Speech Communication xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Intonation modelling using a muscle model and perceptually weighted matching pursuit

Pierre-Edouard Honnet[a], Branislav Gerazov[b], Aleksandar Gjoreski[b], Philip N. Garner[*,a]

[a] Idiap Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592, Martigny, 1920 Switzerland
[b] Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University– Skopje, Skopje, Macedonia

ARTICLE INFO

Keywords:
Intonation modelling
Matching pursuit
Physiology
Weighted correlation
Text-to-speech synthesis

ABSTRACT

We propose a physiologically based intonation model using perceptual relevance. Motivated by speech synthesis from a speech-to-speech translation (S2ST) point of view, we aim at a language independent way of modelling intonation. The model presented in this paper can be seen as a generalisation of the command response (CR) model, albeit with the same modelling power. It is an additive model which decomposes intonation contours into a sum of critically damped system impulse responses. To decompose the intonation contour, we use a weighted correlation based atom decomposition algorithm (WCAD) built around a matching pursuit framework. The algorithm allows for an arbitrary precision to be reached using an iterative procedure that adds more elementary atoms to the model. Experiments are presented demonstrating that this generalised CR (GCR) model is able to model intonation as would be expected. Experiments also show that the model produces a similar number of parameters or elements as the CR model. We conclude that the GCR model is appropriate as an engineering solution for modelling prosody, and hope that it is a contribution to a deeper scientific understanding of the neurobiological process of intonation.

1. Introduction

We are interested generally in speech to speech translation (S2ST). At the time of writing, S2ST is becoming a reality; with both research (e.g., the U-STAR consortium[1]) and commercial (e.g., Skype[2]) systems being available. This is a consequence of the component technologies — automatic speech recognition (ASR), machine translation (MT) and text to speech synthesis (TTS) — becoming quite mature.

In the context of ASR, especially when the goal is to produce text, prosody is normally ignored. By contrast, in the context of TTS, production of appropriate prosody is necessary to approach the naturalness of human speech. Although some applications using TTS do not necessarily require a human sounding voice, many of them would be more attractive if the machine — or communication intermediary — was able to produce natural sounding speech.

In the case of S2ST, not only is a natural voice required, but also one that conveys the intent and nuances of the speaker. This includes the ability to correctly emphasise the words, or groups of words, according to what has been said in the source language. Of course, this places requirements on the MT component, be it a simple mapping or something more complex (Do et al., 2015; Anumanchipalli et al., 2012).

In the present study, we focus on intonation modelling. Intonation modelling can be seen as finding a "good" representation of the intonation signal. The challenges are then: *What should be this representation?* and, *How can its parameters be extracted?*

We recently proposed a model which can be both extracted from a speech signal and recreated in a synthetic speech signal (Honnet et al., 2015). The model is physiologically based and can be seen as a generalisation of the CR model, although differing in some aspects in its definition. Inspired by the work of Kameoka et al. (2010) on the prediction of the CR parameters, we define local components of intonation as impulse responses to critically damped systems. Our first approach consisted of extracting parameters with a standard matching pursuit algorithm, followed by a selection of extracted atoms based on their perceptual relevance. This work was concerned with minimising the reconstruction error and investigating different system orders for the model components.

In a second iteration (Gerazov et al., 2015), the perceptual relevance was integrated directly in the extraction process by modifying the cost function of the matching pursuit algorithm, yielding optimal

* Corresponding author.
E-mail addresses: pierre-edouard.honnet@idiap.ch (P.-E. Honnet), gerazov@feit.ukim.edu.mk (B. Gerazov), aleksandar@gjoreski.mk (A. Gjoreski), Phil.Garner@idiap.ch (P.N. Garner).

[1] http://www.ustar-consortium.com.
[2] https://www.skype.com/en/features/skype-translator/

http://dx.doi.org/10.1016/j.specom.2017.10.004
Received 7 November 2016; Received in revised form 9 August 2017; Accepted 20 October 2017

local decomposition with respect to the perceptual measure used. The ability of the model to reach high perceptual similarity was investigated, and a comparison with the standard CR model was proposed, using a perceptually relevant objective measure.

Both approaches were validated on a rather small but multilingual dataset. In the present paper, we take the opportunity to consolidate our previous work, giving a more in-depth description of the model with a discussion on its physiological credibility; a detailed procedure for extracting parameters and a comparison with the standard CR model are also presented. Additionally, we present a more thorough evaluation, done on a much larger scale with a variety of speakers and languages. Some differences in the cost function for extraction are also introduced in Section 4.4.

In the following sections, after a review of background material, we expand the motivation for our model in terms of muscle modelling and put it in the context of the CR model. We go on to describe how the extraction can be done automatically, and in terms of perceptual metrics known to the linguistics community. Experiments are presented that evaluate the plausibility of the model and place it in the context of the state of the art.

2. Background

The need for correct intonation in TTS systems as well as the more general study of intonation have motivated the creation of different intonation and / or prosody models. In the context of TTS, adaptive systems — almost exclusively statistical parametric speech synthesis (SPSS) — are of great interest in the research community. The current state of the art systems for SPSS are based on hidden Markov models (HMMs) of Tokuda et al. (2002b) and Zen et al. (2009). HMM-based speech synthesis deals with intonation in a framewise manner; each frame from the training speech database has a value — or a null value in the case of an unvoiced frame — and HMM states are trained using these values. At synthesis time, F_0 is generated frame by frame, based on the HMM parameters.

Decision trees allow clustering of different features using different tree structure, thus one can expect that when clustering contextual features with respect to F_0, suprasegmental information in the label will have more impact than segmental information. However, this results in a speech often qualified as “flat” or lacking expressivity, which is due to the oversmoothing of HMMs (Toda and Tokuda, 2005).

There are three main ways of tackling the flatness of HMM-based synthesis at the intonation level: *i*) use a different representation of F_0 in the HMMs, *ii*) postprocess the synthetic intonation coming from HMMs, or *iii*) use an external prosody model that combines with other HMM parameters.

In the early stages of HMM-based synthesis, a multi-space probability distribution (MSD)-HMM was developed by Tokuda et al. (2002a) and became a standard way of handling the fact that speech can be voiced or unvoiced. More recently, some work was done using continuous F_0 and it was shown that continuous F_0 improves the perceived naturalness of synthesis (Yu and Young, 2011; Latorre et al., 2011). This was further improved by hierarchical modelling using a continuous wavelet decomposition to separate the different levels of variation in F_0 (Suni et al., 2013). In this work, the authors exploit the multi stream architecture of an HMM-based TTS framework to cluster these different temporal scale components with different decision trees.

In the second category, an example of what can be done to improve the output of HMM synthesis is given by (Hirose et al. 2011, 2012). Based on the command response (CR) model of Fujisaki and Nagashima (1969), the idea is to estimate the F_0 model commands from linguistic information, and then optimise them according to the F_0 generated by HMMs. By modifying the estimated parameters, it becomes possible to increase the expressivity of the synthetic speech. Another attempt to integrate the CR model in HMM-based TTS was made by Hashimoto et al. (2012), where parameterised F_0, in respect to the CR model, was used for training the HMM intonation features. This improved the quality of the synthetic speech as the model smoothed the F_0 contour before training.

The external prosody models, or intonation models are numerous. They can roughly be divided into models that: *i*) model the surface pitch contour, and *ii*) integrate the underlying physiological mechanisms of pitch production. Most intonation models fall into the first group. The Tone and Break Indices (ToBI) model (Silverman et al., 1992) is not a true surface model, nor is it a physiological one. It is linguistically focused, but is underdetermined and contains annotation and pitch synthesis ambiguities. On the other hand, the Tilt model (Taylor, 2000) is specially tailored for automatic parameter extraction and pitch synthesis. It describes the pitch contour as a sequence of events with specific shapes that can be automatically extracted with an obvious resynthesis step. The INSINT (INternational Transcription System for INTonation) model (Hirst et al., 2000) expands on ToBI and allows for automatic parameter extraction. It models the MOMEL (MOdélisation de MELodie) stylised (Hirst and Espesser, 1993) intonation contour as a sequence of specific F_0 target points. The General Superpositional Model of Intonation (Van Santen and Möbius, 2000), models the pitch contour decomposing it into a sum of a microprosodic segmental perturbation, an accent and a phrase curve. Finally, the Superposition of Functional Contours (SFC) model (Bailly and Holm, 2005), is a data driven approach based on the superposition of intonation prototypes that are directly linked to linguistic information through the use of neural networks.

Only a few models actually try to explain the intonation by investigating its production aspect. The most popular model in this category is the command response (CR) model of Fujisaki and Nagashima (1969). This model decomposes the intonation into additive physiologically meaningful components. The CR model is attractive for two reasons:

1. it has a physiological explanation which tries to account for the underlying mechanisms behind intonation production, and
2. it has a mathematical form, which makes it possible to parameterise.

Extracting the model parameters from an F_0 contour is not trivial, but the opposite resynthesis operation is straightforward.

The qTA (quantitative Target Approximation) model (Prom-on et al., 2009), expands on the CR model, and uses pitch targets as input to the physiological model of pitch production. The StemML (Kochanski et al., 2003), on the other hand, imposes physiological constraints of smoothness and communication constraints specified by target accent templates to the modelling process.

3. Physiologically based intonation modelling

3.1. Motivation

In mimicking the abilities of humans in a machine, it is natural to try to mimic human physiological processes. It is certainly not necessary; this is evidenced by the fact that there are many speech recognition and synthesis methods that use physiologically implausible mechanisms (such as Markov models and windowed frames). However, doing so has two attractive possibilities: The first is the main goal of technological advancement; the second is one of scientific understanding of the underlying processes.

Further, it is clear that there are no fundamental differences between speakers of different languages. We may hence reasonably expect a physiological model not to be language dependent.

3.2. Sources of physiological variation in F_0

A detailed analysis of intonation production is given by Strik (1994). In this work, using electromyographic (EMG) recordings