Title

Author1*[1], Author[2], Author[3]
[1-2] Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, MALAYSIA.
[3] Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, MALAYSIA.

***Abstract*—Speech recognition has been developed through research works over 50 years of advancements. Deep Neural Network is one of the most dominant methods of speech analyzing due to its advantage of minimizing the error rate and optimization problems. This research proposes a profanity recognition model utilizing convolutional neural network. Speech Recognition model select the finest speech signal representation with mel-spectrogram images. Extracted features are fed into a network with several convolutional layers. The CNN model learning parameters are tuned for optimization purpose to achieve higher accuracy model training. The model tested with a novel dataset to recognize foul language from normal conversational speech for the application of film censorship.**

Keywords—Profanity Recognition, CNN, spectrogram, speech recognition.

I. INTRODUCTION

Audio has become an effective tool for shaping one's personality and character. The audio materials could be an independent audio files or an audio of television programs, movies and internet videos. The ease of accessibility to a huge audio files raised the necessity of filtering the audio content. With this in view, all local and foreign films should obtain suitability approval before distribution or public viewing.

The main purpose of this research is to facilitate the task of Profanity detection by proposing to exploit the distinctive power of spectrogram speech features. The model utilizes CNN model for speech recognition.

Speech recognition has been improving drastically from the implementations of basic word recognition to automatic speech recognition that uses continuous speech in the era of Deep Neural Networks, as Deep Learning utilizes artificial neurons to serve high dimensional data [1].

II. METHODOLOGY

This section details the experimental procedure, the choice of the datasets, model development as well as the model verification.

A. Dataset

Although there are instances of audio datasets (e.g. for environmental sound classification [2]), there is still a lack of profanity dataset. In this research, we collect dataset from online videos, and movies including positive and negative samples. Then, it is converted into audio files and processed for noise reduction prior to features extraction.

B. Audio Features Extraction

In this research, we are motivated to use mel-spectrogram based features to represent the noise-like excitations in the soundtracks of the videos. The spectrogram image provides a distinct feature of the data samples [3] to be used as an input to CNN model.

C. Development, Training, and Testing of RNN speech classifier

CNN model for foul words recognition consists of several layers to reduce the computational cost of unwanted spoken terms detection within films. The developed algorithm is then used to be trained for more than a trial to optimize the final model training parameters. The accuracy of the model and loss is checked through a function of TensorFlow named Tensorboard. 75% of the dataset is to be used for the model training purpose. The trained model is to be tested with the remaining 25% of the dataset to evaluate the overall performance of the model. The test phase is done using 5 folds cross validation for a realistic evaluation of the model over the whole dataset. Performance metrics for evaluation includes accuracy, precession, recall and F-measure metrics.

III. CONCLUSION

In this research, we proposed to use speech information extracted from video clips to train supervised classification model and test the feasibility of speech-driven features in the task of profanity recognition. More particularly, mel-spectrogram is employed to construct speech representations of the audio tracks.

ACKNOWLEDGMENT

This research is fully funded by TM R&D, Malaysia.

REFERENCES

[1] A. Halageri, A. Bidappa, C. Arjun, M. M. Sarathy, and S. Sultana, "Speech Recognition using Deep Learning," vol. 6, no. 3, pp. 3206–3209, 2015.

[2] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," *Proc. 23rd ACM Int. Conf. Multimed.*, pp. 1015–1018, 2015.

[3] R. W. Rabiner, Lawrence R and Schafer, "Theory and applications of digital speech processing," in *Pearson Upper Saddle River, NJ*, vol. 13, no. 1973, 2011, pp. 277–284.