Local Speech Rate Extraction, Manipulation, and Caricature

Hartmut R. Pfitzinger

Institut für Phonetik und Sprachliche Kommunikation
Schellingstr. 3, 80799 München, Germany
*e-mail:*hpt@phonetik.uni-muenchen.de

Abstract

An important acoustic characteristic of human speech is its continuous spectral variation in the temporal domain: even two directly successive short-term frequency spectra are in the majority of cases different. But the most remarkable observation on the temporal structure of speech is that the speed of momentary frequency spectrum variation is a composition of very different modulation frequencies, which furthermore are anything but constant.

Tillmann 1980 [4] denominated three of these modulation frequencies: The *A-prosody* is the tier of *slow* speech events (e.g. the succession of stressed syllables or rhythmic feet in a stress-timed language). Its speed lies inside a frequency range where events are clearly perceivable and countable. The second tier is called the *B-prosody* and includes those speech events which are just too fast to count. Nevertheless it is possible to tap synchronously with a finger to each event (e.g. the succession of P-centers or syllables). Finally, the *C-prosody* describes a tier where events are much too fast to count or even to synchronize to. They sound rather like a clatter or rattle (e.g. the taps of the tongue tip when producing an alveolar trill).

These different *prosodies* can be viewed as the modulators of (macro-prosodic and micro-prosodic) pitch, energy, degree of reduction, and voice source properties. The temporal interaction of the *A-, B-,* and *C-prosody* appears to be quite complex from an analytical point of view but since humans decode the temporal structure of speech with ease, we should assume

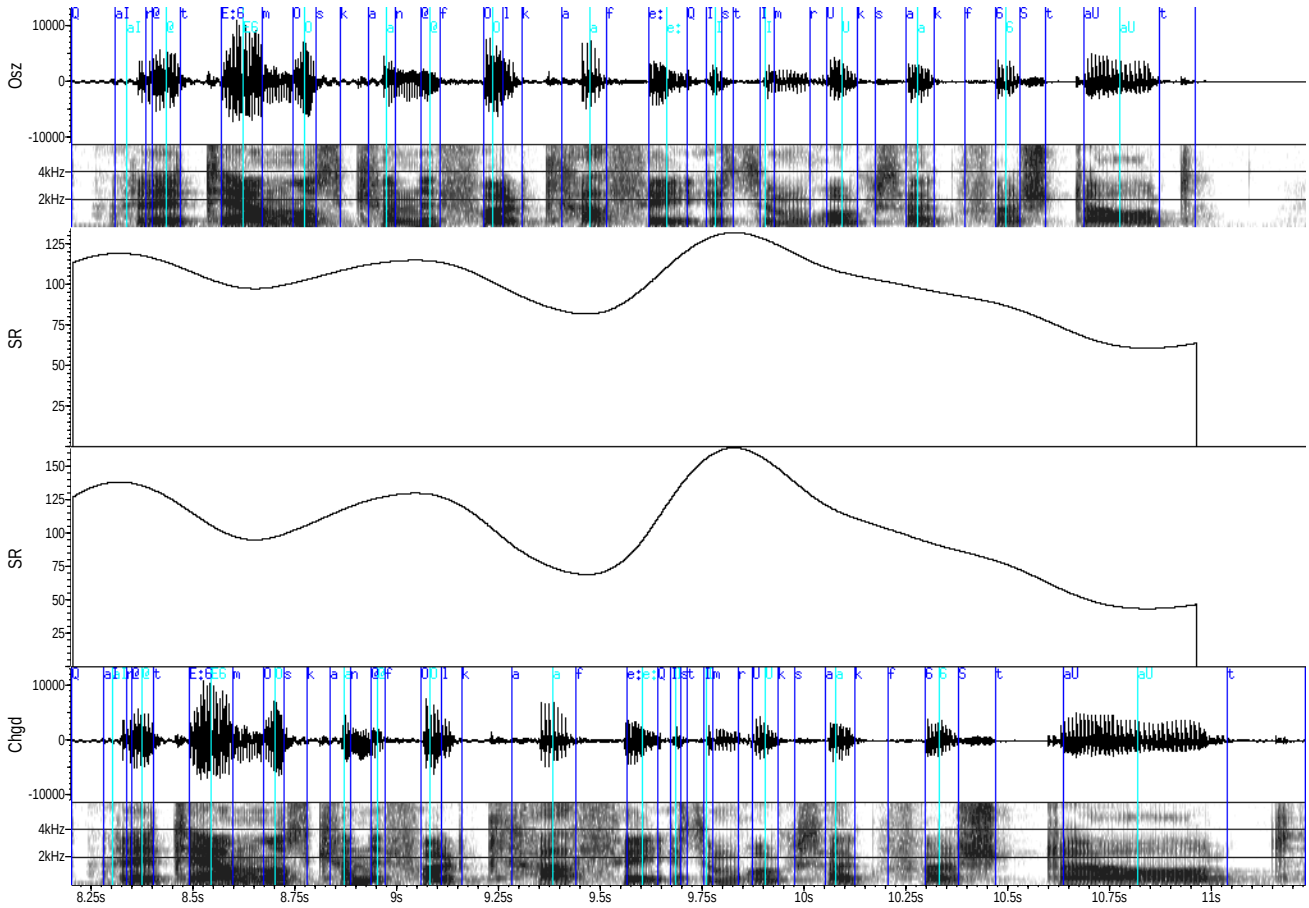

Figure 1: *Osz*: Original utterance “Eine Thermoskanne voll Kaffee ist im Rucksack verstaut” (eng.: *A thermos flask filled with coffee is tucked in the backpack*), *SR*: extracted perceptual local speech rate (PLSR), *SR*: an exaggerated version of the PLSR, and *Chgd*: the accordingly locally time-warped speech signal.

quite simple underlying rules: Local speech rate is not a result of the interaction; instead, it can be viewed as the source and the modulator of the *prosodies*. In the talk we will explain how we define and measure perceptual local speech rate (Pfitzinger 1998/99 [2, 3]) and how we modify it in a given utterance. The relevance of this new prosody for the received impression about a speaker will be demonstrated by caricaturing speakers only through exaggeration (see Fig. 1) of their specific usage of local speech rate modifications.

References

[1] Ladd, D. R.; Campbell, W. N. (1991). Theories of prosodic structure: Evidence from syllable duration. In *Proc. of the XIIth Int. Congress of Phonetic Sciences*, vol. 2, pp. 290–293, Aix-en-Provence.

[2] Pfitzinger, H. R. (1998). Local speech rate as a combination of syllable and phone rate. In *Proc. of ICSLP '98*, vol. 3, pp. 1087–1090, Sydney.

[3] Pfitzinger, H. R. (1999). Local speech rate perception in German speech. In *Proc. of the XIVth Int. Congress of Phonetic Sciences*, vol. 2, pp. 893–896, San Francisco.

[4] Tillmann, H. G.; Mansell, P. (1980). *Phonetik: Lautsprachliche Zeichen, Sprachsignale und lautsprachlicher Kommunikationsprozeß*. Verlagsgemeinschaft Ernst Klett–J. G. Cotta, Stuttgart.