Impact of Dithering in Compressed and Spectrally Distorted Speech for Emotion Recognition

Ritwika Mukherjee
Electronics and Telecommunications Engineering Department
Kalinga Institute of Industrial Technology
Bhubaneswar, India
ritwika.mukherjee1994@gmail.com

Subhabrata Das
Chemical and Biomolecular Engineering Department
National University of Singapore
Singapore
subhabrata@u.nus.edu

***Abstract*— A significant volume of the audio files present over the internet or stored in archives are in a compressed form. In addition, with a growing interest in voice-based applications, it is certain that automatic speech-based emotion recognition systems will soon be implemented in the production scale. MPEG-2 Audio Layer III (MP3) is one of the most popular codecs for audio file compression. However, MP3 compression introduces distortion into spectral and temporal characteristics of the audio waveform. Distorted speech signals result in poor feature extraction, which further downstream results in lowering the emotion recognition accuracy. In this study, we observed a significant loss of emotion recognition accuracy when tested models with compressed data that were trained over uncompressed data. The bit rate of 24 kbit/s and below suffered the maximum distortion in spectral feature and emotion recognition accuracy (ERA). We were able to observe an improvement in the emotion recognition accuracy if only Mel-frequency cepstral coefficient (MFCC) feature was used for building the automatic speech emotion recognition (ASER) model. The impact of dithering on compressed audio signals for emotion recognition experiments was studied. The addition of noise to the spectral valleys can influence the accuracy of the ERA. We observe end-to-end deep neural network models performed better compared to traditional machine learning models under challenging low bit rate scenario.**

Keywords—automatic speech emotion recognition, uniform dithering, MP3 compression, emotion classification

I. INTRODUCTION

Automatic speech emotion recognition (ASER) focuses on naturally recognizing the emotional state of a person [1]. ASER can be applied in different human-machine communication systems, augmenting human judgment, speaker recognition, and verification, biometric security systems, as well as medical and physiological purposes [2]. However, to deploy efficient ASER systems we would require large, clean and well-controlled data sets to execute rigorous experiments. In the recent past, the focus has shifted from "acted" to "in the world" emotional analyses [3, 4]. The testing of the ASER models needs to be performed in real-world scenarios with natural, mixed content and potentially distorted speech signals [3]. A large amount of available online data for developing machine learning models are stored in compressed format due to storage and band-with constraints. However, audio compressions tend to degrade the perceptual quality of the data that could adversely affect the automatic emotion recognition accuracy [5]. With an exponential rise in the affective computing research and its potential near-future industrial applications [6], we wish to study the impact of audio compression on ASER models and the impact of dithering technique on compressed data for automatic emotion recognition accuracy.

Previous studies in the literature cover the subject of ASER models trained with compressed speech audio data. One of the first studies in ASER reported by Garcia *et al.* [7] examined the effect of various speech-related codecs on emotion recognition accuracy of Gaussian Mixture Models (GMMs) using fear-type emotions. Albahri *et al.* [8] studied the effect of bit rates on AMR, AMR-WB and AMR-W+ codecs and reported that accuracy does not always decrease with a decreasing bitrate. In the study, they reported that the variation of emotion classification accuracy depends on different codecs and acoustic feature sets. Unfortunately, both studies focused on the pure recognition results without explaining the underlying spectral errors introduced by compression. Although, it is well established that changes in spectral information introduced during compression can significantly impact the performance of the ASER models [9]. In another study, Siegert *et al.* [10] tested emotional classification accuracy using support vector machines (SVMs) [11] for MP3, AM-WB and SPX coded speech. In these studies, the authors were able to achieve satisfactory unweighted average recall (UAR) results for MP3 with a bit rate of 32 kbit/s or higher. MPEG-2 AudioLayer III, also defined as MP3, is a part of perceptual audio codecs. MP3 is based on human hearing physiology and can retain the good intelligibility of human listeners even at a high compression rate. Recently, professional studios and broadcasters are leaving the MP3 coding tools and starting to deploy formats such as Speex or FLAC that are more suitable for speech [12]. However, the total

amount of data that is present in the MP3 format since the 1990s is huge enough to be considered as a true challenge for research in the ASER domain. Hence, we used MP3 audio codec in this study for compression of our data files to build our ASER models. Problems of spectral valleys caused by MP3 compression can be reduced by adding a small amount of noise to the compressed speech signal [12]. In the case of signals with unnatural valleys in the spectrum, additive noise or dithering can be a simple yet effective method to fill the gaps and transform the corrupt compressed speech signal to a more natural representation [12].

Almost all the previous reports on ASER with a coded speech focused on the telecommunication domain. However, we have a rich pool of "in the wild" data that can be an invaluable resource for training and testing our ASER models. We can also consider testing an already trained classifier, that would reduce implementation time and cost [3]. Secondly, access to a large amount of data also enables us to employ end-to-end deep neural networks (DNNs) that have outperformed other traditional machine learning algorithms in other domains for classification.

The goals of the current article are as follows:

1. ASER models trained on uncompressed audio speech signals and tested on MP3 compressed speech signals

2. Evaluate the impact of dithering on the performance of the ASER models.

The current article evaluates ASER models in the different compression scenarios. Additionally, the authors also compare traditional machine learning algorithms with end-to-end DNN algorithms for classification. To the best of the authors' knowledge, evaluating the effect of dithering for emotion recognition has not been thoroughly investigated before. Also, the study demonstrates a wide range of experimental conditions pertinent to ASER model development, such as bitrate, classification algorithm, training method, data set and feature set. The study aims to add to the current knowledge base of emotion recognition systems by reporting the impact of dithering techniques on emotion recognition accuracy.

II. Experimental design

A. Data sets

In this study, two standard emotion data sets (RAVDESS and TESS) and self-recorded custom data (custom-DB) were used which cover a wide range of acted and spontaneous emotions [3]. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 24 professional actors (12 females, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. The Toronto Emotional Speech Set (TESS) was modelled on Northwestern University Auditory Test No. 6. A set of 200 target words were spoken by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total. Both actresses are native English speakers and have thresholds within the normal range. The custom-DB dataset is an unbalanced noisy dataset created by self-recording samples and converting the raw file to 16 kHz and tagging the emotion to the filename. In order to compare ASER results across data sets a sampling rate of 16 kHz was chosen.

B. Audio Coding Format (Codec)

MPEG-2 Audio Layer III (MP3) is a lossy audio codec, introduced by Fraunhofer Institute in the year 1993 [4]. The audio compression is acquired by perceptual coding: certain regions of the original sound signal, considered beyond the auditory resolution ability, are discarded [6]. Next, the remaining information is stored in an effective manner using Huffman-coding. The bit rates range from 8-320 kbit/s. In the current study, the bit data was encoded at the following bitrates: 256, 192, 128, 96, 64, 24, 16, 8 kbit/s. MP3 is a popular codec and is freely available, allowing for easy reproduction of results presented.

C. Speech emotion recogntion algorithms

- Feature extraction

Previous studies have used the emobase feature set [3] from the openSMILE toolkit. Feature extraction is one of the most key aspects of ASER systems. It converts the speech waveform to parametric representation at a very low data rate. In this study, we have used the features available in the LibROSA library such as mel-frequency cepstral coefficient (MFCC), mel spectrogram frequency (mel), chromagram [13]. LibROSA is a python open-source library for music and audio file analysis. It helps to extract the necessary audio features for developing ASER systems. In audio file processing, the mel-frequency cepstrum (MFC) is defined as the short-term power spectrum of a sound, based in a linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. MFCC are coefficients that collective build an MFC. MFCC is the most widely used feature for emotion recognition from audio speech data sets.

- Classification algorithm

In this study, we used support-vector machine (SVM), multilayer perceptron (MLP) and Long short-term memory (LSTM) as our classification algorithms for building our ASER models. We normalized our data using a mean and standard deviation normalization which was computed on the training set and applied to test set. In machine learning hyper parameter optimization is an act of choosing a set of optimal parameters for a learning algorithm. A hyper parameter is a parameter whose value is used to control the learning process. The traditional way of

performing hyper parameter optimization has been grid search. In grid search we do an exhaustive searching through a manually specified subset of hyper parameter space of a learning algorithm. A grid search was performed on each of the specified models using cross validation on the training set as a performance matrix. The best parameters of each of the classification algorithms were used for the analysis.

MLP is class of feed-forward artificial neural network. The following parameters were set to build the MLP classifier: activation function - ReLU, learning rate (0.001), learning rate (adaptive), batch size (1024), number of hidden layers – 300, validation fraction (0.1) and maximum epoch (50).

We chose SVM with linear kernel and gamma value of 0.001.

LSTM is an artificial recurrent neural network architecture for various classification applications. The first part of the model is a feature extractor. The output is normalized and fed to two uni-directional LSTM layers [3], with a batch size of 64. Other parameters of the end-to-end LSTM based ASER model are as follows: dropout coefficient of 0.5 to prevent over-fitting, learning rate 0.001, epoch 1000.

- Dithering

Dither is the process of addition of low amount of noise to a digital signal to improve the process of audio compression which can lead to loss of information. This is because bit reduction during compression can lead to quantization error. In case of MP3 compression, the major effect on the audio signal is that of the spectral valleys or spectral holes introduced due to lowering of bit rates which can be observed in the frequency domain. With increases depth of spectral valleys, the difference between consecutive MFCC frame becomes larger. To compensate for such bands with no energy which could affect the feature extraction procedure a controlled low volume noise can be added to the signal samples and SER accuracy can be observed.

III. RESULTS AND DISCUSSION

A. *Uncompressed matched scenarios*

In the beginning, we trained our three ASER models (SVM, MLP and LSTM) using high quality uncompressed audio speech files. The input file was in standard Wave Audio File (WAV) format with bit rate of 256 kbit/s and sampling rate of 16 kHz. We combined the two standard data set (RAVDESS and TESS) and self-recorded custom-DB data set to increase the amount and quality of input to build a robust and efficient model. Table 1 illustrates the accuracy of different classification algorithms. As shown in Table I, all three classification algorithms achieved high training accuracy score. However, LSTM gave the highest test accuracy score over MLP and SVM respectively. We also, studied the effect of feature extraction on the ERA. We observed (Table 1), when the LSTM model was trained and tested with only MFCC feature, it gave a test accuracy score of 89.55% compared to only 82.66% when modelled with all the other features (MFCC, mel, chromagram).

TABLE I. TEST ACCURACY FOR VARIOUS MODELS

Feature List	Accuracy for various models		
	MLP	***SVM***	***LSTM***
MFCC,Mel Spectogram, Chromagram	0.8622	0.8688	0.8266
MFCC	0.8866	0.8622	0.89555

B. *Unmatched scenarios*

In this section, we investigate the effectiveness of an algorithm by training it with uncompressed rich audio dataset and testing its accuracy of prediction on compressed audio file. The results presented in Fig. 1 show that there is a significant drop in emotion recognition accuracy for all the classification algorithms across all bit rates compared to uncompressed test accuracy. All the three classification models performed almost similarly upto bit rate of 64 kbit/s. Below the 64 kbit/s bit rate, the traditional machine learning algorithm SVM gave very low emotional recognition accuracy (below 35%). MP3 compression narrows the spectral bandwidth as it decreases its bit rate in order to improve the subjective quality after compression. This causes a major problem where informations contained in high frequency regions are lost during compression. As a result, the partial error rate for these units increases more rapidly than for voiced units, which also steeply moves up the overall error rate [3]. This caused the lower test accuracy score for emotion recognition in the ASER models.

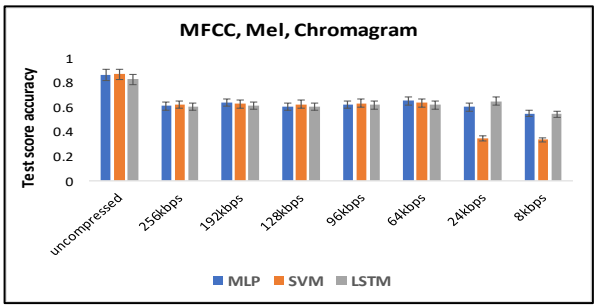

Fig. 1. Effect of compression on accuracy of emotion recognition

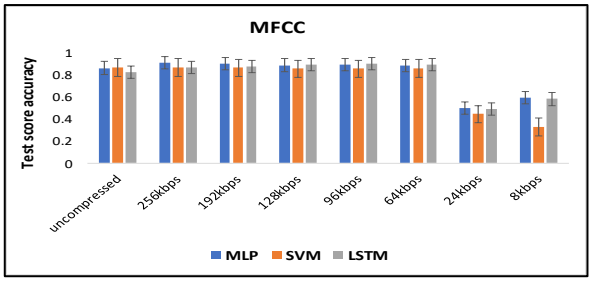

Fig. 2. Effect of MFCC feature extraction on ERA

To improve the drop in the test accuracy score for ERA, we tested the effectiveness of different extraction feature. We found that training the model with uncompressed data with only MFCC feature extraction, resulted in a very high-test score accuracy even after compression (Fig. 2). This is the first report which shows MFCC based feature extraction can improve emotion recognition accuracy with MP3 compressed test data. The relation between MFCCs and MP3 compression is most likely related to the use of the same frequency scale in both, derivation of the MFCC parameters and formulation of the mp3 compression. The ERA was almost similar for all the three classifiers (around 85-90% accuracy) upto bit rate of 64 kbit/s. From 24 kbit/s and below MFCC also was not able to give high ERA as it fell to lower levels.

IV. Conclusion

In this study, we established that emotion recognition is possible with models trained on high quality audio files and
then tested with MP3 compressed audio files. The effect of bit rates, and feature sets were studied on three different classification algorithms. Traditional machine learning algorithm (SVM) could not perform well at low bit rates compared to LSTM model. Training of models with only MFCC feature set gave improved emotion recognition accuracy for all the three models upto bit rate of 64 kbit/s. Impact of dithering was studied in this article. We observed that adding very low amount of noise did not change the ASER accuracy much, but, on the contrary increasing the range decreased the accuracy by 20%. In the current study we reported uniform dithering (UD) for ASER systems. In the future study we want to extend this approach and want to study the effect of spectrally selective dithering (SSD) to automatically detect corrupted bands and add a weighted amount of noise in the spectral domain in order to patch only the affect bands.

Acknowledgment

We would like to thank Dr. Sagarika Mukherjee for her helpful and constructive comments on the manuscript.

References

[1] Y. Huang, K. Tian, A. Wu, and G. Zhang, “Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition,” *Journal of Ambient Intelligence and Humanized Computing,* vol. 10, no. 5, pp. 1787-1798, 2019.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Signal processing magazine,* vol. 18, no. 1, pp. 32-80, 2001.

[3] O. Christopher, T. Andreas, S. Ingmar, and S. Bjorn, "Robust Speech Emotion Recognition under Different Encoding Conditions." pp. 3935-3939.

[4] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds." pp. 5866-5870.

[5] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, “Automatic recognition of facial actions in spontaneous expressions,” *Journal of multimedia,* vol. 1, no. 6, pp. 22-35, 2006.

[6] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals."

[7] N. García, J. Vásquez-Correa, J. Arias-Londoño, J. Várgas-Bonilla, and J. Orozco-Arroyave, "Automatic emotion recognition in compressed speech using acoustic and non-linear features." pp. 1-7.

[8] A. Albahri, M. Lech, and E. Cheng, “Effect of speech compression on the automatic recognition of emotions,” *International Journal of Signal Processing Systems,* vol. 4, no. 1, pp. 55-61, 2016.

[9] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Communication,* vol. 53, no. 9-10, pp. 1062-1087, 2011.

[10] I. Siegert, A. F. Lotz, L. L. Duong, and A. Wendemuth, "Measuring the impact of audio compression on the spectral quality of speech data," 2016, pp. 229-236.

[11] S. Hicsonmez, E. Uzun, and H. T. Sencar, "Methods for identifying traces of compression in audio." pp. 1-6.

[12] B. Michal, P. Petr, M. Petr, and N. Jan, “Dithering techniques in automatic recognition of speech corrupted by MP3 compression: Analysis, solutions and experiments,” 2017.

[13] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python."

C. Dithering

A small amount of uniform random noise was added to each signal sample of the compressed audio signal. The range of the random number was varied with a range R: [1,5]. The already trained model on compressed data was validated with these dithered datasets. It was observed that adding very low amount of noise did not change the ASER accuracy much but on the contrary increasing the range decreased the accuracy by 20% which is due to the sensitivity of MFCC of a signal to additive noise due to power dominance over the actual speech signal.