IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 October 2022, accepted 19 November 2022, date of publication 28 November 2022,
date of current version 1 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3225198

RESEARCH ARTICLE

Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition

BAGUS TRIS ATMAJA [1,2] **AND AKIRA SASOU** [1], **(Member, IEEE)**
[1] National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8560, Japan
[2] Department of Engineering Physics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia (On leave)

Corresponding author: Bagus Tris Atmaja (b-atmaja@aist.go.jp)

This work was supported by the New Energy and Industrial Technology Development Organization (NEDO), Japan, under Project JPNP20006.

ABSTRACT Self-supervised learning has recently been implemented widely in speech processing areas, replacing conventional acoustic feature extraction to extract meaningful information from speech. One of the challenging applications of speech processing is to extract affective information from speech, commonly called speech emotion recognition. Until now, it is not clear the position of these speech representations compared to the classical acoustic feature. This paper evaluates nineteen self-supervised speech representations and one classical acoustic feature for five distinct speech emotion recognition datasets on the same classifier. We calculate the effect size among twenty speech representations to show the magnitude of relative differences from the top to the lowest performance. The top three are WavLM Large, UniSpeech-SAT Large, and HuBERT Large, with negligible effect sizes among them. The significance test supports the difference among self-supervised speech representations. The best prediction for each dataset is shown in the form of a confusion matrix to gain insights into the best performance of speech representations for each emotion category based on the training data from balanced vs. unbalanced datasets, English vs. Japanese corpus, and five vs. six emotion categories. Despite showing their competitiveness, this exploration of self-supervised learning for speech emotion recognition also shows their limitations on models pre-trained on small data and trained on unbalanced datasets.

INDEX TERMS Self-supervised learning, speech emotion recognition, acoustic feature, speech processing, affective computing.

I. INTRODUCTION

Speech processing has recently flourished more than before, thanks to the advancement of deep learning. One of the major advancements in speech processing is how to learn patterns from unlabeled data as in nature. Self-supervised learning (SSL) aims to tackle this problem and has recently shown promising results in automatic speech recognition (ASR). Using small labeled data (1h), SSL has been reported to achieve a smaller word error rate (WER) than models trained on big labeled data (100h) [1]. In [2], the authors observed the performance of SSL to be comparable to that of state-of-the-art ASR with only 3% of the training data. The authors also show that the use of a full training set improves the performance score from the baseline model with labeled data.

Although SSL is claimed to be universal across various domains [3], it is necessary to investigate the effect of utilizing different SSLs for different tasks. For instance, an SSL model may work better on one task than others, while in several general tasks, the other SSLs may perform better. In this paper, we focus our study on evaluating SSL as an acoustic feature extractor to extract speech representations for speech emotion recognition. For the speech domain itself, Yang et al. [4] proposed a Speech processing Universal PERformance Benchmark (SUPERB) to find a universal speech representation for all speech processing tasks listed in their benchmark list.

Instead of universal speech representation, a course to find a subset of universal speech representation for non-semantic

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

tasks has recently begun. The task is aimed at finding a good speech representation for the tasks that do not include ASR or phoneme classification. These two semantic tasks require granularity of the input to be at a word or a phoneme level [5]. The authors proposed TRILL (TRIplet loss network), which is a transfer learning from AudioSet to NOn-Semantic Speech benchmark (NOSS). Speech emotion recognition (SER) is a subset of NOSS that focuses on recognizing affective states within the speech. As in NOSS and SUPERB, the search for a universal representation for SER has begun following those two benchmarks.

Evaluating SSL for SER can be based on the similarities between the two. Most of the speech representations based on SSL are aimed at ASR tasks. While ASR predicts the word (linguistic) content of speech, SER is aimed at recognizing paralinguistic information of the speech. Both tasks are classification tasks given a speech input. Considering processes by which various information manifested in the speech features [6], both linguistic and paralinguistic information may be embedded in the same extracted acoustic features from SSL models. Hence, it is sound to investigate the SSL models that are commonly evaluated and proposed for ASR for SER tasks. Besides that, most available speech datasets are ASR datasets upon which SSL models are built.

In addition to the findings from the previous works, this paper contributes to the following aspects:

- We evaluated the performance of nineteen self-supervised speech representations and one classical acoustic feature for speech emotion recognition on a similar architecture and five datasets to gain insights into the performance of SSL-based speech representations for the SER task;
- We measure the effect size and rank among these twenty speech representations above to show their competitiveness.

Through those contributions, we want to know about the current position of the SSL-based speech representations for the SER task compared to another non-SSL-based acoustic feature (filterbank).

The rest of this paper is organized as follows. The next section II reviews previous works related to this research. Section III describes the research methodology that includes the datasets, self-supervised speech representations as acoustic features, a classifier for mapping acoustic features to labels, and the objective evaluations to analyze the results. Section IV contains tables and figures of the experiment results. Section V discusses the data and findings. Finally, section VI closes this research article with conclusions and future directions on evaluating self-supervised speech representations for the speech emotion recognition problem.

II. RELATED WORK: SSL AND SER

Self-supervised learning (SSL) has recently been adopted in many research areas due to its effectiveness and similarity to nature. SSL learns from labeled and unlabeled data: seeking the pattern of unlabeled data given labeled data in certain intervals to generate pre-trained models. Research on SSLs and SER has been performed independently in most works. The focus of the previous works is either proposing a new SSL for a specific task (e.g., ASR [1]), for several similar tasks (e.g., [7]), for general speech tasks (e.g., [4]), or evaluating the existing acoustic features (including some SSLs) for the SER task (e.g., [8]). The following description describes the difference between our work from the previous works.

In [5], the authors proposed TRILL to achieve state-of-the-art (SOTA) results for three out of six tasks in NOSS (VoxCeleb1, **VoxForge**, **SpeechCommands**, CREMA-D, **SAVEE**, and DementiaBank). The authors then improved TRILL for mobile devices by knowledge distillation known as FRILL [7]. The previous TRILL is based on the modified version of ResNet50, which is expensive to run on mobile devices. In this paper, we do not evaluate FRILL/TRILL since the goal is not to evaluate the speech representations for mobile devices; besides, the representations are also not compatible with the S3PRL toolkit [4].

In [9], the authors reported the use of pre-trained acoustic and linguistic features for continuous SER. The evaluated methods were wav2vec for speech representation and CamemBERT for linguistic representation. The result shows the superiority of fusion between acoustic and linguistic features for SER. In several cases, the use of merely acoustic features may be enough if the representation contains rich affective information. In addition, it is necessary to evaluate beyond the wav2vec feature extractor since there are recent developments in SSLs.

The authors of [10] attempted to find a universal representation for SER. Based on the language-agnostic assumption, the authors proposed a contrastive pretraining-based SSL method by minimizing contrastive loss between a pair of different augmented matrices. The method, known as a contrastive spec, obtained better performance than MFCC, OpenSMILE-based features, and PASE+ [10]. The improvement over the last method is about 1% in terms of accuracy. Compared to other SSL reported in SUPERB [4], the improvement of the latest SSL (HuBERT) now is about 10% for the emotion recognition task. We exclude contrastive spec due to the unavailability of the pre-trained models in SUPERB and its low improvement scores.

Aldeneh et al. [11] proposed a framework to learn to extract paralinguistic embedding. The authors showed that converting synthetic-neutral speech to expressive speech based on that embedding improved the results from acoustic features and other evaluated embeddings. In addition to emotion classification, the learned embedding is also beneficial for detecting speaking style. The baseline method with a convolutional autoencoder could learn a feature transformation that highlights latent paralinguistic embedding in Mel-filterbanks (MFBs). However, no results of the proposed speech embedding have been reported for SER. This representation is also not available in SUPERB.

While previous works focused on the development of new SSLs, Keesing et al. [8] evaluated existing traditional and modern acoustic features on various speech emotion datasets. The authors found that standard features such as Interspeech 2009 feature set (IS09) still perform competitively to neural representations like VGGish [12]. The limitation of the previous work in [8] is that the study only involved nine neural network-based speech embeddings. We extend the work of Keesing et al. [8] by focusing on a larger number and newer SSLs on the common five speech emotion datasets.

III. METHODS

A. DATASETS

This paper employed five emotional speech datasets: IEMOCAP [13], MSP-IMPROV [14], MSP-PODCAST [15], CMU-MOSEI [16], and JTES [17]. IEMOCAP, MSP-IMPROV, MSP-PODCAST, and JTES are selected with four emotion categories: angry, happy/joy, sad, and neutral. CMU-MOSEI predicts six basic emotion categories: happiness, sadness, anger, fear, disgust, and surprise. IEMOCAP, MSP-IMPROV, MSP-PODCAST, and CMU-MOSEI recorded English speakers; JTES recorded Japanese speakers' emotional speech based on a Twitter corpus. The following is a short description of each dataset.

1) IEMOCAP

IEMOCAP is an interactive emotional dyadic motion capture database recorded by the Speech Analysis and Interpretation Laboratory at the University of Southern California. From the original 10,039 utterances with ten emotion categories, we only take a subset of 5,331 utterances with four emotion categories following the previous research [18]. The happy and excitement categories are merged (`exc`). The dataset is divided into five sessions, with two speakers for each session (male and female speakers). The dataset is then split into two parts: training and test. The first four sessions are for training, while the last fifth session is for a test.

2) MSP-IMPROV

MSP-IMPROV is an emotional audiovisual dataset by some authors of the IEMOCAP dataset and other researchers to propose naturalness in dyadic interactions. From the original 8,438 utterances with five primary emotion labels, we take a subset of 7,834 utterances with four emotion labels. The dataset is divided into six sessions, with two speakers for each session. Similar to IEMOCAP, the dataset is split into two parts: training and test: the first five sessions are for training, while the last sixth session is for a test.

3) MSP-PODCAST

MSP-PODCAST is a large-scale natural emotional dataset built from the existing publicly available podcast recordings. From the original 73,042 utterances with nine primary emotion labels, we take a subset of 41,388 utterances with four emotion labels. The original dataset already splits the data into training and test partitions. To match our four emotion categories selection, we choose utterances with a label in one of angry, happy, sad, neutral on both training and test partitions. The validation set provided by the authors of the dataset is merged into a training set; both "Test Set 1" and "Test Set 2" in the original dataset are merged into a test set. The final number for the training set is 32,084 utterances, and for the test, the set is 9,304 utterances.

4) CMU-MOSEI

CMU-MOSEI is both sentiment analysis and emotion recognition corpus with multimodal data (audio, video, text) [16]. The original dataset consists of 23,259 audio files, which are already split into standard training, validation, and test sets via its SDK. From that total number, we only included a total of 22,860 utterances. We merged the training and the validation data into a training set and kept the test data as a test set as it is. Any utterance which does not belong to the training, validation, and test is not included in the experiment. The final number for the training set is 18,189; for the test set is 4,662 utterances.

5) JTES

JTES is a phonetically and prosodically balanced emotional speech corpus of Japanese speakers. From the original 20,000 utterances with six emotion labels, we took a subset of 14,800 utterances with four emotion labels with speaker+text-independent criteria following the previous research [19]. Excitement is referred to as `joy` in this dataset. Of the total 100 speakers (50 males and 50 females) and 50 sentences, speakers 1-45 for each gender were used for training, while the last five speakers were allocated for the test. That splitting criterion also involves text split with text 1-40 for training and the rest of text 41-50 for the test. The speaker and text independent splitting criterion (different speakers and different sentences for training and test) resulted in 14,400 utterances for training and 400 utterances for the test.

Fig. 1 shows the distribution of emotion categories and partitions in the datasets. The portion of validation data is 20% of the number of training data in all datasets. It is shown that we accommodate both balanced and unbalanced datasets, four-emotion and six-emotion datasets, and English vs. non-English datasets in the experiments. The evaluation of SSL-based speech representations is intended to judge their performances in various characteristics of those datasets.

All experiments were conducted in speaker-independent evaluations. These evaluations are chosen to avoid bias due to the model learning the speaker-related information. Furthermore, evaluations of the JTES dataset were performed in the speaker- and text-independent condition, which is more difficult than the speaker-independent condition. By accommodating a speaker-independent condition, we aimed to build a more realistic model where the speakers in the real-life scenario are different from that evaluated in the training phase.

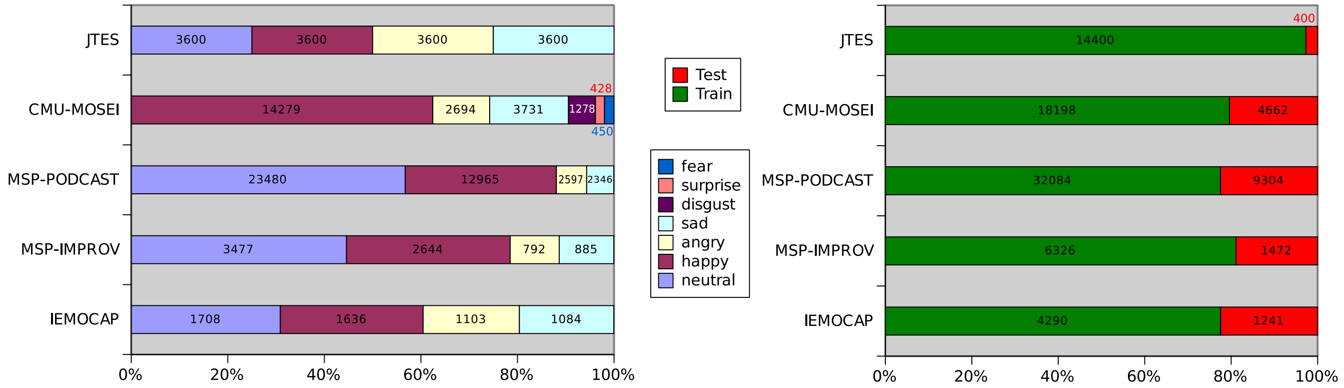

FIGURE 1. Distribution of emotion categories (left) and training/test partitions (right) in the datasets; all partitions are speaker-independent evaluations.

B. THE EVALUATED SELF-SUPERVISED SPEECH REPRESENTATIONS

In this paper, we focus on the evaluation of SSL as an acoustic feature extractor for SER, which attempts to extract affective information from speech. The selection criteria for an SSL to be included in this evaluation is the availability of the corresponding SSL in SUPERB benchmark [4]. Newer SSLs are added if the source code is available and can be incorporated into the S3PRL toolkit. We removed a low-performance SSL (PASE+) from the benchmark and added wav2vec 2.0 XLSR [20], variants of UniSpeech-SAT [21], and variants of WavLM [22].

We used log mel filterbank (FBANK) as a baseline acoustic feature for our SER system, which is also used as a baseline in other systems [4], [21], [22]. The stride and window lengths for the FBANK configuration are 10ms and 25ms. The number of the mel bins is 80 with their deltas and deltas-deltas. Hence, the dimension of FBANK for each frame is 240 (80 × 3). We then used the following self-supervised speech representations to evaluate their performances compared to FBANK in five SER datasets on the same classifier. The short description for each SSL is given below; for more detailed descriptions, please refer to the reference paper for each SSL method.

The first evaluated SSL method is the so-called autoregressive predictive coding (APC) [23]. The APC is intended as a feature extractor for a wide range of downstream tasks by incorporating a language model-like training scheme into an acoustic sequence. APC then is improved by vector quantization (VQ) version, i.e., VQ-APC [24]. The latter improvement aims at limiting the APC model capacity for a general approach to various speech-processing tasks. In the original APC version, it is difficult to quantify the amount of information by changing values in hyperparameters.

Instead of using an autoregressive model, Liu et al. [25] proposed non-autoregressive predictive coding (NPC) for a similar generative modeling purpose. Aside from the non-autoregressive method, NPC replaced RNN with CNN, and future generations with masked reconstruction. Another SSL representation called Mockingjay [26] is a similar technique that also uses masked reconstruction with bidirectional Transformers encoders (self-attention). Mockingjay outperformed FBANK significantly; however, it obtained lower performances than other previous SSL methods [4].

TERA is another generative modeling SSL which also based on Transformers encoders [27]. The main idea in TERA is the change of three orthogonal axes: time, frequency, and magnitude, to learn through a reconstruction of acoustic frames from that changes, controlled via a stochastic policy. TERA showed competitive results only for the ASR task; for other tasks, it is not clear whether it is better than previous SSL methods [4].

A modified version of contrastive predictive coding (CPC) is proposed to pretrain ASR data across languages. The modification consists of stabilization of the training and improving the CPC model. The stabilization was made by replacing batch normalization with channel-wise normalization. The improvement of the CPC model was made by replacing the linear classifier with a 1-layer Transformers network. Modified CPC showed modest performances on SUPERB tasks. It outperformed FBANK and previous SSLs for the following tasks: keyword spotting (KS), emotion recognition, and query by example (QbE).

The wav2vec is an SSL trained on large amounts of unlabeled data to generate speech representations that are fed back to improve model training. The wav2vec outperformed all previous SSLs except for emotion recognition in SUPERB [4]. The VQ technique of wav2vec, as implemented in VQ-APC, has been adopted, which results in no improvement in performance except for intent classification (IC) and speaker identification (SID). Instead, version 2.0 of wav2vec improved the model performance significantly. The wav2vec 2.0 Large outperformed all previous SSLs in the SUPERB tasks [4]. The Large version is trained with a larger network (316M vs. 95M trainable parameters) and on more extensive data (60k hr vs. 960 hr) than the Base model [4]. Similar variants also apply to the HuBERT models.

Motivated by training on extensive acoustic data with a language model over the continuous inputs, hidden-unit BERT (HuBERT) improves the previous SSL methods by masked prediction of hidden units by combining CNN encoders with Transformers. HuBERT Large outperformed all previous SSLs in SUPERB tasks except for the following tasks: KS, QbE, automatic speaker verification (ASV), and speaker diarization (SD) [4]. For those tasks, HuBERT still obtained competitive scores among other SSLs.

UniSpeech-SAT is a speaker-aware pre-training model based on the previous UniSpeech model [21]. This model is a contrastive loss model with multitask learning, which integrates utterance-wise contrastive loss with SSL objective function. On the other hand, the model utilizes an utterance-mixing strategy for data augmentation. The latter method aims at better speaker discrimination (speaker-aware training). UniSpeech-SAT Large dominated the SUPERB tasks score except for the ASR task. For the ASR task, the obtained word error rate (WER) is slightly lower than the previous HuBERT Large.

WavLM is an SSL built based on the HuBERT framework by adding a gated relative position bias and utterance mixing strategy [22]. Trained on the larger datasets (96k hours on WavLM vs. 60k hours on HuBERT), the model outperformed other previous SSLs, excluding UniSpeeh-SAT models on all of the SUPERB tasks. Similar to wav2vec, HuBERT and UniSpeech-SAT with larger models tend to obtain better performances. Most of the obtained SOTA scores on SUPERB tasks are also obtained with the large models. Only for QbE and IC tasks, the WavLM Base attains better performances than the WavLM Large.

For both UniSpeech-SAT and WavLM, we experimented with the Base, Base+, and Large models. The Base and Base+ models have a similar number of trainable parameters (94M) but are trained on different hours of data (960 hr vs. 94k hr). The Large model is trained on 96k hours of data with 316.6M trainable parameters.

The choice of FBANK and nineteen SSL speech representations is based on the previous results [4], [21], [22]. The last two SSLs are the most recent ones; they were built on top of and to improve the HuBERT model. We assume those twenty speech embeddings represent speech representations from a classical approach to the most up-to-date SSL-based speech technologies.

C. CLASSIFIER

For all SSL models, we employed the same classifier, i.e., two linear layers (fully connected network, FCN) with a simple average pooling (from frame to utterance) as used in SUPERB benchmark [4] for emotion recognition (ER) task. The input dimension depends on the acoustic feature type (see Table 2). The first layer contains 256 units. The second layer contains a number of units depending on the number of classes (n_class, e.g., 4 for JTES and IEMOCAP). The different hyperparameter values are employed instead of the original implementation based on the experiment results. These values are shown in Table 1.

TABLE 1. Hyperparameters for the classifier (FCN).

Parameter	Value
Layers	2
Units/nodes	256, n_class
Activation	ReLU
Pooling	Average
Dropout	0.1
Optimizer	Adam
learning rate	0.0001

For training the classifier, we use a batch size of 4 for both training and evaluation, except for the CMU-MOSEI dataset. Due to its large size, we used batch sizes of 2 and 1 for training and evaluation on CMU-MOSEI to avoid out-of-memory (OOM) errors. The training steps are 10000 for all datasets. We reported the test scores based on the model of the best performance on the validation set.

D. OBJECTIVE EVALUATIONS AND EFFECT SIZE

Weighted accuracy (WA) and unweighted accuracy (UA) are the most common metrics for evaluating machine learning tasks. WA is obtained by dividing the number of correct predictions by the number of data. UA is obtained by averaging accuracy for each label. UA is also known as balanced accuracy and unweighted average recall (UAR). While the UA metric is intended to deal with imbalanced datasets, WA is commonly used for balanced datasets. We reported both UA and WA since the datasets contain both unbalanced and balanced data.

To measure the effect of WA and UA for all acoustic features, we calculate **effect size** based on mean absolute deviation (MAD). The intuition behind this calculation is to get insights into the best and the worst performing speech representation by their effect sizes (the first rank to the last rank). This effect size estimates the magnitude of difference among different speech representations for SER. First, we calculate the rank of all acoustic features (based on UA or WA across five datasets) and sort them from the lowest to the highest rank. Then, we calculate the MAD of the ranks. The MAD is the mean absolute deviation of the ranks. The higher the MAD, the more different the ranks are. The effect size is calculated from the ratio of the MAD of the first rank to the MAD of the evaluated rank. While rank, as proposed in the [8], is informative for ranking features from the most to the least predictive, the effect size is more informative to report the strong and weak relationships among features. The greater the effect size, the weaker the relationship between the features.

The effect size is calculated for each SSL based on the following metric:

$$\text{Effect size}_{\text{i}} = \frac{M_{top} - M_i}{\sqrt{\dfrac{\text{MAD}_{top}^2 + \text{MAD}_i^2}{2}}}. \tag{1}$$

where M is the median and MAD is the median absolute deviation. Subscripts i and top indicate the current and the top-ranked acoustic features. Hence, the first rank will have zero effect size, while the last rank will have the highest effect size. Notice that this effect size is independent of the number of evaluated datasets.

For the interpretation of effect size, we propose the following definitions:

1) Small: effect size ≤ 0.2,
2) Medium: $0.2 <$ effect size < 0.8,
3) Large: effect size ≥ 0.8.

This interpretation is based on the [28] for the effect size with mean and standard deviation (standard deviation-based effect size).

E. EXPERIMENTAL PLATFORM

We experimented with the S3PRL toolkit available at https://github.com/s3prl/s3prl (accessed on 29 November 2021) for five different datasets called downstream. The repository contains two original configurations for IEMOCAP ('emotion') and CMU-MOSEI ('mosei'). We evaluated these two datasets and the other three datasets with the previously explained configurations (subsections III-A and III-C) instead of the original configuration given in the repository. These configurations are also stored at https://github.com/bagustris/ssl-ser.

IV. EXPERIMENT RESULTS

We cast our results into two, performance comparison and significance analysis. The first result is the main result to judge the performances of evaluated self-supervised representations for SER. The second result on significance analysis with a critical difference approach (on a side note) is added to know whether one feature set significantly differs (better) than others. Note that this calculation includes several highly-related features (wav2vec, HuBERT, WavLM, and UniSpeech), which are also shown by the results.

We reported performances comparison in WA and UA for the twenty acoustic features over five datasets. WA is beneficial for a balanced dataset (e.g., JTES), while UA is beneficial for an unbalanced dataset (CMU-MOSEI). Aside from the WA and UA, we also calculated the effect size in terms of mean absolute deviation (MAD). For each WA and UA, we also calculated the **effect size** in terms of the mean absolute deviation (MAD), p-value for the Friedman test [29], and a critical difference (CD) from the Nemenyi posthoc test. Note that although we showed CD, we chose effect size as the primary metric for evaluating SSL models since it does not depend on the number of datasets.

Table 2 and Table 4 show the WA and UA scores for the twenty acoustic features over five datasets. Between WA and UA, we choose UA as the main metric for evaluating SSL models since it exhibits the performance of unseen data. The result shows the competitiveness of self-supervised learning methods over a classical handcrafted filterbank feature. It also clearly shows that the recent self-supervised speech representations with large-size training data are superior to the other SSL with smaller training data. The big three SSLs from accuracy calculations are WavLM Large, UniSpeech-SAT Large, and HuBERT Large. However, the order of the importance of all features is not clearly shown using such measurements as average WA or UA over five datasets.

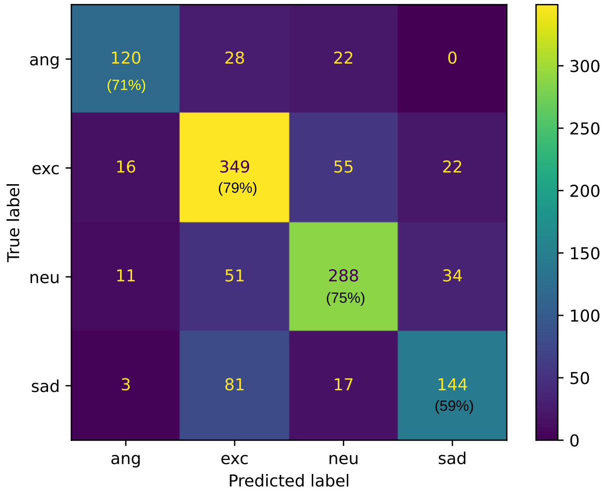

FIGURE 2. Confusion matrix of the best prediction for the IEMOCAP dataset.

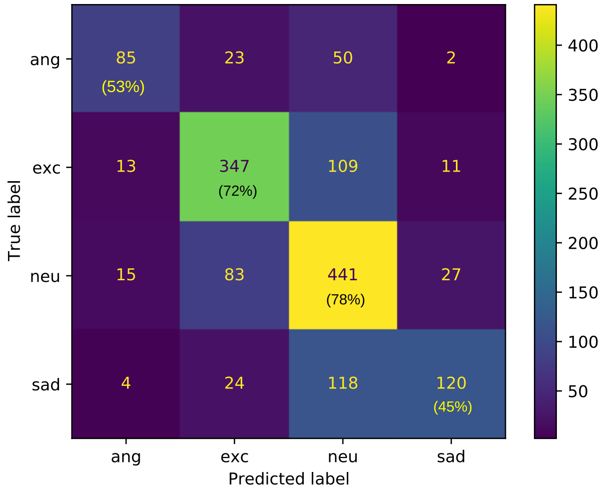

FIGURE 3. Confusion matrix of the best prediction for the MSP-IMPROV dataset.

The Friedman tests were conducted with a significance level of $\alpha = 0.05$. The tests detected significant differences in both WA and UA scores over twenty feature sets on five measurements (datasets). The p-value for WA is 1.72×10^{-10} and for UA is 4.52×10^{-11}.

In Table 3 and Table 5, we calculated the effect size of all features using a systematic method explained in the previous section. First, we computed an average rank, mean, standard deviation (std), median, and MAD of accuracy for each feature from the five datasets. Then, we calculated the effect size of each feature using (1) based on the MAD.

Finally, we reported the experiment results in the format of the best prediction confusion matrix for each dataset. Figs. 2 - 6 shows those pieces of information. The confusion matrix show WA for each emotion category; the average

TABLE 2. Weighted Accuracy (WA %) per feature over five emotional speech datasets (Friedman test $p = 1.72 \times 10^{-10}$).

Feature	Dim.	IEMOCAP	MSP-IMPROV	MSP-PODCAST	CMU-MOSEI	JTES
FBANK	240	30.94	38.45	50.54	62.89	25.00
APC	512	63.50	53.67	63.89	62.68	66.25
VQ-APC	512	61.32	54.35	64.22	63.06	67.00
NPC	512	56.57	50.27	61.59	62.46	58.75
Mockingjay	768	56.65	51.77	60.66	62.74	63.75
TERA	768	59.39	52.24	62.99	63.04	64.50
modified CPC	256	55.60	48.71	60.62	62.40	63.75
wav2vec	512	56.97	50.54	62.75	62.91	64.25
vq-wav2vec	512	49.07	47.69	60.58	62.70	63.75
wav2vec 2.0 Base	768	64.46	58.22	66.76	63.13	71.25
wav2vec 2.0 large	1024	68.65	60.60	66.29	62.57	70.00
wav2vec 2.0 XLSR	1024	64.22	56.79	65.67	62.83	65.25
HuBERT Base	768	66.96	61.14	66.58	63.13	70.75
HuBERT Large	1024	69.46	65.56	69.49	63.73	76.75
UniSpeech-SAT Base	768	67.53	60.26	66.69	63.06	69.50
UniSpeech-SAT Base+	768	70.75	62.16	67.67	63.36	70.00
UniSpeech-SAT Large	1024	71.64	**67.46**	70.21	**63.90**	75.25
WavLM Base	768	67.53	61.48	66.55	63.13	73.00
WavLM Base+	768	68.41	61.62	68.53	62.98	71.50
WavLM Large	1024	**72.60**	67.32	**70.90**	63.88	**78.25**

TABLE 3. Effect of various speech representations calculated from average WA of the test sets from five datasets. The MAD is the median absolute deviation of the ranks. The effect size calculation is based on MAD.

Feature	Mean Rank	Mean	Std	Median	MAD	Effect size
WavLM Large	1.4	70.59	5.44	70.90	3.99	0.00
UniSpeech-SAT Large	1.8	69.69	4.29	70.21	3.21	0.19
HuBERT Large	3.0	69.00	5.00	69.46	3.48	0.38
WavLM Base+	4.9	66.79	3.872	67.67	3.22	0.89
UniSpeech-SAT Base+	6.2	66.61	4.15	68.41	3.47	0.69
WavLM Base	6.5	66.34	4.47	66.55	3.23	1.20
UniSpeech-SAT Base	7.4	65.71	3.72	66.58	2.86	1.24
HuBERT Base	7.6	64.76	4.77	64.46	3.39	1.74
wav2vec 2.0 Large	8.4	65.41	3.71	66.69	3.00	1.19
wav2vec 2.0 Base	9.9	65.62	3.98	66.29	3.23	1.27
wav2vec 2.0 XLSR	11.3	61.99	4.74	63.06	3.32	2.13
APC	12.0	62.95	3.62	64.22	2.51	2.00
VQ-APC	13.2	60.43	4.95	62.99	3.69	2.06
TERA	13.4	62.00	4.84	63.50	3.33	2.01
Mockingjay	14.6	59.48	5.73	62.75	4.58	1.87
NPC	16.0	59.11	4.96	60.66	3.92	2.60
modified CPC	17.6	57.93	4.88	58.75	3.61	3.19
wav2vec	18.0	56.76	7.75	60.58	6.70	1.87
vq-wav2vec	18.2	58.22	6.17	60.62	4.85	2.32
FBANK	18.6	41.56	15.28	38.45	12.12	3.60

value of these WAs is the reported UA. The highest UA for IEMOCAP comes from UniSpeech-SAT Base+. For the other four datasets, the highest UA is from WavLM Large. Hence, the confusion matrix for IEMOCAP is the one obtained by UniSpeech-SAT Base+, while WavLM Large assists other confusion matrices.

V. DISCUSSION

First, we saw a similar pattern of scores between WA (Table 2) and UA (Table 4) for twenty acoustic features. The clearer patterns were shown in Table 3 and Table 5 for the rank of the features for WA and UA. For the interpretation of effect size, the evaluated speech representations can be categorized into three groups: top features with small or negligible effect, features with medium effect from top performance, and features with large effect from the top performance. The big three in the first group are WavLM Large, UniSpeech-SAT Large, and HuBERT Large. The SSLs in the second group with medium effect from the top are WavLM Base+, UniSpeech-SAT Base+, WavLM Base, UniSpeech-SAT Base, HuBERT Base, wav2vec 2.0 Large, and wav2vec 2.0 Base. The rest belong to the third group with large effects from the top performance. The interpretation is based on the UA scores instead of the WA scores.

Next, the data showed that pre-trained models from large-size data achieved superior performances over the small-size data. It can be seen that WavLM Large, along with recent large SSL models (HuBERT and UniSpeech-SAT), dominates top performances for SER tasks than previous models. However, the trend only applies to the recent model with the same size, i.e., HuBERT Large vs. UniSpeech-SAT Large vs. WavLM Large. The first two models obtain small

TABLE 4. Unweighted Accuracy (UA %) per feature over five emotional speech datasets (Friedman test $p = 4.52 \times 10^{-11}$).

Feature	IEMOCAP	MSP-IMPROV	MSP-PODCAST	CMU-MOSEI	JTES
FBANK	25.00	25.00	25.00	16.71	25.00
APC	64.75	46.56	38.55	17.68	66.25
VQ-APC	63.66	48.04	38.01	17.92	67.00
NPC	60.87	36.25	33.80	16.90	58.75
Mockingjay	59.09	38.77	34.75	17.27	63.75
TERA	61.69	43.16	35.57	18.63	64.50
modified CPC	58.78	34.85	33.31	16.73	63.75
wav2vec	61.13	38.94	35.06	17.36	64.25
vq-wav2vec	52.63	33.65	33.29	16.69	63.75
wav2vec 2.0 Base	65.83	52.20	40.82	18.55	71.25
wav2vec 2.0 Large	65.60	53.91	39.09	18.03	70.00
wav2vec 2.0 XLSR	61.21	50.22	36.55	16.67	65.25
HuBERT Base	67.81	57.38	41.88	19.06	70.75
HuBERT Large	68.60	61.64	44.99	21.34	76.75
UniSpeech-SAT Base	69.45	57.00	41.26	17.86	69.50
UniSpeech-SAT Base+	**71.11**	58.03	42.67	19.30	70.00
UniSpeech-SAT Large	68.16	62.11	44.06	22.45	75.25
WavLM Base	68.06	57.36	41.44	19.20	73.00
WavLM Base+	69.47	57.56	42.79	19.85	71.50
WavLM Large	70.83	**63.96**	**50.35**	**23.92**	**78.25**

TABLE 5. Effect of various speech representations calculated from average UA of the test sets from five datasets. The MAD is the median absolute deviation of the ranks. The effect size calculation is based on MAD.

Feature	Mean Rank	Mean	Std	Median	MAD	Effect size
WavLM Large	1.2	57.46	21.38	63.96	16.26	0.00
UniSpeech-SAT Large	3.0	54.66	21.99	61.64	17.20	0.14
HuBERT Large	3.2	54.41	21.28	62.11	16.92	0.11
WavLM Base+	4.2	52.23	21.42	57.56	16.73	0.39
UniSpeech-SAT Base+	4.7	52.22	21.69	58.03	16.99	0.36
WavLM Base	6.2	51.81	21.88	57.36	17.19	0.39
UniSpeech-SAT Base	6.8	51.38	21.31	57.38	16.72	0.40
HuBERT Base	8.4	51.01	21.86	57.00	17.16	0.42
wav2vec 2.0 Large	8.6	49.73	21.10	52.20	16.04	0.73
wav2vec 2.0 Base	9.5	49.33	21.20	53.91	16.61	0.61
wav2vec 2.0 XLSR	11.6	46.93	20.03	48.04	15.17	1.01
APC	12.0	46.76	20.10	46.56	14.99	1.11
VQ-APC	12.6	44.71	19.01	43.16	14.71	1.34
TERA	14.2	45.98	19.81	50.22	15.50	0.87
Mockingjay	14.8	43.35	19.47	38.94	15.47	1.58
NPC	16.2	42.73	18.95	38.77	14.96	1.61
modified CPC	17.0	41.31	18.47	36.25	14.80	1.78
wav2vec	17.6	41.48	19.48	34.85	15.82	1.81
vq-wav2vec	18.6	40.00	18.39	33.65	14.55	1.96
FBANK	19.6	23.34	3.71	25.00	2.65	3.34

effect sizes (≤ 0.2) from the third model on the UA scores. Between models with different sizes, e.g., WavLM Base vs. UniSpeech-SAT Large, the former with smaller training data attained lower performances than the latter trained on the larger data. This result shows the dependency of the training data size on the performance of recent SSL models for the SER tasks. This finding also highlights the necessity of building SSL models that requires less data than the current models. The training data size is equal to the requirement of computing resources to generate the model. For instance, using the same "base" training data as UniSpeech-SAT and WavLM, the next SSL "base" models should attain better performances than the current UniSpeech-SAT and WavLM Large models. The base models are evaluated on 960 hours of data, while the large models are evaluated on 95k hours of data [22].

The necessity of building SSL models that requires less data is in line with the motivation of SSL itself. SSL is motivated by the nature that human does not require an exact number of data and labels to learn. Not only unlabeled data, but humans also can learn from less data than current machine learning requires (e.g., the learning process of infants to recognize emotions). The requirement of high computing costs for training large data is only affordable for big companies and big institutions. For small institutions, training large data is likely undoable. Such an effort has been attempted by the authors of [30]; the proposed distilled model still depends on the previous larger model as a teacher in a teacher-student learning model. This gap has also been shown in past research: while it only needs 3% of training data to match the baseline performance, it needs a complete training set to improve the performance [2].

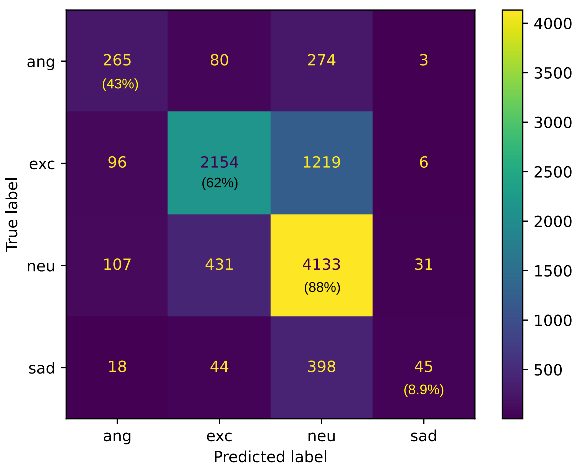

FIGURE 4. Confusion matrix of the best prediction for the MSP-PODCAST dataset.

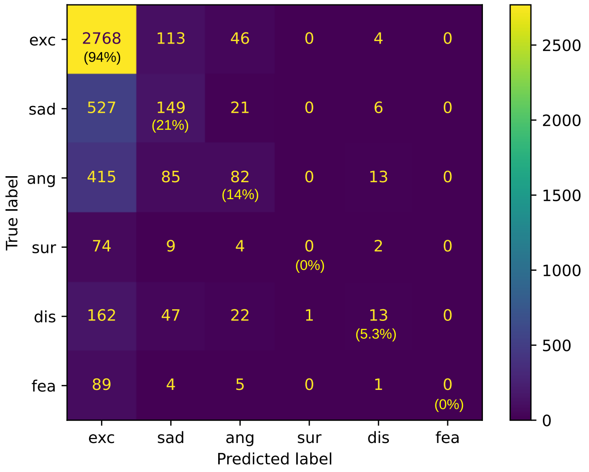

FIGURE 5. Confusion matrix of the best prediction for the CMU-MOSEI dataset.

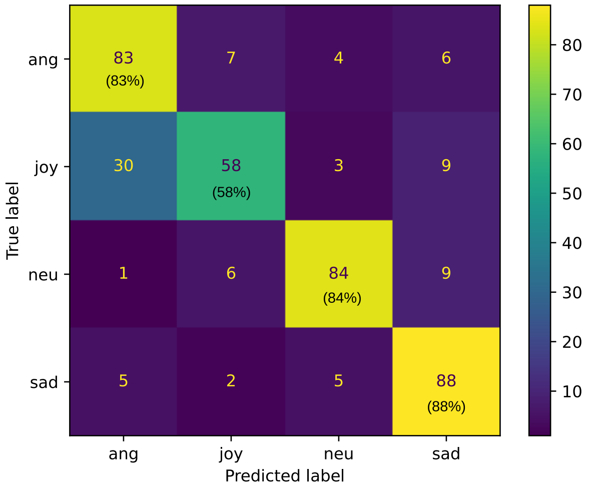

FIGURE 6. Confusion matrix of the best prediction for the JTES dataset.

One of the main issues in SER is the low availability of evaluation of non-English datasets [31]. The majority of SER research has been conducted in the English language. Here, we added a Japanese SER evaluation for comparing the performance across datasets. Surprisingly, the result shows that the highest performances among the five datasets for both WA and UA are the ones from the Japanese dataset. Notice that the scores for WA are the same as UA since the dataset is perfectly balanced. This balanced characteristic helps JTES attain the highest UA scores among other unbalanced datasets. One interesting result is that model trained on multi-language, i.e., wav2vec 2.0 XLSR, showed no better performance than other models trained on mono-language (English). This fact also applies even though for the same wav2vec variants. The previous result on different datasets (IEMOCAP, TESS, SAVEE, EMA, German database EmoDB, and Italian database EMOVO) showed that XLSR obtained better representations for multilingual SER [20]. The lower results obtained by the XLSR model (multiple languages) here for mono language may indicate that the universality of emotion needs to be adjusted for specific languages. Training mono language for mono language is still better than training multiple languages for mono language in the SER task.

In contrast to the well-balanced JTES dataset, CMU-MOSEI is a very unbalanced dataset with six emotion categories. As a result, the performance scores obtained by CMU-MOSEI are the worst among others. SSL-based speech representations fail to predict emotion other than the happy/excitement category (Fig. 5). This failure shows the limitation of deep learning models, including the SSL-based speech representations; they need to be trained on balanced data. While the SSL methods can reduce the number of labeled data, each category's distribution needs to be balanced for better performance.

The unbalanced distribution of each emotion category for the evaluated datasets (except for JTES) prevents further analysis of which emotion the model currently learns better. The need for balancing the datasets opens future research contributions for this analysis. For instance, it is difficult to judge in which category the best model with WavLM Large learns better, as shown in Figs. 2 - 5. Using a well-balanced JTES dataset (Fig. 6), the current model learned best on sad among other emotion categories.

In contrast to the need for well-balanced data, humans can learn well from unbalanced data. SSL, in some portions, is close to the human ability to learn. The nature of the data itself is unbalanced; hence, there is a need to improve SSL method to be able to learn about unbalanced data. This ability will make deep learning models more general and flexible, as well as closer to the human ability to learn.

Another limitation that we did not explore here is the ability of SSL to predict large emotion categories. While we show that the SSL methods achieve comparable performance across datasets with four emotion categories, the result becomes the worst on six emotion categories. Although we suspect that the cause of the low performance is unbalanced data, it is noteworthy to see if the SSL methods can obtain

TABLE 6. Comparison of the best unweighted accuracy (UA %) obtained by SSLs to the scores reported in the literature by non-SSL-based acoustic features.

Method	IEMOCAP	MSP-IMPROV	MSP-PODCAST	CMU-MOSEI	JTES
SER with multitask learning [32]	**78.15**	-	-	-	-
Regional saliency [33]	61.8	53.6	-	-	-
X-Vectors meet emotions [34]	65.95	-	**57.42**	-	-
Multi-stream attention [35]	-	-	-	-	73.4
SSL (ours)	71.11	**63.96**	50.35	**23.92**	**78.25**

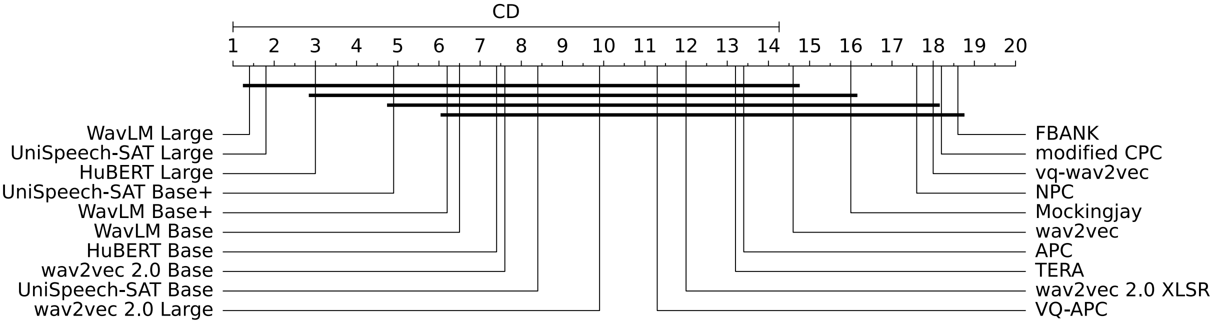

FIGURE 7. Critical difference among self-supervised speech representations and filterbank on WA scores, CD = 13.26 ranks. Groups of features whose performances are not significantly different (Nemenyi test at $p = 0.05$) are connected by bold lines.

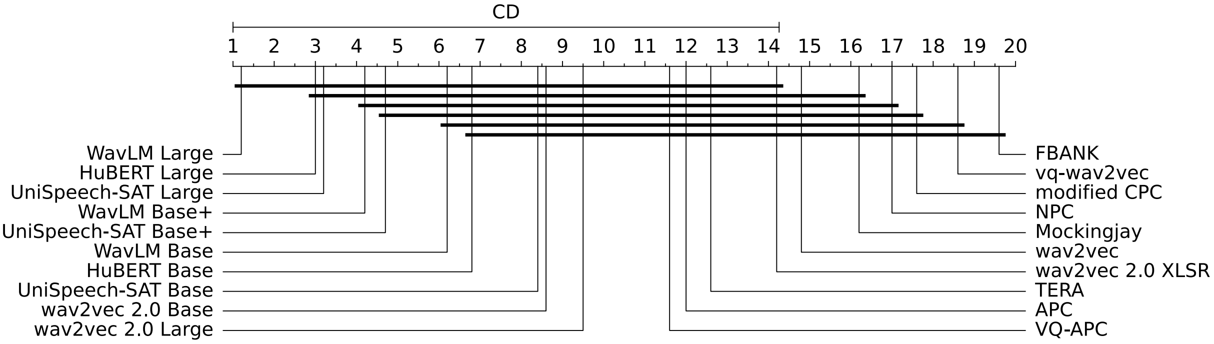

FIGURE 8. Critical difference among self-supervised speech representations and filterbank on UA scores, CD = 13.26 ranks. Groups of features whose performance is not significantly different (Nemenyi test at $p = 0.05$) are connected by bold lines.

good performances on a more prominent number of emotion categories.

Finally, we performed a benchmark test of our best SSL scores compared to the scores reported in the literature under similar conditions. Although this research does not intend to propose a new method, a benchmark of SSL evaluation in this research with non-SSL-based results from the literature will enable us to better understand the current SSL evaluation for SER. We required the benchmarked scores to come from published papers instead of pre-prints for validity. The condition for training and test data is also required to be the same as our method. No other modalities are involved in training data; if possible, no data augmentation is allowed. Table 6 shows a benchmark of our best SSL methods with scores in the literature (in terms of UA).

The benchmark table (Table 6) shows that current SSL methods are competitive with respect to methods reported in the literature. The SSL methods achieve better performances than scores reported in the literature for MSP-IMPROV, CMU-MOSEI, and JTES. For IEMOCAP, a method by multitasking approach (SER and ASR) achieves a better UA score. The reported score is cross-validation from five folds. For MSP-IMPROV, the method utilizing X-Vectors achieves a better UA score. For CMU-MOSEI, we can not find comparable scores with a similar setup. Note that the baseline paper of CMU-MOSEI reported the UA for segmented data [16]; this research evaluated the raw audio data. For JTES, the reported benchmark comes from augmented data. Without data augmentation, our SSL model achieves a better UA score by about 5% improvement.

VI. CONCLUSION

This paper reported an evaluation of SSL-based speech representations on the emotion recognition task. The results showed that all nineteen SSL-based speech representations performed substantially better than the classical filterbank on the emotion recognition task. We calculated the effect size for each SSL and showed their rank from the top (WavLM) to the bottom (vq-wav2vec) for the five-dataset speech emotion recognition task. Both weighted and unweighted accuracy evaluations showed a similar gap in effect size between the top and the worst scores; the gap is large by the definition of the proposed effect size interpretation (meaning a weak relationship between top SSL and filterbank). The top scores were obtained by the recent SSL models trained on large-scale speech data. These three big SSL-based acoustic features (WavLM Large, UniSpeech-Sat Large, HuBERT Large) share negligible effects among them. Evaluations of the five datasets with different characteristics showed the strengths and limitations of the current SSL models. The characteristics include English vs. Japanese language, balanced vs. unbalanced data, and four vs. six emotion categories. The strength is that the SSLs could extract more affective information than conventional FBANK features. The limitations, for instance, are that the SSLs need to be trained on well-balanced large data and that they are not able to learn about a large number of emotion categories.

In future work, we plan to evaluate more datasets to provide a unified SSL-based speech emotion recognition benchmark. As highlighted in the previous discussions, proposing a new SSL method for evaluating SSL-based speech emotion recognition is a research challenge. Future contributions can be made to tackle the limitations of the current SSLs, such as a smaller size of pre-training data, a large number of emotion categories, and unbalanced data.

APPENDIX

A. SIGNIFICANCE ANALYSIS

In addition to the Friedman test for WA and UA, p-values were also evaluated using the Nemenyi test to group features whose performance is not significantly different. Figs. 7 and 8 shows groups of these features with the bold lines. These figures compare features against each other from the Nemenyi posthoc test in critical-difference diagrams [29]. Notice that both figures show that significant differences using critical differences are unreliable in some cases. For instance, both figures suggest that FBANK performance would essentially be the same as HuBERT Base for this small data evaluation. On the other side, the similar SSLs from the related models (WavLM, UniSpeech, HuBERT, wav2vec) are grouped as not significantly different, which represent their relationships. Given the unreliability of significance analysis by a critical-difference diagram (probably due to small datasets) among different features, the previous analysis by the Friedman test on five different datasets should be used as the main analysis for the significance of the results (datasets) across features. The small $p - values$ on Friedman tests indicate that each SSL experiment on each dataset is significantly different from the others. For comparing the performance of features (SSLs), the proposed effect size based on the mean absolute deviation is more reliable than these CD diagrams.

REFERENCES

[1] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7689–7693.

[2] Y. Zhang et al., "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1519–1532, Oct. 2022, doi: 10.1109/JSTSP.2022.3182537.

[3] A. Tamkin, V. Liu, R. Lu, D. Fein, C. Schultz, and N. Goodman, "DABS: A domain-agnostic benchmark for self-supervised learning," in *Proc. NIPS*, 2021, pp. 1–22.

[4] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-T. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-Y. Lee, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, Aug. 2021, pp. 1194–1198.

[5] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. D. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," in *Proc. Interspeech*, Oct. 2020, pp. 140–144.

[6] H. Fujisaki, "Prosody, information, and modeling-with emphasis on tonal features of speech," in *Proc. Workshop Spoken Lang. Process.*, 2003.

[7] J. Peplinski, J. Shor, S. Joglekar, J. Garrison, and S. Patel, "FRILL: A non-semantic speech embedding for mobile devices," in *Proc. Interspeech*, Aug. 2021, pp. 1204–1208.

[8] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech," in *Proc. Interspeech*, Aug. 2021, pp. 3415–3419.

[9] M. Macary, M. Tahon, Y. Esteve, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 373–380.

[10] A. Nandan and J. Vepa, "Language agnostic speech embeddings for emotion classification," in *Proc. Workshop Self-Supervision Audio Speech (ICML)*, 2020.

[11] Z. Aldeneh, M. Perez, and E. Mower Provost, "Learning paralinguistic features from audiobooks through style voice conversion," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 4736–4745.

[12] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.

[13] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[14] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan. 2017.

[15] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct. 2019.

[16] A. Zadeh, P. P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, and L. P. and Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist.*, vol. 1, 2018, pp. 2236–2246.

[17] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *Proc. Conf. Oriental Chapter Int. Committee Coordination Standardization Speech Databases Assessment Techn. (O-COCOSDA)*, Oct. 2016, pp. 16–21.

[18] B. T. Atmaja and M. Akagi, "Speech emotion recognition based on speech segment using LSTM with attention model," in *Proc. IEEE Int. Conf. Signals Syst. (ICSigSys)*, Jul. 2019, pp. 40–44.

[19] B. T. Atmaja and A. Sasou, "Effect of different splitting criteria on the performance of speech emotion recognition," in *Proc. IEEE Region 10th Conf. (TENCON)*, Dec. 2021, pp. 760–764.

[20] Z. Zhang, X. Zhang, M. Guo, W.-Q. Zhang, K. Li, and Y. Huang, "A multilingual framework based on pre-training model for speech emotion recognition," in *Proc. APSIPA Annu. Summit Conf.*, Dec. 2021, pp. 750–755.

[21] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, ''Unispeech-Sat: Universal speech representation learning with speaker aware pre-training,'' in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2022, pp. 6152–6156.

[22] S. Chen et al., ''WavLM: Large-scale self-supervised pre-training for full stack speech processing,'' *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022, doi: 10.1109/JSTSP.2022.3188113.

[23] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, ''An unsupervised autoregressive model for speech representation learning,'' in *Proc. Interspeech*, Sep. 2019, pp. 146–150.

[24] Y. A. Chung, H. Tang, and J. Glass, ''Vector-quantized autoregressive predictive coding,'' in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Oct. 2020, pp. 3760–3764.

[25] A. H. Liu, Y.-A. Chung, and J. Glass, ''Non-autoregressive predictive coding for learning speech representations from local dependencies,'' in *Proc. Interspeech*, vol. 2, Aug. 2021, pp. 3730–3734.

[26] A. T. Liu, S.-W. Yang, P.-H. Chi, P.-C. Hsu, and H.-Y. Lee, ''Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6419–6423.

[27] A. T. Liu, S.-W. Li, and H.-Y. Lee, ''TERA: Self-supervised learning of transformer encoder representation for speech,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2351–2366, 2021.

[28] J. Cohen, ''A power primer,'' *Psychol. Bull.*, vol. 112, no. 1, pp. 155–159, Jul. 1992.

[29] J. Demšar, ''Statistical comparisons of classifiers over multiple data sets,'' *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Dec. 2006.

[30] H.-J. Chang, S.-W. Yang, and H.-Y. Lee, ''DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 7087–7091, doi: 10.1109/ICASSP43922.2022.9747490.

[31] B. T. Atmaja, A. Sasou, and M. Akagi, ''Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion,'' *Speech Commun.*, vol. 140, pp. 11–28, May 2022.

[32] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, ''Speech emotion recognition with multi-task learning,'' in *Proc. Interspeech*, Aug. 2021, pp. 4508–4512.

[33] Z. Aldeneh and E. M. Provost, ''Using regional saliency for speech emotion recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2741–2745.

[34] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, ''X-vectors meet emotions: A study on dependencies between emotion and speaker recognition,'' in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2020, pp. 7169–7173.

[35] Y. Chiba, T. Nose, and A. Ito, ''Multi-stream attention-based BLSTM with feature segmentation for speech emotion recognition,'' in *Proc. Interspeech*, Oct. 2020, pp. 3301–3305.

BAGUS TRIS ATMAJA received the B.E. and M.E. degrees from the Sepuluh Nopember Institute of Technology, in 2009 and 2012, respectively, and the Ph.D. degree in information science with a focus on speech emotion recognition from the Japan Advanced Institute of Science and Technology, in 2021. Then, he was employed as a Docent with the Vibrastic Laboratory, Sepuluh Nopember Institute of Technology. He is currently a Postdoctoral Researcher with the Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST). His research interests include speech processing, including source separation, speech enhancement, and speech (emotion) recognition.

AKIRA SASOU (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in electrical engineering from Tokyo Denki University, in 1994, 1996, and 1999, respectively. He is currently with the Signal Processing Research Group, Department of Information Technology and Human Factor, National Institute of Advanced Industrial Science and Technology (AIST).