Activation vs. valence in noise: Diagnosing the phonetic vulnerabilities of an SER system

Tatiana Limonova & Christopher Carignan, *University College London*

Speech Emotion Recognition (SER) systems are notoriously vulnerable to acoustic noise, posing a significant barrier to their practical deployment. This paper systematically investigates the performance degradation of an XGBoost-based SER system to diagnose its primary failure modes in acoustic-phonetic terms. The model was trained on the RAVDESS dataset (768 utterances, 24 speakers) and evaluated against a speaker-independent test set subjected to Additive White Gaussian Noise (AWGN) at seven Signal-to-Noise Ratios (SNRs) from 20dB to -10dB. Four emotions representing distinct quadrants of the valence-activation space were investigated: angry (high-activation, negative valence), happy (high-activation, positive valence), sad (low-activation, negative valence), and calm (low-activation, positive valence). We extracted 92 global acoustic features, including prosodic measures (F0 statistics, RMS energy, zero-crossing rate) and spectral features (13 MFCCs plus their delta and delta-delta coefficients, spectral centroid).

Results demonstrated catastrophic performance degradation, with overall accuracy declining from 75% (clean) to 48% (20dB), 43% (5dB), and 25% (-10dB SNR). Per-emotion F1-score analysis revealed differential vulnerability: 'calm' maintained the highest robustness (F1: 0.56 at -10dB), while 'happy' and 'sad' collapsed at moderate noise levels (5-10dB SNR, equivalent to café or office environments). Confusion matrix analysis showed systematic activation-based grouping under noise; the model preserved distinctions between high-activation (angry, happy) and low-activation (sad, calm) emotions but failed to distinguish valence within these groups.

Feature importance analysis (XGBoost weight metric) revealed over-reliance on activation-related cues: four of the top five features measured signal energy (`stdEnergy` 27%, `maxEnergy` 21%, `mfcc0_mean` 20%, `meanEnergy` 18%). In contrast, valence-encoding dynamic spectral cues, primarily delta-delta MFCCs capturing formant trajectory velocity and spectral envelope acceleration, proved highly noise-vulnerable. This study concludes that for real-world robustness, SER models must leverage noise-resistant phonetic patterns such as long-term spectral tilt, harmonics-to-noise ratio, voice quality measures (e.g., breathiness, tenseness), and speaker-normalised prosodic contours, rather than relying on fragile, frame-level spectral dynamics and raw intensity features.

Keywords: Affective computing, Speech Emotion Recognition (SER), Noise robustness, Acoustic phonetics, Prosody, XGBoost, Performance degradation