Emergency Detection using audio

Research · November 2020

1 author:

Hirudini Udugama
University of Moratuwa
1 PUBLICATION **0** CITATIONS

SEE PROFILE

Emergency Detection Using Audio

U. K. D. H. T. Udugama
Faculty of Information Technology
University of Moratuwa
Katubedda, Sri Lanka
hirudhini.16@itfac.mrt.ac.lk

***Abstract*— Detection of an emergency can be done through audio. Emergency detection using audio can be processed in three major categories; audio command recognition, emergency detection and sound classification. Filtering sound signals from other surrounding environmental sound signal is an important task. Signals that are related to having the possibility of an emergency should be identified among those sound signals. By classifying acoustic signals, it can be easily understood and detect an emergency situation. For the prevention of an emergency further steps can be taken after communicating emergency situation to relevant parties.**

***Keywords*— Audio Recognition; Emergency Detection; Sound Classification**

I. INTRODUCTION

Smart home concept is a trending market nowadays. Ambient intelligence is an emerging research topic and used in the areas of elderly care, home care and education. Paying attention to acoustics is important because people can monitor their activity. With the emergence of new technologies people tempt to use technological involvement to accomplish basic tasks as well as more comfort and prevention of emergencies. Moreover, among day today activities detection of emergencies is an important fact since after that prevention strategies can be taken to minimize damages.

The system combines an active operating mode with a verbal interface for capturing distress calls and home automation subsystem can be controlled, and proactive mode, a novel algorithm for detection to alert the user in the event of an abnormal acoustic detection. ARM Cortex-A8 CPU, and it is equipped with a LCD monitor, a microphone array, a video camera, and a loudspeaker work as the basic functional unit in LMCU. There are two types of operating modes.

1. Speech monitoring
2. Surveillance monitoring

User can choose from these operation modes [1].

II. AUDIO COMMAND RECOGNITION

Speech Monitoring Mode enables the user to control voice over devices related to home appliances and home automation subsystems. Illuminate the cell phone commands automatically, a list of actions will be converted to a Domotics Gateway code. Speech recognition engine is used to recognize distress calls in an emergency occurs. After the detection, phone call is automatically generated to a relevant person by means of a VoIP stack and an Acoustic Echo Canceller [5]. Low-consuming embedded units are established in the relevant sound domain. All system operations controlled and coordinated by a single central unit. To fulfill the purpose of monitoring and hands free communication an algorithm has been implemented. The results show that enforced power-generalized cepstral constant extraction pipeline improves the word recognition accuracy in quiet and reversed conditions, which grammatical words and phrases will be rejected by the introduction of a "waste phone" within the acoustic model [1].

After the catastrophe, a large number of victims may face serious conditions. It is important to identify the victims and use this method to locate the victims buried among the debris. Victims can be identified by using unmanned aerial vehicles (UAVs). By installing on the UAV there is a sound that the victims can react to, and the victims identify their reactions. By seizing. Victim detection is presented to replace victims in other ways. Compact, low cost, less public restrictions area unit the most important advantages of this technique. By putting in sensors like cameras and GPS modules and integration of mike to the UAV system, its' utility worth is greatly escalated. associate degree Intelligent mechanism is employed and has some characteristics like coming into risk to places wherever anyone can't simply access, discover victims and share location info. Specifically, have a tendency to aim to attain this by employing a quad-copter primarily based UAV [2].

A. THE KINECT DEVICE

Microsoft Kinect is principally use for gesture detection. It's supported Prime sensing element style and is provided with a deep sensing element, a color camera and an electro-acoustic transducer. Depth pictures square measure obtained victimization structural lighting technology. During this technique a beam is felt a grate and divided into totally different beams. The beam is then mirrored by an object within the field of vision (FOV). An infrared red sensing element captures this. victimization triangulation, it will calculate the space of the item. The mike array consists of four microphones which will localize the sound supply [3].

B. EMERGENCY VOICE LEVEL COMBINED RECOGNITION

For any speaker that focuses on limited command-line strategies and is more common in interactive home automation environments, accurate word recognition can be achieved, often at the expense of high computational burden and complexity. The user's emotional state can be combined with audio and video channel interactions to ensure the accuracy of further decisions.

Nowadays, there is a lot of research on affective computing, user emotional modelling and retrieval, and especially emotion (music and speech) data. Recent research has focused on different emotion modelling scenarios, such as fear, stress, and anger, focusing solely on emotional recognition of speech. The proposed approach allows user to consider everyday life-speaking words. The use of widely used verbal commands allows the monitoring mechanism to be fine-tuned and significantly reduced the risk of false alarms. The ability to operate the system in any (digital) microphone capsule, aiming to produce an integrated, intelligent oral sensor that can be integrated with alternative sensing mechanisms in a home monitoring environment. Improving the ability to classify and prioritize them based on the level of speech stress.

The proposed system includes both a general voice / speech recognition functions and a stress assessment system. The identification of the voice is based on the extraction of acoustic signals from short-term spectral features (i.e. Mel frequency capsule coefficients).

Detection of activity is also used to determine moments of silence. The tension-level detector used uses a combination of criteria such as short-term power, instantaneous speed of speech, and fundamental frequency variability. These emotions belong to the same square in the Arousal - valence space used to model human emotions. A series of tests have shown that the mean accuracy of word recognition is close to 90% for stress and stress-free situations. The accuracy of the stress assessment is in the range of 70%, with warnings not identified as a component and the false alarm percentage in the range of 25%. Because of the high correlation between the results and the training dataset, the overall recognition and priority accuracy of the system can be improved with word recordings taken from professional actors. Embedded platforms as part of the intelligent sonic sensor for internal monitoring applications are fundamental to optimizing the real-time performance of the system.

C. AUDIO CONTEXT RECOGNITION

The basis of this method is to visualize each audio context with the aid of a diagram of audio events detected by a supervision classification. During the training part, every context illustrated with the histograms assessed from annotated coaching information. Throughout the check part, individual sound events area unit detected in unknown recordings and a diagram of recording events is formed.

Context recognition is finished by computers. The trigonometric function distance between this bar chart and also the event histograms of every context from the coaching information is calculated. Season frequency coefficient is studied within the frequency-inverse register to manage the importance of the varied events within the bar chart distance calculation. Context recognition is that the method of mechanically determinative the context around a tool. Close data will facilitate wearable devices higher meet users' desires, eg: by dynamic the mode of operation. Compared to image or video sensing, audio has distinctive options. Audio captures data from all directions and is comparatively strong to detector position and orientation, that permits sensing while not distressing the user.

The audio device could give an oversized variety of data which will be associated with location, activity, people, or what that talking concerning. Acoustic ambience and background noise characterize a physical location, such as a car, restaurant or office. Initial hearing tests showed that on average 70% of everyday hearing contexts were detected, and that confusion was present in contexts with predominantly similar hearing patterns. The study suggested that auditory events identified by the auditorium are an important clue to the human perception of the auditory context. However, many of the proposed context recognition systems model the global acoustic features of the audio context rather than sound events.

This approach assumes that totally different contexts, like a street or a restaurant and characterize by the presence of bound sound events. The context may be a diagram of events collected from comment recordings. The projected system is split into 2 phases, noise detection and context recognition. A sound event finding system is employed to detect sound events within the context of testing, and therefore the event result matrix

generated by the detection result matrix is according to the context models. The system is evaluated victimization ten instances that contain identical events [4].

D. MULTIMODAL SPEECH EMOTION RECOGNITION

Speech recognition is a challenging aspect and it is widely relied on formats that use audio features to create a classification that works well. In order to get a better understanding of speech data, a deep dual repetitive coding model that uses simultaneous textual data and audio syntax can be used for this purpose. Because emotional conversation consists of the speech content and sound, this model combines information from these sources to predict the emotional class by encoding information in text and audio sequences using dual Repetitive Neural Networks (RNNs).

This design analyzes speech knowledge from the amplitude to the linguistic level and extends the data of the info to audio-centered formats. In depth tests square measure conducted to research the effectiveness and characteristics of the projected model. Once applied to the IEMOCAP dataset, reflective accuracy from 68.8%, this projected model transcends the previous refined strategies of delivering knowledge to one of four emotional classes (i.e., angry, happy, sad, and neutral) to 71.8%.

In developing emotional intelligence, the first step is to form a sturdy emotional appraisal that performs well in spite of the application. Reconsidering is taken into consideration one of the primary analysis objectives of effective computing. Significantly, speech recognition may be a crucial side of the parallel language field. This field has recently enlarged its application as a result of it may be an essential issue for best human-computer interactions, in addition as fork systems.

The purpose of speech feeling recognition is to predict the emotional content of speech and categorize it per one in all many labels (i.e.; sadness, happiness, moderation, and anger). A range of deep learning techniques are accustomed improve the performance of emotional classification; but, this work continues to be thought-about difficult for many reasons.

First, thanks to the prices related to human intervention, there's low knowledge offered for advanced neural network-based coaching Models. Second, the symptoms of feeling should be learned from low-level speech spells. Feature-based models show restricted talent during this issue.

To overcome these limitations, there is a tendency to propose a model that uses high-level text repetition in addition as low-level audio syntax to use additional of the data contained in low-resource datasets. Given the recent advances in Automatic Speech Recognition (ASR) technology, speech transcription may be performed with extensive talent [5].

The emotional content of the story is clearly emotional words contained during a sentence like "beautiful" and "nice", and it carries robust emotions compared to the normal (non-emotional) words like "person" and "day". Thus, have a tendency to theorize that the speech feeling recognition model are going to be enjoy the incorporation of high-level matter input [6].

III. EMERGENCY DETECTION

In this modality, the system is able to recognize distress calls to provide tele-assistance. This framework gives an answer for emergency detection dependent on sound flag totally coordinated in a low-expending embedded platform. There are two modes.

1. Active operation mode
2. Pro-active mode

In Active Mode, audio signal and voice segments are captured by voice activity detector while detection of commands and distress calls is done by the speech detector in the first operation mode [1].

Earlier perception sensor network (PSN) was developed to recognize information of many peoples' such as WHO, WHAT, and WHERE within its monitoring field. THREE W framework, can be used in several applications. The THREE_ W information also can be sent to remote users by other media formats such as web interface or SMS message. To achieve this objective, the PSN includes multiple units which exists a pan-tilt-zoom camera and a Kinect. To fulfil the purpose of monitoring the entire room multiple PSN units have been used. Objects can be seen from different angles by multiple sensors at simultaneously. Two modules exist in this system.

- ❖ Sound Source Classification (SSC)
- ❖ Sound Source Localization (SSL)

The SSC sound source should be identified as a normal speech that cannot be ignored by either of the emergency classes. The location of the source is estimated by SSL. Once the PSN has identified the emergency, the robot receives a command to notify the emergency and the location. PSN uses SSL result fusion with visual tracking results to identify the person talking to multiple people. Here is a minimal error handling method for visual results across multiple sensors. Events that occur in a

room are identified by audio compilers using a database of five classes, including emergencies [3].

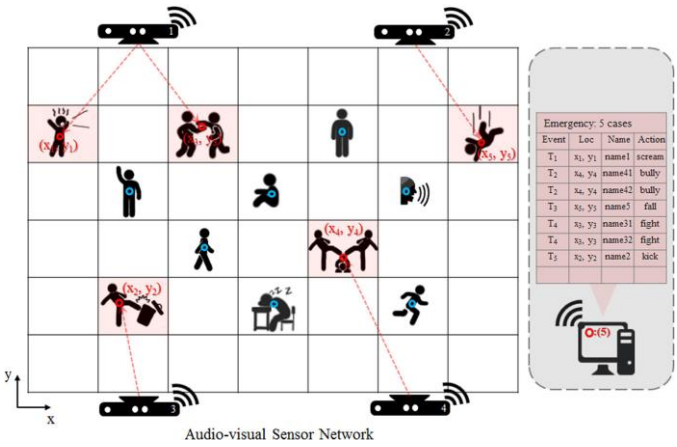

Fig. 1 A basic system overview capable of detecting human-caused multiple emergencies in school environments

A. EMERGENCY DETECTION IN ELDERLY CARE

Using Kinect and also the perception device network (PSN), someone will sight emergencies like screaming. Kinect electro-acoustic transducer array is employed to attain functions of sound supply classification and sound supply localization. To sight that someone is talking between teams, associate audio-visual fusion methodology will be used, combining sound process modules with PSN.

With the windows closed and the radio turned on, sometimes drivers have difficulty detecting vehicle noise and taking appropriate action. This situation is very affecting ambulances. In this case, the technology based on sound signals is used. An ambulance siren can be automatically detected using a radio signal and transmitted and received by an ambulance. Communication between peers is used between cars [7].

The proposed system for Speech-Based System for In-Home Emergency Detection and Remote Assistance contains followings.

- Local Multimedia Control Units (one or more) to monitor the acoustic environment. That contains voice activity detector, an interference removal module, and a speech recognizer.
- Central Management and Processing unit to coordinate the system.

Hand-free communication can be done through a SIP protocol-based VoIP shipment and appropriate acoustic image cancellation algorithm. Reach the overall best performance in terms of F1-Measure. In experiments with noisy words, the implemented PNCC feature extraction pipeline provides significantly higher accuracy than MFCC [8].

Implementing a system to detect urgency or stress levels may respond to common language phrases. A specific account adds those words and assesses the stress level. The application of an intelligent home environment triggers specific impact-priority actions defined in the scope. There is an interactive way to improve circulating intelligence. There is an impactful verbal sensor that integrates with just about any sensor-based application.

B. VOICE ACTIVITY DETECTION

The speaker is bestowed with associate in nursing audio and video activity detector victimization extensive sensors. Its monotonic detectors are supported modeling the temporal variation of audio and video parts victimization hidden mathematician models; which use a post-decision system of reconsideration.

The mel-frequency cepstral constant and vertical mouth gap are the chosen audio and visual options, severally, and amplify each with their 1st order derivatives. The planned methodology for the detection of voice functions victimization audio-visual info is assessed victimization remote recordings from four completely different speakers and underneath different levels of additive white Gaussian noise.

Emerging applications for voice and voice recognition (VAD) systems are close intelligence and sensible areas. These are in distinction to ancient applications (automated speech recognition, mobile phones), that need distance field (FF) sensors for close intelligence applications. As a consequence, captured compilations might suffer from severe close noise and reversal, creating the normal VAD algorithmic program ineffective.

The planned system consists of 2 totally different modes, Audio VAD (A-VAD) and Optical VAD (V-VAD), that care for FF audio sensors. The core of those subsystems depends on the utilization of hidden mathematician models (HMMs): A-WAD uses a combine of HMMs to model speech arrival and speech loss, Associate in nursing V-WAD uses an HMM combine to find vertical lip movements.

The two subsystems meet at totally different call levels to account for the reliability of every outcome. Experiments have shown that this multimedia system VAD will work satisfactorily even underneath adverse sound conditions [9].

❖ VOICE ACTIVITY DETECTOR (VAD)

Voice Activity Detector (VAD) can be used to detect emergency through monitoring the environment. Possible sources of interferences are reduced using an interference cancellation module

- Speech recognition engine based on Pocket- Sphinx - Detects distress calls, able to recognize distress calls as well as voice commands
- Normal Force Normalized Cepstral Coefficients (PNCC) - Increased Strength Against Noise and transformation

When a distress call is detected, system automatically establishes hands-free communication with one of the contacts in the address book. The injured person can then ask for help and explain the reason for asking for help. Evaluates the functioning of the disaster call detection system. Experiments include assessing grammatical refusal abilities and identifying accuracy in noisy and reversed situations.

Automatic detection of alarm sounds helps hearing impaired or distracted people, eg, traveling and contributing to autonomy and safety. The technology is not limited to specific alarms, but can be detected electronically in most MS200.

The main research problem dealing with these works is the selection of a suitable set of audio features. Common features used for these purposes are wavelet properties, Mel-Frequency Cepstral Coefficients (MFCC) and single temporal and frequency characteristics. These works are difficult to generalize because sound classes set is limited and hand-selected.

Thus, a more specialized technique should be developed to create a reliable warning sound detector. With this insight in mind, several previous works have attempted to deal specifically with the task of detecting alarm noise. Because it is difficult to formulate a general model of alarm sounds, many works attempt to identify only the sirens of a particular country's emergency vehicle.

The system operates in real-time by assessing the pitch frequency and comparing it to the siren frequencies defined earlier. Offer generalized alarm detection technology in generalized noisy environments, apart from specific examples used in development to cover most realistic alarm sounds [11].

C. SOUND DETECTION IN MEDICAL TELESURVEY

In medical telesurvey sound detection and classification in a noisy environment and presents a first classification approach. Steps of this procedure are as follows.

- Detection - Significant audio extraction
- Wavelet environment-based algorithm is evaluated in a noisy environment.
- Then the cepstral coefficients based on wavelet are proposed. Results are compared with the most likely parameters [10].

D. DEEP LEARNING BASED EVENT DETECTION

Event detection is part of the environmental identification, and various technologies are announced based on audio signals, image, video and more. Event detection applications include monitoring mobile robots and living environments. However, these methods have a weakness in the area of radio shadows and the lack of information under the action. Therefore, there is some study using multi-modules such as RGB camera or laser camera with deep camera, connected sensors and others audio signal is a strong signal for visual modification and added to the touch. It is therefore used in the field of incident detection and surveillance.

Ultra-wideband (UWB) radar has emerged as a sensor for detecting non-contact and continuous monitoring. But there are problems with the signals and shading areas similar to the two models above. The shaded area and the lack of information problems may complement the use of two sensors simultaneously. So suggest combining two contactless sensors, a microphone and a UWB radar.

Audio signal base model and radar signal base modeling method is a suitable method for combining audio and radar signals to classify the detection of indoor events to solve the problem of similar signal and shadow areas.

The proposed event detection technology has been shown to be superior to single signal event detection. Event detection consists of a classification model that considers each signal's feature extraction and feature vector as input.

The deep learning model was used for both feature extraction and categorization and showed good performance. However, detection of events through one form of signaling generates shadow zones and creates similar signal signaling in different environments, resulting in performance degradation [12].

E. DETECTION OF CONTINUOUS BARKING

The continuous brushing action is unconcealed by audio data and body movements of a dog. This identification technique is predicated on dynamic time deformation (DTW), that has been with success accustomed analyze human audio data. Cyclic body movement was ascertained throughout dog barking. This circular motion is known by an inertial measurement unit (IMU) hooked up to the coat.

An accelerated Fourier electrical device (FFT) is employed to investigate the dog's motion. The projected detection strategies were evaluated exploitation audio and terrorist organization knowledge recorded throughout actual SAR dog coaching sessions. The F-scores of the audio and motion-based bark detection strategies were zero.95 and 0.90, severally. As an experiment, there is a tendency to marked the victim's locations on a map supported body movement.

The study was conducted to boost the flexibility of operating dogs. recommend sound-based and motion-based ways for police investigation continuous bursts. SAR dogs still bark after they square measure found. Audio is a crucial clue for continuous detonating. Additionally, SAR dogs still shake bodies whereas perpetually barking. The motion of the body is another vital clue to the continual burst activity.

Dynamic Time distortion (DTW) was employed in the audio-based methodology to notice continuous bursts. In DTW, the correlation distance between two signals is calculated by extending or catching the audio signal on the time axis. Acceleration frequency is employed in motion-based detection as a result of dogs are ascertained to exhibit cyclic movements throughout continuous burrowing.

An accelerated frequency Fourier electrical device (FFT) is employed to investigate the acceleration frequency. FFT has been wont to analyze foreign terrorist organization information. The correlation distance feature was used for the sample bark audio information calculated exploitation DTW. The F scores for the one and a pair of coaching sessions were 0.89 and 0.98, severally.

The FFT information on the acceleration of the x-z plane was used as an element of the motion. the utmost F-scores for coaching sessions one and a pair of were 0.80 and 0.93, with a 10-second window size and forty thresholds, severally. Compared to two detection strategies, associate in nursing audio-based one can score higher than a motion-based one.

But within the rescue scenario, noise-based detection could also be a lot of prone to noise than motion-based noise. SAR canine maps and locations of victims may be created exploitation motion-based continuous bark detection. The map will tell the placement of the victim and also the mechanical phenomenon of the SAR dog to rescue the staff [13].

IV. SOUND SOURCE LOCALIZATION (SSL)

Kinect's array consists of four microphones in linear shape. Four microphones can simultaneously capture sound signals at a sampling rate of 16 kHz. SSL can be performed on two microphones, but using four microphones increases the localization accuracy over two microphones.

Since the Kinect microphone array has four microphones in the linear array, SSL provides only azimuth angle in front of the sensor. Several ATSUs are placed close to the wall, and zone monitoring is sufficient for our system in front of our Kinect sensors. To cover altitude information or 360 ° asymmetries for SSLs, we can use the microphone array in other formats, such as 3-dimensional cubes or pyramids.

Multiple channel sound signals are extracted at 16,000 Hz and divided into frames, with a duration of 32 ms and an overlap of 20 ms. Multi-channel audio converters into the frequency domain using Fast Fourier Transformer (FFT). Short-term power-based voice activity detection (Gan et al. 2013) is used to determine whether the frame contains only background noise or a target sound source.

Next, the frame of the audio signal for the direction estimation is set by the arrival phase difference (PDOA) algorithm (Nguyen and Choi 2015). In the frequency domain, the arrival phase difference between each pair of microphones is estimated. The PDOA measurement is compared to the expected value in different directions and the direction given by the minimum is determined by the location of the sound source. These estimates were collected for several continuous frames and then clustered to obtain the sound incident direction (asymmetric angle) [14].

V. MALICIOUS AUDIO SOURCE DETECTION

Audio or speech will be simply recorded with the hand held devices of nowadays. It's vital to acknowledge and prove the audio supply once a malicious user employing a movable is recording and enjoying audio from a certified device. Associate in nursing audio watermark will be accustomed determine a malicious audio supply. This audio watermarking system should be sturdy against geometric distortions like continuance modification (TSM), pitch scale, and random cultivation.

Watermark bits are a unit intercalary to the acceptable frequency bands to find a modification within the signal characteristics as they suffer the analog channel. Empirical observations show that if have a tendency to record associate in nursing play audio from an unauthorized supply, the system will find a malicious audio supply. The least common multiple feature-based audio watermark rule, that provides higher performance for geometric distortion caused by DA-AD

conversion, has been selected with associate in nursing audio signal. This rule selects the acceptable embedding band and therefore the embedding multiplier to be accustomed determine malicious audio sources [15].

VI. SOUND CLASSIFICATION

In real time systems sound extraction process is considered as two stages.

- Sound detection
- Sound classification.

Detected sounds are further classified whether it is normal or abnormal. Classification is processed after detection. In recognition step using a statistical study applied to acoustical parameters, user can choose the appropriate parameters that give the best classification results with a GMM system. The entire tele-monitoring system consists of three computers that share information over a control area network bus. The main computer is in charge of data fusion, analyzing the data from the fixed and moving sensors and the sound computer, which constantly monitors the microphone.

The process of sound analysis system is as follows: a message is sent to the main computer when a sound event is analyzed, such as detection time, event type (speech or other noise), localization of the emission source. It should also feature the most classic sound classes. Sound or speech source can be localized by comparing the microphone's noise levels. This can send an alert to the main computer if needed. At the moment, the detection system is only for testing and the detected events are classified by a human operator. This method has been developed for medical monitoring applications within the framework of the DESDIS project, but many of the applications of our audio extraction process include: multimedia document classification, security audio monitoring, medical tele-monitoring. Using speech recognition probability parameters. New parameters such as wavelet-based coefficients have been tested. Identifying a specific alarm keyword is very useful for a data fusion system [16].

VII. COMMUNICATION

The system is integrated to communicate emergencies using VoIP infrastructure. When a distress call is detected, a telephone call is automatically installed and the system enters the state of the call. In this case, the Acoustic Echo Canceller (AEC) ensures proper hands-on communication and the VoIP shipment manages the entire phone call. The home area network is connected to a monitoring unit containing multiple microphones and is capable of communicating with the home automation subsystem. The algorithm has been implemented in a low-consumption embedded platform based on the ARM Cortex-A8 CPU. Two different databases; The effectiveness of ITAAL and A3Novelty effectiveness of standard algorithms has been tested. In the event of a distress call, the user can ask for guidance to receive an alert through hands-free communication, and then the appropriate action can be taken by the addresser [1].

A. COMMUNICATION IN ELDERLY CARE

It is hoped that the system will be able to smoothly communicate with people when considering system which supports man in day today life. For these purposes it is very useful to recognize man's actions in a similar way to man. Understanding of human actions can be took as a pattern recognition of a problem. Human activities identification is divided into two areas.

- Gaining entire body motion data - Can consider different techniques (stereo vision or motion capture or infrared)
- Interpreting human motion – Involves action identification, functional modeling, classification, and component extraction

Second drawback relies on this framework, coaching a model mechanically to partition associated sub-divisions dead reckoning of an action sequence. Audio Visual Content analysis is conferred, that analyzes accounts for audio and video sources and interconnections between those to extraction of high-level linguistics data.

A transmission approach is hoped to supply ends up in higher performance. Data of audio and video play an important role during this task. Though solely visual data doesn't manufacture satisfactory results, this drawback are often avoided through as well as audio information which can contain extra valuable data. For instance, in associate explosion situation will embrace sound of the explosion. The visual content might completely different from one video to a different.

It is hoped that the simplest results are obtained by considering audio and video. This audio and video data combination greatly helps users to access and explore audio elements from the information discourse that means recognition (human action) is additionally thought-about. A linguistics context may be a philosophy that encompasses a significant phase of data. Associate degree instructive question tool is additional sensible than supporting untagged shooting. It's necessary to know the instructive context of multimedia system documents in several areas of multimedia system, like classification, management, compartmentalization and retrieval.

Some are true audio-visual systems, and each strategy is fusion at early fusion (feature level) or delay fusion (decision level). All of those systems try and scan lips in varied ways in which. it's a mouth gap variant that brings helpful data to VAD connected applications. this can be directly derived from the contours of the lips victimization chroma-keys or active appearance models (ANCs), or indirectly supported the looks of the image within the mouth space, principal component analysis (PCA), tissue layer filtration, or distinct trigonometric discrete cosine transformer (DCT). The amount of the mouth isn't enough to work out the activity of the voice. That's the foremost necessary facet concerning visual-only VAD systems. The static PCA coefficients and their initial order temporal distinction are combined with a video feature super vector and shapely by Gaussian mixture models (GMM).

Speech generation may be a two-modal method that transmits each audio and video info. ancient VAD systems think about audio information; Their performance is reciprocally related with the extent and characteristics of the sound. On the opposite hand, strictly visible VAD systems area unit proof against interference noise; but, their performance is subject to many problems, like the camera's angle of vision, poor lighting, and poor quality of the captured pictures.

This system is an audio VAD that mixes the professionals of each ways. FF sensors capture audio and video knowledge. above all, a FF mike and FF video camera area unit accustomed gather the mandatory information. Two classifications of left-right structural HMM-based two-dimensionalization area unit accustomed classify observations as speech or absence. The ultimate call is formed by combining intermediate audio and video module selections.

Based on audio and video speaker, associate economical voice and voice perform detector is conferred that's appropriate for non-invasive intellectual applications. The planned technique, in distinction to the prevailing single-modem approaches, is very economical, even underneath extreme conditions. Each elements implement audio and video and voice perform detectors mistreatment hidden Markoff models.

The elements equipped to the audio and voice activity detector are the Mel frequency cepstra constant of the captured audio signal, and also the input to the audio and voice activity detector is that the movement of the lips detected by the Viola and Jones object detection rule and also the Sobel Filtering Cascade application. The adaptation threshold of each systems has been introduced to trot out completely different environments and a decision-making graded fusion system has been adopted. Emergency detection is done by mensuration noise or sight. Examples embrace associate microcircuit answer for the detection of acoustic signals in emergency traffic, and also the basic mathematical operations like convention and FFT.

Due to size, power, current consumption, and SNR ASICs, the silicon's succeeding application showed sensible properties, however the event team cautioned that any testing should to be done below real conditions. The strategy was used for FFT and once its application the fundamental parameters spectrum was calculated (maximum frequency, minimum, average). The algorithmic program is applied to associate in nursing embedded platform. The results of this technique printed within the article were inadequate as a result of the authors failed to take under consideration the Doppler effect and used the signal filtering. Acoustic protected emergency vehicle detection, supported intelligent transportation systems, uses a mike array that extends across it. Use mathematical strategies like correlation technique, noise delay, least squares method, statistical technique and adaptive filtering. With a comparison of strategies for later detection and use of audio [8].

CONCLUSION

As the popularity of smart homes are increasing, novel facilities are added more and more to provide better comfort to people. With the emergence of new technologies people tempt to use technological involvement to accomplish basic tasks as well as more comfort and prevention of emergencies. Moreover, among day today activities detection of emergencies is an important fact since after that prevention strategies can be taken to minimize damages. After recognizing of audio signals and detecting emergency through classifying sound signals, further it can be communicated for either prevention or minimizing damages. Audio command recognition can be accomplished through Kinect device. Emotion detection is important to rescue victims. In emergency detection VAD plays a major role. Sound source localization and sound classification are some other important tasks. Later on for prevention or minimization of damages can be taken through communicate that emergency situation to relevant parties.

ACKNOWLEDGMENT

This paper reveals some important facts based on some early publications and polishes some point that early speakers had been revealed. The author would glad to thank Dr. S. Sumathipala for his valuable comments and motivation on this task.

REFERENCES

[1] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni and F. Piazza, "An integrated system for voice command recognition and emergency detection based on audio signals," *Expert Systems with Applications*, vol. 42, no. 13, 2015.

[2] Y. Yamazaki, M. Tamaki, C. Premachandra, C. J. Perera, S. Sumathipala and B. H. Sudantha, "Victim

Detection Using UAV with On-board Voice Recognition System," *Proceedings - 3rd IEEE International Conference on Robotic Computing, IRC 2019*, 2019.

[3] Q. Nguyen, S. S. Yun and J. Choi, "Detection of audio-based emergency situations using perception sensor network," *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence, URAI 2016*, 2016.

[4] T. Heittola, A. Mesaros, A. Eronen and T. Virtanen, "Audio context recognition using audio event histograms," *European Signal Processing Conference*, 2010.

[5] G. -, S. Ferdous and F. Makedon, "Multi-modal Person Localization And Emergency Detection Using The Kinect," *International Journal of Advanced Research in Artificial Intelligence*, vol. 2, no. 1, 2013.

[6] G. Sahu, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," 2019.

[7] J. Palecek and M. Cerny, "Emergency horn detection using embedded systems," *SAMI 2016 - IEEE 14th International Symposium on Applied Machine Intelligence and Informatics - Proceedings*, 2016.

[8] E. Principi, D. Fuselli, S. Squartini, M. Bonifazi and F. Piazza, "A speech-based system for in-home emergency detection and remote assistance," *134th Audio Engineering Society Convention 2013*, 2013.

[9] T. Petsatodis, A. Pnevmatikakis and C. Boukis, "Voice activity detection using audio-visual information," *DSP 2009: 16th International Conference on Digital Signal Processing, Proceedings*, 2009.

[10] M. Vacher, D. Istrate and L. Besacier, "Sound detection and classification for medical telesurvey," 2004.

[11] D. Carmel and A. Yeshurun, "Detection of alarm sounds in noisy environments," *25th European Signal Processing Conference, EUSIPCO 2017*, pp. 1839-1843, 2017.

[12] K. Taeho, N. Kyoungjin, K. Jaeha, Y. Jeongnam and C. Joon-Hyuk, "Event Detection Based on Deep Learning using Audio and radar sensors," 2018.

[13] K. Yuichi, O. Kazuaki, F. Takuaki, S. Takahiro and T. Satoshi, "Detection of Continuous Barking Actions from Search and Rescue," 2015.

[14] J. Kheradiya, C. S. Reddy and R. Hegde, "Active Speaker Detection using audio-visual sensor array," *2014 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2014*, no. 2, p. 2, 2015.

[15] B. M. Garlapati and K. R. Kakkirala, "Malicious audio source detection using audio watermarking," *Proceedings - APMediaCast: 2015 Asia Pacific Conference on Multimedia and Broadcasting*, pp. 46-50, 2015.

[16] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane and A. Beghdadi, "Automatic detection and classification of audio events for road surveillance applications," 2018.

[17] P. Murali Krishna, R. Pradeep Reddy, V. Narayanan, S. Lalitha and D. Gupta, "Affective state recognition using audio cues," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 3, pp. 2147-2154, 2019.

[18] N. A. Lili, "VASD: Video action scene detection using audio visual data," *ICCTD 2009 - 2009 International Conference on Computer Technology and Development*, vol. 2, pp. 303-307, 2009.

[19] J. Luo, B. Caputo, A. Zweig, J. H. Bach and J. Anemüller, "Object category detection using audio-visual cues," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5008 LNCS, pp. 539-548, 2008.

[20] C. Okuyucu, M. Sert and A. Yazici, "Audio feature and classifier analysis for efficient recognition of environmental sounds," *Proceedings - 2013 IEEE International Symposium on Multimedia, ISM 2013*, pp. 125-132, 2013.

[21] H. Sun, P. Yang, Z. Liu, L. Zu and Q. Xu, "Microphone array based auditory localization for rescue robot," pp. 606-609, 2011.

[22] I. Dabran, O. Elmakias, R. Shmelkin and Y. Zusman, "An intelligent sound alam recognition system for smart cars and smart homes," *IEEE/IFIP Network Operations and Management Symposium: Cognitive Management in a Cyber World, NOMS 2018*, pp. 1-4, 2018.

[23] H. Lei and O. Valdez, "Special sound detection for emergency phones," *Proceedings - 2013 10th*

International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2013, pp. 816-820, 2013.

[24] H. Christensen, I. Casanuevo, S. Cunningham, P. Green, T. Hain, H. C. Sciences and U. Kingdom, "homeService : Voice-enabled assistive technology in the home using cloud-based automatic speech recognition," *Proceedings of 4th Workshop on Speech and Language Processing for Assistive Technologies(SLPAT)*, pp. 29-34, 2013.

[25] M. Vacher, N. Guirand, J.-f. Serignat, A. Fleury, M. Vacher, N. Guirand, J.-f. Serignat, A. Fleury, N. N. Speech, M. Vacher, N. Guirand, J.-f. Serignat, A. Fleury and N. Noury, "Speech recognition in a smart home : some experiments for telemonitoring To cite this version : HAL Id : hal-00422573 Speech Recognition in a Smart Home : Some Experiments for Telemonitoring," 2009.

[26] D. Hollosi, J. Schröder, S. Goetze and J. E. Appell, "Voice activity detection driven acoustic event classification for monitoring in smart homes," *2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies, ISABEL 2010*, 2010.

[27] B. Fazenda, H. Atmoko, F. Gu, L. Guan and A. Ball, "Acoustic based safety emergency vehicle detection for intelligent transport systems," *Iccas-Sice, 2009*, pp. 4250-4255, 2009.

[28] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J. E. Appell and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 1, no. 40-50, p. 6, 2012.

[29] Q. Nguyen, S. S. Yun and J. Choi, "Audio-visual integration for human-robot interaction in multi-person scenarios," *19th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2014*, 2014.

[30] E. Santos and Y. Zhao, "Automatic Emergence Detection in Complex Systems," pp. 1-24, 2017.

[31] Y. Komori, K. Ohno, T. Fujieda, T. Suzuki and S. Tadokoro, "Detection of continuous barking actions from search and rescue dogs' activities data," *IEEE International Conference on Intelligent Robots and Systems*, 2015.

[32] C. N. Doukas and I. Maglogiannis, "Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, p. 2, 2011.