Multilingual, Cross-lingual, and Monolingual Speech Emotion Recognition on EmoFilm Dataset

Bagus Tris Atmaja* and Akira Sasou*
* National Institute of Advanced Industrial Science and Technology, Japan
E-mail: b-atmaja@aist.go.jp, a-sasou@aist.go.jp

***Abstract*—Research on speech emotion recognition has been actively conducted; most are in monolingual settings. Considering that emotion expressed in speech is universal, it is noteworthy to conduct multilingual emotion recognition across different cultures. This paper contributes to evaluating multilingual, cross-lingual, and monolingual automatic speech emotion recognition (SER) on the same EmoFilm dataset. We first evaluated these three scenarios on a fixed training/test split. For multilingual emotion recognition, we then expanded the evaluation with cross-validation. The results show that the multilingual SER gained the highest performance with 74.86% of balanced accuracy for five categorical emotions, followed by 72.17% and 58.03% for monolingual and cross-lingual evaluations. We reduced the number of training samples to observe its impact and found that the monolingual setting is superior among others on the same number of samples. The results of this study could suggest the potential use of multilingual SER over cross-lingual and monolingual SER in future speech technologies.**

***Index Terms*—speech emotion recognition, pre-trained model, multilingual emotion recognition, EmoFilm**

I. Introduction

There is some universality in emotion perception by humans, particularly from vocal cues. A study by Scherer et al. [1] reported that the accuracy of vocal expression across cultures was 66%. The study was conducted in nine countries in Europe, the United States, and Asia. The vocal emotion is acted in the voices of anger, sadness, fear, joy, and neutrality. Another study [2], a meta-analysis study by Elfenbein and Ambaday, revealed that emotions were universally recognized at better-than-chance levels. The study also examined studies that included both a majority and minority group in the design. Majority group members were poorer at judging minority group members than the reverse.

Scherer's study [1] suggested that language differences between encoders (speakers) and decoders (listeners) may play a major role in recognizing emotion through voices. The difference is significant for Germany and Indonesia but marginally different for the German and Indo-European languages. In Elfenbein and Ambady [2], the accuracy of emotion recognition was higher when emotions were both expressed and recognized by members of the same national, ethnic, or regional group. This result suggests an in-group advantage; members in the same culture have higher recognition than members from different cultures. Both studies suggest that although emotion recognition among humans is universal, there are differences between cultures.

The commonness and differences in human perceptual emotion recognition have been studied (in [3], Figs. 6-7) for building automatic emotion recognition by computers. As in human perception of emotion through voice, computers should be able to distinguish human emotion regardless of the culture. The authors found that the direction of neutral speech to other emotional states is similar for Japanese and Chinese. The difference includes the neutral position, which is slightly different among Japanese, Chinese, and German by Japanese, Chinese, and Vietnamese listeners. Since there is commonness in the perception of emotion from speech, it is possible to build a system that can recognize emotion from speech multiculturally or cross-culturally, operationalized by different countries through languages.

The first study for building a cross-cultural automatic emotion recognition dataset for machines perhaps is the work of Scherer in 2000 [4]. He claimed to build the first large-scale dataset to obtain empirical data on cross-cultural emotion recognition from voices in nine countries on three different continents. The results, still evaluated by human listeners, showed recognition rates that were much better than chance accuracy. Scherer also found that in affective speech, voices other than vocal bursts, segmental and suprasegmental aspects of language affect the encoding and decoding of emotion.

The research of cross-cultural and multicultural emotion then grew with its debates. Other researchers [5] claimed that six basic (facial expressions) emotions are not universal. The claim is based on the study among Westeners and Easteners on Ekman's six basic emotions (happy, surprise, fear, disgust, anger, and sad). The Westeners show a unique pattern of facial movement for each basic emotion, while the Easteners do not (but show unique eye movement). On the other hand, Narimatsu et al. [6] have found some commonalities and differences between English and Japanese visual art labels by neural speakers. It can be argued that aside from emotions being expressed differently among cultures, the perceiver across cultures still could recognize the emotion with some cultural differences.

This paper contributes to the evaluation of multilingual, cross-lingual, and monolingual automatic speech emotion recognition (SER) on the same dataset, i.e., EmoFilm Dataset [7]. To the best of our knowledge, this is the first study to evaluate multilingual, cross-lingual, and monolingual SER simultaneously on this dataset. Given the argument that emotional speech is universal with cultural differences, the performance of all three evaluations (multilingual, cross-lingual, and monolingual) should be comparable under the same test set. The results of this study could suggest the readability of which SER technologies, from these three evaluations, have the potential to be used in future speech technologies.

II. Previous studies

In the previous section, we introduced the classical approach to multilingual emotion recognition. This section discusses recent research on cross-lingual and multilingual speech emotion recognition in chronological order.

Cross-lingual SER

Elbarougy and Akagi [8] in 2013 investigate whether emotions can be estimated cross-lingually. The study proposed three-layer models with semantic primitives in between categorical emotional and acoustic features. To experiment with that proposal, the authors evaluated the Japanese dataset (Fujitsu database) and the German dataset (EmoDB). The results, based on a system with a fuzzy inference system, show small errors in cross-lingual emotion recognition. Both Japanese and German datasets are small and clean (140 and 200 samples); whether the results can be generalized to other languages is questionable.

Neuman and Vu [9] in 2018 conducted experiments with monolingual, cross-lingual, and multilingual SER in English (IEMOCAP dataset) and French (Recola dataset). The best score for French was achieved using a cross-lingual scenario with finetuning on arousal detection. Other dimensions are obtained their best with monolingual (IEMOCAP-valence and Recola-valence) and multilingual (IEMOCAP-arousal). This study is the closest to our work with different datasets and approaches.

Latif et al. [10] in 2019 proposed unsupervised adversarial domain adaptation for cross-lingual SER. The method was evaluated on four datasets: EmoDB, SAVEE (English), EMOVO (Italian), and URDU. They used EmoDB, SAVEE, and EMOVO as source domains and URDU as the target domain and vice versa. The former achieved 65% of UA while the latter achieved 61% of UA (URDU $\rightarrow$ EmoDB). As a comparison, using a multilingual approach achieved 67.3% on URDU and 65.2% on EmoDB (German).

Cai et al. [11] in 2021 proposed a domain adversarial neural network for cross-lingual SER on IEMOCAP and Recola datasets (IEMOCAP $\rightarrow$ Recola and Recola $\rightarrow$ IEMOCAP). Their method outperformed the baseline method but not the monolingual method. The main advantage of their method is no requirement for the target data. No evaluation of multilingual SER is reported, which may obtain better results than the cross-lingual SER.

Multilingual SER

Li and Akagi [12] in 2016 expanded the previous cross-lingual study [8] with a Chinese dataset for multilingual SER. With an SVM as a classifier, the authors evaluated similar three-layer models with the same semantic primitives and acoustic features. The results show that there are small degradations in classification rates from monolingual to multilingual emotion recognition. As in the previous research, the new Casia dataset is also small and clean (198 selected samples), which makes it difficult to draw robust conclusions from the results. The authors then improved their results in 2019 [13] for Japanese and German by 3-4% of UA by combining several acoustic features. The new result, instead, decreases the performance in the Chinese dataset.

Lee [14] in 2019 evaluated two architectures of multitask learning optimized for emotion recognition, three softmax layers and one softmax layer. The first architecture is to predict gender, emotion, and language. The second architecture is to predict language and emotion. The results revealed a better generalization of the second architecture to predict emotion categories with common normalization. The common normalization was performed by normalized features by variances and means from two databases of two languages (English and Japanese).

Zang et al. [15] in 2021 mixed datasets of IEMOCAP, TESS (English), EMA (English), EmoDB, and EMOVO (Italian) and created their training/set partitions based on the newly merged data. The ratio of training:test is set to 4:1. Using wav2vec 2.0 XLSR version (cross-lingual speech representations) with hierarchical grained and feature model (HGFM) for multilingual SER outperformed the baseline with wav2vec 2.0 and HGFM.

Wang et al. [16] in 2022 proposed a multi-gating mechanism with neural architecture search on English (IEMOCAP), German (EmoDB), and French (Cafe) databases. Their proposed methods outperformed the baseline methods for German and French on monolingual evaluations. The result for multilingual evaluations is very similar to the monolingual methods, highlighting the effectiveness of their proposed gating mechanism across different language corpus.

There are several other papers that include "multilingual" terms, but the definition is for the evaluation of several mono languages (e.g., [17], [18]), not a mixed language, nor a multilingual SER model evaluated on mono languages. We exclude these paper since it does not fall in our scope. Furthermore, we provided a fixed split and cross-validation setting in our project repository that other researchers can benchmark in the future.

III. EmoFilm dataset and its partition

The EmoFilm dataset [7] is the focus of this study on evaluating a multilingual speech emotion recognition system, along with its comparison to cross-lingual and monolingual analyses. The emotional speech corpus consists of 1115 samples from English (EN), Italian (IT), and Spanish (SP), extracted from 43 films and 207 speakers. The Italian and Spanish are the dubbed versions of English. The distribution of samples for each language can be seen in Fig. 1, in which Italian has the most portion, followed by English and Spanish.

The dataset includes five categorical emotions: fear, contempt, happiness, anger, and sadness (chance level of 20%).

The average utterance per emotion is 34.3, 41.3, and 35.9 for English, Italian, and Spanish. The distribution of each emotion category is not balanced since the same sentence in the original English is not always considered as emotional as in Italian and Spanish. Table I shows the distribution of each emotion category for each language. Note that the number (including the number in Fig. 1) might be different from the original paper [7]. The counting is based on the latest version (v3) of the EmoFilm dataset available at https://zenodo.org/record/7665999 (accessed on 29 March 2023).

We performed two evaluations for partitioning the dataset into training and test split. Both evaluations are speaker-independent. First, we create a near-balanced fixed split for training and test partition. We evaluated multilingual, cross-lingual, and monolingual emotion recognition with this first evaluation. Second, we evaluated multilingual-only emotion recognition using five-fold cross-validation. The number of samples for each fold is 223 samples (I), which makes it impossible to create a language-balanced partition for each fold due to the nature of the dataset. For the multilingual with the first fixed split, we evaluated the multilingual model on both multilingual samples (a mixed language, 181 samples in the test set) and three monolingual samples (50, 65, and 66 samples in the test set for English, Italian, and Spanish, respectively). The former represents a multilingual model for multilingual samples, while the latter refers to a multilingual model for monolingual samples.

Finally, we downsampled the original samples with 48 kHz to 16 kHz WAV files. This downsampling is needed since the wav2vec2-based model ([19], [20]) to extract acoustic embedding requires 16 kHz WAV files. We provided a Python file to downsample the dataset's files using SoX in the project repository. The repository also contains a CSV file listing files allocated for fixed training/test split and the code to evaluate the cross-validation.

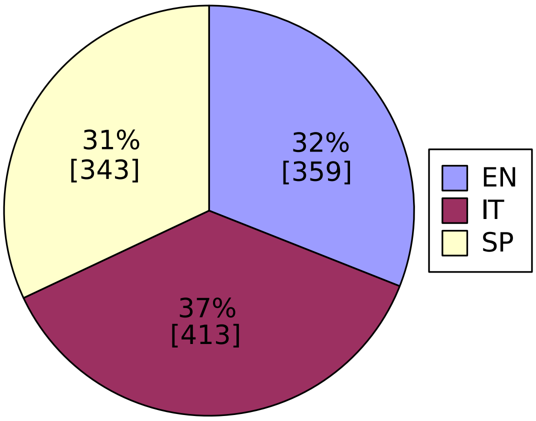

Fig. 1. Percentage [number of samples] of language distribution in EmoFilm dataset; EN: English; IT: Italian; SP: Spanish.

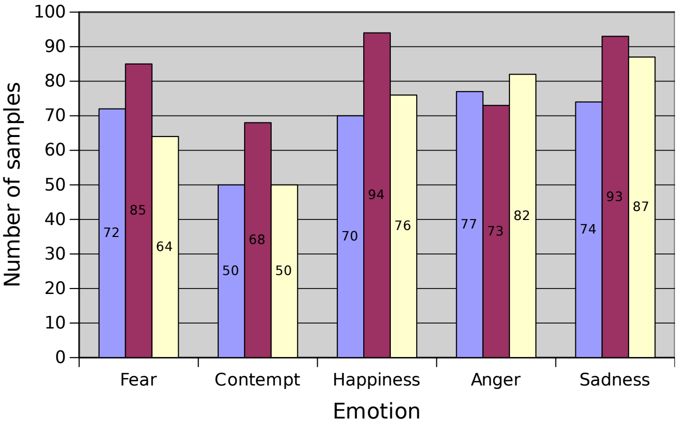

Fig. 2. Distribuion of samples on each emotion category; see Fig. 1 for color code.

TABLE I
DISTRIBUTION OF SAMPLES IN EMOFILM DATASET IN FIXED (TRAIN/TEST) SPLIT AND CROSS-VALIDATION (CV) SPLIT.

Partition	EN	IT	SP	Total
		Fixed split		
Train	293	347	294	934
Test	50	66	65	181
		Cross-validation		
Fold 1	49	122	52	223
Fold 2	100	86	37	223
Fold 3	64	75	84	223
Fold 4	83	72	68	223
Fold 5	47	58	118	223

IV. MULTILINGUAL, CROSS-LINGUAL, AND MONOLINGUAL SER

In this section, we defined the terms multilingual, cross-lingual, and monolingual emotion recognition as shown in Fig. 3. Multilingual emotion recognition is the process of recognizing emotions in multiple languages (e.g., languages A+B+C), either in a set of mixed languages (language A+B) or sets of several mono languages (Fig. 3 left). In this study, we mixed English, Italian, and Spanish to train the speech-emotion recognition (SER) system and tested the system in a set containing sentences from these languages (both in a mixed language or three mono languages). Cross-lingual emotion recognition is the process of recognizing emotions in different languages (language A+B → language C or language B → language C). Monolingual emotion recognition is the process of recognizing emotions in a single language (e.g., language B → language B). The term multicultural and cross-cultural emotion recognition may be better to replace multilingual and cross-lingual since emotion is more related to culture than language. Since different countries with their specific languages operationalize the cultures, we use the term multilingual and cross-lingual in this study following the previous studies.

V. ACOUSTIC FEATURE AND CLASSIFIER

Recent self-supervised learning models have shown promising results in speech emotion recognition [21], [22], [23],

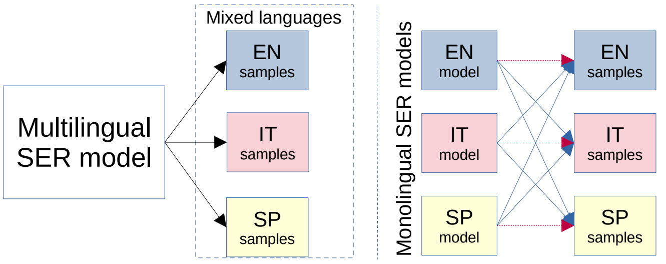

Fig. 3. The difference between multilingual (left), cross-lingual (right: solid blue line), and monolingual SER (right: magenta dot line) in experiments; the models are built using an SVM classifier.

[24]. Following this trend, we evaluated a pre-trained model finetuned on an affective speech dataset [20], [19], which is based on wav2vec2-large-robust model [25] trained on MSP-Podcasts dataset [26]. Models based on wav2vec 2.0 are known to be effective for SER and vocal bursts [27]. The speech embedding is the hidden state of the last layer before the output layer (1024-dim).

We evaluated the support vector machine (SVM) for classification (hence, support vector classification or SVC) for all experiments. This choice is based on the effectiveness of SVM in the previous research on small data [28], [29]. We optimized the C parameter in SVC with a loop in a range [0, 100] with step 0.1. The kernel is RBF with gamma “scale”. Other parameters are left as default. The SVC is implemented in scikit-learn [30] with version `1.0.2`.

The evaluations are evaluated by both unbalanced accuracy (overall accuracy or weighted accuracy, WA) and balanced accuracy (unweighted average recall or unweighted accuracy, UA). WA is suitable for balanced datasets, and UA is suitable for unbalanced datasets. UA is perhaps more suitable in this study since the data is not balanced (both in categorical emotions and languages).

All information to reproduce the experiments is available in the project repository: [https://https://github.com/aistairc/emofilm-mser]. The Python codes are light and can be run on a notebook PC. We used the Iiyama laptop model L140MU (i7-1165G7, 24 GB RAM) with Ubuntu 20.04 OS for experiments. The experiments run for about three minutes for extracting acoustic embedding and within a minute for the classification result. We replicate the methods on the other two PCs to check the consistency.

VI. Results and Discussion

We present our results in two parts. The first part is the multilingual SER evaluation on both a multilingual model for multilingual/mixed data and a multilingual model for monolingual data. The second part is evaluations of cross-lingual SER and monolingual SER.

A. Multilingual results

Table II shows the result of WA and UA of a multilingual model on multilingual data or mixed languages (multi2multi). The first row refers to the fixed training/test split, while the rest are cross-validation results. The results of WA in both cases are generally higher than UA. In this unbalanced data, the UA scores are more reliable since they are the average score of each emotion category. The highest score from cross-validation was achieved on Fold 1 as test data, in which the Italian samples occupy the majority of that test data (Table I). In both cases, the cross-validation attained a lower score (average of five folds) and should be considered as the more reliable score since the data is shifted five times.

TABLE II
Results of multilingual SER model on mixed languages

Split	% WA	% UA
Fixed split	78.45	75.77
Cross-validation		
Fold 1	79.37	78.89
Fold 2	68.61	68.27
Fold 3	69.06	66.04
Fold 4	77.13	77.13
Fold 5	77.13	77.13
Average	72.11	71.91

Table III exhibits the results of a multilingual model on monolingual data (multi2mono). This is similar to Table II but with a different test set for each language. In Table II, we mixed all samples from three languages; in Table III, we separated samples for each language. This evaluation used a fixed-split partition. This result is comparable to Table II, highlighting the robustness of the model. Italian samples attained the highest score in WA. Surprisingly, the highest score for UA was obtained in the Spanish language. A possible assumption for this finding is given as follows. Spanish may be more similar to English than Italian (perhaps because English has borrowed many words from it), and the pre-trained model is trained in English. Although the model for extracting acoustic embedding was trained in English, the portion of English samples in the training set is the lowest among the three languages. These explanations lead to Spanish being the highest score. However, the assumption is rejected since Spanish obtained the lowest scores on language-balanced evaluations (Table V), remaining unclear explanations on why Spanish is gaining the highest score in this multilingual to monolingual evaluation. Nevertheless, the differences in recognition rate occur mainly due to differences in the number of samples for each language. Future research could investigate this phenomenon, particularly by comparing the effect of languages in strict conditions (e.g., same number of samples, same speakers, same emotions, etc.).

TABLE III
Results of multilingual SER model on mono languages (fixed split)

Language	% WA	% UA
EN	78.00	74.52
IT	81.81	72.34
SP	76.92	77.73
Average	78.91	74.86

B. Cross-lingual and monolingual results

Table IV depicts the results of cross-lingual and monolingual speech emotion recognition. Notice that monolingual evaluation can be seen as a special case of cross-lingual evaluation where the source and target are the same language. Hence, the configuration for both is the same. We combined both cases in one table for a better comparison. The diagonal cells are the monolingual SER results. The cross-lingual SER results are in the other cells. The results are in the form of WA / UA. The partition of data follows the fixed split evaluation. If we look at the score of UA, the preferred metric over WA, almost all scores showed low-performance scores ($< 70\%$) for cross-lingual emotion recognition. Only the Italian model tested on the Spanish model gained 70.52% of UA, perhaps due to the similarity between both languages. Monolingual evaluations scored higher than cross-lingual evaluations, particularly on the UA scores. The only exception for this case (monolingual vs. cross-lingual) is the score of WA for the Spanish model, in which the score of WA for Italian data (70.52%) is higher than its own Spanish data (69.15%). Among the three languages in monolingual evaluations, Italian scored the highest (78.41%), perhaps because of the number of samples (on both training and test) and the expressiveness of the Italian language. Although cross-lingual emotion recognition is more challenging than monolingual emotion recognition, one can merge multilingual data to build a multilingual model, which previously performed better than cross-lingual emotion recognition.

TABLE IV
RESULTS (% WA / % UA) OF CROSS-LINGUAL (ALL ROWS EXCEPT THE DIAGONAL MAGENTA) AND MONOLINGUAL (DIAGONAL MAGENTA) SER MODELS; PARTITION: FIXED SPLIT

Language	EN	IT	SP
EN	72.00 / 68.96	62.12 / 52.52	50.76 / 54.86
IT	62.00 / 61.14	77.27 / 78.41	67.69 / 70.52
SP	62.00 / 53.41	77.27 / 65.05	70.76 / 69.15

C. Discussion

The evaluations of multilingual, cross-lingual, and monolingual emotion recognition show the potential benefit of multilingual emotion recognition systems. First, it achieved comparable results to monolingual emotion recognition. The score of average UA was 71.91% in cross-validation multilingual SER (mixed languages) compared to 72.17% in monolingual SER (average of three languages). For a multilingual model to mono languages, the UA was higher, i.e., 74.86%, which is an average score for a multilingual model on three languages. The multilingual model to mono language evaluations may gain this higher score due to more samples in the training data. Cross-lingual SER performed worst since it is the most challenging task. Given the higher score of a multilingual model on mono languages, we found no benefit in using the monolingual model for SER in this study, evidenced by the scores evaluated on the same test set (fixed split). In this fair comparison, multilingual SER gained an average UA of 74.86%, followed by monolingual (average UA 72.21%) and cross-lingual (average UA 58.03%, still higher than chance level 20%). This result is in line with the previous study, in which multilingual SER outperformed cross-lingual and monolingual SER [9] but within different dataset evaluations.

To check that the high score of multilingual SER is due to the number of data, we reduced the number of training samples from the previous 934 samples to 420 samples (160 samples per language), 360 samples (120 samples per language), and 300 samples (100 samples per language). The last number of samples is to match the least sample number (EN 293 samples). Table V shows the result. It can be shown that by reducing the number of samples, the performance of both WA and UA decreased. Interestingly, using only 360 language-balanced samples, we achieved comparable results with 934 language-unbalanced samples. Furthermore, using 420 balanced samples, less than half of the original training samples, we achieved a new higher score of multilingual SER to monolingual samples (multi2mono) with a UA of 75.52%. Another interesting finding in Table V is that Italian SER still obtained the highest scores under language-balanced evaluations. Again, it highlights the expressiveness of Italian culture among English and Spanish.

Table V also shows the inconsistency ranking of languages in the multilingual model to monolingual evaluations. In 420 samples and 360 samples, the best-performing language is English, followed by Italian and Spanish. In this case, the benefit of pre-training wav2vec2 on the English dataset (MSP-Podcasts) as an acoustic feature extractor may be the cause. In the smallest number of samples setting, with 300 samples, Italian gained the highest score, followed by Spanish and English. In this small number of samples, the effect of the classifier may be more dominant than the acoustic feature extractor (perhaps due to the expressiveness of the Italian languages than others).

TABLE V
EFFECT OF REDUCING THE NUMBER OF TRAINING SAMPLES ON MULTILINGUAL SER (OVERLINES SHOW MEAN SCORES)

# Samples	Multi2mono				Multi2multi	
	WA	UA	$\overline{WA}$	$\overline{UA}$	WA	UA
420 samples						
EN	78.00	80.08				
IT	80.30	77.76	74.82	**75.52**	74.59	73.13
SP	66.15	68.71				
360 samples						
EN	74.00	77.39				
IT	78.79	76.33	72.98	74.24	72.83	72.51
SP	66.15	69.00				
300 samples						
EN	66.00	66.52				
IT	75.76	72.37	68.79	68.85	69.06	67.47
SP	64.62	67.67				

There is a tendency that the more samples for a language in the test set, the higher the score. An example is the Italian language which has the most samples in the Fold 1 cross-validation (mixed languages, Table II) and in the fixed split mono language evaluations (Table III). In both cases, the performances were the highest among others. These also can be caused by Italian data having the largest samples in the training set. Future research could be directed to build a language-balanced multilingual speech emotion dataset for observing the effect of language and samples in multilingual SER.

As an extension of this study, we expand our experiments to find insights into the usefulness of the extracted acoustic embedding. We plot the t-SNE visualization in Fig. 4 to show the groups of features according to their labels. As shown in that figure, there is a strong separation between anger and sadness; however, there is confusion between contempt and fear. Happiness is located in between since its features occupy both a specific location (left-top) and other locations (spread). We checked the t-SNE for each language, and the results were similar to the mixed language. This result supports the universality of emotional information from speech.

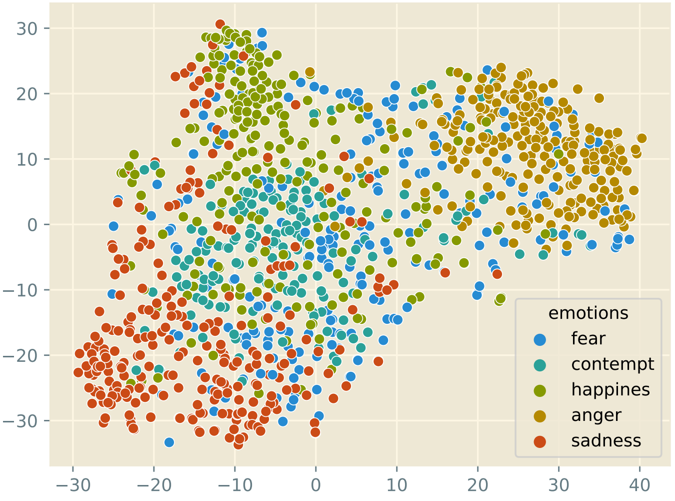

Fig. 4. t-SNE visualization of acoustic embedding from all samples (mixed languages)

VII. Conclusions

The research on multilingual speech emotion recognition has progressively evolved; this study evaluated different cases of speech emotion recognition in speaker-independent multilingual scenarios. It is believed that emotion, including those expressed in speech, is universal across cultures (operated by languages). We support this belief by showing the results of cross-lingual SER, which is higher than the chance level. Multilingual SER achieved better than cross-lingual and monolingual emotion recognition on the same test set. We achieved a score of 74.86% in the multilingual SER model on mono languages and 71.91% in the cross-validation multilingual SER model. The former score is higher than monolingual and cross-lingual evaluations in the same test set (but not on the same number of training sets). The high score of multilingual emotion recognition may be due to its natural benefit from merging all samples from all languages. We experimented with reduced numbers of training samples to prove this hypothesis. We also found a new high score using 420 language-balanced samples for the multilingual model to the monolingual with 75.52% balanced accuracy. However, the fair comparison using a similar number of training samples shows the superiority of monolingual SER among multilingual and cross-lingual.

Nevertheless, the previous scores (multilingual > monolingual > cross-lingual) may also show a fair comparison since all evaluations are done in the same test set. An ideal performance (unbalanced accuracy) would be over 90% for application in real-world scenarios. Hence, a direct implementation of multilingual speech emotion recognition into technology is far from being reliable. The results of this study are promising, but there is still a long way to go before we can achieve a reliable speech emotion recognition system. One way to mitigate this issue is by reducing the number of emotion categories to those that show strong correlations with the extracted acoustic feature/embedding. In this multilingual speech emotion recognition study, we showed that anger and happiness are separable from each other in t-SNE visualization.

Future research could also be directed to investigate the effectiveness of a large number of multilingual SER data by combining several datasets. It also discovered the possibility of training a model simultaneously to predict transcription, Language, age, gender, and emotion from speech [31], [22], [32]. This multitask learning may also help multilingual SER in the future. In this research, we showed that a monolingual setting is still the best-performing configuration under a similar number of training samples. Given the benefit of multilingual data that is available abundantly by nature, it is worth investigating the effectiveness of multilingual over monolingual SER on specific levels (e.g., benchmarking multilingual models to a monolingual dataset which have a large number of samples like English by tracking the performance improvements with the increase of training samples in the multilingual model).

Acknowledgment

This paper is partly based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan. The authors would like to thank the authors of the EmoFilm dataset [7] for sharing their data and Prof. Akagi of JAIST for the valuable discussion.

Ethical Impact Statement

This study explored the feasibility of multilingual speech emotion recognition (language A+B+C → language A+B+C) compared to cross-lingual (language A → language B) and monolingual (language A → language A) speech emotion recognition. One of the evaluations was performed on the same test set (a fixed train/test split) to compare the three evaluations fairly. The study shows the potential benefit of multilingual speech emotion recognition over monolingual and cross-lingual emotion recognition. The results support the previous finding of similar evaluations on the different datasets (English and French) [9]. This study's results are limited to the EmoFilm dataset (English, Italian, and Spanish) and should not be generalized (to other datasets). Instead, it could be used as a reference for future studies. A CSV file for splitting samples into training/test sets will be given in the open repository to enable future benchmarking, along with cross-validation and other settings. The model generated by this study should not be used in real-case scenarios.

References

[1] 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.

[2] 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.

[3] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Emotional Speech Recognition and Synthesis in Multiple Languages toward Affective Speech-to-Speech Translation System," in *2014 Tenth Int. Conf. Intell. Inf. Hiding Multimed. Signal Process.* IEEE, aug 2014, pp. 574–577. [Online]. Available: http://ieeexplore.ieee.org/document/6998394/http://ieeexplore.ieee.org/document/7041623/

[4] 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.

[5] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 19, pp. 7241–7244, 2012.

[6] H. Narimatsu, R. Ueda, and S. Kumano, "Cross-Linguistic Study on Affective Impression and Language for Visual Art Using Neural Speaker," 2022.

[7] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, and B. Schuller, "Categorical vs Dimensional Perception of Italian Emotional Speech," in *Interspeech 2018*, vol. 2018-Septe, no. September. ISCA: ISCA, sep 2018, pp. 3638–3642.

[8] R. Elbarougy and M. Akagi, "Cross-lingual speech emotion recognition system based on a three-layer model for human perception," in *2013 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.* Kaohsiung: IEEE, oct 2013, pp. 1–10.

[9] M. Neumann and N. G. Thang Vu, "Cross-lingual and Multilingual Speech Emotion Recognition on English and French," in *2018 IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2018-April. IEEE, apr 2018, pp. 5769–5773. [Online]. Available: https://ieeexplore.ieee.org/document/8462162/

[10] S. Latif, J. Qadir, and M. Bilal, "Unsupervised Adversarial Domain Adaptation for Cross-Lingual Speech Emotion Recognition," in *2019 8th Int. Conf. Affect. Comput. Intell. Interact. ACII 2019*, 2019.

[11] X. Cai, Z. Wu, K. Zhong, B. Su, D. Dai, and H. Meng, "Unsupervised Cross-Lingual Speech Emotion Recognition Using Domain Adversarial Neural Network," in *2021 12th Int. Symp. Chinese Spok. Lang. Process. ISCSLP 2021*, 2021, pp. 3–7.

[12] X. Li and M. Akagi, "Multilingual Speech Emotion Recognition system based on a three-layer model," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 3608–3612, 2016.

[13] ——, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," *Speech Commun.*, vol. 110, pp. 1–12, 2019. [Online]. Available: https://doi.org/10.1016/j.specom.2019.04.004

[14] S.-w. Lee, "The Generalization Effect for Multilingual Speech Emotion Recognition across Heterogeneous Languages," in *ICASSP 2019 - 2019 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2019, pp. 5881–5885.

[15] 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 *APSIPA Annu. Summit Conf.*, no. December, 2021, pp. 750–755.

[16] Z. Wang, Q. Meng, H. Lan, X. Zhang, K. Guo, and A. Gupta, "Multilingual Speech Emotion Recognition With Multi-Gating Mechanism and Neural Architecture Search," in *SLT*, oct 2022. [Online]. Available: http://arxiv.org/abs/2211.08237

[17] V. Scotti, F. Galati, L. Sbattella, and R. Tedesco, "Combining deep and unsupervised features for multilingual speech emotion recognition," in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12662 LNCS. Springer Science and Business Media Deutschland GmbH, 2021, pp. 114–128.

[18] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "SERAB: A Multi-Lingual Benchmark for Speech Emotion Recognition," in *ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2022, pp. 7697–7701. [Online]. Available: https://ieeexplore.ieee.org/document/9747348/

[19] 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.

[20] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0 (1.1.0)," 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6221127

[21] 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. [Online]. Available: https://ieeexplore.ieee.org/document/9964237/

[22] B. T. Atmaja, Zanjabila, and A. Sasou, "Jointly Predicting Emotion, Age, and Country Using Pre-Trained Acoustic Embedding," in *2022 10th Int. Conf. Affect. Comput. Intell. Interact. Work. Demos.* IEEE, oct 2022, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10085991/

[23] M. Pastor, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Cross-Corpus Speech Emotion Recognition with HuBERT Self-Supervised Representation," in *IberSPEECH 2022*, no. November. ISCA: ISCA, nov 2022, pp. 76–80.

[24] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1. ISCA: ISCA, aug 2021, pp. 521–525.

[25] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Interspeech 2021*, vol. 3. ISCA: ISCA, aug 2021, pp. 721–725.

[26] 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, 2019.

[27] B. T. Atmaja and A. Sasou, "Evaluating Variants of wav2vec 2.0 on Affective Vocal Burst Tasks," in *ICASSP 2023 - 2023 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, jun 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10096552/

[28] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages," in *2018 Int. Conf. Front. Inf. Technol.* IEEE, dec 2018, pp. 88–93.

[29] B. T. Atmaja, Zanjabila, and A. Sasou, "On The Optimal Classifier For Affective Vocal Bursts And Stuttering Predictions Based On Pre-Trained Acoustic Embedding," in *2022 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC).* IEEE, nov 2022, pp. 1690–1695.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825—-2830, 2011.

[31] Z. Kons, H. Aronowitz, E. Morais, M. Damasceno, H.-K. Kuo, S. Thomas, and G. Saon, "Extending RNN-T-based speech recognition systems with emotion and language classification," in *Interspeech 2022*, no. September. ISCA: ISCA, sep 2022, pp. 546–549.

[32] H. Zhao, N. Ye, and R. Wang, "Transferring Age and Gender Attributes for Dimensional Emotion Prediction from Big Speech Data Using Hierarchical Deep Learning," in *2018 IEEE 4th Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput. IEEE Int. Conf. Intell. Data Secur.* IEEE, may 2018, pp. 20–24.