Available online at www.sciencedirect.com
SCIENCE DIRECT
Speech Communication 48 (2006) 941–956
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems

Francisco Campillo Díaz, Eduardo Rodríguez Banga *

Dpto. Teoría de la Señal y Comunicaciones, ETSI Telecomunicación, Universidad de Vigo, Campus Universitario, 36200 Vigo, Spain

Received 8 June 2005; received in revised form 22 December 2005; accepted 28 December 2005

Abstract

In this paper, we focus on improving the quality of corpus-based synthesis systems by considering several candidate intonation contours. These candidates are generated by a unit selection procedure for which cost functions are defined. The consideration of several possible pitch contours adds an additional degree of freedom to the search for appropriate speech units. Objective and subjective tests confirm an improvement in the quality of the resulting synthetic speech.

Keywords: Speech synthesis; Unit selection; Corpus-based; Intonation

1. Introduction

As is widely acknowledged, a crucial stage in TTS is prosodic modelling, due to its important influence on the naturalness of synthetic speech. Normally, three specific features are referred to by the term prosody: intonation, segmental duration, and energy. Since these are intimately related, any TTS must take them into account. However, intonation modelling above all is considered to play a major role in the endeavour to produce natural speech synthesis systems. Our work is mainly focused on intonation modelling for corpus-based speech synthesisers. We describe a model that endeavours to reproduce the way a specific speaker talks that preserves most of his/her intonational richness.

Traditionally, corpus-based text-to-speech systems generate synthetic speech in a two-stage process. First, the target prosody is specified by means of a phonological or phonetic representation and, secondly, a set of speech units that minimise a cost function is finally chosen. Once the target prosody is selected, no alternative prosodic information is generally considered, even when appropriate speech units are not found.

We propose an intonation model that employs unit selection to determine a set of possible pitch contours (N candidates). Next, for each candidate, we look for a suitable sequence of speech units. The selection of the definitive intonation contour

* Corresponding author. Tel.: +34 986 812676; fax: +34 986 812116.
E-mail addresses: campillo@gts.tsc.uvigo.es (F. Campillo Díaz), erbanga@gts.tsc.uvigo.es (E. Rodríguez Banga).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.12.004

is also influenced by the existence of proper speech units. The quality of synthetic speech is improved in two aspects: firstly, in the suitability of pitch contour to the set of speech units and secondly, in the variability in pitch contour introduced by the influence of the segmental information on the intonation model. As a consequence of this additional variability, pitch contour is much less predictable and synthetic speech is not monotonous. This is an important advantage with respect to other intonation models which are highly predictable. In order to mitigate this limitation, some systems introduce a random component in the intonation contour, which, however, we consider to be quite unrealistic. Our model, in contrast, not only generates 'real' variability, but preserves the micromelody in most of the speech segments, as will be seen below. A similar idea of integrating prosody prediction and the unit selection process was previously proposed in (Bulyko and Ostendorf, 2001), although implemented as the composition of weighted finite-state transducers (WFSTs) and described for a constrained domain application.

Our TTS system is called Cotovia (see http://www.gts.tsc.uvigo.es/cotovia for a demonstration). It is a bilingual text-to-speech synthesiser (Galician and Spanish) (Campillo and Banga, 2002) that takes into account several intonation contours and that subordinates its final selection to the existence of a proper sequence of acoustic units (in this case, demiphones). The generation of the candidate pitch contours and the selection of the speech units are made by means of target cost and concatenation cost functions. The cost functions employed for the selection of the speech units are similar to those described in (Black and Campbell, 1995; Hunt and Black, 1996). This article focuses on a technique for generating candidate pitch contours and the design of the corresponding cost functions, and terminates with our conclusions on the improvements furnished by our approach.

2. Intonation

The literature contains descriptions of a great number of intonation models, typically classified as either phonological or phonetic models. In simple terms, phonetic models consider intonation as a mere pitch contour along a sentence or proposition. Phonological models, on the other hand, are more compact and general, and describe intonation as a combination of tonal features that form bigger structural units. These structural units are ultimately associated with the acoustic correlate of the fundamental frequency. In (Botinis et al., 2001) the authors emphasise the multiple correspondence between phonological and phonetic planes. From a practical point of view, this means that the same message can be transmitted with different pitch contours (phonetic realizations).

Intonation is affected by many factors, some of which reflect characteristics of the speaker (geographical origin, cultural level, mood, emotion, age, sex, etc.), and others which are related to linguistic properties (lexical stress, syntactic dependencies, type of proposition, etc.). In corpus-based speech synthesis, a speaker's characteristics are determined by a recording of the speech corpus. With respect to linguistic factors, a TTS can only use those which can be determined from the input text (in a simple format or further annotated using a markup language).

Our intonation model is a classical hierarchical model that considers the accent group (or tonic group) as the basic intonation unit. In this work an accent group is considered to be a sequence of unaccented words ending in an accented word. The intonation group is formed by the concatenation of accent groups. We make the assumption that the intonation group always coincides with the phonic group, which is delimited by pauses in the speech. Finally, a sentence is composed of one or more intonation groups. An important advantage of this approach is that the division in accent groups and intonation groups is quite easy to assess from the input sentence. Another advantage is that it groups together and relates accented and unaccented syllable characteristics, and so avoids the need to define complex relations between syllables.

3. Unit selection intonation modelling

As already mentioned in the Introduction, our intonation model first obtains a set of N candidate pitch contours and then selects one by taking into account the existence of a suitable sequence of speech units. In corpus-based TTS the search for a sequence of speech units is normally made using a Viterbi algorithm that minimises a cost function. This cost function is typically composed of two subcosts: the target cost and the concatenation cost. The target cost measures the suitability of a speech unit to the desired characteristics in terms of factors