International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 3537–3548

Applications of Neural Network in Speech Recognition

Sakshi Singh,
Department of Computer Science and Engineering, Dr A.P.J. Abdul Kalam Technical University, Uttar Pradesh, India
s.singh858473@gmail.com

Abstract

Human speech is laced with lots of information. Humans have used the speech as a form of communication since the stone age. In contemporary times, it has become increasingly important to exploit this composite information. The modern machine learning algorithms have allowed us to acknowledge and manipulate speech. Emotion recognition is another very powerful tool in numerous fields such as Human-computer interaction, Psychology, Mass multimedia etc. Speech Emotion Recognition (SER) has been a long-term challenge to the research community to strive upon. In this paper, we have proposed deep learning model for classifying emotions in speech. We have used both audio and text features and compared them minutely. We have laid down an extensive account on the comparison between the audio-based such as Convolutional Neural Network, Gated Recurrent Unit Networks and text-based models like A Long Short-Term Memory Network and a Gated Recurrent Unit Network. We have used IEMOCAP dataset in our experiments and enlisted our findings.

Keywords— *LSTM, GRU, IEMOCAP, Spectrogram.*

I. INTRODUCTION

With the utilization of Machine Learning and Deep Learning, we can recognize, adjust and decipher human discourse. Computerized reasoning has furnished us with an incredible asset to secure and see rational discourse. Speech Emotion Recognition has essential significance in the field of human-PC cooperation as PCs need to get dialogue and feelings also and react to them in like manner which makes SER a significant test in the territory of sign handling.

Our goal in this is to inspect models and their characteristics for the most suitable arrangement of feelings in the IEMOCAP dataset. This dataset has eight unique feelings, yet we will think about just three among them for our simplicity, to be specific "Happy", "Sad" and "Neutral". We present a full record of the similar investigation of various models and their separate productivity.

We will start the procedure of Speech Emotion Recognition by separating significant highlights from sound and text records. These highlights are properties of syllables, rhythms, stress, pitch, tone, Cepstral Coefficients like Mel Frequency Cepstral, Delta MFCC and Spectrograms. The regular decision is MFCC, yet with the assistance of Convolutional Neural Network, we can make spectrograms increasingly practical. In our proposed model, we have utilized Mel scaled Spectrograms as the chief component. The use of word implanting has increased enormous ubiquity lately, and this has been investigated in incredible detail, for instance, slanting hashtags in web-based social networking sites like Instagram, Twitter and so on.

There are countless calculations and models propounded in the field of Deep learning, including Convolutional Neural Network, Recurrent models, for example, Gated Recurrent Unit and remote momentary memory systems. The raison d'etre of this paper is to explain the models utilized in IEMOCAP dataset.

II. COMPARATIVE ANALYSIS

Speech Emotion Recognition (SER) is the field which forces centre around perceiving feeling from discourse signals. To improve hold on the ideas of profound learning and AI, we will break down the related examination work done previously. More accentuation has been laid on separating between ideal acoustic highlights like fundamental frequency, transfer speed, power, length of discourse as they are focal highlights in speech recognition.

ISSN: 2233-7857 IJFGCN

Generally, Hidden Markov Rule was utilized to distinguish feeling from the discourse. The appearance of deep learning has ended up being an aid in this field as it persuaded the researchers to create a different neural system-based design, which prompted improved execution in this field. In the beginning phases of improvement, the deep neural system exhibited their effectiveness in Speech Emotion Recognition by extricating significant level highlights from the speech sample.

As more events happened in profound learning techniques and simple openness of incredible processing stages, we can grow increasingly complex neural systems, for example, Convolutional Neural Network (CNN), Long Short-Term Memory organize, and so on. Significant examination ventures have been created by utilizing these models. The data is gained from crude sound information using spectrograms or audioscriptors. We have employed a mix of these models on our IEMOCAP dataset to get an ideal examination.".

III. IEMOCAP DATASET

We have utilized IECMOCAP dataset in our proposed work. This dataset was discharged in 2008. IEMOCAP is one of the most utilized databases in the field of SER. This database contains nearly twelve hours of various media data. It likewise includes a translation. This database was formulated by scientists at the University of Southern California. The recorded discussion from ten speakers has been arranged into eight feeling names, for example, outrage, happy, amazed, dissatisfaction, energy, dread, pity and nonpartisan.

The assemblage is part in five meetings. Each contains a discussion between two individuals and no two on-screen characters imbricates between these meetings. Every session if further isolated into numerous articulations having a place with every session.

The IEMOCAP database involves scripted and imagined exchanges as recordings. We have utilized translation and sound information.

We have picked discourse articulations from three feeling classes in particular, glad, angry and unbiased. We have consolidated subsiding named "energized" into "cheerful". We have disposed of chronicles that were designated as uncertain.

We have utilized fivefold cross approval and, in each overlay, we have used information from picked meetings for model preparation. Rest of the conference are being used for testing reason. We have first plotted the length and determined the middle worth and chose the ideal outcome in the nearness of significant quality.

The all-out 5584 sound clippings in the IEMOCAP have shifting lengths. We have chosen the term to be five seconds as it is near the middle estimation of span dissemination. We have standardized the translation as crude information is tainted with futile data. This off-base data may cause variations from the norm in the outcome. We have removed the special symbols and abbreviations.

IV. FEATURE EXTRACTION

We have determined experimentally that Mel spectrogram has given better results than Linear spectrogram in our model. We used for Mel spectrogram function for audio processing, and it is found in-built in Librosa library.

We first calculated the Short-Term Fourier Transform of the audio recording followed by calculation of Fast Fourier transform on every window to go from the time to the frequency domain. After this, we scaled the spectrogram to the Mel-scale and then converted the spectrogram to a dB representation.

In table b, we have depicted the parameters used in the spectrogram. We plotted the spectrogram to get the projection of their difference which depend on the emotions.

The goal behind deep learning is that neural network scrutinizes representation of the features, rather than the waiting for the researcher to design it. Accordingly, we have picked word embeddings as our boundaries and refreshed them throughout preparing. We partitioned every articulation into a rundown of names. After this, we characterized a record for each word, and we got to the implanting for each word from |V|×|D|

ISSN: 2233-7857 IJFGCN

grid. In this lattice, 'i' index compares to the ith line of the framework containing its assertion implanting. Here, the inputs are of |V| dimensions. V is the linguistic archive and D is the dimension of the embeddings.

V. THE ARCHITECTURE OF THE NEURAL NETWORK

A. Convolutional neural network

The convolutional neural systems are practically equivalent to the network example of neurons in the human mind. CNN is a calculation which takes input pictures and apportions learnable highlights to different parts of the image so as to separate one from another. The measure of pre-preparing required is extremely low in CNN, and it is exceptionally productive and exact. Convolutional frameworks have accepted an unusual activity all through the whole presence of significant learning. They are a key instance of a compelling utilization of bits of information gained by thinking about the brain to AI applications. They were moreover a bit of the first significant models to perform well, sometime before significant abstract models were seen as appropriate. Convolutional frameworks were in like manner a bit of the first neural structures to understand critical business applications and remain at the bleeding edge of business uses of significant adjusting today. For example, during the 1990s, the neural framework research bundle at AT&T developed a convolutional orchestrate getting checks. Prior to the completion of the 1990s, this system sent by NCR was scrutinizing 10 per cent of the impressive number of tests in the United States. A while later, a couple of OCR and handwriting affirmation systems subject to convolutional nets were sent by Microsoft.

Convolutional systems were likewise used to win numerous challenges. The current force of business enthusiasm for profound learning started when Krizhevsky (2012) won the ImageNet object acknowledgement challenge. Yet, convolutional arrange shad been utilized to win other AI and PC vision challenges with less effect for quite a long time prior. Convolutional nets were a portion of the first working profound systems prepared with back-proliferation. It isn't altogether clear why convolutional networks succeeded when general back-spread orders were considered to have fizzled. Bigger systems likewise appear to be simpler to prepare. With present-day equipment, huge completely associated systems seem to perform sensibly on numerous assignments, in any event, when utilizing accessible datasets and actuation works that were mainstream during the occasions when completely associated orders were accepted not to function admirably. It might be that the essential obstructions to the achievement of neural systems were mental (specialists didn't anticipate that neural networks should work, so they didn't make a genuine eff ort to utilize neural systems).

Whatever the case, it is blessed that convolutional systems performed well decades prior. From various perspectives, they led for the remainder of profound learning and prepared to the acknowledgement of neural networks when all is said in done. Convolutional systems give an approach to practice neural systems to work with information that has an unmistakable matrix organized geography and to scale such models to enormous size. This methodology has been the best on two-dimensional picture geography. To process successive one- dimensional information, we go close to another incredible specialization of the structure of the neural system: repetitive neural networks.

Convolutional Networks, otherwise called convolutional neural systems, or CNNs, are a specific sort of neural network for preparing information that has known lattice-like geography. Models incorporate time-arrangement information, which can be thought of as a 1-D lattice taking examples at regular periods, and picture information, which can be thought of as a 2-D matrix of pixels. Convolutional systems have been colossally fruitful in useful applications. The name "convolutional neural system" demonstrates that the system utilizes a numerical activity called convolution. Convolution is a specific sort of straight activity. These frameworks are additionally called a channel since they act like the exemplary channels in the picture preparing. Be that as it may, in the convolutional neural system, these channels are introduced, trailed by the preparation technique shape channels, which are progressively reasonable for the given task. To make this strategy increasingly helpful, it is conceivable to include more layers after the information layer. Each layer can be related to various channels. The examination into convolutional organize models continues so quickly that

ISSN: 2233-7857 IJFGCN

another best design for a given benchmark is reported like clockwork to months, rendering it unrealistic to depict in print the best engineering

CNN's fundamentally center around the premise that the information will be involved pictures. This centers the design to be set up in the approach to best suit the requirement for managing the specific sort of information. The profundity doesn't allude to the all-out number of layers inside the ANN, however the third element of an actuation volume.

The most significant supposition about issues that are illuminated by CNN ought not to have highlights which are spatially reliant. As it were, for instance, in a face recognition application, we don't have to focus on where the appearances are situated in the pictures. The main concern is to identify them paying little heed to their situation in the given images. Another significant part of CNN is to get conceptual highlights when info engenders toward the more profound layers. For instance, in picture arrangement, the edge may be distinguished in the first layers, and afterwards the more straightforward shapes in the subsequent layers, and eventually the more elevated level highlights, for example, faces in the following segments.

Convolution use three significant thoughts that can help improve an AI framework: scanty communications, boundary sharing and equivariant portrayals. Besides, convolution gives way to working with contributions of variable size. Conventional neural system layers use grid augmentation by a framework of boundaries with a different radius portraying the association between each information unit and each yield unit. This implies each yield unit connects with each info unit. Convolutional systems, in any case, ordinarily haves parse cooperation's (likewise alluded to as inadequate availability or small loads). This is practised by making the part littler than the info. For instance, when preparing a picture, the information picture may have thousands or a huge number of pixels. However, we can distinguish little, significant highlights, for example, edges with portions that involve just tens or several pixels. This implies we have to store fewer boundaries, which both decreases the memory prerequisites of the model and improves its factual efficiency.

Convolutional systems are maybe the best example of overcoming adversity of naturally motivating artificial knowledge. In spite of the fact that convolutional systems have been guided by numerous different fields, a portion of the key plan standards of neural networks was drawn from neuroscience. The historical backdrop of convolutional systems starts with neuroscientific explores some time before the pertinent computational models were created. Neurophysiologists David Hubel and Torstein Wiesel teamed up for quite a long while to decide a large number of the most fundamental realities about how the mammalian vision framework functions. Their achievements were in the long run perceived with Nobel prize. Their findings that have had the best influence on contemporary profound learning models depended on recording the activity of individual neurons in cats. They saw how neurons in the feline's mind reacted to pictures anticipated in exact areas on a screen before the feline. Their extraordinary disclosure was that neurons in the early visual framework responded most unequivocally to very specific examples of light, for instance, exactly situated bars, however, reacted barely at all to other patterns. Their work assisted with describing numerous parts of cerebrum work that are beyond the extent of this book. From the perspective of profound learning, we can focus on a simplified, animation perspective on cerebrum function. There are numerous diff erences between convolutional networks and the mammalian vision framework. A portion of these diff erences are well known to computational neuroscientists; however, outside the extent of this book. Some of these diff erences are not yet known, on the grounds that numerous fundamental inquiries regarding how the mammalian vision framework functions stay unanswered.

As a concise rundown:

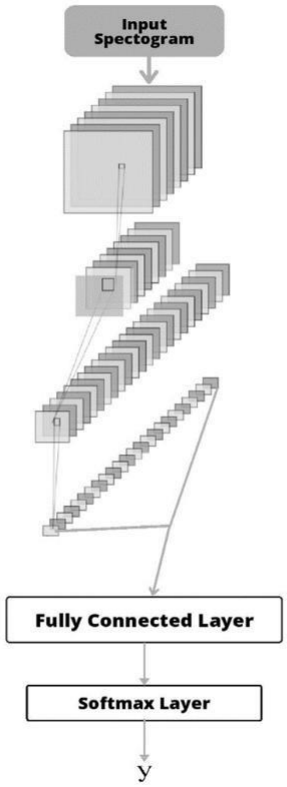
- The natural eye is, for the most part exceptionally low goal, with the exception of a little fix called the fovea. The fovea just watches a region about the size of a thumbnail held at arm's length Recurrent neural networks (RNNs) are models with the capacity to specifically pass data across grouping steps while handling consecutive information for each component in turn. In this way, they can show input as well as yield, which incorporates successions of components that are not free. Further, repetitive neural systems can, at the same time model following and time conditions on

ISSN: 2233-7857 IJFGCN

numerous scales. Most convolutional arranges to get huge full-goal photos as information. The human cerebrum makes a few eye developments called saccades to glimpse the most outwardly salient or task-applicable pieces of a scene. Fusing comparable consideration mechanisms into profound learning models is a functioning exploration bearing. In the setting of deep learning, consideration components have been best for special language handling. A few visual models with foveation components have been grown yet so far have not become the predominant methodology

- The human visual framework is incorporated with numerous different faculties, for example, hearing, and factors like our states of mind and considerations. Convolutional arranges so far are simply visual.
- The human visual framework does significantly more than simply perceive objects. It is able to comprehend whole scenes, including numerous articles and connections among items, and it forms rich 3-D geometric data required for our bodies to interface with the world. Convolutional systems have been applied to a portion of these issues, yet these applications are in their earliest stages.
- While feed forwards IT firing rates catch a significant part of similar data as convolutional organizes highlights, it isn't clear how comparative the middle of the road calculations is.

The cerebrum most likely uses very diff erent initiation and pooling capacities. An individual neuron's actuation most likely isn't very much described by a solitary straight filter reaction. In fact, our animation image of "straightforward cells" and "complex cells" may make a non-existent differentiation; basic cells and complex cells may both be a similar sort of cell however with their "boundaries" empowering a continuum of practices extending from what we call "basic" to what we call "complex.



B. Recurrent neural network

Recurrent neural network (RNNs) are equipped for learning highlights and long-haul conditions from successive and time-arrangement information. Preparing an RNN in an organized manner requires an informational preparation collection of info target sets. It happens commonly because of the enormous number of boundaries that should be enhanced during preparing in RNN over significant stretches of time. The goal is to limit the distinction between the yield and target pairs (i.e., the misfortune esteem) by improving loads of the system. The thought behind RNNs is to utilize serial data.

ISSN: 2233-7857 IJFGCN

RNNs are a class of administered AI models, made of artificial neurons with at least one criticism loops. The input circles are repetitive cycles after some time or arrangement. The capacity to learn successive conditions has permitted RNNs to pick up prominence in applications, for example, discourse acknowledgement, discourse synthesis, machine vision, and video portrayal age. Preparing an RNN is like making a customary Neural Network. We likewise utilize the backpropagation calculation, yet with a little curve. Since the boundaries are shared by unsurpassed strides in the system, the angle at each yield depends not just on the estimations of the current time step, yet besides the past time steps. For instance, to compute the slope at t=4, we would need to backpropagate 3 stages and summarize the inclinations. This is termed as Backpropagation Through Time (BPTT). The vanilla RNNs prepared with BPTT experience issues learning long haul conditions (for example conditions between steps that are far separated) because of what is known as the disappearing/detonating angle issue. There exists some hardware to manage these issues, and particular kinds of RNNs (like LSTMs) were explicitly intended to get around them.

Recurrent neural networks (RNNs) are models with the capacity to specifically pass data across grouping steps while handling consecutive information for each component in turn. In this way, they can show input as well as yield which incorporates successions of components that are not free. Further, repetitive neural systems can at the same time model consecutive and time conditions on numerous scales. A few types of RNNs include –

- Bidirectional RNNs depend on the possibility that the yield at a time may not just rely upon the past components in the arrangement, yet additionally future components. For instance, to anticipate a missing word in an arrangement, you need to take a gander at both the left and the correct setting. Bidirectional RNNs are very straightforward. They are only two RNNs stacked on the head of one another. The yield is then processed dependent on the shrouded condition of both RNNs.

- Deep (Bidirectional) RNNs are like Bi-directional RNNs, just that we currently have numerous layers per time step. By and by this gives us a higher learning limit (yet we likewise need a great deal of preparing information).

- LSTM Networks are very famous nowadays, and we quickly discussed them above. LSTMs don't have an on a very basic level diverse engineering from RNNs, however, they utilize an alternate capacity to process the shrouded state. The memory in LSTMs are called cells, and you can consider them secret elements that take as information the former state and current info. Inside these cells choose what to keep in (and what to delete from) memory. They at that point, join the past express, the current memory, and the information. Things being what they are, these sorts of units are extremely productive at catching long haul conditions.

The capacity of RNNs to perform subjective calculation shows their expressive force. However, one could contend that the C programming language is similarly fit for communicating discretionary projects. But then there are no papers asserting that the innovation of C speaks to a panacea for AI. A crucial explanation is there is no straightforward method of effectively investigating the space of C programs. In particular, there is no broad method to ascertain the slope of a subjective C program to minimize a picked misfortune work. In addition, given any limited dataset, there exist incalculable projects which overfit the dataset, producing wanted preparing yield yet neglecting to sum up to test models.

The key component of a Recurrent Neural Network (RNN) is that the system contains, in any event, one criticism association so that the initiations can stream round in a loop. That empowers the methods to do worldly preparing and learn
arrangements, e.g., perform sequence acknowledgement/generation or transient affiliation/prediction. Recurrent neural system structures can have various structures. One regular sort comprises of a standard Multi-Layer Perceptron (MLP) in addition to included circles. These can misuse the incredible non-straight planning abilities of the MLP, and furthermore, have some type of memory. Others have progressively uniform structures, conceivably with each neuron associated with all the others, and may likewise have stochastic actuation functions. For straightforward models and

ISSN: 2233-7857 IJFGCN

deterministic enactment capacities, learning can be accomplished utilizing comparative angle plummet strategies to those prompting the back-spread calculation for feed-forward systems.

The Universal Approximation Theorem discloses to us that: Any non-direct dynamical framework can be approximated to any precision by a recurrent neural system, without any limitations on the smallness of the state space, given that the system has enough sigmoidal concealed units.

This underlies the computational intensity of repetitive neural networks. However, realizing that an intermittent neural system can inexact any dynamical framework doesn't disclose to us how to accomplish it. Similarly, as with feed-forward neural networks, we, by and large, need them to gain from a lot of preparing information to perform fittingly. We can utilize Continuous Training, for which the system state is never reset during preparing, or Epoch wise Training, which has the system state reset at every age.

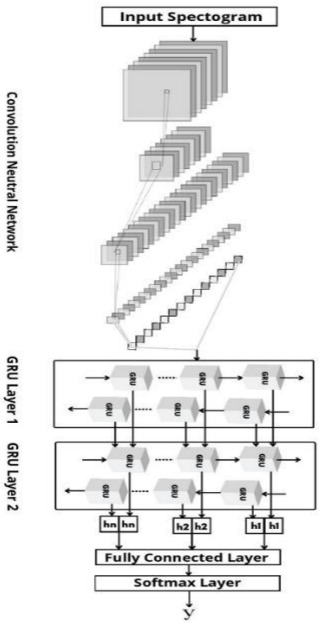
These systems have demonstrated viable in the field of Deep Learning and were what begun a significant part of the current enthusiasm for all types of profound learning neural networks. Such utilization of semi-regulated component learning in in-depth various levelled neural systems is supplanting hand-made element creation in numerous zones. The fast achievement of repetitive neural systems on natural language errands persuades that augmentations of this work to longer messages would be productive. Also, we envision that exchange frameworks could be constructed along comparable standard to the structures utilized for interpretation, encoding prompts and producing reactions while holding the total of discussion history as logical data. This is a significant progressing research territory, with numerous new improvements distributed every year.

C. Bidirectional Gated Recurrent Unit

This unit is an adjusted form of Long Short-Term Memory (LSTM). Disappearing Gradient issue is an obstacle looked by Recurrent Neural Network, and it hinders the learning of great information arrangements. This happens in view of the weight repeating and in the event that the weight repeating is in less amount, at that point, every increase in the inclination turns out to be less. Because of the little propensity, it gets unbending to refresh the load of the system, and there is no significant learning done by the model. To defeat this obstacle; train these vectors with the goal that they can resolve between the pertinent and irrelevant data. We will utilize two bigger Gated Recurrent Unit arrange for our analysis.

D. Combination of Convolutional and Gated Recurrent Unit Neural Network

In this combination model, we process the output of the Convolutional Neural Network by using two-layer Gated Recurrent Unit. Convolutional Neural Network output is of "Height" × "Width" × "Filters" format. We use the width of the production as the time step for Gated Recurrent Unit. Filters and height are used as feature for the GRU.



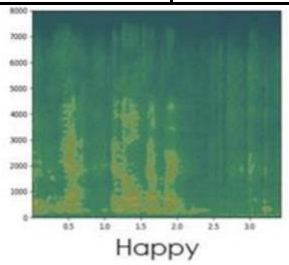
ISSN: 2233-7857 IJFGCN

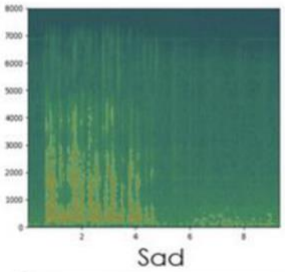
VI. EXPERIMENT

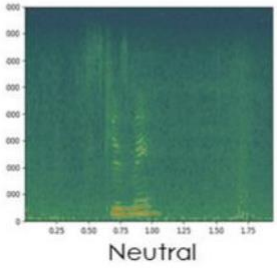
Most of the sample in IEMOCAP dataset are of neutral emotion which makes our dataset imbalanced. Therefore, we will use both weighted and unweighted accuracy in our classification.

A. Result

Out-turn	**Models**		
	Table CNN	***GRU***	***CNN-GRU***
Weighted Accuracy	49.81%	47.23%	51.91%
Unweighted Accuracy	49.95%	43.20%	49.68%







VII. CONCLUSION

In this paper, we reviewed the sound examples present in IEMOCAP dataset. We utilized three sound models for our test and broke down the outcomes acquired. We inferred that Recurrent Neural system end up being wasteful in our investigation. It performed clumsily during highlight extraction and was dominated by Convolutional Neural Network. According to our general perception, the

ISSN: 2233-7857 IJFGCN

IEMOCAP dataset was imbalanced along these lines, we needed to utilize weighted function to forestall overfitting to one class and to separate the extrapolated outcome.

During our trial, we watched the contrast between the working model of Recurrent Neural Network and Convolutional Neural Network about how they process sound records utilizing various structures.

REFERENCES

[1] 1. B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)., vol. 2, pp. II– 1, IEEE, 2003

[2] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Fifteenth Annual

Conference of the International Speech Communication Association,2014.

[3] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in Platform Technology and Service (PlatCon), 2017 International Conference on. IEEE, 2017, pp. 1–5

[4] Aharon Satt, Shai Rozenberg, and Ron Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," Proc. Interspeech 2017, pp. 1089–1093, 2017

[4] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, pages 6645–6649. IEEE

[5] Sepp Hochreiter and Jurgen Schmidhuber. 1997. Lstm can solve hard long time lag problems. In Advances in neural information processing systems, pages 473–479

[6] . Cho, K, van Merrienboer, B, Gulcehre, C, Bougares, F, Schwenk, H & Bengio, Y 2014, Learning phrase representations using RNN encoder-decoder for statistical machine translation. in Conference on Empirical Methods in Natural Language Processing (EMNLP 2014).

[7] Zhou, Qimin and Hao Wu. "NLP at IEST 2018: BiLSTM-Attention and LSTMAttention via Soft Voting in Emotion Classification." WASSA@EMNLP (2018).

[8] S. Yoon, S. Byun and K. Jung," Multimodal Speech Emotion Recognition Using Audio and Text," 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 112-118, doi: 10.1109/SLT.2018.8639583.

[9] S. Yoon, S. Byun and K. Jung," Multimodal Speech Emotion Recognition Using Audio and Text," 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 112-118, doi: 10.1109/SLT.2018.8639583.

[10] Dragos Datcu and L Rothkrantz. 2008.Semantic audio-visual data fusion for automatic emotion recognition. Euromedia'2008

[11] G. Sahu and D. R. Cheriton, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," Tech. Rep. [Online].

[12] Florian Eyben, Martin W¨ollmer, Alex Graves, Bj¨orn Schuller, Ellen DouglasCowie, and Roddy Cowie. 2010a. On-line emotion recognition in a 3-d activationvalence-time continuum using acoustic and linguistic cues. Journal on Multimodal User Interfaces 3(1-2):7–19.

[13] P. Aarabi and S. Zaky. 2001. Robust sound localization using multi-source audio visual information fusion. Information Fusion 2, 3 (2001), 209--223.

[14] M. Andersson, S. Ntalampiras, T. Ganchev, J. Rydell, J. Ahlberg, and N. Fakotakis. 2010. Fusion of acoustic and optical sensor data for automatic fight detection in urban environments. In Proceedings of the 2010 13th Conference on Information Fusion (FUSION). IEEE, 1—8

[15] A. R. Abu-El-Quran, R. A. Goubran, and A. D. C. Chan. 2006. Security monitoring using microphone arrays and audio classification. IEEE Transactions on Instrumentation and Measurement 55, 4 (2006), 1025-- 1032.

[16] Tito Spadini. Sound events for surveillance applications dataset, October 2019.

ISSN: 2233-7857 IJFGCN

[17] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. Springer Series in Statistics. Springer New York Inc., 2001.

[18] Geert Rombouts, Ann Spriet, and Marc Moonen. Generalized sidelobe canceller based combined acoustic feedback- and noise cancellation. Signal Processing, 88(3):571 – 581, 2008.

[19] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(4):320–327, 1976.

[20] Y. Zeng and R. C. Hendriks. Distributed delay and sum beamformer for speech enhancement via randomized gossip. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(1):260–273, 2014.

[21] Iain McCowan. Microphone arrays: A tutorial. Citeseer, 2001

[22] Hamid Krim and Mats Viberg. Two decades of array signal processing research: The parametric approach. Signal Processing Magazine, IEEE, 13:67 – 94, 08 1996.

[23] Jure Murovec, Jurij Prezelj, Luka Curovi ˇ c, and Tadej Novakovi' c. Microphone array based automated environmen- 'tal noise measurement system. Applied Acoustics, 141:106 – 114, 2018.

[24] Simone Scardapane, Michele Scarpiniti, Marta Bucciarelli, Fabiola Colone, Marcello Vincenzo Mansueto, and Raffaele Parisi. Microphone array based classification for security monitoring in unstructured environments. AEU - International Journal of Electronics and Communications, 69(11):1715 – 1723, 2015.

[26] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. Audio surveillance: A systematic review. ACM Comput. Surv., 48(4), 2016

[27] T. D. Räty. Survey on contemporary remote surveillance systems for public safety. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(5):493–515, 2010.

[28] Manjunath Mulimani and Shashidhar G. Koolagudi. Extraction of mapreduce-based features from spectrograms for audio-based surveillance. Digital Signal Processing, 87:1 – 9, 2019.

[29] S. U. Hassan, M. Zeeshan Khan, M. U. Ghani Khan, and S. Saleem. Robust sound classification for surveillance using time frequency audio features. In 2019 International Conference on Communication Technologies (ComTech), pages 13–18, 2019.

[30] G. Fabregat, J. A. Belloch, J. M. Badía, and M. Cobos. Design and implementation of acoustic source localization on a low-cost iot edge platform. IEEE Transactions on Circuits and Systems II: Express Briefs, pages 1–1, 2020.

[31] Tito Spadini, Dimitri Leandro de Oliveira Silva, and Ricardo Suyama. Sound event recognition in a smart city surveillance context. arXiv preprint arXiv:1910.12369, 2019.

[32] Alessia Saggese Nicola Strisciuglio Pasquale Foggia, Nicolai Petkov and Mario Vento. Reliable detection of audio events in highly noisy environments, July 2015.

[33] N. Surampudi, M. Srirangan, and J. Christopher. Enhanced feature extraction approaches for detection of sound events. In 2019 IEEE 9th International Conference on Advanced Computing (IACC), pages 223– 229, 2019.

[34] Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. Signal Processing, IEEE Transactions on, 50(2):174–188

[35] Azlan, M., Cartwright, I., Jones, N., Quirk, T., and West, G. (2005). Multimodal monitoring of the aged in their own homes. In Proceedings of the 3rd Inteernational Conference on Smart Homes and Health Telematics (ICOST'05).

[36] Beal, M. J., Jojic, N., and Attias, H. (2003). A graphical model for audiovisual object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25:828–836.

[37] Carletti, V., Foggia, P., Percannella, G., Saggese, A., Strisciuglio, N., and Vento, M. (2013). Audio surveillance using a bag of aural words classifier. In Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on, pages 81–86.

ISSN: 2233-7857 IJFGCN

[38] Chung, Y., Oh, S., Lee, J., Park, D., Chang, H., and Kim, S. (2013). Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems. Sensors, 13(10):12929–12942.

[39] Couvreur, L., Bettens, F., Hancq, J., and Mancas, M. (2008). Normalized auditory attention levels for automatic audio surveillance.

[40] Cristani, M., Bicego, M., and Murino, V. (2004a). On-line adaptive background modelling for audio surveillance. In Proceedings of International Conference on Pattern Recognition (ICPR 2004), pages 399–402.

[41] Cristani, M., Bicego, M., and Murino, V. (2004b). On-line adaptive background modelling for audio surveillance. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, volume 2, pages 399 – 402 Vol.2.

[42] Dai, C., Zheng, Y., and Li, X. (2005). Layered representation for pedestrian detection and tracking in infrared imagery. In Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on, pages 13–13.

[43] Ellis, D. P. W. (2001). Detecting alarm sounds. In In Proc. Workshop on Consistent and Reliable Acoustic Cues CRAC-2000, pages 59–62.

[44] DiBiase, J., Silverman, H. F., and Brandstein, M. S. (2001). Robust localization in reverberant rooms. In Brandstein, M. and Ward, D., editors, Microphone Arrays, Digital Signal Processing, pages 157–180. Springer Berlin Heidelberg.

[45] Ellis, D. P. W. (2001). Detecting alarm sounds. In In Proc. Workshop on Consistent and Reliable Acoustic Cues CRAC-2000, pages 59–62.

[46] Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., and Theodoridis, S. (2010). Audio-visual fusion for detecting violent scenes in videos. In Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C., and Vouros, G., editors, Artificial Intelligence: Theories, Models and Applications, volume 6040 of Lecture Notes in Computer Science, pages 91–100. Springer Berlin Heidelberg.

[46] Guo, G. and Li, S. Z. (2003). Content-based audio classification and retrieval by support vector machines. Neural Networks, IEEE Transactions on, 14(1):209–215.

[47] Harma, A., McKinney, M., and Skowronek, J. (2005). Automatic surveillance of the acoustic activity in our living environment. In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, page 4 pp.

[48] Hermansky, H. and Morgan, N. (1994). Rasta processing of speech. Speech and Audio Processing, IEEE Transactions on, 2(4):578–589.

[49] Monaci, G., Vandergheynst, P., and Sommer, F. (2009). Learning bimodal structure in audio visual data. Neural Networks, IEEE Transactions on, 20(12):1898–1910.

[50] Mitrovic, D., Zeppelzauer, M., and Breiteneder, C. (2010). Chapter 3 - features for content-based audio retrieval. In Zelkowitz, M. V., editor, Advances in Computers: Improving the Web, volume 78 of Advances in Computers, pages 71 – 150. Elsevier.

[51] Sasou, A. (2011). Acoustic surveillance based on higher-order local auto-correlation. In Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on, pages 1–5.

[52] Raty, T. D. (2010). Survey on contemporary remote surveillance systems for public safety. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(5):493 –515.

[53] Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., and Sarti, A. (2007). Scream and gunshot detection and localization for audio-surveillance systems. In Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on, pages 21 –26.

[54] Zhao, D., Ma, H., and Liu, L. (2010). Event classification for living environment surveillance using audio sensor networks. In Multimedia and Expo (ICME), 2010 IEEE International Conference on, pages 528 –533.

ISSN: 2233-7857 IJFGCN

[55] Zajdel, W., Krijnders, J., Andringa, T., and Gavrila, D. (2007). Cassandra: audio-video sensor fusion for aggression detection. In Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on, pages 200 –205.

[56] Hermansky, H. and Morgan, N. (1994). Rasta processing of speech. Speech and Audio Processing, IEEE Transactions on, 2(4):578–589.

[57] Ito, A., Aiba, A., Ito, M., and Makino, S. (2009). Detection of abnormal sound using multi-stage gmm for surveillance microphone. In Information Assurance and Security, 2009. IAS '09. Fifth International Conference on, volume 1, pages 733 –736

ISSN: 2233-7857 IJFGCN