Prosodic and spectral characteristics of non-lexical units with different communicative loads

Katalin Mády[1] & Uwe D. Reichel[1,2]
[1]HUN-REN Hungarian Research Centre for Linguistics
[2]audEERING GmbH, Germany

Filled pauses and conversational grunts are non-lexical units that occur frequently in spontaneous discourse. Filled pauses (FP) like *uh, um* signalise disfluency, while conversational grunts (CG), e.g. *mhm*, serve as signals for backchanneling, surprise, uncertainty or positive/negative response in many languages, e.g. Hungarian, subject to the current analysis.

FPs can have similar forms to CGs, but the latter carry consciously planned pragmatic information similar to lexical units like *yes, what? really?*, while FPs primarily refer to planning difficulties. The two categories were predicted by feature-based and end-to-end models in Hungarian task-oriented dialogues [5, 3], achieving accuracies up to 0.99. In this talk, acoustic parameters from openSMILE [1, 2] and CoPaSul [4] feature sets are discussed with respect to their differences in pragmatic load.

CGs differred from FPs in three main aspects. (1) Longer duration, higher f0 mean and standard deviation (SD) along with higher interquartile range and SD in energy lead to stronger prominence and carry an overall rising pattern with mostly two rather than one syllable in the *mhm*-like sequence (see Fig. 1). (2) More voicing and more harmonicity (higher cepstral peak prominence and more negative spectral slope indicating less creaky voice), since pitch is relevant for expressing the communicative functions (see Fig. 2). (3) Larger spectral distances between adjacent spectra in voiced segments and higher amplitude for F1 (see Fig. 3). Although CGs are typically produced with closed lips, the presence of a formant structure indicates that speakers apply articulatory techniques similar to lexical speech as opposed to the more diverse FPs that are often produced with neutral vowels lacking an articulatory target.

In sum, CGs show more prosodic, voice quality and articulatory variation which reflects their communicative load, i.e. the need to encode several pragmatic functions and meanings.

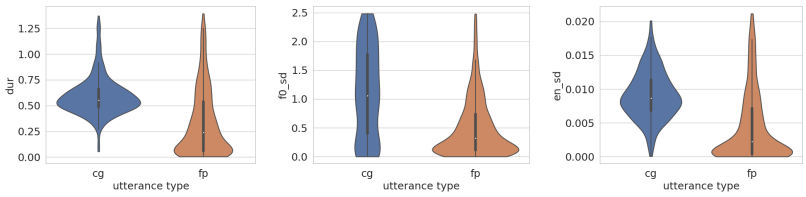

Figure 1: Longer duration (left), higher standard deviation in f0 (mid), and in energy (right).

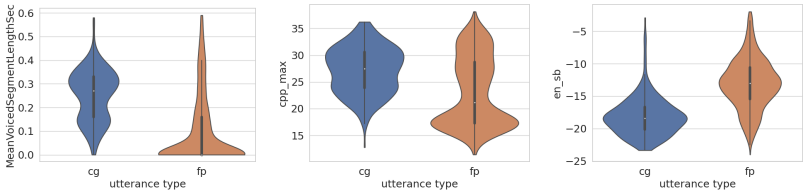

Figure 2: More voicing (left) and higher cepstral peak prominence (mid) and more negative spectral slope (right) for CG.

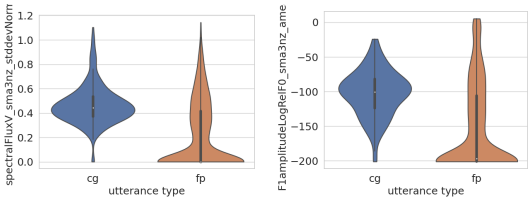

Figure 3: Higher spectral distances between adjacent spectra (left), higher mean F1 (right) for CG.

This work was funded by the National Research, Development and Innovation Office, grants K 135038 and 143075.

References

[1] Florian Eyben. *Real-time speech and music classification by large audio feature space extraction*. Springer, 2015.

[2] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[3] Katalin Mády, Anna Kohári, Péter Mihajlik, and Uwe D. Reichel. The Budapest Games Corpus. In *Proc. Beszédkutatás – Speech Research Conference*, pages 75–77, Budapest, 2023.

[4] U. D. Reichel. *CoPaSul Manual – Contour-based parametric and superpositional intonation stylization*. RIL, MTA, Budapest, Hungary, 2016. https://arxiv.org/abs/1612.04765.

[5] Uwe D. Reichel, Anna Kohári, and Katalin Mády. Acoustics and prediction of non-lexical speech in the Budapest Games Corpus. In *Proc. Beszédkutatás – Speech Research Conference*, pages 91–93, Budapest, 2023.