Available online at www.sciencedirect.com
SCIENCE DIRECT
Speech Communication 48 (2006) 1057–1078
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Tone-Group F_0 selection for modeling focus prominence in small-footprint speech synthesis

Gerasimos Xydas, Georgios Kouroupetroglou *

University of Athens, Department of Informatics and Telecommunications, Division of Communication and Signal Processing, Panepistimiopolis, Ilisia, GR-15784 Athens, Greece

Received 13 July 2005; received in revised form 28 December 2005; accepted 1 February 2006

Abstract

This work targets to improve the naturalness of synthetic intonational contours in Text-to-Speech synthesis through the provision of prominence, which is a major expression of human speech. Focusing on the tonal dimension of emphasis, we present a robust unit-selection methodology for generating realistic F_0 curves in cases where focus prominence is required. The proposed approach is based on selecting Tone-Group units from commonly used prosodic corpora that are automatically transcribed as patterns of syllables. In contrast to related approaches, patterns represent only the most perceivable sections of the sampled curves and are encoded to serve morphologically different sequence of syllables. This results in a minimization of the required amount of units so as to achieve sufficient coverage within the database. Nevertheless, this optimization enables the application of high-quality F_0 generation to small-footprint text-to-speech synthesis. For generic F_0 selection we query the database based on sequences of ToBI labels, though other intonational frameworks can be used as well. To realize focus prominence on specific Tone-Groups the selection also incorporates a level indicator of emphasis. We set up a series of listening tests by exploiting a database built from a 482-utterance corpus, which featured partially purpose-uttered emphasis. The results showed a clear subjective preference of the proposed model against a linear regression one in 75% of the cases when used in generic synthesis. Furthermore, this model provided ambiguous percept of emphasis in an experiment featuring major and minor degrees of prominence.

Keywords: Text-to-speech synthesis; Tone-Group unit-selection; Intonation and emphasis in speech synthesis

1. Introduction

Emphasis is essentially the use of language that humans employ in order to bring to prominence selective parts of speech and mainly convey non-lexical and pragmatic information. It primarily signals contrast (contrastive focus), distinction between new and given information (focus as the missing variable in a proposition), meaning pronunciation and mood or other emotions. Generally, it points

Abbreviations: HRG, Heterogeneous Relation Graph; LR, Linear Regression; MPE, Mean Perceived Emphasis; NLP, Natural Language Processing; TI, Tone Item; TG, Tone Group; TGS, Tone-Group Selection; TtS, Text-to-Speech.

* Corresponding author. Address: University of Athens, Efkalypton 39, Agia Paraskevi, GR-15342 Athens, Greece. Tel.: +30 2107275305; fax: +30 2106018677.

E-mail addresses: gxydas@di.uoa.gr (G. Xydas), koupe@di.uoa.gr (G. Kouroupetroglou).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.02.002

out the most important parts in an utterance. Humans use a collection of different prosodic aspects to denote emphasis when they speak. The most common are pause insertions before and after the emphasized words, duration stretching, intensity, and substantial pitch rate change. The latter has proven to be the most significant factor for the perception of prosody ('t Hart et al., 1990; d'Alessandro and Mertens, 1995; Xub and Sun, 2002).

Human speech communication is emphasized by its nature. Most sentences have at least one focus and this is something that is partially ignored in most prosody modeling works, providing "neutral" or "generic" coverage in preliminary prototypes; however this is not the case in real speech. One of the drawbacks of Text-to-Speech (TtS) synthesis that leads to monotonous prosodic cues is the lack of focus prominence over the corresponding segments of speech. Therefore, emphasis modeling and provision is a mean to increase the expressiveness and thus naturalness of synthetic speech.

A TtS synthesis system mainly consists of two components (Dutoit, 1997; Sproat, 1998): the natural language processing (NLP) and the signal processing. The first one deals with the text-to-prosody part, providing the latter with sufficient segmental and prosodic information to generate an appropriate acoustic signal that "resembles human speech well enough for the human brain to interpret it as such" (Clark, 2003). The generation of the prosodic structure is derived in the synthesis chain from higher-level linguistic analysis of utterances carried by the NLP component. To represent this specification, several intonational frameworks have been proposed by linguists as well as engineers, ranging from qualitative (e.g. ToBI (Silverman et al., 1992)) to quantitative (e.g. Tilt (Taylor, 2000; Dusterhoff and Black, 1997)). They model intonation in terms of segmental anchoring and type, as for example, which syllables deserve a pitch accent and what value, type or shape should that accent be of. To incorporate this intonational description in the acoustic signal, the F_0 modeling component generates a continuous pitch curve from these events (location and type of accent). The resultant degree of naturalness of the synthetic pitch is closely related to the quality of the events. F_0 modeling is of great importance in any signal processing approach, from formant synthesis (defining the F_0 parameter) and diphone-based concatenative synthesis (defining pitch modifications) to unit-selection synthesis, as prosody selection is also of significant factor in the latter (Campbell, 1994).

The rule-based F_0 generation approaches have given place to machine learning ones. The most commonly used statistical method is the Linear Regression (LR) (Black and Hunt, 1996). This offers reasonable pitch generation, especially when the input conditions match the training ones. Objective evaluations have reported correlation between the training and the observed data from 0.6 to 0.8 in generic conditions (Black and Hunt, 1996; Xydas et al., 2005). On the other hand, subjective experiments usually contrast with the good statistical results, as prosody is usually judged as adequate but rarely natural. The modified suprasegmental structure of utterances and the lack of prominence seem to affect the naturalness of the delivered prosody, as well as the normalization of timing during pitch alignment. To overcome this problem, corpus-based F_0 models have been proposed and the recent related research focuses on optimizing (a) model's design in order to achieve adequate data coverage within reasonably sized databases (Black and Lenzo, 2003; Schweitzer et al., 2003) and (b) selection algorithms that not only minimize joining costs but also reveal the semantics of prosody (Bulyko and Ostendorf, 2001; Quazza et al., 2001; Wightman et al., 2000).

1.1. Corpus-based F_0 modeling

Following the natural effects on the segmental quality of corpus-based speech synthesis (Hunt and Black, 1996), corpus-based F_0 modeling (Huang et al., 1996; Malfrere et al., 1998; Meron, 2001; Raux and Black, 2003) attempts to maintain the suprasegmental structure intact thus achieving finest tonal representation. The minimization of the concatenation cost between jointed units affects the overall smoothness of the contour, based on the available inventory and the selection algorithm. However, natural curves are preserved at least over the range of each selected unit. In each case, the delivered speaking style originates and hardly deviates from that of the original human speaker.

In (Huang et al., 1996) the intonation cues of a group of consecutive syllables that form a clause, constitute an F_0 template. The template database is constructed in such a way so that it includes only one instance of each template. In (Malfrere et al., 1998), a sequence of successive words ending in a content word forms a pattern (intonational group).