MEASURING EMOTION PRESERVATION IN EXPRESSIVE SPEECH-TO-SPEECH TRANSLATION

Bagus Tris Atmaja, Toru Shirai, Sakriani Sakti

Nara Institute of Science and Technology, Ikoma, Japan

ABSTRACT

Measuring how well emotions are preserved in speech-to-speech translation is a difficult task. Previous research focused on measuring the similarity between speech embeddings related to emotion or prosody, which may not accurately capture the emotional content of both the source and target utterances. This study proposes a more direct approach by evaluating emotion preservation using metrics derived from speech emotion recognition (SER) models, including balanced accuracy ratio, emotion preservation rate, Cohen's Kappa, and other previous metrics. The results show that about half of the original emotion is preserved in the translation process in the MELD-ST dataset, with metrics such as balanced accuracy ratio, valence-arousal similarity, and pause rate serving as reliable indicators of emotion preservation. We also found that high similarities between emotion embeddings do not necessarily mean emotion preservation, since the same acoustic embeddings used for SER lead to distinct performance. The analysis highlights the challenges of maintaining emotional consistency during speech-to-speech translation.

Index Terms— speech-to-speech translation, emotion preservation, speech emotion recognition, expressive speech

1. INTRODUCTION

Expressive speech-to-speech translation (S2ST) systems aim to preserve not only linguistic content but also emotional expression across language boundaries. While significant advances have been made in maintaining semantic accuracy, the preservation of emotional nuances remains a fundamental challenge that directly impacts the naturalness of cross-lingual communication [1, 2]. Current evaluation approaches for emotion preservation primarily rely on embedding similarities or prosodic feature comparisons, which may not adequately capture the perceptual quality of emotional transfer from the perspective of human listeners [3, 4].

Recent studies have explored various methodologies for assessing emotion preservation in speech translation systems. Some approaches focus on acoustic similarity measures between source and target utterances [5], while others examine prosodic characteristics such as fundamental frequency patterns and temporal dynamics. However, these indirect measures often fail to provide clear insights into whether the translated speech conveys the same emotional message as perceived by human listeners. The lack of standardized evaluation protocols for emotional content preservation has hindered systematic comparison across different S2ST architectures and training methodologies.

This study addresses the limitations of existing evaluation approaches by proposing a direct assessment framework based on speech emotion recognition (SER) models. Our contribution is threefold: (1) we introduce novel metrics including balanced accuracy ratio (BAR), emotion preservation rate (EPR), and Cohen's Kappa to directly measure emotion preservation through categorical emotion recognition, (2) we demonstrate that commonly used embedding-based similarity metrics may not reliably indicate emotion preservation, as evidenced by the contradiction between high embedding similarity and distinct emotion recognition performance using the same features, and (3) we establish a comprehensive evaluation framework that identifies reliable metrics for emotion preservation assessment, i.e., BAR, va-sim, and Pause rate, that showed approximately half of emotion is maintained across language pairs in the MELD-ST dataset. These findings provide crucial insights for developing more emotionally-aware speech translation systems and establish methodological standards for future evaluation studies.

2. DATASET AND SER METHODS

This research used MELD-ST dataset [6], an expressive speech-to-speech translation dataset comprising English-to-Japanese (ENG-JPN) and English-to-German (ENG-DEU) language pairs based on MELD dataset [7]. The number of samples for ENG-DEU pair is 10085, while ENG-JPN pair is 11637 (we excluded samples less than 1 kB). The dataset contains 7 emotion classes: anger, disgust, fear, joy, neutral, sadness, and surprise. We used the original split of training, validation, and test sets to measure the performance of speech emotion recognition (SER) model.

We calculated metrics to measure emotion preservation

Part of this work was supported by JSPS KAKENHI Grant Numbers JP21H05054, JP23K21681, and JP24K0296, as well as JST NEXUS (JPMJNX25C1).

on expressive speech-to-speech translation based on the SER model. We used the same SER model for all language pairs, that is EN in ENG-JPN, JA in ENG-JPN, EN in ENG-DEU, and DE in ENG-DEU. The SER model is based on the audio features extracted from the emotion2vec-plus-large model (1024-dim), fed into a two-layer feedforward network (64 and 32 nodes), with a batch size of 2, a dropout rate of 0.3, and an maximum epoch of 80 using the Nkululeko toolkit [8]. This feedforward model worked well in the previous SER study [9]. We addressed the class imbalance in the dataset by using the cluster centroids method from the imbalanced-learn package [10].

3. EVALUATION METRICS

We evaluated multiple metrics below to judge the preservation of emotion transferred from one language to another.

Balanced Accuracy Ratio (BAR) is the ratio of balanced accuracy between the target language and the source language. BAR is defined as follows:

$$BAR = \frac{UA_{target}}{UA_{source}}, \tag{1}$$

where UA is the unweighted average of the accuracy of each emotion class (balanced accuracy).

Emotion Preservation Rate (EPR) is the ratio of the number of utterances that have the same predicted emotion label in both source and target languages to the total number of utterances. EPR is defined as:

$$EPR = \frac{N_{same}}{N_{total}}. \tag{2}$$

Kappa is the Cohen's Kappa score [11] which measures the agreement between the predicted emotion labels in source and target languages. Kappa is defined as follows:

$$Kappa = \frac{p_o - p_e}{1 - p_e}, \tag{3}$$

where p_o is the observed agreement between the predicted emotion labels in source and target languages, and p_e is the expected agreement by chance; Kappa adjusts for random agreement, whereas previous EPR does not.

emo-sim is the frame-level (averaged) cosine similarity of emotion2vec model (large version) [12], between the source and target languages.

va-sim is the frame-level (averaged) cosine similarity of valence and arousal predicted by a model [13], between source and target language. The basis for calculating emo-sim and va-sim is EmoCtrlTTS-Eval [14].

autoPCP is the utterance-level prosodic similarity between the target and source languages. The extraction of prosody information is based on XLS-R 53 embeddings [15].

Vsim is the cosine similarity of vocal style measured from WaVLM Large embeddings [16] between source and target speech.

Pause is the Spearman correlation of the pause rate between the source and target speech.

Rate is the Spearman correlation of speech rates between source and target speech.

4. EVALUATION RESULTS

4.1. Multiple metrics evaluation

Table 1 presents an evaluation of emotion preservation across multiple complementary metrics for expressive speech-to-speech translation on the MELD-ST dataset. The results reveal distinct patterns across different evaluation dimensions that collectively characterize the effectiveness of emotion transfer between language pairs. We calculated the first five metrics, while autoPCP, Vsim, Pause, and Rate are taken from from previous work [6] (no fine-tuning model).

The discrete emotion recognition metrics (BAR, EPR, Kappa) demonstrate moderate to low preservation rates across both language pairs. The Balanced Accuracy Ratio indicates that approximately 57% of the original emotion recognition performance is maintained in both EN-JA (0.577) and EN-DE (0.568) translations, suggesting comparable emotion preservation capabilities across target languages. However, the Emotion Preservation Rate reveals substantially lower direct emotion label agreement, with EN-JA achieving 21.6% and EN-DE achieving 25.5% exact label preservation. The Cohen's Kappa scores (0.086 for EN-JA, 0.074 for EN-DE) indicate poor inter-rater agreement between source and target emotions, falling within the "slight agreement" range according to conventional interpretation guidelines.

In contrast to discrete emotion metrics, embedding-based measures demonstrate substantially higher preservation rates. However, the interpretation of these results requires careful consideration. The emotion2vec similarity scores achieve remarkably high values (EN-JA: 0.890, EN-DE: 0.939), which appears contradictory given that the same emotion2vec embeddings used as SER features yield low recognition performance. This contradiction suggests that high cosine similarity in the emotion2vec embedding space may not necessarily correspond to preserved emotion semantics, but rather to general acoustic-linguistic similarities that are maintained during translation. The high emo-sim scores may reflect the preservation of non-emotional speech characteristics (e.g., prosodic patterns, spectral features) that are captured by the emotion2vec model but are not discriminative for emotion classification.

The valence-arousal similarity maintains moderate preservation (EN-JA: 0.544, EN-DE: 0.559), which is more consistent with the observed discrete emotion recognition performance. This dimensional approach may provide a more reliable indicator of genuine emotion preservation, as it measures continuous emotional dimensions rather than high-dimensional embedding similarities that may be dominated

Table 1. Evaluation metrics for emotion preservation of expressive speech translation in MELD-ST dataset

Data	BAR	EPR	Kappa	emo-sim	va-sim	autoPCP	Vsim	Pause	Rate
ENG-JPN	0.577	0.216	0.086	0.890	0.544	1.75	0.0034	0.501	-0.09
ENG-DEU	0.568	0.255	0.074	0.939	0.559	2.00	0.0091	0.501	0.09

by non-emotional acoustic features.

Prosodic similarity metrics reveal mixed preservation patterns. The autoPCP scores (EN-JA: 1.75, EN-DE: 2.00) indicate moderate prosodic feature preservation, with German translation demonstrating slightly superior prosodic transfer. Vocal style similarity (Vsim) shows minimal preservation (EN-JA: 0.0034, EN-DE: 0.0091), suggesting that speaker-specific vocal characteristics are largely lost during the translation process. Pause rate correlation maintains moderate preservation (0.501 for both language pairs), while speech rate correlation varies substantially between language pairs (EN-JA: -0.09, EN-DE: 0.09), indicating language-specific temporal adaptation effects.

The convergent patterns observed across multiple metrics provide strong evidence for reliable emotion preservation measurement. Metrics that demonstrate similar relative performance across language pairs (BAR: 0.577/0.568, va-sim: 0.544/0.559, pause: 0.501/0.501) exhibit convergent validity, suggesting they capture fundamental aspects of emotion preservation that are robust across different linguistic contexts. This convergence is particularly valuable because it indicates that these metrics are measuring genuine emotion preservation phenomena rather than artifacts of specific measurement approaches or language-dependent confounds. Notable diverse patterns are found include vocal style similarity (3-fold difference) and speech rate correlation (opposite signs), indicating that these metrics may be not be suitable for emotion preservation measurement.

The high consistency across complementary measurement paradigms—categorical emotion recognition (BAR), dimensional emotion representation (va-sim), and temporal-prosodic features (pause correlation)—strengthens confidence in the observed emotion preservation patterns. Conversely, metrics showing substantial cross-linguistic variation (Vsim, speech rate correlation) may be capturing language-specific acoustic adaptations or measurement noise rather than core emotion preservation characteristics.

4.2. Speech emotion recognition results

Table 2 presents the speech emotion recognition (SER) performance across different language pairs, evaluated using both unweighted accuracy (UA) and weighted accuracy (WA) metrics. The results are stratified by the application of cluster centroids balancing technique to address class imbalance in the dataset.

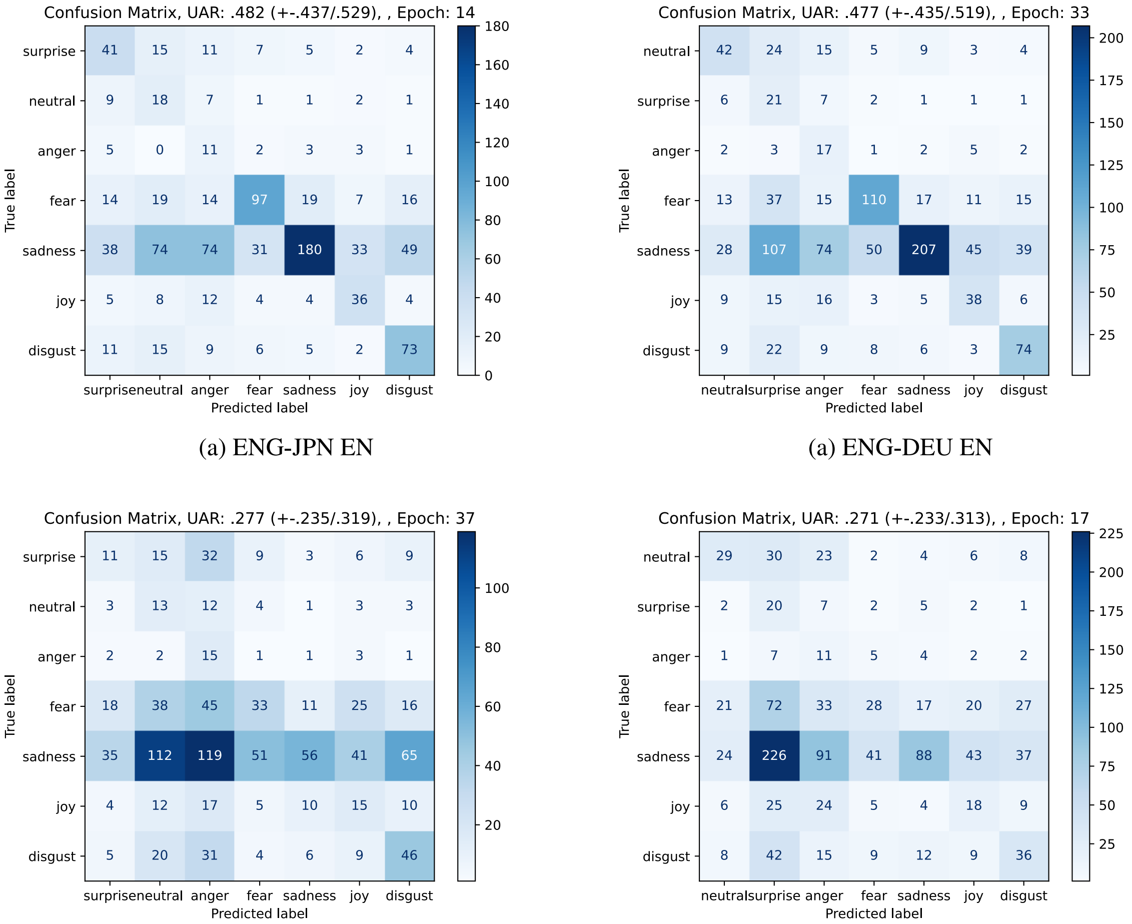
The cluster centroids balancing method demonstrates markedly different effects on UA and WA metrics. With balancing, UA scores are substantially higher across all languages (English: 0.483/0.477, Japanese: 0.277, German: 0.271), while WA scores are correspondingly lower (English: 0.452/0.437, Japanese: 0.182, German: 0.197). Conversely, without balancing, UA scores decrease significantly (English: 0.454/0.414, Japanese: 0.176, German: 0.190), whereas WA scores increase dramatically (English: 0.627/0.623, Japanese: 0.481, German: 0.480). This inverse relationship indicates that balancing effectively addresses class imbalance by improving performance on minority emotion classes, as reflected in higher UA, but potentially compromises overall accuracy weighted by class frequency.

Table 2. Emotion recognition performance on the MELD-ST dataset (UA: Unweighted Accuracy, WA: Weighted Accuracy)

Language Pair	Language	UA	WA
	w/ balancing		
ENG-JPN	EN	0.483	0.452
	JA	0.277	0.182
ENG-DEU	EN	0.477	0.437
	DE	0.271	0.197
	wo/ balancing		
ENG-JPN	EN	0.454	0.627
	JA	0.176	0.481
ENG-DEU	EN	0.414	0.623
	DE	0.190	0.480

Average class accuracy (UA) is important for evaluating the model's ability to recognize emotions across all classes, especially in imbalanced datasets. However, the previous studies on MELD dataset [17, 18, 19] failed to report these metrics, which are crucial for understanding the model's performance on minority classes. They only reported overall accuracy to show the competitiveness of their audio models (kimi-audio, qwen-audio). Although the balanced accuracy obtained in this study is lower than general SER studies, the score (UA: 0.483) is higher than unimodal scores in the baseline MELD model (UA: 0.443) [7] and conversational transformer network (0.469) [20].

Both Japanese and German target languages exhibit remarkably similar SER performance patterns, despite their linguistic diversity. In the balanced condition, both languages achieve comparable UA scores (Japanese: 0.277, German: 0.271) and WA scores (Japanese: 0.182, German: 0.197). Similarly, in the unbalanced condition, both languages demonstrate analogous performance degradation



(a) ENG-JPN EN

(a) ENG-DEU EN

(b) ENG-JPN JA

(c) ENG-DEU DE

Fig. 1. Confusion matrix for each language in the language pairs

patterns with UA scores of 0.176 and 0.190, and WA scores of 0.481 and 0.480, respectively. This consistency suggests that the emotion recognition challenges in speech translation may be more dependent on the translation process itself rather than the specific target language characteristics.

A substantial performance gap exists between source English and target languages across all conditions. The original English source consistently outperforms both Japanese and German by approximately 0.2 points in UA under balanced conditions and maintains superior performance in unbalanced conditions. This disparity likely reflects the inherent challenges in emotion preservation during the speech translation process, where emotional nuances may be attenuated or altered during cross-lingual transfer. By calculating UA ratio, the preserved emotion based on SER model is about half of the original emotion in source language.

Figure 1 presents the confusion matrices for speech emotion recognition across source and target languages, revealing distinct performance patterns that illuminate the challenges of emotion preserving emotion in expressive speech translation. The matrices demonstrate systematic differences between source English and translated target languages that corroborate the quantitative metrics presented in Table 2.

The English confusion matrices for both language pairs exhibit relatively balanced diagonal dominance, indicating moderate but consistent emotion recognition capabilities across most emotion categories. The ENG-JPN EN matrix shows reasonably strong recognition for anger, joy, and surprise, with notable confusion primarily between semantically related emotions (e.g., sadness-neutral, fear-surprise). Similarly, the ENG-DEU EN matrix demonstrates comparable performance patterns, with slight variations in the confusion patterns between specific emotion pairs, confirming the similar BAR scores observed across the two translation directions.

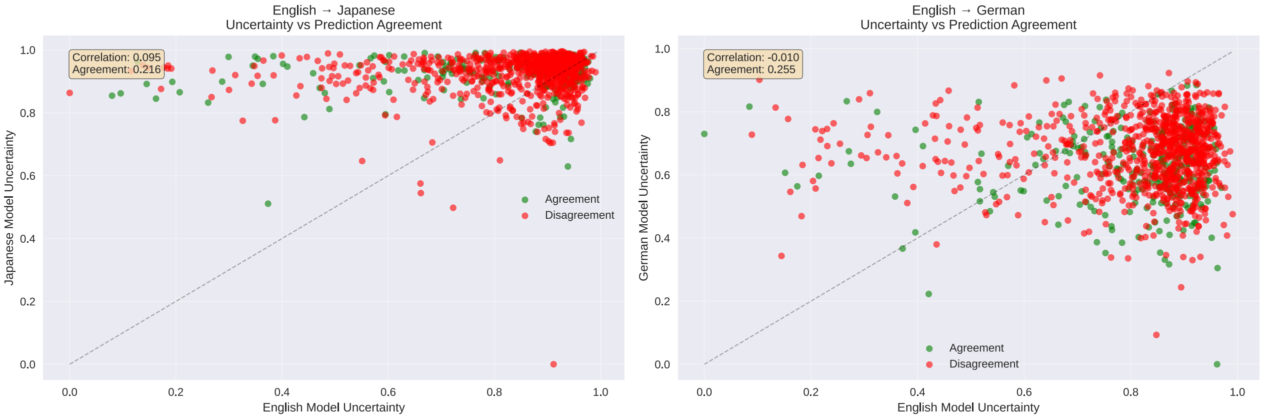


Fig. 2. Uncertainy vs. agreement plot of SER for ENG-JPN model (left) and ENG-DEU model (right)

The translated Japanese and German confusion matrices reveal substantial performance degradation compared to their English counterparts. Both target languages exhibit markedly weaker diagonal elements and increased off-diagonal confusion, consistent with the approximately 50% reduction in UA scores observed in Table 2. The Japanese matrix shows particular difficulty in distinguishing between multiple emotion categories, with significant confusion clustering around neutral emotions. The German matrix demonstrates similar degradation patterns, though with slightly different confusion distributions, particularly affecting discrete emotions like disgust and fear.

Both Japanese and German target languages exhibit analogous degradation patterns despite their linguistic diversity. Both matrices show similar reductions in recognition confidence (lighter diagonal elements) and comparable increases in systematic confusion patterns. This consistency supports the conclusion that emotion recognition challenges in speech translation are primarily process-dependent rather than target language-specific, reinforcing the similar BAR and UA scores observed for both target languages.

The confusion matrices reveal that certain emotions are more susceptible to translation-induced degradation than others. High-arousal emotions (anger, surprise) tend to maintain relatively better recognition in target languages, while lower-arousal emotions (sadness, neutral) exhibit more substantial confusion increases. This pattern suggests that arousal-related acoustic features may be more robust to the speech translation process than valence-related characteristics, providing insight into the dimensional emotion preservation patterns observed in the va-sim metric.

Figure 2 presents the uncertainty versus agreement analysis for both English-Japanese and English-German language pairs, providing a comprehensive visualization of emotion preservation quality at the utterance level. This analysis plots the relationship between model prediction uncertainty (y-axis) and source-target emotion label agreement (x-axis), where each point represents an individual utterance colored by its agreement status (red for disagreement, blue for agreement). The theoretical ideal for effective emotion preservation would manifest as a concentration of samples in the bottom-left quadrant (low uncertainty, high agreement), indicating confident and consistent emotion recognition across languages.

Both English-Japanese and English-German pairs exhibit remarkably similar distributional characteristics, with comparable concentrations of red (disagreement) samples in the high uncertainty regions. This consistency reinforces the conclusion that emotion preservation challenges are primarily attributable to the speech translation process itself rather than specific target language characteristics. The similar uncertainty-agreement patterns across both language pairs corroborate the comparable EPR scores (EN-JA: 0.216, EN-DE: 0.255) and Cohen's Kappa values (EN-JA: 0.086, EN-DE: 0.074) observed in the quantitative analysis.

The predominance of red samples (disagreement) in the uncertainty-agreement space directly validates the low emotion preservation rates observed in the discrete metrics. The scarcity of samples in the desirable bottom-left quadrant aligns with the poor Cohen's Kappa scores (0.086 and 0.074), which fall within the ”slight agreement” range. Similarly, the high concentration of disagreement samples corresponds to the low EPR values (21.6% and 25.5%), indicating that the majority of utterances fail to maintain consistent emotion labels across translation.

The positive correlation between uncertainty and disagreement suggests that emotion preservation failures are not merely random misclassifications but systematic breakdowns in the translation process. High uncertainty regions predominantly contain disagreement samples, indicating that the model recognizes its own limitations in maintaining emotion consistency. This pattern suggests that uncertainty measures could serve as valuable indicators for identifying utterances where emotion preservation is likely to fail, as shown in a previous study [21].

The uncertainty-agreement analysis supports the proposed metric hierarchy by demonstrating that categorical emotion metrics (EPR, Kappa) capture meaningful emotion preservation phenomena that manifest at the utterance level. The visual confirmation of poor emotion preservation across both language pairs validates the primary status assigned to BAR and the supporting role of dimensional metrics like va-sim, while questioning the reliability of high-scoring metrics like emo-sim that appear inconsistent with BAR scores.

5. CONCLUSION

This study evaluated emotion preservation in expressive speech-to-speech translation using multiple complementary metrics, including discrete emotion recognition, balanced accuracy ratio (BAR), emotion preservation rate (EPR), and Cohen's Kappa. The results reveal low scores of emotion preservation rates highlighting the difficulty of preserving emotion from source to target languages. Metrics such as BAR, valence-arousal similarity, and pause rate show similar scores which can be good indicators for quantifying emotion preservation, while EPR can serve as additional quality check. These metrics showed that about half of emotion are preserved in the translation process in the evaluated dataset. The analysis highlights the challenges of maintaining emotion consistency during translation, particularly for lower-arousal emotions.

6. REFERENCES

[1] Ye Jia et al., "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation," *Proc. Mach. Learn. Res.*, vol. 162, pp. 10120–10134, 2022.

[2] Hirofumi Inaguma et al., "UnitY: Two-pass Direct Speech-to-speech Translation with Discrete Units," *Proc. Annu. Meet. ACL*, vol. 1, pp. 15655–15680, 2023.

[3] Adam Polyak et al., "Speech resynthesis from discrete disentangled self-supervised representations," in *Interspeech 2021*, 2021, pp. 3531–3535.

[4] Ann Lee et al., "Textless Speech-to-Speech Translation on Real Data," in *NAACL 2022*, 2022, pp. 860–872.

[5] Seamless Communication et al., "Seamless: Multilingual Expressive and Streaming Speech Translation," 2023, arXiv:2312.05187.

[6] Sirou Chen et al., "MELD-ST: An Emotion-aware Speech Translation Dataset," in *Find. Assoc. Comput. Linguist. ACL 2024*, 2024, pp. 10118–10126.

[7] Soujanya Poria et al., "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," in *Proc. 57th Annu. Meet. ACL*, 2019, pp. 527–536.

[8] Felix Burkhardt et al., "Nkululeko: Machine Learning Experiments on Speaker Characteristics Without Programming," in *Interspeech 2023*, 2023, pp. 2010–2011.

[9] Bagus Tris Atmaja and Akira Sasou, "Multi-label Emotion Share Regression From Speech Using Pre-Trained Self-Supervised Learning Models," in *2024 IEEE Reg. 10 Conf.*, 2024.

[10] Guillaume Lemaitre et al., "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2016.

[11] Jacob Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[12] Ziyang Ma et al., "emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation," in *Findings ACL 2024*, 2024, pp. 15747–15760.

[13] Johannes Wagner et al., "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, 2023.

[14] Haibin Wu et al., "Laugh Now Cry Later: Controlling Time-Varying Emotional States of Flow-Matching-Based Zero-Shot Text-to-Speech," in *SLT 2024*, 2024.

[15] Arun Babu et al., "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[16] Sanyuan Chen et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2021.

[17] Kimi Team, "Kimi-Audio Technical Report," Tech. Rep., 2025.

[18] Yunfei Chu et al., "Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models," 2023, arXiv:2311.07919.

[19] Ruichen Li et al., "Speech Emotion Recognition via Multi-Level Cross-Modal Distillation," in *Interspeech 2021*, 2021, pp. 4488–4492.

[20] Zheng Lian et al., "CTNet: Conversational Transformer Network for Emotion Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 985–1000, 2021.

[21] Bagus Tris Atmaja et al., "Uncertainty-Based Ensemble Learning for Speech Classification," in *2024 27th Conf. Orient. COCOSDA*, 2024, pp. 1–6.