Multi-label Emotion Share Regression From Speech Using Pre-Trained Self-Supervised Learning Models

Bagus Tris Atmaja
AIST
Tsukuba, Japan
https://orcid.org/0000-0003-1560-2824

Akira Sasou
AIST
Tsukuba, Japan
https://orcid.org/0000-0003-1700-0325

***Abstract*—Speech emotion recognition (SER) has been enabling natural human-machine interaction. Previous SER only focused on predicting a single emotion from humans' voices, whereas emotions from voices might have multiple categories. This study aims to predict emotion share from a voice, i.e., the proportion of the different emotion categories expressed in the speech. We evaluated eight pre-trained self-supervised learning (SSL) models from major companies for multilingual emotion share recognition. The models were evaluated on the Hume-Prosody dataset, which contains 51,881 samples of nine different emotion categories. We observed the important factor of wider window size for multilingual emotion share recognition in contrast to the narrow window size used (around 5.5 s). The most effective evaluated pre-trained model for this case is the large version of HuBERT, which achieved a Spearman correlation score of 0.579, showing the feasibility of emotion share recognition using SSL.**

***Index Terms*—Speech emotion recognition, multilingualilty, emotion share, multi-label regression, pre-trained models**

I. Introduction

Speech emotion recognition (SER) has been actively researched in the past two decades [1]. Some advancements in this area include the development of data-driven pre-trained models, multilingual speech emotion recognition, multimodal emotion recognition, and multitask learning. Humans can perceive emotion from voice even though other modalities are not perceived; computers should be able to do so.

Evaluating pre-trained self-supervised learning models appears to be effective in recognizing emotions from vocal bursts [2], yet its implementation for multilingual emotion share recognition has not been confirmed. Vocal bursts may be effective in indicating the appearance of specific discreet emotions; however, their existence is less frequent than speech. Hence, evaluating speech (in addition to vocal bursts) is necessary to gain insights into the effectiveness of recent models for emotion recognition.

Instead of a single emotion, an utterance may consist of several emotions. For example, a speaker may express 'Anger' and 'Sadness' at the same time. This emotion share by nature applies to multiple cultures. By this assumption, it is plausible that a single model of emotion share recognition may work for multiple languages.

This paper is based on results obtained from NEDO project JPNP20006 and JSPS KAKENHI 24K0296.

This study contributes two-fold. First, we attempted to recognize more than one emotion using pre-trained models and measure its feasibility using the Spearman correlation coefficient. Second, we experimented with a broader range of window sizes from 32 ms to 6 s and found that longer window context help in improving the performance score of multilingual emotion share than shorter window size.

II. Research Methods

A. Dataset

The dataset evaluated in this study comes from the ACM Multimedia 2023 COMputational PARalinguistics challengE (ComParE) Emotion Share Sub-Challenge [3]. The data was provided by Hume AI, namely the Hume-Prosody dataset, for addressing the multi-label regression task. The dataset contains more than 5,000 'seed' samples. Seeds consist of various emotional expressions that were gathered from openly available datasets, including VENEC and MELD. There are 51,881 'mimic' samples from 1,004 speakers aged from 20 to 66 years old. The total duration is 41:48:55 h of data, a mean 2.9 s, and a range of 1.2 - 7.98 s. The recordings were collected multilingually in three countries: the United States (English), South Africa (English), and Venezuela (Spanish).

The dataset contains nine categorial emotions that have been selected from the original 48 emotions because of their more balanced distribution across the valence-arousal space [4], i.e., 'Anger', 'Boredom', 'Calmness', 'Concentration', 'Determination', 'Excitement', 'Interest', 'Sadness', and 'Tiredness'. For each emotion, a proportion or 'share' ranging from 1 to 100 has been assigned to the nine emotions based on the raters that rated the emotion for 'seed' samples. The audio files were also normalized to -3dB and resampled to 16 kHz (from raw audio recorded at 48 kHz), 16-bit, mono format.

B. SER vs. Emotion Share

While the previous SER research attempts to recognize emotions from speech by using categorical labels or dimensional scores, this study uses emotion share as the target. The emotion share is a proportion of the emotion categories expressed in the speech. For example, if a speech contains 50% of 'Anger' and 50% of 'Sadness', then the emotion share for 'Anger' is 0.5 and for 'Sadness' is 0.5. The emotion share

in the Hume-Prosody dataset is a continuous value between 0 and 1 (normalized) of nine different emotion categories based on the rating given by multiple raters. The emotion share is more realistic than categorical labels because it is possible that a speech contains more than one emotion category.

III. Pre-trained models as feature extractors

We evaluated nine pre-trained models from three companies for acoustic feature extractors: Microsoft, Facebook, and Audeering. Please note that for Audeering, the actual name for their model is 'wav2vec2-large-robust-12-ft-emotion-msp-dim'. However, we use `wav2vec2-audeering` for simplicity. These models can be obtained from the Hugging Face web page with a URL 'https://huggingface.co/[company]/[model-name]'. The list of pre-trained models is shown in Table I; for accessing the pre-trained models, replace the `company` and `model-name` in the previous URL with the ones provided in the table.

TABLE I
List of pre-trained models and their size

Company	Model name	Size
Facebook	wav2vec2-base [5]	768
	hubert-base-ls960 [6]	768
	hubert-large-ll60k [6]	1024
	wav2vec2-xls-r-300m [7]	1024
	wav2vec2-xls-r-1b [7]	1280
Microsoft	wavlm-base-plus [8]	768
	wavlm-large [8]	1024
Audeering	wav2vec2-audeering [9]	1024

A. Classifier

We adopted a fully connected network as a classifier with two hidden layers. The first layer outputs 256 nodes; the second layer outputs nine nodes (number of categorial emotions). We trained the classifier using the CCC loss function. Hyperparameters used in the classifier are specified in Table II. These parameters are selected based on experiments with different values (e.g., for optimizers are SGD, RMSPROP, Adam, and AdamW).

TABLE II
Hyperparamters used in the classifier

Parameter	Value
# layers	2
# nodes per layer	256, 9
activation function	GeLU
optimizer	AdamW
learning rate	0.001
weight decay	0.001
loss function	CCC loss
window size	32 ms - 6 s
pooling	attention
batch size	8

B. Evaluation Metrics

The main evaluation metric to judge the performance of the model is the Spearman rank coefficient following the previous papers [3]. This metric is similar to the Pearson correlation coefficient but for the rank of the data. Spearman correlation coefficient (SCC) is formulated as follows:

$$SCC = \frac{\text{cov}(\text{R}(X), \text{R}(Y))}{\sigma_{\text{R}(X)}\sigma_{\text{R}(Y)}}. \quad (1)$$

As the second metric, we calculated the CCC score since the loss is based on this metric (CCC loss) and this metric is also a common score in regression-based speech emotion recognition [10], [11]. The CCC is formulated as follows:

$$CCC = \frac{2\rho\sigma_{XY}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2}. \quad (2)$$

The ρ denotes Pearson correlation between X and Y, cov() denotes the covariance of the data, σ^2 denotes variances, and σ denotes the standard deviation of the data. $\text{R}(X)$ and $\text{R}(Y)$ denote the ranked for X and Y, respectively. X and Y denote the predicted and ground truth values, respectively. μ denotes the mean of the data.

The average of SCC and CCC (Avg) is used as the metric to evaluate the performance of the model in different evaluations (e.g., across different pre-trained models) at the development stage for obtaining test scores,

$$Avg = \frac{SCC + CCC}{2}. \quad (3)$$

We also monitored (but not reported) the Pearson correlation coefficient, which shows the same pattern as the Spearman rank coefficient on different features and window sizes. The Pearson correlation is suitable for unranked data, while the Spearman rank coefficient is suitable for ranked data.

We created a repository to reproduce the results of this paper, which is available at https://github.com/bagustris/compare2023_share. Note that the dataset is not included in this repository and must be obtained from the organizer of the ComParE 2023 [3].

IV. Results and Discussion

We organized the results of experiments into three parts: (1) the effect of different pre-trained models, (2) the effect of window size, and (3) test results. The first two parts are evaluated on the development set, while the last part is evaluated on the (hidden) test set.

A. Effect of Different Pre-trained Models

The main question of this research is to find the most robust and suitable pre-trained model for multilingual emotion share recognition. At first, we evaluated 12 pre-trained models, including the recent ones, but we found that some were unsuitable for our task due to very low scores and memory

limitations. Aside from models listed in Table I, we also investigated Unispeech-SAT variants (Base+ and Large) [12], which resulted in low SCC and CCC scores. The scores for UniSpeech-SAT Large are 0.2179 and 0.2583 for SCC and CCC. This low score may show that UniSpeech-SAT cannot extract well affective information across languages since it was designed to disentangled information from different speakers. As for the `wav2vec2-xlsr-2b`, we cannot use it due to memory limitations (we used an RTX3090 with 24GB of VRAM). This feature could be explored in the future.

Table III shows SCC and CCC scores of eight evaluated pre-trained models on the development set of ComParE 2023 Emotion Share Sub-Challenge. All pre-trained models gained comparable results (0.50 - 0.55), showing the models' competitiveness. The HuBERT Large gained the highest average score of SCC and CCC (0.5491), followed by `wav2vec2-audeering` (0.5390), and `wav2vec2-xls-r-300m` (0.5313). The lowest average score is `wav2vec2-base` (0.5119). In this regard, the highest performance of the HuBERT Large is consistent with the previous study [13], although we also evaluated the newer models (WavLM and XLSR).

B. Effect of The Window Size

Windowing is a common technique in signal processing to divide a signal into smaller segments and multiply these segments by a certain function, e.g, rectangular window or Hanning window. The window size is the length of the window in seconds. The size of window size depends on the task since the information to be extract lies in these segments. For instance, the common window size for automatic speech recognition is in the range 20 to 30 milliseconds (ms), e.g., 25 ms [14].

The window size in speech emotion recognition is an important parameter. While the previous studies focus on small (50 - 500 ms) [15] and medium size of window (2 s) [16], in this study, we evaluated the effect of small ($<$ 1 s) medium to large window sizes (1 - 6 s) on the performance of the model. Audio files less than 6 s are padded with zeros. We used the best pre-trained model from the previous experiment (HuBERT Large) and varied the window size from 32 ms to 6 seconds. The results are shown in Table IV. We exclude the results from window sizes less than 1 s since they are not comparable to large window sizes (SCC $\leq$ 0.15). The best average score is 0.5509 (window size 5.5 s), followed by 0.5491 (window size 4 s and 6 s), and 0.5489 (window size 4.5 s). The lowest average score is 0.3740 (window size 1 s). It should be noted that while the common speech processing technique uses a small size of the window (20 $\sim$ 200 ms), the experimental results show that the larger window size is more suitable for emotion share recognition. This phenomenon could be explained by the fact that multi-emotion lies in a longer time span. In other words, it is more effective to observe multi-emotion information in a longer context window. Future studies should investigate this effect of window size on different datasets to seek its generalization.

TABLE III
EFFECT OF DIFFERENT PRE-TRAINED MODELS ON THE PERFORMANCE OF MULTILINGUAL EMOTION SHARE ON THE DEVELOPMENT SET; SCC: SPEARMAN CORRELATION COEFFICIENT, CCC: CONCORDANCE CORRELATION COEFFICIENT

Pre-trained model	SCC	CCC	Avg.
wav2vec2-base	0.5068	0.5170	0.5119
wavlm-base-plus	0.5023	0.5115	0.5069
hubert-base-ls960	0.5129	0.5231	0.5180
wavlm-large	0.5129	0.5059	0.5094
hubert-large-ll60k	0.5530	0.5453	**0.5491**
wav2vec2-audeering	0.5415	0.5365	0.5390
wav2vec2-xls-r-300m	0.5288	0.5339	0.5313
wav2vec2-xls-r-1b	0.5206	0.5263	0.5234

TABLE IV
PERFORMANCE ON DIFFERENT WINDOW SIZES ON THE DEVELOPMENT SET USING HUBERT LARGE MODEL; RESULTS FROM WINDOW SIZES $<$ 1 S ARE NOT SHOWN SINCE THE SCORES ARE VERY SMALL.

Window size	SCC	CCC	Avg.
1	0.3818	0.3662	0.3740
1.5	0.4839	0.4794	0.4816
2	0.5297	0.5105	0.5201
2.5	0.5447	0.5269	0.5358
3	0.5496	0.5372	0.5434
3.5	0.5455	0.5398	0.5426
4	0.5557	0.5445	0.5501
4.5	0.5553	0.5426	0.5489
5	0.5545	0.5436	0.5491
5.5	0.5557	0.5462	**0.5509**
6	0.5503	0.5480	0.5491

V. TEST RESULTS

The previous results reported in this paper are based on the development set, which is openly available in the dataset provided by the organizer of the ComParE 2023 Emotion Share Sub-challenge. In this section, we report the results of the test set, in which the predictions of that set are uploaded to a dedicated web page to get the performance score (Table V). The best pre-trained model (HuBERT Large) in the dev set is used for this experiment. The results are shown in Table V. The average score is 0.5790, which is higher than the average score on the development set (0.5491). This result shows that the model is robust and can generalize to unseen data. However, due to the time limitation of the available web page to obtain the test score, we only submitted two times with two different architectures; another submission was with ensemble learning [17], which got lower than the model reported in this paper but is still higher than the baseline.

Table VI shows a benchmark of different models, including the baselines (first five rows). Note that other approaches incorporate support vector regression (SVR) as the classifier with different acoustic embeddings, while this study used multilayer perceptron (MLP). These results suggest stronger

TABLE V
TES SCORES (SPEARMAN RANK CORRELATION COEFFICIENT, SCC) OF THE BEST DEVELOPMENT MODEL (HUBERT LARGE)

Emotion	SCC
Anger	0.5179
Boredom	0.6309
Calmness	0.6284
Concentration	0.5937
Determination	0.5916
Excitement	0.5441
Interest	0.5277
Sadness	0.5681
Tiredness	0.6090
average	0.5790

TABLE VI
SCC SCORES OF DIFFERENT APPROACHES

#	Approach	Dev	Test
1	wav2vec2 [9]	0.500	0.514
2	auDeep [18]	0.347	0.357
3	DeepSpectrum [19]	0.335	0.331
4	ComParE [20], [21]	0.359	0.365
5	Late Fusion of 1-4 [3]	0.470	0.476
6	Fusion of 9 SVRs [17]	0.524	0.537
7	HuBERT Large + MLP (ours)	**0.549**	**0.579**

monotonic relationships between the processed speech (test) and the reference data (development).

As in the previous studies [13], HuBERT Large obtained the highest performance in our study. In contrast to [22], in which WavLM Large attained higher score than HuBERT Large, we anticipate that this is due to different task (multi-label SER vs. single-label SER), which needs to be confirmed in future studies.

VI. CONCLUSIONS

In this paper, we evaluated eight pre-trained self-supervised learning models from major companies for multilingual emotion share recognition from speech. Emotions, including those expressed in speech, are believed to be universally distinguishable across cultures. However, multi-emotion categories could be expressed in human voices instead of a single emotion. Multilingual share emotion recognition in this study is an attempt to study this phenomenon via voice dataset gathered from the US, South Africa, and Venezuela. Cross-evaluations among different pre-trained models revealed the robustness of the HuBERT Large model. Evaluations on different window sizes showed that *multilingual emotion share could effectively be observed on larger windows* ($\sim$ **5.5 s**) than on shorter windows. This phenomenon is contrast with MFCC processing, in which shorter window size ($\sim$ 10-30 ms) usually perform better than larger windows. On submitting the test results, we obtained improvements from 0.514 to 0.579 in SCC scores. Exploration of the larger size of multilingual pre-trained models could improve the current results further since memory limitation issues blocked us from using the larger model in this study.

REFERENCES

[1] B. W. Schuller, "Speech Emotion Recognition two decades in a Nutshell," *Commun. Acm*, vol. 61, no. 5, 2018.

[2] B. T. Atmaja and A. Sasou, "Evaluating Variants of wav2vec 2.0 on Affective Vocal Burst Tasks," in *ICASSP 2023*. IEEE, jun 2023.

[3] B. Schuller *et al.*, "The ACM Multimedia 2023 Computational Paralinguistics Challenge: Emotion Share and Requests," in *Proc. 30th ACM Int. Conf. Multimed.*, 2023, pp. 7120–7124.

[4] A. S. Cowen and D. Keltner, "Semantic Space Theory: A Computational Approach to Emotion," *Trends Cogn. Sci.*, vol. 25, no. 2, pp. 124–136, 2021.

[5] 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.

[6] 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, no. Cv, pp. 3451–3460, 2021.

[7] A. Babu *et al.*, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech 2022*, vol. 2022-Septe. ISCA: ISCA, sep 2022, pp. 2278–2282.

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

[9] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–13, mar 2023.

[10] B. T. Atmaja and M. Akagi, "Improving Valence Prediction in Dimensional Speech Emotion Recognition Using Linguistic Information," in *Proc. 2020 23rd Conf. O-COCOSDA*. IEEE, nov 2020, pp. 166–171.

[11] B. T. Atmaja, A. Sasou, and M. Akagi, "Speech Emotion and Naturalness Recognitions With Multitask and Single-Task Learnings," *IEEE Access*, vol. 10, pp. 72 381–72 387, 2022.

[12] 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 *ICASSP 2022*. IEEE, may 2022, pp. 6152–6156.

[13] S.-w. Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*, ISCA, pp. 1194–1198.

[14] Abdel-Hamid *et al.*, "Convolutional neural networks for speech recognition," *IEEE TALSP*, vol. 22, no. 10, pp. 1533–1545, 2014.

[15] M. Pereira, S. Chapaneri, and D. Jayaswal, "Analysis of windowing techniques for speech emotion recognition," in *2016 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2016*. Institute of Electrical and Electronics Engineers Inc., jul 2016.

[16] J. Shor *et al.*, "Universal paralinguistic speech representations using self-supervised conformers," in *ICASSP 2022*, 2022, pp. 3169–3173.

[17] B. T. Atmaja and A. Sasou, "Ensembling Multilingual Pre-Trained Models for Predicting Multi-Label Regression Emotion Share from Speech," in *2023 Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*. IEEE, oct 2023, pp. 1026–1029.

[18] M. Freitag *et al.*, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *J. Mach. Learn. Res.*, vol. 18, pp. 1–5, 2018.

[19] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore Sound Classification Using Image-Based Deep Spectrum Features," in *Interspeech 2017*. ISCA: ISCA, aug 2017, pp. 3512–3516.

[20] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimed. - MM '13*. New York, New York, USA: ACM Press, 2013, pp. 835–838.

[21] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Interspeech 2013*, no. August. ISCA: ISCA, aug 2013, pp. 148–152.

[22] B. T. Atmaja and A. Sasou, "Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.