

# The use of generative adversarial network as a domain adaptation method for cross-corpus speech emotion recognition

Muhammad Farhan Fadhil, Amalia Zahra

Department of Computer Science, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

## Article Info

### Article history:

Received Feb 17, 2024

Revised Aug 31, 2024

Accepted Sep 28, 2024

### Keywords:

Cross-corpus SER

Domain adaptation

Generative adversarial networks

SER performance degradation

Speech emotion recognition

## ABSTRACT

The research of speech emotion recognition (SER) is growing rapidly. However, SER still faces a cross-corpus SER problem which is performance degradation when a single SER model is tested in different domains. This study shows the impact of implementing a generative adversarial network (GAN) model for adapting speech data from different domains and performs emotion classification from the speech features using a 1D convolutional neural network (CNN) model. The results of this study found that the domain adaptation approach using a GAN model could improve the accuracy of emotion classification in speech data from 2 different domain such as the ryerson audio-visual database of emotional speech and song (RAVDESS) speech corpus and the EMO-DB speech corpus ranging from 10.88% to 28.77%, with the highest average performance increase across three different class balancing method reaching 18.433%.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Muhammad Farhan Fadhil

Department of Computer Science, BINUS Graduate Program, Master of Computer Science

Bina Nusantara University

27, Kebon Jeruk Raya Street, RT.1/RW.9, Kemanggisan, Palmerah, West Jakarta, Jakarta 11530, Indonesia

Email: muhammad.fadhil009@binus.ac.id

## 1. INTRODUCTION

Human emotion is naturally expressed throughout every process of the human communication. Emotions could be expressed through many types of natural responses such as bodily gestures, facial expressions, sweat glands, language expression, and other responses [1]. One of the non-invasive ways to detect human emotions is to analyze speech signal. Speech emotion recognition or SER is a research topic in the development of artificial intelligence that focuses on analyzing a human speech signal to accurately predict emotions embedded within the speech signal.

The current research advancement in the field of SER, has found that many studies have achieved outstanding results in the case of estimating valence/arousal [2]-[4] and performing the task of emotion classification [5]-[9]. One of the challenges that the SER faces currently is how the models encounter a performance degradation issue when tested with data that comes from a different domain which occurs because of the domain shifting problem [10]. This issue will undeniably be quite problematic when the SER systems eventually are targeted to be implemented in real-world situations.

Domain adaptation is one method that can be used to produce a model that has good adaptability when faced with data characteristics from different domains. One approach that has been reviewed in previous studies is to reduce or eliminate the distribution of source and target domain data to minimize the occurrence of domain shifting [11]. In addition, one approach using supervised domain adaptation utilizes labeled utterances to adapt data in the source domain to the target domain [12]. In the case of SER, one thing

that attracts the attention of researchers in solving this domain adaptation problem is the generative adversarial network (GAN) model. This is because the GAN model can learn and produce data that resembles the distribution of the input data [13].

A cross-lingual SER study was carried out by implementing a GAN-based SER model which was designed in such a way that it did not require data labels in the target domain and obtained the highest unweighted average recall rate (UAR) results in testing URDU target data by training the model using EMO-DB with value 65.2% [14]. A 2021 study tested the conditional cycle emotional generative adversarial network (CCEmoGAN) to increase the variability of data-aware source corpus data on the target corpus by synthesizing the source corpus and obtaining a UAR of 51.13% [15]. In this research, a novel domain adaptation approach is carried out by utilizing a GAN model which was trained by using a small sample of the data taken from the target domain to enrich the training pool of the classifier in the hope of improving the model's performance. Figure 1 illustrates the use of GANs in contribution of cross-corpus SER domain adaptation.

Figure 1(a) illustrates the cross-corpus SER problem that arises when an emotion classification model is trained on a specific source domain. This classification model may exhibit optimal performance when used for emotion classification within the same domain. However, performance degradation occurs when the model is tested on different target domain that has not been previously seen by the classification model. The objective of this research is to propose and evaluate a domain adaptation method that utilizes a GAN model to augment training data by incorporating synthetic data from a target domain to be adapted with the data from the source domain, aiming to minimize performance degradation in cross-corpus SER. The method proposed in this study is depicted in Figure 1(b).

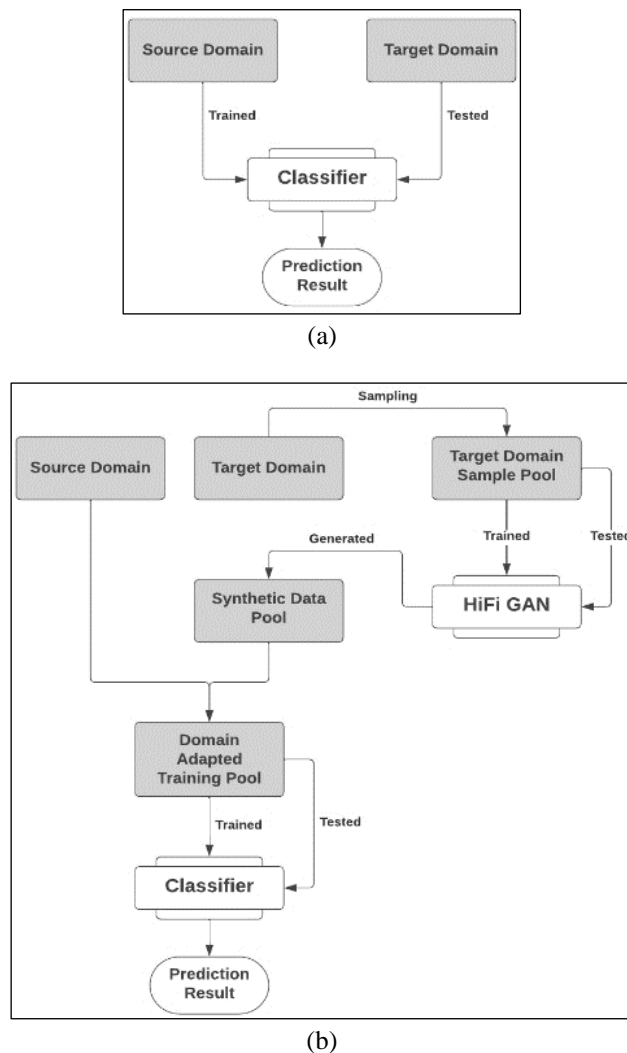


Figure 1. Cross-corpus SER; (a) problems and (b) the proposed domain adaptation approach

## 2. METHOD

This research utilizes the HiFi-GAN model [16] which is used as a data augmentation method in order to enrich the training data pool with the hope that the classification model that will be used will then have prior knowledge of the characteristics of data from the target domain. The classification model that will be used in this research is a one-dimensional deep convolutional neural network (DCNN) model, which is based on the model proposed by Issa *et al.* [17] in 2020 because it has been proven to have satisfactory performance for emotion classification. Figure 2 depicts the research stages carried out in this study. These stages include data fetching, GAN model development, domain adaptation approach, pre-processing, classification model development, and model evaluation.

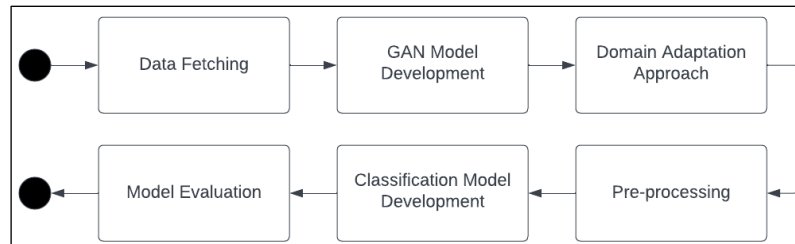


Figure 2. Research process flow

### 2.1. Data fetching

This research is performed by using data that originated from two emotional speech corpora from different languages, namely the ryerson audio-visual database of emotional speech and song (RAVDESS) Dataset in English [18] and EMO-DB in German [19]. Table 1 presents the distribution of the number of utterances from each emotion class from the RAVDESS and EMO-DB datasets. Both corpora have data differences in the availability of emotion classes. For example, the RAVDESS dataset has the emotion class 'sad', whereas EMO-DB does not have that emotion class. To avoid zero-prior knowledge problems in this study, each emotion class that has an availability conflict between them will be eliminated [20].

Table 1. Number of utterances for each emotion class in RAVDESS and EMO-DB

Emotion class	RAVDESS	EMO-DB
Angry	192	127
Anxiety	-	72
Boredom	-	81
Calm	192	-
Disgust	192	46
Fearful	192	69
Happy	192	71
Neutral	96	79
Sad	192	-
Surprised	192	-

### 2.2. Generative adversarial networks model

HiFi-GAN is a GAN model that consists of a single generator and two discriminators which are used to synthesize audio signal data [16]. The HiFi-GAN generator structure shown in Figure 3 is based on a CNN model that receives input in the form of a mel-spectrogram and performs upsampling by applying transposed convolutions until the output sequence reaches the temporal resolution of a raw waveform. The two discriminators used in HiFi-GAN shown in Figure 4, are the multi-period discriminator (MPD) and the multi-scale discriminator (MSD) with each having its own tasks. MPD focuses on capturing the structure of the whole audio signal implicitly [16]. On the other hand, MSD is intended to capture sequential patterns with long-term dependencies. The MSD used in HiFi-GAN is adapted based on the MelGAN model [21] from a previous study.

The Hi-Fi GAN model is used to produce synthetic speech data using real data taken from the target domain on a small sample pool. The model used in this research is a pretrained model UNIVERSAL\_V1 that has been trained using 3 different speech corpus such as LJ Speech Dataset [22], LibriTTS [23], and VCTK [24]. Figure 5 shows the mel-spectrogram comparison between the real and synthetic data of utterances of an emotion.

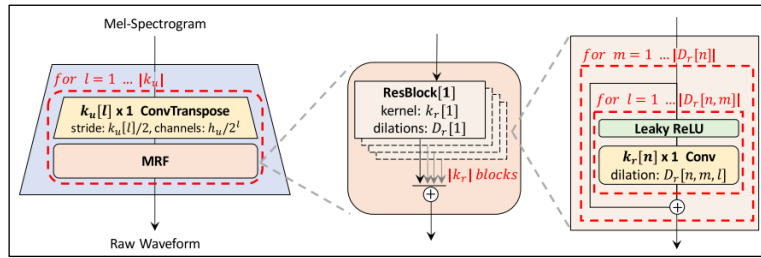


Figure 3. The structure of generator in HiFi-GAN [16]

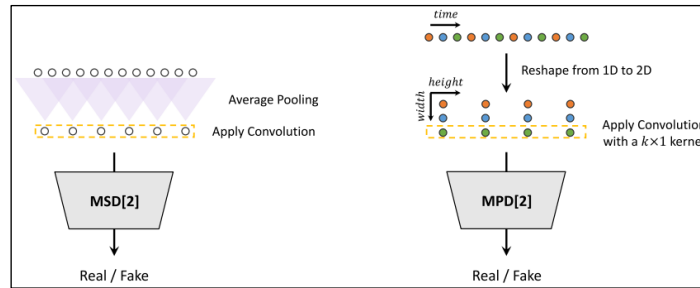


Figure 4. The structure of discriminator in HiFi-GAN [16]

Figure 5(a) presents a visualization of the mel-spectrogram of the original audio signal extracted from the Emo-DB corpus for the neutral emotion. Meanwhile, Figure 5(b) shows the mel-spectrogram visualization of the synthesized data generated by the HiFi GAN model used in this study. In general, the audio signals appear quite similar to their original counterparts throughout the signal. The most significant visual differences can be observed in frequencies above 8192 Hz, indicating the presence of new signal artifacts not present in the original data.

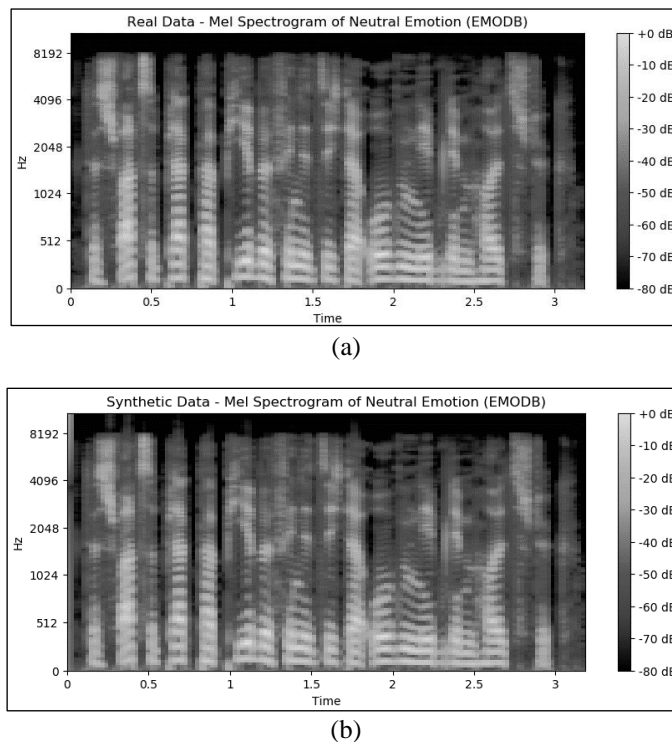


Figure 5. Mel-spectrogram of; (a) real speech data and (b) generated synthetic data of neutral emotion taken from the EMO-DB corpus

The Hi-Fi GAN model is used to produce synthetic speech data using real data taken from the target domain on a small sample pool. The model used in this research is a pretrained model UNIVERSAL\_V1 that has been trained using 3 different speech corpus such as LJ Speech Dataset [22], LibriTTS [23], and VCTK [24]. Figure 5 shows the mel-spectrogram comparison between the real and synthetic data of utterances of an emotion.

**2.3. Domain adaptation approach**

The domain adaptation approach carried out in this research is the use of a training pool that has been adapted to the target domain by enriching the training pool with synthetic data produced by the HiFi-GAN model which then will be used to train the classification model. This is done in the hope that the classification model will be able to have prior knowledge of the characteristics of the target domain data by only utilizing a small portion of the data from the target domain so that it can be used to obtain better emotion classification accuracy.

The domain adaptation approach begins by taking a small data sample from the target domain which is then used to train the HiFi-GAN model. After the model has been successfully trained and has achieved optimal data synthesis results, the data augmentation can be carried out. Synthetic audio signal data from the target domain will then be entered into a final training pool which will later be used to train the classification model for testing against the target domain. Figure 6 is a diagram of the domain adaptation approach proposed in this research. In this figure it is assumed that each source and target domain have 5 emotion classes (color coded) where a small portion of data from the target domain will be taken to be used as training data for the HiFi-GAN model.

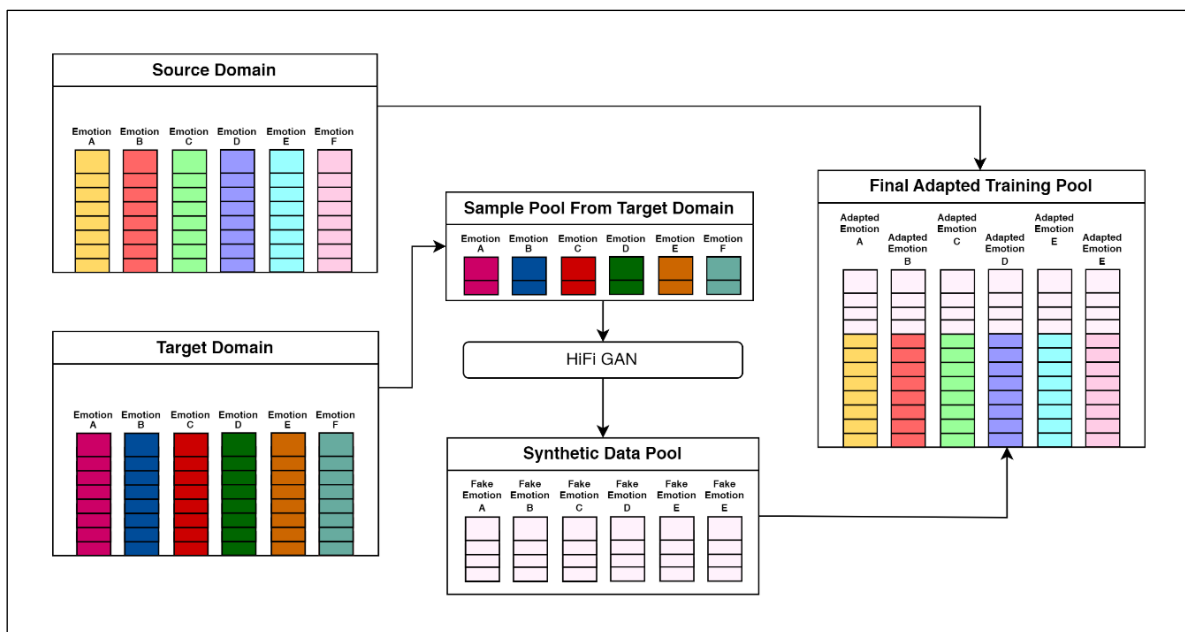


Figure 6. Domain adaptation approach

In this study, the HiFi-GAN model takes a total of 25% of each emotion class from the target domain to be reproduced as synthetic data which then will be adapted with source domain. This study also uses two different class balancing methods which will also affect the total amount data that will be augmented. This is because the class balancing methods in this study are done before the HiFi-GAN model produces the synthetic data.

Prior to training using the adapted training pool, the data synthesis process needs to be conducted first. The availability of utterances in each class will affect the amount of data that needs to be synthesized, as this study utilizes 25% of the total available data in a particular emotion class. Table 2 presents a recapitulation of the total input data used in each emotion class with different class balancing methods. The column labeled REAL indicates the total utterances available in an emotion class, while the column SYNTHETIC (+sEMO/+sRAV) represents the amount of synthesized data to be generated by adjusting the proportion of available utterances in that emotion class, which is 25% of the REAL data total.

Table 2. Data input total with different class balancing method

Class balancing method	Emotion class	EMO-DB		RAVDESS	
		REAL	SYNTHETIC (+sEMO)	REAL	SYNTHETIC (+sRAV)
Not balanced	Neutral	79	20	96	24
	Disgust	46	12	192	48
	Happy	71	18	192	48
	Sad	62	16	192	48
	Angry	127	32	192	48
	Fearful	69	18	192	48
Undersampling	Neutral, disgust, happy, sad, angry, and fearful	46	12	96	24
Oversampling	Neutral, disgust, happy, sad, angry, and fearful	127	32	192	48

## 2.4. Preprocessing

The preprocessing done in this research was carried out through several stages. In the initial stage, the number of utterances in each emotion class needs to be balanced to avoid underrepresented classes which can have a negative impact on model performance [25]. In this study, the imbalance class handling method was carried out using the undersampling, and oversampling method.

The next preprocessing stage is to concatenate the entire speech signal taken from the final adapted training pool for each emotion class into one long signal. This was done because previous studies have been shown to significantly improve the performance of the SER model [26]. Signal concatenation needs to be done before other preprocessing stages are carried out to ensure that the entire speech signal will have a uniform and continuous character.

To ensure that high-frequencies signals are not interfered with random noise, pre-emphasis is performed. Pre-emphasis is carried out mainly on data with quite diverse variations, as well as data that has uncertain signal-to-noise-ratio (SNR) consistency [27]. In the case of cross-corpus, it is safe to say that pre-emphasis is one of the most crucial stages to ensure the model can get optimal results.

Apart from pre-emphasis, another preprocessing stage that can help improve the quality of sound signal data is root mean square (RMS) normalization. RMS Normalization is carried out with the aim of adjusting the overall level of the sound signal [28]. The RMS normalization process is done by calculating the RMS value of the sound signal and using this value to adjust the entire signal until it reaches a target value.

The final step in preprocessing is to perform framing and windowing. In this step, the concatenated speech signal is divided into smaller segments, referred to as frames. Each frame is then subjected to feature extraction, where these features will be used as input to the model for training and classification. In this research, the length of each frame of the audio signal that is cut is 2 seconds. To ensure that there is continuity between the segments of each frame, a windowing function is used. The windowing function used in this step is the Hanning window [29].

The signal features used in this study consist of five commonly used speech signal features in several previous SER studies [17], [30]. These features include Spectral Contrast, Tonnetz, Chromagram, Mel-Spectrogram, and mel-frequency cepstral coefficients (MFCC). These five features, once extracted, are concatenated into a one-dimensional array by calculating the mean of each value along the time axis, and then stacked into a one-dimensional array. However, previous research has shown that the order in which these features are stacked can significantly impact the classification model's performance [30]. Therefore, based on the research by Tanoko and Zahra [30], this study uses the most optimal stacking order found, which is Spectral Contrast, Tonnetz, Chromagram, Mel-spectrogram, followed by MFCC. Table 3 shows the number of coefficients taken from each speech feature.

Table 3. Total number of different features extracted for each frame

Speech feature	Number of coefficient ( <i>n</i> )
Spectral contrast	7
Tonnetz	6
Chromagram	12
Mel-spectrogram	128
MFCC	40

## 2.5. Classification model development

This study utilizes the CNN model used by Issa *et al.* [17] in their 2020 study to perform the classification of 7 emotion classes on the RAVDESS dataset with an accuracy rate of 82.86%. The

implemented model accepts stacked feature arrays with a size of 193. The first convolutional layer uses a filter size of 256 with kernel size of 5 and stride of 1. Then it is taken to the next convolutional layer which uses a filter size of 128 with a kernel size 5 and stride of 1. A dropout layer with a dropout rate of 0,1 is then added followed by a max pooling layer with a pool size of 8. Then, the next convolutional layer uses a filter size of 128 with a kernel size of 5 with stride of 1. Then followed by another dropout layer with a dropout rate of 0,2. Then, the output is then flattened using a flatten layer which then followed by a dense layer with the size of 5 that represents the number of emotion class that the model will try to classify. The topology of the model used in this study is shown in Figure 7.

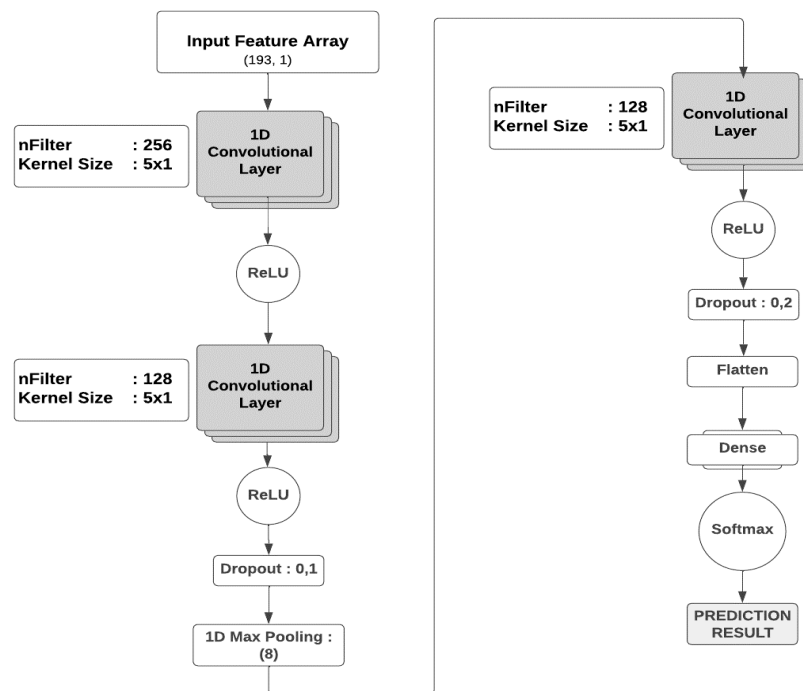


Figure 7. The topology of the classification model

### 3. RESULTS AND DISCUSSION

#### 3.1. Results

Based on the results of the experiments performed, each model tested achieved different performance metrics shown in Table 4 that show the effects of implementing the domain adaptation method proposed in this study. The result shows that the with use of the domain adaptation, all the test results show an improvement on the accuracy of the classification model on cross-corpus SER. The results show that the best performance improvement achieved in the case of adapting EMO-DB to a RAVDESS as a source domain which shows a 28.77% increase in unweighted accuracy in performing 6 emotion classification. The average unweighted accuracy of cross-corpus SER shown in Table 5 also shows that the domain adaptation approach used in this research generally improves the classification performance of cross-corpus SER. The best performance improvement average of cross-corpus SER across three different class balancing methods reaches an 18.293% increase in the case of RAVDESS as a source domain with adapted EMO-DB as a target domain.

#### 3.2. Discussion

Among the three class balancing methods tested, the domain adaptation method generally improves the accuracy of the classification model. However, it is observed that this domain adaptation method works best by having more data available in target domain since the HiFi-GAN model takes 25% of data from the target domain to be adapted into the source domain. One downside to this method is where the target domain needs to have pre-existing data to be taken in order to work. Further research may explore dynamic data handling using this domain adaptation method where the target domain data can start at zero and is added progressively to simulate a real-world situation where the data size may change dynamically.

Table 4. Model evaluation results

Class balancing method	Source domain	Target domain	Evaluation metrics					Domain adapted?	UA difference (%)	
			Unweighted accuracy	Weighted accuracy	Precision	Recall	F1 score			
Not balanced	EMO	EMO	78.17	78.67	79.45	78.17	78.27	NO	+12.23	
	EMO	RAV	25.31	24.48	25.22	25.31	25.91	NO		
	EMO	RAV	37.54	36.94	37.66	37.54	36.85	YES		
	(+sRAV)	RAV	RAV	51.01	49.18	50.09	51.01	50.09		NO
	RAV	EMO	40.49	39.84	40.44	40.49	40.54	NO		
	RAV	EMO	53.41	51.92	52.84	53.41	53.34	YES		
Undersampling	EMO	EMO	81.29	81.90	82.63	81.29	81.60	NO	+10.88	
	EMO	RAV	24.64	24.54	24.76	24.64	24.89	NO		
	EMO	RAV	35.52	35.39	35.29	35.52	34.88	YES		
	(+sRAV)	RAV	RAV	61.31	60.65	60.89	61.31	60.56		NO
	RAV	EMO	34.21	34.14	35.26	35.21	35.23	NO		
	RAV	EMO	47.82	46.83	47.61	47.82	47.49	YES		
Oversampling	EMO	EMO	80.30	79.76	80.34	80.30	80.21	NO	+13.39	
	EMO	RAV	35.82	35.98	36.31	35.82	35.78	NO		
	EMO	RAV	49.21	49.38	50.46	49.21	49.78	YES		
	(+sRAV)	RAV	RAV	52.63	52.28	53.46	52.63	52.90		NO
	RAV	EMO	41.64	40.73	43.77	41.64	41.90	NO		
	RAV	EMO	70.41	69.04	70.31	70.41	70.09	YES		

EMO is EMO-DB corpus, RAV is RAVDESS corpus, and (+sEMO/+sRAV) is marks the source domain that has been adapted with synthetic data from the EMO/RAV domain.

Table 5. Average unweighted accuracy difference of cross-corpus SER with domain adaptation applied

Source domain	Target domain	Average UA difference (%)
EMO	RAV	+12.667
RAV	EMO	+18.433

#### 4. CONCLUSION

This study aims to explore the impact of the implementing a domain adaptation method by utilizing a GAN model in order to tackle the performance degradation of the cross-corpus SER problem. Based on the results, it can be concluded that the domain adaptation approach used in this study could generally improve the performance of cross-corpus SER ranging from 10.88% to 28.77%, with the highest average performance increase reaching 18.433% which was achieved in a test where RAVDESS was used as the source domain and EMO-DB as the target domain.

Although the research results show significant performance improvement, this study has several limitations. One limitation is that it relies on two corpora with labeled data availability for the training process. Future research could explore the development by incorporating unlabeled data. Additionally, the data used in this study was recorded using optimal methods in controlled environments. Therefore, testing on data simulating real-world scenarios, including environmental ambiance and noise, could be conducted for further research.

#### ACKNOWLEDGEMENT

The research presented in this paper, titled "The Use of Generative Adversarial Network as a Domain Adaptation Method for Cross-Corpus Speech Emotion Recognition", was made possible through the financial support provided by Bina Nusantara University.




#### REFERENCES

- [1] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, Mar. 2018, doi: 10.1007/s10772-018-9491-z.
- [2] Z. Xiao, D. Wu, X. Zhang, and Z. Tao, "Speech emotion recognition cross language families: Mandarin vs. Western Languages," in *PIC 2016 - Proceedings of the 2016 IEEE International Conference on Progress in Informatics and Computing*, IEEE, Dec. 2017, pp. 253–257, doi: 10.1109/PIC.2016.7949505.
- [3] M. Li *et al.*, "Contrastive unsupervised learning for speech emotion recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, Jun. 2021, pp. 6329–6333, doi:






- 10.1109/ICASSP39728.2021.9413910.
- [4] S. Ramakrishnan and I. M. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, Mar. 2013, doi: 10.1007/s11235-011-9624-z.
  - [5] G. Elbanna, N. Scheidwasser-Clow, M. Kegler, P. Beckmann, K. El Hajal, and M. Cernak, "BYOL-S: Learning Self-supervised Speech Representations by Bootstrapping," in *Proceedings of Machine Learning Research*, 2021, pp. 25–47.
  - [6] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proceedings - IEEE International Conference on Multimedia and Expo*, IEEE, 2003, pp. 1401–1404, doi: 10.1109/ICME.2003.1220939.
  - [7] O. U. Kumala and A. Zahra, "Indonesian Speech Emotion Recognition using Cross-Corpus Method with the Combination of MFCC and Teager Energy Features," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, pp. 163–168, 2021, doi: 10.14569/IJACSA.2021.0120422.
  - [8] H. Tang, S. M. Chu, M. Hasegawa-Johnson, and T. S. Huang, "Emotion recognition from speech via boosted Gaussian mixture models," in *Proceedings - 2009 IEEE International Conference on Multimedia and Expo, ICME 2009*, IEEE, Jun. 2009, pp. 294–297, doi: 10.1109/ICME.2009.5202493.
  - [9] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, Jan. 2019, doi: 10.1016/j.bspc.2018.08.035.
  - [10] S. M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, IEEE, Sep. 2015, pp. 125–131, doi: 10.1109/ACII.2015.7344561.
  - [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 International Symposium on Chinese Spoken Language Processing, ISCSLP 2021*, IEEE, Jan. 2021, pp. 1–5, doi: 10.1109/ISCSLP49672.2021.9362058.
  - [12] M. Neumann and N. G. Thang Vu, "Cross-lingual and Multilingual Speech Emotion Recognition on English and French," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, Apr. 2018, pp. 5769–5773, doi: 10.1109/ICASSP.2018.8462162.
  - [13] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3693–3697, 2018, doi: 10.21437/Interspeech.2018-1883.
  - [14] S. Latif, J. Qadir, and M. Bilal, "Unsupervised Adversarial Domain Adaptation for Cross-Lingual Speech Emotion Recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019*, 2019, pp. 732–737, doi: 10.1109/ACII.2019.8925513.
  - [15] B. H. Su and C. C. Lee, "A Conditional Cycle Emotion Gan for Cross Corpus Speech Emotion Recognition," in *2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings*, IEEE, Jan. 2021, pp. 351–357, doi: 10.1109/SLT48900.2021.9383512.
  - [16] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.
  - [17] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, May 2020, doi: 10.1016/j.bspc.2020.101894.
  - [18] 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, doi: 10.1371/journal.pone.0196391.
  - [19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology*, ISCA: ISCA, Sep. 2005, pp. 1517–1520, doi: 10.21437/interspeech.2005-446.
  - [20] J. Wang and J. Jiang, "Learning Across Tasks for Zero-Shot Domain Adaptation from a Single Source Domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6264–6279, Oct. 2022, doi: 10.1109/TPAMI.2021.3088859.
  - [21] K. Kumar *et al.*, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
  - [22] K. Ito and L. Johnson, "The LJ Speech Dataset." [Online]. Available: <https://keithito.com/LJ-Speech-Dataset/>. (Date accessed: Jan. 01, 2024).
  - [23] H. Zen *et al.*, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISCA: ISCA, Sep. 2019, pp. 1526–1530, doi: 10.21437/Interspeech.2019-2441.
  - [24] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vetk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019, doi: 10.7488/DS/2645.
  - [25] V. S. Spelman and R. Porkodi, "A Review on Handling Imbalanced Data," in *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018*, IEEE, Mar. 2018, pp. 1–11, doi: 10.1109/ICCTCT.2018.8551020.
  - [26] F. J. Charpentier and M. G. Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers, 1986, pp. 2015–2018, doi: 10.1109/icassp.1986.1168657.
  - [27] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," in *Canadian Conference on Electrical and Computer Engineering*, IEEE, 1995, pp. 1062–1065, doi: 10.1109/cece.1995.526613.
  - [28] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 155–166, Feb. 2005, doi: 10.1109/TMM.2004.840604.
  - [29] F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978, doi: 10.1109/PROC.1978.10837.
  - [30] Y. Tanoko and A. Zahra, "Multi-feature stacking order impact on speech emotion recognition performance," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 6, pp. 3272–3278, Dec. 2022, doi: 10.11591/eei.v11i6.4287.

**BIOGRAPHIES OF AUTHORS**

**Muhammad Farhan Fadhil**    is a graduate student currently studying in Bina Nusantara University under the Department of Computer Science majoring in computer science. His research interest includes speech technology such as speech emotion recognition and signal processing. He can be contacted at email: [muhammad.fadhil009@binus.ac.id](mailto:muhammad.fadhil009@binus.ac.id).



**Amalia Zahra**    is a lecturer at the Master of Computer Science, Bina Nusantara University, Indonesia. She received her bachelor degree in computer science from the Faculty of Computer Science, University of Indonesia (UI) in 2008. She does not have a master degree. Her Ph.D. was obtained from the School of Computer Science and Informatics, University College Dublin (UCD), Ireland in 2014. Her research interests cover various fields in speech technology, such as speech recognition, spoken language identification, speaker verification, and speech emotion recognition. Additionally, she also has interest in natural language processing (NLP), computational linguistics, and machine learning. She can be contacted at email: [amalia.zahra@binus.edu](mailto:amalia.zahra@binus.edu).