Temporal dynamics of acoustic emotion encoding

Yuxin Fan[1], Yufeng Wu[2]
[1]Southeast University, China
[2]City University of Hong Kong, Hong Kong

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/005/000665

Abstract

Static analyses of speech emotion often overlook temporal dependencies. This study examines how the Valence, Arousal, and Dominance (VAD) of a preceding utterance moderate the relationship between acoustic features and the VAD of the preceding utterance. This study fitted linear mixed-effects models to 5,221 utterances from the IEMOCAP corpus. Results showed that the lagged VAD was the strongest predictor among all dimensions, demonstrating significant emotional inertia. Furthermore, the association between acoustic parameters and subsequent VAD was significantly moderated by lagged VAD. These findings confirm that acoustic-emotion associations are dynamic and context-dependent, challenging static models and highlighting the need to incorporate temporal dynamics in emotion recognition systems.

Keywords: speech emotion recognition (SER), affective computing, acoustic features

Introduction

The voice carries a wealth of information that extends beyond linguistic content to convey details of emotions. As a crucial component of interpersonal communication, this paralinguistic channel holds huge research value. One of the most influential patterns for modelling emotion is the Valence-Arousal-Dominance (VAD) model ((Fontaine et al., 2007)), which insists that emotions can be divided into three fundamental dimensions. Valence (V) describes the direction of an emotion, indicating whether it is positive or negative. Arousal (A) describes the intensity of an emotion, referring to the level of physiological and psychological activation. Dominance (D) is a distinct dimension that describes the sense of control experienced during the emotion.

Existing researches has confirmed that acoustic features—such as F0, intensity and spectral slope—can predict VAD scores. However, these analyses share a critical limitation: they are static. While static analysis can capture the relationship between emotions and acoustic parameters at a specific point in time, it fails to capture the dynamics of emotion as it evolves over time, ignoring the contributions of context and temporal sequencing. An individual's emotional is strongly influenced by the preceding conversational content and emotional states. 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.

Methodology

This research used the speech corpus from The Interactive Emotional Dyadic Motion Capture (IEMOCAP). (Busso et al., 2008) IEMOCAP is a widely-used database in affective computing that features dyadic interactions between actors. This study focused on the script sessions to extract clear emotional dynamics.

Based on the GeMAPS feature set (Eyben et al., 2016), this study extracted a total of 26 acoustic parameters across five dimensions: F0, intensity, duration, voice quality parameters (spectral balance), and voice quality parameters (variability). The features were extracted using the parselmouth library (Jadoul et al., 2018) in Python. The VAD scores were sourced from the IEMOCAP database, provided by expert annotators.

Analytical framework

Data processing

The raw dataset consisted of 5255 sentences in total. An initial correlation analysis revealed high multicollinearity among the 26 extracted acoustic parameters. To reduce complexity and improve interpretability, this research used Principal Component Analysis (PCA) (Schuller, 2012).

All acoustic parameters were z-scored and adjusted for speaker effects before PCA. The first 10 components (explaining >80% of variance) were retained and descriptively labelled by their strongest loadings (>.5).

Statistical modelling

To account for these data dependencies, this research used a linear mixed-effects (LME) model approach. The LME framework was chosen for its ability to simultaneously address two sources of non-independence: the clustered nature of the data (multiple sentences per speaker) and temporal dependency.

The model's fixed effect structure included the main effects of the acoustic factors, the main effect of the lagged VAD predictor, and their crucial two-way interactions. The random effect structure was specified to account for by-speaker variation in both baseline VAD levels and sensitivity to the temporal dependency effect.

Results

The LME results revealed highly dynamic associations between emotion and acoustics. First, the VAD of the preceding utterance (lagged VAD) was the

strongest predictor across all dimensions, demonstrating strong emotional continuity.

Crucially, the models confirmed that lagged VAD significantly moderated the associations between subsequent acoustic features and VAD.
Acoustic features themselves remained important predictors even under this strong emotional inertia. After controlling for lagged VAD, multiple acoustic factors were still significantly associated with the three VAD dimensions.

Taken together, our results reveal that the association between acoustic features and VAD is highly contingent on the prior emotional state. For instance, the prior state was found to strengthen, weaken, or in some cases, reverse the direction of a feature's influence. This highlights the complex, dynamic nature of emotional encoding. The following figure 1. provides a clear example of this moderating effect.

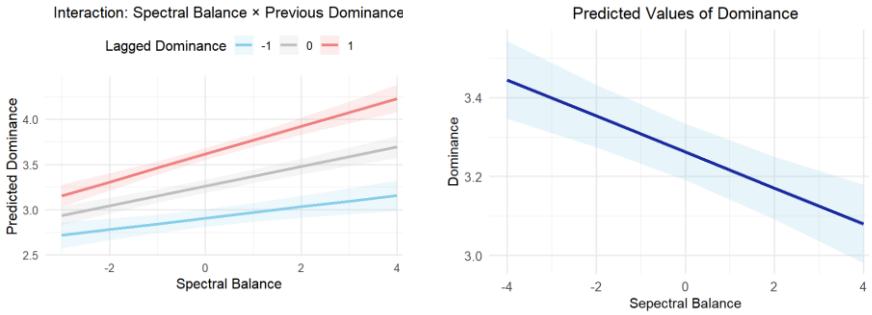

Figure 1. Example of moderating effect.

Discussion

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. 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.

References

Fontaine, J.R.J., Scherer, K.R., Roesch, E.B., Ellsworth, P.C. 2007. The world of emotions is not two-dimensional. Psychological Science 18, 1050-1057.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation 42, 335–359.

Martijn, G. Klaus, S. 2010.Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. Journal of the Acoustical Society of America, 128(3), 1322–1336.

Schuller, B.W. 2012. The computational paralinguistics challenge. IEEE Signal Processing Magazine 29, 97-101.

Eyben, F., et al. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing 7, 190-202.

Jadoul, Y., Thompson, B., de Boer, B. 2018. Introducing Parselmouth: a Python interface to Praat. Journal of Phonetics 71, 1-15.