Vocalizing Affect: A Comprehensive Review and Taxonomic Synthesis of Publicly Available Speech Emotion Recognition Datasets

Himani Thakor
Ph.D. Scholar,
Department of Computer Science
Veer Narmad South Gujarat University, Surat, Gujarat, India
E-Mail: ht3991214@gmail.com

Abstract:

Speech Emotion Recognition (SER) serves as a primary pillar in affective computing, enabling systems to decode and respond to human expressive communication. This review systematically catalogs and analyzes major publicly available speech emotion datasets, establishing a unified taxonomic framework across three primary paradigms: acted/studio-controlled, elicited/interactive, and spontaneous/conversational. We detail the technical specifications, demographic structures, and annotations of eight leading corpora: Berlin EmoDB, RAVDESS, TESS, Surrey SAVEE, CREMA-D, IEMOCAP, MELD, and the MSP-Podcast corpus. Finally, we evaluate core methodological hurdles, including the domain generalization gap, continuous vs. discrete modeling, and annotator subjectivity, while charting future computational paradigms driven by self-supervised speech foundation models, multimodal fusion, and clinical applications.

1. Introduction:

Speech Emotion Recognition (SER) serves as a primary pillar in the field of affective computing, enabling artificial intelligence systems to decode, model, and respond to human expressive communication. In modern human-computer interaction (HCI), the capacity to interpret vocal effect is essential for deploying robust conversational agents in high-stakes domains, including virtual assistants, clinical mental health surveillance, and empathetic customer support platforms. Emotions modify acoustic parameters systematically: physiological shifts in respiration, phonation, and articulation translate directly to changes in voice source properties, spectral distribution, and speech timing.

Vocal emotional displays are fundamentally multi-dimensional and nuance heavy. Key prosodic features, including fundamental frequency (F0) trajectory, amplitude envelope (intensity), formant-shifting, harmonics-to-noise ratio, and temporal phrasing, convey substantial emotional markers. Traditional machine learning architectures and modern deep learning models, such as self-supervised learning (SSL) speech encoders, rely heavily on the quality, scale, and demographic representativeness of their training data to extract these cues. Consequently, the development and curation of public speech emotion datasets have become indispensable to academic and industrial progress in speech processing.

A critical bottleneck in SER design is the tension between acoustic cleanliness and naturalistic emotional expression. Early datasets favored pristine studio environments and acted performances to ensure high recognizability of targeted affect categories. However, these controlled settings often fail to generalize to complex, real-world conversational dynamics. This review systematically catalogs and analyzes the major publicly available speech emotion

datasets, establishing a unified taxonomic framework, detailing individual database profiles, evaluating core methodological hurdles, and charting future horizons for the discipline.

2. Taxonomic Classification of Speech Emotion Databases

To establish structural clarity across the broad landscape of public SER corpora, databases can be classified into three primary structural paradigms based on the recording environment and emotional generation method. Each paradigm exhibits unique trade-offs regarding phonetic variation, emotional purity, speaker diversity, and acoustic signal-to-noise ratio.

2.1 Acted and Studio-Controlled Paradigms

Acted databases represent the historical foundation of speech emotion research. These datasets employ professional or amateur actors tasked with uttering pre-determined, lexically neutral sentences while portraying specific discrete emotions (e.g., happiness, anger, sadness, fear). The recordings are executed in high-quality, acoustic treated spaces—such as anechoic chambers or professional recording studios—guaranteeing optimal signal quality. The key advantage of this paradigm is the minimization of confounding acoustic variables, such as background noise, overlapping speakers, and lexical variation. This allows feature extraction pipelines (such as Mel-Frequency Cepstral Coefficients, or MFCCs) to isolate emotional vocal features cleanly. However, the primary limitation is a lack of ecological validity: acted portrayals are prone to exaggeration, resulting in stereotypical, archetypal emotional cues that rarely reflect the complex, blended, and muted expressions encountered in daily human communication.

2.2 Elicited and Semi-Spontaneous Interactive Paradigms

Elicited databases seek a middle ground by inducing genuine emotional displays in structured laboratory settings. Instead of reading scripted statements, participants engage in dyadic interactions, improvisational roleplay, or cognitive tasks specifically designed to evoke authentic emotional responses (e.g., frustration through a broken interface or happiness through a joke). While recorded under controlled laboratory conditions to maintain decent acoustic fidelity, the emotional content is semi-spontaneous, capturing natural conversational cues, overlapping turns, and realistic speech rate variations. This approach mitigates the theatrical bias of acted databases, although the laboratory setting still places boundary conditions on emotional expression, and the volume of obtainable data remains constrained by the logistics of in-person capture.

2.3 Spontaneous and Conversational 'In-the-Wild' Paradigms

Spontaneous databases represent the state-of-the-art in real-world application modeling. These corpora are harvested from existing naturalistic audio environments, such as media

broadcasts, talk shows, customer service phone calls, conversational TV series, and podcasts. The emotional displays are completely unscripted and genuine, reflecting the highly subtle, blended, and dynamic nature of real-world human affect. These datasets often include multiple speakers per dialogue, capturing multi-party conversational structures. The primary challenge of this paradigm is the severe presence of acoustic noise, including overlapping voices, laughter, background music, compression artifacts, and ambient environmental sounds. Additionally, labeling spontaneous speech is highly complex and subjective, frequently yielding low annotator agreement and requiring multi-rater crowdsourced annotation strategies.

3. Profiles of Prominent Public Databases

The following detailed profiles examine the technical specifications, development methodologies, and structural compositions of eight leading public databases that have defined modern SER research.

3.1 Berlin Database of Emotional Speech (EmoDB)

The Berlin Database of Emotional Speech (EmoDB), developed by Felix Burkhardt, Astrid Paeschke, Miriam Kienast, Walter F. Sendlmeier, and Benjamin Weiss at the Technical University of Berlin (TUBerlin), stands as a classic and foundational benchmark for German-language SER (Burkhardt et al., 2005). Funded by the German Research Community (DFG Grant no. SE 462/3-1), the dataset consists of high-quality audio recorded in the anechoic chamber of the TU-Berlin department of Technical Acoustics. Ten professional actors (5 male, 5 female) simulated seven emotional states (anger, boredom, disgust, fear, happiness, sadness, and neutral) by producing ten standard German utterances (5 short and 5 longer everyday sentences). These sentences are lexically interpretable in any of the applied emotions, ensuring the lexical content does not bias emotional classification.

To ensure the authenticity and academic utility of the corpus, the authors subjected the initial pool of approximately 800 recorded sentences to a rigorous human perception test evaluating recognizability and naturalness. Only utterances that achieved a human recognition rate greater than 80% and were judged as natural by more than 60% of the listening cohort were retained, resulting in a highly refined collection of 535 samples. Recorded at a sampling rate of 16 kHz with corresponding electro-galactogram (EGG) measurements, the database is partitioned into 4 speakers for testing and 6 speakers for training to support speaker-independent evaluation (Burkhardt et al., 2022).

3.2 The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

Developed by Livingstone and Russo (2018) at Ryerson University, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) represents a highly controlled, multimodal reference database. The dataset contains 7,356 files across WAV (audio-only) and MP4 (video and audio-visual) formats, totaling approximately 24.8 GB of media. RAVDESS features 24 professional actors (12 male, 12 female) vocalizing two lexically matched, North American accented English statements: 'Kids are talking by the door' and 'Dogs are sitting by the door'. The actors portrayed eight discrete emotions: calm, joy, sadness, anger, fear, surprise, disgust, and neutrality.

A unique aspect of RAVDESS is its inclusion of two distinct content types: spoken speech and sung speech. Furthermore, emotional portrayals are executed across two calibrated intensity levels (normal and strong/high), facilitating the study of acoustic scaling under varying degrees of emotional arousal. Because recordings are conducted under pristine studio conditions with high-end recording equipment, the signal-to-noise ratio is exceptionally high. Distributed under a Creative Commons Attribution-Noncommercial 4.0 (CC BY-NC 4.0) license, RAVDESS is widely utilized to evaluate speech-to-emotion, face-to-emotion, and hybrid multimodal fusion architectures.

3.3 Toronto Emotional Speech Set (TESS)

The Toronto Emotional Speech Set (TESS), curated by M. Kathleen Pichora-Fuller and Kate Dupuis (2020) of the University of Toronto Mississauga, was developed to study age-related patterns in auditory perception and speech processing. The dataset consists of 2,800 high-quality audio files (WAV format, 268.3 MB) recorded by two native English-speaking actresses from the Toronto area who are university educated and possess musical training and confirmed normal hearing thresholds. The actresses represent two distinct demographic generations: a young actress aged 26 years and an older actress aged 64 years. This unique age division allows researchers to explore how aging affects vocal emotional generation and acoustic signal properties.

The stimuli in TESS were systematically modeled on the Northwestern University Auditory Test No. 6 (NU-6; Tillman & Carhart, 1966). A set of 200 target monosyllabic words are spoken within the standard carrier phrase 'Say the word ' (e.g., 'Say the word back', 'Say the word bar'). The actors portrayed seven basic emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. Recorded individually in a sound-attenuating booth, TESS is hosted openly on the Borealis Canadian Dataverse Repository under a CC BY-NC 4.0 license,

serving as an ideal platform for speaker-independent, gender-focused, and age-comparative acoustic evaluations.

3.4 Surrey Audio-Visual Expressed Emotion (SAVEE) Database

The Surrey Audio-Visual Expressed Emotion (SAVEE) Database, designed by Philip Jackson and Sanaul Haq at the University of Surrey (2011), serves as a standardized multimodal benchmark for British English. The corpus comprises 480 utterances recorded from four native British English male actors (aged 27 to 31, identified as DC, JE, JK, and KL), who were postgraduate students and researchers at the university. To ensure phonetic completeness, the sentences were systematically selected from the standard TIMIT corpus and phonetically balanced across seven emotional categories: anger, disgust, fear, happiness, sadness, surprise, and neutral.

Recorded in a visual media lab, the dataset provides high-quality synchronized audio (44.1 kHz sampling rate) and video (60 frames per second). Crucially, the recordings included physical 3D blue tracking markers on the subjects' faces, providing spatial facial motion data to trace spatial muscle displacements during emotional speech. The quality of emotional expression was validated by 10 human subjects under audio, visual, and audio-visual conditions. Baseline machine learning experiments on SAVEE achieved speaker-independent recognition rates of 61% for audio-only, 65% for visual-only, and 84% for combined audio-visual modalities, demonstrating the substantial advantage of cross-modal feature fusion (Jackson and Haq, 2015).

3.5 Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D)

The Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D), introduced by Cao et al. (2014) in the IEEE Transactions on Affective Computing, represents a substantial advancement in demographic and speaker scaling. Unlike its predecessor datasets, which feature tiny speaker cohorts, CREMA-D comprises 7,442 original clips recorded from 91 professional actors (48 male, 43 female) between the ages of 20 and 74. The actor cohort is highly diverse, spanning multiple racial and ethnic backgrounds including African American, Asian, Caucasian, Hispanic, and Unspecified ensuring that models trained on the corpus are less susceptible to demographic bias.

The dataset features 12 sentences spoken in six basic emotional states (anger, disgust, fear, happy, sad, and neutral) across four distinct emotional levels (low, medium, high, and unspecified). Because of the large number of ratings required to validate this expanded corpus, the authors leveraged large-scale crowdsourcing. A total of 2,443 human raters evaluated 90 unique clips across three separate conditions: audio-only, visual-only, and combined audio-

visual. Human recognition rates were recorded at 40.9% for audio, 58.2% for visual, and 63.6% for multimodal presentations. Notably, 95% of the clips in the dataset received more than 7 independent human ratings, establishing a highly validated, multimodal labeled corpus.

3.6 Interactive Emotional Dyadic Motion Capture (IEMOCAP)

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database, collected at the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC), is widely regarded as the premier academic benchmark for interactive and dyadic speech emotion recognition (Busso et al., 2008). Spanning approximately 12 hours of audiovisual information, IEMOCAP contains rich, synchronized data streams across multiple modalities, including high-resolution video, speech audio, text transcriptions, and precise facial motion capture. The data is structured across five sessions, totaling 150 conversations and 9,953 dialogue turns, performed by 10 professional theater actors (5 male, 5 female).

IEMOCAP is distinguished by its emotional elicitation methodology. Rather than reading isolated statements, actors performed dyadic sessions featuring both structured, scripted scenarios and unscripted improvisational roleplays specifically designed to elicit authentic emotional displays. During these interactions, facial expressions were captured in high-definition using physical motion-tracking markers. The corpus is annotated by multiple human evaluators. It features discrete, categorical emotion labels (anger, happiness, sadness, neutral, frustration, excitement, surprise, and disgust) as well as continuous dimensional labels describing Valence (positive vs. negative), Activation/Arousal (calm vs. excited), and Dominance (weak vs. strong). This rich dual-annotation scheme makes IEMOCAP a cornerstone for advanced multimodal fusion and conversational SER research.

3.7 Multimodal Emotion Lines Dataset (MELD)

The Multimodal Emotion Lines Dataset (MELD), presented by Poria et al. (2019) at the 57th Annual Meeting of the Association for Computational Linguistics (ACL), serves as a major benchmark for multiparty conversational emotion recognition. MELD is an extension and enhancement of the text-only Emotion Lines dataset. It incorporates audio and visual modalities by harvesting video segments directly from the popular television series Friends. The database consists of 1,433 dialogue episodes and 13,708 individual utterances, representing a massive expansion in scale relative to early lab-recorded databases.

MELD is uniquely structured as a multi-party dataset, regularly containing more than two active speakers per dialogue segment, capturing complex conversational structures, speaker-dependent shifts, and contextual cues. Each utterance is meticulously annotated with one of seven discrete emotion labels (anger, disgust, sadness, joy, neutral, surprise, and fear) as well

as sentiment categories (positive, negative, and neutral). The integration of text, acoustic signals, and facial video underlines the dataset's value for evaluating multimodal baselines and demonstrating how contextual and multi-party dynamics influence emotion classification.

3.8 The MSP-Podcast Corpus

The MSP-Podcast Corpus, developed at the Multimodal Speech Processing (MSP) Laboratory at the University of Texas at Dallas (UT Dallas) under the direction of Carlos Busso, represents a monumental, ten-year effort to build the largest naturalistic, spontaneous emotional speech dataset in the scientific community (Busso et al., 2025). Rather than relying on actors or staged laboratory sessions, the MSP Podcast corpus consists of over 400 hours of diverse audio turns (specifically, Version 2.0 contains 264,705 speaking turns, representing 409 hours of audio) harvested from various public audio-sharing websites under Common Licenses permitting academic distribution. This spontaneous speech reflects genuine, unscripted, real-world emotional displays across a massive variety of conversational topics, environments, and speakers.

To compile this massive resource, the researchers developed a sophisticated, machine learning-driven data collection pipeline that automatically pre-screened raw podcast recordings to retrieve emotionally diverse segments. These retrieved segments were then annotated using crowdsourcing, with at least five raters evaluating each turn. The annotation schema is highly rich, incorporating primary categorical emotions, secondary emotional blends (perceiving multiple co-occurring emotions), and continuous emotional attributes for Valence, Arousal, and Dominance (VAD). Furthermore, the dataset contains manual transcripts of the lexical content and precise speaker identification labels, providing an unparalleled, high-quality resource for training deep learning and SSL models capable of generalizing to practical, in-the-wild conversational applications.

4. Quantitative Synthesis and Comparative Analysis

To facilitate systematic comparison across the public SER landscape, the following matrix synthesizes the eight detailed datasets across key architectural dimensions, including language, modalities, speaker cohort, overall size, and annotation framework. The quantitative differences mapped in Table 1 highlight the fundamental tension in dataset design. Early databases like EmoDB and SAVEE focus on minimizing acoustic variables through small speaker cohorts and rigid studio settings, achieving excellent internal consistency at the expense of external validity. In contrast, modern databases like MELD and the MSP-Podcast corpus trade acoustic cleanliness for unprecedented scale and naturalistic expression, shifting

the burden of emotion classification onto advanced neural models capable of parsing noisy signals.

Dataset	Language	Modalities	Speaker Cohort	Size / Duration	Annotation
Berlin EmoDB	German	Audio (16 kHz WAV), electroglottogram	10 professional actors (5M / 5F)	535 utterances (~1.25 hours)	7 discrete categories (> 80% human consensus)
RAVDESS	English	Audio (WAV), Video (MP4), Audio-Visual	24 professional actors (12M / 12F)	7,356 files (spoken & sung)	8 discrete categories, 2 intensity levels
TESS	English	Audio (WAV, 24.4 kHz)	2 actresses (ages 26 and 64 years)	2,800 audio stimuli (NU-6 words)	7 discrete categories, age-comparative
Surrey SAVEE	English	Audio (WAV), Video (MP4), Facial Markers	4 male actors (postgraduates, 27–31)	480 utterances (TIMIT sentences)	7 discrete categories, human-validated
CREMA-D	English	Audio (16 kHz WAV), Video (MP4), Audio-Visual	91 diverse actors (48M / 43F, 20–74)	7,442 clips, multi-racial	6 discrete emotions, 4 intensity levels
IEMOCAP	English	Audio, Video, Facial MoCap, Transcripts	10 professional actors (5M / 5F)	9,953 conversational turns (~12 hrs)	8 discrete emotions, continuous VAD
MELD	English	Audio, Video (MP4), Text Transcripts	Conversational (multiple actors from Friends)	13,708 utterances from 1,433 dialogues	7 discrete emotions, 3 sentiment categories
MSP-Podcast	English	Audio (WAV), Speaker ID, Transcripts	Spontaneous (hundreds of podcast hosts)	264,705 turns (>400 hours, V2.0)	Primary/Secondary categories, VAD

5. Methodological Bottlenecks and Emerging Challenges

5.1 The Domain Generalization and Acted-to-Spontaneous Gap

The primary challenge facing contemporary SER models is domain generalization. Models trained exclusively on acted datasets (such as RAVDESS or EmoDB) routinely experience catastrophic performance degradation, often exceeding a 30% drop in unweighted accuracy—when evaluated on spontaneous speech datasets (such as EMOVOME or MSP-Podcast). This generalization bottleneck stems from the stylistic difference between acted and real-life effect. Actors convey emotions through hyper articulation and exaggerated prosody to maximize categorical distinction. In contrast, spontaneous real-life speech is highly subtle, context-dependent, and frequently blended with neutral or overlapping conversational states. This performance gap highlights the limitations of traditional machine learning pipelines and underscores the necessity of training on diverse, spontaneous corpora to achieve practical utility in real-world applications.

5.2 Continuous (Dimensional) vs. Discrete Modelling

The representation of human emotion has long been divided between discrete categorical theories (such as Ekman's six basic emotions: anger, disgust, fear, happiness, sadness, and surprise) and continuous dimensional frameworks (Valence, Arousal, and Dominance, or VAD). While discrete categories are highly intuitive and straightforward to classify, they are conceptually rigid and cannot represent intermediate emotional states, blended expressions, or dynamic transitions. Modern benchmarks like IEMOCAP and MSP-Podcast address this by annotating both discrete categories and continuous VAD dimensions. Continuous dimensional tracking allows speech models to map emotions as a continuous space, capturing subtle variations in intensity and pleasantness. However, continuous annotation is cognitively demanding for human raters, resulting in higher subjectivity, greater variance, and requiring complex statistical aggregation methods (e.g., Evaluator Weighted Estimator) to establish reliable consensus.

5.3 Annotator Subjectivity and Demographic Bias

Vocal emotional perception is inherently subjective, shaped by cultural background, gender, age, and individual experience. Consequently, ground-truth annotations in public datasets exhibit significant variance. In conversational settings, evaluators frequently assign highly mismatched emotional labels to the same dialogue turn, illustrating the complexity of labeling spontaneous effect. To manage this annotator disagreement, modern datasets are moving away from simple majority-voting toward advanced disagreement models that treat annotator variance as a valuable distribution of emotional perception rather than noise.

Furthermore, early datasets suffer from severe demographic, geographic, and gender biases. Corpora with tiny speaker cohorts (e.g., SAVEE's 4 male speakers, or EmoDB's 10 German actors) cannot adequately capture the phonetic diversity of global speech. Models trained on such narrow cohorts generalize poorly across different accents, age groups, and genders. Modern datasets increasingly prioritize stability and fairness; for instance, TESS incorporates an explicitly age-comparative actress split, and CREMA-D balances its 91-actor cohort across gender, age, and diverse ethnic backgrounds (African American, Asian, Caucasian, and Hispanic) to improve algorithmic fairness and robust speaker-independent validation.

Core Methodological Insight: *The traditional view of treating evaluator variance as 'noise' is being replaced by disagreement modeling. Because vocal emotional perception is fundamentally shaped by an individual's demographic and cultural background, annotator variance represents a rich distribution of perceptual truths rather than a labeling error. Models that capture this distribution demonstrate superior robustness in complex, real-world deployments.*

6. Future Horizons and Computational Paradigms

6.1 Self-Supervised Learning and Pre-trained Speech Foundations

The integration of self-supervised learning (SSL) models—most notably Wav2Vec 2.0, HuBERT, and WavLM—has revolutionized SER. Instead of training feature extraction pipelines from scratch on tiny, labeled corpora, SSL foundation models are pre-trained on massive volumes of unlabeled raw audio (often exceeding tens of thousands of hours, such as LibriSpeech). This pre-training allows the model to learn rich, robust acoustic representations of speech, including phonetic structures, prosodic envelopes, and speaker characteristics. When fine-tuned on target emotion datasets (like IEMOCAP or RAVDESS), SSL models achieve state-of-the-art accuracy while requiring significantly fewer labeled samples. Crucially, pre-trained speech representations act as powerful domain generalizers, substantially mitigating the acted-to-spontaneous transfer gap.

6.2 Multimodal Fusion and Conversational AI

As human communication is inherently multimodal, future progress in SER lies in the tight coupling of speech audio, facial dynamics, and lexical text. Multimodal datasets (including IEMOCAP, MELD, and RAVDESS) have proven that integrating facial expression videos or textual transcripts significantly boosts classification accuracy. For example, the presence of text transcripts allows language models (e.g., fine-tuned BERT or RoBERTa) to capture semantic and contextual sentiment, which directly complements the acoustic features extracted

from the voice. Future architecture will focus on advanced fusion techniques—such as cross-attention mechanisms and late decision fusion—to model multi-party conversations, enabling conversational AI agents to engage in highly empathetic and context-aware human-computer dialogue.

6.3 Clinical and Mental Health Applications

Beyond commercial applications, SER datasets are increasingly leveraged in clinical psychology and psychiatric monitoring. Changes in speech patterns—including speech rate deceleration, flat prosodic intonation, and increased vocal jitter—are known acoustic markers of depressive disorders, anxiety, and post-traumatic stress. By validating speech emotion models on highly accurate, continuous emotional attributes (such as Valence and Arousal), developers can deploy non-invasive, automated monitoring systems to assist healthcare professionals in tracking patient mental health trajectories. The validation of clinical speech databases remains a critical frontier, demanding stringent privacy standards, calibrated diagnostic guidelines, and robust demographic generalization to ensure safety and therapeutic efficacy.

7. References

1. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In *Proceedings of InterSpeech 2005* (pp. 1517–1520).
2. Burkhardt, F., Paeschke, A., Kienast, M., Sendlmeier, W. F., & Weiss, B. (2022). *Berlin EmoDB* (Version 1.3.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7447302
3. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Provost, E. M., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42*(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6
4. Busso, C., Lotfian, R., Sridhar, K., Salman, A. N., Lin, W. C., Goncalves, L., Parthasarathy, S., Naini, A. R., Leem, S., Martinez-Lucas, L., Chou, H. C., & Mote, P. (2025). The MSP-Podcast Corpus. *arXiv preprint arXiv:2509.09791*. https://arxiv.org/abs/2509.09791
5. Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing, 5*(4), 377–390. https://doi.org/10.1109/TAFFC.2014.2360010

6. Jackson, P., & Haq, S. (2011). *Surrey Audio-Visual Expressed Emotion (SAVEE) Database* [Dataset]. University of Surrey. http://kahlan.eps.surrey.ac.uk/savee/
7. Jackson, P., & Haq, S. (2015). *Surrey Audio-Visual Expressed Emotion (SAVEE) Database: Human evaluation and baseline experiments* (CVSSP Technical Report). University of Surrey.
8. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLoS ONE*, *13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391
9. Pichora-Fuller, M. Kathleen, & Dupuis, Kate (2020). *Toronto emotional speech set (TESS)* (Version 1.0) [Dataset]. Borealis: The Canadian Dataverse Repository. https://doi.org/10.5683/SP2/E8H2MF
10. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 527–536). Florence, Italy. https://doi.org/10.18653/v1/P19-1050
11. Tillman, T. W., & Carhart, R. (1966). *An evaluation of Northwestern University Auditory Test No. 6* (Technical Report). USAF School of Aerospace Medicine, Brooks Air Force Base, Texas.