

Multimodal implementation of prominence in overlapping speech

Margaret Zellers¹

¹*Institutionen för lingvistik, Stockholms universitet*

margaret.zellers@ling.su.se

Although, as [1] have argued, speech in dialogue is produced largely by one speaker at a time, conversationalists regularly produce speech in overlap; for example, [2] find that as many as 12% of words and 54.4% of talk spurts produced in two-party telephone conversations are produced at least partially in overlap. Not all overlaps are perceived as such [3], and even if overlapping speech is perceived, it is not necessarily perceived as turn-competitive or otherwise disruptive to the smooth progress of conversation; rather, this can be mediated by prosodic features. For example, research in Conversation Analysis, e.g. [4], has argued that higher fundamental frequency (f0) and increased loudness—that is, increased prosodic prominence—on the part of an incoming speaker are signals that overlapping speech is intended to be turn-competitive. Similarly, [5] report that measures of an incoming speaker’s adaptation of f0 features to his or her interlocutor’s f0 patterns can be used to classify whether episodes of overlapping speech are turn-competitive.

A large body of recent research supports the assertion that prosody and gesture should be treated as a single system within language (cf. discussion in [6]). It has been shown that visual and auditory information are automatically integrated during speech perception [7], and gesture can be used for disambiguation if the speech signal does not provide sufficient information [8]. Overlapping speech might be a case where gesture could be helpful in providing disambiguation, since the multiple speech streams increase the chances that acoustic information is disrupted. On the other hand, there is evidence that prosody and gesture, while functioning in parallel, may convey subtly different information depending on their specific organization [9, 10]. They may also provide information about turn-taking in different timeframes; for example, gestural behaviour has been shown to differ systematically as early as 3 seconds before the offset of a speech turn depending on whether the speaker intends to continue speaking or to stop and let an interlocutor take up a turn [11], while prosodic turn-taking cues have recently been interpreted as late “go-signals” rather than contributors to anticipation of turn ends [12]. Thus it is valuable to investigate the interaction of prosodic and gestural marking of prominence in overlapping speech, and specifically the functional impacts of different feature constellations.

The current exploratory study investigates the multimodal implementation of prominence in overlapping speech in the vicinity of turn ends in German two-party conversation from the DUEL Corpus [13]. The existing annotations for DUEL include annotation of both speakers’ turns with orthographic transcription and laughter labels, and classification of turn-taking behaviour was carried out as part of a different project (cf. [14]). F0 and amplitude envelope features were automatically extracted using Praat [15], and prominent syllables were estimated as local prominences in the amplitude envelope; turns with laughter were excluded from the analysis. Gestures were automatically detected using OpenPose [16]; the current analysis uses the right and left wrist points as representative points to investigate magnitude of gesturing, which is operationalized as velocity. Since the video had a resolution of 25 frames per second, the acoustic data was also extracted at this rate for ease of temporal alignment. Each frame was labelled for its temporal distance from the current speaker’s turn offset and for whether or not the frame involved speech produced in overlap. In the subset (4 speakers) used for the initial analysis, 22.4% of frames involved speech produced in overlap.

Initial results suggest that prosodic prominence features and gesture prominence features are both manipulated in the context of overlapping speech, but that these manipulations are carried out independently. A linear mixed model summed Z-score-normalized values for f0, amplitude, and right- and left-hand velocity finds no overall difference in prominence in turn ends produced in overlap compared to speech produced in the clear. However, investigating the acoustic and gestural activity independently, a more complex picture arises. The amplitude of prominences decreases when approaching a turn end, and this effect is strengthened in overlapping speech (cf. Fig. 1). Meanwhile, velocity of hand movements is higher in speech produced in overlap (cf. Fig. 2), regardless of how near this speech is to an upcoming turn end; as found by [11], movement phases of gesture decrease overall approaching speech offset. No significant results were found for f0 in the initial analysis. These findings suggest that prosodic and gestural prominence are implemented independently of one another for purposes of disambiguation in overlapping speech versus signalling turn-taking intentions.

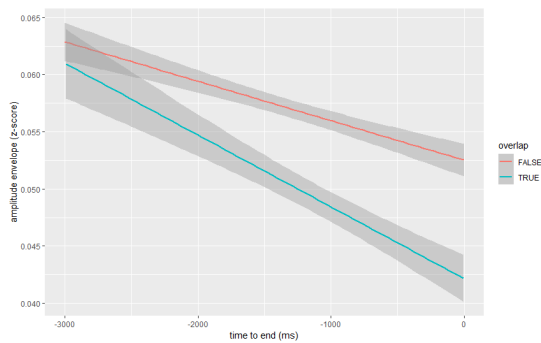


Figure 1: Linear model of Z-score amplitude approaching turn ends in overlap and in the clear.

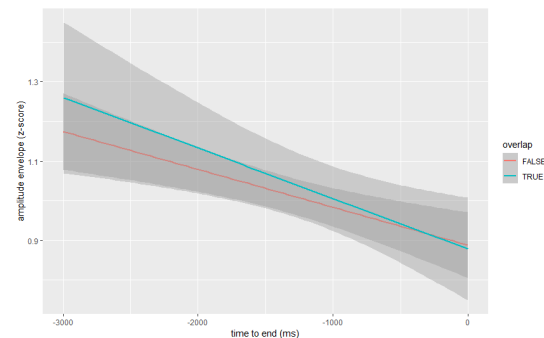


Figure 2: Linear model of Z-score gesture velocity approaching turn ends in overlap and in the clear.

References:

- [1] Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696-735. doi: 10.1353/lan.1974.0010
- [2] Shriberg, E., Stolcke, A., & Baron, D. (2001). Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Aalborg, Denmark, pp. 1359-1362. doi: 10.21437/Eurospeech.2001-352
- [3] Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *Journal of the Acoustical Society of America*, 130(1), 508-513.
- [4] Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1), 1-63. doi:10.1017/S0047404500001019
- [5] Kurtić, E. & Gorisch, J. (2018). F0 accommodation and turn competition in overlapping talk. *Journal of Phonetics*, 71, 376-394. doi: 10.1016/j.wocn.2018.09.006
- [6] Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57, 209-232. doi: 10.1016/j.specom.2013.09.008
- [7] Kelly, S.D., Creigh, P. & Bartolotti, J. (2010). Integrating speech and iconic gestures in a Stroop-like task: evidence for automatic processing. *Journal of Cognitive Neuroscience* 22(4), 683-694. doi: 10.1162/jocn.2009.21254
- [8] Guellaï, B., Langus, A. and Nespors, M. (2014). Prosody in the hands of the speaker. *Frontiers in Psychology* 5, 700.
- [9] Prieto, P., Pugliesi, C., Borràs-Comes, J., Arroyo, E. & Blat, J. (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, 49(1), 41-54. doi: 10.1016/j.wocn.2014.10.005
- [10] Ambrazaitis, G. & House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication* 95, 110-113. doi: 10.1016/j.specom.2017.08.008
- [11] Zellers, M., Gorisch, J. & House, D. (2025) Temporal relationships between speech and hand gestures in the vicinity of potential turn boundaries in German and Swedish conversation. *Language and Cognition*, 17, e57. doi: 10.1017/langcog.2025.10014
- [12] Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final “go-signals”. *Frontiers in Psychology*, 8, 1-10. doi: 10.3389/fpsyg.2017.00393
- [13] Hough, J., Tian, Y., De Ruyter, L., Betz, S., Kousidis, S., Schlangen, D., & Ginzburg, J. (2016). DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, pp. 1784-1788.
- [14] Rossi, M., Schröer, M., Ludusan, B., & Zellers, M. (2023). A multimodal account of listener feedback in face-to-face interactions. In *Proceedings of ICPhS 2023*, Prague, Czechia, pp. 4120-4124.
- [15] Boersma, P. & Weenink, D. (2025). Praat: doing phonetics by computer [Computer program]. Version 6.4.42, retrieved 1 September 2025 from <https://praat.org>
- [16] Cau, Z., Hidalgo Martinez, G, Simon, T., Wei, S. & Sheikh, Y.A. (2019). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1), 172-186. doi: 10.1109/TPAMI.2019.2929257