Transcription of tonal aspects in speech and a system for automatic tonal annotation

Piet Mertens
University of Leuven, Leuven, Belgium

This paper describes a transcription system for tonal aspects in speech, as well as an automatic tonal annotation system using this notational convention. Tonal aspects include pitch levels (tones) and pitch movements, associated with individual syllables. A comprehensive prosodic transcription should also indicate stress, but this aspect is not covered here.

The tonal notation uses 5 **pitch levels**. The levels L (low), M (mid) and H (high) are identified on the basis of pitch intervals in the left context, whereas levels B (bottom) and T (top) are relative to the pitch range of the speaker. In addition, the notation uses 5 intra-syllabic **pitch movements**: level, large rise (R), large fall (F), small rise (r) and small fall (f), where the size (large vs. small) of pitch intervals is relative to the speaker's pitch range. The level pitch movement, noted by the underscore (_), is indicated only in **compound pitch movements**, such as "rise followed by level". Although the notation allows for movements up to any level of complexity, in practice, movements consisting of 3 or more parts (e.g. level-rise-fall) are rare. In order to represent **stress**, a mark for stress is placed before the pitch level of the stressed syllable (as in the IPA notation).

Although the proposed notation shows many resemblances to other systems, including the auto-segmental approach of **ToBI** (Silverman *et al*. 1992, Beckman *et al*. 2005), and **INTSINT** (Hirst *et al*. 1998), there are **major differences** as well. (1) The number of pitch levels. (2) The definition of pitch levels, which is either local (based on pitch changes in the context) or global (relative to pitch range). In both cases it takes into account the speaker's pitch range. (3) The representation of intra-syllabic pitch movements. (4) The distinction between 2 sizes of pitch movements. (5) All pitch movements are aligned with syllables: the pitch level indicates the pitch near vowel onset and, for a compound tone, parts are spread over the syllable rime. Pitch movements spreading over sequences of syllables should be identified on the basis of locations (positions), such as stress and prosodic boundaries. Differences (3), (4) and (5) above also apply to the INTSINT notation.

Existing approaches to **automatic prosody transcription** use various strategies, which may be based on a global analysis of F0 in utterances or on the calculation of a local reference level, such as a "declination line", which varies during the utterance, and relative to which observed pitch variations are interpreted. Apart from these specialized algorithms, generic machine learning techniques have often been applied (Wightman & Ostendorf 1994; Braunschweiler 2005; Ananthakrishnan *et al*. 2008; Campione *et al*. 2000, 2001; Geoffrois 1995; Wagner 2008; Rosenberg 2011). The latter all require a training corpus in which prosodic events are labelled according to some transcription system, such as ToBI or INTSINT. Unfortunately, except for ToBI transcriptions of English, such reference corpora are not available. And so there is a clear need for an automatic transcription system which does not require labelled training corpora. The approach proposed here does not require a training corpus.

The **system for automatic prosodic annotation** includes the following processing steps (see figure 1). 1. The calculation of acoustic parameters for F0, intensity, voicing. 2. The segmentation of syllabic nuclei, on the basis on intensity, but starting from the time interval corresponding to the rime of each syllable. 3. The detection of speech pauses. 4. The pitch stylization, as provided by Prosogram (Mertens 2004), based on the tonal perception model by d'Alessandro & Mertens 1995, in which perceptual thresholds (glissando, differential glissando) are applied to the pitch variation of each syllable. 4. The calculation of the pitch range of the speaker, and the definition of the size of large and small melodic intervals on the basis of this pitch range. 5. The analysis of intra-syllabic pitch movements. 6. The identification of pitch level for each syllable. The approach proceeds bottom-up, can be applied to many languages and is largely theory-neutral. The output may be mapped to other transcription systems.

The system is implemented in the scripting language of Praat (Boersma & Weenink 2011). It has been **tested** on part of the Rhapsodie corpus for French (corresponding to 65 minutes of speech in 38 recordings, by 42 speakers, male and female). It is being **evaluated** on speech samples (ca. 10 min.) for which a reference annotation has been obtained by hand-correcting the output of the automatic tonal annotation.

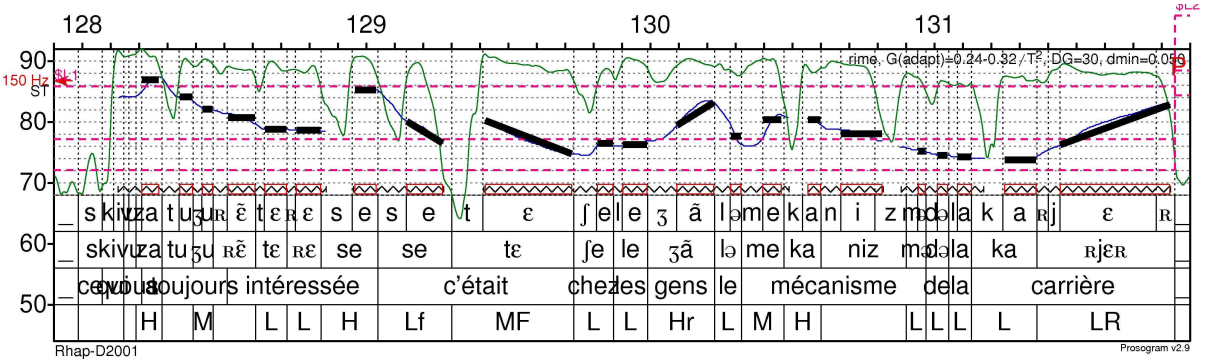

Figure 1. Illustration of the tonal annotation and some processing steps used in the automatic annotation system. Tiers 1 to 3 show the annotation in speech sounds (phonemes), syllables and words, taken from the speech corpus. Tier 4 shows the tonal annotation (see text for details). The upper part of the figure shows the acoustic parameters of intensity (green line), fundamental frequency (thin blue line), voicing (saw-tooth line, in black). The syllabic nuclei obtained by the segmentation appear as rectangles (in red). The stylized pitch contour is plotted as the thick black line. The calculated pitch range is represented by three horizontal lines (using long dashes), corresponding to the bottom and top of the pitch range, and the median of the pitch values (maximum and minimum pitch in each syllable) used by the speaker.

References

Alessandro, C. d'; Mertens, P. (1995) Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language* 9(3), 257-288.

Ananthakrishnan, S.; Narayanan, S. (2008) Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Trans. on Audio Speech and Language Proc.* 16(1), 216-228.

Beckman, M.E.; Hirschman, J. & Shattuck-Hufnagel, S (2005) The original ToBI system and the evolution of the ToBI framework. in: Jun, Sun-Ah (ed.) (2005) *Prosodic Typology.* Oxford University Press. pp. 9-54.

Boersma, Paul & Weenink, David (2011). Praat: doing phonetics by computer [Computer program]. Version 5.2.46, retrieved 7 October 2011 from http://www.praat.org/

Braunschweiler, N. (2005) The Prosodizer - Automatic Prosodic Annotations of Speech Synthesis Databases. *Proceedings Speech Prosody* (Dresden).

Campione, E.; Hirst, D.; Véronis, J. (2000) Automatic Stylisation and Modelling of French and Italian Intonation. in: Botinis (ed) (2000) *Intonation: Analysis, Modelling and Technology*. Kluwer Academic Publishing, pp. 185-208.

Campione, E. & Véronis, J. (2001) Etiquetage prosodique semi-automatique des corpus oraux. *Actes TALN*, Tours, 2-5 juillet 2001.

Geoffrois, E. (1995) *Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole*. Ph.D. Université Paris XI Orsay, 20 decembre 1995.

Hirst, D. & Di Cristo, A. (1998) A survey of intonation systems. In: Hirst, D. & Di Cristo, A. (eds.) (1998) *Intonation Systems. A Survey of Twenty Languages*. Cambridge Univ. Press. 1-44.

Mertens, P. (2004) The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. in B. Bel & I. Marlien (eds.) *Proceedings of Speech Prosody 2004*, Nara (Japan), 23-26 March 2004.

Mertens, P. (2004) Un outil pour la transcription de la prosodie dans les corpus oraux. *Traitement Automatique des langues* 45 (2), 109-130.

Rosenberg, A. (2010) AuToBI - A Tool for Automatic ToBI Annotation. *Proc. Interspeech* 2010.

Silverman, K. ; Beckman, M.; Pitrelli, M.; Ostendorf, M.; Wightman, C. & Price, P. (1992) TOBI: a standard for labeling English prosody. *Int. Conf. on Spoken Language Systems*, 867-870.

Wagner, A. (2008) A comprehensive model of intonation for application in speech synthesis. Ph.D. Uniwersytet im. Adama Mickiewicza w Poznaniu.

Wightman, C.W. & Ostendorf, M. (1994) Automatic labeling of prosodic patterns. *IEEE Trans Speech and Audio Processing* 2, 469-481.