Challenges in automatic annotation and the perception of prosody in spontaneous speech

István Szekrényes
Universtiy of Debrecen, Hungary

Introduction

The proposal aims at demonstrating a rule-based annotation method and an experimental approach for the automatic analysis and the perceptual aspect of speech prosody. The development is inspired by the work of Piet Mertens [5] using its objectives and the psychoacoustic model of tonal perception [2]. The main purpose of our implementation is to extend the *HuComTech* Hungarian multimodal corpus[1] with prosodic labels for the analysis of non-verbal behaviour in human-human interactions [3]. The most recent repairs and adaptations - resulting in the current exibility of the program - are made in collaboration with the *SegOrg* project[2] during the analysis of various kinds of recordings (with 2-14 speakers) from the FOLK [6] German corpus. Details of the algorithm under the name *ProsoTool*[3] are described in [8] [7]. The same algorithm as part of the e-magyar project [4] referring to that project is mentioned as *emPros*[4].

Methodology

The algorithm is implemented as a Praat[1] script including a speaker isolation and an intonation processing module which stylizes and categorizes F0 curves as perceptually relevant melodic sequences labelling the shape (rise, fall, ascending etc.) and the relative (compared to the individual vocal range) and absolute (in Hertz) position of every movement. The input is a speech sound file (in WAV format) and the acoustic representation of turn-taking (in Praat TextGrid) to isolate the voice segments of the speakers excluding overlapping speeches. Based on the F0 distribution of the isolated segments, the algorithm divides the individual vocal range into five levels (see in Figure 1). F0 smoothing and stylisation are performed in every single speech segments resulting the melodic sequences of intonation as it can be seen in Figure 2. The categorizations (the resulting labels) are based one the global F0 distribution and some parameters (the amplitude and the duration of movements) which are also used in the Tilt intonation model [9].

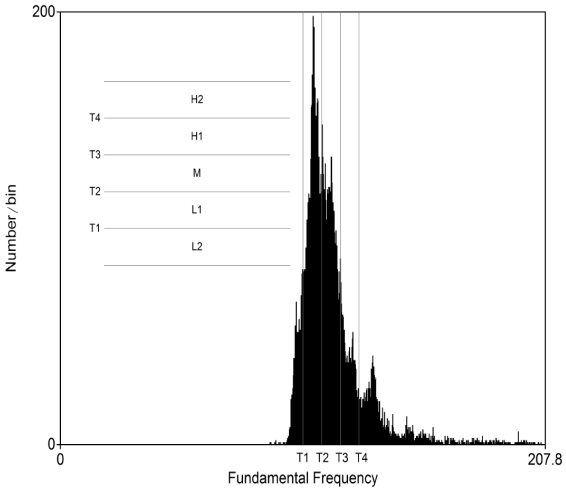

Figure 1: Individual vocal ranges based on F0 distribution

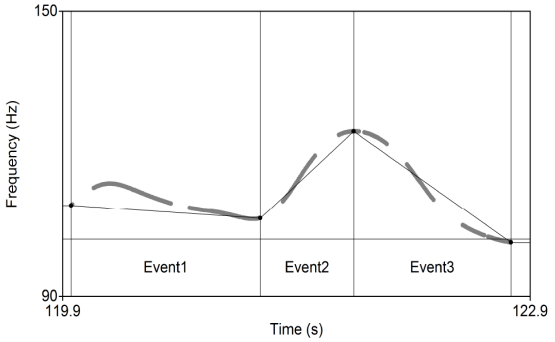

Figure 2: The result of smoothing and stylization

Results

The validation is the most complicated and a still on-going part of the project. In Figure 3, one can see the output (the annotation at the bottom and the measured F0 with transcription above) for an utterance with a prototypical intonation of a Hungarian yes-no question.

[1] https://hdl.handle.net/1839/00-0000-0000-001A-E17C-1@view
[2] http://www1.ids-mannheim.de/prag/muendlichekorpora/segcor.html
[3] https://github.com/szekrenyesi/prosotool
[4] http://e-magyar.hu/hu/speechmodules/empros

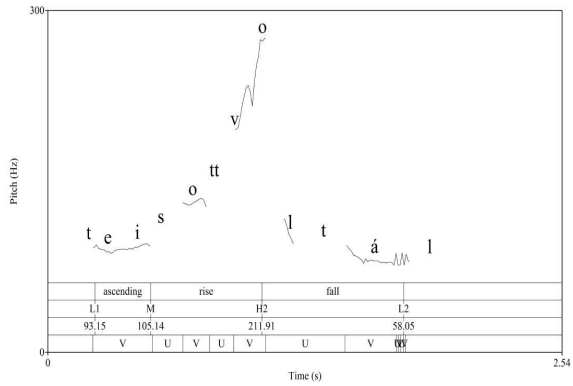

Figure 3: Output for a Hungarian yes-no question: "Te is ott voltál? [Were you there too?]"

In spontaneous speech (see Figure 4), the perceptual alignment of the resulting labels is less evident or verifiable.

For validation, some experiments are also designed to explore the perception of prosody in spontaneous conversations using various conditions (see Figure 5). The results are still in the process of evaluation.

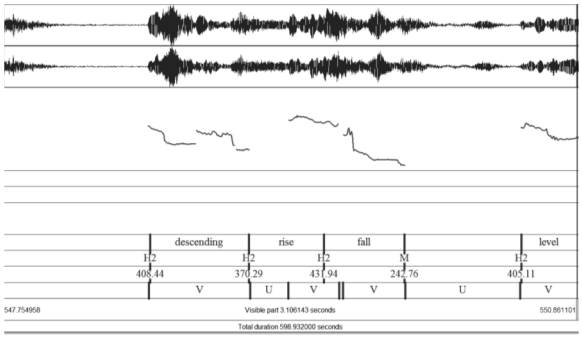

Figure 4: Results for a spontaneous conversation

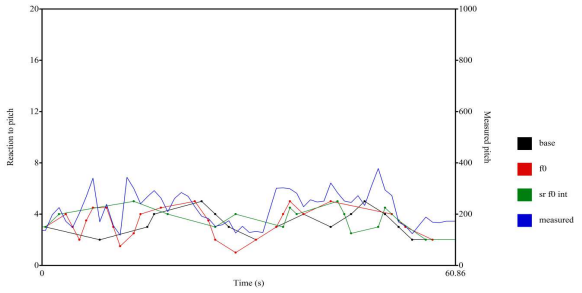

Figure 5: Perception of speech prosody in different conditions

References

[1] David Boersma, Paul & Weenink. *Praat: doing phonetics by computer* [computer program]. version 6.0.22. http://www.praat.org/, 2016. retrieved 15 November 2016.

[2] J. 't Hart. 1976. Psychoacoustic backgrounds of pitch contour stylisation. *IPO-APR*, 11:11–19.

[3] László Hunyadi, András Földesi, István Szekrényes, Alexandra Staudt, Hermina Kiss, Ágnes Abuczki, and Alexa Bódog. 2012. Az ember-gép kommunikáció elméleti-technológiai modellje és nyelvtechnológiai vonatkozásai. In *Általános Nyelvészeti Tanulmányok XXIV: Nyelvtechnológiai kutatások*, Akadémiai Kiadó, Budapest, 265–309.

[4] András Kornai and István Szekrényes. 2011. e-magyar beszédarchívum. In V. Vincze, editor, *XIII. Magyar Számítógépes Nyelvészeti Konferencia* (MSZNY2017), Szegedi Tudományegyetem Informatikai Tanszékcsoport, 103–109.

[5] Piet Mertens. 2004. *The prosogram: Semi-automatic transcription of prosody based on a tonal perception model.* In Proceedings of Speech Prosody.

[6] Thomas Schmidt. 2016. Good practices in the compilation of folk, the research and teaching corpus of spoken German. In John M. Kirk and Gisle Andersen, editors, *Compilation, transcription, markup and annotation of spoken corpora*, Special Issue of the International Journal of Corpus Linguistics [IJCL 21:3], 396–418.

[7] István Szekrényes. 2015. Prosotool, a method for automatic annotation of fundamental frequency. In 6th *IEEE In- ternational Conference on Cognitive Infocommunications (CogInfoCom)*, New York, IEEE 291–296.

[8] István Szekrényes. 2014. Annotation and interpretation of prosodic data in the hucomtech corpus for multimodal user interfaces. *Journal on Multimodal User Inter- faces*, 8:(2):143–150.

[9] P Taylor. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 107(3):1697–1714.