Prosody, Speech Synthesis and Cognitive Linguistics

Tomasz Kuczmarski, Adam Mickiewicz University, Poznań, Poland

Abstract

The current paper discusses a way to employ methods and utilities provided by computational cognitive linguistics in speech technology research. First, the currently predominant methodologies are briefly discussed. It is shown that some problems regarding the naturalness of synthetic voices, such as prosody, are intractable within the ruling framework. Therefore, a special case of Construction Grammar (CxG) – Embodied Construction Grammar is proposed as a method to investigate spoken language holistically. It is suggested that this approach might show some prosodic dependencies which can hardly be studied in another way. At the end of the article some intuitions about the anticipated results are covered.

1. Introduction

Speech synthesis research began in 1779, when Christian Kratzenstein created bellow-operated models of vocal tracts that could produce one of the following long vowels: [aː], [eː], [iː], [oː] and [uː]. Since then, the problem of artificial speech has gained much attention and has been confronted from many viewpoints, conforming to current scientific trends. Many paradigms and resulting synthesizer technologies arose on grounds of numerous linguistic theories.

Linguistics have recently experienced yet another notable paradigm shift, a shift toward a holistic cognitive approach to language, started by Lakoff [1] [2], Fillmore [3] and Langacker [4]. Cognitive linguistics has proven to provide better explanation to problems which the predominant Chomskyan school considered as an irrelevant part of the performance [5].

The only field of linguistics that stayed relatively indifferent is the speech technology. In current speech synthesis research, most of the attention is given to the latter part of the abstract-to-physical scheme. This so-called low level speech synthesis is an orthodoxy started by the discovery of acoustic cues for the perception of phonetic segments and by the work of Klatt [6]. Most of the field was taken over by engineers who are on an everlasting quest for segmental „intelligibility” and utility of final systems.

Nowadays, most state-of-the-art synthesis systems are able to generate voice of near-human quality. However, there is still much to be done in terms of naturalness. This in turn, needs solving problems which are intractable within the currently predominant paradigm. Prosody, being the most important, is influenced by all aspects of language, including pragmatics. It also leans toward the abstract part of the abstract-to-physical rendering scheme. Therefore, prosody should not only be a subject of a holistic research but it also needs a high level speech synthesis framework. The recently emerged computational cognitive linguistics provides the necessary methods and tools.

2. Methods

Among the tools and methods offered by computational cognitive linguistics, frame semantics [7] and construction grammar (CxG) [8] [9] [10] [11] seem to be the most promising in terms of prosody research. Construction grammar is a family of theories and models of grammar. It constitutes a mature holistic and usage-based framework that treats all types of expressions and dimensions of language (syntax, semantics, pragmatics, discourse, morphology and phonology) as equally central to capturing grammatical patterning.

Within this theory a grammatical construction is a pairing of form and content. This approach corresponds to the foundation of general semiotics. The form is any combination of syntactic, morphological, or prosodic patterns, whereas meaning is understood in terms of lexical semantics, pragmatics and discourse structure. Therefore, a grammar consists of intricate networks of overlapping and complementary patterns used for encoding and decoding of linguistic expressions.
Embodied Construction Grammar (ECG) is a special case of CxG, developed at ICSI, UC Berkeley, and the University of Hawaii within the The Neural Theory of Language (NTL) project – an interdisciplinary research effort to answer the question: How does the brain compute the mind? In this formalism designed specifically for integration into a simulation-based model of language understanding, conceptual representations are also constrained to be grounded in the body’s perceptual and motor systems, and more precisely to parametrize mental simulations using those systems. Understanding an utterance thus involves at least two distinct processes: analysis to determine which constructions the utterance instantiates, and simulation according to the parameters specified by those constructions [12].

The so-called topic-comment articulation and its relation to prosody has been widely studied by the Prague school [13] and has shown some interesting dependencies. However, little attention has been given to a multi-level investigation of coherence of prosodic structures and other dimensions of language, such as the predication structure of an utterance and the distribution of predicate and its arguments within that utterance. It is suspected that this approach could yield promising results.

Such research requires a number of preparations. First of all, a dedicated speech corpora should be built. The problem here is that it was indirectly assumed that only “neutral” prosody is to be examined. There is an ongoing discussion whether such “neutral” prosody exists at all. The opponents argue that all speech is uttered in some context, which triggers unpredictable emotional responses and variable pragmatic decisions during utterance planning. All this affects the resulting prosody. Proponents turn to various means in search of arguments. For example, results of the brain activation study showed that recognizing emotional intonations and discriminating their expressiveness leads to a predominant activation of the right hemisphere (RH) with right frontal preponderance only in the absence of linguistic task demands. Inner speech performed in addition and concurrently with the identification/discrimination task gave rise to a balanced RH/LH activation pattern with left frontal preponderance. [14]. This might suggest that “emotional” prosody is only superimposed on the “neutral” prosody and may be somehow “filtered” out.

One way of dealing with this problem would be to record the whole speech corpora in constant conditions, using a strictly defined scenario (reading a text aloud to some audience in case of application in text-reader systems) in as few number of sessions as possible.

Once the speech corpora is ready, a well-defined ECG and the underlying lexical semantics should be built for the language in question, to serve as a parser for input utterances and their initial annotations. The final step is to analyze the speech corpora in terms of its ECG structures and their correspondence to prosodic contours. This might be attained by employing Artificial Neural Networks or Hidden Markov Models. However, an in-depth expert analysis performed by a human is indispensable.
3. Anticipated results and discussion

This unifying approach to prosody analysis seems very promising. However, at this point it is hard to expect any definite results except for some correlations on the level of predicate structure of an utterance, which is a somewhat derivative of the Prague school’s research mentioned above.

Even if some results are attained, the gap between abstract and physical, phonological and phonetic is still to be filled. Since there is no linear relation between the constituents of prosody; rhythm, stress and intonation, and their physical or phonetic counterparts; amplitude, timing and fundamental frequency (f0) of speech signal, even a good model of prosody might turn out to be useless until the transition between phonology and phonetics in the process of speech production is well defined. However, the cognitive framework might become useful also in this case, as was proven by some investigation in cognitive phonetics [15] [16].

Fig. 2. Overview of constructs in ECG.

Bibliography


Author:
MA Tomasz Kuczmarski
Adam Mickiewicz University
ul. Wieniawskiego 1
61-712 Poznań, Poland
tel. (048)509079785
e-mail: tkucz@amu.edu.pl