Integration of Frequency-Domain Features and Deep Learning Techniques to Enhance Recognition of Emotion in Speech

Ghanisht Aggarwal*, Meenu Garg and Dr. Neha Agrawal

Maharaja Agrasen Institution of Technology, Rohini, New Delhi, India, 110086

Abstract: Emotions can play an important role in determining how we think and behave. Emotions compel us to take decisions be it small or big. In order to understand emotions, it is paramount that we understand the critical expressive component. While interacting with people, it is cardinal to provide clues in the form of emotions to interpret and react accordingly. In this work, to tackle the ambiguity of speech, we have adopted an engineering technique based on speech emotion recognition. Formalizing our concern as a multi-class classification model, we compare the performances of different machine learning models by extracting numerable artisanal features of the audio signal and employing them to train six conventional machine learning models. For the various experiment settings in which we tested our models, we document accuracy, f-score, accuracy and recall. We are able to achieve at par performances form Gradient boosting and Random Forest classifiers. Ultimately, we have shown that simpler machine learning based models trained over a few hand-crafted features are able to achieve performances that may be analogous to the current deep learning based state-of-the-art methods.

Check for updates

DOI of the Article: https://doi.org/10.46501/IJMTST0706055

Available online at: http://www.ijmtst.com/vol7issue06.html

As per UGC guidelines an electronic bar code is provided to seure your paper

To Cite this Article: Ghanisht Aggarwal; Meenu Garg and Dr. Neha Agrawal. Integration of Frequency-Domain Features and Deep Learning Techniques to Enhance Recognition of Emotion in Speech. *International Journal for Modern Trends in Science and Technology* 2021, *7*, 0706220, pp. 334-340. https://doi.org/10.46501/IJMTST0706055

Article Info. Received: 15 May 2021; Accepted: 14 June 2021; Published: 22 June 2021

INTRODUCTION

Correspondence is indispensable for human presence and as a rule, we need to manage questionable circumstances. For instance, the expression "This is great" could be said under either upbeat or miserable settings. People can resolve uncertainty by and large since we can productively grasp data from various spheres such as discourse, audio and visual. With the ascent of profound learning calculations, there have been different endeavours to handle the assignment of Speech Emotion Recognition (SER) as in (Z. Huang, M. Dong, Q. Mao and Y. Zhan, 2014)(S. Yoon, S. Byun and K. Jung, 2018)(K. Han, D. Yu and I. Tashev, 2014). Nonetheless, this ascent has made specialists spend more on the intensity of the profound learning models rather than utilizing area information to develop important highlights and build better-performing and better-interpretable models. In our last paper, we investigated the ramifications of hand-made highlights for SER and analysed the presentation of lighter Artificial Intelligence (AI) models with the vigorously information-dependent profound learning models (G. Aggarwal and Y. Khanna). Moreover, we additionally join highlights from the printed methodology to comprehend the relationship between various modalities and help vagueness goals. All the more officially, we represent our errand as a multi-class arrangement issue and utilize the two classes of models to explain that. For both the methodologies, we first concentrate on hand-created highlights from the time area of the sound sign and train the separate models.

In this work, we use Frequency-Domain features such as Pitch, Noise, Speech Energy, Shift, Zero Crossing Rate (ZCR) and Mel-Frequency Cepstral Coefficient. We utilise the prowess of Deep Learning in an intricately assembled Neural Network, with an efficiency of 98% on the test data. The model is evaluated on the Toronto Emotional Speech Set (TESS) dataset under different audiosettings.

LITERATURE SURVEY

In this part, we study a portion of the efforts that have been made in the discipline of Speech Emotion Recognition (SER). The undertaking of SER isn't novel and has been read for a long period of time in literature. Utilizing Hidden Markov Models (HMMs)(L. R. Rabiner and B.H. Juang, 1986**)** are a dominant part of the early methodologies (A. Nogueiras, A. Moreno, A. Bonafonte and J. B. Mariño, 2001) (B. Schuller, G. Rigolland M. Lang, 2003)for distinguishing feeling from discourse. State-of-the-art performances have been observed after the advent of deep neural networks. The amalgamation of features such as Tensor Fusion Networks(Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh and L.P. Morency, 2017)and Low-Rank Matrix Multiplication (Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh and L.P. Morency, 2018)are among the recent proposals to help take SER forward.

This work aims to provide an analysis of lighter machine learning models trained over features that help in reducing obscurity in SER.

DATASET

In this work, we use the TESS (M. K. P. Kate Dupuis)released in 2010 by researchers at the University of Toronto. A set of 200 target words were spoken in the carrier phrase “Say the word _____” 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.Audiometric testing indicated that both actresses have thresholds within the normal range.

METHODOLOGY

Data Pre-processing

Audio: The above dataset for initial analysis offers an insight into the balancing of the dataset. In fact, there are two voices, a young female and an older female. All the information was preprocessed and sample-by-sample standardized, since there were not so many samples, and standardizing across samples would make it difficult to discern the data across classes with little variation. The normalization was based on loudness and length, for which we did zero padding to the heads and tails of all samples to make them length invariant.

Feature Extraction

The features used to train both models, the ML-based models, are now defined.

(a) Pitch: Wave-structures delivered by our vocal cords differ based on our emotion. Pitch is one of the features that help us recognize the same. Although there have been numerous calculations for assessing the pitch signal, we put our trust on utilizing the most widely recognized technique dependent on autocorrelation of centre-clipped frames (M. Sondhi, 1968).Formally, the input signal y[n] is center-clipped to give a resultant signal, $y_{clipped}[n]$:

$$y_{clipped}[n] = \begin{cases} y[n] - C_l, & \text{if } y[n] \geq C_l \\ 0, & \text{if } |y[n]| < C_l \\ y[n] + C_l, & \text{if } y[n] \leq C_l \end{cases} \quad (1)$$

$C_l$is typically almost half the average of the input signal. The distinct nature of input signal is denoted by [·]. The autocorrelation is now calculated for the $y_{clipped}$ signal received, which is further normalised, and the peak values associated with the y[n] input pitch given. Centre-clipping of the input signal was observed to result in more distinct peaks of autocorrelation.

(b) Noise: There are additional excitation signals other than pitch in the emotional state of anger or for stressed speech. This additional excitation is apparent in the spectrum as noise. Calculation of the noise is performed using a median-based filter as described in. First, the median filter is created for a given window size l, given by:

$$y[n] = median(x[n-k : n+k] | k = (l-1)/2) \quad (2)$$

Where, l is an odd value here. Considering the cases where l is supposed to be even, at the midpoint of the organised list, the midpoint is achieved as the mean of two values. Applying this filter to S_h, which is the h^{th} frequency slice of a given spectrogram S, to get a harmonic-enhanced spectrogram frequency slice H^h as:

$$H_i = M(S_h, l_{harm}) \quad (3)$$

where the midpoint filter is M , i is the i^{th} time step and the length of the harmonic filter is l_{harm}.

(c) Speech Energy: Since a speech signal's energy can be linked to how loud it is, we utilize the same to study certain feelings. To reflect speech energy, we use the conventional Root Mean Square Energy (RMSE) which is given by the equation:

$$E = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y[i]^2} \quad (4)$$

RMSE is calculated frame by frame and we take statistical features such as the mean and standard deviation.

(d) Shift: One spectral match condition and four spectral shift conditions were investigated: 2 mm, 3 mm, and 4 mm linear shift, and 3 mm shift with compression, in terms of cochlear distance.

(e) Zero Crossing Rate (ZCR): To calculate of the zero-crossing rate of a signal you need to compare the sign of each pair of consecutive samples. In other words, for a length N signal you need $O(N)$ operations. Such calculations are also extremely simple to implement, which makes the zero-crossing rate an attractive measure for low-complexity applications.

f) MFCC: Mel-Frequency Cepstral Coefficient is used to identify airline reservation, numbers spoken into a telephone and voice recognition system for security purpose. Some modifications have been proposed to the basic MFCC algorithm for better robustness, such as by lifting the log-mel-amplitudes to an appropriate power (around 2 or 3) before applying the DCT and reducing the impact of the low-energy parts.

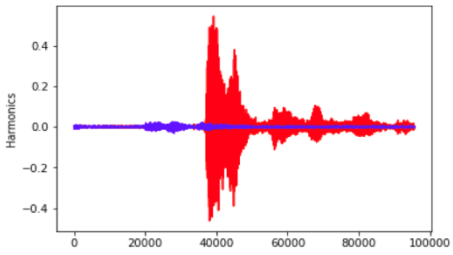

Fig. 1: Harmonics of angry (red) and sad (blue) audio signals

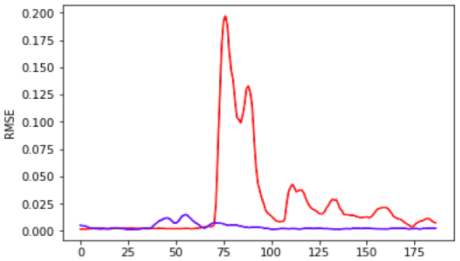

Fig. 2: RMSE plots of angry (red) and sad (blue) audio signals

Deep Learning Neural Network

A deep feedforward neural network (DNN) is an artificial neural network with multiple hidden layers of units between the input and output layers. Similar to shallow neural networks, DNNs can model complex non-linear relationships. DNN architectures generate compositional models, where extra layers enable composition of features from lower layers, giving a huge learning capacity and thus the potential of modelling complex patterns of speech data.

A Recurrent Neural Network (RNN) is similar to a MLP but differs in that it also has feedback connections. Experimental results show that like a MLP, these networks also perform as universal function approximations. Experimental results also show that RNNs are able to approximate simple functions between temporal trajectories in the input space and temporal trajectories in the output space. Although they constitute a generalization of the MLP, nobody has yet proven that they are universal function approximations. So far, the only proof comes from Funahashi, who proved that an RNN, provided that is has enough neurons, can generate any finite time trajectory. An example of a RNN architecture is shown in a figure. The main difference with the MLP consists of the existence of feedback connections between the neurons. This allows these types of neural networks to display all type of dynamical behaviours.

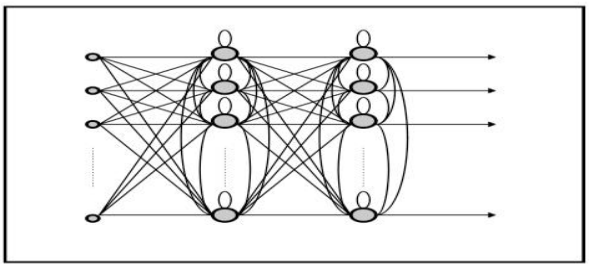

Fig. 3: RNN Architecture

Since the early eighties, researchers have been using neural networks in the speech recognition problem. One of the first attempts was Kohonen's electronic typewriter. It uses the clustering and classification characteristics of the SOM to obtain an ordered map from a sequence of feature vectors. The training was divided into two stages, where the first of these was used to obtain the SOM. Speech feature vectors were fed into the SOM until it converged. The second training stage consisted in labelling the SOM, i.e., each neuron of the feature map was assigned a phoneme label. Once the labelling process was completed, the training process ended. Then, unclassified speech was fed into the system, which was then translated it into a sequence of labels. This way, the feature extractor plus the SOM behaved like a transducer, transforming a sequence of speech samples into a sequence of labels. Then, the sequence of labels was processed by some AI scheme in order to obtain words from it.

Another approach is Waibel's Time Delay Neural Network (TDNN). It used a modified MLP to capture the space deviations and time warpings in a sequence of features. One input layer, two hidden layers and, one output layer were used to classify the different phonemes produced by English native speakers. The weights that defined the TDNN were defined such that the system was somewhat invariant to time warpings in the speech signal. It only recognized speech at a phoneme level and it was not used to make decisions in longer time spans, i.e., it was not directly used for word recognition.

As larger segments of speech are considered, approaches like the Electronic Typewriter or the TDNN become less useful. It is difficult for these approaches to deal with the time warpings, a problem which so far has impeded the neural networks to be successfully employed in the speech recognition problem. To integrate large time spans has become a critical problem and no technique using neural networks has yet been devised to solve this problem in a satisfactory manner.

In order to address the problem stated above, hybrid solutions have been used instead. Usually, after the phoneme recognition block, either HMM models or Time Delay Warping (TDW) measure procedures are used to manipulate the sequences of features produced by the feature extractors. In this thesis, alternate approaches, solely based on neural networks are developed.

Experiments

For the audio environment, we train all classifiers using only the previously mentioned audio function vectors.

Implementation Details

In this section, we describe the implementation details adopted in this work.

- Keras Sequential API is used, where we can add one layer at a time, starting from the input.
- The first is the sequential layer. It takes 512 units it is a positive integer; it specifies dimensionality of the output space and the activation function used in this layer is relu.
- relu is the rectifier (activation function max (0,x)). The rectifier activation function is used to add non linearity to the network.
- Batch normalization is the process to make neural networks faster and more stable through adding extra layers in a deep neural network.
- In last layer, instead of using sigmoid, we will use the Softmax activation function in the output layer. The Softmax activation function calculates the relative probabilities. That means it uses the value of Z21, Z22, Z23 to determine the final probability value.

Evaluation Metrics

Accuracy: This refers to the percentage of correctly graded research samples.

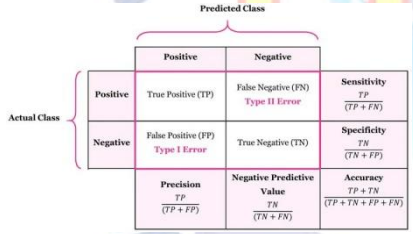

Fig. 4: Explanation of Confusion Matrix

RESULTS

In this section, we will be discussing our findings from the trained deep learning neural network.

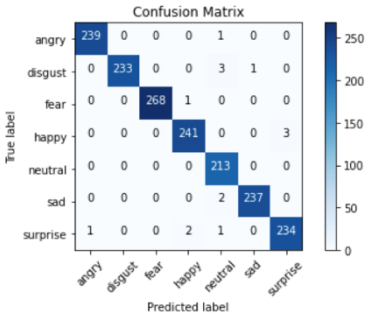

Fig. 4: Confusion Matrix

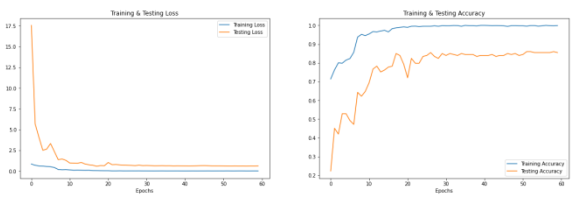

Fig. 4: Performance of Neural Network

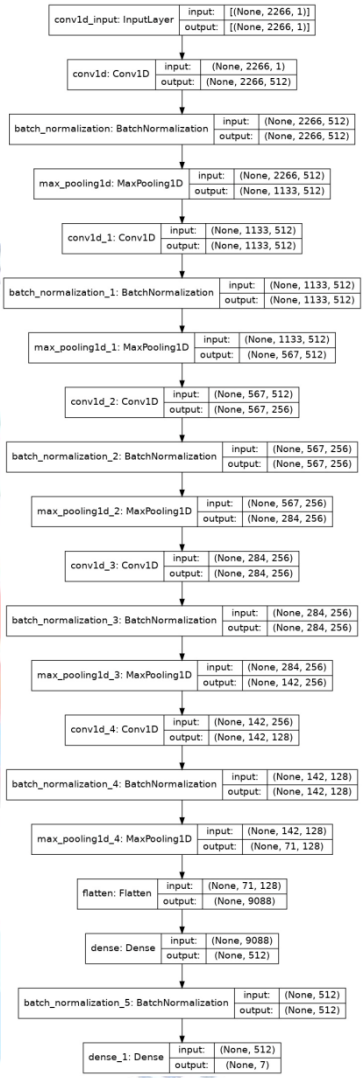

Fig. 5: Layers of the Neural Network

We can observe from the above figure that the state-of-the-art deep neural network has performed well. We reported the above metrics using optimal hyperparameter tuning.

CONCLUSION AND FUTURE WORK

In this research endeavour, we have discussed the task of speech emotion recognition and further worked on the contribution of various modalities for the resolution of uncertainty on the TESS dataset. We study a DL neural network. To resolve the formidable task of reducing ambiguity from the speech dataset, we used

an engineering technique based on speech emotion recognition and curated our problem into a multi-class classification problem by extracting hand-crafted frequency-domain audio features. We have created a comparison between both the ML- and DL-based techniques by showing that Random Forest yielded an accuracy of 57%, and that of a Neural Network is 98%. Including more data may as well help scale the performance of DL models.

REFERENCES

1. Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. LREc, 10, 2010
2. S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 112–118, IEEE, 2018.
3. K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural networks and extreme learning machines," in the Fifteenth annual conference of the international speech communication association, 2014.
4. S. Hoch Reiter and J. Schmid Huber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
5. M. K. P. Kate Dupuis, "Toronto emotional speech set (TESS),"2010. [Online].Available:https://tspace.library.utoronto.ca/handle/1807/24487
6. A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden Markov models," in Seventh European Conference on Speech Communication and Technology, 2001.
7. B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP 03)., vol. 2, pp. II– 1, IEEE, 2003.
8. L. R. Rabiner and B.H. Juang, "An introduction to hidden Markov models," IEEEASSP magazine, vol. 3, no. 1, pp. 4– 16, 1986.
9. Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3687–3691, IEEE, 2013.
10. A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.P. Morency, "Tensor fusion network for multimodal sentiment analysis," arXiv preprint arXiv:1707.07250, 2017.
11. Z. Liu, Y. Shen, V. B. Lakshmi Narasimhan, P. P. Liang, A. Zadeh, and L.P. Morency, "Efficient low-rank multi-modal fusion with modality-specific factors," arXiv preprint arXiv:1806.00064, 2018.
12. M. Sondhi, "New methods of pitch extraction," IEEE Transactions on audio and electroacoustics, vol. 16, no. 2, pp. 262–266, 1968.
13. H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in Speech production and speech modelling, pp. 241–261, Springer, 1990.
14. G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature-based classification of speech under stress," IEEE Transactions on speech and audio processing, vol. 9, no. 3, pp. 201– 216, 2001.
15. D. Fitzgerald, "Harmonic/percussive separation using median filtering," 2010.
16. Y. Amit, D. Geman, and K. Wilder, "Joint induction of shape features and tree classifiers," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 11, pp. 1300–1305, 1997.
17. L. Breiman, "Bagging predictors," Machine learning, vol. 24, no. 2, pp. 123–140, 1996.
18. J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning, vol. 1. Springer series in statistics New York, 2001.
19. J. Platt et al., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," Advances in large margin classifiers, vol. 10, no. 3, pp. 61– 74, 1999.
20. C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.
21. A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in Australasian Joint Conference on Artificial Intelligence, pp. 488–499, Springer, 2004.
22. G. King and L. Zeng, "Logistic regression in rare events data," Political analysis, vol. 9, no. 2, pp. 137–163, 2001. [23] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar,
23. E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference, pp. 18–25, 2015.
24. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in python," No. Oct, pp. 2825–2830, 2011.
25. T. Chen, "Scalable, portable and distributed gradient boosting (gbdt, gbrt or gbm) library for Python, R, Java, Scala, C++ and more. Runs on single machine, Hadoop, Spark, Flink and Dataflow," 2014.
26. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in

continuously spoken sentences," IEEE transactions on acoustics, speech, and signal processing, vol. 28, no. 4, pp. 357– 366, 1980.

27. F. Gouyon, F. Pachet, O. Delerue, et al., "On the use of zero-crossing rate for an application of classification of percussive sounds," in Proceedings of the COST G-6 conference on Digital Audio Effects (DAFX-00), Verona, Italy, 2000.
28. G. Sahu and D. R. Cheriton, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," Tech. Rep. [Online]. Available: http: //tinyurl.com/y55dlc3m
29. G. Aggarwal, Y. Khanna, "Leveraging Machine Learning Techniques in Enhancing Recognition of Emotion in Speech", Elsevier SSRN [Online]. Available: http://ssrn.com/abstract=3842556