Human Vocal Type Classification using MFCC and Convolutional Neural Network

Kriesna B. Pratama
School of Computing
Telkom University
Bandung, Indonesia
kriesnabayup@student.telkomuniversity.ac.id

Suyanto Suyanto
School of Computing
Telkom University
Bandung, Indonesia
suyanto@telkomuniversity.ac.id

Ema Rachmawati
School of Computing
Telkom University
Bandung, Indonesia
emarachmawati@telkomuniversity.ac.id

2021 International Conference on Communication & Information Technology (ICICT) | 978-1-6654-3914-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT52195.2021.9568474

***Abstract*— The range of voices is an essential aspect that a singer needs to know. This knowledge is necessary so that the singer can maximize their singing potential. This study discussed about how to classify someone's vocal range into four classes commonly used in choir using Mel-frequency Cepstral Coefficient (MFCC) for its feature extraction and Convolutional Neural Network (CNN) for the classification. This study emphasized how MFCC and CNN was able to solve human vocal type classification problem. It is assisted by WavAugment for augmentation to maximize the learning process. In this study, the data used were primary so that the data were collected through surveys and experiments conducted directly by the researchers. The data used also affect the classification result, where the data need to be sparse enough to avoid the model being overfitted. The experiment is giving a good result where the training accuracy reaches 91.83% and testing accuracy is 91.14%. This model (specifically the feature extractor) was able to outperform the STFT model that usually has a competitive result with 3.11% in training accuracy and 1.15% in testing accuracy. This study is a multi-disciplinary science that has a strong influence on music, especially in the choir. This study was conducted to improve choir music and computer technology continuity by combining music with computer science.**

***Keywords*— MFCC, CNN, WavAugment, Vocal**

I. Introduction

Singing is prevalent for everyone. Singing is produced from the vocal cords' vibrations, which contain rhythm and tone to form harmony. Every human being has their own way of processing their voice to form a tone so that every one of them has their vocal range. Human vocal type is depending on the extent of the person's vocal range. Human vocal types are grouped into six forms, in which men and women have different classes. These vocal types are usually used in a choir. For female vocal types, it is divided into three classes, namely alto (low), mezzo-soprano (medium), and soprano (high). Meanwhile, male vocals are divided into three classes, namely bass (low), baritone (medium), tenor (high)[1]. This voice division aims to form a harmonization of tones where no notes collide with each other.

Technically, the human voice has a particular frequency. The computer recognizes the voice signal and extracts its frequency. This frequency is expressed in hertz (Hz), where each voice has a varying frequency. Each person's pitch is different which is influenced by the timbre and pitch of the person. The resulting tone frequency can classify human's vocal type into specific classes. Men and women can be classified based on the frequency of voice emitted, and a person's vocal range can be classified in this way.

This research is a multi-disciplinary science where in-depth knowledge of vocal classification is needed to build a capable system of classifying human vocal. The dataset used is primary so that the researcher collects data directly through surveys and experiments on several people who already understand the rules of singing. Several studies have been carried out, especially in classifying voices with various approaches and various data, resulting in a good accuracy[2], [3]. This study aims at how human vocal types can be classified using the commonly used method to classify another data and measure its performance based on accuracy.

II. Literature Review

A. Related Works

Several studies related to human voice processing have been conducted, and some can provide exemplary voice classification performance [4]. However, there is very little study about human vocal type classification, so in this study, we refer to some studies about speech recognition with the same feature extraction method and classification method. In several previous studies (before 2015), sound classification was carried out using Artificial Neural Networks (ANN), considering this method is the most commonly used[5][6].

Research continues to be developed from year to year, and in 2016 Anjali Pahwa tested gender classifications using voice with the Mel-Frequency Cepstral Coefficient (MFCC) as a Feature Extractor with Support Vector Machine (SVM) and Neural Network (NN) as a classifier[4]. This study provided two models which gave different result. The first model considered by the MFCC value gave the result; 96% for women and 90.48% for men. While in the second model that without paying attention to MFCC, it gave a lesser accuracy. They are 95.36% for women and 86.36% for men. MFCC feature extraction is a very popular method to use and many researchers use MFCC as a comparison feature extraction from the models built by the researchers. For example, in the research that pre-trained Convolutional Neural Network (CNN), CNN pre-trained is used as a feature extractor which performance testing is compared to feature extraction from MFCC[7].

More recent research conducted by Toan Pham Van in 2019 examines the classification of a singer's voice from popular music in Vietnam[8]. This study was using the CNN method for classifying the vocalist's voice and MFCC for its feature extraction. This study was giving a good results for the classification, namely 93% for the mean precision. In newer research, MFCC and CNN were still used for signal classification to solved a heart sound classification problem[11]. From those studies, it can be assumed that the current CNN method provides good results as a classifier for audio processing. Audio classification research mainly focuses on its feature extraction where many researchers carry out tests using some other feature extractor such as Short-Time Fourier Transform (STFT)[9] or even their own proposed feature extraction model with classifiers which generally use CNN, ANN, or SVM[10].

The method of classifying human voice itself was also further developed by the researchers. The Convolutional Recurrent Neural Network (CRNN) method was also tested several times to classify human voice. CRNN was tested and compared with the CNN method to measure its accuracy and computation speed[2]. Besides, CRNN itself has also been used as the basis for the method proposed by the researchers, namely Joint Detection and Classification (JDC), to perform music classification tasks[3]. The research that used CRNN method itself showed good results where the accuracy given was able to exceed the accuracy of the CNN method in solving some speech recognition problems.

Based on some of the literature that has been studied, there is still little that highlights the problem of classifying human vowels. Although several studies on classification have been carried out, the classification focuses more on the classification of musical genres or spoken speech than human vocal types[9], [12]. In addition, the tested method in the classification case can still be developed further to improve its accuracy. Therefore, in this study, we propose to test the classification of human vowels using MFCC as the feature extractor and CNN as the classifier.

B. Vocal Type Classification

TABLE I. VOCAL RANGE FOR EACH TYPE OF VOICE[6]

Gender	Vocal Types	Vocal Range	Vocal Range Frequency (Hz)
Male	Bass	E2 – E4	82.41 - 329.63
	Bariton	F2 – F4	87.31 - 349.23
	Tenor	C3 – C5	130.81 - 523.25
Female	Alto	F3 – F5	174.61 - 698.46
	Mezzo-Soprano	A3 – A5	220.00 - 880.00
	Soprano	C4 – C6	261.63 - 1046.50

Based on Table I, the human voice is classified into six types of vocal types in which men and women have different types of voices. This classification is based on how broad a note the person can reach, and usually, the tone used as the benchmark comes from the piano. The chords of the piano produce note that a human can follow, and the lowest and highest pitch limits of a person determine the vocal range of the person.

In this study, we took only four classes of human vocal types: Alto and Soprano for female voice and Bass and Tenor for the male voice. Those four kinds of vocal types are the vocal type that is commonly used in a choir. Meanwhile, Mezzo-Soprano and Baritone are vocal types that are rarely used in the choir, and also, the people from those vocal types usually singing for another role in the choir. For example, people from baritone vocal types can sing for the bass role if the song reachable by those people, or those people can sing in the tenor role if the song if the tone is not too high for them.

III. Proposed Model

A. Scenario Overview

In this section, the researchers will describe the experimental scenario carried out according to Fig. 1. In this experiment, the authors divide the several stages into six parts: data collection, augmentation, feature extraction, data splitting, the learning process through models, and then evaluation of results. The data set was collected through a survey conducted directly by the researchers of several people who can sing. Furthermore, the dataset is augmented to avoid an overfit in the CNN model.

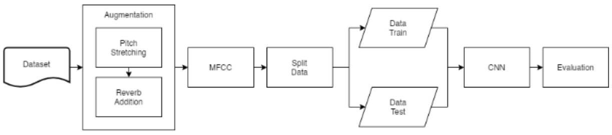

Fig. 1. Scenario Overview

Based on Fig.1, the researchers divide the augmentation process into two parts so that augmentation results are varied through pitch shifting and reverb addition. In pitch shifting, the writer defines the pitch parameter with a limit of 100 to minus 100 so that the data has variation but minimizes overlap between classes. Furthermore, in the reverb addition section, the author uses a parameter value of 50; this is because the author adjusts to the reference source in the research conducted by Facebook in the augmentation process. The original dataset consisted of 83 sound files, but after augmentation of all data, the authors obtained 8,440 datasets.

The next step is feature extraction using the MFCC method. It is divided into six sub-stages: pre-emphasis filtering, frame blocking, windowing, fast Fourier transform, Mel-frequency wrapping, and Discrete Cosine Transform (DCT). Then, the results obtained from the feature extraction process will get the coefficients in acoustic vectors for classification. The researchers divided the train data and test data with a composition of 80% and 20%. Furthermore, the vector-shaped data is processed in the CNN model for the classification process. In the final step, the researcher evaluates the classification results with performance metrics to see the level of accuracy of the model in handling the classification process of human vocal types.

B. Dataset

The data is in the form of people's voices from each vocal type. Due to the lack of similar research, for the result there is no data that is open to the public. Thus, the researcher took the data manually. The data is in the form of a recording in the .wav format. This format is an uncompressed original recording file that made the data augmentation process easier. Thus, the processed data has been directly labeled by people who already have credibility in their fields. Those data were taken online via phone recorder due to Covid-19 pandemics. This method made each data have different lengths and noises. Later in this research, the recording lengths remain uncut within the range of six until ten seconds. The noise was also unremoved to improve the variety of the data.

There are two kinds of data which are; the data taken from the template that has been set before based on table I, and the data taken from the singer's random basic tone data. The data is taken from people from each type of voice who sing do-re-mi-fa-so-la-si-do based on the template but only from one octave. The second kind of data is quite similar to the first data, but the difference is that the data is taken from the basic tone which is comfortable for the singer to increase the variety of the data. Each person chanted the note five until seven times to get about 83 sound files.

C. *Augmentation*

Data augmentation is a process of multiplying data by modifying the data so that the computer detects the results of the augmentation and the original data are a different data. This augmentation is done because the data taken has a limited number of 83 sound files, while the classification model used is CNN which requires many data. The CNN model tends to overfit easily if only a small amount of data is used for the learning process[13]. Thus, the data augmentation process was carried out where each data was augmented separately to balancing the data amount. The total data retrieved by augmentation and the original data is 8,523 data.

In our research, WavAugment was the method that used for data augmentation. WavAugment is a library that implements time-domain augmentation that has been researched by Facebook and published on github.com/facebookresearch/WavAugment[14]. WavAugment using Pytorch- and libsox-based effects to implements transformation on audio. The transformation includes reverb addition and pitch stretching.

D. *Mel-frequency Cepstral Coefficient*

In this study, feature extraction was carried out using Mel-frequency Cepstral Coefficient (MFCC). It is giving output a cepstral coefficient later used as a feature vector for the classification process. MFCC is a feature extraction method commonly used for voice classification problems, and it is commonly used as a comparison method for the proposed models from some studies. This MFCC method is based on the bandwidth and frequency that humans can receive. It cannot perceive frequencies above 1Khz[15]. MFCC uses a voice signal as an input and then processes the signal in several steps.

Step 1: *Pre-Emphasis Filtering*

In this step occurs the process of maintaining high-frequency signals and increasing low-frequency signals. The signal passed through a filter to maintain the high frequencies in those signals, and the signal with a lower frequency was given additional signal energy[4]. This pre-emphasis filter can be calculated using the equation:

$$S'(n) = S(n) - \alpha * S(n-1)$$

In this equation, the α is called the pre-emphasis coefficient which has a value of 0,9 ≤ α ≤ 1.

Step 2: *Frame Blocking*

In this step, the signal that has been through Pre-Emphasis Filtering is being chopped/segmented into blocks called a frame. Those frames are processed in a short time to obtain a stable[6].

Step 3: *Windowing*

Windowing is performed on each frame to minimize signal discontinuity in the beginning and the end of each frame caused by the frame blocking process[15]. The window used is Hamming Window which is multiplied by the frame from the previous process. The equation can calculate hamming Windowing:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$

In this equation, the range of n values is between 0 and N - 1.

Step 4: *Fast Fourier Transform*

Fast Fourier Transform (FFT) converts each frame from the time domain into the frequency domain. In the frequency domain, signals are easier to analyze[14]. The following equations support this statement:

$$Y(w) = FFT\,[h(t) * X(t)] = H(w) * X(w)$$

Step 5: *Mel-Frequency Wrapping*

The voice signal consists of tones of various frequencies, calculated as f and in Hz[16]. Meanwhile, the subjective pitch is measured in "Mel" units. The "Mel" frequency scale is linear for frequencies below 1kHz and logarithmic for frequencies above 1kHz. Mel scale measurement for a predetermined frequency in Hz can use the following formula:

$$mel(f) = 2595 * log\,\log_{10}\left(1 + \frac{f}{700}\right)$$

In this equation, *f* is a linear frequency.

Step 6: *Discrete Cosine Transform (DCT)*

In this step, the log Mel spectrum is converted back into the time domain using DCT. The result is called the Mel-Frequency Cepstral Coefficient[6]. One set of coefficients is called an acoustic vector. The result of this conversion can be calculated using the equation:

$$c(n) = \sum_{k=1}^{K} \log S_k * cos\left(n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right)$$

In this equation, the value of n is the number 1,2… K, while Sk is the output from the filter bank at the k index, and K is the expected coefficient.

E. *Convolutional Neural Network*

In classifying the image, a Neural Network-based method can be applied with the Convolutional Neural Networks(CNN) model, and CNN can provide excellent results in solving classification and image recognition cases[17]. An object's image will be analyzed through several neurons with activation functions, weights, and biases [18]. However, Speech recognition is one of the problems that are quite challenging to apply to the CNN model because there are still not many research activities related to it, especially the classification of human vowels. The CNN model architecture can be illustrated in fig. 2.

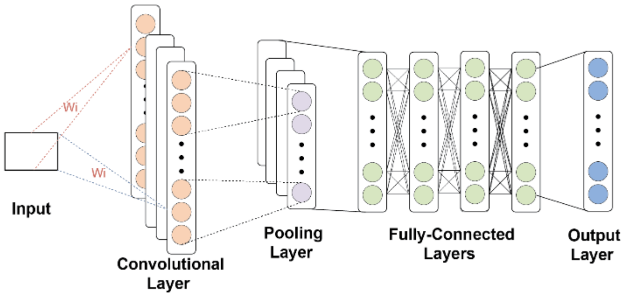

Fig. 2. CNN Architecture [18]

There are three kinds of CNN layers: Convolutional Layer, Pooling Layer, and Fully Connected Layer.

Convolutional layers consist of neurons that are connected to the receptive field of the previous layers. Convolutional Layers contain multiple learnable filters that consist of weights corresponding to the previous layer[19]. Each neuron receives input from a small rectangular section from feature extraction obtained from the previous layer, where the input results are then multiplied by the weight on this layer. The activation function used on this layer is the Rectified Linear Unit (ReLU). The ReLU activation function has competitive results with previous pre-train models[20].

The pooling layer is layer after Convolutional Layer which has the same steps and receptive field. The pooling layer takes inputs from a local section in Convolutional Layer to do down-sampling, producing output from the section[18]. The most frequent pooling layer that is used in classification is Max Pooling. In Max Pooling, the neuron's receptive field from the pooling layer is two-dimensional, giving the maximum output of the receptive field output.

The Convolutional Layer and Pooling Layer are connected from one to another. The output connections from each of the previous layers, forming a stack. At the end of this stack of connected layers, there is a Fully Connected Layer. A fully Connected Layer is connected to each output from the pooling layer. It is giving results in the form of a score from the process that has been done. The output layer on the already built CNN model is using the activation function SoftMax. SoftMax is an activation function used in multi-class modeling with the output in the form of a probability from each class, with the target class with the highest probability[21].

F. The Link Between MFCC and CNN

At this time, CNN is commonly used for solving speech or audio recognition problems. The CNN model used for audio recognition has been built based on frequency[18]. The MFCC produces an acoustic vector coefficient which is the feature of the signal. From the vectors, a small rectangular section is taken by CNN in Convolutional Layer. The feature from MFCC needs to be reshaped first by adding one additional dimension to fit the CNN model. The kernel size is determined first then the filter is determined in each layer.

G. Testing Scenario

The testing conducted using two kinds of comparison: data comparison and comparison between MFCC and STFT, which are modeled by ourselves. The data comparison scenario was comparing two kinds of data that have been taken. Those are the template-based data and random basic tone data. Those comparisons were the template-based data compared with the mix of template-based data and random basic tone data (random basic tone data addition). The performance comparison between those data was tested the model's performance to classify data in their respective class that can be seen in training accuracy and testing accuracy. Performance measurement is also done by measuring the model's ability to adapt to a new kind of data (avoiding overfitted model).

The comparison between MFCC and STFT conducted using the same classifier using the same number of layers, activation function, and optimizer. STFT is used as a comparator for MFCC because it has a time-frequency domain representation that can compete with MFCC. In another research, STFT was able to outperform MFCC in solving some classification problems using different data [22]. From the study, we assumed that STFT was able to become a competitive feature extractor for MFCC.

IV. Result and Discussion

The experiment was completed with two different codes, which the augmentation process and the primary process were in different codes. The CNN model built using ReLu activation function on each layer and SoftMax function on the output layer. There are four layers in the CNN model, with pool size equals two and dropout = 0.2 in each layer. The loss function used is categorical cross-entropy that is suitable for multi-class classification. The model's optimizer using Stochastic Gradient Descent (SGD) with a 0.009 learning rate and 0.9 momentum, and in feature extractor comparison, the optimizer Adam is also used to compare the performance of both feature extractors. The model fitted using 128 batch size and 100 epochs of training.

A. Data Comparison

Those experiments are giving a quite different result between template-based data only and random basic tone data addition. The model's performance was measured based on training and testing accuracy and how the model classified a new kind of data. The result of the experiment can be seen in Table II.

TABLE II. The Result from Data Comparison

Tested Data	Training Accuracy	Testing Accuracy
Template-Based Data Only	100%	100%
Random Basic Tone Data Addition	91.65%	89.50%

The template-based data only gives a result where the model gives high accuracy, 100% training accuracy, and 100% testing accuracy. The model trained with template-based only data predicts the data from each vocal type accurately, even though the voice comes from another person. However, the model was not able to predict correctly from data with another basic tone, although the data was still within the range of the kind of vocal it is supposed to be. It indicated that the model was overfitted, which means that the difference between each vocal type's data was too contrast. Making the model was able to differ the data from each class easily and difficult to recognize a new kind of data. The MFCC produced similar acoustic vectors as features in each class, where the acoustic vectors produced in one class with another class have contrasting differences. Furthermore, those vectors learned by the CNN making the CNN learned similar features from each class. The model became easier to classify the data but encountered difficulties in classifying data outside of those data (new kind of data).

The model was trained using the mix of template-based data and random basic tone data on the second data testing, giving a different result than the first model trained using the template-based only data. This model gives a lesser accuracy which is 91.65% on training accuracy and 89.50% on testing accuracy using SGD optimizer. This model learned from more varied data than the first model, so this model easier to recognize a new kind of data. Some data overlapped from one class to another in terms of frequency, and the augmentation increased its variety using pitch shifting. The process made the overlapping data more numerous and varied, making this model more difficult to classify but gave a better classification performance. Although the model trained using this data giving a lesser accuracy, this model was able to avoid overfitting, so this model's accuracy becomes the main result of this experiment.

B. *Feature Extractor Comparison*

This experiment used a mix of random basic tone data and template-based data using the same CNN parameters. The model performance also measured using testing and training accuracy with the result in Table III.

TABLE III. ACCURACY COMPARISON

Feature Extractor	Optimizer	Training Accuracy	Testing Accuracy	Duration
MFCC	SGD	91.65%	89.50%	11 minutes
	Adam	91.83%	91.14%	11 minutes
STFT	SGD	88.54%	88.35%	8 hours
	Adam	88.75%	89.50%	52 minutes

In table III shows that MFCC + CNN was able to outperform the performance of STFT + CNN. It can be seen in Table III. The Adam optimizer provided a better training and testing accuracy than the SGD optimizer in both feature extractors. Although the testing accuracy was quite competitive, the training accuracy was significantly different in both optimizers. The STFT produces a much bigger input shape for the CNN than MFCC, making the STFT features took a long time for CNN to process. It indicated that MFCC is slightly more efficient than STFT in resolving human vocal type classification. The testing also conducted using a different optimizer which is Adam optimizer. The result is quite the same, where MFCC + CNN still outperformed the STFT + CNN. From this experiment, we conclude that although MFCC is a classical feature extractor that often becomes a comparison model for another proposed model, MFCC still robust enough to resolve this classification problem.

We also performed some measurements using precision, recall, and F1-Score, as shown in Table IV. Based on this table, there are two feature extractions and the percentage level in form performance metrics that we compare to be used in this study. In this experiment, the results show that the MFCC extractor has a better performance than STFT. Therefore we use the MFCC feature extractor in this study.

TABLE IV. PERFORMANCE METRICS COMPARISON

Feature Extractor	Optimizer	Precision	Recall	F1-Score
MFCC	SGD	90%	90%	90%
	Adam	91%	91%	91%
STFT	SGD	86%	86%	86%
	Adam	90%	90%	90%

V. CONCLUSION AND FUTURE WORK

Based on the experimental results, the authors conclude that the MFCC method and the CNN model can solve the problem of classifying the human voice into four parts. This study resulted 91.83% of training accuracy and 91.14% of testing accuracy by using the Adam Optimizer. One of the essential steps to achieve these results needs to be augmented so that the CNN model can avoid overfitting data to work correctly. These results indicate that the combination of the MFCC features extractor and the CNN model provided more performance when it was compared to the combined features of the STFT extractor and CNN models.

Future works on this research might be about classifying human vowels with six classes which might become a more challenging problem, since baritone and mezzo-soprano vocal types are similar to other vocal types. Those problems will become even more challenging if the data are obtained from more people. It will make the augmentation process become less necessary because the data is already variated, so it can become more robust to be classified.

REFERENCES

[1] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, “Vocalset: A singing voice dataset,” *Proc. 19th Int. Soc. Music Inf. Retr. Conf. ISMIR 2018*, pp. 468–474, 2018.

[2] W. Wang, C. Liao, Q. Cheng, and P. Wang, “CONVOLUTIONAL RECURRENT NEURAL NETWORKS FOR MUSIC CLASSIFICATION,” *ACM Int. Conf. Proceeding Ser.*, pp. 1–6, 2017, doi: 10.1145/3371425.3371430.

[3] S. Kum and J. Nam, “Joint detection and classification of singing voice melody using convolutional recurrent neural networks,” *Appl. Sci.*, vol. 9, no. 7, 2019, doi: 10.3390/app9071324.

[4] A. Pahwa and G. Aggarwal, “Speech Feature Extraction for Gender Recognition,” *Int. J. Image, Graph. Signal Process.*, vol. 8, no. 9, pp. 17–25, 2016, doi: 10.5815/ijigsp.2016.09.03.

[5] P. Zwan, P. Szczuko, and A. Czy, “Neural Networks and Rough Sets,” pp. 793–802, 2007.

[6] I. Wijayanto and R. Dwifebrianti, “Jenis Tipe Jangkauan Suara Pada Pria Dan Wanita Menggunakan Metoda Mel-Frequency Cepstral Coefficient,” *Konfrensi Nas. Sist. dan Inform.*, no. October 2013, pp. 2–10, 2013.

[7] K. Choi, G. Fazekas, M. Sandler, and K. Cho, “Transfer learning for music classification and regression tasks,” *Proc. 18th Int. Soc. Music Inf. Retr. Conf. ISMIR 2017*, pp. 141–149, 2017.

[8] T. Van Pham, N. T. N. Quang, and T. M. Thanh, “Deep learning approach for singer voice classification of Vietnamese popular music,” *ACM Int. Conf. Proceeding Ser.*, pp. 255–260, 2019, doi: 10.1145/3368926.3369700.

[9] A. Elbir, H. O. Ilhan, G. Serbes, and N. Aydin, “Short Time Fourier Transform based music genre classification,” *2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet. EBBT 2018*, pp. 1–4, 2018, doi: 10.1109/EBBT.2018.8391437.

[10] T. Kobayashi, Y. Suzuki, and A. Kubota, “Audio feature extraction based on sub-band signal correlations for music genre classification,” *Proc. - 2018 IEEE Int. Symp. Multimedia, ISM 2018*, no. 3, pp. 180–181, 2019, doi: 10.1109/ISM.2018.00-15.

[11] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, "Heart sound classification based on improved MFCC features and convolutional recurrent neural networks," *Neural Networks*, vol. 130, pp. 22–32, 2020, doi: 10.1016/j.neunet.2020.06.015.

[12] D. Nagajyothi and P. Siddaiah, "Speech recognition using convolutional neural networks," *Int. J. Eng. Technol.*, vol. 7, no. 4.6 Special Issue 6, pp. 133–137, 2018, doi: 10.14419/ijet.v7i4.6.20449.

[13] D. S. Park *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 2613–2617, 2019, doi: 10.21437/Interspeech.2019-2680.

[14] E. Kharitonov *et al.*, "Data Augmenting Contrastive Learning of Speech Representations in the Time Domain," *arXiv*, pp. 1–6, 2020.

[15] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," vol. 2, no. 3, pp. 138–143, 2010, [Online]. Available: http://arxiv.org/abs/1003.4083.

[16] R. Hasan, M. Jamil, G. Rabbani, and S. Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients," *3rd Int. Conf. Electr. Comput. Eng. ICECE 2004*, no. December, pp. 28–30, 2004.

[17] E. Zawadzka-Gosk, K. Wołk, and W. Czarnowski, "Deep Learning in State-of-the-Art Image Classification Exceeding 99% Accuracy," in *Advances in Intelligent Systems and Computing*, 2019, vol. 930, doi: 10.1007/978-3-030-16181-1_89.

[18] J. T. Huang, J. Li, and Y. Gong, "An analysis of convolutional neural networks for speech recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2015, vol. 2015-August, doi: 10.1109/ICASSP.2015.7178920.

[19] M. Kubanek, J. Bobulski, and J. Kulawik, "A Method of Speech Coding for Speech Recognition Using a Convolutional Neural Network," *Symmetry (Basel).*, vol. 11, no. 9, p. 1185, 2019, doi: 10.3390/sym11091185.

[20] H. K. Vydana and A. K. Vuppala, "Investigative study of various activation functions for speech recognition," *2017 23rd Natl. Conf. Commun. NCC 2017*, pp. 6–10, 2017, doi: 10.1109/NCC.2017.8077043.

[21] C. E. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv*, pp. 1–20, 2018.

[22] M. Huzaifah, "Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks," *arXiv*, pp. 1–5, 2017.