Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 1105–1126

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Two-stage intonation modeling using feedforward neural networks for syllable based text-to-speech synthesis☆

V. Ramu Reddy *, K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India

Received 31 March 2012; received in revised form 8 February 2013; accepted 9 February 2013
Available online 21 February 2013

Abstract

This paper proposes a two-stage feedforward neural network (FFNN) based approach for modeling fundamental frequency (F_0) values of a sequence of syllables. In this study, (i) linguistic constraints represented by positional, contextual and phonological features, (ii) production constraints represented by articulatory features and (iii) linguistic relevance tilt parameters are proposed for predicting intonation patterns. In the first stage, tilt parameters are predicted using linguistic and production constraints. In the second stage, F_0 values of the syllables are predicted using the tilt parameters predicted from the first stage, and basic linguistic and production constraints. The prediction performance of the neural network models is evaluated using objective measures such as average prediction error (μ), standard deviation (σ) and linear correlation coefficient ($\gamma_{X,Y}$). The prediction accuracy of the proposed two-stage FFNN model is compared with other statistical models such as Classification and Regression Tree (CART) and Linear Regression (LR) models. The prediction accuracy of the intonation models is also analyzed by conducting listening tests to evaluate the quality of synthesized speech obtained after incorporation of intonation models into the baseline system. From the evaluation, it is observed that prediction accuracy is better for two-stage FFNN models, compared to the other models.

Keywords: Intonation models; Prediction accuracy; Text-to-speech synthesis; Feedforward neural networks; Linguistic constraints; Production constraints; Positional; Contextual; Phonological; Articulatory; F_0 of syllable; Tilt

1. Introduction

Prosody plays an important role in improving the quality of text-to-speech synthesis (TTS) system both in terms of naturalness and intelligibility. Prosody refers to duration, intonation and intensity patterns of speech for the sequence of syllables, words and phrases. In this work, we focus on modeling one of the important prosodic parameters i.e., intonation. Intonation plays an important role in human speech communication. Intonation can be defined as the dynamics of fundamental frequency (F_0) contour over time, caused due to vocal folds vibration. The perceptual correlate of F_0 is pitch. Human hearing system is highly sensitive to variations in pitch (Moore, 1989). In speech synthesis, intonation directly affects the overall quality of the synthetic speech. From the speaker's view point, intonation can be used to convey pragmatic and emotional information. Different intonation patterns of syntactically similar sentences can convey dramatically distinct information. Intonation patterns are influenced by various factors, while analyzing

☆ This paper has been recommended for acceptance by B. Moebius.

* Corresponding author. Tel.: +91 7872381435.
E-mail addresses: ramu.csc@gmail.com (V.R. Reddy), ksrao@iitkgp.ac.in (K.S. Rao).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.02.003

the speech at different levels (O'Shaughnessy, 1984). At the lowest level, F_0 (micro-intonation) is affected by local segmental factors which are caused by dynamics of human speech production process. At a higher level, stress patterns, rhythm and melody affect the F_0 contour. The F_0 contour is also affected by attitude, gender, physical and emotional state of the speaker. From the listener's view point, intonation plays an important role in (i) resolving syntactic ambiguity, (ii) segmentation of utterances, (iii) speech perception during noisy environments, and (iv) perceiving the emotional state of the speaker (O'Shaughnessy, 1987). In addition to the above functions, pitch contours also carry the lexical meaning in tonal languages, such as Mandarin Chinese. The speech synthesized without intonation appears to be highly monotonous and robotic, and is not pleasant for listening over longer durations. Hence, while developing speech synthesis systems, acquisition and incorporation of the intonation knowledge is very much essential.

The implicit knowledge of intonation is usually captured by using modeling techniques. In this work, we propose a two-stage intonation model using feedforward neural networks (FFNN) for predicting the intonation patterns of the sequence of syllables. Neural networks are known for their ability to capture the underlying interactions that exist between input and output features (Haykin, 1999; Yegnanarayana, 1999). Neural networks also have the generalization ability to predict intonation patterns reasonably well for the patterns which are not present in the learning phase (Haykin, 1999). In speech signal, the intonation of each unit is dictated by the linguistic and production constraints of the unit (Kumar, 1993; Rao and Yegnanarayana, 2009). In this study, linguistic and production constraints are used to predict the F_0 values of the sequence of syllables. Linguistic constraints are represented by positional, contextual and phonological features, and production constraints are represented by articulatory features. In addition to the above features, tilt parameters are also used to capture the true shape of the intonation patterns. The prediction accuracy of the intonation models is further improved by imposing the other prosodic constraints represented by duration and intensity values of the syllables.

The paper is organized as follows: Section 2 presents an overview of the existing research on acquisition of intonation knowledge using different models. The speech database used for the development of baseline TTS system, and the performance of the intonation models used in the baseline TTS system are discussed in Section 3. Section 4 describes the proposed features used for predicting the F_0 values. Performance of the neural network models along with that of Linear Regression (LR) and Classification and Regression Tree (CART) models is given in Section 5. The performance of the proposed two-stage intonation model using tilt parameters is discussed in Section 6. Section 7 presents the prediction accuracy of the intonation models using individual features. The influence of the other prosodic constraints for predicting the intonation patterns is discussed in Section 8. Summary of this paper is presented in Section 9.

2. Literature review

Many methods have been developed for generation of F_0 contours to build successful TTS systems. In the last 20 years, two major approaches have emerged for modeling intonation: (i) the tone sequence approach which follows the traditional phonological description of intonation and (ii) the superposition approach (Botinis et al., 2001).

Phonological (tone sequence) models interpret F_0 contour as a linear sequence of phonologically distinctive units (tones or pitch accents), which are local in nature. There is no interaction of events in the F_0 contour with each other. Popular phonological models include Pierreihumbert's model and TOBI (Tone and Break Indices). Tone sequence intonation model was initially developed by Pierrehumbert (1980) for American English. The original model was extended by Beckman and Pierrehumbert (1986), and it evolved as Tone and Break Indices (ToBI) for transcribing intonation of American English (Silverman et al., 1992). Tone sequence models have been implemented for English, German, Chinese, Navajo and Japanese (Sproat, 1998; Jilka et al., 1999). Phonological models do not properly represent actual pitch variations. No distinction is made on the differences in tempo or acceleration of pitch movements. The temporal information is also not modeled. Phonological models are not easily ported from one language to another, since the inventory of categories must be thoroughly reviewed by linguistic experts (Buhmann et al., 2000).

Due to availability of large speech corpora, more acoustic-phonetic models have been proposed in recent years for text-to-speech systems. Acoustic-phonetic models are developed using acoustic data. Intonation model for Danish language was developed by Gronnum which is conceptually quite different from the tone sequence model (Gronnum, 1992, 1995). The model is hierarchically organized and includes several simultaneous, non-categorical components of different temporal scopes. The components are *layered*, i.e., a component of short temporal scope is superimposed on a component of a longer scope. Gårding (1983) also developed an intonation model which analyzes the intonation contour of an utterance as the result of the effects of several factors. Acoustic-phonetic (superposition or overlay)