The Role of Voice Source and Filter in Speech Emotion Recognition

Yuhan Huang[1]*, Josef Schlittenlacher*[1]*, Chris Carignan*[1]

[1]Department of Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom

yuhan.huang.22@ucl.ac.uk, j.j.schlittenlacher@ucl.ac.uk, c.carignan@ucl.ac.uk

Abstract

Speech emotion recognition is rooted in the acoustic properties of two fundamental components of speech production: The vocal source (i.e., larynx) and the vocal filter (i.e., vocal tract), each contributing distinct idiosyncratic characteristics to the emotional signal. Prior research has lent considerable support to the source-dominance hypothesis, which holds that emotional information is more strongly encoded in the vocal source than in the vocal filter, with pitch dynamics serving as its primary perceptually salient carrier. To isolate the respective contributions of these two components, the present study used whispered speech as an approximation of filter characteristics, and modal speech as a representation of the integrated source-filter signal. A comprehensive acoustic feature set (88 features, eGeMAPS) was applied to both speech modes and fed into an interpretable machine learning framework (XGBoost) for emotion classification, with the aim of separating the influence of the combined source-filter (modal) from that of the primary filter (whisper). Sixty professional actors were recruited and each recorded 20 short sentences across seven emotion categories in both modalities in an anechoic chamber. The relative contributions of source and filter were inferred by quantifying the reduction in model accuracy under cross-mode conditions (e.g., training on modal speech, testing on whispered speech) relative to within-mode baselines (e.g., training on modal speech and testing on modal speech). The results supported the source-dominance hypothesis; performance dropped considerably more across modalities (42.3% versus 20.4%). These findings potentially inform feature selection strategies for emotion recognition by selectively prioritising source or filter components.

Index Terms: Emotion recognition, Cross-mode prediction, Source-filter theory, Whispered speech, XGBoost classification