VOICE-BASED FRAUD DETECTION USING SPEECH EMOTION RECOGNITION

Fonseka H.F.N.K.

Dr. Rasika Rajapaksha

fonsekah_cs16011@stu.kln.ac.lk rasikar@kln.ac.lk

***Abstract*— In this research, audio signal analysis approach has been used for deception detection. This novel approach will overcome the language barrier. Main functionalities are implemented based on analyzing the vocal tone. This enables the user to avoid the shortcomings of the manual fraud detection procedures and other related investigation procedures. By applying this user can predict whether the speaker is genuine or not. The response can be given in any visualization method for a better experience.**

Keywords— Audio signal analysis, Speech Emotion Recognition, Mel Spectrogram

I. INTRODUCTION

Calls are a common way for criminals to access people's personal and financial information without being suspicious. In present, money lost due to fraud calls has been increased into billions of dollars. Scammers use various methods to trick people into giving up sensitive information like bank details, credit/debit card details and passwords. Therefore, identifying scammers is incredibly important for public safety and financial sector stability[1].

Considering the Insurance industry, it has become extremely popular among lots of people and one of the main difficulties in this sector is Insurance Fraud. Insurance Fraud refers to deliberate acts with the intention of misleading lead the insurer to obtain financial benefits. Common frauds include providing false facts on the insurance application and submitting claims for injuries or damages that have never happened. Most of the times people who commit these crimes are organized as teams or criminal organizations, who carry off massive sums through fraudulent businesses. When it comes to ordinary people who tend to deceive these companies by paying fewer subscriptions, and simulated disasters to make money. So, we can understand the value of speech emotion recognition systems are increasing but due the limitations of resources there are only few researches in the domain of audio-based fraud detection that has been already done[2] .

There are different types of algorithms for fraud detection yet most of those algorithms are based on Natural Language Processing (NLP) and primarily support only the English language. Even though there are some tools developed using NLP to support multiple languages, the performance of these systems are not enough to use in practical scenarios. Hence the existing systems are not being particularly useful to a multilingual society. The outcome of the research includes development of an algorithm using deep learning to classify calls using speech emotion recognition which support many languages.

II. BACKGROUND AND LITERATURE REVIEW

Recently the attention of the researchers has been increased in the Speech Emotion Recognition (SER) mechanism which attempts to recognize emotional states by analyzing the speech signals[3]. Since emotions play an important role in human communication, applications can be developed using SER. For example, designing robots, mobile services, etc. There are diverse ways to represent emotions. For example, we can consider Ekman's six basic emotion model which is able to identify anger, disgust, sadness, fear , happiness, and surprise[4].

To obtain an accurate model, first step is to collect relevant data to train the Machine Learning (ML) model. Several open-source databases can be found on the internet the first option to consider is RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) database. It includes emotional data of 24 different actors and those actors were asked to read two different sentences in many ways in North American English. The spoken styles they used are neutral, calm, happy, surprised, sad, fearful, angry and disgusted. It includes a total of 1440 files with these emotions[5]. The second database is CREMA-D (Crowd-sourced emotional multimodal actors' dataset) database which is similar to RAVDESS. In this, 91 different actors recorded their speech for 12 different sentences with 6 different emotions namely neutral, happy, surprised, sad and disgust it includes 7442 files in total[6]. IEMOCAP (Interactive Emotional Dyadic Motion Capture) database contain a large amount of emotional data and includes audio-visual recordings of five sessions between female and male. It has a recording of 10 speakers total in 12 hours of data [7]. A research paper that was published in 2018 by the Department of Electrical and Computer Engineering at Seoul National University proposed a system for Multimodal Speech Emotion Recognition Using Audio and Text[8]. They proposed a model that uses low-level audio signals as well as high-level text transcription to improve information included within low-resource datasets towards a greater degree. They also proposed a novel deep dual recurrent encoder model that simultaneously improves text and audio data emotion recognition from speech and their model accurately identifies emotion classes.

The recent technological advancements in signal processing, machine learning, and linear algebra have resulted in high-performing voice recognition systems[9]. The applications can be observed in the research[10], as well as in commercial[11]. Modern engines use various models, statistical speaker-dependent models, such as mixtures of Gaussian distributions[12] and speech signal representations

using super vectors[13] or i-vectors[14] combined with powerful machine learning algorithms.

There is an enormous amount of literature that uses machine learning approaches for classification[15] are few among them. The SVM is the most commonly used classifier for global features[16]. Some other classifiers, such as decision trees[17] and K-nearest neighbor (KNN)[18], have also been used in speech emotion recognition. These approaches require very high-dimensional handcrafted features chosen empirically.

Another system was proposed using Dempster-Shafer theory and Bayesian inferencing for superimposed fraud detection in mobile communication networks[19]. They used a supervised neural network to detect fraud in mobile telecommunication networks[20]. The notion of voice call graphs to represent voice calls from domestic callers to foreign recipients and used a Markov Clustering-based method for isolating dominant fraud activities from these international calls[21]. Although many approaches have been proposed to detect telecom fraud, some issues still deserve further consideration.

III. METHODOLOGY

The primary objective of the research is to create a web application based on SER to detect fraud. The proposed system's SER ability will cater to a wide range of people including people with low computer literacy. The proposed theories and mechanisms are capable of analyzing the audios and visualize the emotions with percentages. For the development of a speech emotion recognition model either an existing dataset, or a dataset provided by a specific area, or a sector can be used depend on the availability of the resources. Using any of these options, domain-specific speech recognizer can be built.

RAVDESS dataset is one the best datasets we can use to create the model it includes audio clips of twenty-four actors and includes more spoken style compared to other datasets. Since the objective is to analyze the waves of the audio it should be in wav format and all the audios need to be in the same sampling rate. (Ex: 16000Hz). If the dataset is from organization that has Realtime data, those also has to be converted into the previously mentioned formats.

As it is evident in the literature survey that there has not been done an SER-based Fraud detection system in any domain. There are some functional requirements that has been identified for fraud detection. First, the proposed model should be able to identify the emotions of the speaker. Here, special focus has been given to identify the high-pitched voices and silences to detect whether the audios are suspicious. Next, it must be able to work for different languages other than English. Then the listener should be able to record the call and get the result in real-time. User Interface should be simple and easy to manage.

As non-functional requirements, realtime result has to be given in less time which means model needs to be very efficient. Since the application records and saves privacy should be ensured. When using a database security of the collected data must be ensured.

Here I have used RAVDESS emotional dataset to train and test the model. First few things were followed up before start training the model. The original size of the RAVDESS dataset is around 24GB but I had to use a smaller portion of it due to the lack of computational power and resources. Data preprocessing is very important because the whole system will be dependent on the training data set. I did the preprocessing by identifying the basic features that helps to improve the accuracy of the ML model. Since we are analyzing the waves of the audio all the audios must be in .wav format. This was obtained using the PyAudio library which convert the sample rates into the specific value.

Next was to implement a script to handle this task. Before beginning to extract features, it's a must to understand and analyze the data. First, let's consider these audio files and how to analyze them step by step. The first step is to understand the data we have already downloaded from Kaggle website and GitHub. Sound waves have a sampling rate that indicates the discrete intervals and quality of the audios. Each sample is the amplitude of the wave at some point of time interval or duration, where the bit depth decides how detailed the sample will be. When considering signal processing, sampling indicates the reduction of a continuous signal into a series of discrete values. The number of samples taken in a fixed amount of time is called the sampling frequency or rate. If the audio has a high sampling frequency, then it provides less information loss. But there is s higher computational expense and low sampling. The advantage is they are fast and less expensive for computation.

Sound can be represented as an audio signal having parameters like bandwidth, frequency, etc. Normally, the audio signal can be described as a function of Time and Amplitude. Some devices are built to help you get these sounds and represent those in computer-readable formats. For example, wav (Waveform Audio File) format, WMA (Windows Media Audio) format, mp3 (MPEG-1Audio Layer 3) format, etc. Generally, audio processing involves the acoustics feature extraction relevant to the task and some decision-making tasks such as detection, classification, etc. Fortunately, we have many useful python libraries which make this task reliable and easier.

Spectrograms

It is very rare to take raw audios directly as inputs in deep learning models. The common practice is to convert those audios into a spectrogram. The spectrogram is simply a snapshot of an audio wave and since it is an image, we can input it to a CNN-based architecture which is developed for image handling.

Using Fourier Transforms, Spectrograms can be generated from sound signals. It decomposes the signal into its component frequencies and then the amplitude of every frequency can be displayed. A Spectrogram divides the sound signal duration into smaller time components and then applies the Fourier Transform to those segments, to find the frequencies in that segment. Then it combines the Fourier Transforms for every component into a single plot. Unfortunately, this spectrogram doesn't have much information to observe. This happens because of the way humans perceive sound. Normally humans can hear a narrow

range of frequencies and amplitudes and these frequencies are known as 'pitch'. Humans do not perceive frequencies linearly and humans are more sensitive to differences between low frequencies and high ones.

Mel Spectrograms

A Mel Spectrogram is a bit different from a regular Spectrogram that plots Frequency vs Time. It uses the Mel Scale without using Frequency on the y-axis and Decibel Scale instead of Amplitude which indicates colors.

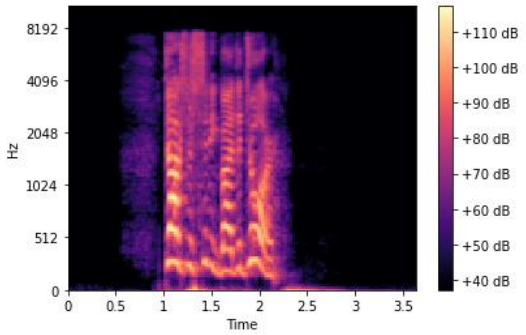

Figure 1 After modifying with Mel-spectrogram

For deep learning models, it is better to use this than using a simple Spectrogram. By modifying it with the Decibel Scale instead of Amplitude we can get a clear result. So, it's better to use the Mel spectrogram than a simple spectrogram.

Now let's consider the backend implementation of the program. As mentioned before the first thing is data preprocessing. Figure 1 shows the data I have collected from Kaggle. It has 24 actors saying some sentences with emotions. Librosa python library was used to extract speech features. Soundfile library has ability to read audio files and os, NumPy, glob, and utils packages for other functionalities like loading data, Calculations, etc. Then to train and test the model we can use sklearn and here are the libraries we can import,

sklearn.model_selection.train_test_splitsplit – This splits matrices or arrays into random train and test subsets. Allowed inputs are pandas data frames, lists, NumPy arrays, or SciPy-sparse matrices and we can decide the size of the test data size and train data size.

sklearn.neural_network.MLPClassifier - This model optimizes the log-loss function using stochastic gradient descent or Limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm.

To train the model, first loaded the dataset with labels and then declared the training sets and testing set. Conv1D was used because its mostly used for the time series data specially for the audio classification. The reason to not choose Conv2D or Conv3D is in Conv2D kernel moves in two directions and it is suitable for image data. In Conv3D kernel moves in three directions and it is mostly used in 3D image data like MRIs and CT scans. Padding has three statuses. "valid", "same" and "casual". First one which is "valid" indicated no padding. "same" is padding with zeros either up or down, or left or right of the input.in that case output has the same dimensions as the input. "casual" is such that current output does not depend on the next input. Rectified linear unit or Relu is the most used activation function in deep learning because it is easy to understand. If the function receives negative input, it gives zero and it the positive inputs are there, then the output returned is the value which was given as the input itself. Dropout is regularization method that can be used to avoid overfitting from the neural network. It randomly disables neurons including their connections correspondingly. This forces all the neurons to learn without depending on a single neuron and generalize better. Root mean squared propagation or rmsprop is used to accelerate the optimization process. It is considered as a gradient descent optimization algorithm. To improve the capability of the algorithm it decreases the number of evaluations to reach the optima.

Next, the implementation of the python flask web application. For the python program I selected PyCharm as an IDE, but this can be implemented using any other python supported IDEs like Visual studio code, Brackets etc. PyCharm IDE gives the privilege to work on an Anaconda environment itself. In brief PyAudio library has been used to record audios because it supports on a variety of platforms and Wave library supports wav sound format and allows to read and write wav files. Another important library is sys which is for system-specific parameters and functions. This module provides access to variables that are maintained or used by the interpreter. Web application can be developed using flask, HTML, and CSS and to provide the analytics such as pie charts or bar charts, matplotlib, and mpld3 can be used. Flask was used because it allows to connect python backend with HTML and CSS. Now let us consider the general imports of main.py. First one is NumPy which is used to create arrays. Normally in Python we have list data type which works as a array but it is not very accurate. NumPy arrays works 50X faster than lists. Pandas module is used to analyze data. It has functions to clean, explore, analyze and manipulate data. Here I have used pandas to create data frames. The python library OS allows to interact with the operating system. This mainly helps to interact with file system. Matplotlib.pyplot is a library that makes it as MATLAB. It includes collection of functions that will help us to create figures, plots, add labels to the plots etc. The next import is dataframe_image which is used to convert dataframes made using pandas into image formats.

Audio imports are the speech_emotion_recognition library which was created manually. Speech_emotion_recognition library will be discussed later this chapter. The math module is a built-in module can be used for mathematical function or tasks. Flask module has many other imports including render_template, session, request, flash, redirect and response. It's a web application framework based on Jinja2 template engine and Werkzeug WSGI toolkit. WSGI or web server gateway interface is considered as a common interface between web application and severs. Werkzeug is a toolkit that implements response, requests and utility functions which allows a web framework to be built on that toolkit. Jinja2 is a template engine which combines web template system with the data source to render a web page. Flask is also referred as microframework which helps to keep the application scalable and simple.

Finally, the predicted emotions will be into a csv and that will be accessed from main.py.So, the technology stack used in the application is,

Programming (Python, HTML, CSS, JavaScript)

Frameworks (Flask, TensorFlow)

Environments (Anaconda, Google Colab)

and as for deployment, we can use Microsoft Azure or Amazon Web Services and for the database development, we can use MySQL, Firebase Databases, etc.

IV. Experiments and Results

The optimized system was validated using available data with different techniques. Mainly I focused on static and dynamic testing for validation. Under the static validation, we have done an informal review, code walkthrough, inspection, technical review, and static code review. Initially tested the system with model accuracy which can be done by the python libraries and the testing data set that was partitioned from the data collected.

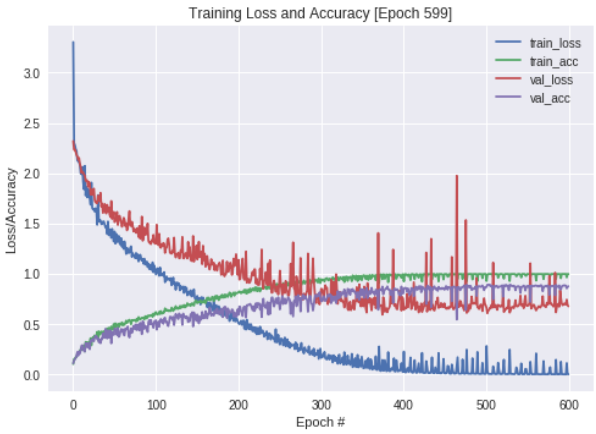

Figure 2 Result after training the ML model

Secondly, after implementation I tested it with audios which was collected from the GitHub project which included some truth audios and lie audios. Expected results was given by the ML model and the performance also could be measured from the dataset because it was a labeled dataset.

The best approach is to test the model performance in the industry with real-time cases and the data that is already classified by the relevant companies and organizations. Since they already have the data records of their solved cases, using this system we can test and validate the accuracy and decide how reliable and efficient it is.

V. Conclusion

This application can predict whether a person is speaking the truth in a particular conversation. In the usual context the fraud detecting application are based on Natural Language Processing but due to the language barrier, there are some limitations those applications. Even the google speech assistant is not being able to catch many languages well. Also, the researches I considered in the previous chapters are based on NLP and signal analysis. There were no certain applications for deception detection using audio signal analysis. Even in the Polygraph test which is used for criminal investigations need a physical contact with the person who is being investigate. So, this research focuses on the idea speech signal analysis and by developing it with a suitable real-time dataset and better technology will help to minimize the fraudulent activities in the industry.

VI. References

[1] R. Derrig, 'Insurance Fraud', Journal of Risk and Insurance, vol. 69, pp. 271–287, Sep. 2002, doi: 10.1111/1539-6975.00026.

[2] J. Jung and B.-J. Kim, 'Insurance Fraud in Korea, Its Seriousness, and Policy Implications', Frontiers in Public Health, vol. 9, Nov. 2021, doi: 10.3389/fpubh.2021.791820.

[3] K. Han, D. Yu, and I. Tashev, 'Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine', p. 5.

[4] M. Piórkowska and M. Wrobel, 'Basic Emotions', 2017. doi: 10.1007/978-3-319-28099-8_495-1.

[5] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, 'The Emotional Voices Database: Towards Controlling the Emotion Dimension in Voice Generation Systems', arXiv:1806.09514 [cs, eess]

[6] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, 'CREMA-D: Crowd-sourced emotional multimodal actors dataset', IEEE transactions on affective computing, vol. 5, pp. 377–390, Oct. 2014, doi: 10.1109/TAFFC.2014.2336244.

[7] C. Busso et al., 'IEMOCAP: interactive emotional dyadic motion capture database', Lang Resources & Evaluation, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.

[8] S. Yoon, S. Byun, and K. Jung, 'Multimodal Speech Emotion Recognition Using Audio and Text', arXiv:1810.04635 [cs]

[9] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, 'Support vector machines using GMM supervectors for speaker verification', IEEE Signal Processing Letters, vol. 13, no. 5, pp. 308–311, May 2006, doi: 10.1109/LSP.2006.870086.

[10] Bergstra, James, Frédéric Bastien, Olivier Breuleux, PascalLamblin, Razvan Pascanu, Olivier Delalleau, GuillaumeDesjardins et al. Theano: Deep learning on gpus withpython. In NIPS 2011, BigLearning Workshop, Granada,Spain. 2011

[11] R. Connaughton, A. Sgroi, K. W. Bowyer, and P. J. Flynn, 'A cross-sensor evaluation of three commercial iris cameras for iris biometrics', presented at the CVPR Workshops, Jan. 2011.

[12] Schoerkhuber, Christian, and Anssi Klapuri. Constant-Qtransform toolbox for music processing. 7th Sound andMusic Computing Conference, Barcelona, Spain. 2010.

[13] S. Safavi, H. Gan, I. Mporas, and R. Sotudeh, 'Fraud Detection in Voice-Based Identity Authentication Applications and Services', 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016

[14] S. Safavi and I. Mporas, 'Improving Performance of Speaker Identification Systems Using Score Level Fusion of Two Modes of Operation', in Speech and Computer, vol. 10458, A. Karpov, R. Potapova, and I. Mporas, Eds. Cham: Springer International Publishing, 2017, pp. 438–444. doi: 10.1007/978-3-319-66429-3_43.

[15] Department of Computer Science, Sri Krishna arts and Science, Coimbatore, Tamilnadu, India., S. Sandhya*,

N. Karthikeyan, Department of Computer Science, Sri Krishna arts and Science, Coimbatore, Tamilnadu, India., R. Sruthi, and Department of Computer Science, Sri Krishna arts and Science, Coimbatore, Tamilnadu, India., 'Machine Learning Method for Detecting and Analysis of Fraud Phone Calls Datasets', IJRTE, vol. 8, no. 6, pp. 3806–3810, Mar. 2020, doi: 10.35940/ijrte.F8751.038620

[16] OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit'.224088060_OpenEAR__Introducing_the_Munich_opensource_emotion_and_affect_recognition_toolkit.

[17] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, 'Emotion recognition using a hierarchical binary decision tree approach', Speech Communication, vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011, doi: 10.1016/j.specom.2011.06.004.

[18] Young, Steve, Evermann, Gunnar, Gales, Mark, Hain,Thomas, Kershaw, Dan, Liu, Xunying (Andrew), Moore,Gareth, Odell, Julian, Ollason, Dave, Povey, Dan, Valtchev,Valtcho, and Woodland, Phil. The HTK book. Vol. 2. Cam-bridge: Entropic Cambridge Research Laboratory, 1997

[19] Zito, N. Wilbert, L. Wiskott, and P. Berkes. Modular toolkit for Data Processing (MDP):a Python data processing framework. Frontiers in neuroinformatics, 2, 2008.

[20] Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal componentanalysis. SIAM Journal on Matrix Analysis and Applications, 31(3):1100–1124, 2009.

[21] Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin, and B. Thirion. A supervisedclustering approach for fMRI-based inference of brain states. Patt Rec, page epub aheadof print, April 2011. doi: 10.1016/j.patcog.2011.04.006

[22] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, 'Emotion recognition using a hierarchical binary decision tree approach', Speech Communication, vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011, doi: 10.1016/j.specom.2011.06.004.

[23] Young, Steve, Evermann, Gunnar, Gales, Mark, Hain,Thomas, Kershaw, Dan, Liu, Xunying (Andrew), Moore,Gareth, Odell, Julian, Ollason, Dave, Povey, Dan, Valtchev,Valtcho, and Woodland, Phil. The HTK book. Vol. 2. Cam-bridge: Entropic Cambridge Research Laboratory, 1997

[24] Zito, N. Wilbert, L. Wiskott, and P. Berkes. Modular toolkit for Data Processing (MDP):a Python data processing framework. Frontiers in neuroinformatics, 2, 2008

[25] Rokhlin, A. Szlam, and M. Tygert. A randomized algorithm for principal componentanalysis. SIAM Journal on Matrix Analysis and Applications, 31(3):1100–1124, 2009.

[26] Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin, and B. Thirion. A supervisedclustering approach for fMRI-based inference of brain states. Patt Rec, page epub aheadof print, April 2011. doi: 10.1016/j.patcog.2011.04.006.