SPECTROGRAM-BASED CLASSIFICATION OF SPOKEN FOUL LANGUAGE USING DEEP CNN

Abdulaziz Saleh Ba Wazir [1], Hezerul Abdul Karim [1], Mohd Haris Lye Abdullah [1], Sarina Mansor [1], Nouar AlDahoul [1], Mohammad Faizal Ahmad Fauzi [1], John See [2]

Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia [1]
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia [2]

***Abstract*—Excessive content of profanity in audio and video files has proven to shape one's character and behavior. Currently, conventional methods of manual detection and censorship are being used. Manual censorship method is time consuming and prone to misdetection of foul language. This paper proposed an intelligent model for foul language censorship through automated and robust detection by deep convolutional neural networks (CNNs). A dataset of foul language was collected and processed for the computation of audio spectrogram images that serve as an input to evaluate the classification of foul language. The proposed model was first tested for 2-class (Foul vs Normal) classification problem, the foul class is then further decomposed into a 10-class classification problem for exact detection of profanity. Experimental results show the viability of proposed system by demonstrating high performance of curse words classification with 1.24-2.71 Error Rate (ER) for 2-class and 5.49-8.30 F1-score. Proposed Resnet50 architecture outperforms other models in terms of accuracy, sensitivity, specificity, F1-score.**

***Index Terms*— Foul language, Speech detection, Censorship, Spectrogram, CNN.**

I. INTRODUCTION

Filtering audio content has become one of the society's major concern since many young viewers has portable and instantly accessible source of screen time. Such audio content could be a standalone audio files or a part of a video content (e.g. movies). The responsibility of providing a zero-foul language content rely on broadcasting companies and video creator. Censorship has been a bottleneck situation for filtering language content to suit viewers as it cost large amount of manpower, money, and time. In addition, human factor such as fatigue also contribute to the failure in securing safe audio content. Hence, an automated detection and censorship models are required for audio and speech content.

Deep learning-based speech recognition has received a surge in interest in recent years. Speech recognition systems works by detecting different type of utterances, which are isolated word, connected word, continuous speech, and spontaneous speech, , [1]. This work attempts to develop an intelligent censorship system to detect unwanted speech content accurately. In the literature, intelligent speech recognition models solely focus on the recognition of inoffensive language [2]–[5], while this experiment highlights the classification of foul words by mean of convolutional neural network (CNN) models. The contribution of this work is experimentally investigating the use of CNNs applications, proving that using of shallow and deep CNNS with spectrogram images of spoken terms can serve the purpose of speech censorship.

Speech and acoustic recognition methods use (i) acoustic features extraction, and (ii) classifiers. Feature extraction includes the Mel-Frequency Cepstral Coefficients (MFCCs) [6],[7], spectral features [8],[9], Linear Prediction Coefficients (LPF) [10], Linear Spectral Frequencies (LSF) [11],[12], Discrete Wavelet Transform (DWT) [13]–[15], and spectrogram image [16]–[19]. On the other hand, classifiers include the Hidden Markov Model (HMM) [7], K-nearest neighbor (KNN) [20], Support Vector Machine (SVM) [21], Convolutional Neural Network (CNN) [16], [17], Recurrent Neural Networks (RNNs) [22]–[24]. Image classification field drastic improvements in recent years have led researchers to use image recognition techniques for speech/sound classification and detection.

Transfer learning is defined as the information and knowledge transfer from the source domain to the target domain. This technique has been studied and employed in various machine learning applications. Previous studies have also implemented the use of transfer learning for acoustic processing and classification, where the performance of an acoustic model with limited amount of data have been improved by retraining another acoustic model trained on a large dataset, producing higher accuracy.. Transfer learning of CNN models has been analysed for acoustic event classification (AED) [25], and environmental sounds classification [26], where sounds are plotted into spectrogram images as input to CNNs for classification purpose. Therefore, this work proposes to investigate the use of speech spectrogram images and CNNs for foul language detection application.

978-1-7281-9320-5/20/$31.00 ©2020 IEEE

The first aspect in this experiment is data rarity of offensive language. Therefore, a new database of profanity speech was collected with different accents to design a realistic and robust detection system. All recording used are real and not generated by simulated acoustic mixes or synthetic signals. Secondly, the imbalance of data samples as spoken foul language data samples existence is way lesser compared to normal conversational speech.

In our work, speech spectrogram images are computed for each speech segment to produce the spectrogram frames that are fed to CNN. 2-class classification task was carried out, with target classes defined as foul (comprising 9 foul words with derivatives) and normal conversational speech. Another task of 10-way classification was performed on 9 foul subclasses and one normal class.

This paper is structured as follows: Section 2 details the database used for the experiment. Section 3 briefs on the proposed method. Section 4 presents the experimental settings, while Section 5 reports the results and discussion. Finally, Section 6 draws conclusions of the current work.

II. DATASET

The dataset used is a compilation of offensive language collected through recordings from different locations with different backgrounds to measure the robustness of the proposed model during system evaluation. The 9 classes of profanities are “Asshole”, “Balls”, “Bastard”, “Bitch”, “Cock”, “Cunt”, “Dick”, “Fuck”, and “Pussy”, which are categorized as foul language and content to be censored based on Malaysian Film Censorship Act 2002 [27]. The following are the notations of offensive words for a decent representation of the samples throughout this paper: “A: asshole”, “B: balls”, “Ba: bastard”, “Bi: bitch”, “C: cock”, “Cu: cunt”, “D: dick”, “F: fuck”, and “P: pussy”.

Normal class samples were collected from Freesound, a dedicated sharing site for audio samples [28]. Normal class is represented as “N” in the following sections of this paper.

A. Data Annotation

Data were manually labelled in two stages; the first stage was to label the entire dataset into two-category classification (foul vs. normal) and the second stage was to decompose foul class into the 9 classes for a specific annotation. The subclasses definition of offensive data can help in detecting the exact profanity, as opposed to a two-category classification (foul vs. normal).
The dataset is divided into three sets: training, validation, and test sets. Table I shows the distribution of 3105 foul language samples and 5100 normal samples across the training, validation and test sets for each main class and subclasses. For example, foul class of 3105 samples contain at least 345 data points of each of the 9 subclasses under foul category. Though there are large normal samples available, we choose to use just a portion of the available normal dataset to mitigate the difference in data samples caused by the imbalance of data between normal samples as a class and each of the profanity subclasses. For example, training set for each profanity subclass contain 207 samples and total of 1863 samples, while normal train set contain 3060 samples. The dataset was arranged into 5 folds for cross-validation.

III. METHODS

The present method used convolutional neural networks (CNNs) and spectrogram images for the classification of foul language and normal speech using transfer learning. This section briefs on the general principles of the proposed CNN models and spectrogram for this applicative work.

A. Convolutional Neural Network

CNN is made of stacked of convolutional and fully connected layers with subsampling or pooling layers in between [29]. CNN works by passing many types of filters sliding in horizontal and vertical lines over an image to pick up different signals. These signals help in mapping the different portions of image features and to train classifiers on the target application.

The proposed CNN models for the classification of foul language are well-known Deep CNNs including Alexnet [30], VGG16 [31], Googlenet [32], and Resnet50 [33]. As transfer learning is the transfer of knowledge from source domain (Image classification of Imagenet) to target domain (foul language classification from normal speech), the four CNN architectures knowledge are transferred either by finetuning all layers or freezing some layers’ weights. The knowledge transfer broadens the use of image recognition applications.

TABLE I: Data samples distribution of all classes

	Train	**Validation**	**Test**	**Total**
Foul/ (subclass)	1863/ (207)	621/ (69)	621/ (69)	3105
Normal	3060	1020	1020	5100

B. Spectrogram

The target speech is characterized by spectral content that is usually represented by a vector of consecutive spectral coefficients. Spectral feature vectors are formed using

analysis window slid over a portion of its size with another overlapping window to ensure spectral features extraction of all frames. The total number of spectral coefficient vectors representing a given speech varies based on the duration and characteristics of target speech [34].

The spectrogram images of speech are computed based on the total duration of the speech clips, the duration of each spectrogram frame, the time shift between each spectrogram frame, and the number of frequency bands. Fig. 1 presents samples of foul language spectrograms of two different foul words.

Spectrograms that are formed of vectors of coefficients are fed into CNNs that consider the temporal dimensions of images input. The convolution process then extracts full features of the spectral content in both frequency and time domains [29], [34].

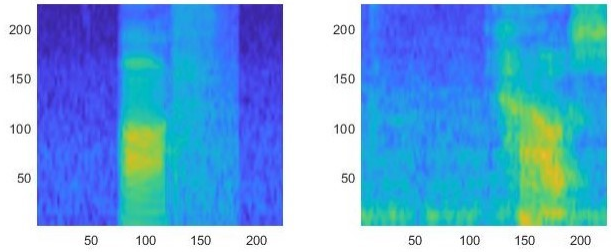

Fig 1: Spectrogram samples of foul language

IV. EXPERIMENTAL SETUP

This section details the experimental procedure that comes after data collection and annotation explained in Section 2. Fig. 2 presents the proposed system for foul language classification. The spectrogram of each speech in the training dataset is computed before fed as input to CNN model. The trained model was then used to classify the speech in the test dataset for offensive language prediction.

A. Spectrogram Computation

The database is 16 bits PCM and 1-channel samples at 16-kHz, turned into a series of feature vectors. 40 log Mel-frequency spectrogram coefficients are used in defining the visual representation of the speech energies as spectrum of frequencies. Foul and normal speech spectrograms have been computed with the following parameters: 1 second segment duration, 0.025 frame duration, 0,010 overlap window between frames, and 40 frequency bands. For Alexnet model, the spectrogram image size dimensions are 227-by-227, while the image dimensions are 224-by-224 for VGG16, Googlenet, and Resnet50.

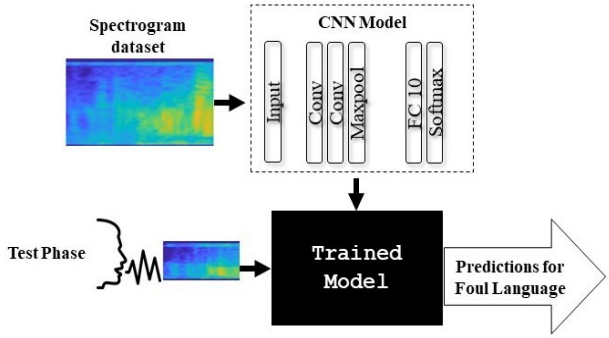

Fig 2: Illustration of foul language model

B. Training Algorithm Settings

The training of the target model (foul language classification model) from source model using transfer learning is carried out in two different ways:

- Fine-tune all layers for Alexnet, where the parameters of all the convolutional layers are fine-tuned along with the fully connected layer. This is due to the architecture of Alexnet that is kind of shallow CNN of 25 layers including 5 convolutional layers.
- Freeze the first 10 layers of VGG16, Googlenet, and Resnet50, where the weights and biases of the first 10 layers were transferred from the pre-trained networks on Imagenet. However, the other layers of the models are fine-tuned during models training stage. Freezing the weights and biases of the first 10 layers significantly speed up deep networks training considering the depth of VGG16, Googlenet, and Resnet50 networks with 41, 144, and 177 layers, respectively. In addition to also prevent the earlier layers from overfitting to the small new data set of our experiment.

All models were trained using the momentum method of Adaptive Moment Estimation (Adam) while cross-entropy was used as the loss function. The models were trained with batch size of 28 input matrices for 25 epochs. The models were trained and tested with five different folds for realistic averaging results of the whole data samples. The models were trained firstly on 2-class problem (foul vs normal), then were trained on 10-class problem to specifically define the type of foul language detected.

C. Metrics

Classification of offensive language model performance was evaluated based on two metrics: detection error rate (ER) and F1-score, since dataset is uneven between foul and normal classes. F1-score is defined as follows:

$$F1\text{-}score = 2\times(P\times R)/(P+R) \quad (1)$$

where precision P and recall R are calculated based on N_{tp}, N_{fp}, and N_{fn}, which are the total number of true positives, false positives, and false negatives in all segments:

$$P = N_{tp} /(N_{tp} + N_{fp}) \quad (2)$$
$$R = N_{tp} /(N_{tp} + N_{fn}) \quad (3)$$

ER is defined as the rate of wrongly detected spoken terms over the total input of speech terms. The four models were validated and tested with 5 different folds for precise evaluation metrics. The average results of folds $k=1$ through 5 is reported in the next section.

V. EXPERIMENTAL RESULTS

In this work, we investigate four different deep CNN models for the classification of foul language for automated censorship purpose. All results reported used the average of 5 cross validation to produce a realistic performance on the whole dataset used.

A. Performance of 2-class Models

Table II illustrates the performance of 2-class model with the proposed deep CNNs in terms of ER and F1-score with 621 foul language utterances and 1020 normal speech utterances. As noted in Table II, All the four proposed and retrained CNN models achieved a good performance in the classification of foul and normal language with only small ER differences (maximum ER of 1.77 for Alexnet to minimum ER of 0.87 for Resnet50). However, the four model's performance slightly drop for normal speech class (maximum ER of 2.84 for VGG16 and minimum of 1.61 for Resnet50). This attest to the wider range existence of normal conversational speech compared to foul language dataset that consist of only 9 different subclasses. F1-score is the measure that comprises both precision and recall of model performance for uneven dataset. The proposed models achieve the best performance on offensive language classification for 2-class problem with high F1-score of 97.03%, 97.07%, 98.02%, and 98.13% using Alexnet, VGG16, Googlenet, and Resnet50, respectively.

B. Performance of 10-class Models

The performance evaluation of proposed models on all 10 classes (9 profanity classes vs normal class) is detailed in Fig.3, which shows the comparison between models based on F1-score. The test on the models was carried out with imbalanced data and showed interesting pattern. The 9-class profanity model were tested with 69 utterances per each of the 9 foul classes and 1020 utterances for the normal conversasion. This experiment was done to specify the exact foul language utterances instead "foul".

The 9-class problem results show that the proposed models produced good F1-score that implies good sensitivity and specificity in detecting offensive language for these classes. On the other hand, Resnet50 model showed significant performance over the three systemd in detecting all the classes except for class "B" where Googlenet showed a slight difference of 97.98% F1-score compared to 97.0% F1-score of Resnet50. Interestingly, Alexnet and VGG16 shows a similar performance in detecting some of the foul languages (e.g. class "Ba" and "Bi") with F1-score of 89.31%, and 87.16% respectively.

Classes "C", "D" and "F" achieved the lowest metrics with high F1-score of 81.84% (Alexnet), 83.74% (Alexnet) and 80.89% (VGG16) respectively for proposed CNN. The three classes exhibit similar results due to the similarity of the acoustic features for the three classes and the similar phonemes pronunciation. This resulted in low accuracy and low sensitivity for the three classes in the problem of 1-class profanity decomposition to 9 classes. This is also due to high variations of samples (e.g. "F" + "-ing").

Table III. compares the overall performance of the models for 2-class and 10-class problem based on the average ER and average F1-score. Resnet50 model for 2-class shows the best performance with 1.24 ER and 98.54% F1-Score. Resnet50 also outperformed the other models for 10-class problem with 5.49 ER, and 94.20% F1-score. This suggest that Resnet50 provides better performance recognizing temporal data than other deep CNN models used in this work. Previous study has suggested the use of Alexnet and Googlenet for the classification of environmental sounds using spectral acoustic features [26]. Our experiment confirms the use of Alexnet and Googlenet based on the acoustic features for spoken foul language classification by proving the good performance achieved. Current work also added a suggestion of using Resnet50 for speech and sound classification based on the acoustic features as Resnet50 clearly outperformed Alexnet and Googlenet used in the previous research for the task of acoustic features classification.

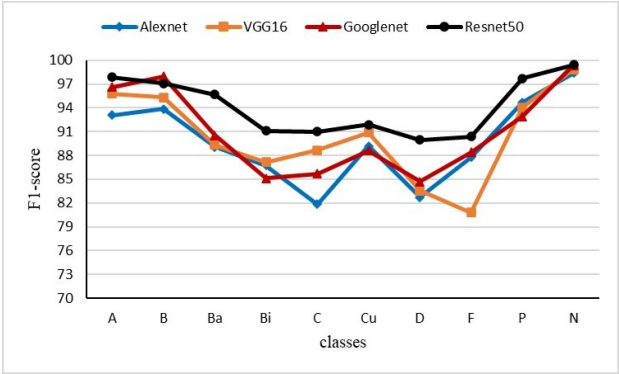

Fig 3: F1-score comparison for 10-class models (foul language notations highlighted in section 2)

TABLE II: Average performance metrics of proposed CNN models for 2-class foul classification based on 5 folds

Class	Alexnet		VGG16		Googlenet		Resnet50	
	ER	F1-score	ER	F1-score	ER	F1-score	ER	F1-score
Foul	1.77	97.03%	1.01	97.07%	1.01	98.02%	**0.87**	**98.13%**
Normal	2.57	98.19%	2.84	98.29%	1.63	98.89%	**1.61**	**98.96%**

TABLE III: Overall performance metrics of CNN models for 2-class and 10-class foul classification over 5 folds

Class	Alexnet		VGG16		Googlenet		Resnet50	
	ER	F1-score	ER	F1-score	ER	F1-score	ER	F1-score
2-class	2.17	97.61%	1.93	97.68%	1.32	98.46%	**1.24**	**98.54%**
10-class	8.30	89.74%	8.46	90.43%	8.47	90.99%	**5.49**	**94.20%**

Table III. also shows that the proposed 2-class system exhibits better performance compared to 10-class model for the same CNN architecture and the same folds of data. For example Alexnet achieved 2.17 ER in 2-class problem and 8.30 ER for 10-class problem. Additionally Alexnet F1-score drops from 97.61% for 2-class model to 89.74% for 10-class model. This can be illustated using confusion matrix in Table IV. Notably, some profanities were misclassified with another curse words of the 9 classes while in 2-class experiment all falls under one "foul" category. Table IV shows a confusion matrix of 10-class with Alexnet using one of the folds to highlight the misclassification between profanities.

This experiment was conducted on specific dataset of foul language and normal conversational speech and has proven to produce good results when tested on any derivative of the profanities. However, performance of the system may change if wider range of spoken words were tested. In such case, the models need to be adapted to the new conditions of variability in the new training dataset of either foul or normal classes. Several strategies can be done to tackle this gap, either by retraining the system fully or partially with more normal samples or leveraging the transfer learning technique for CNN models.

VI. CONCLUSION

This paper proposed the use of deep CNN models in detecting foul language from speech. Collected dataset was manually labelled with 2 and 10 annotations. Model was trained fully and partially to classify an incoming stream to detect targeted speech of foul language. The proposed models performed well in both 2-class and 10-class problem with ER ranging from 1.24 to 2.71 for 2-class, and 5.49 to 8.30 for 10-class. Resnet50 outperforms the other models for all experiments based on ER and F1-score.

This work has shown that convolutional neural networks, that are specifically used for image recognition, could be adapted to the application of classification and detection of unwanted speech's spectral images. Our work has proven the feasibility to use the same technology (i.e. CNNs) for both visual and speech recognition and censorship. Most of entertainment shows on TV or any video sharing websites and apps contain visual, acoustic, and speech content, and companies that required to censor their content would like to provide functionality for recognition of images/videos and speech/sounds. Hence, employing the same model technology for both applications of censorship would significantly reduce the development cost.

TABLE IV. Confusion Matrix of one-fold using Alexnet

Predicted Class

True Class	A	B	Ba	Bi	C	Cu	D	F	P	N
A	**64**	0	3	0	0	1	0	0	0	1
B	0	**69**	0	0	0	0	0	0	0	0
Ba	0	0	**69**	0	0	0	0	0	0	0
Bi	0	1	1	**55**	0	0	12	0	1	0
C	0	1	0	0	**62**	2	0	4	0	0
Cu	0	0	0	0	0	**68**	0	0	0	0
D	0	0	0	10	0	1	**51**	0	4	3
F	0	0	0	0	2	2	0	**62**	0	3
P	0	0	1	0	0	1	0	0	**67**	0
N	2	1	4	4	2	1	4	2	2	**998**

7. ACKNOWLEDGEMENT

This project funded by TM R&D, Malaysia. Project title "Automated Detection of Visual and Speech Contents for

Film Censorship Using Deep Learning", project number (MMUE/180029).

8. REFERENCES

[1] P. Y. Santosh K.Gaikwad, Gawali W.Bharti, "A Review on Speech Recognition Technique," *Int. J. Comput. Appl.*, vol. 10, no. 3, pp. 16–24, 2010.

[2] D. Amodei *et al.*, "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin," in *Proceedings of the 33rd International Conference on Machine Learning.*, 2016, vol. 48.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-To-End Attention-Based Large Vocabulary Sspeech Recogniton," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4945–4949.

[4] C. Chiu *et al.*, "State-Of-The-Art Speech Recognition WITH Sequence-To-Sequence Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4774–4778.

[5] P. Warden, G. Brain, and M. View, "Speech Commands : A Dataset for Limited-Vocabulary Speech Recognition," *arXiv:1804.03209v1*, pp. 1–11, 2018.

[6] J. Salamon and J. P. Bello, "Unsupervised Feature Learning For Urban Sound Classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 171–175.

[7] A. Pieropan, G. Salvi, K. Pauwels, and H. Kjellstrom, "Audio-visual classification and detection of human manipulation actions," in *IEEE International Conference on Intelligent Robots and Systems*, 2014, no. Iros, pp. 3045–3052.

[8] Y. Zhang, W. Chan, and N. Jaitly, "Very Deep Convolutional Networks For End-To-End Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4845–4849.

[9] K. Han, Y. He, D. Bagchi, E. Fosler-lussier, and D. Wang, "Deep Neural Network Based Spectral Feature Mapping for Robust Speech Recognition," in *Interspeech*, 2015, pp. 2484–2488.

[10] W. Chou, M. G. Rahim, E. Buhrke, M. R., and F. Tsang, "Hierarchial Subband Linear Predictive Cepstral FeatureS For HMM-Based Speech Recognition," US 6,292,776 B1, 2001.

[11] E. E. Elif Bozkurt, ,Cigdem Eroglu Erdem, and A. Tanju Erdem, "Use of Line Spectral Frequencies for Emotion Recognition from Speech," in *20th International Conference on Pattern Recognition*, 2010, pp. 3708–3711.

[12] D. F. Silva, V. M. A. De Souza, G. E. A. P. A. Batista, and R. Giusti, "Spoken Digit Recognition in Portuguese Using Line Spectral Frequencies," *Intel. Artif.*, vol. 15, pp. 241–250, 2012.

[13] Y. Wei and S. Wang, "Specific Two Words Lexical Semantic Recognition Based on the Wavelet Transform of Narrowband Spectrogram," in *International Conference on Electronics Instrumentation & Information Systems (EIIS)*, 2017, pp. 1–6.

[14] Bibi Zahra Mansor, Hamid Mirvaziri, and Faramarz Sadeghi, "Designing and Implementing of Intelligent Emotional Speech Recognition with Wavelet and Neural Network," vol. 7, no. January, pp. 26–30, 2016.

[15] R. K. Singh, R. Saha, P. K. Pal, and G. Singh, "Novel Feature Extraction Algorithm using DWT and Temporal Statistical Techniques for Word Dependent Speaker's Recognition," in *Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2018, pp. 130–134.

[16] H. Zhang, I. Mcloughlin, and Y. Song, "Robust Sound Event Recognition Using Convolutional Neural Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 559–563.

[17] M. Stolar, M. Lech, R. S. Bolia, and M. Skinner, "Acoustic Characteristics of Emotional Speech Using Spectrogram Image Classification," in *Signal Processing and Communication Systems, ICSPCS.*, 2018, pp. 1–5.

[18] R. Hyder, S. Ghaffarzadegan, Z. Feng, J. H. L. Hansen, and T. Hasan, "Acoustic Scene Classification Using a CNN-SuperVector System Trained with Auditory and Spectrogram Image Features," in *Interspeech*, 2017, no. November, pp. 3037–3077.

[19] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in *International Conference on Platform Technology and Service (PlatCon)*, 2017, pp. 1–5.

[20] R. Banaeeyan, H. A. Karim, S. Mansour, and J. See, "Acoustic Pornography Recognition Using Fused Pitch and Mel-Frequency Cepstrum Coefficients," in *International Conference on Advanced Science, Engineering and Technology MMU Engineering Conference, MECON. (MECON)*, 2019.

[21] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker Dependent Speech Emotion Recognition using MFCC and Support Vector Machine," in *International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, 2016, pp. 1080–1084.

[22] A. S. Ba Wazir, H. A. Karim, M. H. Lye Abdullah and S. Mansor, "Acoustic Pornography Recognition Using Recurrent Neural Network," in *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2019, pp. 144–148.

[23] A. S. Mahfoudh BA WAZIR and J. Huang CHUAH, "Spoken Arabic Digits Recognition Using Deep Learning," in *IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, 2019, pp. 339–344.

[24] Laffitte Pierre, Yun Wang, D. Sodoyer, and L. Girin, "Assessing the performances of different neural network architectures for the detection of screams and shouts in public transportation," *Expert Syst. Appl.*, vol. 117, pp. 29–41, 2019.

[25] P. Arora and R. Haeb-umbach, "A Study on Transfer Learning for Acoustic Event Detection in a Real Life Scenario," in *19th International Workshop on Multimedia Signal Processing (MMPS)*, 2017, pp. 1–6.

[26] A. Petef *et al.*, "Classifying environmental sounds using image recognition networks Classifying environmental sounds using image recognition networks," *Procedia Comput. Sci.*, vol. 112, pp. 2048–2056, 2017.

[27] "FILM CENSORSHIP ACT 2002," Malaysia, Act 620, 2006.

[28] F. Font, G. Roma, and X. Serra, "Freesound Technical Demo," in *Proceedings of the ACM International Conference on Multimedia. ACM*, 2013, pp. 411–412.

[29] W. Rawat, "Deep Convolutional Neural Networks for Image Classification : A Comprehensive Review," *Neural Comput.*, vol. 29, pp. 2352–2449, 2017.

[30] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR),* 2015.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[34] S. Karpagavalli and E. Chandra, "A Review on Automatic Speech Recognition Architecture and Approaches," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 9, no. 4, pp. 393–404, 2016.