INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Cross-Lingual Cross-Age Group Adaptation for Low-Resource Elderly Speech Emotion Recognition

*Samuel Cahyawijaya**, *Holy Lovenia**, *Willy Chung**, *Rita Frieske, Zihan Liu, Pascale Fung*

The Hong Kong University of Science and Technology
{scahyawijaya, hlovenia, whcchung}@connect.ust.hk

Abstract

Speech emotion recognition plays a crucial role in human-computer interactions. However, most speech emotion recognition research is biased toward English-speaking adults, which hinders its applicability to other demographic groups in different languages and age groups. In this work, we analyze the transferability of emotion recognition across three different languages–English, Mandarin Chinese, and Cantonese; and 2 different age groups–adults and the elderly. To conduct the experiment, we develop an English-Mandarin speech emotion benchmark for adults and the elderly, BiMotion, and a Cantonese speech emotion dataset, YueMotion. This study concludes that different language and age groups require specific speech features, thus making cross-lingual inference an unsuitable method. However, cross-group data augmentation is still beneficial to regularize the model, with linguistic distance being a significant influence on cross-lingual transferability. We release publicly release our code at `https://github.com/HLTCHKUST/elderly_ser`.

Index Terms: speech emotion recognition, cross-lingual adaptation, cross-age adaptation, elderly, low-resource language

1. Introduction

Understanding human emotion is a crucial step towards better human-computer interaction [1, 2]. Most studies on speech emotion recognition are centered on young-adult people, mainly originating from English-speaking countries [3, 4, 5, 6]. This demographic bias causes existing commercial emotion recognition systems to inaccurately perceived emotion in the elderly [7]. Despite the fast-growing elderly demographic in many countries [8], only a few studies with a fairly limited amount of data work on emotion recognition for elderly [9, 10, 11], especially from non-English-speaking countries [12].

To cope with the limited resource problem in the elderly emotion recognition, in this work, we study the prospect of transferring emotion recognition ability over various age groups and languages through the utilization of multilingual pre-trained speech models, e.g., Wav2Vec 2.0 [13]. Specifically, we aim to understand the transferability of emotion recognition ability using only speech modality between two languages and two age groups, i.e, English-speaking adults, English-speaking elderly, Mandarin-speaking adults, and Mandarin-speaking elderly. To do this, we develop BiMotion, a bi-lingual bi-age-group speech emotion recognition benchmark that covers 6 adult and elderly speech emotion recognition datasets from English and Mandarin Chinese. Additionally, we analyze the effect of language distance on the transferability of emotion recognition ability from Mandarin and English using YueMotion, a newly-constructed Cantonese speech emotion recognition dataset.

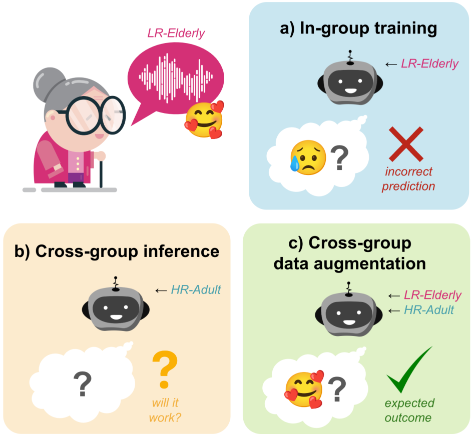

Figure 1: *Elderly speech emotion data in low-resource languages is extremely rare, which makes an in-group trained model perform poorly (a). We explore two ways to improve it: cross-group inference (b) and cross-group augmentation (c).*

We analyze the transferability of emotion recognition ability in three ways, i.e., 1) cross-group inference using only the source group data for training to better understand the features distinction between each group, 2) cross-group data augmentation to better understand the transferability across different groups, and 3) feature-space projection to better understand the effect of transferability across different language distance. Through a series of experiments, we conclude that:

- Very distinct speech features are needed for recognizing emotion across different language and age groups, which makes cross-lingual inference an ill-suited method for transferring speech emotion recognition ability.
- Despite the different important speech features across groups, data augmentation across different groups is still beneficial as it helps the model to better regularize yielding higher evaluation performance.
- Language distance plays a crucial role in cross-lingual transferability, as more similar languages tend to share more similar speech characteristics, thus allowing better generalization to the target language.

2. Related Work

2.1. Emotion Recognition

Various efforts have explored emotion recognition solutions for around two decades ago through different modalities [14, 15,

* Equal contribution.

10.21437/Interspeech.2023-327

Table 1: *Statistics of datasets covered in BiMotion.*

Dataset	Language	Age Group	#Train	#Valid	#Test
CREMA-D	English English	Elderly Adults	150 5000	42 750	300 1200
CSED	Mandarin	Elderly	200	52	400
ElderReact	English	Elderly	615	355	353
ESD	Mandarin English	Adults Adults	15000 15000	1000 1000	1500 1500
IEMOCAP	English	Adults	7500	1039	1500
TESS	English English	Elderly Adults	699 700	200 201	500 500

16, 17, 18, 19]. Most studies focus on methods that can estimate the subject emotion effectively through a specific modality while some others focus on combining multiple modalities to better estimate the subject emotion. The transferability of emotion recognition across different cultures and language groups has also been previously studied [20, 21, 22] showing that there are significant differences across languages and cultures. Nevertheless, it is only evaluated on a small-scale model that owns limited world knowledge. In this work, we extend this analysis to pre-trained speech models and further explore the transferability to cover an extremely low-resource group, i.e., the elderly in the low-resource language group.

2.2. Pre-trained Speech Model

Large pre-trained speech models achieve state-of-the-art performance for various speech tasks in recent years [13, 23, 24]. These models have been trained in an unsupervised manner on large-scale multilingual speech corpora. The representation of these models has been shown to be effective, achieving state-of-the-art performance in various downstream tasks including automatic speech recognition, speaker verification, language identification, and emotion recognition [25, 26, 27, 28, 29]. In this work, we adopt the XLSR Wav2Vec 2.0 model [13, 24], which has been pre-trained on large-scale multilingual speech corpora. The multilingual representation learned by the model will be crucial for evaluating the transferability of emotion recognition ability across different languages.

3. Benchmark and Dataset

3.1. BiMotion Benchmark

To test the transferability of emotion recognition ability across different languages and age groups, we develop a bilingual speech emotion recognition benchmark namely BiMotion, that covers two languages, i.e., English and Mandarin, and two age groups, i.e., Elderly (age $\geq$60 y.o) and Adults (age 20-59). We construct our benchmark from a collection of six publicly available datasets, i.e., CREMA-D [9], CSED [12], ElderReact [10], ESD [30], IEMOCAP [3], and TESS [11]. For datasets with an official split, i.e., ESD and ElderReact, we use the provided dataset splits. For the other datasets, we generate a new dataset split since there is no official split provided. The statistics of the datasets in BiMotion are shown in Table 1.

3.2. YueMotion Dataset

To strengthen our transferability analysis, we further introduce a new speech emotion recognition dataset covering adults in Cantonese named YueMotion. The utterances in YueMotion are recorded by 11 speakers (4 males and 7 females) each with a personal recording device. During the recording session, each speaker is asked to say out loud 10 pre-defined sentences used in daily conversations with 6 different emotions, i.e., happiness, sadness, neutral, disgust, fear, and anger. The YueMotion dataset is used to further verify our hypothesis regarding the transferability of emotion recognition ability. The detailed statistics of YueMotion dataset are shown in Table 2.

Table 2: *Statistics of the YueMotion dataset.*

Age Group	Gender	#Train	#Valid	#Test	#All
Adults	Male Female	120 210	36 63	84 147	240 420
Total Samples		**330**	**99**	**231**	**660**

4. Methodology

We employ two methods to evaluate the transferability of emotion recognition ability, i.e., cross-group data augmentation and cross-group inference.[1] The detail of each method is explained in the following subsection.

4.1. Cross-Group Inference

We define a set of language L=$\{l_1,\ l_2,\ \ldots,\ l_n\}$ and a set of age group A=$\{a_1,\ a_2,\ \ldots,\ a_m\}$. We define training and evaluation data from a language l_i and age group a_j as $X^{trn}_{l_i,a_j}$ and $X^{tst}_{l_i,a_j}$, respectively. To understand the transferability from the source group to the target group, we explore a method called cross-group inference. Following prior works on zero-shot accent/language adaptation [31, 32, 33], we explore cross-lingual and cross-age inference settings in our experiment. Specifically, given a training dataset $X^{trn}_{l_i,a_j}$, we select three out-of-group test datasets: 1) datasets that are not in l_i language ($X^{tst}_{\neg l_i,a_j}$), 2) datasets that are not in a_j age-group ($X^{tst}_{l_i,\neg a_j}$), and 3) datasets that are not in l_i language and a_j age-group $X^{tst}_{\neg l_i,\neg a_j}$. We conduct cross-group inference by training the models on a specific group training set and evaluating them on the three out-of-group test sets. With this method, we can analyze the effect of cross-group information on a specific language and age group.

4.2. Cross-Group Data Augmentation

Prior works [34, 35] have shown that cross-domain and cross-lingual data augmentation method effectively improves the model performance in textual and speech modalities, especially on low-resource languages and domains. We adopt the data augmentation technique to the speech emotion recognition task by utilizing cross-lingual and cross-age data augmentation. Our cross-group data augmentation injects a training dataset $X^{trn}_{l_i,a_j}$ from a language group l_i and an age group a_j with data from other language and age groups producing a new augmented dataset. Given $X^{trn}_{l_i,a_j}$, there are three different data augmentation settings possible, i.e., cross-lingual data augmentation on the same age group resulting in X^{trn}_{L,a_j}, cross-age data augmentation on the same language resulting in $X^{trn}_{l_i,A}$, and cross-lingual cross-age data augmentation resulting in $X^{trn}_{L,A}$. We then fine-tune the model on the augmented dataset and evaluate it on $X^{tst}_{l_i,a_j}$, the test set with a language l_i and age group a_j.

[1] To enhance readability, we use colored underlines to mark the term cross-group data augmentation and cross-group inference onwards.

Table 3: *Evaluation results of cross-lingual and cross-age groups inference on BiMotion.* ***Cross-all*** *denotes cross-age cross-lingual inference.* ***eng****,* ***zho****,* ***eld****, and* ***adu*** *denote English, Mandarin, elderly, and adults, respectively.*

Training Data	Test Data				Avg.
	l_i=eng a_j=eld	l_i=eng a_j=adu	l_i=zho a_j=eld	l_i=zho a_j=adu	
Cross-all ($X^{trn}_{\neg l_i,\neg a_j}$)	27.48	9.71	24.75	28.71	22.66
Cross-age ($X^{trn}_{l_i,\neg a_j}$)	40.40	27.69	26.69	14.56	27.34
Cross-lingual ($X^{trn}_{\neg l_i,a_j}$)	14.54	32.42	16.06	57.54	30.14
Baseline ($X^{trn}_{l_i,a_j}$)	**55.12**	**70.28**	**57.00**	**97.00**	**69.85**

Table 4: *Evaluation results of cross-lingual and cross-age groups data augmentation on BiMotion.* ***Cross-all*** *denotes cross-age cross-lingual data augmentation.* ***eng****,* ***zho****,* ***eld****, and* ***adu*** *denote English, Mandarin, elderly, and adults, respectively.*

Training Data	Test Data				Avg.
	l_i=eng a_j=eld	l_i=eng a_j=adu	l_i=zho a_j=eld	l_i=zho a_j=adu	
Cross-all ($X^{trn}_{L,A}$)	**68.06**	**77.24**	52.21	**97.40**	**73.73**
Cross-age ($X^{trn}_{l_i,A}$)	66.96	75.47	**59.93**	97.00	74.84
Cross-lingual (X^{trn}_{L,a_j})	54.55	74.13	54.84	95.82	57.34
Baseline ($X^{trn}_{l_i,a_j}$)	55.12	70.28	57.00	97.00	69.85

Through this method, we can analyze the influence of cross-group augmentation on a specific language and age group.

4.3. Feature-Space Projection

Given a fine-tuned emotion recognition model θ, we extract the speech representation by taking the high-dimensional features before the last linear classification layer. By using θ, we extract the speech features from data points and project them into a 2-dimensional space with UMAP [36] for visualization. With this approach, we can approximate the effect of augmenting cross-group data points to the training coverage, which will determine the prediction quality of the model to the evaluation data in a specific language l_i and age-group a_j.

5. Experimental Settings

5.1. Baseline Models

For our experiment, we utilize the XLSR-53 Wav2Vec 2.0 [13, 24] model which is pre-trained in 53 languages including English, Mandarin, and Cantonese.[2] We compare the performance of the cross-group data augmentation and cross-group inference settings with a simple in-group training baseline where given a language l_i and an age-group a_j, the model is fine-tuned only on X_{l_i,a_j} and evaluated on the test set of the corresponding language and age-group.

5.2. Training Settings

To evaluate the transferability of emotion recognition ability, we conduct an extensive analysis in the BiMotion benchmark. We explore two different cross-group transfer methods, i.e., cross-group data-augmentation (§4.2) and cross-group inference (§4.1). Specifically, we fine-tune the pre-trained XLSR-53 Wav2Vec 2.0 model using the in-group and the cross-group augmented datasets, i..e., X_{l_i,a_j}, X_{a_j}, X_{l_i}, X_{all}, and evaluate on all the test set covering both the in-group and the out-of-group test data, i.e, $X^{test}_{l_i,a_j}$, $X^{test}_{\neg l_i}$, $X^{test}_{\neg a_j}$, and $X^{test}_{\neg l_i,\neg a_j}$.

To further strengthen our analysis, we conduct the transferability experiment on the YueMotion dataset. We use the Cantonese elderly and Cantonese adults data on the YueMotion as the test set, and explore cross-group data augmentation and cross-group inference for those test datasets. To measure the effectiveness of the cross-group transferability on YueMotion, we utilize the training datasets from BiMotion and compare the cross-group transfer methods with the baseline trained only on the YueMotion training dataset.

Hyperparameters We use the same hyperparameters in all experiments. For fine-tuning the model we use the following hyperparameter setting, i.e., AdamW optimizer [37] with a learning rate of 5e-5, a dropout rate of 0.1, a number of epochs of 20, and early stopping of 5 epochs.

5.3. Evaluation Metrics

Some speech utterances may consist of multiple emotion labels, e.g., happiness and surprise, which makes emotion recognition a multilabel problem. By framing the problem as multilabel, the occurrence of each emotion becomes sparse, thus leading to an imbalanced label distribution. To consider the imbalance distribution, we use binary F1-score to compute per-label evaluation performance and take the weighted F1-score over different emotion labels, which is an average of the F1 scores for each class weighted by the number of samples in that class.

6. Results and Discussion

6.1. Transferability via Cross-Group Inference

The results of our cross-group inference experiment are shown in Table 3. Based on our experiments, cross-lingual, cross-age, and cross-age cross-lingual inferences are not beneficial for improving the performance of the target group compared to the baseline trained on the specific language and age group, instead, they hinder the performance by a huge margin. For instance, the best cross-age inference comes from **English-Adults** to **English-Elderly**, however, it still hampers the performance of the in-group **English-Elderly** baseline by ∼15% weighted F1-score. While for cross-lingual inference, the results are much worse with more than $\sim 30\%$ lower weighted F1-score than the corresponding in-group training baseline. This result suggests that there are some differences in speech features used to recognize emotion between different languages and different age groups, which makes transfer through cross-group inference not suitable for speech emotion recognition.

6.2. Transferability via Cross-Group Data Augmentation

Despite having different characteristics, cross-group transfer through data augmentation could provide richer and more diverse samples of emotion-induced speech and help to avoid overfitting problems, especially in a low-resource setting. Based on our cross-group data augmentation results in Table 4, the **English-Elderly** and **English-Adults** groups benefit the most from the cross-group data augmentation, gaining ∼12% and ∼5% weighted F1-score respectively. While for the Mandarin groups, the improvement is quite small. **Mandarin-Elderly** gains ∼3% weighted F1-score with data augmentation from **Mandarin-Elderly**, while the result for the **Mandarin-Adults** remains the same, which is probably due to the limited

[2] We use the model checkpoint from `https://huggingface.co/facebook/wav2vec2-large-xlsr-53`.

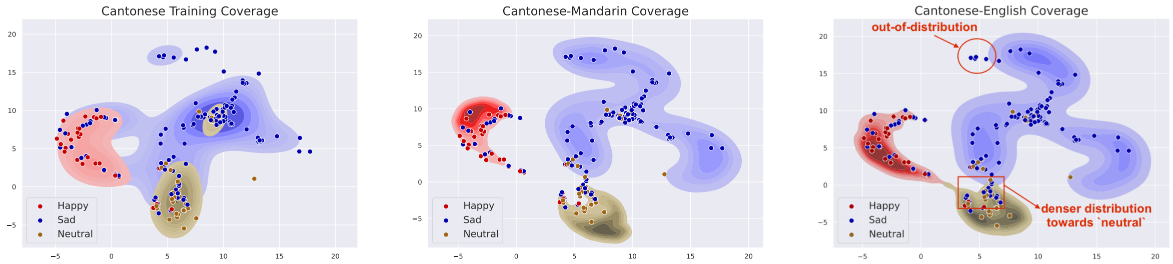

Figure 2: *Cross-group data augmentation on the Cantonese-Adults data. Blue, red, olive regions represent the density plot of the training data for happy, sad, and neutral emotions, respectively.*

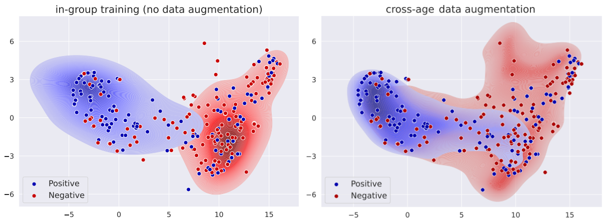

Figure 3: *Cross-group data augmentation on the Mandarin-Elderly data. Blue, and red regions denote the density of the training data for positive and negative emotions, respectively.*

data available from the **Mandarin-Elderly** group.

For cross-lingual data augmentation, the effect is apparent with the only improvement coming from the **English-Adults** group, while the performance in other groups decreases. We conjecture this is due to the huge linguistic differences between English and Mandarin [21, 38]. Further analysis regarding the effect of language distance is discussed on §6.3. Furthermore, a combination of cross-lingual and cross-age data augmentation tends to improve the performance even higher. Specifically, the performance on **English-Elderly** and **English-Adults** increase by ∼13% and ∼17% weighted F1-score, respectively. The performance on **Mandarin-Adults** group improves marginally by ∼0.4% weighted F1-score. On the other hand, the performance of **Mandarin-Elderly** decreases by ∼5% weighted F1-score compared to the in-group training baseline. This might be caused by distributional shift due to the large amount of augmentation from other groups with respect to the amount of data in the **Mandarin-Elderly** group.

We further analyze the cross-group behavior further through feature-space projection using the model trained on all BiMotion training data. As shown in Figure 3, cross-age data augmentation improves the training coverage of the model resulting in a better generalization on unseen evaluation data, despite the speech feature differences among different age groups. This shows the potential of leveraging cross-age data augmentation for modeling low-resource groups such as the elderly.

6.3. Effect of Language Distance on Transferability

We further analyze the effect of language distance on cross-lingual data augmentation. Specifically, we explore the effect of cross-lingual data augmentation from Mandarin and English to Cantonese. To balance the amount of data between English and Mandarin, we only utilize the ESD dataset [30] which contains 15K training, 1K validation, and 1.5K utterances for each English and Mandarin. For the Cantonese dataset, we utilize the newly constructed speech emotion dataset, YueMotion. As shown in Table 5, the performance on the **Cantonese** data increases when the model is augmented with **Mandarin** data and decreases when the model is augmented with **English**. This follows the linguistic similarity of languages; Mandarin and Cantonese are more similar as both come from the same language family, i.e., Sino-Tibetan, compared to English that comes from the Indo-European language family. Similar pattern is also observed in the Mandarin test data when the Cantonese-Mandarin trained model outperforms the Cantonese-Mandarin-English model by ∼3% weighted F1-score.

Table 5: *The effect of cross-lingual transferability in YueMotion.*

Training Data	Test Data		
	l_i=yue	l_i=eng	l_i=zho
Cross-lingual (yue+eng+zho)	47.59	**92.61**	93.75
Cross-lingual (yue+eng)	43.59	92.51	-
Cross-lingual (yue+zho)	**51.60**	-	**97.00**
Baseline (yue)	45.33	-	-

We also analyze the cross-lingual behavior further through feature-space projection with the Cantonese-Mandarin-English trained model. As shown in Figure 2, the original Cantonese training data cannot cover all the test data. When Mandarin data is added, the training coverage improved which covers more test samples for each label. When English data is added, the training coverage is improved, but not without shifting the training distribution, e.g., the out-of-distribution on the sad label, causing the model to perform worse on the Cantonese evaluation data.

7. Conclusion

We investigate the transferability of speech emotion recognition ability to achieve a better generalization for the elderly in low-resource languages. We construct an English-Mandarin speech emotion benchmark for adults and the elderly, namely BiMotion, from six publicly available datasets. We also construct, YueMotion, a low-resource speech emotion dataset for adults in Cantonese to analyze the effect of language distance in cross-lingual data augmentation. Based on the experiments, we conclude: 1) significantly distinct speech features are necessary to recognize emotion across different language and age groups, 2) although the speech features may vary across different groups, cross-group data augmentation is still beneficial to better generalize the model, and 3) language distance substantially affects the effectiveness of cross-lingual transferability. Our analysis lays the groundwork for developing more effective speech emotion recognition models for low-resource groups, e.g., the elderly in low-resource languages.

8. References

[1] R. Beale and C. Peter, "The role of affect and emotion in HCI," in *Affect and Emotion in Human-Computer Interaction*. Springer Berlin Heidelberg, 2008, pp. 1–11.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Nov. 2008.

[4] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, May 2018.

[5] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *32nd AAAI*, 2018.

[6] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th ACL*. ACL, Jul. 2019, pp. 527–536.

[7] E. Kim, D. Bryant, D. Srikanth, and A. Howard, "Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults," in *Proc. 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 638–644.

[8] D. Rouzet, A. C. Sanchez, T. Renault, and O. Roehn, "Fiscal challenges and inclusive growth in ageing societies," no. 27, 2019.

[9] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, Oct. 2014.

[10] K. Ma, X. Wang, X. Yang, M. Zhang, J. M. Girard, and L.-P. Morency, "Elderreact: A multimodal dataset for recognizing emotional response in aging adults," in *2019 International Conference on Multimodal Interaction*, ser. ICMI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 349–357.

[11] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (tess)," 2020.

[12] M. C. Lee, S. C. Yeh, S. Y. Chiu, and J. W. Chang, "A deep convolutional neural network based virtual elderly companion agent," in *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, Jun. 2017.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[14] J. Martinez-Miranda and A. Aldea, "Emotions in human and artificial intelligence," *Computers in Human Behavior*, vol. 21, no. 2, pp. 323–341, Mar. 2005.

[15] D. Bertero and P. Fung, "Multimodal deep neural nets for detecting humor in tv sitcoms," in *2016 IEEE SLT Workshop*, 2016.

[16] L. Cen, M. Dong, H. L. Z. L. Yu, and P. Ch, "Machine learning methods in the application of speech emotion recognition," in *Application of Machine Learning*. InTech, Feb. 2010.

[17] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *IEEE ICASSP*, 2017.

[18] W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung, "Multimodal end-to-end sparse model for emotion recognition," in *Proceedings of the NAACL-HLT 2021*. Online: ACL, Jun. 2021, pp. 5305–5316.

[19] S. Latif, H. S. Ali, M. Usama, R. Rana, B. Schuller, and J. Qadir, "Ai-based emotion recognition: Promise, peril, and prescriptions for prosocial path," 2022.

[20] S. Hareli, K. Kafetsios, and U. Hess, "A cross-cultural study on emotion expression and the learning of social norms," *Frontiers in Psychology*, vol. 6, Oct. 2015.

[21] N. Lim, "Cultural differences in emotion: differences in emotional arousal level between the east and the west," *Integrative Medicine Research*, vol. 5, no. 2, pp. 105–109, Jun. 2016.

[22] I. Iosifov, O. Iosifova, O. Romanovskyi, V. Sokolov, and I. Sukailo, "Transferability evaluation of speech emotion recognition between different languages," in *Advances in Computer Science for Engineering and Education*. Springer International Publishing, 2022, pp. 413–426.

[23] 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," 2021.

[24] 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," in *Proc. Interspeech 2022*, 2022.

[25] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on Speaker Verification and Language Identification," in *Proc. Interspeech 2021*, 2021, pp. 1509–1513.

[26] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[27] H. Lovenia, S. Cahyawijaya, G. Winata, P. Xu, Y. Xu, Z. Liu, R. Frieske, T. Yu, W. Dai, E. J. Barezi, Q. Chen, X. Ma, B. Shi, and P. Fung, "ASCEND: A spontaneous Chinese-English dataset for code-switching in multi-turn conversation," in *Proc. 13th LREC*. ELRA, Jun. 2022, pp. 7259–7268.

[28] S. Cahyawijaya, H. Lovenia, A. F. Aji, G. I. Winata, B. Wilie, , R. Mahendra, C. Wibisono, A. Romadhony, K. Vincentio, F. Koto, J. Santoso, D. Moeljadi *et al.*, "Nusacrowd: Open source initiative for indonesian nlp resources," 2022.

[29] W. Dai, S. Cahyawijaya, T. Yu, E. J. Barezi, P. Xu, C. T. Yiu, R. Frieske, H. Lovenia, G. Winata, Q. Chen, X. Ma, B. Shi, and P. Fung, "CI-AVSR: A Cantonese audio-visual speech datasetfor in-car command recognition," in *Proc. 13th LREC*. ELRA, 2022.

[30] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[31] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, and P. Fung, "Learning Fast Adaptation on Cross-Accented Speech Recognition," in *Proc. Interspeech 2020*, 2020.

[32] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *Proc. 37th ICML*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020.

[33] G. I. Winata, A. F. Aji, S. Cahyawijaya, R. Mahendra, F. Koto, A. Romadhony, K. Kurniawan, D. Moeljadi, R. E. Prasojo, P. Fung, T. Baldwin, J. H. Lau, R. Sennrich, and S. Ruder, "Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages," 2022.

[34] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, and S.-H. Lai, "Auggan: Cross domain adaptation with gan-based data augmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[35] J. Singh, B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "Xlda: Cross-lingual data augmentation for natural language inference and question answering," 2019.

[36] L. McInnes, J. Healy, N. Saul, and L. Grossberger, "Umap: Uniform manifold approximation and projection," *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[38] C. Fu, T. Dissanayake, K. Hosoda, T. Maekawa, and H. Ishiguro, "Similarity of speech emotion in different languages revealed by a neural network with attention," in *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*. IEEE, Feb. 2020.