Temporal Dynamics of Acoustic Emotion Encoding: A Mixed-Effects Modeling Analysis of Valence, Arousal, and Dominance

Yuxin FAN, Yufeng WU

Acoustic features play a central role in expressing both informational and emotional meaning. While previous work has shown that acoustic features reliably predict valence, arousal, and dominance (VAD) in static analyses, the possible contribution of conversational dynamics insufficiently investigated. This study shows how the VAD score of a preceding utterance influences the dynamic relationship between acoustic parameters (prosodic, spectral, and voice quality measures; e.g., F0, intensity, spectral slope, HNR) and the VAD of the subsequent utterance in spoken English dialogues.

A total of 26 GeMAPS-based acoustic parameters were extracted from 5,221 utterances in the Interactive Emotional Dyadic Motion Capture (IEMOCAP) campus. These features were subjected to principal component analysis (PCA) with varimax rotation, resulting in 10 interpretable acoustic factors. To test the temporal dynamics of emotional encoding, linear mixed-effects models (LMEs) were fitted, analyzing associations between acoustic parameters and VAD ratings, with the VAD of the preceding utterance specified as a key moderator. The model included random intercepts and random slopes for the lagged VAD effect by speaker, thus capturing inter-speaker variability.

Beyond this temporal dependency, acoustic main effects remained significant: Valence was negatively associated with 3 factors, Arousal with 7 factors, and Dominance with 8 factors. LME results revealed dynamic associations between emotion and acoustics. Lagged VAD was the strongest predictor across all dimensions (all $p < .001$), demonstrating strong emotional continuity. Moreover, the models confirmed that lagged VAD significantly moderated these associations. Lagged Valence interacted with 4 factors, strongest for Intensity Level ($\beta = 3.459$, $p < .001$); lagged Arousal with 3 factors, strongest for Speech Fluency ($\beta = 3.921$, $p < .001$); and lagged Dominance with 2 factors, also strongest for Speech Fluency ($\beta = 3.575$, $p < .001$). These results underscore the role of temporal dependencies in shaping acoustic–emotion associations

Overall, the results confirm that the perception of emotion in an individual utterance depends not only on its immediate acoustic cues but is also significantly shaped by an “emotional inertia” from the preceding utterance. The influence of acoustic features can thus become stronger, weaker, or even reverse direction depending on their interaction with this prior emotional state. These findings highlight the limitations of context-independent models and demonstrate that incorporating emotional inertia is essential for future speech emotion recognition systems to capture the continuous and dynamic nature of emotion in dialogue with greater accuracy and human-likeness.