Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks

Agustinus Bimo Gumelar
Dept. of Electrical Engineering,
Faculty of Electrical Technology,
Institut Teknologi Sepuluh Nopember
Fakultas Ilmu Komputer
Universitas Narotama
Surabaya, Indonesia
bimogumelar@ieee.org

Afid Kurniawan
Fakultas Ilmu Komputer
Universitas Narotama
Surabaya, Indonesia
afidk@fasilkom.narotama.ac.id

Adri Gabriel Sooai
Dept. of Computer Science
Universitas Katolik Widya Mandira
Kupang, Indonesia
adrigabriel@ieee.org

Mauridhi Hery Purnomo
Dept. of Electrical Engineering,
Dept. of Computer Engineering,
Faculty of Electrical Technology,
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
hery@ee.its.ac.id

Eko Mulyanto Yuniarno
Dept. of Electrical Engineering,
Dept. of Computer Engineering,
Faculty of Electrical Technology,
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ekomulyanto@ee.its.ac.id

Indar Sugiarto
Dept. of Electrical Engineering
Petra Christian University
Surabaya, Indonesia
indi@petra.ac.id

Agung Widodo
Fakultas Ilmu Komputer
Universitas Narotama
Surabaya, Indonesia
agung.widodo@narotama.ac.id

Andreas Agung Kristanto
Dept. of Psychology
Universitas Mulawarman
Samarinda, Indonesia
andreasagungk@gmail.com

Tresna Maulana Fahrudin
Fakultas Ilmu Komputer
Universitas Narotama
Surabaya, Indonesia
tresna.maulana@narotama.ac.id

***Abstract*— It is well known that human voice at the perceptual level consists of multimodality information. Therefore, a modality can be shared via neural emotion network through the independent stimuli processes. The expression and identification of emotions are significant steps for the human communication process. This process is biologically adaptive in a continuous manner, and for this reason, human voice identification becomes useful for classifying and identifying an effective specific characteristic between them. In this paper, we propose to identify the difference between the six essentials of human voice emotion. They will be generated using prosodic and spectral features extraction by utilizing Deep Neural Networks (DNNs). The result of our experiment has obtained accuracy as much as 78.83%. It presupposed that the higher intensity of emotions found in the sound sample would automatically trigger the level of accuracy as same as a higher one. Moreover, gender identification was also carried out along with the approximate accuracy at 90%. Nevertheless, from the learning process with the composition of 80:20, the training-testing data has obtained an exact accurate result by 100%.**

Keywords—human voice emotion, prosodic feature, spectral feature, DNNs

I. INTRODUCTION

Human characteristics have interacted with emotional beings. In particular, the emotions will influence all thoughts and behaviors. The emotions can be transmitted into positive and negative aspects e.g. happy, terrified, sad, etc. The definition of emotions, according to Drever, is a complex state of living things. They involve changes within the whole body, including breathing, pulse, chemical release from the gland, and others. On the mental side, there is also a change in attraction or concern and becomes a trigger for a definitive form of behavior [1]. The emotions are part of human development, and there are some aspects in common to be expressed. The human's ability to recognize emotions is called emotional intelligence. It is the ability to use emotional knowledge to understand and analyze current emotions that are being experienced [2].

Previous research [3] has argued that there are five basic emotions, such as anger, sadness, fear, disgust, and happiness. These emotions are recognized in terms of prosodic and sound qualities. In linguistic studies and theories, the prosodic include intonation, stress, and speech rhythms. Whereas, the sound refers to tone, energy, and tempo [4]. The nature of emotions is very subjective because it is a complex physiological phenomenon. This phenomenon results from interactions between the effects of biochemical reactions in the human body and the surrounding environment. In this research paper, the attention will be focused on prosodic and spectral features.

The human voice is an embodiment within itself in many social aspects [5]. Its pattern is an essential part of our identity and one of the most basic attributes that allow human to interact and communicate with other people. As well as in the extent of the physical environment or digital media [6]. To understand the pattern of the human voice, the process of identifying vocal emotions and facial expressions of others often occurs within the threshold of consciousness and is also processed automatically as well. This research was conducted within the aim of strengthening the alleged emotions of the

978-1-7281-0300-6/19/$31.00 ©2019 IEEE

human voice. In particular, we use a machine learning system that can identify six essentials emotion based on the Plutchik models [7].

This paper is arranged as follows: Section I explains the importance of human voice leads to a source of identifying human vocal emotion and emotion aspect on it. Section II is briefly present relevant work of human voice emotion classification and machine learning algorithm which are used in voice identification. Section III describes two feature extractions used in this study which is namely prosodic and spectral features. The dataset we used (RAVDESS) and the preprocessing technique. Some discussion about interesting results that derive from the experimental simulation and model we used to train based on Convolutional Neural Network (CNN) is elaborated in Section IV. At last, in Section V, we provide some conclusions and directions for future work.

II. Human Voice Emotion

In the everyday life of human relations, we have an emotional experience in various feelings such as despair, joy, sadness. The presence of the emotions may occur at the same time in unconsciously. By detecting emotional information, we need a passive device or sensor which can capture data on physical conditions or user's behavior without having to interpret inputs. It is as same as the analogy between a video camera device and a microphone which are capable of capturing speech sounds.

A. Plutchik's Model of Human Emotion

Emotions are a fundamental aspect of our daily life. They are embedded in the extent of our psychological reaction. Some interactions are formed continuously between emotions, behaviors, and thoughts so that they influence each other. Emotions are the primary source of information in terms of communicating and interacting with people. The difference in each person and culture can modify the expressions of others and also interpret an emotion. Besides, it is also a challenge for researcher to detect and identify the emotions.

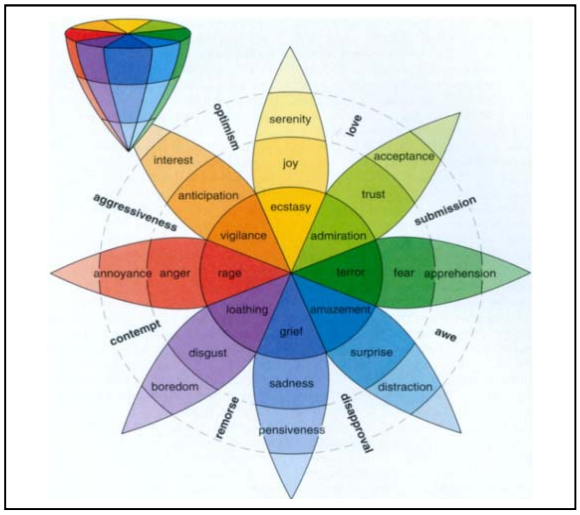

Fig. 1. Plutchik's human emotional relations model [7]

According to American Psychologist, Robert Plutchik has made theories of emotional classification with the taxonomy of human emotions which is divided into the eight most basic emotions [7]. All information consists of various levels of intensity and combination of emotions that will produce new emotions, namely secondary emotions. These basic emotions will be associated with a number of emotions represented in the center of the wheel on Fig.1, which are considered as manifestations of the extreme intensity of these basic emotions.

The way human recognizes the emotions is by capturing the expressions through sound. Nevertheless, the emotions of the human voice are dynamic. Thus, a specific tool is needed to strengthen our hypotheses on emotions that are being expressed in day-to-day conversation. In the process of interaction and communication between humans, the human voice becomes a source of emotional signaling. It is believed to be very reliable and humans can recognize vocal emotional expressions in speech. This ability will lead to a better understanding of effective signals created among humans.

B. Identify Voice Emotion based on Machine Learning

The determination of patterns and the process of detection of emotion based on the human voice has become a vital part of the interdisciplinary field of affective computing. In a paper by [8], propose a kernel-based method for multi-modal information analysis. In this work, they find patterns in speech and examine the relationships of the kernels produced uses speech and visual data. In another paper, the researcher captured speech features by using bispectral and bicoherence [9]. They use an ELM classifier for emotion recognition and feature selection. Single HMM is trained for each emotion and classified based on the model which is the closest of the resemblance.

III. Feature Extraction

Feature extraction determines what key features of the input are beneficial to learning, while the nature of the selected model determines the exact method by which the model learns to make predictions for the problem at hand. Feature extraction is particularly important in audio processing since raw audio input is extremely noisy and complicated. The features used in the experiments are extracted from RAVDESS [10]. The total number of 4320 samples corresponding to six emotional classes are used in this study. In some literature, various features that can be extracted from conversations have been proposed to detect emotional patterns. The most commonly used features are tone frequency, log energy, formant, sub band energy, MFCC (Mel-frequency Cepstrum Coefficient), sub-harmonic addition (SHS), and several dynamic features such as tone frequency speed/acceleration, or duration ratio of each the voice part that is voiced.

A. Prosodic Feature

Prosodic characteristics are achieved by modulation of various acoustic features that are perceived by listeners. They are suitable to follow the objective versus subjective terminology of those common features. Prosodic characters can be used to obtain some information, including emotions, word or sentence constraints, speaker characteristics, and language characteristics [11]; also the results for voice pattern recognition. Moreover, the prosodic features comprehend within a fundamental frequency (F0) which is a base for pitch (also intonation or melody), an objective measure duration as a subjective length, the intensity is denoted as loudness, and spectral structure is referred to as a timbre. In these contents,

there are intonation, rhythm, and stress. Each signal is indicating a complex perception entity and can be expressed using three main parameters i.e. duration, energy, and pitch.

Prosodic features can be extracted in a time interval of multiple frames. As mentioned in [12], the time interval is taken from a gap between two pauses present in a sequence. This approach is related to the pause to stop and use all utterances as a time for integration. Intonation and statistical calculations in intensity have been calculated from the *F0* value; resulting in derivatives, which its first and second are used for emotion recognition. The prosodic value towards the end of the sentence can give a signal for emotion recognition. While increasing *F0* value is related to the rest of the sentence with a surprise, the falling *F0* value may be related to sadness. At this stage, the Mel-Frequency Cepstral Coefficients (MFCC) feature extraction are coefficients that represent sound like a short-term power spectrum. Followings, the pitch feature is used as this can be complemented to the MFCC features. The process is carried out as follows:

1. Windows size = 0.025 (25ms),
2. step = 0.01 (10ms)
3. Cepstral number = 40,
4. Number of filters = 26,
5. nfft = None,
6. Lowest frequency = 0 Hz,

Highest frequency = None (no limits)

The initial stage of speech recognition is to extract features [13]. To perform feature extraction, we used the MFCC (Mel Frequency Cepstral Coefficient) method. The MFCC takes the sensitivity of human perception and its relationship to sound frequency as a source of information to examine. Another advantage is that this method is fairly easy to implement, hence becoming a widely used method for speech recognition [14][11].

B. *Spectral Feature*

The results of a recent research have shown that statistically, speech levels which use segmental spectral features contain rich information about expressiveness and emotion [15]. In this paper, the extraction of spectral features from speech signals is carried out. The fundamental of emotions in human voices has spectral features; which had a critical role in it. One set of other relevant sound frequency features is to describe in the form of a sound signal spectrum. Shown in the 95th percentile of the power spectral distribution is a strong value of Spectral roll off point [16].

TABLE I. SPECTRAL ROLLOFF POINT PSEUDOCODE

Spectral Feature Extraction using SRP pseudocode
w : array [shape=(n,)] or None for audio time series srat : number > 0 [scalar] audio sampling rate of `w` Smag : array [shape=(d, t)] or None for spectrogram magnitude n_fft : int > 0 [scalar] for FFT window size hop_length : int > 0 [scalar] for STFT freq : None or np.ndarray [shape=(d,) or shape=(d, t)] Center freqs for spectrogram bins. If `None`, then FFT bin center frequencies are used. Otherwise, it can be a single array of `d` center frequencies, roll_percent : float [0 < roll_percent < 1]

```
def spectral_rolloff(d=None, srat=22050, Smag=None,
n_fft=2048, hop_length=512, freq=None, roll_percent=0.95):

if not 0.0 < roll_percent < 1.0:
    raise ParamErr('roll_percent must in the range (0, 1)')

  S, n_fft = _spectrogram(w=w, Smag=Smag, n_fft=n_fft,
hop_length=hop_length)

  if not real object(Smag):
    raise ParamErr('Spectral rolloff  point is only defined '
                'with real-valued input')
  elif any(Smag < 0):
    raise ParamErr('Spectral rolloff is only defined '
                'with non-negative energies')
  # Compute the center frequencies of each bin
  if freq is None:
    freq = fft_frequencies(srat=srat, n_fft=n_fft)

  # Make sure that frequency can be broadcast
  if freq.ndim == 1:
    freq = freq.reshape((-1, 1))

  tot_energy = csum(S, axis=0)

  threshold = roll_percent * tot_energy[-1]

  ind = where(tot_energy < threshold, nan, 1)

  return nanmin(ind * freq, axis=0, keepdims=True)
```

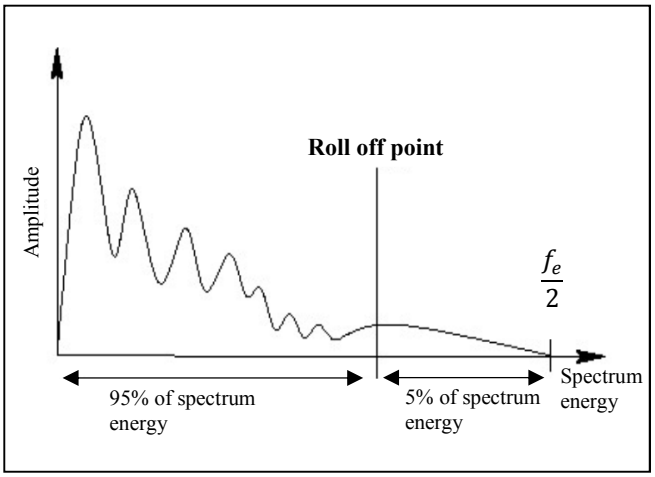

Fig. 2. Spectral Roll off Point [16]

The spectral roll off point in the Fig.2 can be considered as a slope measure in spectral form. It can be used to distinguish between the voiced speech and voiceless one. This has also been used in the classification of musical genres [17]

C. *RAVDESS Datasets*

On these experiments, we used speech emotion and vocal expression dataset in North American English language accent from the Department of Psychology at the Ryerson University, Toronto, Canada [10]. The dataset contains 4904 files of emotional speech in eight basic human emotional categories i.e. angry, disgust, fearful, happy, calm, sad, and neutral. It was an acted recording process by 24 people; 12 women and 12 men, and professional actors.

IV. Experimental Simulation

Before performing feature extraction, the datasets are labeled first.

A. Datasets Labeling Process

. Labeling refers to the provisions of RAVDESS as follows:

1. File type (01 = Audio-Video, 02 = video only without sound, 03 = sound only)
2. Voice channel (01 = sayings, 02 = song)
3. Type of emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = scared, 07 = disgust, 08 = surprised)
4. Emotional intensity (01 = normal, 02 = strong) NOTE = For the type of neutral emotion strong intensity is not available.
5. Sayings (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door")
6. Repetition (01 = First Repetition, 02 = Second Repetition)
7. Actors (01-24, odd numbers are male actors, even numbers are female actors)

Moreover, the label is adjusted to the file name (variable name = item) according to the provisions of RAVDESS (shown above). The settings of labeling initialization showed in Table I as follows:

TABLE II. Labeling Initialization

item[6:-16]	*item[18:-4]%2*	*feeling_list[]*
02	0	female_calm
02	1	male_calm
03	0	female_happy
03	1	male_happy
04	0	female_sad
04	1	male_sad
05	0	female_angry
05	1	male_angry
06	0	female_fearful
06	1	male_fearful

The feature extraction process is carried out using libROSA [18] to simplify the process. The resampling method used is Kaiser Fast with a sampling rate of 22050 Hz x 2 (stereo) so that the amount is 44,100 KHz. The duration of the sample taken is 2.5 seconds, starting after 0.5 seconds (offset = 0.5). After resampling, then the MFCC features are taken. The feature extraction process is done by repeating the entire dataset. The results of feature extractions are stored in the form of multi-dimensional arrays as shown in Table II.

TABLE III. Result of Feature Extraction

Feature
0 [-28.284271247461902, -28.284271247461902, -28...
1 [-28.284271247461902, -28.284271247461902, -28...
2 [-28.284271247461902, -28.284271247461902, -28...
3 [-28.284271247461902, -28.284271247461902, -28...
4 [-28.284271247461902, -28.284271247461902, -28...

B. Emotion Classification

Emotional classification stage is the stage to classify RAVDESS data into 8 types of emotions i.e. neutral, calm, disgust, surprised, happy, sad, angry, and scared. The classification stage is done by giving codes to the name of the audio file with the following conditions:

1. File type (01 = full-AV, 02 = only video, 03 = audio only).
2. Voice channel (01 = sayings, 02 = song).
3. Type of emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = scared, 07 = disgust, 08 = surprised). As per the problem limits, code 01, 07 and 08 are not used.

The last two digits describe the gender of the recorded voice (odd number = male, even number = female).

C. Deep Neural Network

Deep Neural Network is used to classify the form of learning. Classification can be done by supervision, which means the labeling classification part is done first on input data. Nevertheless, without supervision, it does not require labeling part. The algorithm is on behalf of the development of Artificial Neural Networks, by adding hidden layers between inputs and outputs. The process of building artificial neural networks is done by forming a tensor first. Tensor is a multidimensional array containing the generalizations of vectors and matrices as a result of data labeling [19]. The further step is to construct the layer. In fact, there is no exact formula regarding the number of layers that must be used. The amount is determined from trial-and-error until an optimal performance value is found. [20]. In this study, the types of activation we used are Relu and Softmax, considering that both the functions are suitable for speech recognition [21].

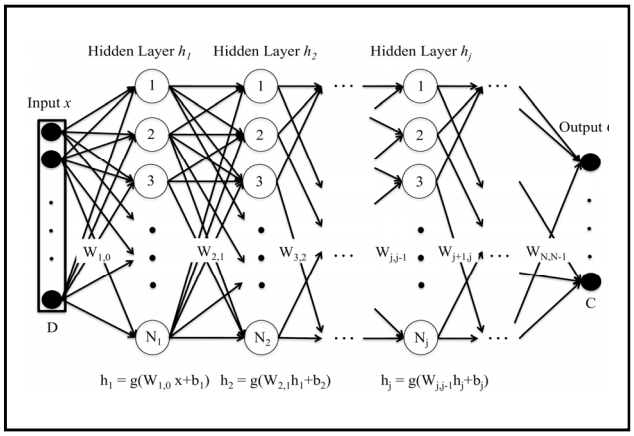

Fig. 3. A Standard Feedforward DNN Architecture [22]

D. CNN Based Architecture

At this stage, the results of extracting the MFCC feature are labeled first; then a model is made to be included in the Deep Neural Network. Then the training process is carried out on the available datasets, continuing checking its accuracy. If the training results which based on the chart of loss and accuracy level are considered as inappropriate, the

adjustment process is carried out on the number of layers and the number of MFCC features. Thus, the training process is repeated.

Another possible approach is training a convolutional neural network for emotion classification. In this approach, the features are passed as input to a 2-layer network composed of one convolutional layer with 32 3x3 filters (and stride 1x1) and one max pooling layer. The output of this convolutional layer is then passed as input to eight different fully connected layers followed by 8 output layers. Each of these output layers predicts the probability of the corresponding dimensions. This architecture is depicted in Fig. 3.

In the next process, the training data is separated from testing data. Separation is done randomly with the percentage of training data up to 80% (random numbers). Furthermore, the data are later converted into numerical data using the LabelEncoder feature from the scikit-learn library (sklearn)[23]. After the training data and testing data have been processed and prepared, then we made a model containing 18 layers. The type of model that will be created is Sequential and uses the Deep Learning Convolutional Neural Network.

```
from sklearn.utils import shuffle
rnewdf = shuffle(newdf)
rnewdf[:10]

rnewdf = rnewdf.fillna(0)

"""## Training-Testing Data Separation ##"""

newdf1 = np.random.rand(len(rnewdf)) < 0.8
train = rnewdf[newdf1]
test = rnewdf[~newdf1]
```

Fig. 4. Training-testing data separation

```
trainfeatures = train.iloc[:, :-1]
trainlabel = train.iloc[:, -1:]
testfeatures = test.iloc[:, :-1]
testlabel = test.iloc[:, -1:]
from keras.utils import np_utils
from sklearn.preprocessing import labelEncoder
X_train = np.array(trainfeatures)
y_train = np.array(trainlabel)
X_test = np.array(testfeatures)
y_test = np.array(testlabel)
lb = labelEncoder()
y_train =
np_utils.to_categorical(lb.fit_transform(y_tra
in))
y_test =
np_utils.to_categorical(lb.fit_transform(y_tes
t))
```

Fig. 5. Conversion into Numerical Data

```
x_traincnn = np.expand_dims(X_train, axis=2)
x_testcnn = np.expand_dims(X_test, axis=2)
model = Sequential()

model.add(Conv1D(256, 5, padding='same',
input_shape=(216,1)))
model.add(Activation('relu'))
model.add(Conv1D(128, 5, padding='same'))
model.add(Activation('relu'))
model.add(Dropout(0.1))
model.add(MaxPooling1D(pool_size=(8)))
model.add(Conv1D(128, 5, padding='same',))
model.add(Activation('relu'))
model.add(Conv1D(128, 5, padding='same',))
model.add(Activation('relu'))
model.add(Conv1D(128, 5, padding='same',))
model.add(Activation('relu'))
model.add(Dropout(0.2))
model.add(Conv1D(128, 5, padding='same',))
model.add(Activation('relu'))
model.add(Flatten())
model.add(Dense(10))
model.add(Activation('softmax'))
opt = keras.optimizers.rmsprop(lr=0.00001,
decay=1e-6)
```

Fig. 6. Layer and Model Making

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_1 (Conv1D)            (None, 216, 256)          1536
_________________________________________________________________
activation_1 (Activation)    (None, 216, 256)          0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 216, 128)          163968
_________________________________________________________________
activation_2 (Activation)    (None, 216, 128)          0
_________________________________________________________________
dropout_1 (Dropout)          (None, 216, 128)          0
_________________________________________________________________
max_poolingid_1 (MaxPooling1(None, 27, 128)            0
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 27, 128)           82048
_________________________________________________________________
activation_3 (Activation)    (None, 27, 128)           0
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 27, 128)           82048
_________________________________________________________________
activation_4 (Activation)    (None, 27, 128)           0
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 27, 128)           82048
_________________________________________________________________
activation_5 (Activation)    (None, 27, 128)           0
_________________________________________________________________
dropout_2 (Dropout)          (None, 27, 128)           0
_________________________________________________________________
conv1d_6 (Conv1D)            (None, 27, 128)           82048
_________________________________________________________________
activation_6 (Activation)    (None, 27, 128)           0
_________________________________________________________________
flatten_1 (flatten)          (None, 3456)              0
_________________________________________________________________
dense_1 (Dense)              (None, 10)                34570
_________________________________________________________________
activation_7 (Activation)    (None, 10)                0
=================================================================
Total params: 528,266
Trainable params: 528,266
Non-Trainable params: 0
```

Fig. 7. Result from the model

E. Evaluation Phase

The Confusion Matrix is used to evaluate the results of the experiment. The Confusion Matrix compares the expected results with the actual results. Confusion Matrix in the form of a table containing the four combinations of guesswork and actual results i.e. True Positive, False Positive, False Negative, True Negative [24].

TABLE IV. CONFUSION MATRIX

		Actual Value	
		Positive (1)	**Negative (0)**
Prediction Value	**Positive (1)**	***True Positive (TP)***	***False Positive (FP)***
	Negative (0)	***False Negative (FN)***	***True Negative (TN)***

On behalf with the Confusion Matrix, there are functions of Recall, Precision, and F-Measure.

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$F-measure = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}} \quad (3)$$

From the learning process, the final result is 78.83% level of accuracy. The learning process takes around 1-2 seconds per epoch. Learning results are depicted from the form of loss charts, accuracy charts, and the Confusion Matrix. Both loss and accuracy charts, as well as the Confusion Matrix, can be seen in the following picture:

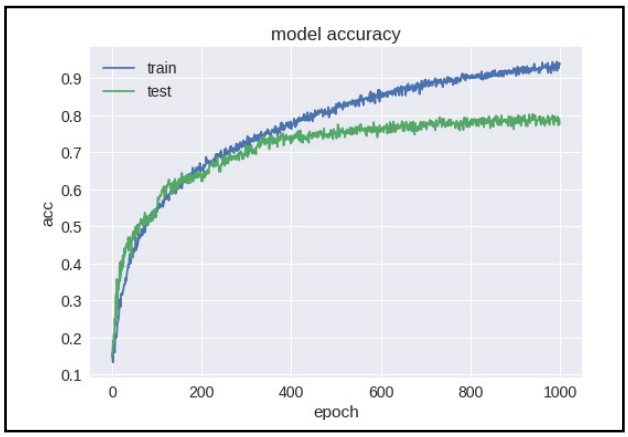

Fig. 8. The Accuracy Level Chart

From Fig.3, it is known that the level of accuracy is increasing along with the number of epochs that have been passed. The Confusion Matrix form can be seen below:

TABLE V. THE CONFUSION MATRIX

	F1-score							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
CNN	0.69	0.77	0.70	0.65	0.76	0.66	0.52	0.70
Decision Tree	0.29	0.56	0.42	0.42	0.53	0.43	0.29	0.34
Random Forest	0.53	0.70	0.58	0.51	0.76	0.64	0.44	0.56

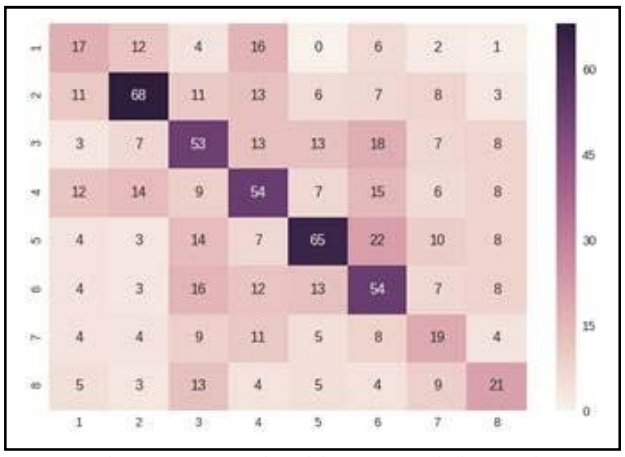

Fig. 9. Heatmap of Decision Tree Confusion Matrix

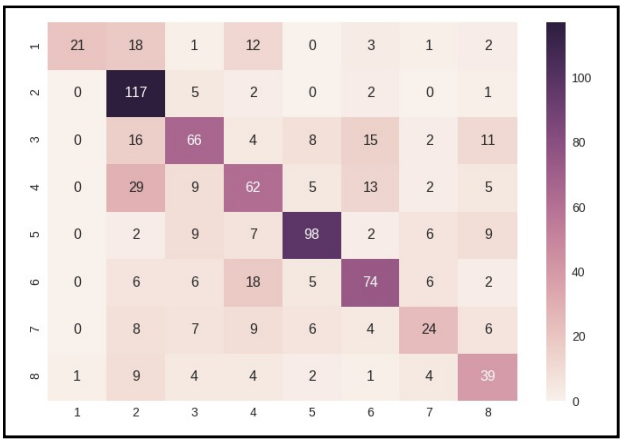

Fig. 10. Heatmap of Random Forest Confusion Matrix

Fig. 11. Heatmap of CNN Confusion Matrix

F. Changing the Composition of Training-Testing Data

To ensure that the model made is good enough, another experiment is carried out by changing the composition of the training data and testing data into 50:50 scale with 1 (one) type of emotion. Theoretically, the results should have 100% accuracy with losses close to or equal to 0. The results can be seen in the Loss chart, accuracy chart, and Confusion Matrix as follows on Fig.10. This accuracy level is comparable by the works of Huang [25], which also used the same RAVDESS dataset, but only reached 65% using the CNN model. However, in the study of Huang [25], it is recently reported that they had peaked the accuracy level up to 85%, but only

limited to ~500 epochs. In addition, the initial model of their RAVDESS data is slightly adjusted so that it may befits their study.

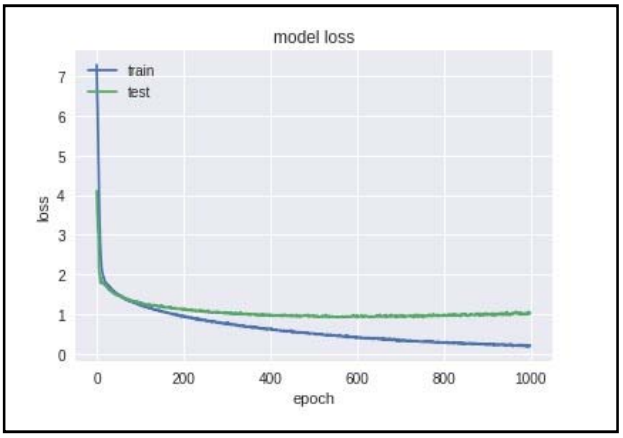

Fig. 12. The Loss Chart (80:20 data composition)

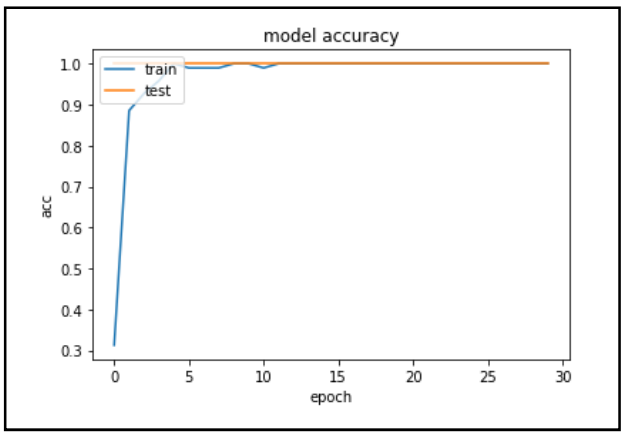

Fig. 13. The Accuracy Level Chart (80:20 data composition)

G. Emotion Prediction

In this experiment, we used a file named 03-01-05-02-01-02-19.wav. By following the nomenclature previously mentioned, it is known that the file has the type 05 of emotion (angry) and gender 19% 2(mod2)=1 (male). As a result, predictions can be made correctly, the prediction is conducted once again with the different type of emotion and gender; using a file named 03-01-04-02-01-02-20.wav. The selected file has the type 04 of emotion (sad) and gender 20% 2(mod2)=0 (female). Based on the following picture, the predictions can be done correctly.

V. Conclusion

The deep learning algorithm is one of a hungry data methodology. This method has been very successful in terms of processing human speech recognition. By the estimation of emotions, the biggest challenges lie in the availability of the number of references and the form of voice data patterns. The amount of available data to conduct the learning process can affect the results of accuracy, which more data will produce a better level of accuracy. The results of this study will have a much more decreased accuracy if the sound samples used have low emotional intensity. The accuracy value obtained is 78.83%.

We believe we could increase this accuracy value much better if more datasets are involved. Most other papers make use of clean data (no noise), which makes these particular results compelling. Yet, one future project utilizes the relations used in the detection to recreate emotions in speech. Furthermore, we would like to experiment more with feature selection methods such as nature-inspired algorithm e.g. Bee Swarm Optimization and other more complex models such as a deeper LSTM.

References

[1] J. Drever, *A Dictionary of Psychology*. [Baltimore] Penguin Books.

[2] J. D. Mayer and P. Salovey, *Emotional Development and Emotional Intelligence*. BasicBooks, 1997.

[3] M. D. Pell and S. A. Kotz, "On the time course of vocal emotion recognition," *PLoS One*, vol. 6, no. 11, 2011.

[4] E. Rodero, "Intonation and emotion: Influence of pitch levels and contour type on creating emotions," *J. Voice*, vol. 25, no. 1, pp. e25–e34, 2011.

[5] D. Sidtis and J. Kreiman, "NIH Public Access," *Integr. Psychol. Behav. Sci.*, vol. 46, no. 2, pp. 146–159, 2012.

[6] A. B. Gumelar, M. H. Purnomo, E. M. Yuniarno, and I. Sugiarto, "Spectral Analysis of Familiar Human Voice Based On Hilbert-Huang Transform," in *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, 2018, pp. 311–316.

[7] R. Plutchik, "The Nature of Emotions Human emotions a fact that have deep evolutionary roots , may explain their and provide tools complexity for clinical practice," vol. 89, no. 4, pp. 344–350, 2015.

[8] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. Multimed.*, vol. 14, no. 3 PART1, pp. 597–607, 2012.

[9] M. Hariharan *et al.*, "A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal," *Expert Syst. Appl.*, vol. 69, pp. 149–158, 2016.

[10] S. R. Livingstone and F. A. Russo, *The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english*, vol. 13, no. 5. 2018.

[11] C. Koniaris, M. Kuropatwinski, and W. B. Kleijn, "Auditory-model based robust feature selection for speech recognition," *J. Acoust. Soc. Am.*, vol. 127, no. 2, pp. EL73–EL79, 2010.

[12] S. Edition, *Extraction of Prosody for Automatic Speaker , Language , Emotion and Speech Recognition*. .

[13] V. Chernykh and P. Prikhodko, "Neural Networks."

[14] B. J. Mohan and N. Ramesh Babu, "Speech recognition using MFCC and DTW," *2014 Int. Conf. Adv. Electr. Eng. ICAEE 2014*, 2014.

[15] S. R. Krishna and R. Rajeswara, "SVM based Emotion Recognition using Spectral Features and PCA," vol. 114, no. 9, pp. 227–235, 2017.

[16] Pinquier, "Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle," no. October, 2004.

[17] T. Li and M. Ogihara, “Construction and evaluation of a robust multifeature speech/music discriminator,” in *In Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, USA*, 2005, pp. 197–200.

[18] B. McFee *et al.*, “librosa/librosa: 0.6.3,” Feb. 2019.

[19] I. Goodfellow, B. Yoshua, and C. Aaron, *Deep Learning*. 2016.

[20] F. Chollet, *Learn VIP !! 2017 -Deep Learning with python -Keras - book builds understanding through intuitive explanations and practical examples*. .

[21] E. Franti, I. Ispas, V. Dragomir, M. Dasc, E. Alu, Zoltan, and I. C. Stoica, “Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots,” *Rom. J. Inf. Sci. Technol.*, vol. 20, no. 3, pp. 222–240, 2017.

[22] A. Lozano-Diez, R. Zazo, D. T. Toledano, and J. Gonzalez-Rodriguez, “An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition,” *PLoS One*, vol. 12, no. 8, pp. 1–22, 2017.

[23] Pedregosa Fabian *et al.*, “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[24] Allison Ragan, “Taking the Confusion Out of Confusion Matrices – Towards Data Science,” *Oct 10, 2018*, 2018. .

[25] A. Huang and P. Bao, “Human Vocal Sentiment Analysis,” pp. 1–16, 2019.