*International Journal of Innovative Computing, Information and Control*
Volume **19**, Number **6**, December **2023**
 ISSN 1349-4198
pp. **1983–1994**

# STRESS DETECTION OF CHILDREN THROUGH SPEECH SIGNALS IN MULTI-SPEAKER ENVIRONMENT USING DEEP LEARNING

Phie Chyan[1], Andani Achmad[1,*], Ingrid Nurtanio[2] and Intan Sari Areni[1]

[1]Department of Electrical Engineering
[2]Department of Informatics
Hasanuddin University
Jalan Poros Malino Km. 6, Gowa 92171, Indonesia
chyanp21d@student.unhas.ac.id; {ingrid; intan}@unhas.ac.id
*Corresponding author: andani@unhas.ac.id

Received April 2023; revised July 2023

Abstract. *Stress is a psychological problem that can affect anyone, including children. Detecting stress in children is a complex problem because they are generally unaware of the psychological problems they are experiencing and have verbal limitations that affect their communication skills with their parents. One of the biomarkers that can be used to detect stress is the voice (speech signal). The use of speech in stress detection has advantages in terms of convenience for the subject and ease of acquisition. This study proposes a stress detection model through speech signals in a multi-speaker environment. This model accepts audio input from the classroom environment, where there is noise and many speakers' voices overlap. The audio acquired is then separated using a speech separation algorithm based on an RNN architecture, producing output as segregated speech. The speech is then extracted for features and fed to the stress detection model based on CNN architecture, which predicts the speaker's stress status. The experimental results show that the proposed model is capable of speech separation with up to five speakers and predicts the stress status of the subject with an average accuracy of 95.6%.*
**Keywords:** Stress detection, Child mental stress, Speech separation, Speech signal, Deep learning

1. **Introduction.** Psychological stress is a change in a person's emotional state as a response to the pressures faced in everyday life. According to the psychological review, stress or distress is a condition that triggers a person to express negative emotions such as anger, sadness, panic, fear, and anxiousness [1]. Several studies related to psychological stress show that prolonged stress is closely correlated with decreased cognitive ability, motivation, decision-making skills, and spatial awareness of a person [1-4].

Children are a group that is also vulnerable to stress. Based on data from WHO (World Health Organization), the global prevalence of children with stress-related health problems is estimated to be around 13 percent of the population of children worldwide which, when observed statistically, is more or less balanced with the prevalence of mental problems in adults, which reaches 20 percent of the adult population [5]. This fact may indicate that mental problems experienced in childhood can be carried over into adulthood if no intervention is made to address the underlying problems. The main problem with stress detection in children is that detecting stress in children is more challenging than detecting stress in adults. Children, especially in pre-and early school age, generally have limited verbal abilities, which limit their ability to communicate with their parents or caregivers about the various problems they face; moreover, a child is generally not aware of stress

DOI: 10.24507/ijicic.19.06.1983

itself as a psychological condition that can have a negative effect on them [6,7]. Stressors or stimuli that trigger stress in children can come from any source, such as family problems, difficulty following lessons, to bullying problems at school. Stress that occurs at this age which is still classified as a golden age, if left untreated, can cause problems that interfere with the growth and development of children [8,9].

Voice (speech) is one of the biomarkers that can be used to detect a state of psychological stress because voice output is a psychophysiological response resulting from the cooperation of approximately one hundred muscles connected by a network of cranial nerves and spinal nerves and is an integrative part of the psychophysiological stress system in humans [10,11]. Psychological stress experienced by a person affects how the body works, such as the emergence of muscular tension, increased respiratory rate, and excessive saliva production, which then affect voice reproduction, such as voice pitch, intonation, speech prosody, and many other sound parameters [12]. Using voice signals to detect stress eliminates the need for sensors or devices to be worn on the body, making it more convenient and safe for children. Voice can also be easily obtained from the microphone, making it possible to build a large database to support the stress detection model. However, the process of selecting and extracting sound features needed to support the model is crucial to obtain good accuracy [13].

Many studies on stress detection through speech have been conducted in recent years. However, this study uses voice acquired in a controlled or stationary environment, where each subject's voice is recorded individually with minimal or no ambient sound [10,14-19]. Although this method can achieve high accuracy, it is unlikely to identify the activity or event that triggers the stressful state (stressor) because the moment of the stressful event might have passed by the time the subject's voice was acquired. In addition, acquiring individual subject voices will increase the number of voice recordings that must be stored and analyzed. To overcome these problems, this study proposes a stress detection model that acquires voice directly from the environment where children learn and play. This environment is generally a non-stationary environment where many sound sources come from the surroundings, for example, the voices of other children or adults and noise from the environment. The solution to the problems presented is the primary focus of this study, which is to detect stress through the speech acquired from a noisy multi-speaker environment where the speaker's voices overlap.

Deep learning is used in a wide range of fields due to its ability to analyze large amounts of data and extract meaningful patterns. With the support of deep learning, areas such as computer vision and signal processing are experiencing very rapid development [20]. In this study, we use a deep learning approach to separate speech signals obtained from a multi-speaker environment. The results of separated speech are in the form of segregated speech from each child in the acquired sound recordings. Then we extract the speech samples to find discriminant features in detecting stress and use these features to train a model that can predict the subject's stress status. This proposed model, which is capable of detecting stress in a multi-speaker non-stationary environment, is the main contribution of this study. The results of this study help caregivers or parents better understand the stress that occurs in children and the potential underlying stressors so that stress mitigation can be carried out more effectively.

The remainder of the paper is organized as follows. Section 2 discusses the methodology used in this study, and the proposed model. Section 3 presents the experimental results, and Section 4 concludes this paper.

## 2. **Model and Methodology.**

To detect stress through sound acquired from a multi-speaker environment, we propose a system model as shown in Figure 1.

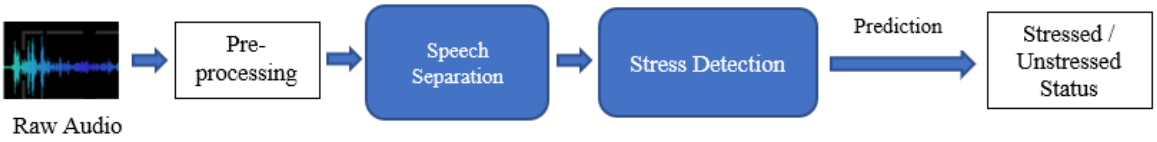


FIGURE 1. Model system diagram

Stress detection using speech in real-world applications requires the acquisition of sound directly from the environment of everyday human interaction. This method aims to get an overview of the causes (stressors) of the stress experienced. Regarding the interaction environment for children, the classroom at school is where children spend time. This environment is generally non-stationary because there are many sound sources with different frequency properties. Besides that, speech is naturally a form of non-stationary signal. In an interactive environment like this, there is more than one subject speaking simultaneously, so a speech separation approach is needed to deal with this problem. The problem referred as the cocktail party problem is the ability to track the voice of each specific subject when many subjects speak simultaneously, especially in an environment with noise [21,22]. Although this problem is easy for humans because the human senses can separate signals originating from multiple sources and focus on recognizing and tracking one particular source, it is a challenging task for computers [23].

In this study, we acquired sound directly from the classroom through a single-channel high-gain microphone placed in the middle of the room. The classroom size is 16 $m^2$ with five 1st grade elementary school students who are the subjects involved in this study. All the students and their parents agreed to participate in this study. This research was assisted by a child psychologist who designed and implemented various activities using the Trier Social Stress Test method, which aims to emulate stress caused by stressors children often face in everyday life at school [2,24]. Each child will do activities that involve public speaking and answering challenging math problems. Throughout the activity, the psychologist will observe the behaviour and gestures of each child for signs of stress. Before and after the activity, a saliva sample from each child will be taken to measure whether there is an increase in the hormone cortisol (stress hormone). Based on the theory, the cortisol hormone will increase 9 to 15 times from normal baseline levels during periods of stress and will last for several hours afterwards [25]. The results of the psychologist's observations, which were double-validated with the subject's cortisol hormone count, were used to label each child's speech as stressed or unstressed on the audio recording.

We use a deep learning approach to perform speech separation, producing segregated speech from each subject. Each sound recording is then cut into sound samples between 1 and 2 seconds long. After the data validation, we obtained 500 sound samples, each consisting of 223 samples labelled stressed and 277 samples labelled unstressed. Furthermore, for a more detailed discussion, Section 2.1 will explain in detail the speech separation model we use, and Section 2.2 will elaborate on the stress detection model in detail.

2.1. **Speech separation.** Before being fed into the speech separation module, the speech acquired first goes through the preprocessing phase for noise reduction and silence removal. The model shown in Figure 2 is derived from the latest development of the speech separation model based on dual path RNN (Recurrent Neural Network) [26-28]. Input in the form of an audio signal (waveform) containing speech mixture $x \in R^T$ is entered into the encoder network and will produce an output of $N$ dimensional $z$ of size $T' = (2T/L) - 1$ where $T$ is input length and $L$ is the encoding compression factor, which then produce Formula (1)

$$z = E(x) \tag{1}$$

where $E$ is a 1D convolutional layer network with kernel size of $L$ and stride of $L/2$, subsequent by non-linear ReLU (Rectified Linear Unit) activation function, $x$ is speech mixture from audio signal, and $z$ is resulting latent representation. The 3D tensor $\nu = [\nu_1, \ldots, \nu_R] \in R^{N \times K \times R}$ is obtained by concatenating each chunk on a single dimension. Then $\nu$ will be sent to the separation network $Y$, which is made up of $b$ RNN blocks [27]. The even blocks $B_{2i}$ will be used along chunk dimensions of size $K$, while odd blocks $B_{2i-1}$ will be used along the time dependent dimension. The RNN block used contains Multiply and Concat (MULCAT) blocks with two sub networks. It then uses two separate bidirectional LSTM (Long Short-Term Memory) networks to multiply its outputs and combines its inputs to produce a module output using Formula (2)

$$B_i(\nu) = P_i([M_i^1(\nu) \odot M_i^2(\nu), \nu]) \tag{2}$$

where $P_i$ is a learned linear project that converts an input's dimension ($\nu$) into the dimension of the output obtained by concatenating the product of the two LTSMs denoted by $M_i^1$ and $M_i^2$, $\odot$ is the element-wise product operation, and $B_i$ is resulting RNN block. An overview of the RNN block pairs can be seen in Figure 3. In this method after each pair of blocks processed requires model to reconstruct the original audio. The 3D tensor via PreLU initialized at 0.25. The decoder is a $1 \times 1$ convolution with a $C$ output channel. Each pair of blocks will be decoded using the same PreLU and decoder parameters. We use the add and overlap operators to convert back the tensor to audio. This operator is used to reverse the chunking process and add overlapping frames from the signal.

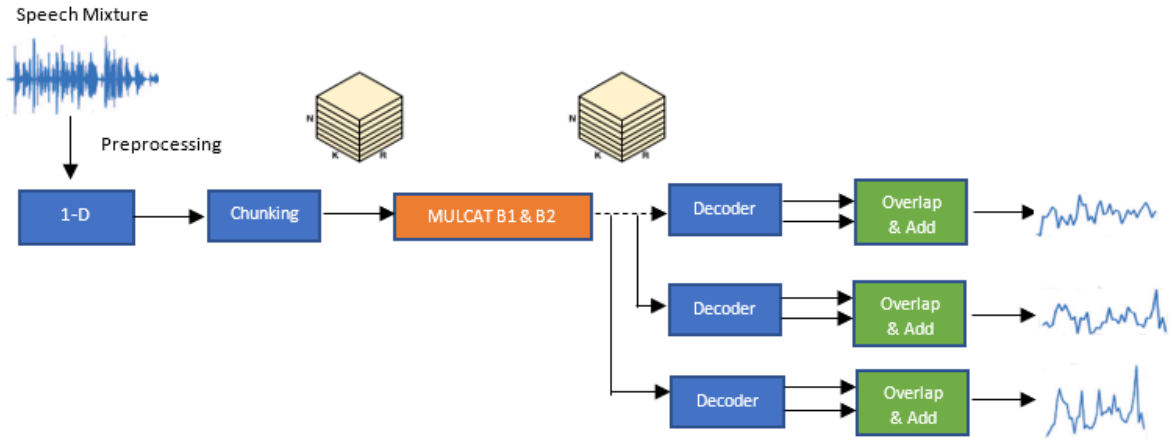


FIGURE 2. Speech separation model

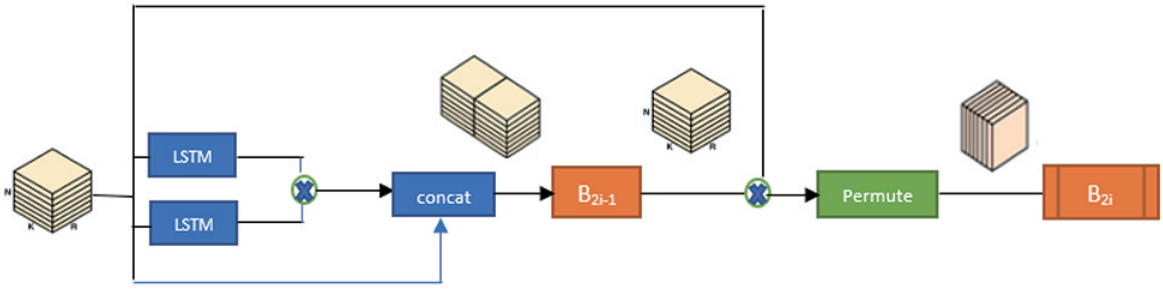


FIGURE 3. Multiply and Concat (MULCAT) block

For the speech separation model in this non-stationary environment, we use the open-source dataset LibriMix, derived from LibriSpeech (clean speech) and WHAM! Noise. We generate two to five speech mixtures, namely LibriMix-2 to LibriMix-5. Each dataset consists of 20 hours of training and 10 hours of validation and test sets. Each speech separation model will be trained with these datasets, which correlate with the number of possible output channels (number of students in the class).

Because the number of speakers from the sound acquired can vary up to 5 speakers, the selection process begins by using a model trained using the largest number of sounds $C$, which is five. The detection of each output channel is carried out. If an empty channel (silence) is obtained, the model selection is repeated using the model with the $C-1$ output channel. The process will continue to be repeated until all the channels are filled, representing the correct $C$ output model for the speech mixture. Figure 4 shows a flowchart of the model selection algorithm to match the number of speakers in the audio mixture. For implementation, we use batch size 2 with ADAM optimizer. Input kernel size 4. The architecture uses 6 MULCAT blocks where each LSTM layer contains 128 neurons. To evaluate the model, the Si-SNRi metric (Scale-invariant Signal-to-Noise Ratio improvement) is used to measure the quality of the speech separation results [29].

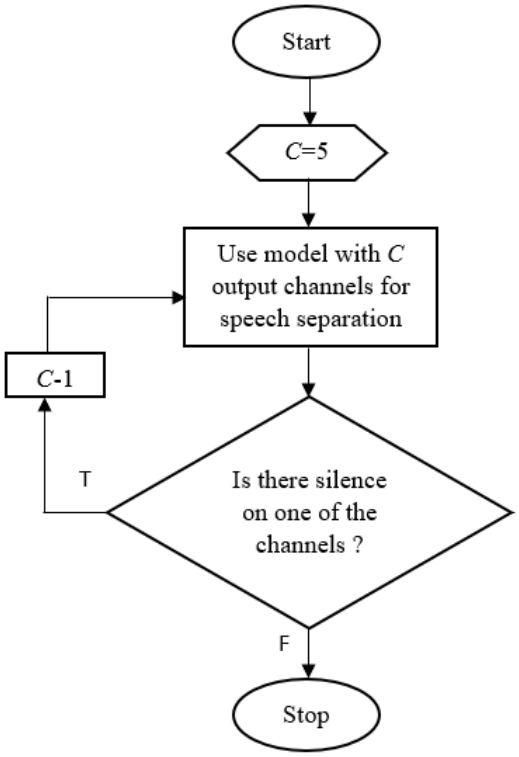


FIGURE 4. Flowchart of model selection to match the number of speakers in audio sample

2.2. **Stress detection.** To build this stress detection model, in addition to using our own built dataset, we also use audio data from the open source TESS (Toronto Emotional Speech Set) and SAVEE (Surrey Audio-Visual Expressed Emotion) datasets, which contain 2,800 and 480 audio files, respectively. Both datasets consist of audio files that represent various emotional speeches. Following the review of the psychological theory that stress is a change in psychological reactions that are manifested in various negative emotions [30], based on this theory, we conduct a relabeling process for the dataset used.

The original dataset files labelled sad, angry, disgusted, and fearful were relabeled with the stress label, while files labelled pleasant, pleasantly surprised and neutral were relabeled with the unstressed label. After relabeling and adding our own built dataset, 1,973 data were labelled with stressed labels and 1,807 with unstressed labels.

To enrich the dataset with better generalization capabilities for various noise signal disturbances, we perform a data augmentation process to create new synthetic sample data. Each original audio file is added with two additional variations. The first variation is the original audio file with additional noise injection. The second variation is the original audio with a modified pitch. Thus, 11,340 sound sample files are obtained after this augmentation. Figure 5 shows the waveform of the original audio data and the two additional different versions resulting from the augmentation process.

For the model to perform the classification process, it is necessary to perform feature extraction to convert the audio signal to a format the model can comprehend. Our stress

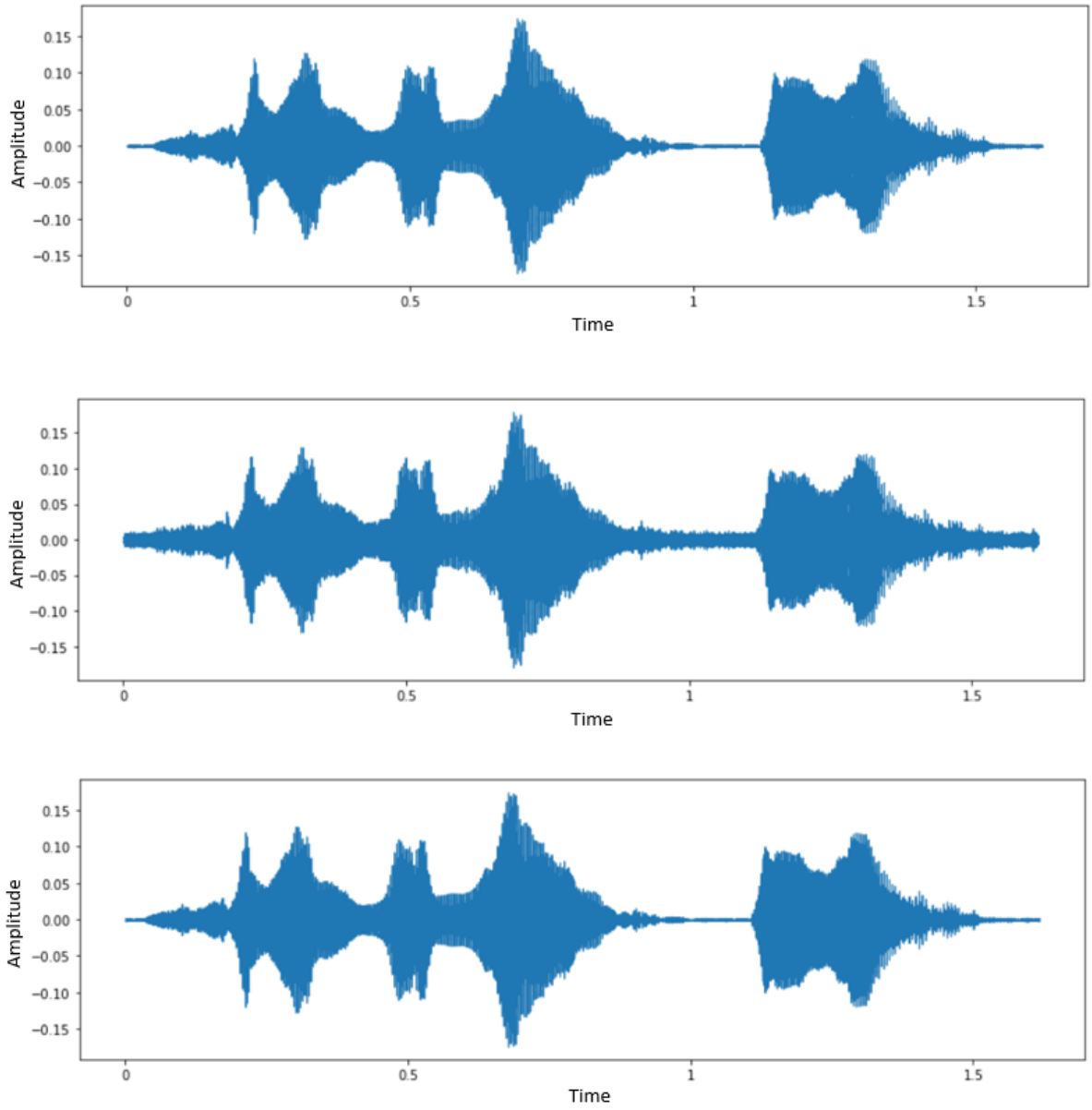


FIGURE 5. The original speech waveform (top), noise added modification (mid), and pitch shifting modification (bottom)

detection model, as shown in Figure 6, is based on the CNN (Convolutional Neural Network) architecture with the Convolution 1D (Conv 1D) sequential model, which consists of 4 convolution layers and two dense layers followed by a flattened layer. Each convolution layer has a max pooling layer to get maximum pattern variations. From each audio sample, various features in the signal domain are extracted related to the speech signal's attributes. This signal domain consists of the time, frequency, and cepstral domain [31,32]. Time domain features refer to characteristics or properties of an audio signal that are derived from the waveform. The frequency domain feature refers to the analytic space in which mathematical functions or signals are conveyed in terms of frequency, rather than time, which result from conversion from the time domain using the Fourier transforms. The cepstral domain is a feature in the cepstral domain obtained by the inverse Fourier transform of the logarithm spectrum of the signal. Cepstrum is often used as a feature vector to represent human voices and musical sounds. So to get the properties of speech, cepstral is a feature widely used, such as for speech recognition needs.

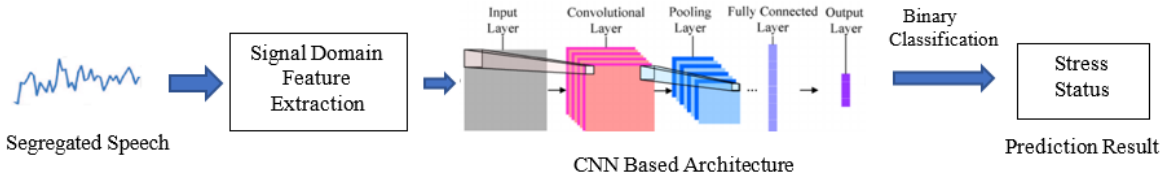


FIGURE 6. Stress detection model

For feature extraction needs to be fed to the model, we will experiment using the cepstral domain features and also a combination of the cepstral domain features with the time domain and frequency domain features to see if there is an increase in accuracy to be obtained by combining the three signal domain features. Table 1 shows the audio features used in each signal domain for feature extraction.

TABLE 1. The audio features used in signal domain

| **Signal domain feature** | **Audio features used** |
|---|---|
| Time domain | Amplitude Envelope (AE) |
| | Zero Crossing Rate (ZCR) |
| | Root Mean Square (RMS) |
| Frequency domain | Spectral Centroid (SC) |
| | Spectral Rolloff (SR) |
| | Spectral Bandwidth (SB) |
| Cepstral domain | Mel Spectogram |

The model was trained using a 75 : 25 random training – testing split for the training phase. The softmax activation model is used in a dense layer with two neurons. CNN was trained for 40 epochs with a batch size of 32 using the Adam optimizer. In addition to calculating the model's accuracy, three additional metrics are used to evaluate model performance: Precision, Recall, and F1 Score. The Precision (P) metric compares True Positives (TP) and the number of data predicted to be positive. Recall (R) compares the True Positive (TP) and the number of positive data. The F1 Score (F1) is the harmonic average of precision and recall.

3. **Experiment Result.** This section will explain the experimental results, performance evaluation of the model's speech separation and stress detection.

3.1. **Speech separation model result.** We evaluate the speech separation model using the SI-SNRi metric to see the quality of the sound separations obtained. The model trained with the sound sample $C_m$ can train audio samples with the number of speakers $C$ as long as $C_m \geq C$. Suppose the number of speakers in the model used exceeds the number of speakers in the audio sample. In that case, the separation results will produce an unused channel containing a silent signal. Utilizing a model that corresponds to the number of speakers in the sound sample will provide more optimal performance; therefore, for sound samples when $C$ is unknown, the model selection algorithm outlined in Section 2 will be utilized, which will select the right model based on channel activity detection. Table 2 shows the obtained SI-SNRi values. Based on these results, the separation performance is better when the model is used on par with the number of speakers in the audio sample.

TABLE 2. SI-SNRi score of various models used for the number of speakers in the audio sample

| Used model | The number of speakers in the audio sample | | | |
|---|---|---|---|---|
| | 2 | 3 | 4 | 5 |
| 2-speaker model | 17.5 | – | – | – |
| 3-speaker model | 12.3 | 13.6 | – | – |
| 4-speaker model | 9.6 | 10.8 | 9.7 | – |
| 5-speaker model | 6.5 | 8.7 | 8.4 | 7.5 |

3.2. **Stress detection model result.** We tested the model using each feature in the signal domain. Besides that, we also tried to use a combination of the three signal domain (time, frequency and cepstral) features to see whether combining all signal domain features for the model can improve the accuracy of the stress detection model. The dataset used for the training and test procedures combines the open source dataset and our own built dataset, which aims to provide the model with better generalization ability and applicability in various speech acquisition environments. The results obtained in Figure 7 show that the cepstral domain feature is a discriminant feature for stress detection with an average accuracy of 95.6% and an average F1 Score of 94.8%. The result also confirms that cepstral-based audio features such as the Mel spectrogram correlate with how the human ear perceives sound and are very good at representing various speech signal properties. Based on the results, audio features in the time and frequency domains are less accurate for sound stress detection. Likewise, when the two feature groups are combined with cepstral-based audio features, they do not significantly affect the model's performance.

We also conducted tests to compare the performance of models trained using open-source datasets and trained using our own built dataset. Both models were trained using cepstral domain features as extracted features. The results, as shown in Figure 8, show that using an open-source dataset produces model performance with an average accuracy of 97.4% and an average F1 Score of 96.5% which is significantly better than using our own built dataset with an average accuracy of 88.2% and an average F1 Score of 90%.

The difference in performance obtained is mainly because the open source dataset used comes from audio samples acquired in a controlled room where there is no noise or other sound sources, whereas, in our own built dataset, the audio was acquired directly from the child's learning environment which is a room that is relatively noisy with multiple speakers can talk simultaneously. In addition, the performance of the speech separation model also influences the performance of the stress detection model because the speech

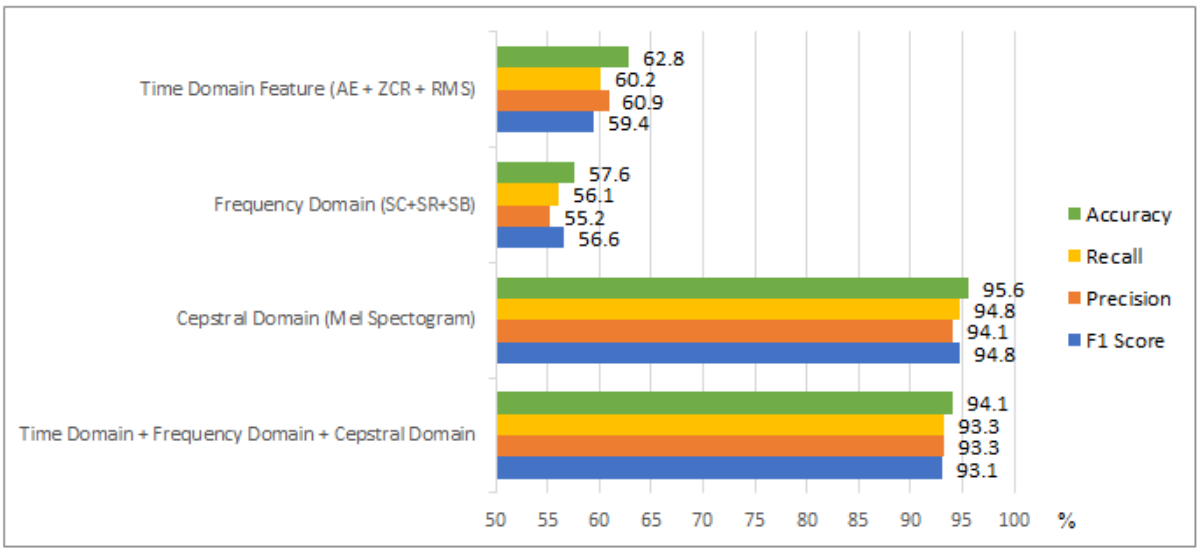


FIGURE 7. Model evaluation results using various audio signal domain features

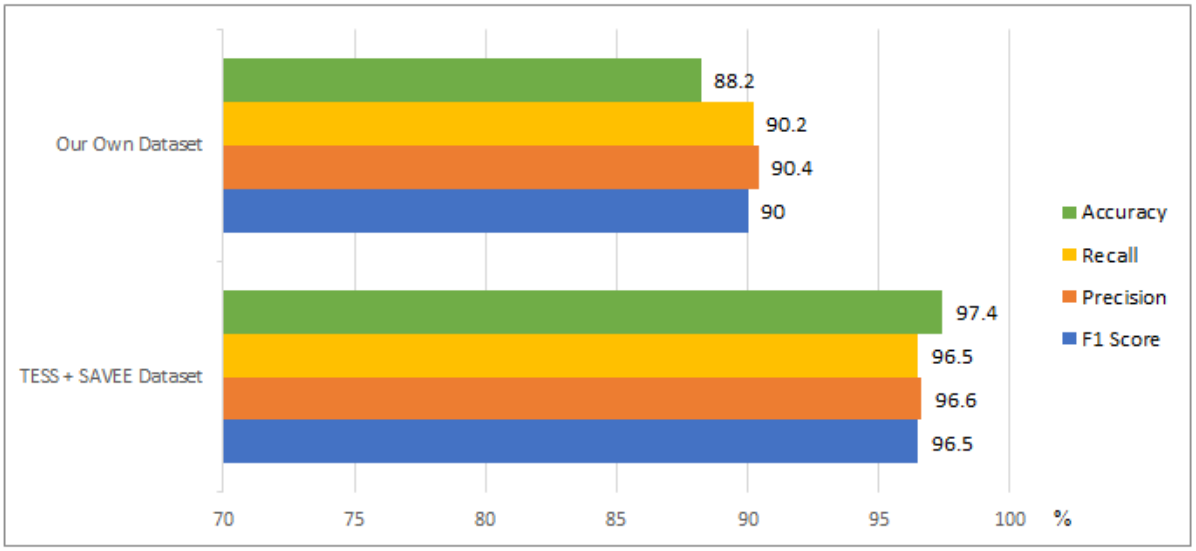


FIGURE 8. The comparison between the model's performance evaluation using open source dataset and our own built dataset

sample in our own built dataset is derived from segregated speech as a product of the speech separation model performed.

4. **Conclusion.** This study proposes a stress detection model through speech in a multi-speaker environment. Audio is acquired directly in the children's learning environment, a non-stationary environment with much noise and many speakers' voices overlapping. The proposed model performs speech separation to produce segregated sounds. The results of the segregated voice samples then are extracted using cepstral audio features that are discriminant in detecting stress. Then, the results are fed to the model to predict the stress status of the subject. Based on the experimental results, it was found that the proposed model can detect stress in subjects with high accuracy. Using a combination of open source datasets and our own built dataset, we obtain an average accuracy of 95.6% and an average F1 Score of 94.8%.

We have yet to achieve real-time detection in this proposed model due to high computational costs, especially in speech separation. Also, for the same reason, the number of overlapping voices that can be separated in this model is a maximum of 5 in the acquired voice sample. In future work, we will explore the possibility of optimizing the speech separation algorithm so that the model can achieve real-time detection, which allows the model to detect and monitor the psychological condition of the subject directly through conversation.

**Acknowledgment.** The grants from Indonesia's Ministry of Education, Culture, Research, and Technology's "Penelitian Disertasi Doktor" Scheme 2023 support this work.

## REFERENCES

[1] J. A. Healey and R. W. Picard, Detecting stress during real-world driving tasks using physiological sensors, *IEEE Trans. Intell. Transp. Syst.*, 2005.
[2] S. Wemm and E. Wulfert, Effects of acute stress on decision making, *Physiol. Behav.*, vol.176, no.3, pp.139-148, 2017.
[3] P. Morgado and J. Cerqueira, The impact of stress on cognition and motivation, *Front. Behav. Neurosci.*, 2018.
[4] H. Yaribeygi, Y. Panahi, H. Sahraei, T. P. Johnston and A. Sahebkar, The impact of stress on body function: A review, *EXCLI J.*, vol.16, pp.1057-1072, 2017.
[5] World Health Organization (WHO), *Mental Health*, 2021.
[6] Y. Choi, Y. M. Jeon, L. Wang and K. Kim, A biological signal-based stress monitoring framework for children using wearable devices, *Sensors (Switzerland)*, vol.17, no.9, pp.1-16, 2017.
[7] T. Y. Kim, L. Měsíček and S. H. Kim, Modeling of child stress-state identification based on biometric information in mobile environment, *Mob. Inf. Syst.*, vol.2021, 2021.
[8] M. Bucci, S. S. Marques, D. Oh and N. B. Harris, Toxic stress in children and adolescents, *Adv. Pediatr.*, vol.63, no.1, pp.403-428, 2016.
[9] N. Garmezy, A. S. Masten and A. Tellegen, The study of stress and competence in children: A building block for developmental psychopathology, *Child Dev.*, vol.55, no.1, pp.97-111, 1984.
[10] G. M. Slavich, S. Taylor and R. W. Picard, Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations, *Stress*, vol.22, no.4, pp.408-413, 2019.
[11] S. Paulmann, D. Furnes, A. M. Bøkenes and P. J. Cozzolino, How psychological stress affects emotional prosody, *PLoS One*, vol.11, no.11, pp.1-21, 2016.
[12] K. Pisanski and P. Sorokowski, Human stress detection: Cortisol levels in stressed speakers predict voice-based judgments of stress, *Perception*, vol.50, no.1, pp.80-87, 2021.
[13] P. Chyan, A. Andani, I. Nurtanio and I. Areni, A deep learning approach for stress detection through speech with audio feature analysis, *The 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE2022)*, pp.269-273, 2022.
[14] K. Tomba, J. Dumoulin, E. Mugellini, O. A. Khaled and S. Hawila, Stress detection through speech analysis, *Proc. of the 15th International Joint Conference on e-Business and Telecommunications (ICETE2018)*, 2018.
[15] C. A. Jason and S. Kumar, An appraisal on speech and emotion recognition technologies based on machine learning, *Int. J. Recent Technol. Eng.*, vol.8, no.5, pp.2266-2276, 2020.
[16] A. König et al., Measuring stress in health professionals over the phone using automatic speech analysis during the COVID-19 pandemic: Observational pilot study, *J. Med. Internet Res.*, vol.23, no.4, pp.1-14, 2021.
[17] N. Matsuo, S. Hayakawa and S. Harada, Technology to detect levels of stress based on voice information, *Fujitsu Sci. Tech. J.*, vol.51, no.4, pp.48-54, 2015.
[18] H. Han, K. Byun and H. G. Kang, A deep learning-based stress detection algorithm with speech signal, *Proc. of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia (AVSU'18)*, pp.11-15, 2018.
[19] J. Kejriwal, Š. Beňuš and M. Trnka, Stress detection using non-semantic speech representation, *2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*, pp.1-5, 2022.
[20] S. M. Noe, T. T. Zin, P. Tin and I. Kobayashi, Automatic detection and tracking of mounting behavior in cattle using a deep learning-based instance segmentation model, *International Journal of Innovative Computing, Information and Control*, vol.18, no.1, pp.211-220, 2022.

[21] Y. Qian, C. Weng, X. Chang, S. Wang and D. Yu, Past review, current progress, and challenges ahead on the cocktail party problem, *Front. Inf. Technol. Electron. Eng.*, vol.19, no.1, pp.40-63, 2018.
[22] Y. Li, F. Wang, Y. Chen, A. Cichocki and T. Sejnowski, The effects of audiovisual inputs on solving the cocktail party problem in the human brain: An fMRI study, *Cereb. Cortex*, vol.28, no.10, pp.3623-3637, 2018.
[23] J. R. Hershey, Z. Chen, J. Le Roux and S. Watanabe, Deep clustering: Discriminative embeddings for segmentation and separation, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2016)*, pp.31-35, 2016.
[24] M. A. Vallejo, L. Vallejo-Slocker, E. G. Fernández-Abascal and G. Mañanes, Determining factors for stress perception assessed with the Perceived Stress Scale (PSS-4) in Spanish and other European samples, *Front. Psychol.*, vol.9, no.1, 2018.
[25] K. E. Hannibal and M. D. Bishop, Chronic stress, cortisol dysfunction, and pain: A psychoneuroendocrine rationale for stress management in pain rehabilitation, *Phys. Ther.*, vol.94, no.12, pp.1816-1825, 2014.
[26] Z. Chen, Y. Luo and N. Mesgarani, Deep attractor network for single-microphone speaker separation, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2017)*, vol.2, no.1, pp.246-250, 2017.
[27] Y. Luo, Z. Chen and T. Yoshioka, Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation, *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2020)*, pp.46-50, 2020.
[28] E. Nachmani, Y. Adi and L. Wolf, Voice separation with an unknown number of multiple speakers, *The 37th Int. Conf. Mach. Learn. (ICML2020)*, pp.7121-7132, 2020.
[29] A. Wijayakusuma, D. R. Gozali, A. Widjaja and H. Ham, Implementation of real-time speech separation model using time-domain audio separation network (TasNet) and dual-path recurrent neural network (DPRNN), *Procedia Comput. Sci.*, vol.179, pp.762-772, 2021.
[30] E. S. Epel et al., More than a feeling: A unified view of stress measurement for population science, *Front. Neuroendocrinol.*, vol.49, no.12, pp.146-169, 2018.
[31] H. P. Shi, J. H. Cao and X. Liu, Blind source separation for non-stationary signal based on time-frequency analysis, *Proc. of 2011 4th Int. Conf. Intell. Networks Intell. Syst. (ICINIS2011)*, pp.45-48, 2011.
[32] A. Défossez, G. Synnaeve and Y. Adi, Real time speech enhancement in the waveform domain, *arXiv.org*, arXiv: 2006.12847, 2020.

## Author Biography

**Phie Chyan** received the bachelor's degree in Electrical Engineering from Atma Jaya Makassar University, Indonesia, 2004; the master of Computer Science from Gadjah Mada University, Jogjakarta, Indonesia, 2011. His main research interests include computer vision, multimedia processing, and expert system. Now he is pursuing his Ph.D. degree in Hasanuddin University. He is the staff of Department of Informatics, Faculty of Information Technology, Atma Jaya Makassar University.

**Andani Achmad** received a bachelor's degree in 1986, a master's degree in 2000 and a doctorate in 2010 from Hasanuddin University in Indonesia. His primary topic of study is Electrical Power Engineering. He is a Professor at Hasanuddin University's Department of Electrical Engineering. He is current head of the doctorate program in electrical engineering, Hasanuddin University.

**Ingrid Nurtanio** received the bachelor's degree in Electrical Engineering from Hasanuddin University, Makassar, Indonesia in 1986. She received her master of technology from Hasanuddin University, Makassar, Indonesia in 2002. She received her doctoral degree from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia in 2013. Her research interest is digital image processing, computer vision and intelligent system. Currently, she is the staff of Department of Informatics, Faculty of Engineering, Hasanuddin University. She is a member of IAENG and IEEE.

**Intan Sari Areni** received B.E. and M.E. degrees in Electrical Engineering from Hasanuddin University (UNHAS), Makassar (1999) and Gadjah Mada University (UGM), Jogjakarta (2002), respectively, and received a Doctorate degree from Ehime University, Japan in 2013. Currently, she is a Professor at the Department of Electrical Engineering, Hasanuddin University. Her research interests include multimedia signal processing, telecommunication, biomedical engineering, computer vision, and powerline communication system. She is a member of IEEE and IAENG.