

Data augmentation and enhancement for multimodal speech emotion recognition

Jonathan Christian Setyono, Amalia Zahra

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

Article Info

Article history:

Received Oct 18, 2022

Revised Jan 14, 2023

Accepted Mar 10, 2023

Keywords:

Attention mechanism

Data augmentation

Deep learning

Imbalanced dataset

Pre-trained transformer model

Speech emotion recognition

Transfer learning

ABSTRACT

Humans' fundamental need is interaction with each other such as using conversation or speech. Therefore, it is crucial to analyze speech using computer technology to determine emotions. The speech emotion recognition (SER) method detects emotions in speech by examining various aspects. SER is a supervised method to decide the emotion class in speech. This research proposed a multimodal SER model using one of the deep learning based enhancement techniques, which is the attention mechanism. Additionally, this research addresses the imbalanced dataset problem in the SER field using generative adversarial networks (GAN) as a data augmentation technique. The proposed model achieved an excellent evaluation performance of 0.96 or 96% for the proposed GAN configuration. This work showed that the GAN method in the multimodal SER model could enhance performance and create a balanced dataset.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Jonathan Christian Setyono

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

Bina Nusantara University

Jakarta, Indonesia

Email: jonathan.setyono@binus.ac.id

1. INTRODUCTION

Humans are social creatures where one of their fundamental needs is interaction with each other. This study identifies emotions in voice or speech. According to Akçay and Oğuz [1], conversation or speech is one of the most natural ways to express oneself to humans. Hence it naturally becomes one of the types of interactions that is analyzed using computer technology. For this reason, it is a significant problem to determine the emotions inside a speech. In a publication by Sarma *et al.* [2], humans expressed their emotions in speech through implicit or indirect ways such as intonation or tone of voice. The speech emotion recognition (SER) method detects emotions in speech by examining various aspects. SER is a way to explore the human emotional state by using a computer to investigate a speech signal [3]. SER has many implementation cases in our lives, such as in a company call center to detect user satisfaction and in an emergency call center to detect the user's emotional condition so that responders can help provide the correct response and aid users [4].

SER is a supervised method to determine the class of emotions possessed in speech. Based on Akçay and Oğuz [1], there are three main types of SER classifiers which are classical classifiers [5]–[10], classifiers based on deep learning [11]–[14], and deep learning based enhancement techniques [15]–[19]. In this research, the newest type of SER classifier, which is the deep learning based enhancement techniques as SER classification method, is explored because there is still paper regarding this technique recently [20]. SER needs a data source to predict emotions. Three categories of data sources are simulated, elicited, and

spontaneous [21]. Firstly, a simulated dataset means that there are scripts that the actors are obliged to follow. Secondly, an elicited dataset implies that there are scenarios that the actors need to improvise from them. Thirdly, a spontaneous dataset means that the data were collected from real-life situations. A dataset is created from one or more types of data sources. This study used one of the most common English speech datasets which is the interactive emotional dyadic motion capture (IEMOCAP) [21]. IEMOCAP is comprised of two data source types which are simulated and elicited [22]. According to Lieskovská *et al.* [21], IEMOCAP has 10 subjects (5 female and 5 male) with 10,039 speech utterances and 4 modalities which are audio, video, text, and motion capture of face (MOCAP).

In building the SER model, a few aspects must be considered such as the classifier method, the modalities, and the speech data. According to Khalil *et al.* [3], a deep learning SER classifier fuses feature extraction, and feature classification into one phase resulting in a more efficient way than classical classifiers such as artificial neural networks (ANN). Sarma *et al.* [2] proved that using a deep learning classifier and combining it with the attention mechanism could improve the model's accuracy. The model configuration of time delay neural network (TDNN) long short term memory (LSTM) attention improved the weighted accuracy (WA) from 59,5%, which used the TDNN-LSTM model, to 66,3%. To further support the previous work, [23] used a transformer model based on an attention mechanism to achieve the WA of 68,1%. Kumar *et al.* [24] achieved better accuracy by using bidirectional encoder representations from transformers (BERT) as a pre-trained transformer model with the WA of 71,7%. An SER model is unimodal if the model only uses one modal such as audio. On the contrary, a SER model is multimodal if the model uses two or more modals such as audio and text. Based on N and Patil [25], using a multimodal SER could improve the unweighted accuracy (UA) of the IEMOCAP dataset from 65,9% for text-only modal (self-attention-LSTM) to 72,82% using cross-modal attention for multimodal SER which uses text and audio modal. The inputted speech data will determine the SER model's performance. Chatziagapi *et al.* [26] introduced generative adversarial networks (GAN) as a data augmentation technique for SER. Firstly, Chatziagapi *et al.* [26] created an imbalanced dataset scenario on the IEMOCAP dataset by deleting 80% of samples from the angry, sad, and happy emotion classes and retaining the neutral emotion class samples. Then, GAN is used to augment new speech samples to balance the dataset. A balanced dataset means that the count of speech samples for each emotion class is the same for every emotion class. This technique improves the SER model accuracy of unweighted average recall (UAR) from 52,3% to 54,6% and F-score from 52,7% to 55% [26].

Based on the previous studies, two problems are explored. Firstly, in the SER field, there is an imbalanced dataset condition where there are one or more underrepresented emotion classes. Those emotion classes have lesser utterances than the other classes, which resulted in worse performance for the SER model [26]. The second problem is there still is a chance to improve the SER model performance when using deep learning on the IEMOCAP dataset [24]. Therefore, the research is conducted using IEMOCAP as the speech dataset and used one of the deep learning based enhancement techniques which is the attention mechanism inside the multimodal SER model. The proposed work improves the SER model accuracy with three contributions: i) this research uses multimodal SER consisting of audio and text modals instead of using unimodal SER [26], ii) this research inferences new audio files using GAN from raw audio instead of spectrogram [26], iii) this research changes BERT for the text modal inside the SER model [24] into A Lite BERT (ALBERT) and uses GAN as a data augmentation technique.

2. METHOD

2.1. Overview

This work referenced [24] research method of SER phases which used four main steps presented in Figure 1. There are two types of inputs used in this experiment. First, the speech or audio files for the audio modal. Second, the transcript files for the text modal. The final output of the method is the predicted emotion class that will be used to evaluate the model performance.

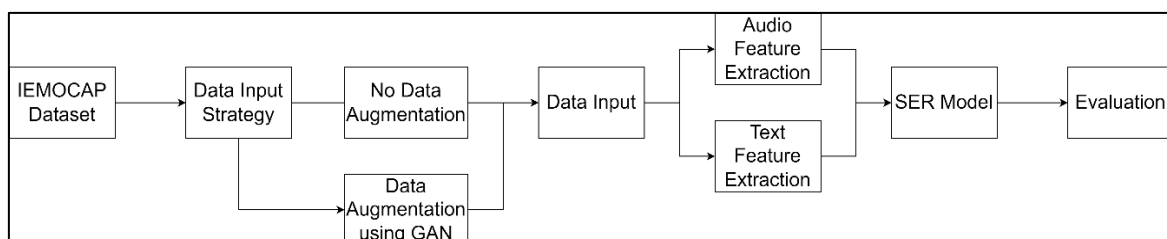


Figure 1. Method overview

Based on Figure 1, the first step is the data input and preprocessing phase. Then, the second step is feature extraction phase for audio and text modal. The third step is the SER model training phase. The fourth or last step is the evaluation phase of the proposed SER model. This proposed work compares multiple data input strategies using no data augmentation technique and GAN as a data augmentation technique to produce SER work that reaches better accuracy.

2.2. Data input and preprocessing

In the data input and preprocessing phase, the IEMOCAP dataset is prepared to be used in the next stage. According to Busso *et al.* [22], the IEMOCAP dataset has ten emotion classes. The ten emotion classes are: Neu=neutral, Hap=happiness, Sad=sadness, Ang=anger, Sur=surprise, Fea=fear, Dis=disgust, Fru=frustration, Exc=excited, and Oth=other [22]. Following the previous works of SER that used the IEMOCAP dataset [20], [24], [26], this research used four emotion classes. The four emotion classes are neutral, happiness (combined with excitement), anger, and sadness. The total samples from the four emotion classes are 5,531, consisting of 1,103 for anger, 1,636 for happiness, 1,708 for neutral, and 1,084 for sadness. There are two flows utilized to prepare the data. The first flow uses data augmentation. The second one does not.

2.2.1. No data augmentation

Three configurations of data input from IEMOCAP dataset are used with no data augmentation which is shown in Table 1. These data configurations used no GAN data augmentation techniques. Hence, every configuration in Table 1 could be said as the original dataset configuration.

Table 1. Data input configurations from IEMOCAP with no data augmentation

Configurations	Total utterances	Samples of each emotion class
Two sessions [24]	2.108	366 Ang, 605 Hap, 746 Neu, 391 Sad
Full samples	5.531	1.103 Ang, 1.636 Hap, 1.708 Neu, 1.084 Sad
4.000 samples	4.000	1.000 Ang, 1.000 Hap, 1.000 Neu, 1.000 Sad

In Table 1, the first configuration taken from [24] used two out of five sessions from the IEMOCAP dataset. Meanwhile, the full samples configuration used all the samples from the IEMOCAP dataset, and the third configuration used 1.000 utterances for every emotion class in the IEMOCAP dataset. Two sessions [24] configuration served as the comparison baseline.

2.2.2. Data augmentation using GAN

This experiment used HiFi-GAN as the data augmentation technique [27]. HiFi-GAN achieved the best mean opinion score (MOS) of 4.36 compared to other GAN techniques such as WaveNet and MelGAN [27]. A better MOS score indicates that the generated audio is more like real audio or human-quality audio. HiFi-GAN also produces audio faster than real-time audio by 167.9 times [27]. The data augmentation steps using HiFi-GAN in this research can be seen in Figure 2.

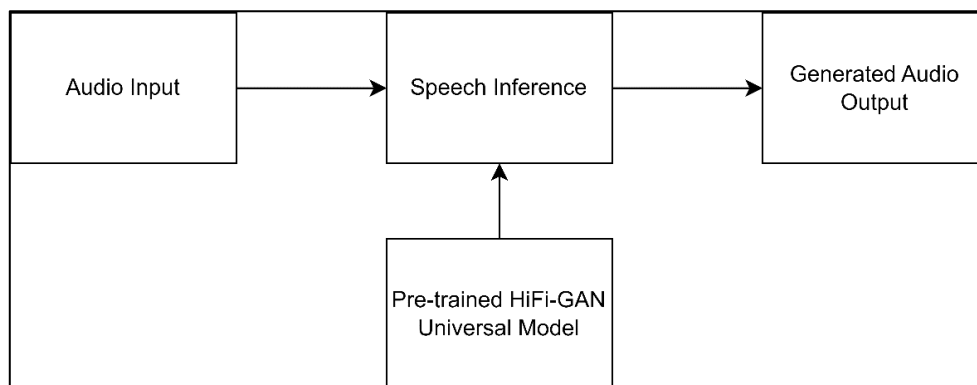


Figure 2. HiFi-GAN data augmentation

Based on Figure 2, the first step of HiFi-GAN data augmentation is selecting the audio input from the IEMOCAP dataset. Six configurations of data input from the IEMOCAP dataset are presented in Table 2. After inputting the data, the second step is to do speech inference. HiFi-GAN provided UNIVERSAL_V1 pre-trained model to be used as transfer learning with the IEMOCAP dataset. In this work, the speech inference used raw audio files with .wav file extension to synthesize audio files.

Table 2. Data input configurations from IEMOCAP with data augmentation using GAN

Configurations	GAN Samples	Original Samples
One class GAN	1,000 samples of one emotion class	1,000 samples of three emotion classes
Two class GAN	1,000 samples of two emotion classes	1,000 samples of two emotion classes
Three class GAN	1,000 samples of three emotion classes	1,000 samples of one emotion class
Four class GAN	1,000 samples of four emotion classes	0 samples of four emotion classes
Two sessions [24]+two class GAN	300 samples of anger & sadness	2.108 samples
Full samples+two class GAN	500 samples of anger & sadness	5.531 samples

The one to four class GAN configurations in Table 2 used the 4,000 sample configuration from Table 1. The GAN technique replaced one or multiple emotion classes with generated samples. Meanwhile, the two last configurations used the two sessions [24] and full samples configurations from Table 1. The GAN technique duplicated the original data for the two last configurations.

2.3. Feature extraction

Following data insertion, the feature extraction followed the selected features by [24]. There are two types of feature extraction for the multimodal SER model. Firstly, audio feature extraction for the audio modal. Secondly, text feature extraction for the text modal. Based on Kumar *et al.* [24], the three audio features that are used for the multimodal SER model are 128-dimensional mel-spectrogram, 40-dimensional mel-frequency cepstral coefficients (MFCC), and 12-dimensional chroma vectors. Peeters [28] indicated that using the chroma vectors feature, a model could capture the regularity of a speech utterance that cannot be captured by using spectral features such as mel-spectrogram and MFCC.

As the text source, [24] used transcription files from the IEMOCAP dataset. The files contain each spoken word or utterance that is used as the input data. Then, the transcription files are cleaned from stopwords. Pre-trained transformer model of BERT is used to extract the text features [24]. This work also added the use of ALBERT to compare the multimodal SER performance. According to Lan *et al.* [29], ALBERT has three differences from BERT that makes it a better model. The advantages are factorized embedding parameterization, cross-layer parameter sharing, and inter-sentence coherence loss. Besides improving performance, ALBERT also has a smaller model, which leads to lower graphical processing unit (GPU) or tensor processing unit (TPU) usage and faster training speed [29]. BERT or ALBERT produced the text feature as an encoded input tokens vector with two special tokens of classification (CLS) and separator (SEP). This experiment used BERT base and ALBERT base configuration [29].

2.4. Multimodal SER model training

Before the model training, the data from the previous phase is divided into two parts. The train data used 80%, and the test data used 20% of the total data. Every configuration is trained with 100 epochs and used early stopping when the model accuracy does not improve. Apart from two sessions [24]'s that used 64 batch sizes, the other configurations used 128 batch sizes, which is the most optimal parameter according to the hyperparameter tuning. The training of the multimodal SER model consisted of three parts [24], as shown in Figure 3.

Based on Figure 3 as proposed in [24], in the audio modal, the audio or SER phase accepts mel-spectrogram, MFCC, and chroma vectors from the audio feature extraction step. The inputs are processed separately using gated recurrent unit (GRU) and attention mechanism to extract the parts that contain the most significant emotional information. The speech vectors are utilized in the multimodal emotion recognition phase. The text emotion recognition phase accepts the encoded speech token generated by BERT or ALBERT to produce encoded hidden vectors that will be used in the multimodal emotion recognition phase. The last part, which is the multimodal emotion recognition phase, processed each of the outputs from the audio and text modal by concatenating them into one final vector. The final vector will determine the predicted emotion class. The Adam optimizer and the sparse categorical cross-entropy loss function are used to train the multimodal SER model [24].

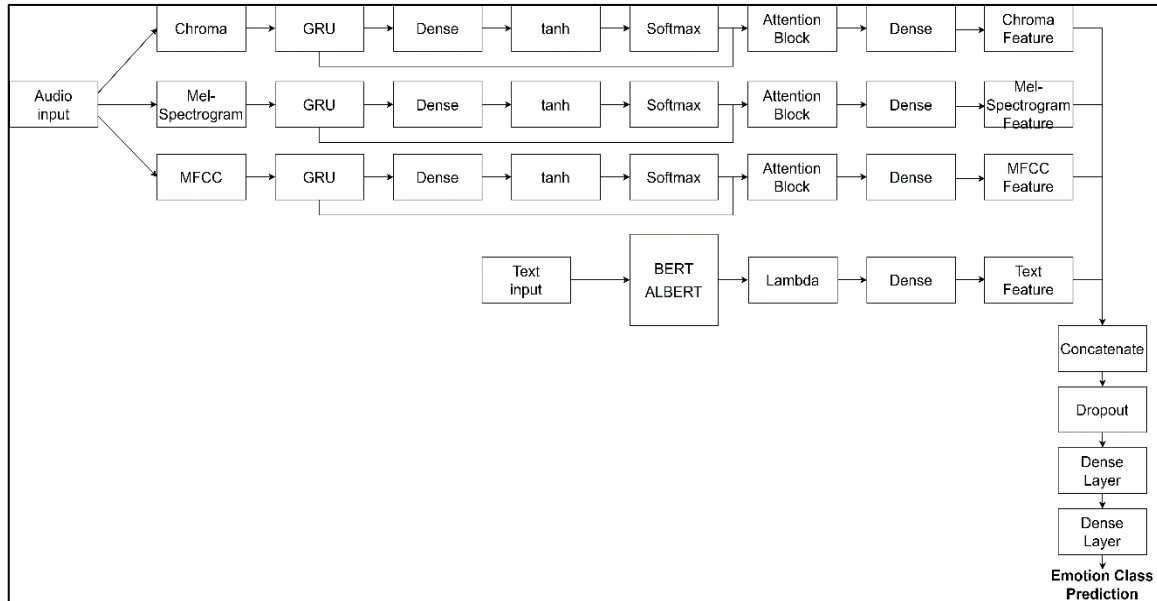


Figure 3. Multimodal SER model

2.5. Evaluation

This research utilized four multi-class evaluation metrics because the proposed SER model used the IEMOCAP dataset to classify four emotion classes. Two evaluation metrics used are accuracy and F1-score. F1-score are calculated using the unweighted average. Meanwhile, accuracy is scored by using both the weighted and unweighted average. There are two steps in the evaluation phase. First, each configuration is scored by the chosen evaluation metrics. The second step is to analyze the evaluation score by comparing every data input configuration.

3. RESULTS AND DISCUSSION

3.1. HiFi-GAN data augmentation result

As mentioned, HiFi-GAN is utilized in this work to either replace or duplicate the audio data from the IEMOCAP dataset. Figure 4 compares the mel-spectrogram of the real audio data and the augmented data from HiFi-GAN. The mel-spectrogram is taken for one of the samples from the anger emotion class from the IEMOCAP dataset.

Figure 4(a) shows the original mel-spectrogram. Meanwhile, Figure 4(b) shows the generated mel-spectrogram using HiFi-GAN data augmentation. Based on Figure 4(b) can be found that HiFi-GAN [27] could produce an audio file that is similar to the real human-quality audio of Figure 4(a). Both displayed figures used the Ses01F_impro01_F012_anger speech file. Hence, the data augmentation technique using HiFi-GAN to tackle an event where the data source is imbalanced is an option for the SER field.

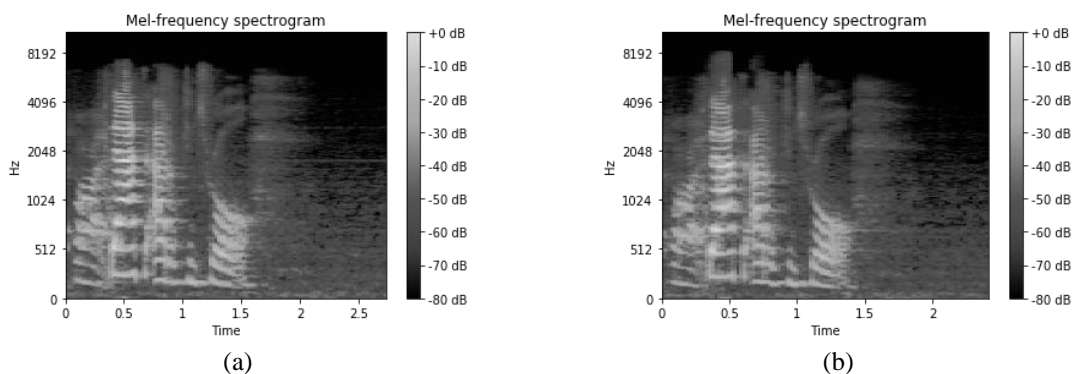


Figure 4. Comparison between (a) real audio and (b) HiFi-GAN generated audio

3.2. SER model training time

This proposed work used both BERT and ALBERT in the text modal of the SER model. Based on Lan *et al.* [29], using ALBERT instead of BERT could improve the training time of the SER model. This claim is valid in this research. By choosing ALBERT, the training time per epoch is shorter than BERT, as presented in Table 3.

According to Table 3, ALBERT has an 18 seconds training time per epoch. Meanwhile, BERT has a 12 seconds training time per epoch. Hence, ALBERT has a faster training time per epoch by 6 seconds compared to BERT. This could save a lot of time when training bigger SER models.

Table 3. Training time per epoch for ALBERT and BERT

Text modal configuration	Training time per epoch (seconds)
ALBERT base	12
BERT base	18

3.3. SER model performance

Based on the completed training, this research proved that the use of GAN as a data augmentation technique could improve multimodal SER performance. The SER performance evaluation can be seen in Table 4. The shown results are the best performance for each configuration.

Aside from the two sessions [24]'s configuration, the results presented in Table 4 displayed the best performance achieved by each configuration. The SER configuration of two class GAN attained the best overall performance with a score of 0.96 or 96% for every evaluation metric. This means that the configuration achieved great multi-class classification performance for the IEMOCAP dataset. This configuration used GAN for the Happiness and Sadness emotion classes. Meanwhile, data from the IEMOCAP dataset are used for anger and neutral. The obtained score improves the weighted and UA performance from two sessions [24] significantly. The configuration resulted in an improvement of WA of 0.25 or 25% and an UA of 0.22 or 22%. The SER configuration of three class GAN came second. The configuration has each evaluation metrics' score of 0.94 or 94%. This configuration used GAN for three emotion classes except for the happiness emotion class. But the use of GAN in this research cannot replace the IEMOCAP data completely as presented by the seventh SER configuration that achieved only the score of 0.73 or 73%.

Table 4. Multimodal SER model performance comparison for accuracy

No	SER configuration	Transformer model	Use of GAN	GAN percentage (%)	Weighted accuracy	Unweighted accuracy	F1-score
1	Two sessions [24]	BERT	No	-	0.71	0.74	0.74
2	Full samples	ALBERT	No	-	0.69	0.72	0.72
3	4.000 samples	ALBERT	No	-	0.73	0.72	0.72
4	One class GAN	ALBERT	Yes	100	0.88	0.89	0.89
5	Two class GAN	ALBERT	Yes	100	0.96	0.96	0.96
6	Three class GAN	ALBERT	Yes	100	0.94	0.94	0.94
7	Four class GAN	ALBERT	Yes	100	0.73	0.73	0.73
8	Two sessions [24]+two class GAN	BERT	Yes	100	0.72	0.71	0.71
9	Full samples+two class GAN	ALBERT	Yes	50	0.76	0.76	0.76

The last two SER configurations created a balanced dataset by duplicating a certain percentage of the selected emotion classes. The twelfth SER configuration of two sessions [24]+two class GAN could not achieve better performance than the original SER configuration of two sessions [24] by having a lower score of 0.03 or 3% for UA, precision, recall, and F1-score. On the other side, the full samples+two class GAN SER configuration could attain higher performance than the full samples configuration. The UA score increased by 0.07 or 7%, and the other evaluation metrics' scores increased by 0.04 or 4%.

Based on the result, the researchers thought that the increase in performance by using the GAN technique happened because the GAN method creates a new audio file that had a few differences from the original audio file from the IEMOCAP dataset, which can be seen in Figure 4. This phenomenon is caused by the pre-trained HiFi-GAN model, which learned the features of speech before the inference phase using the selected IEMOCAP data. Hence, this captures pattern similarity generally and generates samples of an emotion class that has a few differences from the original IEMOCAP samples.

4. CONCLUSION

This study aimed to explore the use of the data augmentation technique, which is GAN, to replace or duplicate emotion class samples from the IEMOCAP dataset to tackle the imbalanced data problem. From the result, this research can conclude that the GAN method in the multimodal SER model could enhance performance. The best SER model configuration, which is two class GAN for the Happiness and Sadness emotion classes with ALBERT, achieved WA, UA, and an F1-score of 0.96 or 96%. Likewise, it is crucial to have a balanced dataset when creating the SER model and could be attained by using GAN as a data augmentation method. Further works can explore the use of GAN on other speech emotion datasets, and the use of other GAN configurations. Furthermore, the explanation of the SER model performance increase using GAN as a data augmentation method is still not explored.

ACKNOWLEDGEMENTS

The author thanks Bina Nusantara University for the funding obtained in conducting this research.




REFERENCES

- [1] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.
- [2] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion Identification from Raw Speech Signals Using DNNs," in *Interspeech 2018*, Sep. 2018, doi: 10.21437/interspeech.2018-1353.
- [3] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/access.2019.2936124.
- [4] Mustaqeem and S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019, doi: 10.3390/s20010183.
- [5] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion Recognition in Speech Using Neural Networks," *Neural Computing and Applications*, vol. 9, no. 4, pp. 290–296, Dec. 2000, doi: 10.1007/s005210070006.
- [6] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, doi: 10.1109/icassp.2003.1202279.
- [7] M. Borchert and A. Dusterhoft, "Emotions in Speech - Experiments with Prosody and Quality Features in Speech for Use in Categorical and Dimensional Emotion Recognition Environments," in *2005 International Conference on Natural Language Processing and Knowledge Engineering*, doi: 10.1109/nlpke.2005.1598724.
- [8] C. Busso, S. Lee, and S. Narayanan, "Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009, doi: 10.1109/tasl.2008.2009578.
- [9] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Multi-stage classification of emotional speech motivated by a dimensional emotion model," *Multimedia Tools and Applications*, vol. 46, no. 1, pp. 119–145, Jul. 2009, doi: 10.1007/s11042-009-0319-3.
- [10] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011, doi: 10.1016/j.specom.2011.06.004.
- [11] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1–2, pp. 7–19, Dec. 2009, doi: 10.1007/s12193-009-0032-6.
- [12] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, Sep. 2014, doi: 10.21437/interspeech.2014-57.
- [13] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, doi: 10.1109/icassp.2016.7472669.
- [14] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2016, doi: 10.1109/slt.2016.7846319.
- [15] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, doi: 10.1109/acii.2013.90.
- [16] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards Speech Emotion Recognition 'in the Wild' Using Aggregated Corpora and Deep Multi-Task Learning," in *Interspeech 2017*, Aug. 2017, doi: 10.21437/interspeech.2017-736.
- [17] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2017, doi: 10.1109/icme.2017.8019296.
- [18] S. E. Eskimez, Z. Duan, and W. Heintzman, "Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, doi: 10.1109/icassp.2018.8462685.
- [19] S. Sahu, R. Gupta, G. Sivaraman, and C. Espy-Wilson, "Smoothing Model Predictions Using Adversarial Training Procedures for Speech Based Emotion Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, doi: 10.1109/icassp.2018.8462065.
- [20] S. Amiriparian *et al.*, *On the Impact of Word Error Rate on Acoustic-Linguistic Speech Emotion Recognition: An Update for the Deep Learning Era*. 2021.
- [21] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulik, "A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism," *Electronics*, vol. 10, no. 10, pp. 1–29, May 2021, doi: 10.3390/electronics10101163.




- [22] C. Busso *et al.*, “IEMOCAP: interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Nov. 2008, doi: 10.1007/s10579-008-9076-6.
- [23] L. Tarantino, P. N. Garner, and A. Lazaridis, “Self-Attention for Speech Emotion Recognition,” in *Interspeech 2019*, Sep. 2019, doi: 10.21437/interspeech.2019-2822.
- [24] P. Kumar, V. Kaushik, and B. Raman, “Towards the Explainability of Multimodal Speech Emotion Recognition,” in *Interspeech 2021*, Aug. 2021, doi: 10.21437/interspeech.2021-1718.
- [25] K. D. N and A. Patil, “Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks,” in *Interspeech 2020*, Oct. 2020, doi: 10.21437/interspeech.2020-1190.
- [26] A. Chatziagapi *et al.*, “Data Augmentation Using GANs for Speech Emotion Recognition,” in *Interspeech 2019*, Sep. 2019, doi: 10.21437/interspeech.2019-2561.
- [27] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis.” arXiv, Oct. 23, 2020, doi: 10.48550/arXiv.2010.05646.
- [28] G. Peeters, “Musical key estimation of audio signal based on hidden markov modeling of chroma vectors,” 2006, pp. 127–131.
- [29] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.” arXiv, Feb. 08, 2020, doi: 10.48550/arXiv.1909.11942.

BIOGRAPHIES OF AUTHORS



Jonathan Christian Setyono    is currently a student at Bina Nusantara University majoring in Computer Science Master Program. He is enrolled in Master Track of Computer Science study program at Bina Nusantara University. He is currently under Mrs. Amalia’s guidance for this research. His research interests include business intelligence and speech recognition. He can be contacted at email: jonathan.setyono@binus.ac.id.



Amalia Zahra    is a lecturer at the Master of Information Technology, Bina Nusantara University, Indonesia. She received her bachelor’s degree in Computer Science from the Faculty of Computer Science, University of Indonesia (UI) in 2008. She does not have a master’s degree. Her PhD 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, speech emotion recognition, and so on. Additionally, she also has interest in natural language processing (NLP), computational linguistics, machine learning, and artificial intelligence. She can be contacted at amalia.zahra@binus.edu.