Empower Typed Descriptions by Large Language Models for Speech Emotion Recognition

Haibin Wu*, Huang-Cheng Chou†, Kai-Wei Chang*, Lucas Goncalves‡, Jiawei Du*,
Jyh-Shing Roger Jang*, Chi-Chun Lee†, and Hung-yi Lee*
* National Taiwan University, Taiwan
E-mail: f07921092@ntu.edu.tw, kaiwei.chang.tw@gmail.com, r11922185@ntu.edu.tw, jang@mirlab.org, hungyilee@ntu.edu.tw
† National Tsing Hua University, Taiwan
E-mail: huangchengchou@gmail.com, cclee@ee.nthu.edu.tw
‡ The University of Texas at Dallas, USA
E-mail: goncalves@utdallas.edu

***Abstract*—Training speech emotion recognition (SER) requires human-annotated labels and speech data. However, emotion perception is complex. The pre-defined emotion categories are not enough for annotators to describe their emotion perception. Devoted annotators will use natural language rather than traditional emotion labels when annotating data, resulting in typed descriptions (e.g., "Slightly Angry, calm" to notify the intensity of emotion). While these descriptions are highly valuable, SER models, designed as classification models, cannot process natural languages and thus discard them. To leverage the valuable typed descriptions, we propose a novel way to prompt ChatGPT to mimic annotators, comprehend natural language typed descriptions, and subsequently adjust the given label of the input data. By utilizing labels generated by ChatGPT, we consistently achieve an average relative gain of 3.08% across all settings using 15 speech self-surprised learning models on the SUPERB, which provides a potential way to integrate the power of LLMs to improve the performances of SER.**

I. Introduction

Speech Emotion Recognition (SER) aims to discern emotional cues from speech inputs, representing a pivotal technology for human-computer interaction (HCI) systems. Over the years, significant advancements have been made in SER, making it a promising field of research and application [1]. One of the critical factors in SER is the annotation of ground truth labels used to train SER models. Typically, emotional corpora are annotated through perceptual evaluations. In these evaluations, annotators provide their ratings by completing questionnaires with a fixed set of options after listening to or watching a stimulus (speech or video). However, this conventional approach has its limitations. The fixed number of options can lead to bias in the labels [2], as annotators can only select the best available option even when it doesn't precisely describe the emotional content of the stimulus. This constraint can result in labels that do not fully capture the nuances of genuine human emotions.

To overcome this limitation, modern emotional datasets [3]–[5] have adopted an alternative approach by allowing annotators to select "other" and provide their descriptions of the emotions in their own words (natural language), resulting in **typed descriptions** (some examples are illustrated in Figure 2). This method can attenuate the problems caused by forced-choice paradigms and give a richer, more accurate representation of the emotional content. Typed descriptions, such as "slightly hopeful," can offer valuable insights that go beyond the pre-defined categories. Despite their potential, these typed descriptors are often discarded by most prior studies due to the limitations of current SER systems, which are designed as classification models that cannot process natural language inputs. As the typed descriptions are highly valuable, the inclusion and processing of these typed descriptions could significantly enhance the performance and accuracy of SER systems. Unfortunately, existing SER models ignore this data due to their inability to handle natural language effectively.

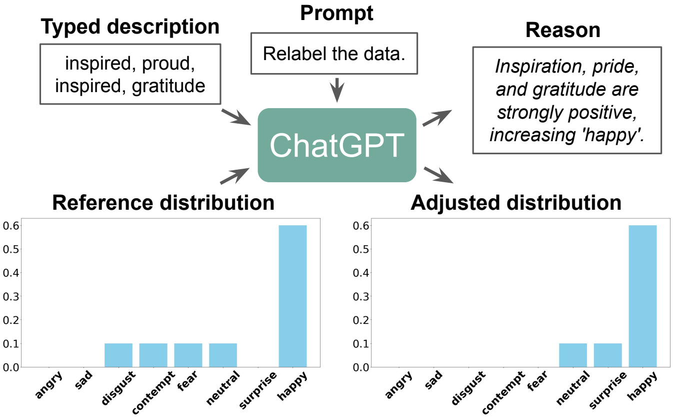

Fig. 1: Labeling process using ChatGPT. Three inputs are **Typed description**, **Reference distribution**, and **Prompt**. Two outputs are **Reason** and **Adjusted distribution**. Notice that the reference distribution is calculated by the number of votes for emotion classes. In the raw annotations of an example, there are instances of disgust, contempt, fear, neutrality, and happiness (*6), resulting in values of 0.6 for happiness and 0.1 for each of the remaining appearing emotions.

LLMs have demonstrated promising performance across a wide range of NLP tasks, including sentiment analysis [6]. Inspired by this success, we propose a novel approach to improve SER systems by utilizing LLMs to process and integrate typed descriptions. Specifically, we prompt ChatGPT to mimic annotators, comprehend the natural language typed descriptions, and adjust the given labels of the input data accordingly as illustrated in Figure 1. This approach harnesses the sophisticated understanding and text-generation capabilities of ChatGPT to enrich the emotional labels used in SER tasks.

In this study, we validate our proposed method on the MSP-PODCAST dataset, as it is the largest public English emotion

TABLE I: The Prompt for ChatGPT.

Objective:
As a knowledgeable assistant psychologist, your role is to analyze the given words and reference labels. You generate emotion label distributions. The emotions to consider are: 'angry,' 'sad,' 'disgust,' 'contempt,' 'fear,' 'neutral,' 'surprise,' and 'happy.' The order of emotions is very important. Please provide 8-dimensional emotion distributions for these 8 emotion classes based on the user input.
Input format:
The user input has two parts separated by #: The first part is the description. The second part is 8-dimensional reference emotion distribution, 'angry,' 'sad,' 'disgust,' 'contempt,' 'fear,' 'neutral,' 'surprise,' and 'happy.' The order of reference emotion is very important. The input has the format "descriptions#reference emotion distribution". Also give the reason for each data point why you want to change the reference emotion distribution. When given the answer, you should focus 25% on the "descriptions" and 75% on the "reference emotion distributions".
Example:
I will give you one example: User Input: Concerned,Interest#0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0. Generated Labels: {'angry': 0.1, 'sad': 0.2, 'disgust': 0.2, 'contempt': 0.3, 'fear': 0.0, 'neutral': 0.2, 'surprise': 0.0, 'happy': 0.0, "reason": ""}
Output format:
Reminder for the given data: It's very important to output the JSON format with an index.
Refine and Iterate :
I will give you 30 data points each time. Each data is separated by "—". It's very important. It's very important to make sure that you complete every response for 30 data points each time. Please reminder it. Output the JSON file that contains adjusted emotion label distributions based on reference distributions and detailed reasons why you adjust the reference emotion distributions for each word by each word. It's very important. It's very important that the JSON output file must contain the reference distributions and reasons. It's very important that do not contain the reference distributions and words. It's very important that use 15 to 20 words to explain the reason you want to change the reference distributions. It's very important that the sum of label distributions equals 1. It's very important to make sure that you explain the reasons for each word in descriptions.

dataset, which contains 34.37% of utterances with natural language descriptions. By leveraging labels generated by ChatGPT, we have consistently achieved an average relative improvement of 3.08% across almost all settings using 15 speech self-supervised learning models on the SUPERB benchmark. This improvement demonstrates the potential of integrating LLMs like ChatGPT to enhance the performance of SER systems. To the best of our knowledge, this study is the first to utilize ChatGPT to process typed descriptions to improve SER systems. Our findings suggest a new direction for future research and applications in the field of SER, showcasing the value of natural language descriptions and the power of LLMs in enhancing emotion recognition from speech.

II. Related Works

The integration of natural language descriptions into Speech Emotion Recognition (SER) has been a relatively unexplored area. One of the most closely related works is Chou et al. [7]. In their study, Chou et al. [7] combined polarity information derived from typed descriptions with the pre-defined emotions (e.g., happy or frustrated) provided by individual annotators. This was done within a label distribution framework to create a more comprehensive representation of the emotional content in spoken sentences. Finally, The authors then trained multi-task learning SER models using established three-label learning methods (soft-label, multi-label, and distribution-label) and demonstrated improved performance when typed descriptions were incorporated and evaluated on the MSP-Podcast corpus. However, the Chou et al. [7] approach has its limitations. The LIWC 2015 toolkit relies on predefined dictionaries to identify and interpret words. These dictionaries, although robust, are inherently limited in scope, often omitting many of the nuanced words found in typed emotional descriptions. For example, "Haaapy", indicating really happy, is not in the dictionaries. These limitation prevents the full utilization of the rich emotional content present in these natural language descriptions.

Different from the study by Chou et al. [7], our approach leverages the capabilities of large language models (LLMs) like ChatGPT to directly process and understand typed descriptions. By guiding the LLMs to consider both the natural language descriptions and the distributional labels calculated from predefined emotion categories, our method dynamically adjusts the labels. This approach not only captures a wider range of emotional nuances but also provides better generalization capabilities compared to the dictionary-based methods used in the LIWC toolkit.

III. Empower typed descriptions by ChatGPT

A. Resource

The MSP-PODCAST [4] collected spontaneous and diverse emotional speech from various real-world podcast recordings with a commercial license. In this work, we focus on primary emotion. Raters choose from nine categorized emotions in the primary emotion: angry, sad, happy, surprised, fear, disgusted, contempt, neutral, and "other." Each utterance has at least 5 raters, and the raters can type their own words to describe their emotion perception, named *typed description*. We exclude the other class and formulate the task as an 8-class emotion recognition task. We use the release version 1.11 of the database, including 84,030 utterances in the train set, 19,815 in the development set, 30,647 in the test1 set, and 14,815 in the test2 set. We combine the test1 and test2 as the test set. The database collects voices from more than 2,404 unique speakers.

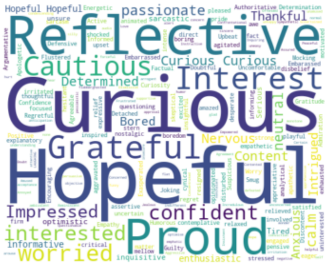

Fig. 2: The figures show typed descriptions of the MSP-PODCAST for an example.

B. Typed Descriptions

Fig. 2 displays a word cloud generated from typed descriptions collected in the MSP-PODCAST dataset. Word clouds are visual representations that highlight the frequency of words in a text corpus, with larger font sizes indicating higher frequencies. In this context, the word cloud provides insight into the types of words or phrases annotators use to describe their emotional responses in the dataset.

Incorporating typed descriptions can be valuable for understanding the nuances of human emotion. However, SER, primarily based on classification models, cannot process natural language and consequently overlooks these valuable typed descriptions.

C. Why Using ChatGPT for Relabeling

ChatGPT [8] exhibits a remarkable ability to comprehend and analyze natural language. [9] had used ChatGPT to do effective sentiment analysis. Hence, we utilize ChatGPT to mimic annotators, summarizing their thoughts to re-label the data with typed descriptions. While GPT models have been previously utilized for data labeling tasks, our approach stands out due to its innovative application in generating a distribution of labels instead of assigning a single label. We show that this approach leads to consistent improvements across all experimental settings, as shown in Table IV.

D. Prompt ChatGPT

We design a prompt to transform the released version of GPT-4 Turbo, a variant of ChatGPT, into a knowledgeable assistant **psychologist**. Its primary function is to generate a distribution across emotion labels based on the input-typed descriptions from annotators.

As shown in Figure 1, three inputs are provided to ChatGPT: the typed descriptions, reference distributions, and a well-designed prompt. When we prompt ChatGPT to refer to the distribution label, it fails to provide the distribution unless we supply the reference distribution. The format of the output emotion label is also a distribution. Guided by the prompt, the ChatGPT can adjust or maintain the reference distribution based on the typed descriptions. In the prompt, we also let ChatGPT explain why it changes or doesn't change the reference distributions. Without this, ChatGPT might default to laziness, consistently avoiding modifying the reference distributions. Table I in Appendix shows the well-designed prompt, which contains five parts: objective, input format, example, output format, and refine and iterate. In the objective part, we clearly describe the goal of the task. Then, we define the input format, including descriptions and reference emotions. Afterward, we provide an example that provides the template the ChatGPT can follow. Finally, we ask the ChatGPT to output the file in JSON format. Notice that the current version of the prompt is the 14th version. In the refine and iterate part, we show more rules that can enhance the accuracy of the output of the ChatGPT. We encourage the community to provide the designed prompt.

We choose the MSP-PODCAST (P) dataset to verify the efficacy of our proposed prompt method in utilizing typed descriptions to improve SER, as it is the largest dataset and has the highest percentage (6.08%) of typed descriptions among all other datasets. Figure 3 and 4 show the changes in label distributions between the original labels and the re-label one. The ChatGPT increased the number of fear and happiness and decreased the other emotions. In addition, Table III shows ten examples, including typed descriptions and reasons provided by ChatGPT.

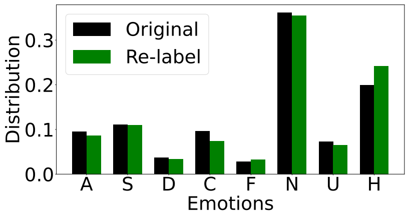

Fig. 3: The figures show comparisons of original and re-labeled label distribution. Emotion includes anger (A), sadness (S), happiness (H), surprise (U), fear (F), disgust (D), contempt (C), and neutral (N).

Model	Loss
Autoregressive Predictive Coding (APC) [10]	Generative loss
VQ-APC [11]	Generative loss
Non-autoregressive Predictive Coding (NPC) [12]	Generative loss
Mockingjay [13])	Generative loss
TERA [14]	Generative loss
DeCoAR 2 [15]	Generative loss
WavLM [16]	Discriminative loss
Hubert [17]	Discriminative loss
wav2vec 2.0 (**W2V2**) [18]	Discriminative loss
Data2Vec [19]	Discriminative loss
XLS-R [20]	Discriminative loss
VQ wav2vec (**VQ-W2V**) [21] [15]	Discriminative loss
wav2vec (**W2V**) [22]	Discriminative loss
Contrastive Predictive Coding (CPC) (**M CPC**)[23])	Discriminative loss

TABLE II: Summary of SSLMs

IV. Experimental setup

A. SSLM-based Codebase

1) Framework: Self-supervised learning (SSL) is a promising direction for developing speech models. This approach entails training a large model with large-scale unlabeled data to obtain robust and general representations. After pre-training, one can achieve nearly SOTA performance on downstream tasks by employing the fixed SSLMs alongside task-specific lightweight prediction heads [24]. Furthermore, SSLMs significantly enhance SER and demonstrate SOTA performance, as evidenced in [1].

We develop a comprehensive codebase highly depending on S3PRL [1] [24] to leverage 15 speech-supervised learning models as feature extractors and trains lightweight heads for exhaustive evaluation.

2) Self-supervised learning models: We leverage two mainstream categories of SOTA SSLMs (in S3PRL), pre-trained using generative losses and discriminative losses, summarized in Table II.

3) Label Representation: Inspired by **Semantics Space Theory** [25], we gather numerous annotations and compute a distribution-like (soft label) representation, to capture the high-dimensional nature of emotion perception more accurately. Notice that these distribution-like labels are the same as the **reference distribution** used for ChaptGPT as the reference label. Let's assume we gather five annotations from five raters for one sample. These annotations comprise neutral (N), anger (A), anger (A), sadness (S), and sadness (S). Subsequently, we compute the label distributions, which in this instance are represented as (N, A, S, H) = (0.2, 0.4, 0.4, 0.0) for training SER systems. Additionally, in order to enhance SER performance, we employ the label smoothing technique proposed by [26] to refine the vector, utilizing a smoothing parameter of 0.05. This approach assigns a small probability to emotional classes with zero values.

B. Evaluation Metric

We use the macro-F1 score [27] to evaluate the SER performance, considering recall and precision rates simultaneously. For the distribution-like multi-label training target, we select target classes by applying thresholds on the ground truth. A prediction is deemed successful if the proportion for a class surpasses $1/C$, where C represents the number of emotional classes, aligning with the settings employed in prior research [28]. For instance, consider a four-class

[1] https://github.com/s3prl/s3prl

TABLE III: Relabeled examples of typed descriptions with ChatGPT. ChatGPT also provides the reason for changing the reference distribution. **Change** denotes whether the reference distribution label is changed or not.

Change	Index	Typed Descriptions	Reason
Yes	01	calm,Slightly Angry,calm	Increased disgust to reflect slight anger.
No	01	Tranquil	Maintained high neutral for tranquil's peacefulness without strong emotions.

emotion recognition task, and the emotion classes contain neutral, anger, sadness, and happiness. Assume we consider the predictions for three different models: (0.2,0.35,0.35,0.1), (0.1,0.45,0.45,0.0), and (0.45,0.1,0.0,0.45). The three predictions are transformed into (0,1,1,0), (0,1,1,0), and (1,0,0,1), respectively, using the threshold. In these cases, only the first two predictions are fully corrected.

C. Training Details

We use the AdamW optimizer [29] with a 0.0001 learning rate, and the batch size is 32. We choose the best models according to the lowest value of the class-balanced cross-entropy loss on the development set. We use the Nvidia Tesla v100 GPUs with 32 GB memory for all results. The total of GPU hours is around 100 hours. According to [24], [30], [31], SSLMs usually result in consistent results and consume large computations. All results in the work are single-run. We also verify it by running experiments for small SSLMs, and the standard deviation is only less than 1% on average.

V. Results and Analysis

A. Main Results

We mainly use SSLMs as our backbone models to train SER systems in the work. Fig. 4 shows a randomly selected example to compare the original distribution and relabeled distribution by ChatGPT (data sample "PODCAST_1631_0043_0001.wav"). We observe that ChatGPT effectively comprehends the typed descriptions conveying positive emotions, thereby assigning greater weight to the "happy" emotion category. More examples can be found in Fig. F5a to Fig. F5d and Table III and Table F5 in Appendix.

Table IV presents the macro-F1 scores of the experiment along with the effects of incorporating data labeled by ChatGPT. We denote "W/O ChatGPT labels" and "W/ ChatGPT labels" to signify results without and with ChatGPT labels, respectively, while maintaining all other settings the same. We note the following observations: (1) The experiments involved 16 models, resulting in an average relative performance gain of 3.08%. (2) Particularly noteworthy is the case of CPC, which exhibits a substantial 9.45% relative improvement.

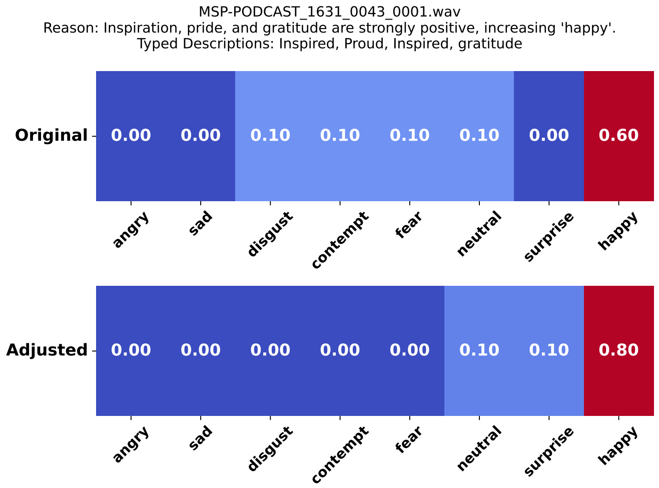

Fig. 4: Original and adjusted distributions. The original distribution, determined by tallying the votes for each emotion class, is compared with the adjusted distribution resulting from ChatGPT's re-labeling. Take the first one as an example; in the raw annotations of the example, there are instances of disgust, contempt, fear, neutrality, and happiness (*6), resulting in values of 0.6 for happiness and 0.1 for each of the remaining emotions.

TABLE IV: The table presents macro-F1 scores using the integration of labels from ChatGPT.

SSLM	W/O ChatGPT Labels	W/ ChatGPT Labels	Relative Gain
WavLM	0.350	0.353	0.77%
W2V2 R	0.331	0.335	1.08%
XLS-R-1B	0.331	0.341	3.09%
Data2Vec-A	0.329	0.338	2.74%
Hubert	0.342	0.350	2.22%
W2V2	0.321	0.325	1.28%
VQ-W2V	0.292	0.300	2.74%
W2V	0.301	0.305	1.58%
CPC	0.265	0.290	**9.45%**
DeCoAR 2	0.308	0.317	3.14%
TERA	0.295	0.306	3.52%
Mockingjay	0.275	0.298	8.49%
NPC	0.275	0.290	5.75%
VQ-APC	0.296	0.310	4.94%
APC	0.298	0.307	3.19%
FBANK	0.186	0.186	0.00%
Average	0.298	0.307	3.08%

B. Illustration for Re-labeled Data

Figure 4 shows one random example to compare the original distribution and relabeled distribution by ChatGPT (data samples "PODCAST_1631_0043_0001.wav"). We observe that ChatGPT effectively comprehends the typed descriptions conveying positive emotions, thereby assigning greater weight to the "happy" emotion category.

VI. Conclusion and Future Work

The findings of our study reveal a novel and effective approach to Speech Emotion Recognition (SER) by harnessing the capabilities of large language models (LLMs) such as ChatGPT to empower typed descriptions. By prompting ChatGPT to simulate annotators' cognitive processes and interpret typed descriptions of emotional perceptions, our method provides a more nuanced and accurate representation of emotional content compared to conventional classification models that rely solely on pre-defined categories. The consistent average relative gain of 3.08% in performance across 15 self-supervised learning models on the SUPERB codebase demonstrates the robustness and applicability of our methodology. Our research marks the first instance of leveraging ChatGPT to process and utilize typed descriptions to improve SER systems. In future work, we will use other databases containing typed descriptions, such as MSP-IMPROV [5] and NNIME [32] and other LLMs to investigate the generalization of the proposed method.

References

[1] J. Wagner *et al.*, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023. DOI: 10.1109/TPAMI.2023.3263585.

[2] J. A. Russell, "Forced-choice response format in the study of facial expression," *Motivation and Emotion*, vol. 17, no. 1, pp. 41–51, Mar. 1993. DOI: 10.1007/BF00995206.

[3] C. Busso and S. Narayanan, "Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, Sep. 2008, pp. 1670–1673.

[4] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, Oct. 2019. DOI: 10.1109/TAFFC.2017.2736999.

[5] C. Busso *et al.*, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017. DOI: 10.1109/TAFFC.2016.2515617.

[6] W. Zhang *et al.*, "Sentiment analysis in the era of large language models: A reality check," in *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 3881–3906. [Online]. Available: https://aclanthology.org/2024.findings-naacl.246.

[7] H.-C. Chou *et al.*, "Exploiting Annotators' Typed Description of Emotion Perception to Maximize Utilization of Ratings for Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7717–7721. DOI: 10.1109/ICASSP43922.2022.9746990.

[8] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[9] K. Kheiri and H. Karimi, *SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning*, 2023. arXiv: 2307.10234 [cs.CL].

[10] Y.-A. Chung *et al.*, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

[11] Y.-A. Chung, H. Tang, and J. Glass, "Vector-quantized autoregressive predictive coding," *arXiv preprint arXiv:2005.08392*, 2020.

[12] A. H. Liu, Y.-A. Chung, and J. Glass, "Non-autoregressive predictive coding for learning speech representations from local dependencies," *arXiv preprint arXiv:2011.00406*, 2020.

[13] A. T. Liu *et al.*, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 6419–6423.

[14] A. T. Liu *et al.*, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[15] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv preprint arXiv:2012.06659*, 2020.

[16] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[17] W.-N. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[18] A. Baevski *et al.*, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[19] A. Baevski *et al.*, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*, PMLR, 2022, pp. 1298–1312.

[20] A. Babu *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[21] A. Baevski *et al.*, "Vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[22] S. Schneider *et al.*, "Wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[23] A. v. d. Oord *et al.*, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[24] S.-w. Yang *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[25] A. S. Cowen and D. Keltner, "Semantic Space Theory: A Computational Approach to Emotion," *Trends in Cognitive Sciences*, vol. 25, no. 2, pp. 124–136, 2021, ISSN: 1364-6613. DOI: https://doi.org/10.1016/j.tics.2020.11.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136466132030276X.

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

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

[28] 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*, Graz, Austria, Sep. 2019, pp. 11–15. DOI: http://dx.doi.org/10.21437/SMM.2019-3.

[29] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7.

[30] H.-S. Tsai *et al.*, “Superb-sg: Enhanced speech processing universal performance benchmark for semantic and generative capabilities,” *arXiv preprint arXiv:2203.06849*, 2022.

[31] T.-h. Feng *et al.*, “Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 1096–1103.

[32] H.-C. Chou *et al.*, “NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 292–298. DOI: 10.1109/ACII.2017.8273615.