Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 179 (2021) 789–797

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Computer Science and Computational Intelligence 2020

Model Comparison in Speech Emotion Recognition for Indonesian Language

Reinert Yosua Rumagit [a,*], Glenn Alexander[b], Irfan Fahmi Saputra[c]

[a,b,c]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

Abstract

Speech emotion recognition has become one of the active researches in machine learning for the past few years. There are already applications that use speech emotion recognition as its feature. This paper's purpose is to examine the difference in performance of model using multilayer perceptron (MLP), support machine vector (SVM), and Logistic Regression (LR) with Mel-frequency cepstral coefficients (MFCCs) on Indonesian language. Recording of various people's voices are used as the dataset, which is collected using a peer-to-peer method. Emotions in the recording are classified as happy and sad. For the experiment, the authors used Precision, Recall, F1-Score, and Accuracy for the measurement to find the best model. Among three models, LR model has the perfect accuracy which is 100%. LR and MLP have the best precision rate for happy emotion and have the best recall rate for sad emotion.

Keywords: Emotion Recognition; Indonesian Language; SVM; Logistic Regression; MLP;

1. Introduction

During recent years, speech recognition plays a crucial role in helping others for ease of use. Several popular technology companies such as Google, Samsung, and Apple have been applying speech recognition to translate human speech into sentences for their users to navigate through their products easily. By having speech recognition, it opens new branches to the research field which specifically speech emotion recognition (SER).

* Corresponding author. Tel.: +6287887350892
E-mail address: reinert.rumagit@binus.ac.id

1877-0509
10.1016/j.procs.2021.01.098

Speech emotion recognition is a machine learning field that enables the model to classify human emotions based on speech itself. Emotions have several modalities that can be perceived by speech, psychological signals, facial expression, etc. Compared to other modalities, speech signals proves to be much more versatile and easy to acquire[1]. In Indonesia, researches that are available about this topic is quite a few, and some of the available ones showing low accuracy compared to the research outside Indonesia. On the previous research that use Indonesian language, their data collection method relies on natural conversation in talk shows as their source of data[2]. However, gathering data through talk shows may cost bias in terms of speaker's lack of ability to show their fully extent emotion due to public image and being watched by the crowds. Therefore, this paper is composed with different collection method which is adopted from RAVDESS rules. Data is gathered through volunteers that is being handed out with predetermined sentences and to be repeated per emotion. There are three classifying methods that are used in this research for analyzing two specified emotions (happy, sad) using SVM, Logistic Regression (LR), and Multi-Layer Perceptron (MLP). Predecessor research that was conducted using SVM gives an accuracy of 68.31%[2]. By using three classification models, it will give the opportunity to compare which dataset performs better. The algorithm in SVM was used for classifying emotions with linear kernel. On the other hand, it could also be done in MLP which is one of the classes of neural networks. And lastly, LR puts into consideration as its simple capability and economic performance compared to the others.

2. Related Work

There are a lot of studies that has already done related to speech emotion recognition in the past. Speech emotion recognition system can be classified into four phases [3], which are analysis, feature extraction, modelling, and testing. Analysis phase is used for extracting information specific to the speaker such as excitation source, vocal tract, and behavior feature, which can be describe as the speaker identity. Feature extraction phase is used to convert the data into useful parametric representation which can be classified and analyzed. Modelling phase is used to process the data to generate speaker model using feature-oriented feature vectors. Testing phase is used to test the accuracy of the model generated from the previous phase.

According to Fahmi et al.[4], there are two research groups which processed the audio data differently in the feature extraction phase. The first research group used statistical representations as the audio data features, for example the research from Wunarso and Soelistio[5]. In their study, they use Indonesian language dataset (I-SpeED dataset) and build a model of speech emotion recognition which can identify emotion such as happiness, sadness, and anger from the corpus built by collecting valid voices from 38 different people in which produce 3420 voice or audio data. The features that are used are average amplitude, voice volume, frequency, and duration of speech. Using SVM and ANN as the model classification algorithm, the accuracy for SVM resulted in 76% and ANN resulted in 66%. The second group used temporal representations as the audio data features like the research from Basu et al[6]. In their study, they use German language dataset (Berlin dataset) and build a speech emotion recognition for seven emotion which are anger, pleasure, boredom, neutral, anxiety, disgust, and fear. For English dataset, there is also Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) which is very popular for speech emotion recognition research[16]. They used MFCC with 13 channels since MFCC is the most popular and important feature extraction in speech processing applications for extracting features [7,8,9] and using a combination of CNN and LSTM which resulted in the average accuracy up to 80%. There is also a study which stated that using MFCC feature with reduced dimensions and row normalization offer much higher accuracy when using SVM model[10].

In terms of model to be used for speech emotion recognition, there are many variation and preferences. SVM is the most popular model to be used since it is reported as the best at identifying emotion in different classifier[11]. However, sometimes MLP model offer better accuracy in some cases[12] and according to Atmaja and Akagi[13], deep MLP even outperform other modern neural network architectures with the same number of layers if given proper parameters. There are also previous studies of speech emotion recognition for Bahasa Indonesia (Indonesian language). Since there is no good audio dataset for Bahasa Indonesia, they must gather the dataset by themselves. Some research gathers the audio dataset from talk show or movie from the internet[14], while other gather the audio dataset by asking directly to people (peer-to-peer method)[5]. However, the second method of data collection is prone

to have an imbalanced dataset which can be solved by using synthetic minority over-sampling technique (SMOTE)[15]. Peer-to-peer method also has the advantage that emotions are clearly displayed in the dataset.

3. Methodology

Fig. 1. Overview of Methodology

Based on Figure 1, this research has 5 phases within the methodology, which mentioned down below:

- Data Collection phase, which responsible for collecting data from friends and relatives based on predetermined sentence and emotions (happy, sad).
- Pre-processing Data phase handles the preparation and quality aspect of the data.
- Feature Extraction phase, which involves extracting features based on Mel-frequency, MFCC, and chroma.
- Classification phase, train and test the model based on SVM, MLP, and Logistic Regression
- Evaluation phase, Creating a report about the result of classification.

3.1. Data Collection Phase

Dataset consists of happy and sad emotions that were requested to the writer's friends and relatives. Data participants were given out two sentences to speak and each of them needs to be spoken in happy and sad emotion. The data were manually gathered and saved in the format of .wav with the sample rate of 16 MHz then all the audios will be renamed according to RAVDESS file naming[16]. There were 31 actors and actresses involved in this data collection and result 117 data for 2 emotion which distributed by 56 goes to happy and 61 labeled as sad.

According to the distribution, the data that counts as happy emotion is slightly lower than the sad one. It happens due from some audio files that were firstly labeled are not fit to be categorized as happy and vice versa during the data relabeling process.

3.2. Pre-processing Data Phase

In this phase, all data were pre-processed. This is the phase where the data is to be cleaned to verify that the data were consistent and ready to use for the model. Below are the sequences in audio pre-processing phase:

- Data Conversion, converting all the available data into one uniform format .wav with 16 MHz sample rates.
- Data Labeling, renaming all the data based on RAVDESS file naming and label the data into the specified emotion.
- Data Relabeling, relabel all data emotions into the corrected ones to satisfy one's closest emotion. This stage took two people which are the authors of this research to judge whether the audio is behaving as requested emotions[17].

- Data Balancing, process of balancing data due to imbalance data between happy and sad. Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing number of cases or data artificially in a balanced way. How SMOTE works is that by selecting existing examples which close in feature space. Then, drawing lines between all the examples. Then SMOTE will create artificial examples within the lines.
- Data Scaling, data here will be normalized by using RobustScaler that is retrieved from scikit. RobustScaler is a normalization process that based on percentiles to center and scaling the statistics. This type of normalization fits more for small data due to its wide marginal scaling compared to MinMaxScaler.

3.3. Feature Extraction Phase

In this section, there are three features used for extracting. Those are mel-spectrogram, chroma, and MFCC.

- Mel-spectrogram is a spectrogram where the frequencies are converted to the mel scale. How it works is that by taking samples of emotion over time to represent audio signal. Then, the audio signal is mapped from time domain into frequency domain using fast Fourier transform then shifted frequency and amplitude to form spectogram[18].
- Chroma feature is applied to show the vocal content inside audio files. As traverses the helix, it also defines the angle of pitch rotation[8].
- MFCC is commonly used in processing speech. It can record the phonetical crucial characteristics of speech. Variation in human ear's demanding with frequency, makes it as the basis for MFCC. Each sound with an actual frequency, individual's pitch is measured on a scale called the 'mel' scale[19].

3.4. Classification Phase

This phase is divided into three models, i.e. creating model using SVM, MLP, and Logistic Regression method. The proportion that was used for training and test size was 80% for training and 20% goes to test. Classification phase was done using k-fold cross validation with 10 folds.

3.4.1. Support Vector Machine

Support Vector Machine (SVM) is one of supervised machine learning model that linearly separable binary sets. The goal of this model is to calculate and create a hyperplane that classifies all training vectors. After creating hyperplane, the next step is to determine maximum margin between data point and hyperplane which can be called as support vectors. The formula for finding hyperplane can be seen from Equation 1 and example of SVM also provided in Fig.2.

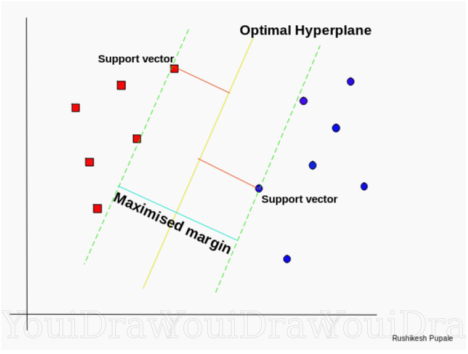

Fig. 2. Linear SVM Example

$$g(\vec{x}) = \vec{w}^T\vec{x} + \omega_0 \quad (1)$$

From Equation 1, Hyperplane can be formed and classifies the data based on $g(\vec{x}) \geq 1,\ \forall\vec{x} \in class\ 1$ and $g(\vec{x}) \leq -1,\ \forall\vec{x} \in class\ 2$. Every data that has value $g(\vec{x}) \geq 1$, it belongs to class 1. Meanwhile, for data that holds value of $g(\vec{x}) \leq -1$ belongs to class 2. Based on Equation 4, it is an equation to avoid explicit mapping to the higher dimension while n defines the size of dimensional mapping. x_i declared as data points and x_j is the given output. Linear kernel function comes into place in the testing process by handling multi-dimensional which computationally expensive.

$$k(x_i, x_j) = (x_i * x_j + 1)^n \quad (2)$$

There are some steps needed to process the data using SVM. First, after data is already balanced by SMOTE and scaled using RobustScaler, data will be plotted according to features extracted. Then, hyperplane is formed based on Equation 2. Last is determining the maximum margin of hyperplane from each class nearest data point. Testing was done using python Linear SVC library. Parameters used during testing were: C: 1, loss: hinge, penalty: l2, multi_class: ovr, class_weight: none, max_iter: 1000.

Parameters were tuned manually without grid search. Multi class is using ovr or can also be called One Versus Rest because the other method which is crammer singer is not actually better in accuracy and more costly in performance rather than ovr. For the iteration process, options were available ranging from 1000, 1500, and 2000. Increasing iteration does not affect much on increasing accuracy hence 1000 was picked.

3.4.2. Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is a part of feedforward artificial neural network which consists of input layer and output layer which are connected. MLP may have multiple hidden layers in between the mentioned layers beforehand, as seen in Fig.3. MLP requires for us to adjust its parameters, biases, or even weights of the model to minimize the risk of errors[8]. Activation function that is used during the experiment was ReLu which less computationally expensive and helps in minimizing the risk of vanishing gradient.

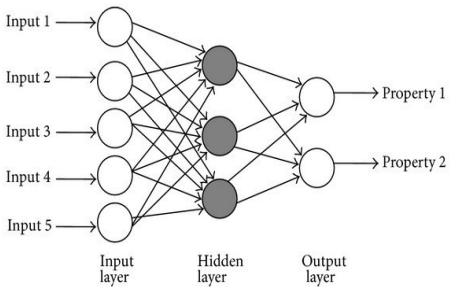

Fig. 3. Multi-Layer Perceptron Example

There are several steps that needs to be done for MLP to give specific emotions output from the input. First, after the data is being balanced by SMOTE, data will proceed to normalization stage using RobustScaler to improves data stability. Then, data will be trained through feedforward method. Input layer will proceed to hidden layer that was set to 200 hidden layers. Activation that is used during testing is ReLu as previously mentioned that can be seen through equation y below. The way ReLu works is by giving an output x if x is positive and 0 otherwise. Lastly, from hidden layers, output value from hidden layer will feedforward to output for getting emotion output. This process will iterate for n times depending on the tester's input.

$$f(x) = \max(0, x) \quad (3)$$

Library that was used for testing is from python library All the parameters are being tuned manually without grid search. Alpha was tested within range 0.01 - 0.07 and 0.05 turned out to be preferable in accuracy. Solver is being tested based on two types 'sgd' and 'adam'. Sgd is slightly longer to execute than adam and the gap between those two in terms of accuracy is not far off. During testing, hidden layer sizes keeps on changing to check which one is better and ranging from 200, 500, 1000, and 1287. Thus, 200 hidden layer sizes picked due to better performance and avoiding waste of hidden layers because greater the number does not seem to affect much on accuracy. Parameters used during the testing process were alpha: 0.05, batch_size: auto, activation: relu, solver: adam, hidden_layer_sizes: (200,), learning_rate: adaptive, max_iter: 1000.

3.4.3. Logistic Regression

Logistic regression in a simple term is binary probability that outcomes with dual answer (True/False, 1/0, Yes/No) if given a set of independent variables[20]. This model is preferable for dependent variable (categorical) data since the data used have small size of output (happy and sad). Because probability needs answer on a positive, exponent is used in the Equation 4:

$$\frac{p}{1-p} = \exp\left(b_0 + b_1 x + \cdots + b_n x\right) \tag{4}$$

Having logarithmic equation, it allows the line to be curved. p is defined as equation 5 below which $e^{-(b_0+b_1 x)}$ means exponential function with b_0 is the intercept from the linear regression equation. And $b_n x$ or $b_1 x$ is the regression coefficient multiplied by value of the predictor.

$$p = \frac{1}{1 + e^{-(b_0+b_1 x+\cdots+b_n x)}} \tag{5}$$

Testing process was done manually by authors without using grid search. Scikit python library is used for the model and parameters is being tuned from the library function. During testing, iteration process is ranging from 1000, 5000, and 10000. High iteration is used because the model itself is not computationally expensive and fast in processing the data. Here are some parameters used during testing: multi_class: multinomial, class_weight: none, solver': 'saga', max_iter: 10000.

3.5. Evaluation Phase

To evaluate the Speech Emotion Recognition models, 4 kinds of measurement are applied for the model. Confusion matrix helps to give general output of the data. Accuracy, Precision, Recall, F1-Score are picked to show out better statistic insight from each data. Accuracy is a ratio of all true predicted emotions to the total predicted emotions. Precision (P) is a measurement that calculates the amount of true selected emotion (i.e. happy) divided with the total prediction of selected emotion data. Recall (R) itself is a calculation which measures the amount of true specified emotion (i.e. happy) divided by total of actual specified emotion data. F1-Score is the measurement that fusions precision and recall which can be seen from Equation 6 below.

$$F = \frac{2.P.R}{P+R} \tag{6}$$

4. Experiment Result

The experiment is conducted using 3 models which are MLP, SVM, and LR models. After the dataset has been preprocessed using MFCC for feature extraction, SMOTE for balancing dataset, and RobustScaler for data normalization, the dataset is split into 10-fold for cross validation.

4.1. MLP Model

Table 1 describes the result when using MLP model for the classification. Fold 2 has the best result in which the model has 83% rate for correctly predict happy emotion from predicted happy class and 86% rate for correctly predict sad emotion from predicted sad class. The model also has 83% rate for correctly predicted happy emotion from the actual happy class and 86% rate for correctly predicted sad emotion from the actual sad class. Average accuracy from 10-fold cross validation is 73% (this is the average from 10th run of the code since the MLP model has different result each run, with 77.05% being the highest average accuracy and 70.58% being the lowest average accuracy).

Table 1. MLP Confusion Matrix

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Precision (%)	Happy	100	83	83	83	83	71	100	67	67	80
	Sad	64	86	83	83	67	60	70	67	100	100
Recall (%)	Happy	33	83	83	83	71	71	40	67	100	100
	Sad	100	86	83	83	80	60	100	67	75	50
F1-Score (%)	Happy	50	83	83	83	77	71	57	67	80	89
	Sad	78	86	83	83	73	60	82	67	86	67
Accuracy (%)		69.23	84.62	83.33	83.33	75	66.67	75	66.67	83.33	83.33

4.2. LR Model

Table 2 describes the result when using LR model for the classification. Fold 9 has the best result in which the model has 100% precision rate for happy emotion from all predicted classes and 100% precision rate for sad emotion from all predicted classes. The model also has 100% rate for correctly predicted happy emotion from the actual happy class and 100% rate for correctly predicted sad emotion from the actual sad class. Average accuracy from 10-fold cross validation is 76.22%.

Table 2. LR Confusion Matrix

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Precision (%)	Happy	100	100	80	80	75	71	67	67	100	78
	Sad	64	78	71	71	75	60	67	100	100	67
Recall (%)	Happy	33	67	67	67	86	71	40	100	100	88
	Sad	100	100	83	83	60	60	86	50	100	50
F1-Score (%)	Happy	50	80	73	73	80	71	50	80	100	82
	Sad	78	88	77	77	67	60	75	67	100	57
Accuracy (%)		69.23	84.62	75	75	75	66.67	66.67	75	100	75

4.3. SVM Model

Table 3 describes the result when using SVM model for the classification. Fold 9 has the best result in which the model has 80% precision rate for happy emotion from all predicted classes and 100% precision rate for sad emotion from all predicted classes. The model also has 100% rate for correctly predicted happy emotion from the actual happy class and 88% rate for correctly predicted sad emotion from the actual sad class. Average accuracy from 10-fold cross validation is 71.47%

Table 3. SVM Confusion Matrix

		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Precision (%)	Happy	50	67	60	75	62	67	60	83	80	80
	Sad	56	71	57	100	50	67	71	83	100	100
Recall (%)	Happy	33	67	50	100	71	86	60	83	100	100
	Sad	71	71	67	67	40	40	71	83	88	50
F1-Score (%)	Happy	40	67	55	86	67	75	60	83	89	89
	Sad	63	71	62	80	44	50	71	83	93	67
Accuracy (%)		53.85	69.23	58.33	83.33	58.33	66.67	66.67	83.33	91.67	83.33

4.4. Evaluation

Evaluation of the experiment involves comparison between each model classification report and accuracy. Table 4 describes the comparison of each emotion (happy and sad) precision, recall, f1-score from the best fold of each model classification report. Both MLP and LR has the best precision value for happy emotion, which is 100%, and recall value for sad emotion, which is also 100%. However, LR model is better than MLP model in terms of F1-Score because the recall rate for happy emotion and precision rate for sad emotion in LR model are higher than MLP model, which are counted for F1-Score calculation. From the three models, LR has the best average accuracy from 10-fold cross validation, which is 76.22%.

Table 4. Model Comparison and Classification Report

Model	Happy			Sad			Accuracy (%)
	Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)	
MLP	100	78	80	67	100	88	84.62
LR	100	100	100	100	100	100	100
SVM	80	100	89	100	88	93	91.67

5. Conclusion

In this research, the authors collect the dataset using peer-to-peer method to 31 people. After labeling and validating whether the emotion is clearly shown in the dataset, the dataset that can be used is 30 people voice. Since the dataset is not balanced, the authors used SMOTE for balancing the dataset and RobustScaler for data normalization. After that, the dataset is split into 10-fold for cross validation. From the experiment, the authors evaluated the result using classification report of each model. From Table 4, LR model has the perfect accuracy which is 100%. LR and MLP have the best precision rate for happy emotion and have the best recall rate for sad emotion. This mean that both LR and MLP can predict happy emotion from speech accurately and can identify sad emotion from speech. From all 3 models, the best average accuracy from 10-fold cross validation is LR models, which has 76.22% average accuracy. For future work, LSTM model will be used, gathering new dataset from different sources, and more emotion will also be used (such as anger, neutral, etc.).

References

1. Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Ali Mahjoub M, Cleder C. Automatic Speech Emotion Recognition Using Machine Learning. In: Social Media and Machine Learning. IntechOpen; 2019. p. 1–16.
2. Lubis N, Lestari D, Purwarianti A, Sakti S, Nakamura S. Emotion recognition on Indonesian television talk shows. 2014 IEEE Work Spok Lang Technol SLT 2014 - Proc. 2014;466–71.
3. Gupta K, Gupta D. An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system. Proc 2016 6th Int Conf - Cloud Syst Big Data Eng Conflu 2016. 2016;493–7.

4. Fahmi, Jiwanggi MA, Adriani M. Speech-Emotion Detection in an Indonesian Movie. 2020;(May):185–93.
5. Wunarso NB, Soelistio YE. Towards Indonesian speech-emotion automatic recognition (I-SpEAR). Proc 2017 4th Int Conf New Media Stud CONMEDIA 2017. 2017;2018-Janua:98–101.
6. Basu S, Chakraborty J, Aftabuddin M. Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. Proc 2nd Int Conf Commun Electron Syst ICCES 2017. 2018;2018-January(Icces):333–6.
7. Motamed S, Setayeshi S, Rabiee A, Sharifi A. Speech emotion recognition based on fusion method. J Inf Syst Telecommun. 2017;5(1):50–6.
8. Suhail MSK, Guna Veerendra Kumar J, Mahesh Varma U, Vege HK, Kuchibhotla S. Mlp model for emotion recognition using acoustic features. Int J Emerg Trends Eng Res. 2020;8(5):1702–8.
9. Winursito A, Hidayat R, Bejo A. Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition. 2018 Int Conf Inf Commun Technol ICOIACT 2018. 2018;2018-January:379–83.
10. Fernandes V, Mascarehnas L, Mendonca C, Johnson A, Mishra R. Speech emotion recognition using mel frequency cepstral coefficient and SVM classifier. Proc 2018 Int Conf Syst Model Adv Res Trends, SMART 2018. 2018;200–4.
11. Idris I, Salam MSH, Sunar MS. Speech emotion classification using SVM and MLP on prosodic and voice quality features. J Teknol. 2016;78(2–2):27–33.
12. Dhaouadi S, Abdelkrim H, Saoud S Ben. Speech Emotion Recognition: Models Implementation Evaluation. Proc Int Conf Adv Syst Emergent Technol IC_ASET 2019. 2019;256–61.
13. Atmaja BT, Akagi M. Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition. 2020; Available from: http://arxiv.org/abs/2004.02355
14. Kasyidi F, Lestari DP. Identification of four class emotion from Indonesian spoken language using acoustic and lexical features. J Phys Conf Ser. 2018;971(1).
15. Lasiman JJ, Lestari DP. Speech Emotion Recognition for Indonesian Language Using Long Short-Term Memory. 2018 Int Conf Comput Control Informatics its Appl Recent Challenges Mach Learn Comput Appl IC3INA 2018 - Proceeding. 2019;40–3.
16. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). PLoS ONE. 2018. 1–35 p.
17. Lotfian R, Busso C. Curriculum learning for speech emotion recognition from crowdsourced labels. IEEE/ACM Trans Audio Speech Lang Process. 2019;27(4):815–26.
18. Dörfler M, Grill T, Bammer R, Flexer A. Basic filters for convolutional neural networks applied to music: Training or design? Neural Comput Appl. 2020;32(4):941–54.
19. Bhuyan AK, Nirmal JH. Comparative study of voice conversion framework with line spectral frequency and Mel-Frequency Cepstral Coefficients as features using artficial neural networks. Proc - 2015 Int Conf Comput Commun Syst ICCCS 2015. 2016;230–5.
20. Jacob A. Modelling speech emotion recognition using logistic regression and decision trees. Int J Speech Technol [Internet]. 2017;20(4):897–905. Available from: http://dx.doi.org/10.1007/s10772-017-9457-6