*Proceedings of the 2023 Winter Simulation Conference*
*C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.*

# EMOTION CLASSIFICATION THROUGH SPEECH DATA ANALYSIS

Luzalen Marcos
Kristiina Valter Mai

Resilience Engineering Lab
Department of Electrical, Computer and
Biomedical Engineering
Toronto Metropolitan University
350 Victoria Street
Toronto, Ontario M5B 2K3, CANADA

Abdolreza Abhari

Distributed Systems and Multimedia
Processing (DSMP) Lab
Department of Computer Science
Toronto Metropolitan University
245 Church Street
Toronto, Ontario M5B 1Z4, CANADA

## ABSTRACT

Good quality healthcare services require effective communication between the patient and the healthcare provider. This work will help improve the areas of healthcare systems automation and optimization by applying Speech Emotion Recognition (SER) in health consultations to prevent miscommunication between patients and healthcare providers. Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) was used to compare the performances of different machine learning models in classifying emotions. Before feeding the raw dataset to the models, exploratory data analysis was done to determine features that should be considered for future analysis. Our results showed that depending on the emotion, there are some syllables in the text that were emphasized or took time to be pronounced by the speaker. After data analysis, the dataset was fed into different models and determined that the Support Vector Machine (SVM) is a machine-learning model for SER.

## 1 INTRODUCTION

Oral communication is essential for everyday interaction. With globalization removing geographical boundaries, people are able to travel or immigrate from one place to another. Although it is beneficial for knowledge transfer, it may lead to miscommunication because English may not be the first language in some countries. The way they speak may be misinterpreted as rude or offensive. To prevent this, there have been technological tools to help remove these language barriers. For instance, Google Translate translates languages from one another. However, some of the translated words may not be appropriate for the context. Another technological tool is emotion detection through texts using machine learning models such as Natural Language Processing (NLP). For example, Hussainalsaid et al. (2015) applied a Sentiment Analyzer and bi-gram analysis to determine whether web documents have positive or negative labels based on the emotion in a document. Another example is Grammarly (2023), a tool that assists in grammar and punctuation correction using texts especially used in a professional and academic setting. This tool is especially helpful for writers who are not native English speakers. Recently, it added a tone detector feature. This feature determines how an email or other forms of written communication sounds to a receiver or reader. This will give an idea to the sender on how their emails will sound to the receiver. Even though it is advantageous online, in-person interaction such as having a conversation with someone is still difficult for some people. Albert Mehrabian designed the 7-38-55 model for communication (Coke 2018). This means that people get most of the context in a conversation from the tone of voice (38%) and body language (55%) than spoken words (7%). In different fields, such as healthcare, oral communication is essential to ensure that clients or customers will receive proper services. Health consultations require

interaction and communication between the healthcare provider and patients. However, miscommunication often occurs during consultations and can cause frustrations between the providers and patients (Morgan 2013). Bartlett et al. (2008) mentioned that there are patients that are in undesirable situations because they could not communicate their needs (Serour et al. 2009). A study conducted by Kwame and Petrucka (2021) mentioned that there are communication barriers between patients and healthcare providers. They proposed person-centred care and communication continuum model (PC4), wherein both the provider and patient are recognized as persons. This would allow both entities to share their concerns regarding the health of the patient. Although this may seem effective, this would require the healthcare providers to do training to achieve the PC4 model, and it would still not solve the problem (Serour et al. 2009). Thus, interventions are needed to prevent patients and healthcare providers to be in bad situations (Bartlett et al. 2008).

A system should be placed between the healthcare provider and patient to assess the emotion of the conversation to prevent any misinterpretation coming from both parties is suggested. To achieve this, a Speech Emotion Recognition (SER) model can be embedded in the device. SER research has gained attention in recent years to analyze speech data. Recognizing the emotion of the speaker could prevent miscommunication between healthcare providers and patients. Since SER is comparatively new than other research areas, it is necessary to do dataset analysis and what factors could affect emotion classification. In this paper, we determine if it is possible to classify emotions through speech using machine learning models. We will introduce some of the state-of-the-art approaches that have been done with SER and then, implement the machine learning models to determine which model accurately distinguishes emotions through speech. The contribution of this paper includes: (1) determining other important characteristics of sounds that can be used as a potential feature for SER through data exploratory analysis and (2) comparing current state-of-the-art SER models to classify emotions and determine which model could be used as a baseline model.

## 2 RELATED WORKS

### 2.1 Sound Signal Characteristics

Signals such as sound have different characteristics, namely, loudness, pitch and quality. Each sound characteristic can contribute to the emotional meaning that a speaker is trying to convey. Babu et al. (2021) mentioned that information on pitch contains the emotion of the speaker.

Sim et al. (2002) utilized signal characteristics such as pitch and frequency for analysis. Segmentation for each sound signal was done for 0.05 s intervals to explore sound characteristics. For instance, if the person is angry, "the pitch contour is steeper than normal state" (Sim et al. 2002). At the end of their analysis, the authors emphasized the importance of data preprocessing as information may lost if the data is not properly fed into the models. Lastly, features such as formant, pitch and pitch slope did not give pleasing results and should be used with other features.

Aryani et al. (2018) analyzed the relationship of sound characteristics of a word with its affective meaning. Numerical analysis of words was done using Phonological Affective Potential (PAP). The acoustic features of words and their corresponding PAP were compared. Their results showed that the words with "voiceless consonants and the lowest sound intensity sound more arousing and give a negative connotation. Further, a slight sound could affect the emotion" (Aryani et al. 2018).

Van Zijl et al. (2014) analyzed sound characteristics to determine how performers such as violinists impart emotions of the music piece to the audience. They collected data from violinists and did an Analysis of Variation (ANOVA) in terms of articulation, timbre and tempo. Results showed that when the emotion is sad, the music is played slowed. In addition, Timbre does not affect the emotion, but tempo and dynamics affect the emotions that the audience will receive from the performers.

### 2.2 Current Approaches on SER

Different machine-learning models have been applied in SER (Issa et al. 2020). Mel-Frequency Cepstral Coefficients (MFCC) were used as features and were fed into Convolutional Neural Network (CNN) were used for gender and emotion classification by Mishra and Sharma (2020). MFCC is first obtained by filtering the signal based on a window. Then, Fast Fourier Transform (FFT) separates each signal into a single spectrum. Each spectrum was analyzed to avoid data loss. The results from the previous step were fed into a Mel-Scale Filter Bank. Then, the log was obtained and used as an input to Discrete Cosine Transform (DCT). The CNN model of Mishra and Sharma (2020) replaced a Fully Connected Network (FCN) Layer with a global average pooling layer. Their results showed that their CNN model can classify gender and emotion using MFCC as features and outperform Support Vector Machines (SVM).

On the contrary, different CNN Models and SVM were compared by Vrebcevic et al. (2019) to determine the best model for classification. Their results showed that the more complex the CNN, the more difficult for the CNN to classify emotions. In addition, their results showed that SVM outperformed all of the CNN models included on the paper. The authors suggested to discover more models for better classification in SER (Vrebcevic et al. 2019).

Ayvaz et al. (2022) extracted MFCC features from speech data spoken by Turkish speakers. Spatial pattern recognition advantages of MFCC was used along with Deep Neural Network (DNN) for the classification of real-time speeches. Their results showed that the MFCC is a good feature to use for SER applications and can lead to discovering "voice fingerprints" (Ayvaz et al. 2022).

Parthasarathy and Tashev (2018) tested different features with CNN models for SER. The CNN models included in the analysis are CNN with baseline features and CNN with only spectral features. Their results showed that CNN-based models performed better than FCN. Further, frame energy and speech presences probability features can help improve the accuracy. Further, the authors suggested in future work to use annotation distribution to improve the accuracy (Parthasarathy and Tashev 2018).

Meyer et al. (2021) presented different CNN-based models for SER. Then, they proposed their own model that combines CNN, Bi-directional long short-term memory (BiLSTM) and Fully connected layers. Log-Mel Spectrogram were used as feature and their results showed that their proposed model outperformed the baseline models (Meyer et al. 2021).

Mao et al. (2014) applied feature learning to determine the features that will help CNN models classify emotions. The authors used an unlabeled dataset and fed it into a modified Sparse Auto-Encoder (SAE). Their results showed that Salient Discriminative Feature Analysis (SDFA) provided better accuracy (Mao et al. 2014).

CNN was implemented for speech-to-emotion recognition (Sengodan et al. 2021). MFCC was used as a feature that will be extracted from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDNESS) dataset (Gokilavani et al. 2022). After that, they used data augmentation through pitch manipulation and adding noise to the dataset. Then, the data was split into 80% training set and 20% test set. Even though they were able to obtain 87.84% accuracy, the authors suggested to add statistical features to improve model accuracy (Sengodan et al. 2021).

## 3 METHODOLOGY

### 3.1 Dataset Description

For this paper, Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) (Lok 2019) dataset was used as input because it contains more dataset compared to other speech emotion dataset (Cao et al. 2014). Other datasets that were considered in this study were the Berlin Database of Emotional Speech (EmoDB) (Burkhardt et al. 2005), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo 2018) and Toronto Emotional Speech Set (TESS) (Pichora-Fuller and Dupuis 2020). EmoDB dataset has speakers speaking in German sentences while the speakers in CREMA-D are saying English sentences. In addition, RAVDNESS, which contains 1440 audio files, has fewer datasets

than CREMA-D, which has 7442 audio files. Moreover, the Toronto Emotional Speech Set (TESS) has female speakers only while the CREMA-D has both male and female speakers. Moreover, the CREMA-D dataset has diverse speakers from different backgrounds speaking English sentences and would help avoid overfitting (Lok 2019). The dataset contains 7442 voice clips from 48 actors and 43 actresses ages 20 to 74. The number of emotions in the CREMA-D dataset is visually represented in Figure 1. From the figure, there are 1271 files for all emotions except neutral emotion which has only 1087 files.

As shown in Table 1, the file name contains information about the *.wav* file. The audio file, where the speaker says, *"Don't Forget a Jacket"* is represented as *DFA* in the file name. The emotion type is indicated as ANG for angry, DIS for Disgust, FEA for fear, SAD is sad, HAP for Happy and NEU for neutral. The XX represents the level of emotion that the speaker felt when speaking. For this project, the ethnic background of actors and actresses is ignored. Further, the level of emotion will not be considered for analysis.

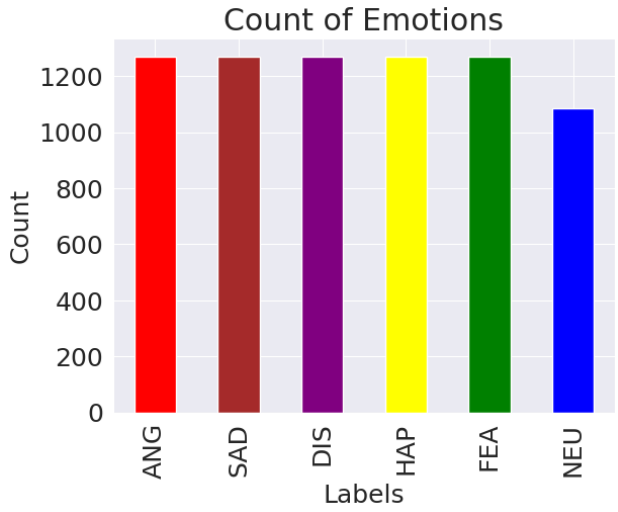



Figure 1: Visual presentation of the number of emotions.

Table 1: CREMA-D dataset description.

| | |
|---|---|
| **File Name** | "1001_DFA_ANG_XX.wav" |
| **Subject ID** | 1001 |
| **Sentence** | "Don't Forget a Jacket" |
| **Emotion** | Angry |
| **Level of Emotion** | XX - unspecified |

### 3.2 Approach

The overall approach for SER is shown on Figure 2. The study is implemented using GPU from Google Pro Plus. After the execution of the project, around 4 GB of RAM is used. From the image, raw data was fed into data preprocessing where features such as zero crossing rate, chroma features, which is pitch-based, MFCC, Log-Mel Spectrogram and Tonnetz were extracted. To achieve this, the Librosa is used. Librosa (McFee et al. 2015) is a Python library that can be used for information retrieval from audio files.

After feature extraction, the data is split into a 70% training and 30%test dataset. The training dataset was fed into the CNN model for classification. The test dataset was also fed into the model after training to evaluate the performance of the model. For this paper, CNN was used as a baseline model and compared

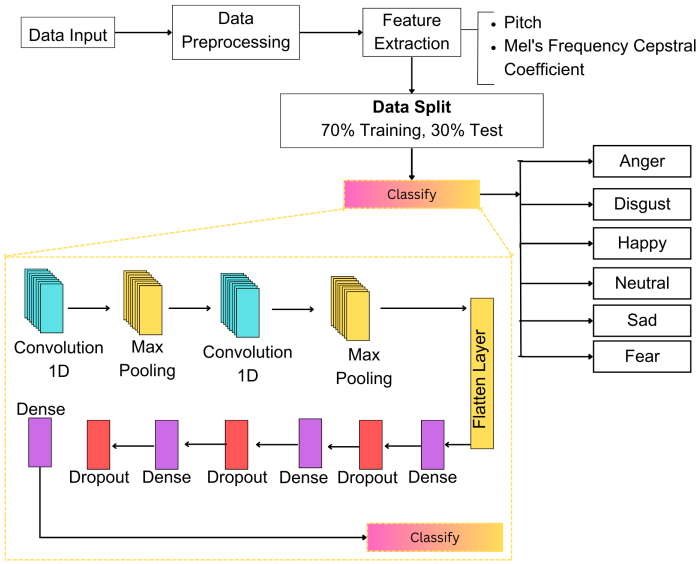



Figure 2: Overall approach of SER.

with SVM, Logistic Regression, and LSTM-CNN models. The CNN Model, in the insert of Figure 2, used for the model uses categorical cross-entropy as a loss function with a learning rate of 0.0001, 50 epochs and 128 as batch size. For the CNN with LSTM, the extracted features are first fed into the LSTM first then, the output goes to the CNN model.

The performance of different models was evaluated using accuracy, precision, recall and F1-score (Harikrishnan N. B 2020). *TP* means true positive, *TN* means true negative, *FP* means false positive, *FN* means false negative. Accuracy measures how the model was able to correctly identify if the signal matches with the right signal while precision measures how the model was able to distinguish the right and wrong emotion correctly. Recall is different from Precision because it determines the number of times the model could predict the correct and incorrect emotions. F1-Score measures the accuracy of the model. This means that the number of correct predictions is assessed throughout the whole dataset (Harikrishnan N. B 2020).

Different feature extraction techniques have been applied in this paper. One of them is Zero Crossing Rate (ZCR) (Delina et al. 2021), which is mathematically represented by

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} |sgn[x_i(n)] - sgn[x_i(n-1)]| \tag{1}$$

In equation (1), $W_L$ means the "ratio between audio samples and the number of characters and $x_i$ is the frame number at a specific audio sample" wherein the sign function, $sgn[x_i(n)] = -x_i(n) < 0$ and $sgn[x_i(n)] = x_i(n) \geq 0$ (Delina et al. 2021). Jothimani and Premalatha (2022) described ZCR as the number of times the signal changes from positive to zero to a negative signal and vice versa.

The purpose of Chroma and Tonnetz are almost the same because they measure the harmony and pitch of speech data (Tanoko and Zahra 2022) (Singh and Prasad 2023). Chroma in Short-Time Fourier Transform (STFT), in (2) (Muller 2015), is used as a feature for this paper. For the parameters in the formula, m represents the time frame for the specific kth-coefficient, w is the discrete-time window with length N and i is the real-world discrete signal. STFT is represented mathematically as

$$y(m,k) = |\chi(m,k)|^2 \tag{2}$$

For this paper, although their purpose are the same, there are still some differences on the graphs and extracted values of Chroma and Tonnetz. Thus, we applied both for this experiment.

The purpose of Mel Frequency Cepstral Coefficients (MFCC), in (3) (Patnaik 2023), measures how the user perceives sound (Jothimani and Premalatha 2022).

$$Mel(f) = 2595 \times log10\left(1 + \frac{f}{700}\right) \tag{3}$$

The Mel Spectrogram can be calculated using (4), where $PTB$ is the Mel Spectrum Power to Decibels and $S$ is the spectrum (Kuo et al. 2021).

$$PTB(S) = 10 \times \log_{10}(S) \tag{4}$$

From the paper of Jothimani and Premalatha (2022), Root Mean Square (RMS) in (5) contains the power of an audio signal. $X(n)$ is the speech signal and $N$ is for "number of samples in the frame of the speech signal" (Jothimani and Premalatha 2022). For this paper, RMS is applied to obtain the loudness graph of different emotions for speech.

$$RMS = \sqrt{\frac{\sum_{n=1}^{N} X^2(n)}{N}} \tag{5}$$

## 4 RESULTS & DISCUSSION

### 4.1 Exploratory Data Analysis

To determine other features that could be used for future SER analysis, we plot the sound signal for each emotion for the same subject saying *"Don't Forget a Jacket"*. The reason for this is that classification models can incorrectly classify emotions. Thus, it is essential to first observe the signals and determine if an emotion has a distinct characteristic from other emotions. From observation in Figure 3, the amplitude of the anger emotion sound signal is higher than the other emotions, followed by happy. This analysis is verified by the values in Table 2, wherein the sound characteristics of the Angry emotion have the highest values than other emotions.

As depicted in Figures 4 and 5, if the spectrograms for each emotion were analyzed individually, the emotions may look similar Hence, we overlapped the wavelength and magnitude of raw speech data to the spectrogram for better visualization and analysis. From the spectrogram, it shows that the emotion is stressed on the first syllable of the word *jacket*. To elaborate, **ja** has the highest frequency, and magnitude and is distinct in all emotions. It took longer for a person who has angry, fear, disgusted and happy emotions to pronounce **ja**.

Table 2: Sound characteristics of different emotions in CREMA-D.

| Emotion | Amplitude | Intensity | Pitch Frequency |
|---|---|---|---|
| **Angry** | 32580 | -22.869 | 1600 |
| **Fear** | 3211 | -40.566 | 800 |
| **Disgust** | 4430 | -39.256 | 444.444 |
| **Happy** | 9761 | -31.128 | 516.129 |
| **Sad** | 2596 | -43.112 | 615.385 |
| **Neutral** | 4121 | 0.48 | 1000.0 |

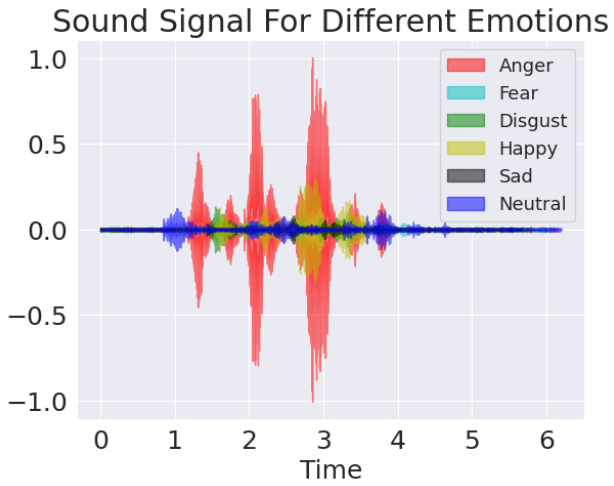
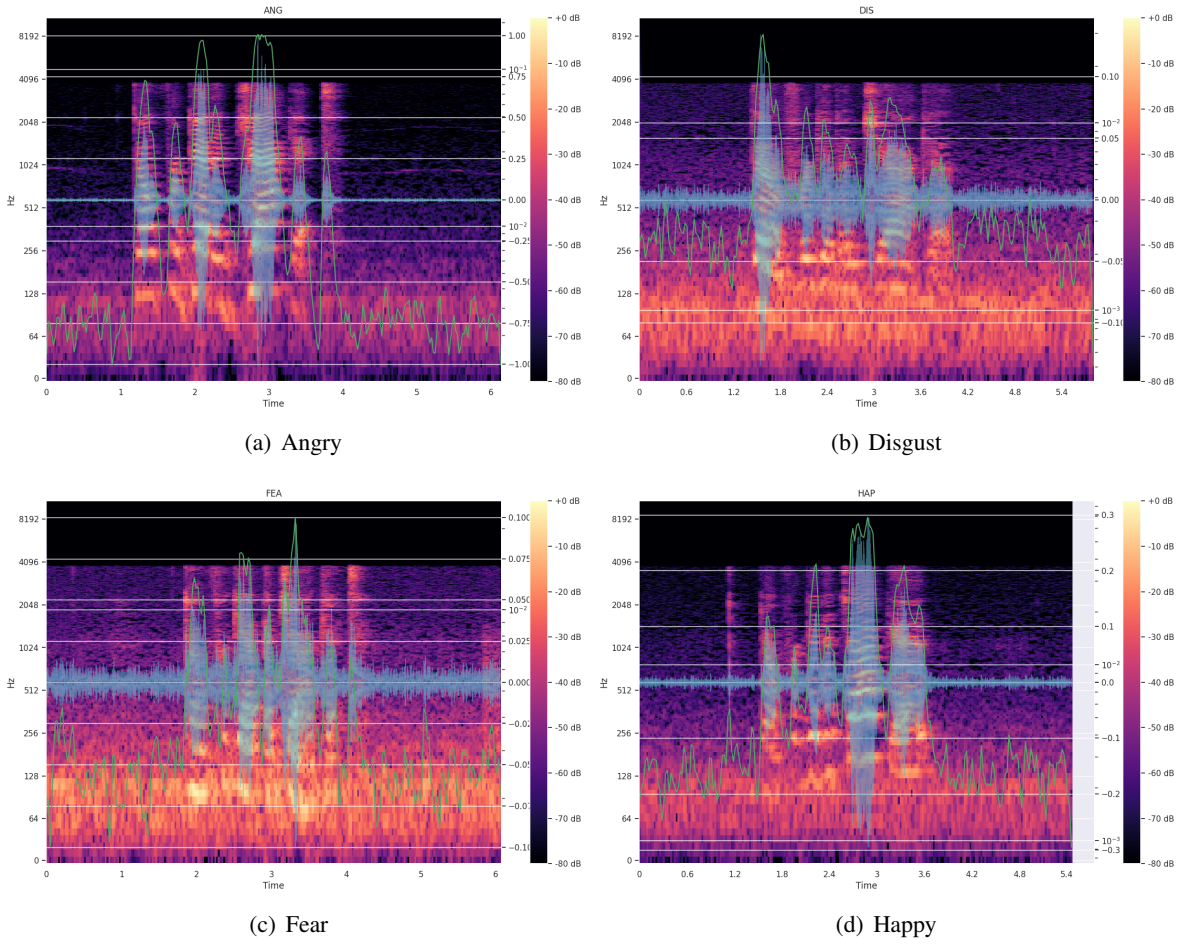


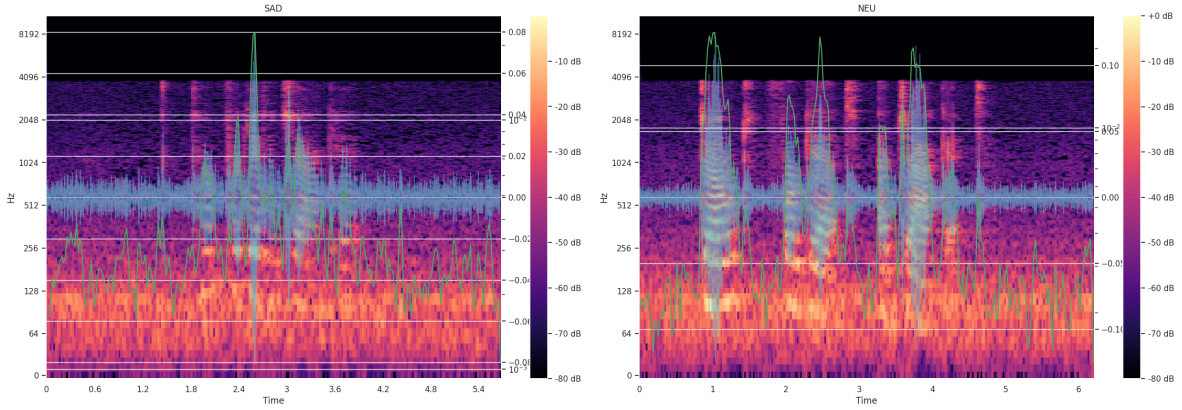

Figure 3: "Don't Forget a Jacket" wave signal for different emotions.




(a) Angry (b) Disgust

(c) Fear (d) Happy

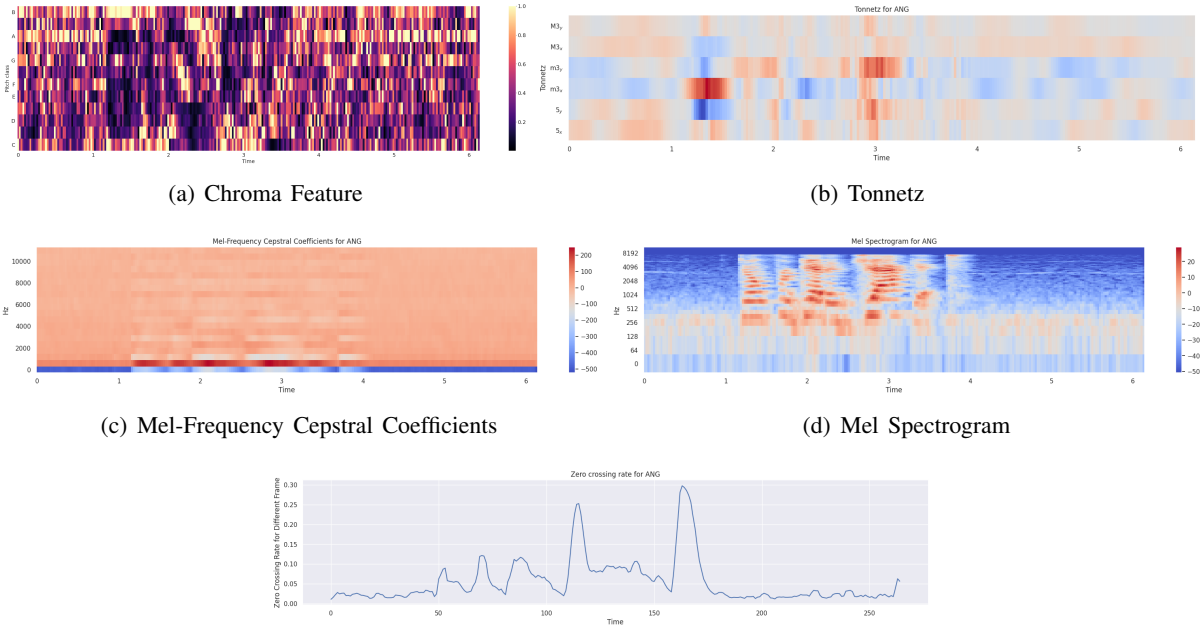
Figure 4: "Don't Forget a Jacket" wave signal, spectrogram and loudness graphs for Angry, Disgust, Fear and Happy emotions.




(a) Sad (b) Neutral

Figure 5: "Don't Forget a Jacket" signal, spectrogram and loudness graphs for Sad and Neutral emotions.

In Figures 6, the Zero Crossing Rate (ZCR), the ZCR has a length of 256. For Chroma and Tonnetz features, there are some parts in the audio that has higher frequency shown as a darker color on the graph. The MFCC indicates that speech data is mostly in the low-frequency range as it shows more positive value on the image while high frequency on the graph is noticeable when the speaker said the word *"jacket"*. The Mel Spectrogram also shows that the model is able to distinguish the pitch of a speaker. When the speaker says *"jacket"*, the graph showed that it took some time for the first syllable of ***ja****cket* to be pronounced along with a high pitch and frequency.




(a) Chroma Feature (b) Tonnetz

(c) Mel-Frequency Cepstral Coefficients (d) Mel Spectrogram

(e) Zero Crossing Rate

Figure 6: "Don't Forget a Jacket" sentence Chroma and Tonnetz feature for angry emotion.

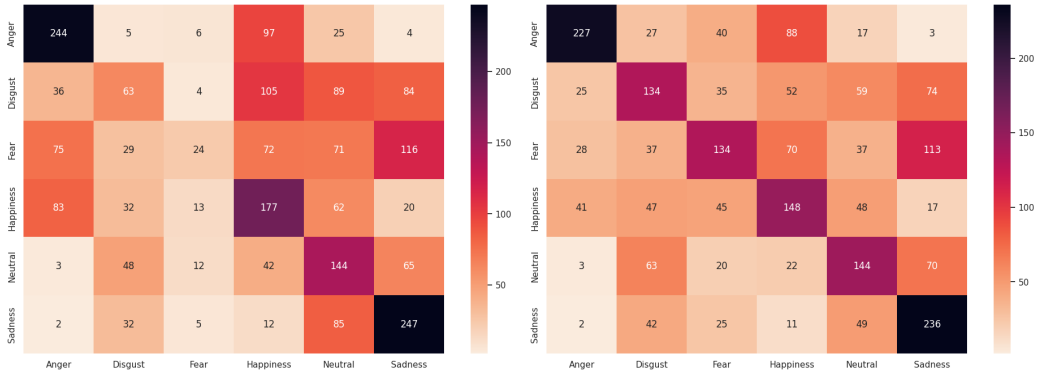
### 4.2 Emotion Classification Models

From the subsequent tables and figures, SVM outperformed CNN and other baseline models. In Table 3, SVM got the highest accuracy with 45% followed by CNN and CNN LSTM with 41% and 40%, respectively. This means that SVM was able to correctly classify the correct emotion from the dataset. From the table, even though the execution time of Logistic Regression is less than SVM, it has lower accuracy.

Table 3: Accuracy and execution time comparison of different models.

| Model | Accuracy (%) | Time (s) |
|---|---|---|
| CNN | 41 | 41.9469 |
| SVM | 45 | 9.1169 |
| Logistic Regression | 34 | 0.9103 |
| CNN+LSTM | 40 | 427.9815 |

The confusion matrices in Figure 7 visually represent the results of comparing the predicted or classified emotion from the machine learning models versus the ground truth or actual labels of the speech data. From the images, the diagonal elements on each matrix represent correctly predicted emotions. It can be seen that the diagonals of SVM Classification, CNN has the highest values which support the initial analysis from Table 3.




(a) CNN

(b) SVM

(c) Logistic Regression

(d) CNN with LSTM

Figure 7: Classification analysis of different models.

In the confusion matrix SVM classification, the Happiness column was only able to classify 37.8% from 397 samples, which is low compared to other emotions in the same confusion matrix. The Disgust emotion has the lowest correctly predicted scores which are 30.14% from 209 samples and 29.70% from 377 samples from the confusion matrices of CNN and CNN with LSTM, respectively. The Logistic Regression model was only able to correctly identify 19.21% of Neutral emotion from 151 samples. The results are not surprising because, from Figure 3, the waveforms of Disgust, Happy and Neutral are closely similar to each other and may have confused the models. Thus, it would be useful to extract features in a smaller time frame or use distinct features such as pitch frequency and Regions of Interest (ROI) from the speech data for each emotion. In relation to this, the Neutral emotion has the least number of data from Figure 1 that lead to unbalanced data. Dataset normalization and parameter modification would be useful to remove any bias during model training and testing.

SVM, from Tables 4 and 5, was able to identify Anger emotion by having an F1 score of 0.62. Importantly, the higher the precision, the lower the recall and vice versa. For instance, CNN-LSTM has a 0.72 precision, which means that CNN-LSTM was able to correctly identify that the samples presented to it are Anger. On the other hand, it has a 0.48 recall for Anger which means that 48% of all the samples was Anger emotion.

Table 4: Performance comparison of different models for anger, disgust and fear emotions.

| | **Anger** | | | **Disgust** | | | **Fear** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Models** | Prec | Recall | F1 | Prec | Recall | F1 | Prec | Recall | F1 |
| CNN | 0.51 | 0.71 | 0.59 | 0.32 | 0.25 | 0.28 | 0.40 | 0.22 | 0.28 |
| SVM | 0.70 | 0.56 | 0.62 | 0.38 | 0.35 | 0.37 | 0.45 | 0.32 | 0.37 |
| Logistic Regression | 0.53 | 0.48 | 0.51 | 0.31 | 0.30 | 0.31 | 0.28 | 0.10 | 0.15 |
| CNN-LSTM | 0.72 | 0.48 | 0.58 | 0.30 | 0.32 | 0.31 | 0.33 | 0.05 | 0.08 |

Table 5: Performance comparison of different models for happiness, neutral and sadness emotions.

| | **Happiness** | | | **Neutral** | | | **Sadness** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Models** | Prec | Recall | F1 | Prec | Recall | F1 | Prec | Recall | F1 |
| CNN | 0.30 | 0.42 | 0.35 | 0.38 | 0.45 | 0.41 | 0.56 | 0.41 | 0.48 |
| SVM | 0.38 | 0.43 | 0.40 | 0.41 | 0.45 | 0.43 | 0.46 | 0.65 | 0.54 |
| Logistic Regression | 0.28 | 0.41 | 0.33 | 0.19 | 0.09 | 0.12 | 0.34 | 0.63 | 0.44 |
| CNN-LSTM | 0.30 | 0.59 | 0.40 | 0.35 | 0.43 | 0.38 | 0.50 | 0.55 | 0.52 |

## 5 CONCLUSION

To avoid miscommunication between healthcare providers and patients, a device could be used to help remove the language barrier. This paper focuses on the potential of SER application to help solve the mentioned problem. It compares the classification performance of basic machine learning models using CREMA-D as a dataset. From the results, SVM seems to outperform advanced machine learning models such as CNN. Results showed that in terms of accuracy, SVM obtained 45%, followed by CNN at 41% and CNN with LSTM at 40%. The results are verified using a confusion matrix wherein the diagonal of SVM is higher than the other models. Data preprocessing may have affected the results due to unbalanced data. SVM can be used to get insight for data exploratory analysis, but combining different machine learning models would help improve the accuracy of the model. For the contributions, different machine learning

models were compared and it was found that SVM and CNN-LSTM are good baseline models. For the second objective, ROI on a pitch graph could be used as an additional feature for SER.

The Exploratory Data Analysis section of this paper showed that the sound signal for each emotion has a distinct characteristic, and thus, ROI for each emotion signal as a feature. Moreover, speech rate and the time it takes for the user to speak a phoneme should be considered for feature extraction. In addition, different machine learning models should be integrated to improve speech emotion recognition. Different audio preprocessing techniques are also needed to be applied to avoid unbalanced data.

## REFERENCES


Aryani, A., M. Conrad, D. Schmidtke, and A. Jacobs. 2018, June. "Why 'Piss' Is Ruder Than 'Pee'? The Role of Sound in Affective Meaning Making". *PLOS ONE* 13(6):e0198430.

Ayvaz, U., H. Gürüler, F. Khan, N. Ahmed, T. Whangbo, and A. Akmalbek Bobomirzaevich. 2022. "Automatic Speaker Recognition Using Mel-Frequency Cepstral Coefficients Through Machine Learning". *Computers, Materials & Continua* 71(3):5511–5521.

Babu, P. A., V. Siva Nagaraju, and R. R. Vallabhuni. 2021, June. "Speech Emotion Recognition System With Librosa". In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, 421–424. Bhopal, India: IEEE.

Bartlett, G., R. Blais, R. Tamblyn, R. J. Clermont, and B. MacGibbon. 2008, June. "Impact of Patient Communication Problems on the Risk of Preventable Adverse Events in Acute Care Settings". *Canadian Medical Association Journal* 178(12):1555–1562.

Burkhardt, F., M. Kienast, A. Paeschke, and B. Weiss. 2005. *Berlin Database of Emotional Speech General Information*. Berlin: Technical University of Berlin. http://emodb.bilderbar.info/docu/.

Cao, H., D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. 2014, October. "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset". *IEEE Transactions on Affective Computing* 5(4):377–390.

Coke, T. 2018. "BODY TALK". *RSA journal* 164(4 (5576)):48–48. ISBN: 0958-0433.

Delina, M., H. Nasbey, A. Gunawan, and S. Muhasyah. 2021, October. "Digital Text Security with Steganography Least Significant Bit and Audio Feature Extraction". *Journal of Physics: Conference Series* 2019(1):012108.

Gokilavani, M., H. Katakam, S. A. Basheer, and P. Srinivas. 2022, January. "Ravdness, Crema-D, Tess Based Algorithm for Emotion Recognition Using Speech". In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1625–1631. Tirunelveli, India: IEEE. https://ieeexplore.ieee.org/document/9716313/.

Grammarly 2023. "Tone Detector and Tone Suggestions | Grammarly". https://grammarly.com/tone.

Harikrishnan N. B 2020, June. "Confusion Matrix, Accuracy, Precision, Recall, F1 Score". https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd.

Hussainalsaid, A., B. Z. Azami, and A. Abhari. 2015. "Automatic Classification of the Emotional Content of URL Documents Using NLP Algorithms". In *Proceedings of the 18th Symposium on Communications & Networking*, CNS '15, 56–59. San Diego, CA, USA: Society for Computer Simulation International. event-place: Alexandria, Virginia.

Issa, D., M. Fatih Demirci, and A. Yazici. 2020, May. "Speech Emotion Recognition With Deep Convolutional Neural Networks". *Biomedical Signal Processing and Control* 59:101894.

Jothimani, S., and K. Premalatha. 2022, September. "MFF-SAug: Multi-Feature Fusion With Spectrogram Augmentation of Speech Emotion Recognition Using Convolution Neural Network". *Chaos, Solitons & Fractals* 162:112512.

Kuo, J.-Y., Z.-M. Chen, and H.-C. Lin. 2021, December. "Constructing Speech Emotion Recognition Model Based on Convolutional Neural Network". In *2021 28th Asia-Pacific Software Engineering Conference Workshops (APSEC Workshops)*, 52–56. Taipei, Taiwan: IEEE.

Kwame, A., and P. M. Petrucka. 2021, December. "A Literature-Based Study of Patient-Centered Care and Communication in Nurse-Patient Interactions: Barriers, Facilitators, and the Way Forward". *BMC Nursing* 20(1):158.

Livingstone, S. R., and F. A. Russo. 2018, May. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of : Facial and Vocal Expressions in North American English". *PLOS ONE* 13(5):e0196391. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391.

Lok, Eu Jin 2019. "CREMA-D". https://www.kaggle.com/datasets/ejlok1/cremad.

Mao, Q., M. Dong, Z. Huang, and Y. Zhan. 2014, December. "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks". *IEEE Transactions on Multimedia* 16(8):2203–2213.

McFee, B., C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto. 2015. "Librosa: Audio and Music Signal Analysis in Python". 18–24. Austin, Texas.

Meyer, P., Z. Xu, and T. Fingscheidt. 2021, January. "Improving Convolutional Recurrent Neural Networks for Speech Emotion Recognition". In *2021 IEEE Spoken Language Technology Workshop (SLT)*, 365–372. Shenzhen, China: IEEE.

Mishra, P., and R. Sharma. 2020, October. "Gender Differentiated Convolutional Neural Networks for Speech Emotion Recognition". In *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 142–148. Brno, Czech Republic: IEEE.

Morgan, S. 2013. "Miscommunication between patients and general practitioners: implications for clinical practice". *Journal of Primary Health Care* 5(2):123.

Muller, D. M. 2015. "Short-Time Fourier Transform and Chroma Features".

Parthasarathy, S., and I. Tashev. 2018, September. "Convolutional Neural Network Techniques for Speech Emotion Recognition". In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 121–125. Tokyo: IEEE.

Patnaik, S. 2023, March. "Speech Emotion Recognition by Using Complex MFCC and Deep Sequential Model". *Multimedia Tools and Applications* 82(8):11897–11922.

Pichora-Fuller, M. Kathleen and Dupuis, Kate 2020. "Toronto Emotional Speech Set (TESS)". https://borealisdata.ca/citation?persistentId=doi:10.5683/SP2/E8H2MF.

Sengodan, T., M. Murugappan, and S. Misra. (Eds.) 2021. *Advances in Electrical and Computer Technologies: Select Proceedings of ICAECT 2020*, Volume 711 of *Lecture Notes in Electrical Engineering*. Singapore: Springer Nature Singapore.

Serour, M., H. A. Othman, and G. A. Khalifah. 2009, April. "Difficult Patients or Difficult Doctors: an Analysis of Problematic Consultations". *Electronic Journal of General Medicine* 6(2):87–93.

Sim, K.-B., C.-H. Park, D.-W. Lee, and Y.-H. Joo. 2002, June. "Emotion Recognition Based on Frequency Analysis of Speech Signal". *International Journal of Fuzzy Logic and Intelligent Systems* 2(2):122–126.

Singh, V., and S. Prasad. 2023. "Speech Emotion Recognition System Using Gender Dependent Convolution Neural Network". *Procedia Computer Science* 218:2533–2540.

Tanoko, Y., and A. Zahra. 2022, December. "Multi-Feature Stacking Order Impact on Speech Emotion Recognition Performance". *Bulletin of Electrical Engineering and Informatics* 11(6):3272–3278.

Van Zijl, A. G. W., P. Toiviainen, O. Lartillot, and G. Luck. 2014, September. "The Sound of Emotion". *Music Perception* 32(1):33–50.

Vrebcevic, N., I. Mijic, and D. Petrinovic. 2019, May. "Emotion Classification Based on Convolutional Neural Network Using Speech Data". In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1007–1012. Opatija, Croatia: IEEE.

## AUTHOR BIOGRAPHIES

**LUZALEN MARCOS** is a Ph.D. Electrical and Computer student in the Department of Electrical, Computer and Biomedical Engineering at Toronto Metropolitan University and a member of the Resilience Engineering Lab. Her research interests include environmental sustainability, data science, natural language processing, and human-computer interaction. Her email address is lmarcos@torontomu.ca.

**ABDOLREZA ABHARI** is a professor in the Department of Computer Science at Toronto Metropolitan University and director of DSMP lab. He holds a Ph.D. in Computer Science from Carleton University. His research interests include web social networks, data science, AI and agent systems, network simulation, and distributed systems. His email address is aabhari@torontomu.ca.

**KRISTIINA VALTER MAI** is a professor in the Department of Electrical, Computer and Biomedical Engineering at Toronto Metropolitan University and director of Resilience Engineering Lab. She holds a Ph.D. in Biomedical Engineering from the University of Toronto. Her research interests include environmental sustainability, biomedical engineering and human-computer interaction. Her email address is kvmai@torontomu.ca.