# EMO-Codec: An In-Depth Look at Emotion Preservation Capacity of Legacy and Neural Codec Models with Subjective and Objective Evaluations

Wenze Ren*, Yi-Cheng Lin*, Huang-Cheng Chou†, Haibin Wu*, Yi-Chiao Wu
Chi-Chun Lee†, Hung-yi Lee*, Hsin-Min Wang‡, Yu Tsao‡
* National Taiwan University, Taiwan † National Tsinghua University, Taiwan ‡ Academia Sinica, Taiwan
E-mail: {r11942166, r12942075}@ntu.edu.tw

***Abstract*—Neural codecs reduce speech data transmission latency and serve as the underlying tokenizer for speech language models (speech LMs). Preserving emotional information in these codes is essential for effective communication and contextual understanding. However, there is a lack of research on emotional loss in existing codecs. This study evaluates both neural and legacy codecs using subjective and objective methods on emotion datasets such as IEMOCAP. Our study identifies which codecs best preserve emotional information at various bitrates. We found that training a codec with both English and Chinese data had limited success in retaining emotional information in Chinese. Additionally, resynthesizing speech using these codecs degrades speech emotion recognition (SER) performance, especially for emotions such as sadness, depression, fear, and disgust. Human listening tests confirmed these findings. This study will guide the development of future speech technologies to ensure that new codecs maintain the emotional integrity of speech.**

## I. Introduction

Audio codecs are initially introduced to compress speech signals into fewer bits to achieve low-latency communication. They generally include an encoder and a decoder, which respectively compress a speech signal into codes and reconstruct the codes back into the speech signal. Given their success, researchers are exploring how speech perception tasks can benefit from the advanced capabilities of large language models (LLMs), which have demonstrated remarkable performance in text modeling and surpassed human capabilities in a variety of tasks [1]. Recent advances in speech language models (speech LMs) use the codes of speech codecs as discrete tokens [2], [3]. Speech LMs are distinct in their ability to extract rich information, such as emotions, from spoken languages [4]. In addition to capturing content information, they delve into the nuances of the speaker's identity and emotion that cannot be fully captured by text alone.

Since codec models are widely used to reduce communication latency and serve as tokenizers for speech LMs, their codes should preserve signal integrity, including emotional information. For instance, when an individual communicates with a virtual assistant through voice commands, the emotional information embedded in the speech signal can provide valuable context for the assistant to deliver a more empathetic and tailored response [5]. However, speech codecs used in communication pipelines inadvertently distort or discard critical emotional cues during the compression process. As a result, the virtual assistant may struggle to accurately perceive the user's emotional state, thereby affecting the interaction. Therefore, preserving the emotional information in the speech signal is crucial for the effectiveness of speech LMs.

Many advanced neural codec models have been developed using different techniques [6]. However, existing evaluations of these codecs have mainly focused on signal-level metrics, overlooking critical paralinguistic elements such as emotion [7]. While Codec-SUPERB [8] endeavors to compare emotion preservation in speech reconstructed by neural codecs, its evaluation was limited to a single dataset in a single language. Similarly, Siegert et al. [9] studied the intelligibility of codec-compressed emotional speech but neglected to evaluate the emotional content of the speech. Previous work compared speech distorted by legacy codec compression algorithms and evaluated speech emotion recognition (SER) performance through human perception [10] and automatic methods such as Gaussian mixture model [11] or Support Vector Machine (SVM) [12]. However, these studies are limited to legacy codecs, which have poor performance compared to neural codecs. The SER models used are also less accurate than today's advanced models.

There is an urgent need for a comprehensive comparative analysis of the ability of codec models of different languages on different downstream SER systems under staged or real-world, multilingual, and multi-speaker dataset conditions. We consider various factors of the codec, such as bitrate, pretraining dataset language, and architecture. Different evaluations are conducted to comprehensively assess these factors and their impact on the accuracy of pre-trained SER systems and human emotion perception. Our goal is to provide the research community with valuable insights to guide the design of new codecs. Our study comprehensively evaluates the efficacy of 14 neural and 3 legacy codecs in preserving emotional information across 15 different SER models on 6 datasets, revealing their potential for enhancing affective computing in real-world applications. The main flow of Emo-Codec is shown in Fig. 1.

Our main contributions are as follows:

- Emo-Codec provides comprehensive performance benchmarks for 14 codec models and 3 legacy codecs in 6 emotion datasets, highlighting their ability to preserve emotional information.
- Descript Audio Codec (DAC) series [13] consistently outperforms other codecs in SER at the same bitrate. Additionally, AcademiCodec [14] and SpeechTokenizer [15] show considerable performance in low-bitrate scenarios.
- Training a codec using Chinese and English data resulted in limited improvements in preserving Chinese emotional information compared to training a codec using only English data.
- Of all the emotions investigated, some negatively valenced emotions, such as sadness, depression, fear, and disgust, showed higher performance declines than others.

## II. Methodology

This section first provides an overview of our rationale for conducting a large-scale evaluation and then explains each part in detail.

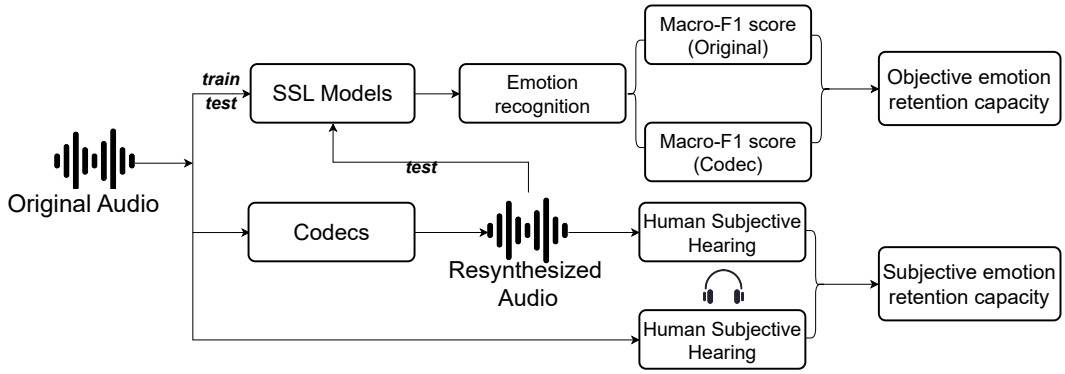



Fig. 1: The pipeline of Emo-Codec. We only use the original training audio from emotion datasets to train the emotion recognition model and fine-tune the SSL model at the same time. Then, we evaluate the testing set's original audio and audio resynthesized using various codecs on SER models. We calculate the $F_1$ score difference of SER to obtain the objective emotion retention capacity. We also conduct human subjective listening tests using original audio and resynthesized audio to obtain the subjective emotion retention capacity.

### *A. Rationale for the evaluation*

We evaluate both neural and legacy codecs to ensure a comprehensive analysis. The legacy codecs serve as established benchmarks, enabling us to measure the progress achieved with neural codecs. For the neural codecs, we select models based on their functionality and design principles. These codecs incorporate a range of innovative architectural designs and sophisticated methods to ensure optimal performance and versatility in processing various types of speech data. Specifically, we target models explicitly tailored for speech LM tokenization. Additionally, we consider models trained on mixed Chinese and English data (for example Funcodec) to ensure comprehensive language coverage and take into account the nuances and emotional variances inherent in both languages. We evaluate all codec models at similar bitrates to ensure a fair comparison.

For the SER models, we use different representations of self-supervised learning (SSL) models (shown in Table I) to train the SER model because the SSL paradigm has achieved state-of-the-art performance in SER tasks [16], [17]. We employ a variety of emotion datasets to increase diversity in languages, dataset collection methods (real world, improvised act, scripted act), and speakers.

### *B. Codec models*

We carefully selected seven cutting-edge, high-fidelity neural codec models for comparative analysis, as shown in Table II. **Encodec** [18] serves as the baseline. We used the 2, 4, 8, 16, and 32 layer settings to compare with other codecs at similar bitrates. Building on Encodec, **AudioDec** [19] introduces a novel approach that employs group convolution to accelerate and streamline operations. **AcademiCodec** [14] uses group-residual vector quantization to reduce codebook usage while maintaining comparable performance. We used a universal version of this model. **FunCodec** [20] proposes a frequency domain codec that achieves comparable performance with lower computational and parameter complexity. Additionally,

TABLE I: SSL models used in this study and their number of parameters (Parm.

| Model | Parm. (M) | Model | Parm. (M) |
|---|---|---|---|
| Wav2Vec 2.0 XLS-R-1B | 965 | VQ-Wav2Vec | 34 |
| WavLM large | 317 | Wav2vec-large-960h | 33 |
| Wav2Vec 2.0 Large Robust | 317 | TERA | 21 |
| HuBERT Large | 317 | NPC | 19 |
| Wav2Vec2 large 960h | 317 | VQ-APC | 5 |
| Data2vec large 960h | 313 | APC | 4 |
| DeCoAR 2.0 | 90 | Modified CPC | 2 |
| Mockingjay | 85 | | |

TABLE II: Codecs evaluated in this study and their transmission bitrates. The column **ID** shows the identification number of all codecs, **kbps** represents the bitrate in kilobits Per Second, and **sr** shows the sampling rate in kHz.

| ID | Codec | Codec Configuration | kbps | sr |
|---|---|---|---|---|
| U | AudioDec | symAD_libritts_24000_hop300 | 6.4 | 24 |
| C | AcademiCodec | large universal | 2 | 16 |
| S | SpeechTokenizer | hubert_avg | 4 | 16 |
| D1 | DAC | DAC_16k | 6 | 16 |
| D2 | | DAC_24k | 24 | 24 |
| E1 | Encodec | Encodec_24k | 1.5 | 24 |
| E2 | | | 3 | 24 |
| E3 | | | 6 | 24 |
| E4 | | | 12 | 24 |
| E5 | | | 24 | 24 |
| F1 | Funcodec | en_libritts_16k_nq32ds320 | 16 | 16 |
| F2 | | en_libritts_16k_nq32ds640 | 8 | 16 |
| F3 | | zh_en_16k_nq32ds320 | 16 | 16 |
| F4 | | zh_en_16k_nq32ds640 | 8 | 16 |
| N | Soundstream | Soundstream | 6 | 16 |
| M1 | MP3 | - | 6 | - |
| M2 | | - | 24 | - |
| M3 | | - | 192 | - |
| O1 | Opus | - | 6 | - |
| O2 | | - | 24 | - |
| A1 | AAC | - | 6 | - |
| A2 | | - | 24 | - |
| A3 | | - | 192 | - |

**SpeechTokenizer** [15] introduces a unified speech tokenizer tailored for speech LMs, integrating HuBERT units as semantic teachers in the first layer of RVQ. **Descript Audio Codec (DAC)** [13] leverages advanced Snake activation from BigVGAN [21] and utilizes a novel complex STFT discriminator at multiple time scales to further enhance audio fidelity. We used the 16k and 24k sample rate models. Lastly, **SoundStream**[22] uses RVQ in its encoder to represent audio signals more efficiently and compactly, resulting in higher quality reconstruction at lower bitrates[1].

To more comprehensively compare the ability of codecs to retain emotional traits, we also selected three legacy codecs for comparison: **MP3** [23], which is widely used for its efficient compression and good sound quality at different bitrates; **Opus** [24], which is known for its adaptability to a wide range of audio and low latency; and **AAC** [25], which provides high fidelity and is commonly used for streaming and

[1]We used the implementation from: https://github.com/kaiidams/soundstream-pytorch

TABLE III: Datasets used in this study and their setting. **Anno** represents the annotation process used; **P** and **S** represent primary and secondary labeling scenarios, respectively. **Speaker** represents the number of speakers in the dataset.

| Dataset | Language | Setting | Anno | Speaker |
|---|---|---|---|---|
| CREAM-D | English | Acted | P | 91 |
| IEMOCAP | English | Acted | S | 10 |
| IMPROV | English | Acted | S | 12 |
| MSP-PODCAST | English | Real-world | P | 2172+Unknown |
| BIIC-PODCAST | Chinese | Real-world | P | Unknown |
| NNIME | Chinese | Acted | S | 43 |

broadcasting. These legacy codecs help us benchmark neural codecs against established standards.

### C. Speech Emotion Recognition Datasets

To evaluate the codecs, we used six public datasets partitioned by EMO-SUPERB [26]. These datasets are classified according to their source (acted or real-world) and language (Chinese or English), as shown in Table III. The datasets have two annotation scenarios: *Primary* (P) requires each annotator to select only one emotion during annotation, while *Secondary* (S) allows annotators to select multiple emotions for a clip.

## III. Experimental Setup

### A. Speech Emotion Recognition Models

In contrast to previous SER methods that perform single-target prediction , we adopt a distribution-like representation to model the multi-dimensional complexity of emotions, as suggested by [27]. Furthermore, to enhance the performance of the SER model, we incorporate label smoothing [28] into the emotion distribution, effectively regularizing the classifier layer by setting the smoothing parameter to 0.05.

We adopted the SER model architecture from the S3PRL [29] toolkit, which is based on a CNN-Self Attention network consisting of three Conv1d layers, a self-attention pooling layer, and two linear layers. We used a fixed learning rate of $10^{-4}$ and the AdamW optimizer [30] to train the SER model until the loss on the development set stopped decreasing for 5 epochs. We used class-balanced cross-entropy loss [31] to mitigate the impact of imbalanced labels in emotion.

### B. Evaluation Metrics

We used the macro-$F_1$ score [32] as the evaluation metric for the SER task. Since the output of the SER model is a probability distribution, the emotion prediction is successful if the corresponding probability of the ground-truth label exceeds the threshold $\frac{1}{n}$ for the $n$-class SER model, following [27], [33].

### C. Human Subjective Evaluation

To determine whether humans perceive emotions differently in synthesized versus original speech, we conducted human subjective listening tests. We randomly selected 45 audio samples from the well-known emotion dataset IEMOCAP. We used three codecs, namely Encodec, DAC, and the legacy codec Opus, to resynthesize the audio at bitrates of 6kbps and 24kbps.

We hired evaluators from the Prolific platform to evaluate the resynthesized audio. Each audio sample was evaluated by 5 male and 5 female evaluators. We required evaluators to be from the US and must have a past assignment acceptance rate above 90%. Each evaluator was asked to answer three questions for each audio sample: (1) select one or more emotions of the speaker from the pre-defined emotions, the same process as in IEMOCAP; (2) rate speech quality based on everyday speech communication on a scale of 1 to 5; (3) rate speech quality according to the degree of distortion on a scale of 1 to 5.

Annotators of the IEMOCAP dataset originally used both video and audio to label multiple emotions. However, our study only focused on audio. This difference may affect emotion perception. Therefore, we also asked these evaluators to re-annotate the emotions of the original audio to make a fair comparison.

After collecting annotations and ratings, we report average values for speech quality and distortion factors. In terms of emotion, we calculate the distributional label for each audio sample, and binarize the label using a threshold in the same way as introduced in Sec. III-B to measure the macro-$F_1$ score. We take the new-collected labels elicited by audios as the ground truth and calculate the accuracy, and define this as *subjective emotion capacity*.

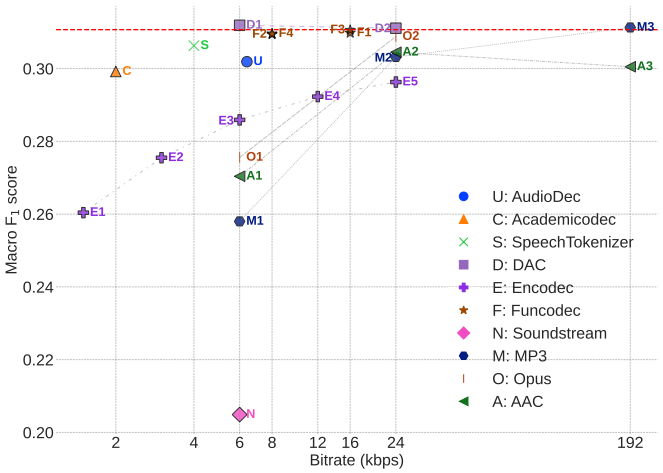



Fig. 2: SER performance (macro-$F_1$) for different codecs on the IEMOCAP dataset. The red dashed line represents the SER performance of the original audio (the topline).

## IV. Results and Discussion

### A. Impact of Bitrate on Machine Emotion Recognition

Fig. 2 shows the emotion preservation capabilities of individual codecs at different bitrates on the IEMOCAP dataset. Across most codecs, whether neural or legacy, there is a consistent trend: emotion preservation increases as the codec bitrate increases. It highlights the direct impact of bitrate on the amount of voice information that can be transmitted and retained. A higher bitrate is beneficial to retaining more detailed emotional information. While lower bitrates often compromise the retention of emotional information, some codecs exhibit excellent emotion retention even at extremely low bitrates, such as **SpeechTokenizer** and **Academicodec**. This indicates that these codecs are particularly effective at preserving the integrity of emotional information despite the limitations of lower bitrates. Of all codecs, Soundstream is the worst at retaining emotional information, showing the worst emotion recognition performance.

Legacy codecs (**MP3**, **Opus**, and **AAC**) generally demonstrate effective retention of emotional information at higher bitrates, e.g., **Opus**_24kbps and **MP3**_192kbps perform well, approaching the topline performance. However, all three legacy codecs perform poorly at low bitrates. Of the three legacy codecs, **AAC** is the worst. In comparison, neural codec **DAC** consistently outperforms all neural and legacy codecs at the same bitrate of 24kbps, highlighting DAC's advanced design and effectiveness in retaining emotional information. The superior performance of DAC may be attributed to two tricks: the snake activation function and balanced data sampling during training. DAC also performs well at 6kbps. In addition, **Funcodec**'s

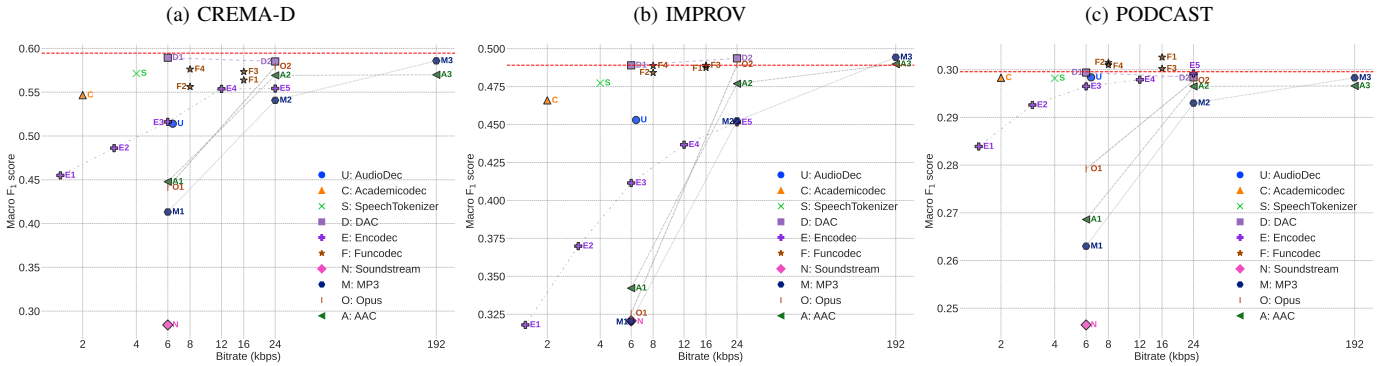



Fig. 3: SER performance (macro-$F_1$) for different codecs on the (a) CREMA-D, (b) IMPROV, and (c) PODCAST datasets. The red dashed line represents the SER performance of the original audio (the topline).

performance under different bitrates is also good and stable. Our results demonstrate that neural codecs are more suitable for retaining the integrity of emotional information in bitrate-constrained environments.

### *B. Trend Across English Datasets*

Next, we evaluate the emotional information retention of codecs on three additional English datasets (CREMA-D, IMPROV, and PODCAST). Obviously, the same trends as in Fig. 2 can also be observed in Fig. 3. This analysis extends the analysis on the IEMOCAP dataset in Sec. IV-A, highlighting consistent trends in the emotional information retention capabilities of codecs.

### *C. Variability Across Chinese Datasets*

Fig. 4 and Fig. A1 in the Appendix show the emotion preservation capabilities of various codecs at different bitrates on the Chinese datasets BIIC-PODCAST and NNIME. Consistent with the trend in the English datasets, higher bitrate codecs generally retain emotional information better in the Chinese datasets; the **DAC** still provides superior performance compared to other codecs of the same bitrate; neural codecs **SpeechTokenizer** and **Academicodec** also maintain good performance at lower bitrates; and legacy codecs are always weaker than neural codecs.

While the overall trend is consistent, there are some differences when comparing the performance of certain codecs on the English and Chinese datasets in detail. In the case of **Encodec**, a significant increase in emotional information retention with increasing bitrate can be observed in the English datasets. However, this trend slows down in the Chinese dataset NNIME and even decreases in the BIIC-PODCAST dataset. Compared with English, emotional information in Chinese has different phonetic and tonal characteristics, and the neural codec may not have been specifically optimized for these differences, causing the retention of emotional information to slowly increase or even decrease as the bitrate increases.

Furthermore, specific codecs are trained on English-only data, such as **funcodec_en_libritts** (F1 and F2), while conversely **funcodec_zh_en** (F3 and F4) is trained on mixed English and Chinese data. From Fig. A1, we can see that on the Chinese dataset BIIC-PODCAST, the codecs F3 and F4 trained on mixed English and Chinese data are better at preserving Chinese emotional information than their counterpart codecs F1 and F2 (F4 vs F2, F3 vs F1) trained on English-only data. However, F3 did not beat F1 on the Chinese dataset NNIME. This observation shows the importance of research on codec training data settings in order to accommodate universal representations of different emotional information across languages.

### *D. Specific Emotion Losses*

As can be seen from Table IV, speech resynthesized through a codec may lose some important emotional information, especially for emotions that are challenging for SER models, such as disgust or

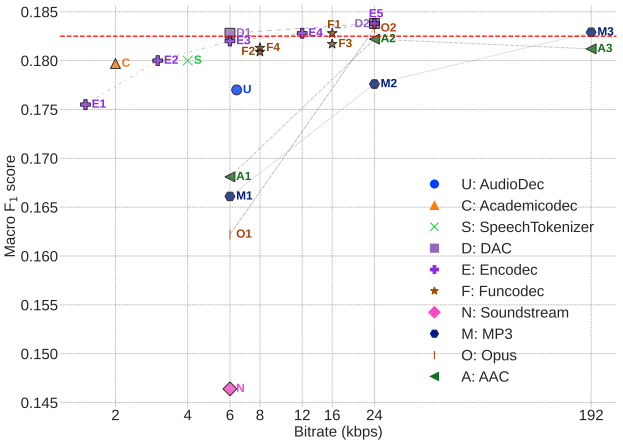



Fig. 4: SER performance (macro-$F_1$) for different codecs on the NNIME Chinese datasets. The red dashed line represents the SER performance of the original audio (the topline).

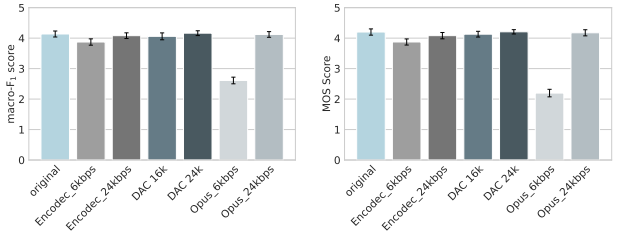
fear. For instance, the $F_1$ scores for the fear emotion category on the IMPROV and IEMOCAP datasets are only 1.86% and 9.58%, respectively. The $F_1$ scores for fear in the resynthesized speech dropped by 57.22% and 18.45%, respectively, Significantly higher than most other emotions. This shows that when the original audio has low emotion recognition performance, any additional distortion from the codec will have a pronounced impact. For the SER model, fear is inherently a difficult emotion to identify. The substantial performance drop highlights the inability of the codec to retain the emotional information of fear, which poses a greater challenge to the SER model. In addition to fear, other emotions such as depression and sadness also experience SER performance degradation when speech is resynthesized through these codecs. This highlights the codecs' shortcomings in retaining complex and subtle emotional information. While codecs are proficient in compressing and resynthesizing speech, they struggle to preserve the nuanced emotional cues needed for accurate emotion recognition, especially for more complex emotions. This demonstrates the need for further optimization and enhancement of codec models to better preserve the emotional integrity of resynthesized speech.

### *E. Human Subjective Listening Tests*

This section provides the subjective evaluation results. The original audio samples are randomly selected from the IEMOCAP dataset. Fig. 5 shows the Mean Opinion Score (MOS) of the original audio and the codec-resynthesized audio. The original audio has the highest MOS. The 24k bitrate codecs (Encodec_24kbps and DAC 24k) result in a slight drop in MOS, but still maintain a score close to that of

TABLE IV: SER performance (macro-$F_1$) degradation of resynthesized speech relative to the original speech. Positive values mean improvement in SER performance. Emotions considered include depression (P), frustration (T), anger (A) sadness (S), disgust (D), excitement (E), fear (F), neutral (N), surprise (U), and happiness (H). The **Ori** row represents the performance of the original audio. The bold text indicates the emotion with the highest, the second-highest, and the third-highest level of average degradation.

| | IMPROVS | | | | | | | | | | IEMOCAP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | P | T | A | S | D | E | F | N | U | H | T | A | S | D | E | F | N | S | H |
| Ori. | 38.16 | 53.34 | 47.91 | 57.33 | 14.35 | 57.07 | 1.86 | 84.04 | 29.12 | 69.21 | 69.27 | 59.70 | 61.36 | 00.89 | 55.10 | 9.58 | 67.54 | 25.19 | 49.28 |
| U | -50.49 | -11.88 | -8.68 | -19.15 | -35.13 | 2.05 | -70.65 | -0.51 | -6.90 | -3.37 | -1.93 | -3.06 | -5.62 | -30.80 | -4.20 | -7.76 | -2.67 | -12.10 | -6.99 |
| C | -42.27 | -8.22 | -5.96 | -13.03 | -25.01 | 1.37 | -73.60 | -0.48 | 4.85 | -2.45 | -1.81 | -4.79 | -8.90 | -25.63 | -5.03 | -20.83 | -6.77 | -8.13 | -7.44 |
| S | -23.12 | -5.39 | -5.61 | -5.93 | -24.68 | 0.36 | -48.77 | -0.31 | -3.41 | -1.88 | -1.37 | -1.67 | -4.07 | -3.81 | -2.62 | -11.04 | -1.78 | -7.79 | -4.36 |
| D1 | -7.58 | -1.75 | -1.37 | -2.06 | -3.44 | 0.64 | -23.91 | -0.16 | -1.76 | -0.47 | -0.31 | -0.19 | -0.72 | -8.84 | -0.47 | -2.97 | -0.40 | 0.22 | -0.72 |
| D2 | -1.84 | -0.80 | -1.21 | -0.37 | -9.82 | 0.65 | -21.38 | -0.08 | 1.28 | -0.14 | -0.02 | -0.40 | -0.40 | -27.97 | -0.31 | -1.81 | -0.13 | 0.16 | -0.30 |
| E1 | -88.21 | -42.40 | -22.08 | -68.06 | -70.39 | -7.00 | -94.65 | -16.25 | -28.31 | -6.26 | -4.91 | -10.47 | -37.24 | -19.09 | -12.82 | -41.13 | -28.06 | -34.14 | -18.71 |
| E5 | -43.75 | -6.24 | -2.94 | -19.94 | -39.15 | 1.84 | -57.34 | -1.22 | -3.94 | -3.39 | -1.43 | -3.90 | -12.28 | -10.62 | -5.65 | -18.05 | -6.68 | -10.58 | -5.62 |
| F1 | -12.57 | -0.54 | -1.63 | -3.10 | -10.31 | 1.15 | -26.03 | -0.07 | -0.07 | -0.82 | -0.72 | -0.47 | -1.12 | -5.97 | -1.73 | -4.23 | -0.55 | -5.07 | -1.48 |
| F3 | -11.86 | -1.29 | -2.03 | -3.25 | -11.97 | 1.08 | -30.95 | -0.10 | 3.63 | -0.61 | -1.02 | -0.98 | -1.34 | -24.16 | -1.78 | -9.88 | -1.08 | -1.00 | -1.70 |
| M1 | -76.03 | -62.86 | -48.19 | -57.12 | -88.15 | -14.26 | -94.21 | -15.82 | -4.42 | -8.45 | -16.58 | -20.97 | -23.86 | 1.29 | -17.02 | -54.93 | -14.26 | -33.99 | -28.78 |
| M3 | -8.07 | -5.33 | -7.00 | -2.82 | -1.36 | -11.41 | -40.98 | 0.49 | -9.42 | -2.08 | -0.21 | -0.10 | -3.99 | -28.14 | -1.54 | 0.28 | -0.63 | -5.75 | -5.41 |
| O1 | -87.75 | -44.84 | -27.26 | -68.92 | -81.77 | -12.65 | -92.32 | -15.93 | -35.08 | -7.74 | -10.67 | -14.65 | -19.91 | -19.66 | -14.21 | -37.22 | -7.63 | -33.98 | -21.71 |
| O2 | -23.29 | -5.32 | -6.04 | -6.27 | -5.44 | -6.98 | -49.64 | 0.51 | -11.87 | -2.24 | -0.28 | -1.18 | -4.34 | -18.19 | -1.82 | -6.20 | -1.46 | -7.46 | -7.00 |
| A1 | -57.40 | -61.04 | -52.67 | -38.45 | -94.45 | -11.11 | -95.81 | -14.95 | -15.95 | -6.56 | -10.71 | -15.80 | -19.79 | -12.93 | -17.42 | -49.14 | -11.81 | -30.28 | -25.50 |
| A3 | -15.53 | -6.24 | -8.17 | -5.52 | -8.68 | -10.30 | -38.00 | 0.21 | -19.70 | -1.83 | -1.29 | -3.92 | -10.34 | -46.57 | -3.02 | -11.84 | -2.71 | -17.44 | -9.60 |
| Avg | **-36.65** | -17.61 | -13.39 | -20.93 | **-33.98** | -4.30 | **-57.22** | -4.31 | -8.74 | -3.22 | -3.55 | -5.50 | -10.26 | **-18.74** | -5.98 | **-18.45** | -5.77 | **-13.82** | -9.69 |




(a) Communication (b) Distortion

Fig. 5: Subjective quality assessment of speech (MOS) based on (a) everyday speech communication and (b) distortion, following ITUT P835 [34]. Higher scores indicate higher quality.

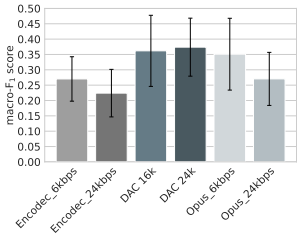



Fig. 6: Subjective emotion retention capacity (macro-$F_1$) for six codecs.

the original audio. In contrast, the MOS for the 6k bitrate codecs drops significantly, except for DAC 16k (as shown in Table II, its bitrate is 6kbps), which maintains a high MOS. Opus_6kbps has the lowest MOS in this evaluation. These results align with the objective evaluation results in the previous sections.

Fig. 6 shows the human SER performance on the codec-resynthesized audio. Again, we can see that DAC 24k and DAC 16k perform well, and **DAC 24k** achieves the best performance. Encodec_24kbps is worse than its lower bitrate counterpart Encodec_6kbps, and Opus_24kbps is also worse than its counterpart Opus_6kbps. Surprisingly, Opus_6kbps performs well in this human SER evaluation. The reason still needs further study. The discrepancy between subjective human listening tests and the objective assessments presented in Fig. 2 is emphasized, and this inconsistency suggests that objective metrics may not fully capture elements of human perception of emotional content in language. Further research into the perceptual aspects of audio quality and emotion recognition is recommended to address this issue.

## V. Discussion and Limitation

Existing neural codecs are mainly trained on English and Chinese datasets. However, there are thousands of spoken languages around the world. Whether existing codecs can be generalized to other languages and how to train codecs that preserve multilingual emotion information are unsolved issues.

While subjective evaluations provide valuable insights, current evaluators are limited to a small number of people. Future research should aim to include a more diverse group of evaluators to account for different emotional perceptions across age groups, cultures, and backgrounds.

Due to limited scope, this work did not consider how the interactivity and context of a conversation might affect the effectiveness of emotion preservation. Future research should explore how codecs perform in interactive dialogue systems, where context and conversational history play crucial roles in emotion recognition and response generation. We will also implement more comprehensive subjective testing protocols and refine objective evaluation metrics to better align with the framework for assessing human perceptual differences.

## VI. Conclusion

This work provides insights into the emotional information preservation capabilities of neural codecs. We provide different perspectives to evaluate codec performance. We confirm that codecs with higher bitrate preserve more emotional information. DAC performs the best among all compared neural codecs, while AcademiCodec and SpeechTokenizer can preserve a considerable amount of emotional information at limited bitrates. Legacy codecs perform worse than neural codecs at low bitrates. Furthermore, a codec trained with Chinese and English data may have a limited improvement in Chinese emotion information preservation capacity compared to a codec trained with only English data. We find that the resynthesis of speech by neural codecs reduces emotional information such as sadness, depression, fear, and disgust. Our future work will train neural codec models to preserve emotional information across diverse language usage contexts.

## ACKNOWLEDGMENT

NSTC supported this research under Grants 112-2634-F-002-005. We thank the anonymous reviewers for their valuable comments.

## REFERENCES

[1] H. Touvron *et al.*, *Llama 2: Open Foundation and Fine-Tuned Chat Models*, 2023. arXiv: 2307.09288 [cs.CL].

[2] C. Wang *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[3] D. Yang *et al.*, "Uniaudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.

[4] C.-y. Huang *et al.*, *Dynamic-SUPERB: Towards A Dynamic, Collaborative, and Comprehensive Instruction-Tuning Benchmark for Speech*, 2023. arXiv: 2309.09510 [eess.AS].

[5] S. Guha and R. Iqbal, "DESCo: Detecting Emotions from Smart Commands," in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2022.

[6] H. Wu *et al.*, *Towards audio language modeling – an overview*, 2024. arXiv: 2402.13236.

[7] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "Warp-Q: Quality Prediction for Generative Neural Speech Codecs," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[8] H. Wu *et al.*, *Codec-SUPERB: An In-Depth Analysis of Sound Codec Models*, 2024. arXiv: 2402.13071 [eess.AS].

[9] I. Siegert *et al.*, "Emotion Intelligibility within Codec-Compressed and Reduced Bandwidth Speech," in *Speech Communication; 12. ITG Symposium*, 2016.

[10] O. Niebuhr and I. Siegert, ""High on Emotion "? How Audio Codecs Interfere With the Perceived Charisma and Emotional States of Men and Women," in *33. Konferenz Elektronische Sprachsignalverarbeitung, ESSV 2022*, 2022.

[11] A. Albahri and M. Lech, "Effects of band reduction and coding on speech emotion recognition," in *2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2016.

[12] A. F. Lotz *et al.*, "Audio Compression and its Impact on Emotion Recognition in Affective Computing," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2017*, 2017, pp. 1–8.

[13] R. Kumar *et al.*, "High-fidelity audio compression with improved RVQGAN," *Advances in Neural Information Processing Systems*, 2024.

[14] D. Yang *et al.*, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[15] X. Zhang *et al.*, "Speechtokenizer: Unified speech tokenizer for speech large language models," *arXiv preprint arXiv:2308.16692*, 2023.

[16] J. Wagner *et al.*, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2023.

[17] E. Morais *et al.*, "Speech Emotion Recognition Using Self-Supervised Features," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[18] A. Défossez *et al.*, "High Fidelity Neural Audio Compression," *arXiv preprint arXiv:2210.13438*, 2022.

[19] Y.-C. Wu *et al.*, "Audiodec: An Open-Source Streaming High-Fidelity Neural Audio Codec," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[20] Z. Du *et al.*, "FunCodec: A Fundamental, Reproducible and Integrable Open-source Toolkit for Neural Speech Codec," *arXiv preprint arXiv:2309.07405*, 2023.

[21] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[22] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, *Soundstream: An end-to-end neural audio codec*, 2021. arXiv: 2107.03312 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2107.03312.

[23] K. Brandenburg and G. Stoll, "Iso-mpeg-1 audio: A generic standard for coding of high-: Quality digital audio," 1994. [Online]. Available: https://api.semanticscholar.org/CorpusID:58871544.

[24] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, *High-quality, low-delay music coding in the opus codec*, 2016. arXiv: 1602.04845 [cs.MM]. [Online]. Available: https://arxiv.org/abs/1602.04845.

[25] M. Bosi, K. Brandenburg, S. R. Quackenbush, *et al.*, "Iso/iec mpeg-2 advanced audio coding," *Journal of The Audio Engineering Society*, vol. 45, pp. 789–814, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:60516290.

[26] H. Wu *et al.*, "EMO-SUPERB: An In-depth Look at Speech Emotion Recognition," *arXiv preprint arXiv:2402.13018*, 2024.

[27] H.-C. Chou *et al.*, "Minority Views Matter: Evaluating Speech Emotion Classifiers with Human Subjective Annotations by an All-Inclusive Aggregation Rule," *IEEE Transactions on Affective Computing*, pp. 1–15, 2024. DOI: 10.1109/TAFFC.2024.3411290.

[28] C. Szegedy *et al.*, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] S.-w. Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021.

[30] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[31] Y. Cui *et al.*, "Class-Balanced Loss Based on Effective Number of Samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

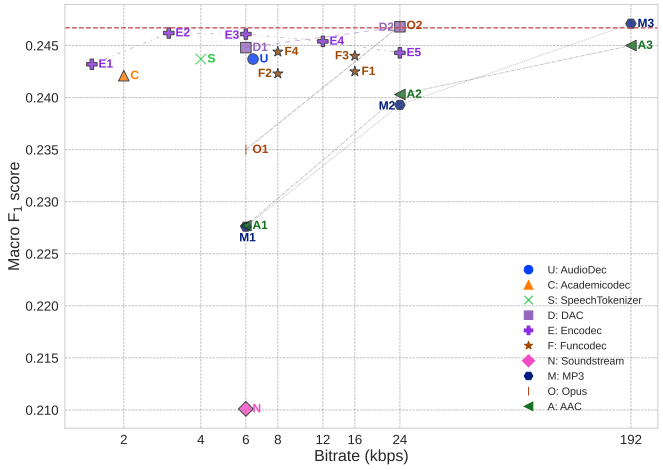
[32] J. Opitz and S. Burst, "Macro f1 and macro f1," *arXiv preprint arXiv:1911.03347*, 2019.

[33] P. Riera *et al.*, "No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems," in *Proc. SMM19, Workshop on Speech, Music and Mind 2019*, 2019.

[34] ITU, "Itu-t p.835, subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *International Telecommunication Union*, 2003.

# Appendix EMO-Codec: An In-Depth Look at Emotion Preservation Capacity of Legacy and Neural Codec Models With Subjective and Objective Evaluations

## I. Emotion recognition performance (macro-$F_1$) on different codec with Chinese datasets.




(a) BIIC-PODCAST

Fig. A1: SER performance (macro-$F_1$) for different codecs on the BIIC-PODCAST Chinese datasets. The red dashed line represents the SER performance of the original audio (the topline).

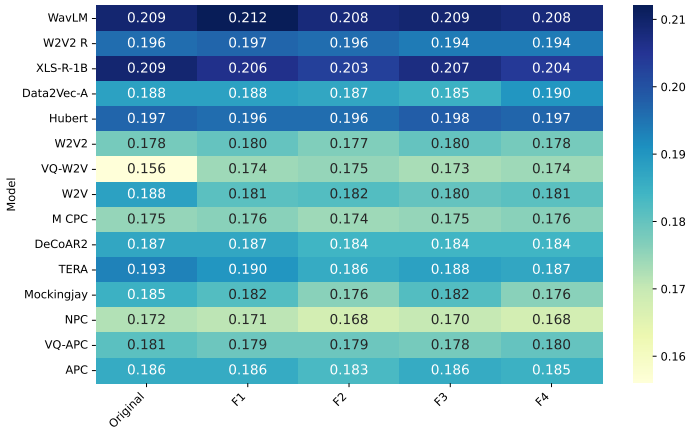



Fig. A2: SER performance (macro-$F_1$) on speech samples from the NNIME dataset resynthesized by various Funcodec models.