The Effects of Normalisation Methods on Speech Emotion Recognition

Tshephisho Joseph Sefara
Next Generation Enterprises and Institutions
Council for Scientific and Industrial Research
Pretoria, South Africa
tsefara@csir.co.za

***Abstract*—Speech emotion recognition systems require features to be extracted from the speech signal. These features include Time, Frequency, and Cepstral-domain features. To normalise features, it is a challenging task to select an appropriate normalisation algorithm since the algorithm may impact classification accuracy. This paper presents the effects of different normalisation methods applied to speech features for speech emotion recognition. Speech features are extracted from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset and normalised before training machine and deep learning algorithms such as Logistic Regression, Support Vector Machine, Multilayer Perceptron, Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM). The CNNs and LSTMs obtained 72% for both accuracy and F_1score outperforming standard machine learning algorithms. Feature normalisation improved both accuracy and F_1score by more than 14% using CNN and LSTM.**

Keywords— machine learning, neural networks, emotion recognition, normalisation method, speech emotions

I. INTRODUCTION

Speech is a fundamental means of communicating words and emotions. As a result, speech processing applications such as speech recognition, speaker recognition [1-3], and human-machine interfacing, can benefit from the inclusion of a reliable method of human emotion recognition through speech. While the speech community is growing with recent development of speech-based applications ranging from speech recognition applications, speaker recognition applications [1-3], text-to-speech applications [4-6], e-learning applications [7-10], to gender and language recognition applications.

Speech emotion recognition can be defined as the detection of emotional states (features) of a speaker from speech and extraction of selected features for the final classification. In some cases, the process is assisted by speech recognition system that contributes to the classification using linguistic information. These two techniques deal with a challenging task since emotional states are different from person to person. Speech emotion recognition systems can be helpful in intelligent assistance [11], criminal investigation [12], health care systems [13], and many more domains.

Speech emotion recognition requires features to be extracted from the speech signal and converted to an appropriate format for easier processing. These features include Time, Frequency, and Cepstral-domain features. It is a challenging task in speech emotion recognition to select efficient normalisation algorithm since the algorithm may impact the performance of the emotion recognition classifier by eliminating the effectiveness of emotional discrimination. This paper investigates the effects of different normalisation algorithms on classification performance and uses the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [14] dataset. The contributions of this work can be summarised as follows,

- We provide a discussion of emotions recognition systems.
- We show the effects of feature normalisation on emotion recognition and how normalisation impact classification performance.
- We compare the machine learning techniques and deep learning approaches.
- We propose novel multimodal models for predicting speech emotions together with recommended normalisation methods.

The rest of the paper is organised as follows: Section II discusses the background including approaches to emotion recognition and feature normalisation. Section III discusses the data, features, algorithms used to build the learning models, and evaluation procedure. Section IV discusses the experimental results, and the paper is concluded in Section V with future work.

II. BACKGROUND

This section discusses the literature review on approaches to emotion recognition and current methods of feature normalisation.

A. Approach to Emotion Recognition

Muszynski et al. [15] study the relationship between perceived and induced emotions of movie audiences. Authors find that perceived and induced emotions are sometimes inconsistent with each other. They further show that movie dialogues, perceived emotions, and aesthetic highlights are discriminative for movie induced emotion recognition besides spectators' behavioural reactions. Wang et al. [16] introduce Fourier parameters as features to improve speech emotion recognition rate over methods using Mel Frequency Cepstral Coefficients (MFCC) features to train a support vector machine

(SVM). Kurpukdee et al. [17] compare the recognition effect of speech recognition technique using convolutional long short-term memory recurrent neural network (LSTM-RNN) and hybrid model SVM coupled with convolutional LSTM-RNN, and as a result, SVM inputted with the output of convolutional LSTM-RNN performed better than conventional convolutional LSTM-RNN. Zamil et al. [18] use a Logistic Model Tree (LMT) classifier to build emotion recognition system using only MFCCs as features. Basu et al. [19] created a recurrent neural network (RNN) model that uses the output of a convolutional neural network (CNN) model to detect emotions in speech while Zheng et al. [20] combine CNN with random forest to create a hybrid model that uses CNN to extract features and RF to classify speech emotions. While [19] obtained test accuracy of 80% after 500 epochs, the testing loss or decay started to increase after 100 epochs from 1 to 2.5, this shows Basu et al. [19] overfitted the model. Neumann et al. [21] show that incorporating representations learnt by an unsupervised autoencoder that is trained on a large dataset, into a CNN-based emotion classifier, leads to consistent improvements in recognition accuracy. Huang et al. [22] propose speech emotion recognition considering verbal and non-verbal speech sounds since non-verbal sounds within an utterance play an important role for people to recognise emotion. Authors use sequence-to-sequence learning models such as LSTM and BLSTM coupled with SVM that is used to determine the verbal and the nonverbal intervals. Yoon et al. [23] exploited textual and acoustic data of an utterance for the speech emotion recognition using attention mechanism to combine acoustic and textual data, and two bi-directional long short-term memory (BLSTM) for obtaining hidden representations of the utterance. Authors improved accuracy by 6.5% compared to the baseline system in classifying the four emotion categories. While Li et al. [23] proposed the combining use of Dilated Residual Network and Multi-head Self-attention for feature learning in speech emotion recognition. Multi-head Self-attention models dependencies between different positions in the suprasegmental feature sequence, focusing on emotion-salient parts of speech in feature learning.

Ooi et al. [24] propose an architecture of intelligent audio emotion recognition that utilises both prosodic and spectral features while Lotfidereshgi et al. [25] propose a method for automatic recognition of speech emotions based on the Liquide State Machine (LSM) that operates directly on the speech signal and does not require feature extraction. Moreover, good results are obtained outperforming Tawari et al. [26] but not better than Jin et al. [27] who also show that MFCCs play an important role in speech emotion recognition. Manamela et al. [28] show recognition of emotions for low-resourced language.

B. Feature Normalisation

This section discusses the types of normalisation methods from scikit-learn[1] that can be applied to speech features.

a) Z-score or standard scaler removes the mean and scales the data to unit variance. However, the outliers may have an influence when computing the empirical mean and standard deviation which shrinks the range of the feature values. Therefore standard scaler may not produce balanced feature scales in the presence of outliers. [16] [17] and [26] use z-score normalisation on speech emotion recognition. The equation is given as follows: for a given speech feature *X*, the mean *E(X)* and standard deviation *std(X)* are computed and the standardised feature is estimated as:

$$\hat{X} = \frac{S - E(X)}{std(X)} \tag{1}$$

b) Minimum and Maximum scaler (MMS) transforms features by scaling each feature to a given range. The transformation is calculated as follows:

$$X_{scaled} = scale \times X + min - arg\,min(X) \times scale$$

$$\begin{aligned} where \\ scale &= \frac{max - min}{arg\,max(X) - arg\,min(X)} \\ min, max &= feature\ range \\ X &= features \end{aligned} \tag{2}$$

c) Maximum Absolute Scaler (MAS) differs from the *Minimum and Maximum scaler* such that the absolute values are mapped in the positive range. This scaler scales and translates each feature individually such that the maximal absolute value of each feature in the training set will be close to the given maximum range.

d) Robust Scaler (RS) scale features using statistics that are robust to outliers. This scaler removes the median and scales the data according to the quantile range. The centring and scaling statistics of this scaler are based on percentiles and are therefore not influenced by few numbers of very large marginal outliers. For each feature *X*, RS is formulated as:

$$RS = \frac{X - Q_1(x)}{Q_3(x) - Q_1(x)} \tag{3}$$

where Q_1 and Q_3 represent first and third interquartile respectively.

e) Power Transformer (PT) is a family of parametric, monotonic transformations that are applied to make data more Gaussian-like. Power Transformer implements the Yeo-Johnson [29] transforms. The power transform finds the optimal scaling factor to stabilise variance and minimise skewness through maximum likelihood estimation.

f) Quantile Uniform Transformer (QUT) transform features using quantiles information. QUT apply a non-linear transformation such that the probability density function of each feature will be mapped to a uniform distribution.

g) Quantile Gaussian Transformer (QGT) same as QUT such that the output distribution follows Gaussian distribution instead of a uniform distribution.

h) Normaliser rescales the vector for each sample to have a unit norm, independently of the distribution of the samples. For feature *a*, *b* and *c* Cartesian coordinates, the scaled value for $\hat{a}$ is formulated as:

[1] https://scikit-learn.org

$$\hat{a} = \frac{a_i}{\sqrt{a_i^2 + b_i^2 + c_i^2}} \tag{4}$$

III. Methods

This section discusses the data, feature extraction, methods, and evaluation procedure. The system architecture of a speech emotion recognition system is illustrated in Fig. 1 where we first use the RAVDESS database to extract speech feature vectors, and secondly we apply appropriate feature normalisation methods. Lastly, we use normalised features to train machine and deep learning models such as Logistic Regression (LR), SVM, Multilayer Perceptron (MLP), CNN, and LSTM-RNN.

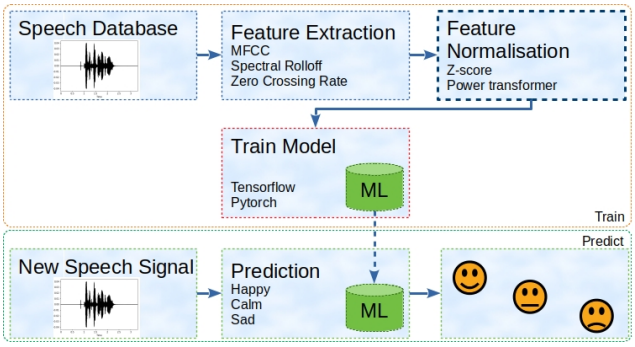


Fig. 1: The process of training emotional speech recogniser.

A. Data

The RAVDESS [14] is a validated multimodal database of emotional speech and song. The database consists of 24 professional actors (12 male and 12 female), vocalising same utterances in a neutral North American accent. Speech consists of 8 emotions: neutral, calm, happy, sad, angry, fearful, surprise, and disgust expressions. The audio files are mono waveform channel converted to 32000 sample rate per second. The songs are not used in this paper.

According to Fig. 2, the speech signals have different amplitude since each signal contains different emotions. It is clear that the signal with higher amplitude may contain anger. This may be different in other languages. The emotions, happy and sad, have similar amplitude, it is interesting to distinguish the two emotions using machine learning and to know which features are important.

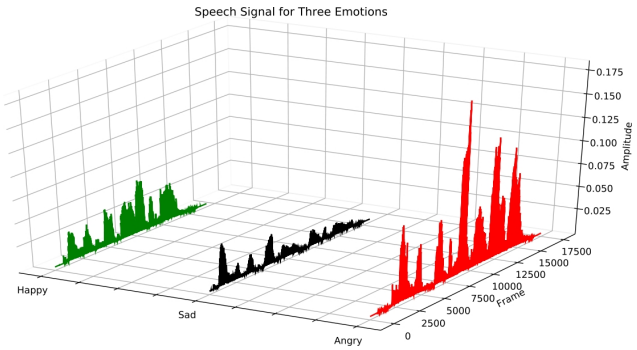


Fig. 2: Raw speech signal uttered in English with three emotion types, *happy*, *sad*, and anger.

Emotions may be expressed differently across cultures and people. We use unsupervised learning approach to investigate the data to see how many emotions can be extracted from the data. In Fig. 3, we cluster the data using k-means clustering. Typically, k-means clustering is an unsupervised algorithm that make inferences from data sets using only input vectors. The objective of k-means is to group similar data points together and discover underlying patterns. To achieve this objective, k-means looks for a fixed number (k) of clusters in a data set. We set $k=8$ to be the number of emotions. Principal Component Analysis (PCA) is used to reduce the dimensionality of the data by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Fig. 3 shows data clustered using PCA and k-means illustrate cluster centres.

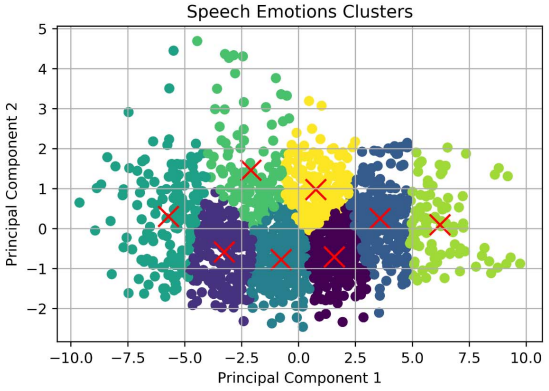


Fig. 3: PCA and k-means clustering of raw speech signal.

B. Feature Extraction

A human voice consists of different discriminative features that have the potential to uniquely identify human beings. Feature extraction is one of the most significant aspects of emotion recognition, this step generates feature vectors that represent each speech signal. We extract a total of 34 short-term features (frame size is set to 50 ms at a rate of 25 ms with the Hamming window) using pyAudioAnalysis [30] and these features are illustrated in Fig. 4 grouped into three domains, namely, Time, Frequency, and Cepstral-domain features [30]. The feature vector consists of minimum, maximum, standard deviation, average, and median. Thus, the size of the feature vector is 170.

1) Time-domain features include Zero Crossing Rate (ZCR), Energy and Entropy of Energy, are extracted directly from the raw audio samples. Speech signals are broadband signals and the interpretation of average ZCR is therefore much less precise. However, rough estimates of the spectral properties are obtained using a representation based on the short-time average ZCR. The ZCR measures the number of times in a given time frame that the amplitude of the speech signals passes through a value of zero. Several features of a speech signal can be selected using energy and zero crossing. A complete definition of computations is given in [31]. ZCR can be calculated as follows:

$$Z_n = \sum_{m=-\infty}^{\infty} |\ \mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\ |\ w(n-m)$$

$$\begin{aligned} &where \\ &\quad \mathrm{sgn}[x(n)] = 1 \quad x(n) \geq 0 \\ &\qquad\qquad\ = -1 \quad x(n) < 0 \\ &and \\ &\quad w(n) = 1/(2N) \quad 0 \leq n \leq N-1 \\ &\qquad\ = 0 \qquad otherwise \end{aligned} \tag{5}$$

2) Frequency-domain features are based on the magnitude of the Discrete Fourier Transform (DFT), these include Spectral Spread, Spectral Centroid, Spectral Flux, Spectral Entropy, Spectral Rolloff, Chroma Deviation and Chroma Vector [30].

3) Cepstral-domain features include MFCCs that result after the inverse DFT is applied on the logarithmic spectrum. MFCCs are popular audio features extracted from speech signals for use in recognition tasks and widely used for emotion and speech recognition [32]. MFCCs are determined with the help of a psychoacoustically motivated filter bank, followed by logarithmic compression and discrete Cosine transform. Suppose the output of an M-channel filterbank is *Y(m), m = 1, ..., M,* the MFCCs are obtained using the following equation:

$$c_n = \sum_{m=1}^{M} \left[\log Y(m)\right] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \tag{6}$$

where *n* is the index of a cepstral coefficient.

Feature ID	Feature Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9-21	MFCCs	MFCCs form a cepstral representation where the frequency bands are distributed according to the mel-scale.
22-33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

Fig. 4: Acoustic Features [30].

C. Model Architecture

This section discusses machine learning models implemented on the RAVDESS dataset.

a) Logistic Regression is a highly accurate and robust method that uses *multinomial logistic regression* method to generalise logistic regression to multi-class problems [33]. We implement the LR model defined by the following equation:

$$(Y_i = 1|X_i) = \frac{exp(\beta_0 + \beta_1 X_i)}{1 + exp(\beta_0 + \beta_1 X_i)} \tag{7}$$

where X_i represent the predictors, *Pr* is the probability where response variable Y_i=*1*, and β_i represent the parameters of the model.

b) Support Vector Machines are supervised machine learning models with associated learning algorithms that analyse data and recognise patterns, used for classification and regression analysis. We implement the radial basis function SVM kernel defined by the following equation [16]:

$$f(x) = exp(-\gamma \|x - x'\|^2) \tag{8}$$

where gamma γ = 1/(n_feature=170) is a positive parameter and penalty parameter *C*=10000.

c) Multilayer Perceptron has more than two hidden layers that may use custom or random initialisation and stochastic gradient descent to initialise and optimise the weights. MLPs take a lot of time to train and are computationally expensive since they can handle complex tasks. We train the MLP classifier using Sequence models on Tensorflow. The first layer is a fully connected layer followed by *Dropout* layer, followed another fully connected layer followed by *Dropout* layer and the output layer is a fully connected layer activated by *softmax*. The fully connected layers are activated by *rectified linear unit*, and *Dropout* layers use a probability of 0.5. The implementation stack is shown in Fig. 5.

Layer (type)	Output Shape	Param #
dense_45 (Dense)	(None, 128)	21888
dropout_36 (Dropout)	(None, 128)	0
activation_18 (Activation)	(None, 128)	0
dense_46 (Dense)	(None, 64)	8256
dropout_37 (Dropout)	(None, 64)	0
activation_19 (Activation)	(None, 64)	0
dense_47 (Dense)	(None, 8)	520

Fig. 5: Tensorflow MLP implementation.

d) Convolutional Neural Network are regularised versions of MLPs since they take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. We train the CNN classifier using Sequence models on Tensorflow. The first layer is the CNN layer activated by *rectified linear unit* followed by *global average pooling* layer, then the output layer is a fully connected layer activated by *softmax*. The implementation stack is shown in Fig. 6.

Layer (type)	Output Shape	Param #
conv1d_9 (Conv1D)	(None, 1, 128)	108928
dropout_39 (Dropout)	(None, 1, 128)	0
global_average_pooling1d_9 (	(None, 128)	0
dense_49 (Dense)	(None, 8)	1032

Fig. 6: Tensorflow CNN implementation.

e) LSTM are recurrent neural networks with multiple hidden layers. This structure allows LSTM models to capture temporal information. Moreover, the LSTM model outperformed the state-of-the-art algorithms to classify voicing or silence with noise in movies [34]. The implementation stack is shown in Fig. 7.

Layer (type)	Output Shape	Param #
lstm_9 (LSTM)	(None, 128)	153088
dropout_38 (Dropout)	(None, 128)	0
dense_48 (Dense)	(None, 8)	1032

Fig. 7: Tensorflow LSTM implementation.

D. Evaluation

The performance of the models is affected by the quality of the speech signal, the size of the training data, and most importantly the type of learning algorithm employed. The following evaluation measurements are used to evaluate the performance of the trained models:

a) Accuracy represents the total number of correctly predicted examples from all the examples given. The formulation is calculated as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{9}$$

b) Precision measures how many of the positively predicted examples are relevant. The formulation is calculated as follows:

$$Precision = \frac{tp}{tp + fp} \tag{10}$$

c) Recall measures how good a model is at predicting the positives. It is also called true positive rate. The formulation is calculated as follows:

$$Recall = \frac{tp}{tp + fn} \tag{11}$$

d) F_1 score is the harmonic mean of precision and recall. The formulation is calculated as follows:

$$F_1 score = 2 \times \frac{precision \times recall}{precision + recall} \tag{12}$$

e) Confusion matrix evaluates the quality of the classifier.

where:

- *tp* = True positive - a number of positive examples predicted positive.
- *tn* = True negative - a number of negative examples predicted negative.
- *fp* = False positive - a number of negative examples predicted positive.
- *fn* = False negative - a number of positive examples predicted negative.

The data is partitioned into 10% (72 samples) for validation and 10% (72 samples) for testing and 80% (1296 samples) for training using the full dataset of 1440 samples (24 speakers × 60 recordings).

IV. Results and Discussions

The models are grouped into two groups. The first group of machine learning algorithms consists of LR and SVM, while the second group of deep learning algorithms consists of MLP, CNN, and LSTM. The objective is to compare the performance of standard machine learning techniques and deep learning techniques. Table I shows the accuracy and Table II shows F_1 score results of the models after testing on 10% of the data. Both tables show almost the same results since F_1 score is averaged using the weights of each class. We use 1000 epochs to train MLP, CNN, and LSTM, on a batch size of 128 units. We observe LR did not perform well under all normalisation

methods compared to other models. SVM obtained a higher accuracy of 69% performed better than MLP with 63% while LSTM outperformed other models with 71% accuracy for PT normalisation method. Deep learning models, CNN and LSTM, obtained the highest accuracy and F_1 score of 72% using Z-score normalisation method. As a result, deep learning models outperformed standard machine learning models using LSTM and SVM results under all normalisation methods.

Table I shows the accuracy when normalisation occurred and when has not occurred. We observe that when normalisation occurred, the accuracy is always higher than when there is no normalisation excluding when *Normaliser* method is used. From these results, we observe that certain normalisation methods increase the performance of certain models. Hence, normalisation affects the classification of emotion recognition.

Confusion matrices of the best models, in our case CNN and LSTM, are shown in Fig. 8. Confusion matrix helps to evaluate the quality of a classifier. We observe both CNN and LSTM are good at classifying *calm* emotion while LSTM classified *neutral* emotion better but these two models confused *calm* and *neutral* emotions on one instance. We observe both models confusing *disgust* and *angry* emotions on three instances while CNN confused *surprise* and *happy* emotions on three instances but LSTM was better in this case. We observe both models correctly classified *fearful* emotion with 100%.

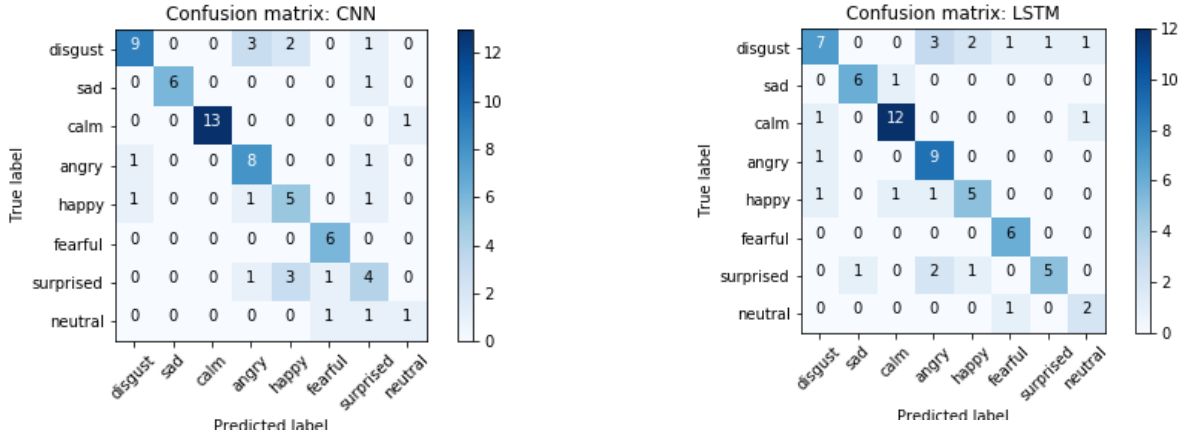
We pair models and show the top six normalisation methods in Table III based on accuracy. LSTM and CNN perform better when paired with the Z-score normalisation method. MLP and SVM perform better when paired with the RS normalisation method while LR performs better when paired with the MMS normalisation method. We observe Z-score and PT normalisation methods being in the top four but obtained lower accuracy for LR. From these results, Z-score, PT, RS, and MMS are good normalisation methods for speech emotions recognition.

TABLE I. ACCURACY AFTER TRAINING THE MODELS.

Model	Normalisation Methods									
	Original	***Z-score***	***MMS***	***MAS***	***RS***	***PT***	***QUT***	***QGT***	***Normaliser***	***Improvement***
LR	0.49	0.53	0.60	0.58	0.54	0.58	0.58	0.54	0.11	0.15
SVM	0.58	0.66	0.60	0.64	0.68	0.69	0.63	0.68	0.36	0.11
MLP	0.53	0.61	0.63	0.57	0.67	0.63	0.56	0.57	0.48	0.14
CNN	0.57	**0.72**	0.71	0.70	0.64	0.69	0.63	0.68	0.46	0.15
LSTM	0.56	0.72	0.67	0.68	0.71	0.71	0.64	0.68	0.46	0.16

TABLE II. F_1SCORE AFTER TRAINING THE MODELS.

Model	Normalisation Methods									
	Original	***Z-score***	***MMS***	***MAS***	***RS***	***PT***	***QUT***	***QGT***	***Normaliser***	***Improvement***
LR	0.45	0.52	0.60	0.58	0.54	0.58	0.58	0.53	0.03	0.15
SVM	0.59	0.66	0.60	0.64	0.69	0.69	0.63	0.68	0.34	0.10
MLP	0.49	0.61	0.62	0.56	0.67	0.61	0.55	0.57	0.45	0.18
CNN	0.56	0.72	0.71	0.70	0.64	0.70	0.63	0.68	0.41	0.16
LSTM	0.58	0.72	0.67	0.68	0.71	0.71	0.64	0.68	0.43	0.14



(a) CNN (b) LSTM

Fig.8: Confusion matrices for CNN and LSTM on a test data set of 72 samples.

TABLE III. PROPOSED COMBINATION

Model	Normalisation Method
LR	MMS
SVM	PT, RS, QGT, Z-score
MLP	RS, MMS, Z-score, PT
CNN	Z-score, MMS, MAS, PT
LSTM	Z-score, PT, RS, QGT, MAS

V. CONCLUSION AND FUTURE WORK

This paper presented the effects of different normalisation algorithms on emotion recognition. Machine learning and deep learning algorithms are trained and compared for each normalisation methods. We observed that feature normalisation method affects the performance of the models. We observed Z-score, PT, RS, and MMS normalisation methods being better compared to other methods on all the models. The Normaliser normalisation method performed poor on all the models. We observed deep learning techniques outperforming standard machine learning techniques. Hence, deep learning techniques are among best models at predicting speech emotions, with proper optimisation better results can be obtained. We proposed best paring options between models and normalisation methods based on the observed results.

In concluding this work, the future work will focus on investigating the most important speech features using machine learning algorithms.

REFERENCES

[1] T. B. Mokgonyane, T. J. Sefara, M. J. Manamela and T. I. Modipa, "The effects of data size on text-independent automatic speaker identification system," in *2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, KwaZulu-Natal, South Africa, 2019.

[2] T. B. Mokgonyane, T. J. Sefara, M. J. Manamela and T. I. Modipa, "Development of a text-independent speaker recognition system for biometric access control," in *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, Western Cape, South Africa, 2018, pp. 128-133.

[3] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela and P. J. Manamela, "Automatic speaker recognition system based on machine learning algorithms," in *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, Bloemfontein, 2019, pp. 141-146.

[4] T. J. Sefara and M. J. Manamela, "The development of local synthetic voices for an automatic pronunciation assistant," in *Southern Africa Telecommunication Networks and Application Conference 2016 (SATNAC2016)*, George, 2016, pp. 142-146.

[5] T. J. Sefara, M. J. Manamela and T. I. Modipa, "Web-based automatic pronunciation assistant," in *Southern Africa Telecommunication Networks and Application Conference 2017 (SATNAC2017)*, Barcelona, 2017, pp. 112-117.

[6] T. J. Sefara, T. B. Mokgonyane, M. J. Manamela and T. I. Modipa, "HMM-based speech synthesis system incorporated with language identification for low-resourced languages," in *International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, KwaZulu-Natal, South Africa, 2019.

[7] T. J. Sefara, P. T. Malatji and M. J. D. Manamela, "Speech synthesis applied to basic mathematics as a language," in *South Africa International Conference on Educational Technologies*, Pretoria, 2016, pp. 243-253.

[8] T. B. Mokgonyane, T. J. Sefara, P. J. Manamela, M. J. Manamela and T. I. Modipa, "Development of a speech-enabled basic arithmetic m-learning application for foundation phase learners," in *2017 IEEE AFRICON*, Cape Town, South Africa, 2017, pp. 794-799.

[9] P. T. Malatji, T. J. Sefara and M. J. D. Manamela, "Creating accented text-to-speech English voices to facilitate second language learning," in *South Africa International Conference on Educational Technologies*, Pretoria, 2016, pp. 234-242.

[10] P. T. Malatji, M. J. Manamela and T. J. Sefara, "Second language learning through accented synthetic voices," in *South Africa International Conference on Educational Technologies*, Pretoria, 2017, pp. 106-116.

[11] A. Tawari and M. Trivedi, "Speech based emotion classification framework for driver assistance system," in *2010 IEEE Intelligent Vehicles Symposium*, San Diego, CA, USA, 2010, pp. 174-178.

[12] H. Alipour, D. Zeng and D. C. Derrick, "AdaBoost-based sensor fusion for credibility assessment," in *2012 IEEE International Conference on Intelligence and Security Informatics*, Arlington, VA, USA, 2012, pp. 224-226.

[13] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi and M. Shimura, "Usage of emotion recognition in military health care," in *2011 Defense Science Research Conference and Expo (DSR)*, Singapore, 2011, pp. 1-5.

[14] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one,* vol. 13, 2018.

[15] M. Muszynski, L. Tian, C. Lai, J. Moore, T. Kostoulas, P. Lombardo, T. Pun and G. Chanel, "Recognizing induced emotions of movie audiences from multimodal information," *IEEE Transactions on Affective Computing,* pp. 1-17, 2019.

[16] K. Wang, N. An, B. N. Li, Y. Zhang and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Transactions on Affective Computing,* vol. 6, pp. 69-75, 2015.

[17] N. Kurpukdee, T. Koriyama, T. Kobayashi, S. Kasuriya, C. Wutiwiwatchai and P. Lamsrichan, "Speech emotion recognition using convolutional long short-term memory neural network and support vector machines," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1744-1749.

[18] A. A. A. Zamil, S. Hasan, S. M. J. Baki, J. M. Adam and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames," in *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 2019, pp. 281-285.

[19] S. Basu, J. Chakraborty and M. Aftabuddin, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*, 2017, pp. 333-336.

[20] L. Zheng, Q. Li, H. Ban and S. Liu, "Speech emotion recognition based on convolution neural network combined with random forest," in *2018 Chinese Control And Decision Conference (CCDC)*, 2018, pp. 4143-4147.

[21] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7390-7394.

[22] K. Huang, C. Wu, Q. Hong, M. Su and Y. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5866-5870.

[23] R. Li, Z. Wu, J. Jia, S. Zhao and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6675-6679.

[24] C. S. Ooi, K. P. Seng, L.-M. Ang and L. W. Chew, "A new approach of audio emotion recognition," *Expert Systems with Applications,* vol. 41, pp. 5858-5869, 2014.

[25] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5135-5139.

[26] A. Tawari and M. M. Trivedi, "Speech Emotion Analysis: Exploring the Role of Context," *IEEE Transactions on Multimedia,* vol. 12, pp. 502-509, Oct 2010.

[27] Y. Jin, P. Song, W. Zheng and L. Zhao, "A feature selection and feature fusion combination method for speaker-independent speech emotion recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4808-4812.

[28] P. J. Manamela, M. J. Manamela, T. I. Modipa, T. J. Sefara and T. B. Mokgonyane, "The automatic recognition of Sepedi speech emotions based on machine learning algorithms," in *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, Durban, 2018, pp. 1-7.

[29] I.-K. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika,* pp. 954-959, 2000.

[30] T. Giannakopoulos, "pyAudioAnalysis: An open-source Python library for audio signal analysis," *PloS one,* vol. 10, 2015.

[31] R. G. Bachu, S. Kopparthi, B. Adapa and B. D. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced Techniques in Computing Sciences and Software Engineering*, Springer, 2010, pp. 279-282.

[32] A. Rajasekhar and M. K. Hota, "A study of speech, speaker and emotion recognition using Mel frequency cepstrum coefficients and support vector machines," in *2018 International Conference on Communication and Signal Processing (ICCSP)*, 2018, pp. 0114-0118.

[33] F. E. Harrell, "Ordinal logistic regression," in *Regression modeling strategies*, Springer, 2015, pp. 311-325.

[34] F. Eyben, F. Weninger, S. Squartini and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 483-487.

[35] S. Yoon, S. Byun, S. Dey and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2822-2826.

[36] X. Ke, Y. Zhu, L. Wen and W. Zhang, "Speech emotion recognition based on SVM and ANN," *International Journal of Machine Learning and Computing,* vol. 8, 2018.