Journal of Signal Processing, Vol.30, No.5, pp.161-171, September 2026

PAPER

Evaluating Multilingual Self-Supervised Learning Models for Multilingual Speech Emotion Recognition

Bagus Tris Atmaja[1] and Akira Sasou[2]

[1]Nara Institute of Science and Technology, Ikoma, Japan
[2]National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan
E-mail: bagus.tris@naist.ac.jp, a-sasou@aist.go.jp

Abstract **Multilingual speech emotion recognition (SER) represents a challenging task that bridges monolingual and cross-lingual approaches, benefiting from larger datasets while leveraging in-group cultural advantages. While previous studies focused on architectural solutions, recent advances in self-supervised learning (SSL) for automatic speech recognition have produced multilingual pre-trained models with superior performance. This study systematically evaluates 13 SSL models, including five models specifically trained on multilingual data. We conducted experiments on the EmoFilm dataset containing English, Italian, and Spanish samples using both fixed split and five-fold cross-validation scenarios in speaker-independent settings. Our results demonstrate that multilingual SSL models, particularly XLS-R variants, significantly outperform previous state-of-the-art monolingual SSL models such as UniSpeech-SAT and WavLM for multilingual SER tasks. The findings suggest that multilingual pre-training provides substantial advantages for cross- and within-language emotion recognition, with XLS-R achieving the highest performance across all evaluation scenarios. This work provides the first comprehensive comparison of multilingual SSL models for SER and establishes new benchmarks for multilingual speech emotion recognition research. Additionally, we evaluated the improvements for each language as well as for each emotion category.**

Keywords: speech emotion recognition, self-supervised learning, multilingual, cross-lingual, XLS-R

1. Introduction

Speech emotion recognition (SER) has evolved significantly in the last two decades [1], yet most research has focused on monolingual approaches [2–4]. While emotional expressions are considered universal across cultures [5–7], the computational challenge of multilingual SER remains largely unresolved. Traditional approaches have relied on conventional acoustic features and architectures that struggle to capture the complex cross-linguistic patterns inherent in emotional speech across different languages and cultures.

Recent advances in self-supervised learning (SSL) have revolutionized automatic speech recognition and related tasks, with models like XLS-R [8] demonstrating superior performance through multilingual pretraining on massive datasets. However, the potential of these multilingual SSL models for emotion recognition tasks has been underexplored, particularly in systematic comparative studies. Previous multilingual SER research has achieved modest performance, with the best reported results reaching only 68.0% unweighted accuracy [9] using conventional eGeMAPS features.

While SSL models trained on multilingual data should, in theory, outperform monolingual models on multilingual tasks, there is limited evidence on this [10]. Moreover, previous studies either focused on architectural solutions [11–14] or evaluated limited SSL models without a comprehensive comparison of multilingual versus monolingual pre-training approaches. Nor has the effect of model size on performance been thoroughly investigated.

The proposed study expands the model pool, refines evaluation protocols, and introduces a reproducible benchmarking framework (S3PRL-SER) from previous studies. It shifts the focus from architectural tweaks to the impact of multilingual pretraining, offering a more nuanced understanding of cross-lingual emotion generalization. It also introduces language-

specific and emotion-specific performance diagnostics, which were less emphasized in the earlier study.

This study addresses the gap through the first comprehensive evaluation of 13 SSL models for multilingual SER, including five specifically trained on multilingual data. Our key findings reveal that multilingual SSL models, particularly XLS-R variants, significantly outperform state-of-the-art monolingual SSL models (UniSpeech-SAT and WavLM) by substantial margins. Specifically, we achieve 88.95% unweighted accuracy with XLS-R 2B, which represents a breakthrough performance that surpasses previous multilingual SER results by over 20 percentage points.

We conducted rigorous experiments on the EmoFilm dataset, which includes English, Italian, and Spanish samples, using both a fixed split and 5-fold cross-validation in a speaker-independent setting. Our analysis goes beyond overall performance to investigate language-specific improvements and emotion-category performance, showing that multilingual pre-training offers significant benefits for cross-language emotion recognition.

Our findings demonstrate *that the choice between multilingual, cross-lingual, and monolingual SSL models is critical* for multilingual SER applications. We show that XLS-R models exhibit strong English bias (+27.9% improvement) while maintaining competitive performance across all languages, with important implications for real-world deployment. The performance-consistency analysis reveals trade-offs between average improvement and cross-language stability, providing guidance for model selection in different scenarios.

All experimental configurations are made *publicly available* through our S3PRL-SER toolkit[1], enabling standardized evaluation of SSL models for SER tasks. This work establishes new benchmarks for multilingual speech emotion recognition and provides a foundation for future research in this domain.

2. Dataset and Partition

The EmoFilm dataset [15] is one of the multilingual SER datasets that are publicly available (restricted access) via a request for academic purposes only. The dataset consists of 1115 samples from English (EN, 343 samples), Italian (IT, 413 samples), and Spanish (SP, 359 samples) movies. The Italian and Spanish are the dubbed versions of the English movies. Five emotion categories are studied in this dataset: anger, fear, happiness, sadness, and contempt (at the 20% chance level). There is no official split of the data into training/test sets for evaluating speech emotion recognition tasks.

We evaluated two scenarios with a fixed training/test split and cross-validation (Table 1). Both are **speaker-independent** scenarios. *The first fixed split is for benchmarking our evaluated SSLs with previous methods, while the second cross-validation (CV) is aimed at evaluating the robustness of SSLs across different training and test sets and to resolve ambiguous rankings that may arise from a single fixed split.* We split the entire dataset into five balanced folds, each fold containing 223 samples. In both scenarios, we allocated 20% of the training samples to dev(elopment), i.e., 185 samples for the fixed split and 177 for CV. We also designed each fold in CV to contain near-language-balanced samples (standard deviations of 1.85, 3.54, and 2.65 for EN, IT, and SP, respectively). By these criteria, we would like to maximize the outputs of SSLs, since the distribution of each language's samples matters during training.

Table 1 Distribution of samples in the EmoFilm dataset in the fixed split and cross-validation split

Partition	EN	IT	SP	Total
Fixed split				
Train + Dev	293	347	294	934
Test	50	66	65	181
Cross-validation				
Fold 1	71	65	87	223
Fold 2	67	73	83	223
Fold 3	69	75	79	223
Fold 4	70	72	81	223
Fold 5	66	74	83	223

The original audio files with 44.1 kHz in WAV format were downsampled to 16 kHz, also in WAV format. This lower sampling rate was required since all evaluated SSLs are only compatible with a 16 kHz sampling rate. The downsampling was performed using the SoX audio tool, and no other pre-processing was performed except for this audio sampling rate manipulation. Metadata in CSV files for fixed split and CV were created to help S3PRL process the audio files for acoustic feature extraction and classification.

Fig. 1 shows the distribution of samples in the cross-validation evaluation. It can be seen that the distribution of samples in each fold is nearly class-balanced for the training and test data (supported by the small standard deviations). A recent study revealed that dataset balancing could hurt model performance [16]; hence, we left the data unbalanced in the training, development, and test sets.

3. Evaluated SSL Models

We evaluated 13 SSL models available in the S3PRL toolkit [17]. The list of evaluated SSL models and their S3PRL Upstream IDs is shown in Table 2. The S3PRL toolkit is a collection of pre-trained SSL models and their corresponding configuration files. The toolkit also provides a unified interface to extract

[1] https://github.com/bagustris/s3prl-ser/

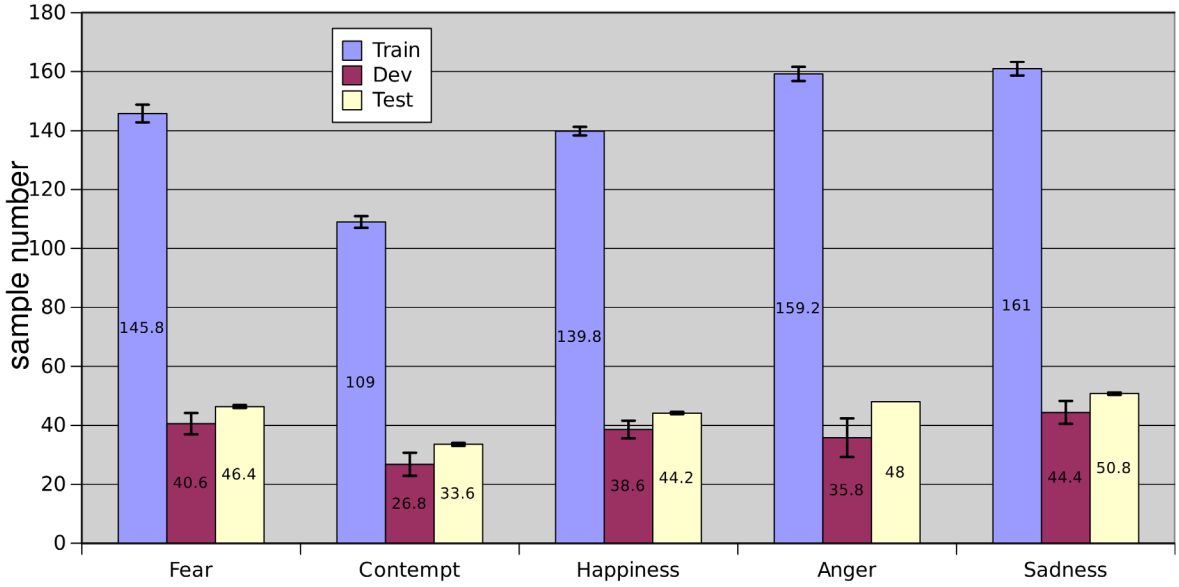

Fig. 1 Average distribution of samples for emotion categories with their standard deviation for training, dev, and test sets in five-fold cross-validation of EmoFilm dataset

acoustic features from the SSL models and to train and evaluate the downstream task with deep-learning models from PyTorch.

The general signal processing pipeline for SSL-based SER operates on a raw speech waveform $\mathbf{x} \in \mathbb{R}^T$ sampled at 16 kHz, where T denotes the total number of samples. A CNN-based feature extractor f_θ first maps the waveform to a sequence of latent feature vectors:

$$\mathbf{Z} = f_\theta(\mathbf{x}) = [z_1, z_2, \ldots, z_L], \quad z_l \in \mathbb{R}^{d_f} \tag{1}$$

where L is the number of output frames and d_f is the CNN feature dimension. These latent features are subsequently processed by a Transformer encoder g_ϕ to yield contextualized representations:

$$\mathbf{C} = g_\phi(\mathbf{Z}) = [c_1, c_2, \ldots, c_L], \quad c_l \in \mathbb{R}^{d_c} \tag{2}$$

where d_c is the Transformer hidden dimension. The SSL models differ primarily in their pre-training objectives. The Wav2Vec 2.0-based models (including XLSR-53 and XLS-R variants) are trained with a contrastive loss over randomly masked time steps:

$$\mathcal{L}_{\text{ctr}} = -\log \frac{\exp\big(\text{sim}(\mathbf{c}_t, \mathbf{q}_t)/\kappa\big)}{\sum_{\tilde{\mathbf{q}} \in \mathcal{Q}_t} \exp\big(\text{sim}(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa\big)} \tag{3}$$

where $\mathbf{q}_t$ is the quantized target representation at masked time step t, $\mathcal{Q}_t$ is a set of distractors (negatives) that includes the true quantized target, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and κ is a temperature hyperparameter [8, 18, 19]. HuBERT-based models (HuBERT, mHuBERT, LightHuBERT, WavLM, and UniSpeech-SAT) instead adopt a masked prediction objective:

$$\mathcal{L}_{\text{mask}} = -\sum_{t \in \mathcal{M}} \log p\big(\hat{z}_t \mid \mathbf{X}_{\setminus \mathcal{M}}\big) \tag{4}$$

where $\mathcal{M}$ is the set of masked time steps, $\hat{z}_t$ are discrete pseudo-labels obtained by offline k-means clustering of acoustic features, and $\mathbf{X}_{\setminus \mathcal{M}}$ denotes the unmasked input context [20–22].

The choice of 13 SSL models is based on the previous study [2]. In that study, the authors evaluated 19 SSL models and one conventional acoustic feature (filter bank). We took five SSL models from that study, two of which were the best-performing SSL in that study (UniSpeech-SAT Large and WavLM Large). The other three are Wav2Vec 2.0 Large (used as a baseline) [18], Wav2Vec 2.0 XLSR (XLSR-53) [19], and HuBERT Large [20]. We added the larger Wav2Vec 2.0 Large (LV-60 + CV + SWBD + FSH) [18] to observe the effect of enlarging training data. The other seven SSLs are chosen based on the multilingual training (XLS-R 300M, XLS-R 1B, and XLS-R 2B [8] and mHuBERT [23]) and their performances on the SUPERB emotion recognition (ER) benchmark[2] (LightHuBERT Stage1, HuBERT Base Robust MGR, data2vec Large [24]).

The purpose of choosing multilingual pre-trained models is to evaluate their effectiveness and benefits compared to models trained in a monolingual setting for this multilingual SER task. In this regard, we evaluated five multilingual pre-trained models (mHuBERT and XLSR variants) with eight other models trained in monolingual settings. A previous study

[2]https://superbbenchmark.org/

Table 2 List of evaluated SSL models with their S3PRL Upstream IDs

No	SSL Name	S3PRL Upstream ID
1.	Wav2Vec 2.0 Large #1 * [18]	wav2vec2_large_ll60k
2.	Wav2Vec 2.0 Large #2 ** [18]	wav2vec2_large_lv60_cv_swbd_fsh
3.	XLSR-53 [19]	xlsr_53
4.	XLS-R 300M [8]	xls_r_300
5.	XLS-R 1B [8]	xls_r_1b
6.	XLS-R 2B [8]	xls_r_2b
7.	HuBERT Large [20]	hubert_large_ll60k
8.	LightHuBERT Stage1 [25]	lighthubert_stage1
9.	HuBERT Base Robust MGR [26]	hubert_base_robust_mgr
10.	mHuBERT [23]	mhubert
11.	UniSpeech-SAT Large [22]	unispeech_sat_large
12.	data2vec Large [24]	data2vec_large_ll60k
13.	WavLM Large [21]	wavlm_large

* trained on ll60k

** trained on lv60-cv-swbd-fsh

shows the potential benefit of using a multilingual pre-trained model [27] for multilingual SER, while another study did not find its effectiveness (XLSR model) for the Japanese SER [2]. As in those previous studies, we extracted the representations from the last layer of the SSL models to the decoder(128 or 257 nodes of the MLP).

The following are brief descriptions of each model.

Wav2Vec 2.0 is the improved version of Wav2Vec [28] with a large amount of data for training and is trained with a masked prediction method (masks the speech input in the latent space) instead of direct contrastive learning. The evaluated large models are trained on 60k hours of data (ll60k) and 60k hours of data plus Common Voice, Switchboard, and Fisher (lv60-cv-swbd-fsh). Both models are trained in the English-only language (monolingual).

XLSR-53 is a multilingual pre-trained model trained on 53 languages using the wav2vec 2.0 framework. The model is trained on three datasets: Common-Voice (CV), BABEL, and Multilingual LibriSpeech (MLS) [19]. The trained languages, multilingual, include English (En), Spanish (es), French (fr), Italian (it), Kyrgyz (ky), Dutch (du), Russian (ru), Swedish (sv), Turkish (tr), Tatar (tt) and Chinese (zh) for CV; Bengali (bn), Cantonese (zh), Georgian (ka), Haitian (ht), Kurmanji (ku), Pashto (ps), Tamil (ta), Turkish (tr), Tokpisin (tp), Vietnamese (vi) for BABEL; and Dutch (du), English (en), French (fr), German (de), Italian (it), Polish (pl), Portuguese (pt), Spanish (es) for MLS. The amount of data is 56k hours.

XLS-R added the number of datasets for training in the previous XLSR-53 with VoxPopuli (23 European languages) and VoxLingual107 (107 languages). Voxpopuli includes : English (en), German (de), Italian (it), French (fr), Spanish (es), Polish (pl), Romanian (ro), Hungarian (hu), Dutch (nl), Czech (cs), Slovenian (sl), Finnish (fi), Croatian (hr), Slovakian (sk). The total amount of data comprises 436K hours of public datasets. The number of languages, multilingual, is 128. Three evaluated variants of large XLS-R models in this research are XLS-R 300M with 317M of parameters, XLS-R 1B with 917M parameters, and XLS-R 2B with 2162M parameters. The total amount of data is 436k hours.

HuBERT is an SSL that improves wav2vec 2.0 by masked prediction of hidden units by combining CNN encoders with Transformers. As in wav2vec 2.0, Hubert Large is trained on the ll60k (60k hours) dataset, which is monolingual English data.

LightHuBERT is the light version of HuBERT Base by creating a large, over-parameterized neural network that contains multiple sub-networks and designing two-stage distillation strategies to leverage the contextualized latent representations from HuBERT. In this evaluation, we evaluated the LightHuBERT model from stage 1. As in HuBERT, it only contains English data (monolingual). The amount of data is 960 hours for pre-training and 10 hours for fine-tuning.

HuBERT Base Robust MGR is an improved version of HuBERT that utilizes domain adversarial training (DAT). The dataset for training is LibriSpeech 960 hours, which is a monolingual English-only dataset. The training includes augmented data by adding noise from the MUSAN dataset, Gaussian noise, and reverberation.

mHuBERT is the multilingual version of HuBERT. The training data is multilingual, including English, Spanish, and French, and contains 4.5k hours of data each (totaling 13.5k hours of data) from subsets of the VoxPopuli dataset.

UniSpeech-SAT is a speaker-aware pre-training model built on the HuBERT framework. The large version is trained on diverse data from LibriSpeech, VoxPopuli, and GigaSpeech datasets. Although the raw datasets include multilingual data, including European languages (VoxPopuli), only English is used in the training stage. Hence, it is a monolingual English

model. The amount of training data is 96k hours.
Data2Vec is a multimodal model that uses the same learning method for speech, text, and image. The core idea is to predict contextualized latent representations of the input data based on a masked view of the input in a self-distillation setup instead of predicting modality-specific targets. Data2Vec Large is trained on LibriLight 60k hours of data (LL60k); hence, it is a monolingual English model.
WavLM is an SSL based on the HuBERT framework by jointly learning masked speech prediction and denoising in pre-training. The Large version is trained on 94k hours of data from LibriLight, VoxPopuli, and GigaSpeech. WavLM Large is a monolingual model for English.

4. Methods

Fig. 2 shows the flow of our experiments. At first, we pre-processed the EmoFilm dataset to match the requirements of S3PRL (downsampling and creation of metadata). Then, we trained 13 SSL models on that dataset with a fixed training/test split. A single-layer MLP with 128 and 256 nodes is evaluated as the classifier. The 256-node MLP (MLP 256) is the default configuration for the CMU-MOSEI dataset in the previous study [2]. Since the EmoFilm dataset is relatively small compared to CMU-MOSEI, we reduce the number of nodes in the MLP to 128 as an alternative architecture. Prior to classification, the frame-level SSL representations $c_l \in \mathbb{R}^{d_c}$ are first projected by a trainable linear connector to an intermediate dimension $H \in \{128, 256\}$:

$$h_l = \mathbf{W}_c\, c_l + \mathbf{b}_c, \quad h_l \in \mathbb{R}^H \tag{5}$$

where $\mathbf{W}_c \in \mathbb{R}^{H \times d_c}$ and $\mathbf{b}_c \in \mathbb{R}^H$. The projected frame-level vectors are then aggregated by temporal mean pooling to form an utterance-level representation:

$$\bar{\mathbf{h}} = \frac{1}{L} \sum_{l=1}^{L} h_l, \quad \bar{\mathbf{h}} \in \mathbb{R}^H \tag{6}$$

Finally, a linear output layer maps the pooled representation to emotion class scores:

$$\hat{\mathbf{y}} = \mathrm{softmax}\big(\mathbf{W}_o\, \bar{\mathbf{h}} + \mathbf{b}_o\big) \tag{7}$$

where $\mathbf{W}_o \in \mathbb{R}^{C \times H}$ and $\mathbf{b}_o \in \mathbb{R}^C$ are trainable parameters, and $C = 5$ is the number of emotion categories. We refer to this architecture as "MLP H" (MLP 128 or MLP 256) where H denotes the connector output dimension. The full pipeline (connector(Eq. 5), mean pooling (Eq. 6), and output layer (Eq. 7)) is trained end-to-end on top of the frozen SSL encoder. The model is trained by minimizing the cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log \hat{y}_c \tag{8}$$

where $y_c \in \{0, 1\}$ is the one-hot ground-truth emotion label. The model was trained with the Adam optimizer and a 0.0001 learning rate, a batch size of two, and for 10000 steps. The five best-performing SSLs in the fixed training split are selected to be evaluated in the cross-validation experiments. The cross-validation experiments are conducted with five-fold cross-validation, with two architectures as in the fixed split evaluations.

We measured the performance in weighted accuracy (WA, overall accuracy) and unweighted accuracy (UA, average recall per emotion category), defined respectively as:

$$\mathrm{WA} = \frac{\sum_{c=1}^{C} \mathrm{TP}_c}{N} \times 100\% \tag{9}$$

$$\mathrm{UA} = \frac{1}{C} \sum_{c=1}^{C} \frac{\mathrm{TP}_c}{N_c} \times 100\% \tag{10}$$

where TP_c is the number of correctly classified samples for emotion class c, N_c is the total number of samples in class c, and $N = \sum_{c=1}^{C} N_c$ is the total number of test samples. Since the dataset and all folds are not balanced in terms of the number of samples per emotion category, the UA is the preferred metric. The reported WA and UA scores are a single run for the fixed training split (seed: 1337) and an average of five folds for CV evaluations.

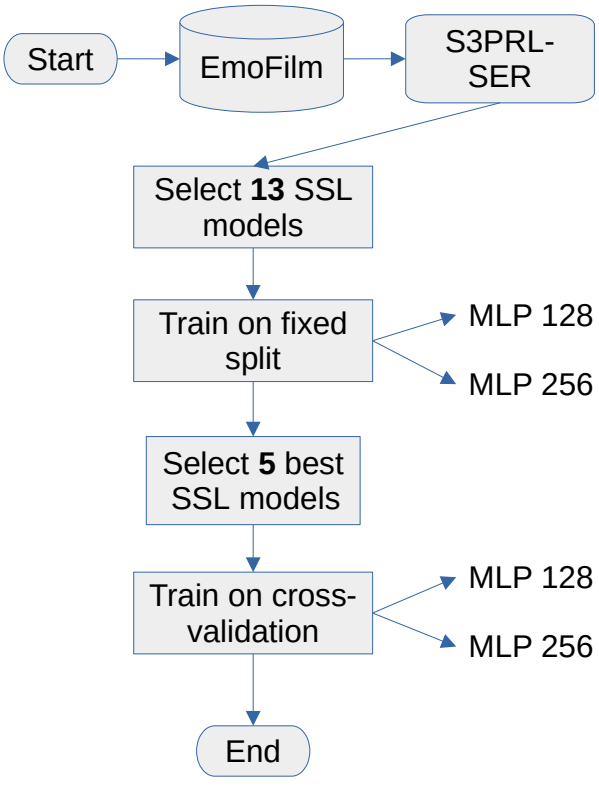

Fig. 2 Flowchart of the experiments

We created a new repository for these experiments and future multilingual SER studies (https://github.com/bagustris/s3prl-ser/). The fixed experiment is located in the `emofilm` directory under `s3prl/downstream/`; the cross-validation experiments are located in the `emofilm-cv` directory. The repository, forked from the original S3PRL toolkit[3], is proposed to standardize the evaluation of the available SER datasets as the downstream tasks with the upstream of the S3PRL toolkit. The S3PRL-SER toolkit is aimed at providing a unified interface to evaluate the performance of SSL models on SER tasks in a simple, consistent, and configurable manner. The SER researchers are encouraged to contribute to the toolkit by adding new SER datasets, new SSL models, or reporting bugs and issues.

5. Results and Discussions

5.1 Fixed split and cross-validation evaluations

We divided our results into two parts according to the experiments: a fixed split (Tables 3 and 4) and cross-validation (Table 5). Although run on the GPU, the results of the fixed split are deterministic for each run, thanks to the initialization of the seed number. For the cross-validation, we reported an average of five folds. The computation time for a single SSL lasts about 2 hours on a single RTX 3090 GPU with i9-10850K or AMD EPYC 7702P CPUs (we run on two different machines). We did not perform hyperparameter optimization due to computation time constraints. Two different nodes (128 and 256) are evaluated based on the initial experiments.

The trends shown in Table 3 by MLP 128 and in Table 4 by MLP 256 are different. The UA scores sort the lists on both tables. In 128 nodes, the worst-performing SSL is mHuBERT, while the best is XLS-R 1 B. In 256 nodes, the worst-performing SSL is Data2Vec Large, while the best is XLS-R 2B. It is difficult to justify which SSL performs better under fixed split evaluation, particularly on small datasets. However, by comparing the two architectures, we obtained that the big five SSLs are consistent among them, i.e., UniSpeech-SAT Large, WavLM Large, XLS-R 300M, XLS-R 1B, and XLS-R 2B. We evaluated these five SSLs in the cross-validation experiments.

Table 5 displays the average performance scores of the cross-validation experiments. The UA score also sorts the list. Although the order of SSLs on the two architectures is different, the top-performing SSL is the same, i.e., XLS-R 2B. This result supports the result of a fixed split with 256 nodes. In the previous fixed split evaluation, it is difficult to judge which one is better between XLS-R 1B and XLS-R 2B. Given the more robust score in CV evaluation, it can be judged that XLS-R 2B generally performs better than XLS-R 1B and other SSLs. A large amount of data for training (two billion) on diverse languages is the main reason for the better performance of XLS-R 2B. This finding proves the benefit of training multilingual datasets for multilingual speech emotion recognition.

The XLS-R variants also performed well on the fixed split evaluations. The top three performing SSLs on both 128 nodes and 256 nodes were achieved by XLS-R 300M, XLS-R 1B, and XLS-R 2B (best). The other two multilingual SSLs, mHuBERT and XLSR-53, gained comparable performances to non-multilingual SSL models. We assumed that this poor performance is due to the amount of training data in both models. Interestingly, the previous study stopped using XLSR-53 only (Wav2Vec2 XLSR) for the multilingual SSL model [2] and found the results are not as competitive as other SSL models like UniSpeech-SAT Large and WavLM Large. Motivated by the success of the use of the multilingual model for multilingual SER [27], we evaluated other XLS-R variants and found that XLS-R 300M, XLS-R 1B, and XLS-R 2B performed well on this multilingual SER task.

In this multilingual SER evaluation, variants of XLS-R performed better than WavLM Large and UniSpeech-SAT Large, which attained state-of-the-art scores on the previous studies for the emotion recognition task [2, 21, 22]. Since the top three performing SSLs in this study are not evaluated in the previous studies, these three XLS-R variants (300M, 1B, and 2 B) may also perform well on monolingual SER tasks. In future studies, it is worth evaluating these top-performing SSLs in common monolingual datasets like IEMOCAP, MSP-IMPROV, MSP-Podcasts, and JTES.

5.2 Performance of best SSL models on each target language

We used results from the big five models with 128 nodes to analyze the effect of the multilingual model on each target language. Our cross-language analysis (Table 6 and Fig. 3 top) reveals significant language-specific advantages among SSL encoders for emotion recognition. We choose Wav2Vec2 Large #1 (128 nodes) as a baseline since the XLS-Rs are variants of this model with much larger parameters and cross-lingual focus. Although the XLS-R 2B model obtained the highest UA on multilingual evaluation (Table 5), individual language evaluation showed that UniSpeech-SAT, XLS-R 300M, and 1B achieved the highest score for IT, EN, and SP. The improvement heatmap demonstrates that XLS-R models exhibit strong English bias, with XLS-R 300M achieving +27.9% unweighted accuracy improvement over the

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

Table 3 Performance scores of the evaluated SSLs on the fixed split with 128 neurons MLP (sorted by %UA)

SSL model	%WA	%UA
mHuBERT	70.72	67.16
data2vec Large	72.93	72.51
XLSR-53	78.45	72.64
Wav2Vec 2.0 Large #1	78.45	73.21
Wav2Vec 2.0 Large #2	78.45	73.59
HuBERT Base Robust MGR	75.14	74.46
HuBERT Large	77.35	75.54
LightHuBERT Stage1	77.90	76.72
UniSpeech-SAT Large	81.22	79.52
WavLM Large	83.98	82.72
XLS-R 2B	84.53	83.42
XLS-R 300M	85.64	85.57
XLS-R 1B	85.64	86.24

Table 4 Performance scores of the evaluated SSLs on the fixed split with 256 neurons MLP (sorted by %UA)

SSL model	%WA	%UA
data2vec Large	74.03	72.42
XLSR-53	79.01	73.89
HuBERT Base Robust MGR	74.59	74.63
Wav2Vec 2.0 Large #1	80.66	75.36
mHuBERT	78.45	75.54
LightHuBERT Stage1	77.35	75.92
HuBERT Large	77.90	77.99
Wav2Vec 2.0 Large #2	78.45	79.59
UniSpeech-SAT Large	83.43	80.30
WavLM Large	83.43	81.46
XLS-R 300M	85.64	83.11
XLS-R 1B	84.53	83.80
XLS-R 2B	85.64	84.16

Table 5 Average performance on five-fold cross-validation with a single-layer MLP (sorted by %UA)

SSL Model	%WA	%UA
128 nodes		
UniSpeech-SAT Large	86.10	85.29
XLS-R 300M	86.10	85.59
WavLM Large	86.55	86.62
XLS-R 1B	87.00	87.17
XLS-R 2B	89.24	**88.95**
256 nodes		
UniSpeech-SAT Large	82.42	82.15
WavLM Large	82.06	82.52
XLS-R 300M	87.44	87.11
XLS-R 1B	87.44	87.56
XLS-R 2B	87.53	87.61

wav2vec2 baseline for English while showing minimal gains for Italian (+5.5%) and negative performance for Spanish (-6.1%). Conversely, UniSpeech-SAT Large demonstrates the most consistent cross-language performance, ranking first for Italian (+9.5%), third for Spanish (-0.7%), and fifth for English (+19.4%). Overall, English has the most benefit from multilingual SSL models (XLS-R 300M, 1B, and 2B), with an average improvement of +25.6% across all three models, while Spanish has disadvantages from multilingual SSL models, with an average decrease of -2.3% across all three models (perhaps due to high score on baseline Wav2Vec2, see Table 6).

Table 6 Unweighted accuracy (UA %) performance of SSL models across languages on 5-fold CV-128 nodes

SSL Model	EN	IT	SP
Wav2Vec 2.0 Large #1	60.51	74.79	85.56
UniSpeech-SAT Large	79.93	**84.33**	84.86
WavLM Large	84.25	80.49	84.50
XLS-R 300M	**88.44**	80.31	79.50
XLS-R 1B	82.03	82.47	**86.14**
XLS-R 2B	88.01	82.31	84.96

The performance-consistency analysis reveals (Fig. 3 bottom) distinct trade-offs between average improvement and cross-language stability among SSL encoders. UniSpeech-SAT Large exhibits the highest consistency ($\sigma = 0.082$) with moderate average improvement (+9.4%), making it optimal for multilingual deployment scenarios. In contrast, XLS-R 2B achieves the highest average performance (+11.5%) but with lower consistency ($\sigma = 0.118$), indicating substantial language-dependent variance. Notably, Spanish emotion recognition shows resistance to SSL improvements across all models, with the wav2vec2 baseline achieving 85.6% UA that most SSL encoders fail to surpass. This suggests that the strong monolingual Spanish baseline creates a performance ceiling that current multilingual SSL approaches struggle to overcome, highlighting the importance of language-specific SSL encoder selection rather than adopting universal multilingual models for cross-language emotion recognition tasks.

5.3 Performance of best SSL models on each emotion category

Fig. 4 provides a systematic evaluation of state-of-the-art self-supervised learning models across discrete emotion categories for each language (results are also averages from five folds with 128 nodes). The analysis reveals substantial heterogeneity in model performance across emotional dimensions, with XLS-R 2B emerging as the optimal architecture for three of the five evaluated emotions (Anger, Happy, and Neutral),

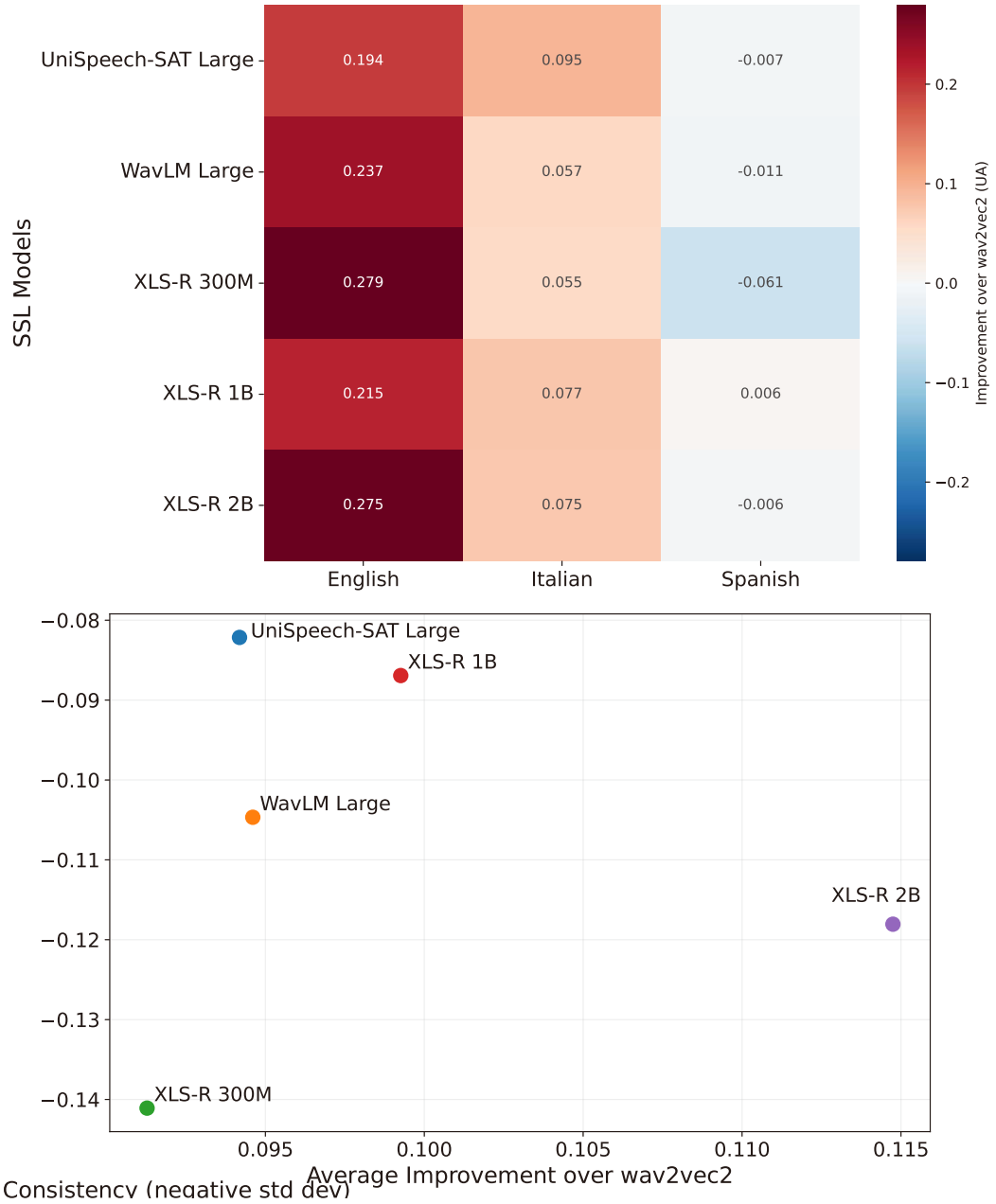

Fig. 3 Performance of best SSL models on each language (top) and their consistency (bottom): Scores are averages from five folds. *Top*: Each cell in the heatmap shows the UA improvement of a given SSL model over the Wav2Vec 2.0 Large #1 baseline, computed as $\Delta\mathrm{UA}_{\mathrm{lang}} = \mathrm{UA}_{\mathrm{lang}}^{\mathrm{model}} - \mathrm{UA}_{\mathrm{lang}}^{\mathrm{baseline}}$ (decimal fraction, e.g. 0.279 $\equiv$ +27.9%) and averaged across five folds with the 128-node MLP; positive values (warm/red colors) indicate gains and negative values (cool/blue colors) indicate degradation relative to the baseline. *Bottom*: Each point represents one SSL model, with the x-axis showing the average ΔUA across all three languages (EN, IT, SP) and the y-axis showing the *negative* standard deviation $-\sigma$ of ΔUA across languages (i.e., higher values indicate more consistent cross-language behavior); models closer to the **upper-right** are preferred, as they combine great average improvement with stable cross-language performance.

achieving particularly pronounced improvements in anger detection (overall +55.8% in all languages) and moderate gains in neutral state recognition (+13.9%). WavLM Large demonstrates superior performance for contempt recognition with a modest but meaningful improvement of +6.9% (all emotions), while UniSpeech-SAT Large exhibits optimal performance for sadness detection, albeit with a marginal decrease of -1.0% compared to baseline methods. The substantial performance variance across emotions, particularly the exceptional gains in anger recognition, indicates that SSL models may be more adept at capturing the acoustic correlates of high-arousal emotional states compared to low-arousal or complex emotional expressions. In contrast, all SSL models struggle with sad emotion (low arousal) recognition.

Compared to the previous study on multilingual SER using SSLs [10], we found that bigger XLSR variants (XLS-R 1B, XLS-R 2B) performed better than smaller models (XLS-R 53, XLS-R 300M). The previous study found that XLSR-53 performed best among the evaluated 12 models. Their top-four models are XLSR-53, XLS-R 1B, XLS-R 300M, and WavLM Large; while our findings suggest that the size of the model linearly correlates with performance improvements in multilingual SER tasks. Hence, our top-four performing models are XLS-R 2B, XLS-R 1B, XLS-R 300M, and WavLM Large.

6. Conclusions

In this study, we evaluated 13 SSL models, including those trained on multilingual data, for multilingual speech emotion recognition tasks. The study is mainly designed to observe the effect of training multilingual (speech) data, e.g., variants of XLS-R, on multilingual SER. *The results show that SSL models trained in diverse languages generally performed better than SSL models trained on monolingual data for the multilingual SER task.* In addition to the required diverse languages, we assumed the matter of the amount of training data. The larger data from monolingual training (wav2vec 2.0) did not improve the performance, but the larger data from multilingual did. Three variants of XLS-R with 300M, 1B, and 2B of training data outperformed other SSLs, including SSLs that previously attained state-of-the-art performance on the monolingual SER task. Evaluation of the fixed split partition and cross-validation showed that XLS-R 2B might perform best on this multilingual task. We achieved a high average score of 88.95% unweighted accuracy for five-fold cross-validation with a single-layer XLS-R 2B. Future studies could be directed to evaluate these top-performing SSL models on other multilingual SER datasets and to observe whether these models also perform well on monolingual SER tasks.

Acknowledgment

This paper is partly based on results obtained from projects JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan, and JSPS Kakenhi 24K02967. The authors would like to thank the creators of the EmoFilm dataset [15] for sharing their dataset.

References

[1] B. W. Schuller: Speech emotion recognition: Two decades in a nutshell, Commun. ACM, Vol. 61, No. 5, 2018, ISSN 0001-0782.

[2] B. T. Atmaja and A. Sasou: Evaluating self-supervised speech representations for speech emotion recognition, IEEE Access, Vol. 10, pp. 124396–124407, 2022, ISSN 21693536.

[3] N. Scheidwasser-Clow, M. Kegler, P. Beckmann and M. Cernak: SERAB: A multi-lingual benchmark for speech emotion recognition, ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 7697–7701, IEEE, 2022, ISBN 978-1-6654-0540-9.

[4] S. Haq and P. J. B. Jackson: Speaker-dependent audio-visual emotion recognition, Int'l Conf. Audit. Speech Process., pp. 53–58, 2009.

[5] K. R. Scherer: A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology, 6th Int. Conf. Spok. Lang. Process. ICSLP 2000, , No. Icslp, 2000.

[6] K. R. Scherer, R. Banse and H. G. Wallbott: Emotion inferences from vocal expression correlate across languages and cultures, J. Cross. Cult. Psychol., Vol. 32, No. 1, pp. 76–92, 2001, ISSN 00220221.

[7] H. A. Elfenbein and N. Ambady: On the universality and cultural specificity of emotion recognition: A meta-analysis, Psychol. Bull., Vol. 128, No. 2, pp. 203–235, 2002, ISSN 00332909.

[8] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau and M. Auli: XLS-R: Self-supervised cross-lingual speech representation learning at scale, Interspeech 2022, Vol. 2022-Septe, pp. 2278–2282, ISCA, ISCA, 2022, ISSN 19909772.

[9] S. Latif, J. Qadir and M. Bilal: Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition, 2019 8th Int. Conf. Affect. Comput. Intell. Interact. ACII 2019, 2019, ISBN 9781728138886.

[10] F. Luthman: Multilingual Speech Emotion Recognition Using Pretrained Models Powered by Self-Supervised Learning, Ph.D. Thesis, KTH Royal Institute of Technology, 2022.

[11] V. Scotti, F. Galati, L. Sbattella and R. Tedesco: Combining deep and unsupervised features for multilingual speech emotion recognition, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Vol. 12662 LNCS, pp. 114–128, Springer Sci-

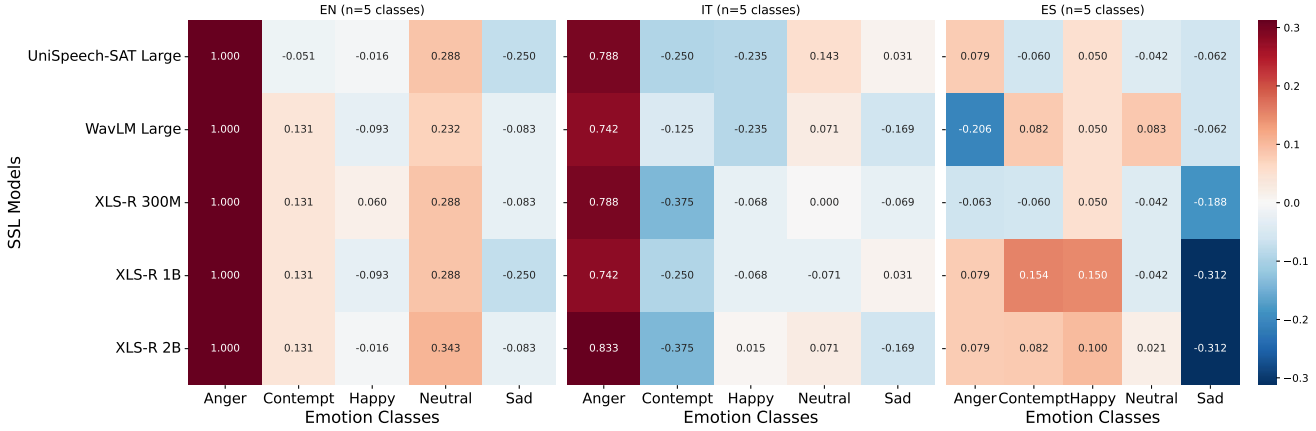

Fig. 4 UA improvement of best SSL models on each emotion category across languages: The figure consists of three heatmap panels, one per language (EN, IT, SP); each cell shows the per-emotion UA improvement of a given SSL model over Wav2Vec 2.0 Large #1, computed as $\Delta\mathrm{UA}_{e,\mathrm{lang}} = \mathrm{UA}^{\mathrm{model}}_{e,\mathrm{lang}} - \mathrm{UA}^{\mathrm{baseline}}_{e,\mathrm{lang}}$ (decimal fraction, rows = SSL models, columns = emotion categories, max=1, min=-1), and averaged across five folds with the 128-node MLP. Warm/red cells indicate that the SSL model correctly classified a higher proportion of samples for that emotion than the baseline, while cool/blue cells indicate a drop; thus, the figure reveals for which emotion categories and languages each SSL model offers the greatest recognition benefit.

ence and Business Media Deutschland GmbH, 2021, ISSN 16113349.

[12] A. Keesing, Y. S. Koh and M. Witbrock: Acoustic features and neural representations for categorical emotion recognition from speech, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, Vol. 1, pp. 521–525, ISCA, ISCA, 2021, ISBN 9781713836902, ISSN 19909772.

[13] M. Gerczuk, S. Amiriparian, S. Ottl and B. W. Schuller: EmoNet: A transfer learning framework for multi-corpus speech emotion recognition, IEEE Trans. Affect. Comput., 2021, ISSN 19493045.

[14] A. Keesing, I. Watson and M. Witbrock: Convolutional and recurrent neural networks for spoken emotion recognition, Proc. 18th Annu. Work. Australas. Lang. Technol. Assoc., pp. 104–109, 2020.

[15] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird and B. Schuller: Categorical vs dimensional perception of Italian emotional speech, Interspeech 2018, pp. 3638–3642, ISCA, ISCA, 2018, ISSN 19909772.

[16] R. C. Moore, D. P. W. Ellis, E. Fonseca, S. Hershey, A. Jansen and M. Plakal: Dataset balancing can hurt model performance, ICASSP 2023 - 2023 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 1–5, IEEE, 2023, ISBN 978-1-7281-6327-7.

[17] 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, Interspeech 2021, pp. 1194–1198, ISCA, ISCA, 2021.

[18] A. Baevski, H. Zhou, A. Mohamed and M. Auli: wav2vec 2.0: A framework for self-supervised learning of speech representations, Adv. Neural Inf. Process. Syst., 2020, ISSN 10495258.

[19] A. Conneau, A. Baevski, R. Collobert, A. Mohamed and M. Auli: Unsupervised cross-lingual representation learning for speech recognition, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, Vol. 1, pp. 346–350, 2021, ISBN 9781713836902, ISSN 19909772.

[20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov and A. Mohamed: HuBERT: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Trans. Audio, Speech, Lang. Process., Vol. 29, pp. 3451–3460, 2021, ISSN 2329-9290.

[21] S. Chen, C. Wang, Z. Z. Chen, Y. Wu, S. Liu, Z. Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Y. Qian, Y. Y. Qian, J. Wu, M. Zeng, X. Yu and F. Wei: WavLM: Large-scale self-supervised pre-training for full stack speech processing, IEEE J. Sel. Top. Signal Process., Vol. 16, No. 6, pp. 1505–1518, 2021, ISSN 19410484.

[22] 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, ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 6152–6156, IEEE, 2022, ISBN 978-1-6654-0540-9.

[23] A. Lee, H. Gong, P. A. Duquenne, H. Schwenk, P. J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu and W. N. Hsu: Textless speech-to-speech translation on real data, NAACL 2022 - 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf., pp. 860–872, 2022, ISBN 9781955917711.

[24] A. Baevski, W.-n. Hsu, Q. Xu, A. Babu, J. Gu and M. Auli: data2vec: A general framework for self-supervised learning in speech, vision and language, Proc. 39th Int. Conf. Mach. Learn., 2022.

[25] R. Wang, Q. Bai, J. Ao, L. Zhou, Z. Xiong, Z. Wei,

Y. Zhang, T. Ko and H. Li: LightHuBERT: Lightweight and configurable speech representation learning with once-for-all hidden-unit BERT, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, Vol. 2022-Septe, No. September, pp. 1686–1690, 2022, ISSN 19909772.

[26] K. P. Huang, Y. K. Fu, Y. Zhang and H. Y. Lee: Improving distortion robustness of self-supervised speech processing tasks with domain adaptation, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, Vol. 2022-Septe, pp. 2193–2197, 2022, ISSN 19909772.

[27] 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, APSIPA Annu. Summit Conf., December, pp. 750–755, 2021, ISBN 9789881476890.

[28] S. Schneider, A. Baevski, R. Collobert and M. Auli: wav2vec: Unsupervised pre-training for speech recognition, Interspeech 2019, pp. 3465–3469, ISCA, 2019, ISSN 19909772.

Bagus Tris Atmaja received degrees of bachelor's and master's of engineering physics from the Sepuluh Nopember Institute of Technology in 2009 and 2012, and a Ph.D. from the Japan Advanced Institute of Science and Technology, Japan. He is now an assistant professor in the Division of Information Science at the Nara Institute of Science and Technology, Japan.

Akira Sasou 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 Department of Information Technology and Human Factor, Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Japan.

(Received January 15, 2026; revised April 16, 2026)