A Control Device to Monitor Domestic Violence using Speech Analysis

Tanmay Debnath
Undergraduate Student
Pandit Dwarka Prasad Mishra Indian Institute of Information Technology
Design and Manufacturing
Jabalpur, Madhya Pradesh, India

ABSTRACT
This paper demonstrates the implementation of MFCC and HMM modules in voice simulated areas for the prevention and monitoring of domestic violence in real-time. This is based on the system for automatic speech recognition using the hidden Markov model Toolkit (HTK) and Mel-Frequency Cepstral Coefficients (MFCC). The paper also holds an account for improved sound recognition provision using google recognition. The expected billing amount is also presented in this paper, for an approximate view of the product pricings. The device is theorised to function in all environment scenarios. The report has been presented in a detailed manner with all underlying components. The device is purely based on user experience and considering real-life scenarios and test cases.

Keywords
HMM, MFCC, User Experience, Domestic Violence, product Design and Development.

1. INTRODUCTION
The World Health Organisation (WHO) estimated that 35% of women worldwide have experienced some form of violence [2]. The prevalence of reported Domestic Violence among Indian states varies from 6% in Himachal Pradesh to 59% in Bihar [1]. There are various kinds of abuses that Indian women are entitled to. For the simplicity of research, in this case, only a single set of parameters has been addressed. The set includes – Verbal and Physical abuse. The participants stated in the above cited thesis have a mean age of 31 of varied marital statuses. The years invested in the marriage varies from 12 to 24 years. After a close investigation, it has been deduced that the most common thing that connects these cases are the sound energy generated due to the pain of the victims. The automatic analysis of auditory scenes, i.e., detection and classification of audio events or audio context is a growing topic of research. For example, smart house concepts are currently being developed, involving automatic systems for domestic events detection using audio and video data streams. This paper utilises the concepts of *MFCC* (Mel Frequency Cepstral Coefficients) and *HTK* (Hidden Markov Model Toolkit) for the speech recognition and implementation purposes. Upon research and applied results, it has been found that the accuracy of the speech recognition in normal terms for a software-based library, like *Sphinx*, is quite low. The accuracy has to be increased using recursive and intensive learning algorithms. HTK is based on the Hidden Markov Model (HMM).

2. IMPLEMENTATION WORKFLOW:
The main function of the device is to detect the sound and hence identify the words in the sound such that it can take necessary actions against the abuser. The details of the functions have been listed as such:

2.1 Main Function:
The device would work as the medium of detection of the sound waves based on the MFC Coefficients and using the same to learn the Markov Model (HMM). The device has been modelled as such to stay hidden from the eyes of the suspect.

2.2 Subordinate Function:
The device would be held responsible for communication with the authorities regarding lodging a complaint against the abuser for the protection of the consumer. A one-way communication has been implemented that would send information regarding the intensity of sound detected and based on the intensity of the sound and the degree of abuse, the authorities can take actions.

Indian judicial system demands proof for any domestic violence case and this device supports the notion. The device uses the use, keep and throw mechanism with the audio files generated. The device translates the speech-to-text from the audio files generated. The device would then manually search for the abusive words. If the words are not found, the file world automatically get deleted and hence saving memory. If the abusive words are detected, then the file would be saved as a wav file format which would serve as a valuable proof regarding the violence or any crime committed against the victim. After the reporting of the violence, the data can be retrieved by the officials. However, the data will not be accessible by the user. The generated file would contain the date and time inscribed in it and hence would be available for forensic data analysis, regarding the authenticity proof status query.

2.3 Design Workflow:
A descriptive form of the design workflow has been depicted in the form of an illustration.

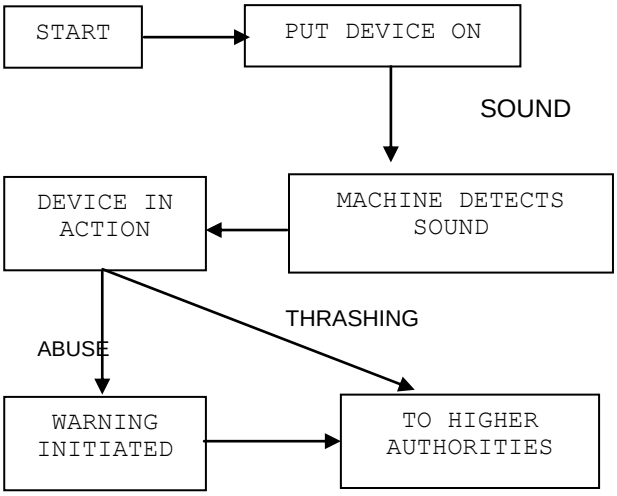

Figure 1: Function workflow of the proposed device

The device has a workflow designed to fit into the normal lives of people and has been designed for the non-complexity architectural analysis. The implementation in real-life would be done in the process described below.

2.3.1 Start

The start of the workflow. Here, the device has been reset to its primitive state. The programming has been installed and the memory has been set to zero state. All necessary modules are implemented at this stage and no further hardware requirements shall be needed in the following next steps of workflow. In another section, the requirements for the manufacturing of this product has been depicted.

2.3.2 Put device ON

The device has been properly installed in to the environment of the household. The device has been trained by using Hidden Markov Models to recognise the sound in household conditions. Since, this is the primary release of the device hence the device has been set to limited usage and household functionalities.

2.3.3 Machine detects sound

At this stage of interaction, the machine detects the presence of sound. The sound here acts as an input for the action of the device. The device has been setup to listen to conversations but only to activate the mechanism only when the threshold of normal speech condition is not fulfilled. According to literature review, the normal sound level in a middle-class Indian family is 51.5dB [3]. The device carries an offset of ±5dB, considering normal situation of differences in any cases. The test cases show the variability index to be transformed and adjusted to the sufficient value for the detection.

2.3.4 Device in action

At this stage, the device is activated and is ready to cope with the abuses and hitting of the abuser on the victim. The device from this segment would work based on the intensity of input. It has been observed that when the intensity of argument happens then due to human nature, the pitch and energy of the sound increases. This feature forms the basis of the manufacturing of the device. The device would calculate the sound energy present in the delivered speech using Python software modules library. Specialised libraries have been designed to study the frequency and energy modulations of the speech in the environment.

```
sample_rate, audio = wavfile.read(TRAIN_PATH)
print("Sample rate: {0}Hz".format(sample_rate))
print("Audio duration: {0}s".format(len(audio) / sample_rate))

Sample rate: 44100Hz
Audio duration: 1.415986394557823s
```

Figure 2: The snippet of code generates the sample rate and audio duration of the wav file generated.

Fig 6 shows the audio sample size that would be used in the entire discussion in this paper. 44,100 Hz is considered to be a reasonable choice for audio processing considering practical environmental situations. Signal would be received from both ends of the spectrum of 22.05kHz, which falls under the normal hearing capability of the humans.

2.3.5 Warning initiated (abuse detection)

When the device detects any form of abuse then it initiates a series of actions which leads to the initiation of warning for the abuser. The device would contain a dictionary containing all the possible abuses in literature. This would form as the source for the detection of the abuses which would be enlisted in the recorded file.

The system would then turn to the final stage of communication and interaction. The user would have no control over this decision. This decision is taken because Indian housewives have the tendency not to inform the higher authorities regarding the abuse. Hence, this step ensures the addressing of such a crime.

2.3.6 To higher authorities

There certain situations may arise when the matter can't be solved in the home premises and hence the authorities need to take out a hand to help the victim. When the device detects scream of high intensity or hitting/thrashing sound then the authorities would be informed regarding the event. The device has a built-in module of Bluetooth which would be in constant communication with a mobile phone to ensure if the crime happens then the cell phone can automatically lodge a complaint. The device would be linked to the personal profile of the consumer using valid National ID. Hence, the address would be automatically available to the higher authorities. Moreover, provisions have been made to verify the address during the installation of the device in the homes itself for the security and authenticity of the user.

3. FUNCTIONING MODULES

The device uses two basic program libraries for its offline functioning and processing. This paper would utilise the concept of MFCC and HTK for the speech pattern recognition and understanding the situation of the environment.

3.1 SPHINX (pocketsphinx)

Sphinx is the python library which would be used in this paper. Precisely the version called pocketsphinx shall be implemented. The Sphinx-4 framework has been designed with a high degree of flexibility and modularity. Sphinx-4 framework consists of three primary modules: The Front-end, the Decoder and the Linguist. One or more input signals are taken by the Front-end and parameterizes them into a sequence of Features. The language model is translated by Linguist, along with the information of pronunciation from the

dictionary and the structural information from one or more sets of Acoustic Models into a search graph [4].

```
from pocketsphinx import LiveSpeech

for phrase in LiveSpeech():
    print(phrase)
```

Figure 3: Code showing the import of pocketsphinx library and usages of LiveSpeech function for the capturing of live sound.

LiveSpeech system recognises the voice in real-time and hence prints the output in real-time. The LiveSpeech contains a built-in dictionary which contains all the words generally spoken in English. The dictionary would serve here as the wordlist, as discussed in the previous section. The dictionary has been edited with sufficient abusive words, that generally are the 'cuss' and hence are not discussed in details.

3.1.1 Working Principle

The pocketsphinx works on the principal of generation of Mel-Frequency Cepstral Coefficient. The input features consisted of independent streams of MFCC features, delta and delta-delta MFCCs and power. The acoustic model uses Hidden-Markov Models with a 5-stateBakis topology and semi-continuous output probabilities [5]. Because audio is a non-stationary process, the FFT will produce distortions. To overcome this, it has been assumed that the audio is a stationary process for a short period of time. Because of that the signal is divided into short frames. Each audio frame will be the same size as the FFT. Also, the overlapping of frames is desired. It is done such that the frames have some correlation between them and because the information is lost on the edges of each frame after applying a window function. The window function divides the normalised audio into tiny segments of stationary audio samples. Upon normalisation of the audio sample, the audio is found to have quite disturbed and uneven frequency spectrum. This is the first hint towards the spectrum energy of the sound generated in a scream. After the normalisation and windowing phase, the generated wave frames go through a more elaborate process that includes Fast Fourier Transformation (FFT), Mel-Frequency warping, generation of Mel filter bank and finally the generation of output as the Mel-coefficients [6,7]. The features are extracted from this step and hence the necessary information is acquired [10, 11]. The explanation of the processes is beyond the scope of the aim of this paper. The papers explaining the mechanism has been cited.

3.2 Hidden Markov Model (HMM)

HTK is a major toolkit used to build Hidden Markov Models (HMM) and to model time series [4]. There are two main phases in a recogniser: training and testing/recognition. The main task of the recognizer is to map **a** phoneme vector and the wanted underlying symbol. Unlike the regular speech recognition system there is no need to worry about one-to-one mapping of symbols to speech as they can give rise to similar speech sounds, and boundary identification problem **is** avoided too by restricting the task to single phoneme recognition. Thus, this will provide the ground-work for people who want to implement complex continuous speech recognition systems [8]. PyHTK has been used, which includes a recently developed Python library and the corresponding Python based ASR pipeline, as well as the new features in HTK 3.5.1 that PyHTK supports. PyHTK enables much easier use of the HTK tools, especially for the design and training of complex ANN architectures. It also contains a distributed SGD training framework using a special type of weighted model averaging [9]. During its history, HTK has integrated many HMM based techniques for the improvement in the speech recognition. Sphinx does not have a great accuracy rate. Using the HMM modules increases the accuracy in the detection of actual spoken words. Recent HTK involves the native support for Artificial Neural Network-based acoustic models. This section introduced the basic model used for increasing the accuracy of speech detection on offline scale. Further explanation is beyond the scope of this paper and has been cited for further detailed information.

4. COMPONENTS:

The components have been selected based on the credibility to function and run a high computation scaled operations in a real-time environment. This section introduces the components used for the invention.

4.1 Raspberry Pi zero W

Based on the requirements of the design and economic feasibility of the product, raspberry pi has been chosen. The product requires a high computation mechanism, since it has to work real-time. This was achievable only with microcontrollers. Raspberry Pi Zero W has been chosen based on the product requirements and the specifications of the microcontroller. The detailed specifications are mentioned. The Raspberry Pi is a low cost, credit-card sized computer that plugs into a computer monitor or TV, and uses a standard keyboard and mouse.

4.1.1 Technical Specifications

The Raspberry Pi Zero W extends the pi family. It has all the features of Raspberry Pi Zero, with additional connectivity the raspberry pi zero has the following specifications:

I. 802.11 b/g/n Wireless LAN
II. Bluetooth 4.1 (BLE)
III. 1GHz Single-core CPU
IV. 512 MB RAM
V. Micro USB Power
VI. Mini HDMI and USB On-The-Go ports
VII. HAT-compatible 40-pin header

The raspberry pi has undergone extensive compliance testing and meets the following European Standards:

I. Electromagnetic Compatibility Directive (EMC) 2014/30/EU
II. Restriction of Hazardous Substances (RoHS) Directive 2011/65/EU [15]

4.2 Microphone (sound recording)

The primary sound recording device used for the recording of the sound and storing the generated data in the memory. Any microphone can be utilised and it can be a user's choice to put the microphone according to choose. For this paper, a local branded microphone, AE Microphone (3.5 mm) has been preferred. Lavalier microphones has been chosen and is preferred for this task because of their flexibility in usages and the range of detection of sounds is admissible, considering the family-size of a middle-class family.

4.3 Audio Driver

Audio driver is used as an interface for the microphones and CPU of the microcontroller. The audio driver has multiple channel input and no output which make it appropriate for the project demonstration. The channel output isn't preferred because the channel is desired only as a source. This would help us in gaining more sound and hence have a better analysis over the receipted conversations and make processing for the analysis of the sound.

5. SYSTEM ARCHITECTURE:

This section introduces the system architecture and how the components are put into action. The proposed model is simplest in its way of design and connections. The model shall be compressed in the form of a box for the practical application and environment. The most probable answer for the compression of the circuit to a more modular format is to design the PCB board, rather than considering the usage of pre-designed PCB computer boards. For the purpose of presentation, the paper utilises the PCB computer board. If the generic PCB board has to be designed, then the most preferred one would the semiconductor MCU boards. They have built-in interfaces for interaction and the boards can be decoded for further research.

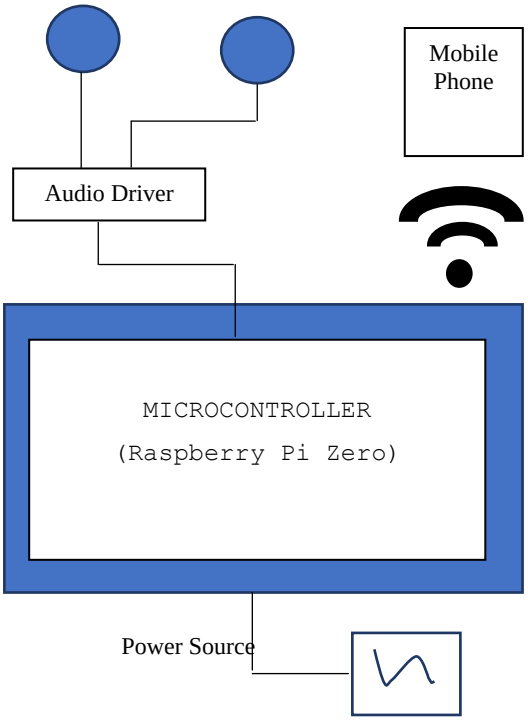

Figure 4: System architecture design for the product. This system contains the minimal arrangement and hence the minimal requirements

The flexibility of the structure is shown in fig 4. The system can function well from a well-hidden place and hence not to come under the vigilance of the abuser. This has been given a protective shield such that it doesn't break even after a fall. However, it is advised to keep the device in a hidden place and hence to make the best out of it. In an unlikely situation of breaking down, the system architecture might break and hence no requirement of buying an entirely different system.

6. DESIGN OF THE PRODUCT:

The prepared design is based on the system architecture designed and data mining of various aspects of the electronics. For the product an audio driver shall be used as well which would be connecting the microphone to the data receiving centre of the pi board. Audio driver has a variety of ranges from the professional tools to project versions. For this paper, the custom version with the dimensions: 28mm X 28mm X 8mm has been used. This dimension is based on the drivers used in small audio projects and are readily available in the market.

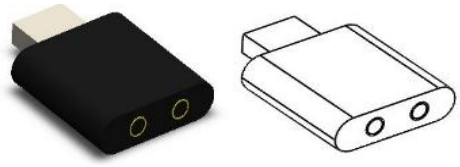

Figure 5: Audio driver representation as a 3D model and wireframe.

The design of the product has been taken as a consideration on data mining.

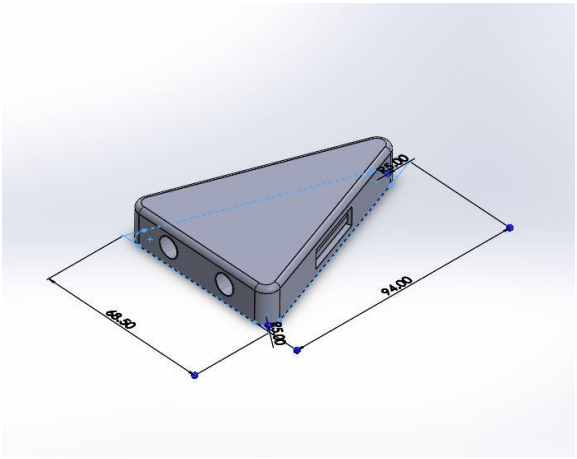

Figure 6: Outlining figure of the finished product depicting the dimensions. The dimensions made on the basis of the net additive dimensions of the components and also considering the international standards of components.

Figure 7: The finished product

The product has been designed considering the technical aspects of the product. For further extension, the designed product has been made as simple as possible for the user to use and understand the concepts behind the same. For primary stages, the design has been considered as the presentable one and upon recursive analysis the designs can be changed and dimensionally improvised.

7. CONCLUSION

The paper concludes with a new invention for the selective action under Domestic Violence. The product aims to curve the present statistics of the rising problems of domestic violence and hence truly respect the place and position of a person in the society.

8. FUTURE SCOPE

There are plenty of room for improvisation of the product and making another version of this demo is quite a viable option.

i. The offline functioning algorithm can be improved. The CMU Sphinx models can be improved by using better HMM models and implementing with further improvised ANN (Artificial Neural Networks).

ii. The design of the product can be aesthetically improved. Considering better test cases, after the initial attempt and understanding the deeper realms of the problems for implementation.

iii. New modules can be designed specifically for the product that would determine the compactness of the product and hence make it a more feasible solution.

iv. The material of the object can be made fire-proof that would be helpful in case of severe emergency.

v. The product can be digitalised and can be made to act as a special function. Dates and time can be shown and can be used as a 'Clock' in disguise. Modes can also be added, such as for party scenes or night time scenes can be implemented.

The main objective of this product is to demonstrate the possibility and utilisation of state-of-the-art technology to implement a viable solution for the prevention of crimes.

9. REFERENCES

[1] Bhandari, S., Hughes, J.C. (2017). Lived Experiences of women facing Domestic violence in India. Journal of social Work in Global Community, vol.2, Issue1. doi: 10.5590/JSWGC.2017.02.1.02.

[2] Avdibegovic, E., Brkic, M., Sinanovic, O. (2017). Emotional Profile of women victims of domestic violence. doi: 10.5455/msm.2017.29.109-113

[3] King G, Roland-Mieszkowski M, Jason T, Rainham DG. J Urban Health. 2012 Dec; 89(6):1017-30. doi: 10.1007/s11524-012-9721-7. PMID: 22707308

[4] Prasanna G. and Ramadass N. (2014). Low cost Home automation using offline speech recognition. International Journal of Signal Processing Systems Vol. 2, No. 2, Dec 2014.

[5] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar and A. I. Rudnicky, "Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, 2006, pp. I-I.

[6] Chakraborty K., Talele A., Upadhya S. (2014). Voice Recognition using MFCC Algorithm. International Journal of Innovative Research in Advanced Engineering., Vol. 1 Issue 10. ISSN: 2349-2163

[7] Dhingra, S.D., Nijhawan, G., Pandit, P. (2013). Isolated speech recognition using MFCC and DTW. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 8. ISSN: 2320-3765

[8] A. G. Veeravalli, W. D. Pan, R. Adhami and P. G. Cox, "A tutorial on using hidden Markov models for phoneme recognition," *Proceedings of the Thirty-Seventh Southeastern Symposium on System Theory, 2005. SSST '05.*, Tuskegee, AL, USA, 2005, pp. 154-157.

[9] C. Zhang, F. L. Kreyssig, Q. Li and P. C. Woodland, "PyHTK: Python Library and ASR Pipelines for HTK," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 6470-6474.

[10] Weimin Huang, Tuan Kiang Chiew, Haizhou Li, Tian Shiang Kok and Jit Biswas, "Scream detection for home applications," *2010 5th IEEE Conference on Industrial Electronics and Applications*, Taichung, 2010, pp. 2115-2120.

[11] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci and A. Sarti, "Scream and gunshot detection in noisy environments," *2007 15th European Signal Processing Conference*, Poznan, 2007, pp. 1216-1220.

[12] (The product website) Source: https://www.raspberrypi.org/products/raspberry-pi-zero-w/