THE EMOTIONAL PORTRAYAL OF AN ORDINARY TALK

Neda Mousavi[1], *Felix Burkhardt*[2,3]
[1]*HAVAS Media, Germany,* [2]*audEERING GmbH, Germany,* [3]*TU Berlin, Germany*
neda.mousavi@havasmedia.com, fburkhardt@audeering.com

Abstract: This study examines how automatic emotion recognition models capture emotional expression in ordinary speech. Using German autobiographical narratives, the results of linguistic and acoustic models are compared at both the speaker and segment levels. While linguistic emotionality is relatively homogeneous and expressed locally, acoustic emotionality shows greater variability between speakers and stronger temporal persistence. The results suggest that dynamic patterns in model-derived emotionality are modality-dependent and reflect the different roles of lexical content and prosody in ordinary speech.

1 Introduction

Speech-based emotion analysis attempts to derive affective tendencies from spoken language and is motivated by applications in the fields of human-computer interaction, health-related technologies, customer satisfaction analysis, education, security and defense, and related areas of application[1, 2]. Within this field, emotion is typically conceptualized either as a set of discrete categories, such as anger, joy, or sadness [3], or as a continuous multidimensional construct characterized by valence, activation, and dominance [4]. Early research on emotion recognition in speech focused largely on acted or highly elicited emotional speech, which provided a controlled basis for modeling clear and prototypical affective expressions. More recently, the focus of research has expanded beyond the analysis of explicitly emotional or highly elicited data to include the study of more natural speech, in which emotional expressions are often subtle, context-dependent, and intertwined with ordinary communicative content [2, 5]. In daily conversation, emotions are rarely expressed in isolation or exaggeratedly. Instead, affective signals are embedded in ordinary speech and are shaped by the situational context, the speaker's style, and the dynamics of the interaction [6].

This paper examines the emotional coloring of ordinary speech and investigates how affective cues in such data are reflected in the results of models for recognizing emotions in speech. We analyze German speech drawn from an elicited corpus produced using a mood induction procedure [7], which encourages narrative-style speech resembling everyday conversational storytelling. The recordings were not collected with predefined emotional categories in mind, nor were participants instructed to express particular emotional states. The study investigates how emotion recognition models respond to ordinary speech, in which affective cues are implicit rather than explicitly structured. It compares linguistic and acoustic models in terms of how affective signals conveyed through semantic content and prosody are reflected across speakers, and examines the resulting emotional patterns both at an aggregated speaker level and in their temporal dynamics.

2 Data and Method

The data consist of recorded speech from 30 monolingual German speakers (aged 18–40), all with an academic background. Speakers are identified using anonymized IDs, where the prefix

F or M indicates female or male speakers, respectively, followed by a numeric speaker identifier. Recordings were conducted in the acoustic laboratories of Goethe University Frankfurt and Martin Luther University Halle-Wittenberg, with the researcher present. The dataset represents a subset of a larger corpus, *Pertsch*, designed to investigate speech production across multiple speaking tasks [8], from which the present task was selected. Participants were instructed to recall a childhood summer vacation and narrate it in the form of a personal story.

Emotion analysis in the present study follows two complementary dimensions: a semantic dimension based on linguistic content and a prosodic dimension based on vocal characteristics, in line with multidimensional approaches to emotional expression in communication [9]. The analysis was conducted using the nkululeko toolkit [10]. First, the original recordings were converted into the nkululeko data format to generate an initial metadata table containing one entry per speaker. All audio files were then normalized to a uniform format (16 kHz, mono, WAV) and segmented based on voice activity detection (VAD). The audio file transcriptions were generated for each segment using the Whisper ASR model [11] accessed through nkululeko. For acoustic emotion recognition, a categorical model was trained on external emotional speech corpora (EmoDB [12] and EMOVO [13]) to classify segments into five emotional categories (happy, angry, sad, scared, neutral) and subsequently applied to the segmented data. These two databases have been shown to be comparable in the way emotions are portrayed [14].

In parallel, emotional content was inferred from the linguistic content of each segment using the `joeddav/xlm-roberta-large-xnli` large language model integrated into nkululeko, which estimates the semantic compatibility between narrative content and candidate emotion labels in a zero-shot classification setting. The final dataset includes segment-level metadata such as start time, duration, speaker information (speaker ID, gender, age), transcription text, and emotion-related outputs, including predicted emotion labels and corresponding probability estimates for both the linguistic and acoustic modalities.

3 Results

The overall emotional distributions differ noticeably across the linguistic and acoustic modalities when examined at the level of discrete category assignments. In the linguistic modality, predictions are distributed across multiple emotion categories, with happy (42.1%) and angry (28.2%) being the most frequent, followed by neutral (18.7%), scared (7.7%), and sad (3.3%). In contrast, the acoustic modality is dominated by neutral (47.5%) and sad (40.9%) classifications, while happy (2.7%) and scared (8.9%) occur infrequently, and angry is not predicted at all. This pronounced weighting toward neutrality and sadness in the acoustic modality is expected, given that the acoustic classifier was fine-tuned on acted emotional speech and is therefore optimized to detect strong and explicitly marked vocal expressions. By comparison, the broader categorical distribution observed in the linguistic modality suggests that, in this storytelling task, affective information is recognized mostly through lexical–semantic content than through prosodic cues.

However, caution is required when relying on categorical patterns alone, as they illustrate the limitations of discrete emotion labels in ordinary speech, where affective meaning emerges implicitly and may span multiple emotional tendencies. For this reason, and in line with previous work advocating probability-based representations to capture affective variation in natural speech [1], the analysis shifts focus to probabilistic emotion profiles derived from established acoustic and linguistic emotion recognition models. Figure 1 presents the overall emotion probability profiles averaged across all speakers and segments for both models.

In contrast to categorical comparisons, which reflect only the most probable emotion per segment, these probability distributions illustrate how affective information is distributed across

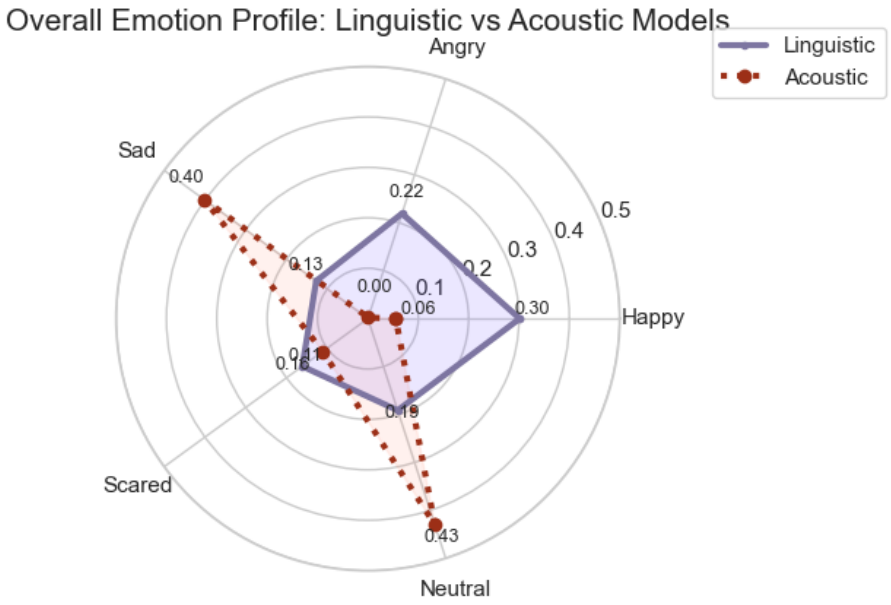

Figure 1 – Mean emotion probability distribution across all speech segments, comparing linguistic and acoustic models.

emotion categories. However, moving beyond this aggregated view, speaker-level inspection, including exploratory radar plots for individual speakers, shows that both the pattern and relative strength of emotional expression vary across speakers, with differences evident in the distribution of emotion probabilities across categories and modalities (figure 2).

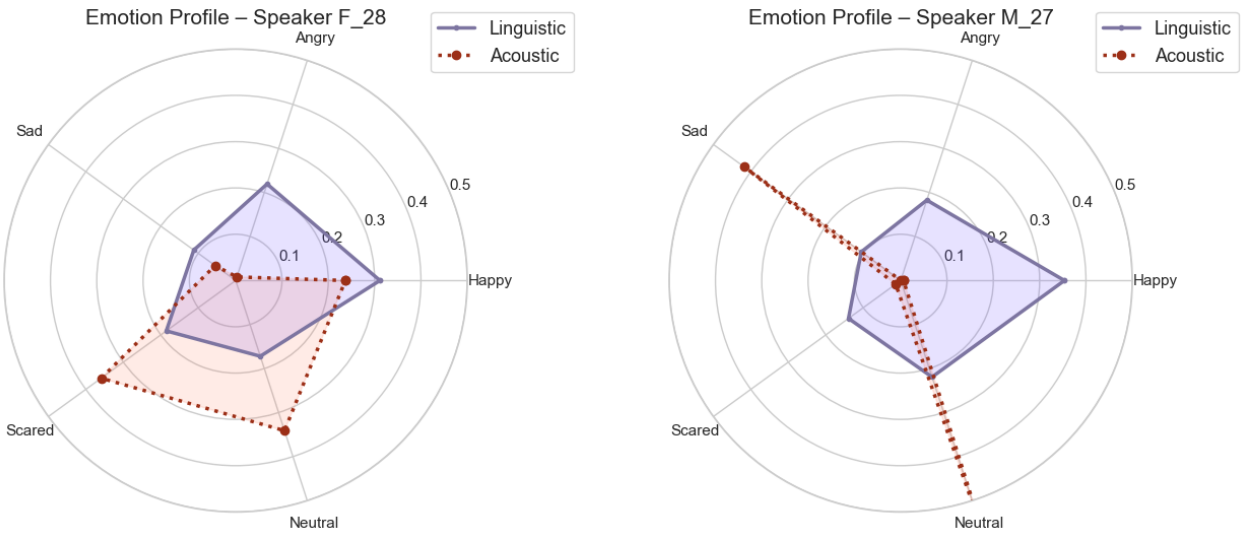

Figure 2 – Example speaker-level emotion probability profiles, illustrating how emotional expression varies across speakers and differs between linguistic and acoustic modalities.

3.1 Between-speaker variability in emotional expression across modalities

Between-speaker variability in emotional expression can be examined from complementary perspectives, capturing both stable speaker-level tendencies and dynamic patterns that unfold over time. As a first step, we adopt a static, aggregated approach to characterize how speakers differ in their overall emotional profiles across modalities.

For each speaker, we calculated an average emotion probability profile in each modality by averaging segment-level probabilities across the five emotion categories. This yields two five-

dimensional vectors per speaker—one based on linguistic predictions and one based on acoustic predictions—summarizing the average distribution of emotional expression for each individual.

The findings show that the extent of between-speaker variability differs markedly across modalities. The linguistic model exhibits very low between-speaker variance (0.0022), indicating that speakers are comparatively similar in the emotional meanings conveyed through lexical–semantic content. This suggests that linguistic cues in the narratives are employed in a broadly consistent manner across individuals, with limited speaker-specific divergence. In contrast, the acoustic model shows substantially higher between-speaker variance (0.0203), indicating that prosodic emotional expression is far more heterogeneous. In other words, while speakers encode affect with relatively similar emotional weighting in *what they say*, they differ considerably in *how emotion is realized in their voice*.

A complementary static perspective is provided by examining *emotionality*, operationalized as $1 - P(\text{Neutral})$, which captures the extent to which a speaker tends to express non-neutral affect (Figure 3). At this level, linguistic emotionality appears relatively homogeneous across speakers, supporting the view that lexical–semantic content conveys emotional meaning in a largely stable, speaker-independent manner. Acoustic emotionality, in contrast, exhibits a much wider spread: while some speakers maintain consistently neutral-sounding prosody throughout their narratives, others show markedly higher levels of vocal expressiveness.

Statistical testing confirms these observations. A Kruskal–Wallis test reveals significant between-speaker variation in both modalities, but with sharply different magnitudes. Linguistic emotionality shows modest speaker differences ($H = 53.50$, $p = 0.0037$), consistent with its low variance. The acoustic modality, by contrast, exhibits very strong between-speaker variation ($H = 191.78$, $p < 10^{-25}$), aligning with the substantially higher variance observed. The findings indicate that the prosodic channel carries a markedly larger component of individual expressive style than the linguistic channel.

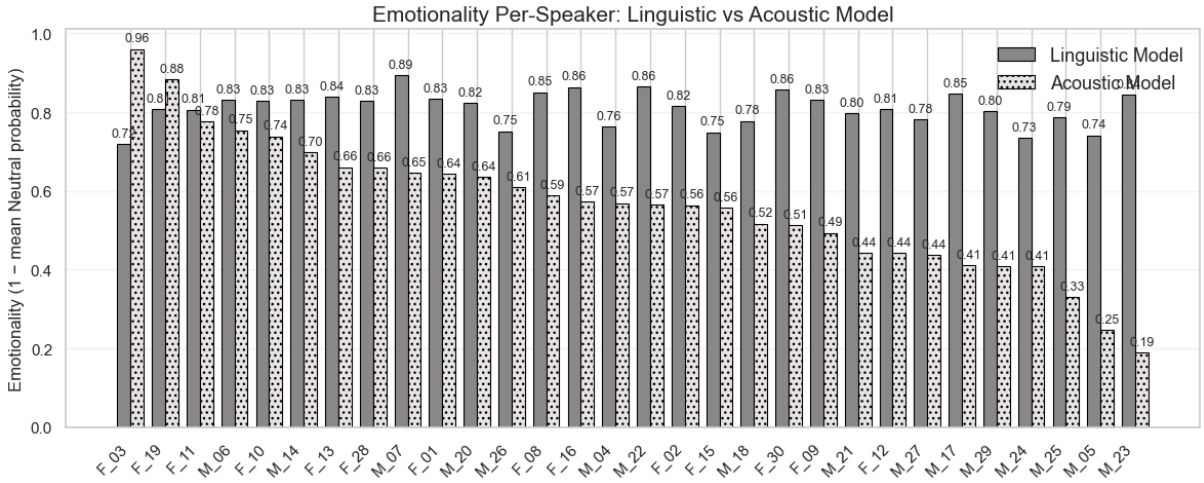

Figure 3 – Emotionality Per-Speaker

Visual inspection of Figure 3 further suggests a potential gender-related pattern in prosodic emotional expression: female speakers tend to appear more frequently toward the higher-emotionality end of the acoustic distribution, whereas linguistic emotionality appears comparatively uniform across genders. To evaluate this observation, emotionality distributions for male and female speakers were compared using the non-parametric Mann–Whitney U test. The results show no significant gender difference in the linguistic modality ($U = 30934, p = 0.191$), indicating comparable lexical–semantic emotional expression across genders. In contrast, the acoustic modality reveals a highly significant gender effect ($U = 21673; p < 10^{-11}$), confirm-

ing that female speakers express stronger non-neutral affect in their prosodic patterns than male speakers.

3.2 Within-speaker temporal dynamics of emotional expression

Beyond speaker-level averages, emotional expression unfolds dynamically over time and may differ in how strongly it persists or changes across successive speech segments. Previous work has approached emotional dynamics from different methodological perspectives, for example, by modeling relative changes in emotional states over time using ranking-based measures [15], or by examining differences in emotion recognition performance across distinct phases of an interaction [16]. More broadly, emotion research has emphasized that affective processes are inherently dynamic rather than static [17, 18]. As a first illustration of this dynamic perspective, Figure 4 shows time series for emotionality for two example speakers and illustrates how emotional expression fluctuates within individual narratives across successive speech segments.

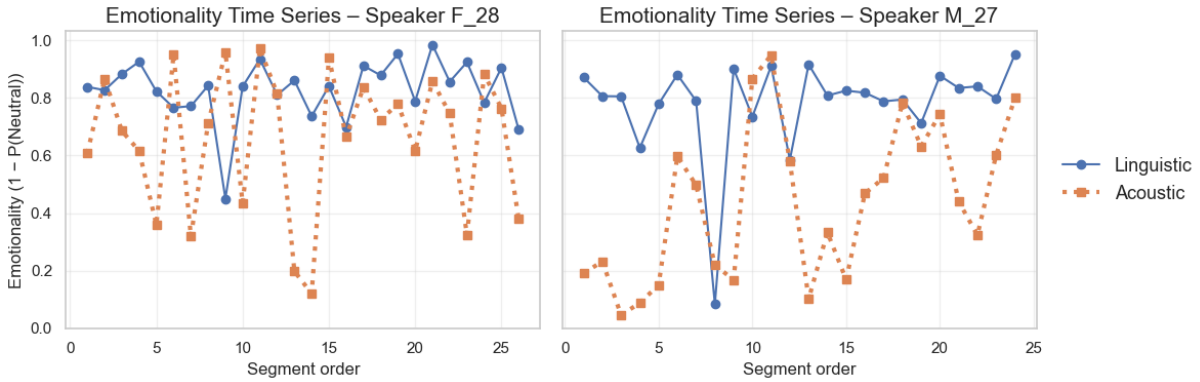

Figure 4 – Emotionality time series for the two example speakers shown in the previous radar plots

In the present study, temporal dependency is operationalized as local autoregressive structure, following approaches in emotion dynamics research that quantify short-term temporal organization using first-order autoregressive (AR(1)) models [19]. For each speaker and modality, emotionality at segment t was regressed on emotionality at segment $t-1$, yielding an autoregressive coefficient (β) indexing temporal persistence, and a residual variance capturing short-term innovation. Models were estimated separately for each speaker after ordering segments by onset time; speakers with fewer than ten adjacent segment pairs were excluded to ensure stable estimation. Because speech segments vary in duration, the resulting autoregressive coefficient is interpreted as reflecting dependency between successive segments rather than persistence over fixed or uniform time intervals.

The results show clear modality-dependent differences in temporal organization. Linguistic emotionality exhibits a uniformly weak autoregressive dependence in all speakers, suggesting that the affective meaning conveyed by lexical content is expressed locally and quickly resets between narrative segments. Acoustic emotionality, on the other hand, shows a stronger and more heterogeneous autoregressive structure as well as greater innovativeness, suggesting that the prosodic expression of emotions tends to persist after its activation, but is highly variable. These results suggest that the temporal structure in emotion recognition results is primarily shaped by the channel of expression and does not reflect a stable, speaker-specific trait, with prosody capturing individual dynamic expression and linguistic content reflecting more transient affective cues.

Figure 5 visualizes speaker-level local temporal dependency in emotionality as estimated using first-order autoregressive (AR(1)) models, comparing linguistic and acoustic modalities.

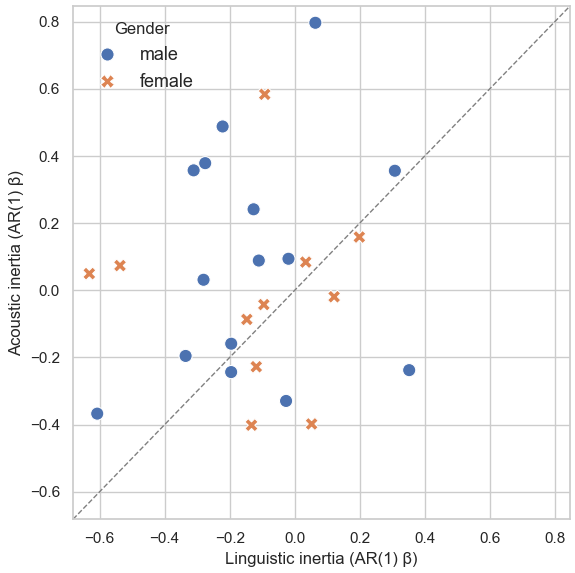

Figure 5 – Speaker-level comparison of linguistic and acoustic emotionality trajectories based on first-order autoregressive (AR(1)) modeling.

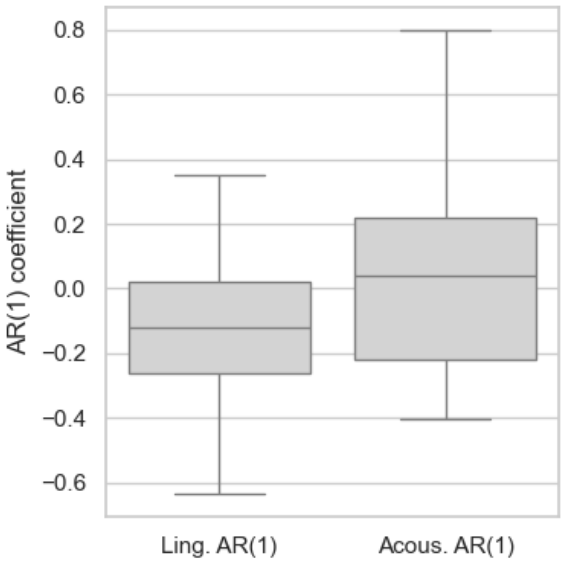

Figure 6 – Distribution of first-order autoregressive (AR(1)) coefficients estimated from linguistic and acoustic emotionality trajectories across speakers.

Each point represents one speaker. The x-axis shows the autoregressive coefficient (β) for linguistic emotionality, while the y-axis shows the corresponding coefficient for acoustic emotionality. The dashed diagonal line indicates equality between modalities ($\beta_{\text{acoustic}} = \beta_{\text{linguistic}}$). Points above the diagonal indicate speakers whose acoustic emotionality is more temporally persistent than their linguistic emotionality, whereas points below the diagonal indicate the opposite pattern. The overall distribution shows that most speakers lie above the diagonal, indicating stronger autoregressive dependency in the acoustic modality. Linguistic AR(1) coefficients cluster tightly around zero, reflecting weak and relatively homogeneous temporal dependency across speakers. In contrast, acoustic AR(1) coefficients span a wider range, including both strongly positive and negative values, pointing to substantial inter-speaker variability in the temporal organization of prosodic emotional expression (Figure 6). Together, these patterns suggest that affective content conveyed through lexical–semantic cues is expressed locally and shows limited carry-over across adjacent segments, whereas prosodic emotional expression is more likely to persist across segments in a speaker-specific manner.

Speaker-level comparisons across modalities confirm these differences in temporal dynamics. Based on 26 speakers with sufficient segment counts in both modalities, paired Wilcoxon signed-rank tests show that acoustic emotionality exhibits significantly stronger temporal persistence than linguistic emotionality ($W = 91.0$, $p = 0.031$), indicating a greater tendency for prosodic emotional expression to carry over across adjacent speech segments. At the same time, innovation magnitude, indexed by the root mean squared error (RMSE) of the AR(1) models, is markedly larger in the acoustic modality ($W = 22.0$, $p < 0.001$), indicating stronger segment-to-segment fluctuations. No association was observed between linguistic and acoustic persistence at the speaker level (Spearman's $\rho = 0.09$, $p = 0.66$), suggesting that temporal dependency does not generalize across modalities within individuals.

4 Discussion

This study explored how automatic emotion recognition models characterize emotional expression in ordinary speech, where affective cues are implicit rather than deliberately produced and no ground-truth emotion labels are available. As a result, the analyses rely on model-derived

estimates rather than verified emotional states. By contrasting linguistic and acoustic model outputs, the study reveals a clear modality-dependent pattern in both the strength and temporal organization of inferred emotionality.

Emotionality derived from linguistic content shows weak and largely uniform temporal dependency across speakers, suggesting that affective meaning conveyed through words is expressed locally and does not form sustained temporal patterns within narration. In contrast, acoustic emotionality exhibits stronger temporal persistence alongside greater short-term variability, indicating that prosodic expression often carries over across adjacent segments while remaining dynamically modulated in a speaker-specific manner. The absence of a systematic relationship between linguistic and acoustic autoregressive coefficients further suggests that temporal structure in model-derived emotional expression is shaped by the expressive channel rather than reflecting a stable speaker-specific characteristic.

The results should be interpreted with several limitations in mind. The absence of reference emotion labels prevents direct evaluation of recognition accuracy and requires interpreting model outputs as comparative signals rather than ground-truth emotions. In addition, the acoustic model was trained on acted emotional speech and applied to ordinary narration, which may introduce domain mismatch effects. The linguistic model operates in a zero-shot setting based on ASR transcripts and is not specifically designed for emotion recognition, making its outputs sensitive to semantic content, transcription quality, and label formulation. Finally, the modest size and narrow scope of the dataset, which focuses on monologic autobiographical narration and does not include other communication situations characteristic of ordinary talk, limit the generalizability of the findings.

5 Code Availability

The analysis code used in this study is available at `https://github.com/felixbur/essv26_emotionalportrayal/tree/main`.

References

[1] MITRA, V., A. ROMANA, D. T. TRAN, and E. AZEMI: *Modeling speech emotion with label variance and analyzing performance across speakers and unseen acoustic conditions*. *arXiv preprint arXiv:2503.22711*, 2025.

[2] LOTFIAN, R. and C. BUSSO: *Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings*. *IEEE Transactions on Affective Computing*, 10(4), pp. 471–483, 2017.

[3] EKMAN, P.: *An argument for basic emotions*. *Cognition & emotion*, 6(3-4), pp. 169–200, 1992.

[4] POSNER, J., J. A. RUSSELL, and B. S. PETERSON: *The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology*. *Development and psychopathology*, 17(3), pp. 715–734, 2005.

[5] MARIOORYAD, S., R. LOTFIAN, and C. BUSSO: *Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora.* In *Interspeech*, pp. 238–242. 2014.

[6] DOUGLAS-COWIE, E., L. DEVILLERS, J.-C. MARTIN, R. COWIE, S. SAVVIDOU, S. ABRILIAN, and C. COX: *Multimodal databases of everyday emotion: facing up to complexity.* In *Interspeech*, pp. 813–816. 2005.

[7] PARADA-CABALEIRO, E., G. COSTANTINI, A. BATLINER, M. SCHMITT, and B. W. SCHULLER: *Demos: An italian emotional speech corpus: Elicitation methods, machine learning, and perception*. *Language Resources and Evaluation*, 54(2), pp. 341–383, 2020.

[8] MOUSAVI, N.: *Individual Speech Rhythm in Persian and German: An Acoustic–Cognitive Approach of Rhythm Production in Different Speaking Tasks*. Phd thesis, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Germany, 2025.

[9] NIU, M., A. ROMANA, M. JAISWAL, M. MCINNIS, and E. MOWER_PROVOST: *Capturing mismatch between textual and acoustic emotion expressions for mood identification in bipolar disorder*. In *Interspeech*. Interspeech, 2023.

[10] BURKHARDT, F. and B. T. ATMAJA: *Nkululeko 1.0: A python package to predict speaker characteristics with a high-level interface*. *Journal of Open Source Software*, 10(115), p. 8049, 2025.

[11] RADFORD, A., J. W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision*. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

[12] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. F. SENDLMEIER, B. WEISS ET AL.: *A database of german emotional speech.* In *Interspeech*, vol. 5, pp. 1517–1520. 2005.

[13] COSTANTINI, G., I. IADEROLA, A. PAOLONI, M. TODISCO ET AL.: *Emovo corpus: an italian emotional speech database*. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, pp. 3501–3504. European Language Resources Association (ELRA), 2014.

[14] BURKHARDT, F., A. HACKER, U. REICHEL, H. WIERSTORF, F. EYBEN, and B. SCHULLER: *A comparative cross language view on acted databases portraying basic emotions utilising machine learning*. In *Proceedings of LREC*. ELRA, 2022.

[15] HAN, W., H. LI, F. EYBEN, L. MA, J. SUN, and B. SCHULLER: *Preserving actual dynamic trend of emotion in dimensional speech emotion recognition*. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pp. 523–528. 2012.

[16] BÖCK, R. and I. SIEGERT: *Recognising emotional evolution from speech*. In *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies*, pp. 13–18. 2015.

[17] HUANG, Z. and J. EPPS: *Prediction of emotion change from speech*. *Frontiers in ICT*, 5, p. 11, 2018.

[18] SCHERER, K. R.: *What are emotions? and how can they be measured? Social science information*, 44(4), pp. 695–729, 2005.

[19] KRONE, T., C. J. ALBERS, P. KUPPENS, and M. E. TIMMERMAN: *A multivariate statistical model for emotion dynamics. Emotion*, 18(5), p. 739, 2018.