studies on large sets of utterances, or based on previous knowledge. The extraction of these rules requires labour of expert phoneticians. In the most prominent attempt in the rule-based duration modelling category, proposed by Klatt (1976), rules which were derived by analyzing a phonetically balanced set of sentences, were used in order to predict segmental duration. These rules were based on linguistic information such as positional and prosodic factors. Initially a set of intrinsic (starting) values was assigned on each phone which was modified each time according to the extracted rules. Models of this type and similar to this were developed in many languages such as French (Bartkova and Sorin, 1987), Swedish (Carlson and Granstrom, 1986), German (Kohler, 1988) and Greek (Epitropakis et al., 1993; Yiourgalis and Kokkinakis, 1996), as well as in several dialects such as American English (Allen et al., 1987; Olive and Liberman, 1985) and Brazilian Portuguese (Simoes, 1990). The main disadvantage of the rule-based approaches is the difficulty to represent and tune manually all the linguistic factors, such as the phonetic, the morphological and the syntactic ones, which influence the segmental duration in speech. As a result, it is very difficult to collect all the appropriate (or even enough) rules without long-term devotion to this task (Klatt, 1987). Consequently the rule-based duration models are restricted to controlled experiments, where only a limited number of contextual factors are involved in order to be able to deduce the interaction among these factors and extract the corresponding rules (Rao and Yegnanarayana, 2007).

Data-driven methods for the task of phone duration modelling were developed after the construction of large databases (Kominek and Black, 2003). Data-driven approaches overcame the problem of the extraction of manual rules by employing either statistical methods or artificial neural network (ANN) based techniques which automatically produce phonetic rules and construct duration models from large speech corpora. Their main advantage is that this process is automated and thus significantly reduces the efforts that have to be spent by phoneticians.

Several machine learning methods have been used in the phone duration modelling task. The linear regression (LR) (Takeda et al., 1989) models are based on the assumption that among the features which affect the segmental duration there is linear independency. These models achieve reliable predictions even with small amount of training data but do not model the dependency among the features. On the other hand, decision tree models (Monkowski et al., 1995) and in particular classification and regression tree (CART) models (Riley, 1992), which are based on binary splitting of the feature space, can represent the dependencies among the features but cannot insert constraints of linear independency for reliable predictions (Iwahashi and Sagisaka, 2000). Another technique which has been used on the phone duration modelling task is the sums-of-products (SOP), where the segment duration prediction is based on a sum of factors and their product terms that affect the duration (van Santen, 1992, 1994). The advantage of these models is that they can be trained with a small amount of data. Bayesian networks models have also been introduced on the phone duration prediction task. These models incorporate a straightforward representation of the problem domain information and despite their time consuming training phase, they can make accurate predictions even when unknown values come across in some features (Goubanova and King, 2008; Goubanova and Taylor, 2000). Furthermore, instance-based algorithms (Lazaridis et al., 2007) have been used in phone duration modelling. In instance-based approaches the training data are stored and a distance function is employed during the prediction phase in order to determine which member of the training set is closer to the test instance and predict the phone duration. In a recent study (Yamagishi et al., 2008), the gradient tree boosting (GTB) (Friedman, 2001, 2002) approach was proposed for the phone duration modelling task as an alternative to the conventional approach using regression trees. The GTB algorithm is a meta-algorithm which is based on the construction of multiple regression trees and consequently taking advantage of them.