Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language xxx (2016) xxx-xxx

www.elsevier.com/locate/csl

Hierarchical representation and estimation of prosody using continuous wavelet transform

Antti Suni[a,b], Juraj Šimko[a], Daniel Aalto[c,d,a], Martti Vainio[a,*]

[a] *Institute of Behavioural Sciences, University of Helsinki, Finland*
[b] *Department of Signal Processing and Acoustics, Aalto University, Finland*
[c] *Communication Sciences and Disorders, Faculty of Rehabilitation Sciences, University of Alberta,* Canada
[d] *Institute for Reconstructive Sciences in Medicine (iRSM), Misericordia Hospital, Edmonton, Canada*

Received 23 February 2015; received in revised form 11 October 2016; accepted 6 November 2016

Abstract

Prominences and boundaries are the essential constituents of prosodic structure in speech. They provide for means to chunk the speech stream into linguistically relevant units by providing them with relative saliences and demarcating them within utterance structures. Prominences and boundaries have both been widely used in both basic research on prosody as well as in text-to-speech synthesis. However, there are no representation schemes that would provide for both estimating and modelling them in a unified fashion. Here we present an unsupervised unified account for estimating and representing prosodic prominences and boundaries using a scale-space analysis based on continuous wavelet transform. The methods are evaluated and compared to earlier work using the Boston University Radio News corpus. The results show that the proposed method is comparable with the best published supervised annotation methods.

Keywords: Phonetics; Prosody; Speech synthesis; Wavelets

1. Introduction

Two of the most primary features of speech prosody have to do with chunking speech into linguistically relevant units above the segment and the relative salience of the given units; that is, boundaries and prominences, respectively. These two aspects are present in every utterance and are central to any representation of speech prosody. Arrangement of prominence patterns and placement of boundaries reflect the hierarchical structure of speech, i.e., gradual nesting of units, segments within syllables, syllables within (prosodic) words, words within phrases, phrases within utterances and beyond (Tseng et al., 2005). Borders between adjoining units of higher order – words, phrases – present affordances for prosodic breaks of different types and strengths. Attention of the listener can be selectively drawn to individual units within the hierarchy; prominent syllables mark lexical stress, prominent words signal focus, etc.

* Corresponding author.
E-mail address: antti.suni@helsinki.fi (A. Suni), juraj.simko@helsinki.fi (J. Šimko), aalto@ualberta.ca (D. Aalto), martti.vainio@helsinki.fi (M. Vainio).

http://dx.doi.org/10.1016/j.csl.2016.11.001
0885-2308/ 2016 Published by Elsevier Ltd.

In speech, boundaries are usually signalled by a local reduction in one or more signal characteristics (such as intensity or pitch) at a border spanning several hierarchical levels. In a complementary fashion, prominence is typically associated with an increase in some or all of these signal properties, typically associated with a particular hierarchical level.

This simple insight suggests that these prosodic constituents could be represented within a uniform methodology that identifies both prominence and boundaries as complementary phenomena manifested in speech signals. Such a methodology would be beneficial to both basic speech research and speech technology, especially speech synthesis and recognition. At the same time, to be useful for data oriented research and technology, the annotation system should strive towards being unsupervised as opposed to the systems that rely on humans, either directly labelling speech data (Silverman et al., 1992) or providing a manually labelled training set used for training the system.

Ideally, the system should approach human-like performance but without the variability of human labellers caused by complex interactions between the top-down and bottom-up influences. In order to achieve that we propose here a system based on Continuous Wavelet Transform (CWT) that (1) approximates human processing of a complex signal relevant for identifying prominence and boundaries, and (2) is capable of representing the speech signal in a manner that captures the hierarchical nature of prosodic signalling.

In this paper we present a hierarchical, time-frequency scale-space analysis of prosodic signals (e.g., fundamental frequency, energy, duration) based on the CWT. The presented algorithms can be used to analyse and annotate speech signals in an entirely unsupervised fashion. The work stems from the need to annotate speech corpora automatically for text-to-speech synthesis (TTS) (s4a, 2014) and the subject matter is partly examined from that point of view. However, the presented representations should be of interest to anyone working on speech prosody.

Wavelets extend the classical Fourier theory by replacing a fixed window with a family of scaled windows resulting in scalograms, resembling the spectrogram commonly used for analysing speech signals. The most interesting aspect of wavelet analysis with respect to speech is that it resembles the perceptual hierarchical structures related to prosody. In scalograms, speech sounds, syllables, (phonological) words, and phrases can be localised precisely in both time and frequency (scale). This would be considerably more difficult to achieve with traditional spectrograms. Furthermore, the wavelets give natural means to discretise and operationalise the continuous prosodic signals.

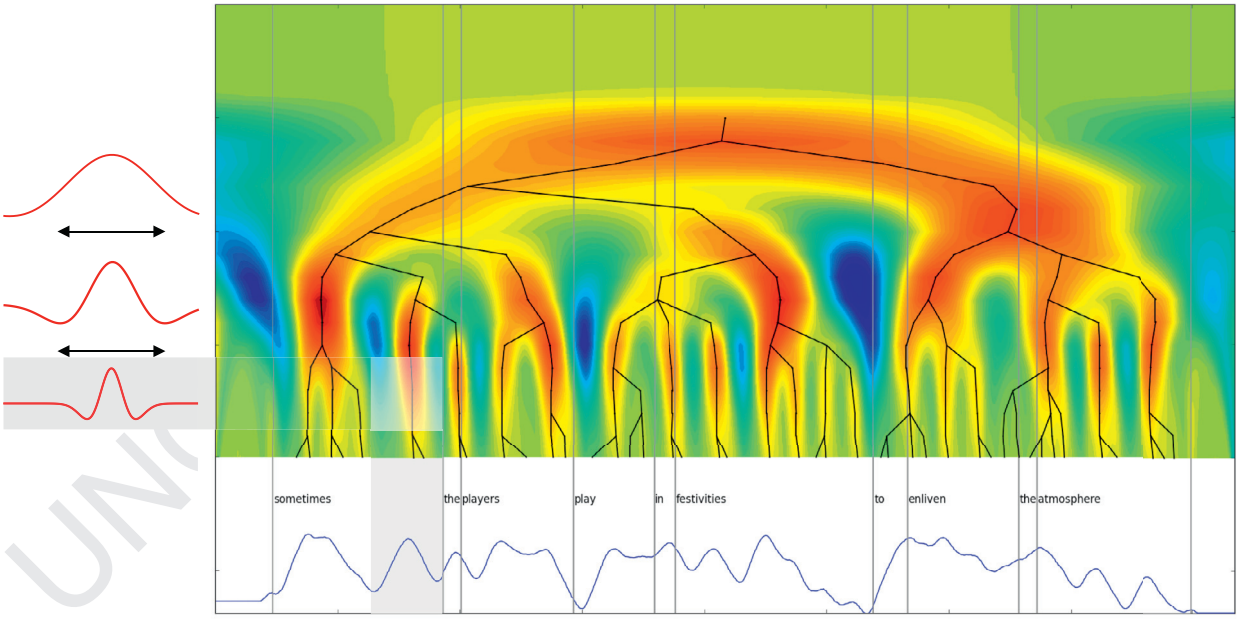

Fig. 1. An illustration of a CWT based analysis of a composite prosodic signal combining energy, f_0 and word duration (bottom) of an English utterance 'Sometimes the players play in festivities to enliven the atmosphere". The hierarchical tree structure is highlighted in black. Mother wavelets corresponding to syllables, prosodic words, and phrases are depicted on the left. See text for more detail. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Please cite this article as: A. Suni et al., Hierarchical representation and estimation of prosody using continuous wavelet transform, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.11.001