International Journal of Computer Science and Engineering Communications,
Volume.5, Issue.4 (2017): Page.1688-1695
www.ijcsec.com

Detection of Non-Speech Human Sounds for Surveillance

Rahna.K.M[1]and Baby.C.J[2]

[1]PG Scholar,[2]Assistant Professor,

[1,2]Department of Computer Science and Engineering,

[1,2]Royal College of Engineering & Technology, Thrissur, India.

rahnarazak@gmail.com[1],babycj1120@gmail.com[2]

Abstract: *The objective of this research is to develop feature extraction and classification techniques for the task of Acoustic Event Detection (AED) in unstructured environments, which are those where adverse effects such as noise, distortion and multiple sources are likely to occur. The objective is to design a system that can achieve human-like sound recognition performance on a range of hearing tasks in different circumstances. The research is important, as the field is commonly overshadowed by the more popular area of Automatic Speech Recognition (ASR), and typical AED systems are often based on techniques taken directly from this. However, the direct application presents difficulties, as the characteristics of acoustic events are less well defined than those of speech, and there is no sub-word dictionary available like the phonemes in speech. Therefore, it is relevant to develop a system that can accomplish well for this challenging task.*

Keywords-*acoustic event detection, Feature extraction, Classification, Deep belief networks.*

I. Introduction

Audio information retrieval has been a popular research subject over the last decades and being a subfield of this area where acoustic event classification has a considerable amount of share in the research. There are a number of reasons which set this research apart from the traditional topic of ASR. Firstly, the characteristics of acoustics events differ from those of speech, as the frequency content, duration, and profile of the sounds have a much wider variety than those of speech alone. Secondly, no sub-word dictionary exists for sounds in the same way that it is possible to decompose words into their constituent phonemes. And finally, factors such as noise, distortion and multiple overlapping resources [6] are possible in unstructured environments, whereas speech recognition research has ignored. The purpose of this paper is to propose an accurate system which performs well in real world noisy condition to detect non speech human sounds such as scream, shout, conversation and noise. There are several acoustic event detection systems are existed, in which classifiers such as the Gaussian Mixture Model (GMM) [1], Hidden Markovian Model (HMM) [2][3], Support Vector Machine (SVM) [4], and Artificial Neural Networks (ANN) [5] are used for anomaly detection such as scream, gunshots, noise and explosions in audio signals. The main contribution of this paper is to enable the system to work and predict outcomes correctly in all noisy real world conditions using Deep Belief Network (DBN).

DBN is a generative graphical model, or alternatively a type of deep neural network, composed of multiple layers of hidden units, with connections between the layers but not between units within each layer. DBNs can be viewed as a composition of simple, unsupervised networks such as Restricted Boltzmann Machines (RBMs) or auto encoders, where each sub network's hidden layer serves as the visible layer for the next. This also leads to a fast, layer-by-layer unsupervised training procedure, where contrastive divergence is applied to each sub-network in turn, starting from the lowest pair of layers.

Deep Neural Networks (DNNs) have recently become a popular technique for regression and classification problems. Their capacity to learn high-order correlations between input and output data proves to be very powerful for automatic speech recognition use DNNs to classify the sounds into screams, shouts and other categories. Determining suitable acoustic features for non speech human sound detection is a crucial part. Here uses Mel-frequency cepstral coefficients, are coefficients that collectively make up an MFC, a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition. As for the classifiers, Deep Neural Networks (DNNs), for instance, a combination of Restricted Boltzman Machines (RBMs) and Deep Belief Networks (DBNs), applied on acoustic MFCC features. We set the task as a 4-class classification problem into screams, shout, conversation and noise. This paper organized as follows: Section I contains complete description about this paper, Section Includes the details of proposed method, Section III describes the conclusion.

II. Proposed Method

The detection of no speech human sounds system for defense and surveillance, mainly consist of 3 stages: Noise reduction, feature extraction, and classification. The block diagram of this system is shown in Fig. 2.1.

In the first stage, the pre-processed audio signal is given to noise reduction and the feature is extracted. According to the features, the audio signal is classified as scream, shout, conversation and noises. The steps involved as follows:

3.1 Pre-processing

Input data is usually preprocessed before being fed into the next phase, i.e., feature extraction. Preprocessing techniques are signal processing operations such as filtering, normalization, transformation, trimming, alignment, windowing, offset correction, smoothing etc and depend on the application. Audio pre-processing offers some practical advantages: no waste of time on the processing of silence intervals and no need to process very long sound chunks.

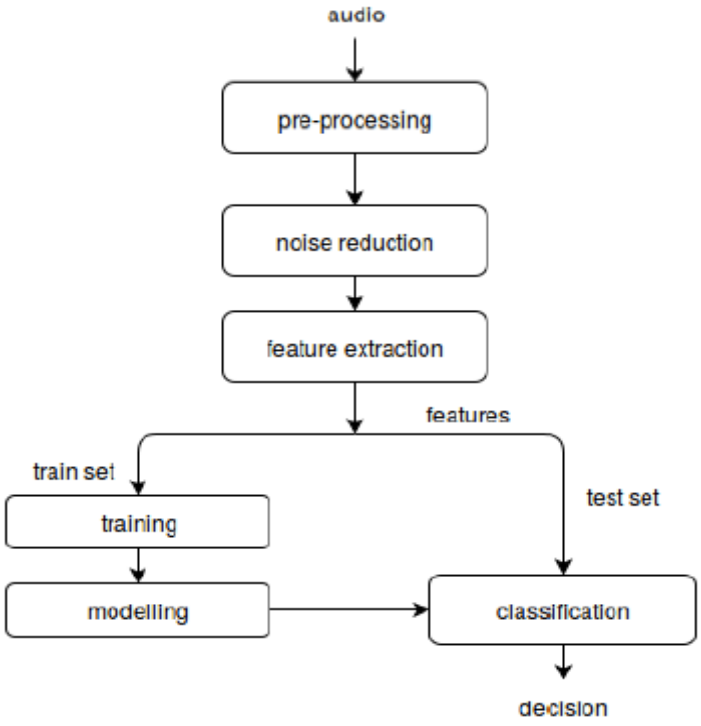

Fig. 2.1 Proposed System

3.2 Noise Reduction

The real-world effects, such as noise, reverberation, and multiple sound sources naturally occur in environments such as meeting rooms, offices and outdoor environments, and humans are amazingly adept at overcoming these issues. For example, the human brain is able to focus on a single speaker in a room full of competing conversations, which would leave most state-of-the-art speech recognition systems struggling. Such a situation is commonly referred to as the cocktail party problem [7]. Many state of the art systems are based on existing AER techniques, which typically have a high recognition accuracy in clean conditions, but poor performance in realistic noisy environmental conditions. If a classifier is trained with varying noise frequencies, from low to high, the rate of detection of pure audio will be less. The noise level of the training database has a significant impact on the performance of the system which allows selecting the most appropriate noise level of the training database for a targeted false rejection rate [8]. This problem can be eliminated by noise cancellation, which worked out.

If a system is subjected to train in different noise conditions, it does not work well in clean conditions and also the training parameters are high. Hence it is difficult to deploy. The noise reduction algorithm removes noises from the input signal. So the feature extraction is done to relevant sounds which eliminate the cocktail party problem and false rejection rate.

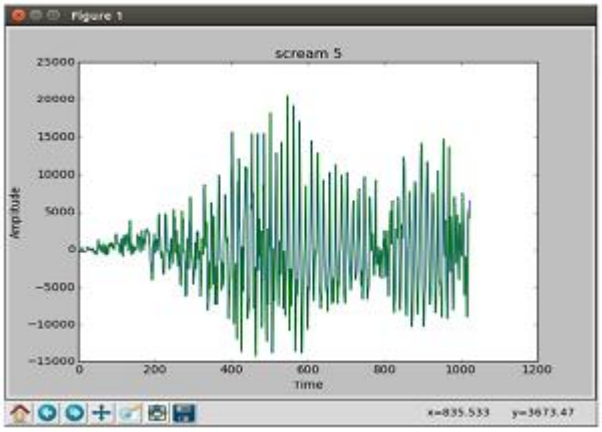

Fig 3.1 Input signal

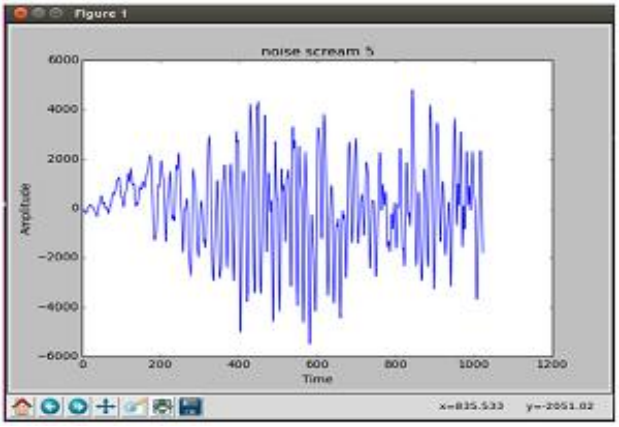

Fig 3.2 Signal after noise removal

3.3 Feature Extraction

Features are higher level representations compared to raw data representations, for example, corners instead of pixels, frequencies instead of raw temporal samples. The Fourier transform is the most popular signal processing method for transforming a time series into a representation of its constituent frequencies. For continuous audio, the Short-Time Fourier Transform (STFT) is often used, which uses a window function, such as the Hamming window, to split the signal into short, overlapping segments, before applying the Fourier transform.

MFCC

It is important to capture the variation of the frequency content over time, in order to fully characterize the sound. Mel Frequency Cepstral Coefficients(MFCC) features represent the frequency content of the sound at a particular instance in time and therefore need to be combined with complex recognizers , which can model the temporal variation of these stationary features. The coefficients are derived from the Mel frequency cepstrum which is a representation of short-time power spectrum of a sound. As the vocal tract shapes the envelope of this spectrum, MFCCs tend to represent the filtering of the sounds by vocal tract. A Mel-frequency cepstrum differs from a regular one as it is linearly scaled in the Mel scale to mimic the human auditory system better.

Energy Spectrum

It is formed by summing the STFT points across frequency in equal-sized blocks. The Energy Spectrum for time step n and frequency index j is:

$$A[n,j] = \sum_{k=0}^{N_f} w_{jk}\, X[n,k]$$

Where $X[n,k]$ are the squared-magnitudes from the Npoint STFT, $N_f = \frac{N}{2} + 1$ is the number of non-redundant points in the STFT of a real signal, and the W_{jk} define a matrix of weights for combining the STFT samples into the more compact spectrum.

3.4 Training

As a supervised system needs to learn the properties of the problem, it requires analysis of examples. Training phase corresponds to the process of learning from labeled data, i.e., training data. It can also be considered as detection of decision boundaries which distinguish different classes in the feature space.

3.5 Classification

The classification has to be performed on the test data, i.e., the data which has not been available to the training phase. The test data represents the observations unseen to the system and the system's performance of generalization is based on the evaluation of the classification phase.

DNN

Deep Neural Networks (DNN) intend to reproduce the mechanism with which the human brain processes information. It involves a network of individual's cells, called units. The term "deep" owes to the fact that the cells are organized in multiple layers stacked onto each other, forming a deep architecture. Conceptually, the units represent hidden causes or factors to the input data. Their output represents the probability of the associated factor to have caused the occurrence of the input data. DNN is defined as a multilayer perceptron with many hidden layers, whose weights are fully connected and are often initialized using either an unsupervised or a supervised pretraining technique. Deep Neural Networks (DNNs) have recently contributed to significant progress in the fields of computer vision, speech recognition, natural language processing and other domains of machine learning. The superior performance of neural networks can be largely attributed to more computational power and the availability of large quantities of labeled data. Neural network models with millions of parameters can now be trained on distributed platforms to achieve good generalization performance [1].

Deep Neural Networks yield the best ratio of sound classification accuracy across a range of computational costs, while Gaussian Mixture Models offer a reasonable accuracy at a consistently small cost, and Support Vector Machines stand between both in terms of compromise between accuracy and computational cost [9].

We use Restricted Boltzmann machine (RBM)s stacked into Deep Belief Networks (DBN)s, and turned into DNNs. DNN is a hybrid deep networks, where the goal is discrimination which is assisted, often in a significant way, with the outcomes of generative or unsupervised deep networks. This can be accomplished by better optimization and regularization of the deep networks for supervised learning. The goal can also be accomplished when discriminative criteria for supervised learning are used to estimate the parameters in any of the deep generative or unsupervised deep networks.

RBM

RBMs are stochastic neural networks with one layer of hidden units which have undirected connections with visible units. The interaction between visible and hidden units is modeled through an Energy function, which associates a scalar energy to each configuration of visible-hidden variables, given by:

$$E(v,h;\theta) = \sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}v_j h_j + \frac{1}{2}\sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j$$

Where $w_{ij}$are the weights between the visible v_i and the hidden h_j units, V is the number of visible units and H is the number of hidden units, b_i and h_j are bias terms, and represents the parameters of our model. The joint distribution over the visible units v and the hidden units h is given by:

$$p(v,h;\theta) = exp\frac{\left(-E(v,h;\theta)\right)}{Z}$$

where Z is a normalization factor. The probability over the visible units is obtained by marginalizing out the hidden units:

$$p(v;\theta) = \sum_{h} exp\frac{\left(-e(v,h;\theta)\right)}{Z}$$

In order to learn the parameters of the model, the greedy learning method where the layers of the network are learned one by one, freezing the weights of all the layers that have already been learned. The greedy algorithm:

1. Learn w_0 assuming all the weight matrices are tied.
2. Freeze w_0 and commit ourselves to using w_0^T to infer factorial approximate posterior distributions over the states of the variables in the first hidden layer, even if subsequent changes in higher level weights mean that this inference method is no longer correct.
3. Keeping all the higher weight matrices tied to each other, but untied from w_0, learn an RBM model of the higher-level data that was produced by using w_0^T to transform the original data.

Deep Belief Networks

The probabilistic generative models composed of multiple layers of stochastic, hidden variables. The top two layers have undirected, symmetric connections between them. The lower layers receive top-down, directed connections from the layer above. First RBM learns a hidden representation of the data, and each RBM after that one takes the hidden layer of the previous one and learns a hidden representation of it. In the end, the deeper layers represent more abstract concepts, or features, associated with the input data.

To classify sounds in to detect screams and shouts, which need to create a model that performs well at classifying. Therefore, one solution is to turn our DBN, after we have learned its parameters, into a DNN. Because DNNs are discriminative classifiers, they learn to model the frontier between classes, through a discriminative rule based on gradient descent and error propagation. The most common algorithm used in DNN training is called the back-propagation algorithm. Practically, the last layer of N sigmoid units is added to perform the calculation of the probability for the input to belong to a

class, N being the number of classes we want to classify. Then feed generatively initialized network with labeled data, and use the back-propagation algorithm to minimize the classification error.

III. Conclusion

AED systems are usually deployed on embedded hardware, which imposes many constraints such as noises, multiple overlapping sources and computational cost. To perform well in noisy conditions, the recognition system must have been explicitly trained in similar noise and SNR conditions. However, this is unrealistic for real-world applications, as the conditions where the system will finally be deployed are often not known in advance. One solution is to perform multi-conditional training, where the system is trained on a variety of different noise conditions, at different SNRs, so that acoustic models contain some knowledge of how the signals might be received in a novel noise environment. However, this requires a large amount of training data, and often reduces the recognition accuracy under clean conditions, due to the reduced discrimination between the noisy acoustic models**.** Hence a noise reduction algorithm is implemented along the DBN. The problem of multiple overlapping sources can be eliminated by using DNN. DNNs perform better than GMM, SVM and HMM for this problem. But the computational cost of DNN is high. In future, to reduce computational cost, using dedicated software like TensorFlow, Keros which are dedicated for neural nets. Finally, will try to increase the number of classes to detect a larger and more specific set of events (siren, gun shots, explosions, etc.) which is helpful both in defense and surveillance.

REFERENCES

[1] American Pierre Laffitte, David Sodoyer, Charles Tatkeu, Laurent Girin, "Deep Neural Networksfor Automatic detection of scream and shouted speech in subway trains",*IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),Shanghai,pp*. 6460-6464, 2016 .

[2] AnnamariaMesaros, Toni.Heittola, AnttiEronen, and T. Virtanen, "Acoustic event detection in real-life recordings", *18th European Signal Processing Conference, (Aalborg, Denmark), pp*. 12671271, 2010.

[3] D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similarsources",*IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, pp*. 1-5, 2015.

[4] Anil Sharma and SanjitKaul,"Two-staged supervised learning based method to detectscreams and cries in urban environment*", IEEE ACM Trans.Audio,speech and language processing., vol.* 24, no. 22, Feb. 2016.

[5] OguzhanGencoglu, Tuomas Virtanen, HeikkiHuttunen: "Recognition of acousticevents using deep neural networks*", EUSIPCO pp*. 506-510,2014.

[6] Rahna K M and Baby C J,"A survey on scream detection methods*",International Conference on Advanced Computing and Communication Systems (ICACCS), pp*.1948-1952, 2017.

[7] A.S. Bregman, "Auditory scene analysis: The perceptual organization of sound", *TheMIT Press*, 1994.

[8] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 1306–1309.

[9] SiddharthSigtia, Adam M. Stark, SachaKrstulovic, Mark D. Plumbley, "Automatic environmental sound recognition"performance versus computational cost*", IEEE/ACM Trans Audio,Speech And Language*,2016.

ABOUTAUTHOR(S)

Rahna K M pursued Bachelor of Technology from CUSAT University, in 2013. She is currently pursuing Master of Technology under APJ Abdul Kalam Technological University, Kerala, India. Her main research work focuses on Signal Processing and Machine Learning.

Baby C J pursued Bachelor of Technology from Calicut University, in 2013 and Master of Technology under CUSAT University, in 2015. She has published several papers in reputed journals. Her main research work focuses on Natural Language processing, Machine learning and Data mining.