# Learning Models for Emotion Analysis in Crime Scene

K Zin Lin
University of Information Technology
*kzinlin78@gmail.com*

***Abstract— Detecting the emotional state of a person is the important challenging in the audio signal processing research area. The characteristics of the people's voice have several features such as timbre, silence, loudness, energy and voiced tone. In the observation, people express their emotions by changing with different positions and actions depend on their feelings. Emotional meanings of speech are implicitly and automatically recorded after the circumstances, importance and other surrounding details of an event have been analyzed. In this proposed system, the audio features such as pitch, short time energy and loudness are used to make a rules and Hidden Markov Model (HMM) is to categorize the emotion of culprit to get the truth of the crime and to provide the police officer for interviewing.***

## I. INTRODUCTION

Emotion detection is the process of recognizing human emotion, most normally from facial representations as well as from verbal representations. When there is no chance to use the facial image, audio-based emotion detection is inspiring in the research area and it is becoming widespread for researcher. Moreover, the research work such as speaker identification, speech and music classification and audio annotation in soccer video, film can use the audio signal to provide the important information and good accuracy.

But actually, there is an even worse type of criminal liar. That's the person who not only tells untruths to law enforcement and the courts, but also gets others to do so as well. By doing like this, they intend to induce a witness to give false testimony. People who are suborders of perjury are much rarer than the cowardly liar or criminal perjurer. That's because it takes a great deal of persuasion to get another to lie to authorities. One typically hears of this sort of crime - suborning perjury - perpetrated by thugs, since it's always an attempt to get others to lie to protect the real criminal.

Human can express their emotions through speech and this is the main reason why emotion detection is used to speech signal. Emotions come from the inner-feelings of human like happiness, sadness, cheer, depress etc. Wherever people hide their feelings, emotions can't be covered. So, it is an important task to detect the emotions of human from conversation. The people tone and correlated speech models are demonstrated by a number of audio features, they are pitch, loudness or energy, quality of sound, and mood [6]. This propose system is intended to investigate the emotion from asking questioning in the court of law area by using the speech features.

The main challenge of this proposed system is to develop such system which can detect the emotion in such a manner so that it is time saving, efficient accuracy in all test cases In [5], for recognition and evaluation of people's emotions, an algorithmic methodology was presented by P. B. Dasgupta by providing the support of tone of voice and and to setup the real time data for training and testing cases. As the emotion using speech or voice of the human was detected, the recorded dataset have different characteristics like rate, frame-rate, frequency and pitch etc. The proposed system is worked on original voice so different operation was performed to calculate or detect the emotion like separate the original voice and white voice or background voice.

In this paper, the following sections are structured as follows. In portion II, the related work was performed and illustrates how an audio clip is exemplified by pitch and energy and offers an overview of HMM in Section III. In Section IV, the proposed system architecture is expressed and in Section V, experimental study is submitted. Finally, the proposed system is concluded in Section VI.

## II. RELATED WORK

Various types of speech categorization and emotion detection research work are discussed in this review section. The category of audio speech signal is classified automatically corresponding to the features in accorded categorization system.

In [1], anger, neutral, sadness and joy emotions are experimented to get the good quality for emotion detection. The speech signals were considered in the prepared dataset by using the peak-to-peak gap that are achieving from the graphical interpretation of these signals. The parameters that influence on the accuracy of the emotion categorization are linked to get the objective.

In [2], the patient's weight is monitored to control the consequences of diseases and murder depends on the patient's underweighting and overweighting. It was discovered by Body Mass Index from the speech signal and remote monitoring is presented in this paper.

K. Tomba, J. Dumoulin and their group [3] focused on speech assessment to identify applicants stress for the duration of human resources screening interviews. To detect stress deep learning is utilized in speech by implementing the classification features such as the mean of energy, the mean intensity and Mel-Frequency Cepstral Coefficients (MFCCs) To obtain good accuracy scores for stress detection, Neural Networks is used.

Stress cannot be precisely defined in [4]. As response against stress can be changed among people, the results are difficult to understand, every people having a certain behavior towards it. In addition, stress can be expressed with different forms, like mental or emotional. Stress is specified such as a people state in different situations that may affect fear or mental encounter.

speech processing. To enhance human-computer interactions, the intended methodology has been established by cooperating with advanced artificial intelligence systems. In

this paper, normal, angry and panicked emotional states are considered.

N. Hossain, R. Jahan and T. T. Tunka [7] classified happy, sad and angry emotion in this system and for the feature extraction Cepstral Coefficient was used. A fixed valued k-means clustering was used to classify the features for detecting emotions. The data set is taken from eight female and seven male speakers and identified the above state of three emotions.

A hierarchical structure for binary decision tree maps inputs speech data into one the emotion classes via following layer of binary classification. It is very simple method and concentrated at the upper level of the tree to reduce accumulation of error. Lee et. al. [8] used this method for emotion detection.

## III. Background

### A. Audio Feature Extraction

Extracted audio feature vectors that are characterized by audio content are the foundation of audio analysis set of rules. In order to obtain high accuracy for audio categorization, the most important thing is to select the best features and to robust for circumstance changing. These audio features can catch the temporal and spectral characteristics of audio signal.

Audio clip-level features are computed based on the frame-level features and used a clip as the classification entity in our proposed system. For features such as pitch, short-time energy (STE) and loudness, means and standard deviation of all data is calculated as basic clip-level features which are demonstrated how many effectiveness for separating speech, voice, and unvoiced etc.

Pitch is an acoustic phenomenon in the field of an auditor designates melodious timbres to relative places on a musical scale created mainly on their experience of the frequency of shaking. Pitch can be measured by Hertz, but it is developed on the particular perception of a signal. Signal fluctuations can be quantified to achieve a frequency in cycles per second. It is unbiased of the strength or breadth of the sound signal. Rapid oscillations were indicated by a high-pitched sound, although, a low-pitched sound corresponds to gentler oscillations. Speech and musical notes have the complex pitch and they corresponds to the repetition rate of periodic or the communal of the time interval between similar replicating occurrences in the sound waveform.

STE is widely used and the easiest of other audio features. It is also known as volume and it is a consistent pointer for silence detection. Normally STE is estimated by the rms (root mean square) of the signal magnitude within each frame. Voiced sound is determined when the energy is soaring in the ratio of energy method,.

Loudness is usually jumbled with physical measures of sound strength such as sound pressure level (in decibels) or intensity or power. A subjective perception of sound pressure is defined by loudness and can be characterized as the characteristic of auditory phenomenon. Signal can be arranged depend on a level varying from calm to loud. Sound pressure level has been identified by a logarithmic measure of the operative pressure of a sound relative to a reference value and is often evaluated with decibel (dB) units.

### B. Hidden Markov Model (HMM)

HMM is well-defined as the finite state machine with fix number of circumstances. It is arithmetical processes to characterize the spectral properties of voice signal. It was two types of probabilities. There should be a set of inspection or states and there should be a specific state changes, which will define that model at the given state in a certain time.

In HMM, the conditions are not visible directly. They are buried but the output is visible which is dependent on the conditions. Output is generated by probability distribution over the states. The information is offered about the sequence of states, but the parameters of conditions are still hidden.

Extracted features are identified as a series of criminal scene or events and used as the input for the HMM. The temporal pattern of the crime scene should be defined by using these features from audio files. It exemplifies a set of rules and the probabilities of making a change from one state to another state. The usage of HMM in crime case is to train one HMM for each emotion class.

## IV. Proposed System Design

In this proposed system, one of four emotion classes are classified from each audio clip. We should figure out six HMM models for classification to extract the emotion. HMM models are being constructed by using the outputs of feature extraction. The emotions that will be defined by using HMM models are anger, disgust, fear and sadness.

First, the background noise is removed from audio file for feature calculation in this proposed system. The sampling frequency is 22 kHz and bit rate is 128 kbps in each audio file and are considered for audio input file and mono channel is used.

The following step is to evaluate the attributes in two levels: frame-level and clip-level by using Pitch, STE and Loudness which are shown to be helpful for separating audio classes. In this step, each clip is used 0.5 overlap in 1sec audio clips to examine in the audio stream and split 20ms frame with non-overlapping.

Not only for getting the good accuracy and efficiency but also data clearance, the preprocessing state is required in signal and image processing. For this purpose, background noise will be removed in this proposed system and then, audio features will be extracted and employs to be modeled HMM. All of the features have some information but the useful information amount will be different. Some features may contain more data than another's. For training case, time consuming is important so the small feature vector should be used for speedy and the most powerful features should be applied for getting good accuracy and reducing error rate. Therefore, each audio class is modeled by using HMM to get the clean data and effective features. It can also increase the emotion recognition rate.

In this proposed system, there are six HMM models are being constructed. First, HMM1 is constructed by using only one features pitch, HMM2 and HMM3 are also constructed by using STE and Loudness respectively. In addition, HMM4, HMM5 and HMM6s are also modeled with (Pitch, STE), (Pitch, Loudness), and (STE, Loudness) respectively. Finally, this HMM training models are used to detect the emotion and can be compared which model is good for defining emotion.

In modeling the HMM1, the choice of the two parameters, the mean of pitch and the standard deviation of pitch are used. According to the tested result, the means and standard deviation of pitch level is raising in "Sad" but lower in "Disgust". The following Fig. 1. is expressed how to extract pitch level from audio files.

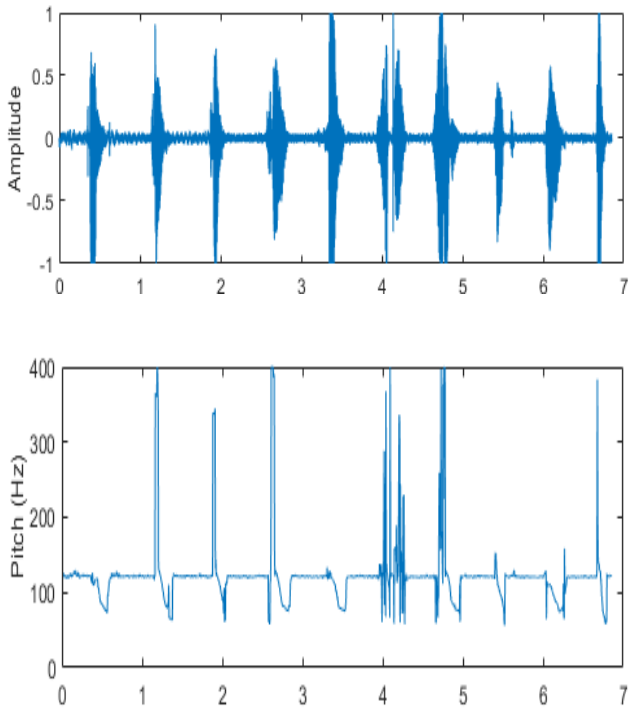


Fig. 1. Extract Pitch frequency of audio signal.

Not only the total spectral power of an audio signal is mostly defined by short time energy but also the volume or loudness is represented by STE. When the sound signal such as speech, silence, voiced and unvoiced are classified and defined in voice activity detection, this feature is widely used. In Fig. 2, the short time energy is used to discriminate the fear emotional state and disgust emotional state.

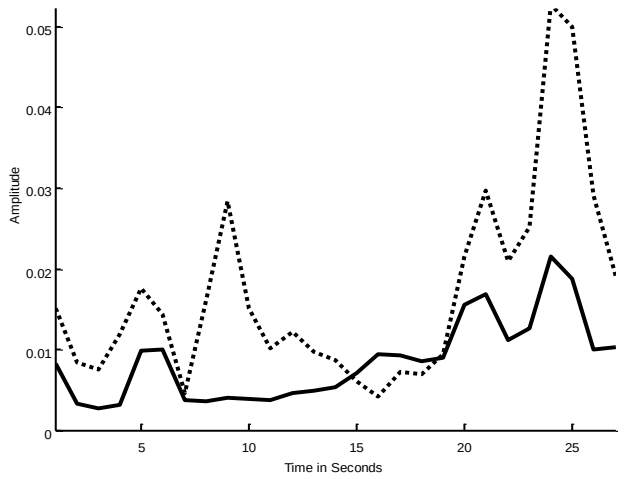


Fig. 2. Short Time Energy for Fear and Disgust Emotion

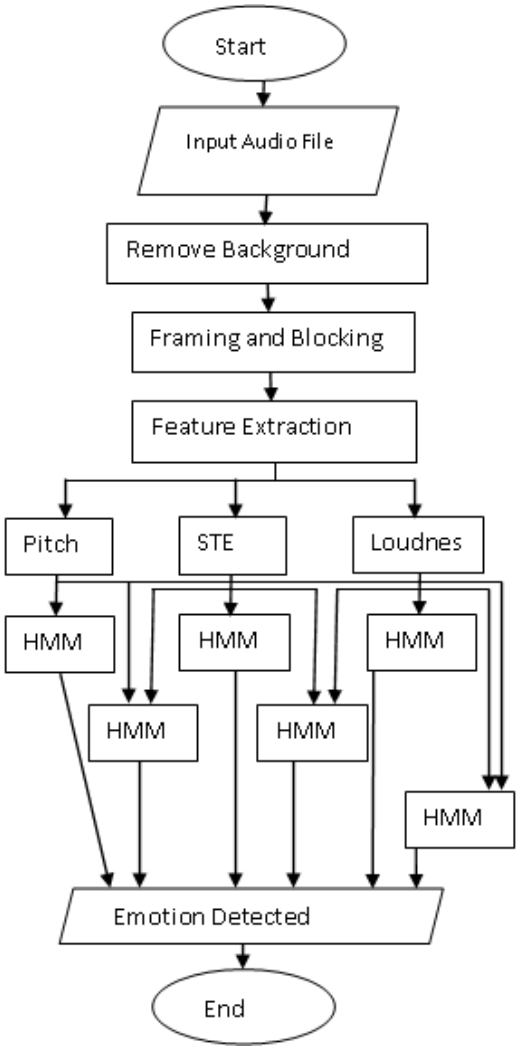


Fig. 3. Proposed System Design.

## V. Experimental Evaluation

### C. Data Setup

The developed HMM models are authenticated by audio files which are recorded from the people's emotion by using the headphone. In total we have used 2000 files for training and testing. In these files there are different emotions and they come from different people, 15 males and 20 females using Myanmar language. These people are under 20 years old and the audio files are recorded with phone, sampled at 22 kHz, and consisted by four emotions. MATLAB R2017a version is used for all experiment in this proposed system.

As the preliminary investigation, 1000 audio files are applied as the training data containing different emotions such as anger, sad, fear and disgust. To gain the ground truth, each file is 10 sec long and emotion is defined by people. In total 2000 audio files, 1000 audio files are divided into training and other 1000 audio files are used for testing. The experiments will be performed with HMM models that are

modeled by extracted features. The total period of all audio files is around 5 hours.

There are many advantages in investigation of human emotions through audio feature analysis and it was proved by doing research in developing conversational and influence skills, particularly face-to-face human communication is not possible or desired [9]. When legal rules or decisions are based on unsupported or mistaken notions of how people behave, justice may be compromised in crime. The human emotion detection via audio signal processing could be various in practical in real world [10, 11].

This proposed system is expressed to momentarily show how much efficient, HMM models are used to extract for emotion. According to a holdout cross-validation, the data set will be carried for training and testing. To check the performance of the detection accuracy is calculated by the number of correctly recognized audio files over total number of files in respective emotion records.

A real-world dataset is experimented in this system and it consists at least over 1 hour of recorded audio files with four different emotions. According to a 2-fold cross-validation, training and testing data are used. The proposed HMM models are used to extract emotional state and compare the result which models are the best to define different emotions. The data for training and testing is expressed in Table I. The features employed for constructing HMM models and the accuracy of extracted emotion are summarized in Table II.

TABLE I. NUMBER OF AUDIO FILES FOR EACH EMOTION

| Emotions | Training | Testing | Total time (hour) |
|---|---|---|---|
| Anger | 300 | 350 | 1.8 |
| Disgust | 200 | 225 | 1.2 |
| Fear | 300 | 250 | 1.5 |
| Sadness | 200 | 175 | 1.0 |

In table I, four emotions will be examined by using 1000 training data set and 1000 testing data set. 300 files are used for training and 350 files for testing to define anger emotional state. In these 300 audio files, each file is 10 s long and recorded from 30 people with anger emotion. For disgust emotional state, 200 audio files are used for training and 225 audio files from different users are tested. For fear and sadness, the training data is 300 and 200 audio files and the testing data is 250 and 175 files respectively.

TABLE II. COMPARISON OF THE RESULT OF HMM MODELS

| Emotions | Accuracy (%) of HMM models | | | | | |
|---|---|---|---|---|---|---|
| | ***HMM 1*** | ***HMM 2*** | ***HMM 3*** | ***HMM 4*** | ***HMM 5*** | ***HMM 6*** |
| Anger | 75.3 | 68.4 | 80.8 | 79.2 | 90.1 | 85.6 |
| Disgust | 70.5 | 60.6 | 52.4 | 82.2 | 75.5 | 73.4 |
| Fear | 85.5 | 87.9 | 60.1 | 92.5 | 80.3 | 76.2 |
| Sadness | 95.5 | 80.1 | 82.5 | 90.3 | 80.5 | 86.3 |

Emotion detection is a very important work for the human being to survive, make decisions and protect his well-being; intuition, perception, understanding and communication are influenced by emotion. There are many benefits to alarm the organism when encountered with important situations. People represent their emotion by using facial or body expression, at that time emotion such as disgust is confused and generally not very accurate to identify through the voice.

The accuracy of HMM models for all emotion classes are compared in Table II. Overall classification accuracy is exceeded average of 90%. Here HMM5 that is constructed by using pitch and loudness can give the accuracy of 90 percent for anger emotional state. For fear and sadness emotional state, HMM4 and HMM1 can define over 90 percent accuracy. But the accuracy of only 82 percent is achieved in disgust state and it is a little bit lower than other emotional states.

The future work of this research is to investigate and analyze the best decision parameter sets for constructing the HMM models and more complicated emotional states should be considered such as surprised, happy, unhappy and cheerful, etc. Performance of disgust emotion could not be satisfied thus conducting more features may solve this problem. Furthermore, only three features were analyzed for detecting various emotional states in this proposed system. So other traditional features could be assessed to advance the investigation and recognition accuracy.

## VI. CONCLUSION

In this research work, the classification of emotion is intended to provide the investigator in criminal case. Those emotions can impact not only the juror's ability to make rational decisions and also the victim's emotion such as sadness, anger, disgust and fear. This work has to focus on those emotions more specifically and the concluding results can help the investigator when the suspect tells the untruth or made up as a liar. We should analyze how specific emotions affect decision-making. Future direction of this work will include to increase more emotional states and make the more useful experimental dataset with incorporation of another audio features.

## REFERENCES

[1] A. Davletcharova, S. Sugathan, B. Abraham and A. P. James, "Detection and Analysis of Emotion From Speech Signals," in Procedia Computer Science, pp. 529–551, June 2015.

[2] B. J. Lee, B. Ku, J.-S. Jang, J. Y. Kim, A nobel method for classifying body mass index on the basis of speech signals for future clinical applications: A pilot study, Evidence-Based Complementary and Alternative Medicine 2013.

[3] K. Tomba, J. Dumoulin, E. Mugellini, O. A. Khaled and S. Hawila, "Stress Detection Through Speech Analysis", in Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018), vol. 1: DCNET, ICE-B, OPTICS, SIGMAP and WINSYS, pp. 394-398, 2018.

[4] T. Johnstone, "The effect of emotion on voice production and speech acoustics", Thesis, 2017.

[5] P. B. Dasgupta, "Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing", International Journal of Computer Trends and Technology (IJCTT), Vol. 52, No. 1, 2017.

[6] T. R. Agus, C. Suied, S. J. Thorpe and D. Pressnitzer, "Characteristics of human voice processing", IEEE International Symposium on Circuits and Systems (ISCAS), pp. 509-512, 2010.

[7] N. Hossain, R. Jahan and T. T. Tunka, "Emotion Detection from Voice Based Classified Frame-Energy Signal Using K-Means Clustering", International Journal of Software Engineering & Applications (IJSEA), Vol. 9, No. 4, pp. 37-44, 2018.

[8] C.C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, 'Emotion Recognition Using a Hierarchical Binary Decision Tree Approach," Speech Commun., Vol. 53, No. 9-10, pp. 1162-1171, Nov 2011.

[9] J. I. Acad, "Voice Fingerprinting: A Very Important Tool against Crime (Review)", Vol. 34, Jan- March 2012.

[10] N. Dave, "Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition", International Journal for Advance Research in Engineering and Technology, July 2013.

[11] J. M. Salerno and L. C. Peter-Hagene, "The Interactive Effect of Anger and Disgust on Moral Outrage and Judgements", in Psychological Science, pp. 2069-2078, Aug 2013.