**Bridging the Gap: Exploring a Middle-Way Approach for Prosodic Annotation in Speech and Gesture**

Author1, Author2

1 Affiliation 2 Affiliation

Prosodic annotation plays a critical role in linguistic research, allowing us to delve into the subtleties of speech and gesture across diverse languages and contexts. However, the manual annotation process is often time-consuming and labor-intensive, resulting in limited data availability. This scarcity hampers the development of reliable computational models for comprehensive prosodic annotation, covering various annotation schemes and languages. To tackle this challenge, our project aims to adopt a middle-way approach, harnessing the expertise of skilled annotators alongside the power of computational tools.

Various methods for prosodic annotation exist, with some widely known and accepted, such as ToBI, while others may seem idiosyncratic or tailored to specific tasks (examples?). Regardless, the majority of these methods rely on manual labor, with annotators honing their skills through years of training. Though multiple annotators often annotate the data to minimize errors, their agreement typically reaches only about 60%, leaving a surprisingly high percentage open to discussion, which itself introduces the potential for errors and manipulation.

The issue of tedious manual annotation leads to a lack of available data, hindering the development of computational models capable of robust general prosodic annotation across diverse languages and schemes (see Ananthakrishnan & Narayanan, 2008; Rosenberg, 2010 as examples for ToBI on Standard American English). Thus, it becomes crucial to seek a middle-ground solution where models are developed for specific prosodic annotation tasks using a particular annotation method, combining the controlled efforts of skilled annotators with the detection power of computational tools.

In our project, we aim to establish connections between prosodic marking and automatic detection of focus. To achieve this, we will analyze audiovisual data from German and Catalan speakers producing various focus types. The perceived prominence of the focus types in these data was annotated on a scale 0 to 3 for both prosodic (DIMA, Kügler et al., 2015, 2019, 2022) and gestural prominence (Rohrer et al., 2020).

To capture acoustic markers of prominence, such as F0 (max peak, mean, range), intensity (max peak, mean, range), and duration, we will extract data from the annotated syllables. These acoustic markers will serve as predictors in a Bayesian ordinal model to rate prominence. Additionally, for gestural analysis, we will employ OpenPose (Cao et al., 2019) to track limb motion and identify related hand movements closest to the syllables and acoustic peaks (i.e., gestures aligned with speech, cf. Kadavá et al., 2023). Features like gesture acceleration, deceleration, and attack, among others, will be extracted, and a Bayesian model will be computed to analyze gestural prominence. Considering language-specific differences, our models will account for variations in acoustic and gestural features of prominence between Catalan and German (cf. Cole et al., 2019 for differences between Spanish, French, and English in speech).

Our goal is to identify the most predictive features of prominence in speech and gesture for German and Catalan, effectively bridging the gap between manual and computer-aided annotation. By establishing these links, we aim to alleviate the challenges posed by tiresome manual annotation. Subsequently, we plan to employ our findings to build a classifier for automatic focus type assignment to utterances, which will undergo verification by human annotators.

# References

Ananthakrishnan, S., & Narayanan, S. S. (2008). Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 216–228. https://doi.org/10.1109/TASL.2007.907570

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields* (arXiv:1812.08008). arXiv. https://doi.org/10.48550/arXiv.1812.08008

Cole, J., Hualde, J. I., Smith, C. L., Eager, C., Mahrt, T., & Napoleão de Souza, R. (2019). Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *Journal of Phonetics*, *75*, 113–147. https://doi.org/10.1016/j.wocn.2019.05.002

Kadavá, Š., Ćwiek, A., Stoltmann, K., Fuchs, S., & Pouw, W. (2023). *Is gesture-speech physics at work in rhythmic pointing? Evidence from Polish counting-out rhymes*. OSF Preprints. https://doi.org/10.31219/osf.io/67fzc

Kügler, F., Baumann, S., Andreeva, B., Braun, B., Grice, M., Neitsch, J., Niebuhr, O., Peters, J., Röhr, C. T., Schweitzer, A., & Wagner, P. (2019). Annotation of German Intonation: Dima Compared with Other Annotation Systems. *Proceedings of the 19th International Congress of Phonetic Sciences*, 1297–1301.

Kügler, F., Baumann, S., & Röhr, C. T. (2022). Deutsche Intonation, Modellierung und Annotation. In C. Schwarze & S. Grawunder (Eds.), *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion: Konzepte, Probleme, Lösungen* (pp. 23–54). Narr Francke Attempto Verlag.

Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., Jannedy, S., Michalsky, J., Niebuhr, O., & Peters, J. (2015). DIMA: Annotation guidelines for German intonation. *Proceedings of the 18th International Congress of Phonetic Sciences*, 317–320. https://kops.uni-konstanz.de/handle/123456789/32841

Rohrer, P., Vilà-Giménez, I., Florit-Pons, J., Esteve-Gibert, N., Ren, A., Shattuck-Hufnagel, S., & Prieto, P. (2020, September 8). The MultiModal MultiDimensional (M3D) labelling scheme for the annotation of audiovisual corpora. *GESPIN 2020*. The 7th Gesture and Speech in Interaction, Stockholm, Sweden.

Rosenberg, A. (2010). AuToBI --- A Tool for Automatic ToBI Annotation. *INTERSPEECH 2010*. 10th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan.