A Comprehensive Analysis of Multimodal Speech Emotion Recognition

Kesava Rao Bagadi
Assistant Professor, ECE, MGIT, Hyderabad, India
bkesavaraoece@mgit.ac.in

Abstract. Emotion recognition is critical in dealing with everyday interpersonal human interactions. Understanding a person's emotions through his speech can do wonders for shaping social interactions. Because of the rapid development of social media, single-modal emotion recognition is finding it difficult to meet the demands of the current emotional recognition system. A multimodal emotion recognition model from speech and text was proposed in this paper to optimize the performance of the emotion recognition system. This paper, explore the comprehensive analysis of speech emotion recognition using text and audio. The results show that enhancement of accuracy compared to either audio or text. Here, results were obtained using the deep learning model i.e. LSTM. The experiment analysis is done for RAVDESS and SAVEE datasets. This implementation is done by python programming.

1. Introduction
In this AI era, one of the challenges is emotional analysis. It can be used for social media analysis, such as reviewing user conversations to better understand the audience. For Example, in Call center applications, understanding the emotions of customers can help communicate more effectively. Emotion intelligence is used as important for human-computer interaction systems to provide appropriate responses based on an emotional state of a user [1]. A lot of machine learning models have been developed for SER, with the models being trained to predict an emotion among candidates, such as happy, sad, angry, or neutral, for any speech.

Emotion identification through the text is a cumbersome task that goes beyond conventional sentiment analysis: instead of simply detecting neutral, positive, or negative feelings from text, the goal is to identify a set of emotions characterized by a higher granularity. For instance, feelings like anger or happiness could be included in the classification. As recognizing such emotions can turn out to be complex even for the human eye, machine learning algorithms are likely to obtain mixed performances. It is important to note that nowadays, emotion recognition from facial expression tends to perform better than from textual expression. Indeed, many subtleties should be taken into account to perform an accurate detection of human emotions through text, context-dependency being one of the most crucial. This is the reason why using advanced natural language processing is required to obtain the best performance possible.

The role of recognizing emotions in expressions, in particular, is the most prominent issue in the paralinguistic field. Emotion detection through speech aims to predict the speech’s emotional context and assign it into one of many categories. Various in-depth learning approaches are in use to improve emotion classifiers accuracy; but, this process is a challenging one. For a variety of reasons, it is still regarded as difficult. First and foremost, less data for complex neural network training is available and because of the costs associated with human labor, models are available. Many scholars have also
conducted regress experiments in this field [5–8]. Given interaction among speech & text, the combination can be used to influence the efficiency of the emotional recognition system on social media. The contribution of multimodal emotion analysis to the final emotional state can be evaluated.

One more difficulty is conventional speech representation models, i.e HMM and GMM [9, 10], can be models with fewer background details. Human emotions, on the other hand, typically switch slowly and rely heavily on background knowledge. Deep neural networks (DNN) can easily expose the latent inside structure of data and can extract more abstract features at a high level that are merely useful for detecting the emotion and recognition of emotion as compared to conventional feature extraction approaches. Deep learning is thought to be capable of getting large-scale training content’s essential features [11, 29]. These days, LSTM [11, 12] and CNN [13] are highly used to learn raw data high-level features in single-mode emotion detection, and their outputs outperform experimental methods. It is accepted to conclude that in-depth neural networks can already perform well in the detection of multimodal emotion, and related research studies are currently being conducted. To overcome these constraints, the researchers suggested a two-level model solution that combined text transcription and speech signals. The emotional terms in a statement reflect the emotional content of the speech. In the above sentence [11], for example, "lovely" along with "awesome," holds powerful emotions (non-emotional) terms like "guy" and "day."

Text input is not the only option; audio and video are both options. Data features (for example, text representation), audio features (for example, MFCC, spectrogram, etc.), and visual features must all be extracted (e.g. object detection and classification). Here this paper has, a comprehensive analysis of multimodal emotion intelligent analysis using both acoustic and textural cues also SAVEE datasets were presented. Most of the research work was done in one database IEMOCAP dataset. But, here analyze the performance analysis of SER for both datasets such as RAVDESS and SAVEE.

The paper is structured as Sec-2 explains the state-of-art in the literature in the emotion recognition task. Sec-3 gives a clear picture of the dataset that needs to be used in this work also the preprocessing steps that need to be applied before feature extraction. Section 4 describes the existing models and implementation details. The comparison analysis proposed models by the researchers is reported in Sec-5, followed by the conclusions and also a future scope of this comprehensive analysis in Section 6.

2. Literature Review

Here in this section, glance at recent work done in the speech emotion recognition (SER) field. The task of SER is old and followed in the literature in recent times. In speech emotion recognition problems, different machine learning-based approaches are used [12, 13, 14]. Several neural network-based architectures were recently suggested by researchers to enhance speech emotion recognition’s accuracy. Early research used neural networks to retrieve more information from speech and shown their efficacy in motion intelligence [16]. Highly tough neural-based models are suggested as advanced deep learning approaches.

Convolutional neural network-based models are equipped using audio features or sometimes spectrograms such as elementals parameters and MFCCs obtained from raw audio signals [15, 17, 18]. Higher-complexity models are generated by combining these neural network-based models [20, 21]. Recently, multimodal models for SER that combine audio and text have received a lot of attention [8, 9, 14, 15]. Since audio and text signals contain different information, it has been a major challenge to develop models that efficiently isolate and integrate information from each modality.

Several research looked at the usage of CNNs to identify whole speech spectrogram arrays or sometimes small spectrogram bands to distinguish speech emotions [16,17,18,19]. From short frames of speech, we knew SER was studied by [19]. The overall accuracy was 60.53 %t (6 emotions from the eNTERFACE db) and also 59.7 % (7 emotions from the SAVEE db) – (db here refers database). An equivalent but better methodology resulted in an overall accuracy of 64.78% (IEMOCAP results for five classes) [18]. Using voice spectrograms, various concatenated constructs integrating CNNs and RNNs were fully trained on EMO-DB data [19]. For seven emotions, the strongest structure yielded an average accuracy of 88.01 % and a recall of 86.86 %t. [16] used CNN to show affect-
salient characteristics, and that was used by the BRNN to distinguish 4 emotions from IEMOCAP. This method produces 64.08 % accuracy that is weighted and 56.41 % accuracy that is un-weighted. In the fact that these methods are convincing, also there is a space for progress. There are many reasons and one among those reasons for the comparatively low precision is that the speech libraries widely consumed in SER studies are very limited to show sufficient in-depth network structure preparation.

In the previous studies, attention mechanisms were frequently used to combine the information [8, 9], where the hidden states obtained separately from the audio and text signals were used as each other’s query or key-value pair. Another line of study has concentrated on the use of variant machine learning methods in conjunction with a few models based on neural networks. One of the authors used the meta entity method of learning and auxiliary tasks such as gender and naturalness to assist the model that is based on a neural network to learn huge features from a provided dataset [21]. Also, one more study looked at transfer learning approaches by using outside evidence from different domains [18]. In terms of material, researchers have looked into both language knowledge and acoustic features developed other network-based approaches for recognizing phrases of the emotional clue then evaluated the verbal signals emotional relevance from both continuous words and sequences [23, 24]. But, neither of these kinds of experiments used input from both speech patterns and text sequences at the same time to distinguish emotions during the whole learning neural network-based model.

3. DATASET and Its Pre-Processing

The RAVDESS was chosen because it is a new audio-visual emotional database in North American English [22]. Developed at the SMART Lab of Ryerson University, Canada, the DB consists of validated emotional speech and songs the lexically-matched recitations, a total of 24 experienced actors spoke with the North American accent. The audio section contains seven emotional states (calm, joyful, sad, frustrated, afraid, astonish, and disgust) and five tracks (fearful, calm, angry, happy, and sad). There are 2452 audio recordings in the corpus, divided into two emotional groups (normal, strong). Two hundred forty-seven people assessed the DB for the ranking, and a group of 72 people presented test-retest results. The database is free to use under a Creative Commons license. It should be remembered that the images from this dataset were not used due to a lack of thermal/infrared images and their capture in a regulated environment when we needed them.

| Emotion  | Speech Count | Song Count | Total Count |
|----------|--------------|------------|-------------|
| Neutral  | 98           | 92         | 188         |
| Calm     | 192          | 184        | 376         |
| Happy    | 192          | 184        | 376         |
| Sad      | 192          | 184        | 376         |
| Angry    | 192          | 184        | 376         |
| Fearful  | 192          | 184        | 376         |
| Disgust  | 192          | 0          | 192         |
| Surprised| 192          | 0          | 192         |
| Total    | 1440         | 1012       | 2452        |

Table1: RAVDESS Dataset details

SAVEE dataset was created as a prerequisite for developing an automated emotion recognition system. The archive contains videos of four male actors expressing seven distinct sentiments, a total of 480 British English utterances.

4. Methodology
Multimodal Emotion Recognition is a relatively new field that seeks to incorporate text, voice, and video inputs. This field has grown in popularity as social networks have provided researchers with access to large amounts of data. Recent studies have been exploring potential metrics to measure the coherence between emotions from the different channels. Speech emotion recognition aims to automatically recognize a person's emotional or physical condition from his voice. The emotional state of a person hidden in his speech is a key factor in human communication and relations as it provides feedbacks in communication while not altering linguistic contents. First Signal processing later feature extraction and signal processing are three-part of speech emotion recognition parts and these three parts use an acoustic filter to separate initial audio signals into meaningful units. The sensitive point in speech emotion recognition is feature extraction since features must effectively describe the human speech’s emotional content by not relying on the speaker or even the lexical content. To conclude, emotion grouping will assign a function matrix to each emotion label.

**Audio And Text for speech Emotion Recognition**

The following points describe audio signal preprocessing before audio features extraction.

4.1 Data Pre-processing

The dataset is not balanced, according to preliminary frequency analysis. The emotions "fear" and "surprise" were under-represented, and up-sampling techniques were used to address the problem. We then combined examples from the “happy” and “excited” classes because “happy” was under-represented and the two emotions are very similar. Furthermore, we discard examples labeled as “others”; they corresponded to examples labeled ambiguous even for a human.

4.2 Feature Extraction

We now describe the handcrafted features used to train both, the ML- and the DL-based models.

| Audio Features          | Pitch, Harmonic, Speech Energy, Silence, Central moments |
|-------------------------|---------------------------------------------------------|
| Text Features           | Term Frequency-Inverse Document Frequency (TFIDF)        |

Table2: Features used in Multimodal Speech Emotion Recognition

4.3 Deep Learning Models

This block contains a summary of the deep learning models used. Typically, Deep Neural Networks (DNNs) are designed in a completion fashion and they are expected to “figure out” features completely on their own. However, training such a model can take a lot of time as well as computational resources. To minimize the computational overhead, we directly feed the handcrafted features as input to these models and compare their performance with the traditional end-to-end trained counterparts.

LSTMs [21] were introduced for long-range context capturing in sequences. Unlike MLP, it has feedback connections that allow it to decide what information is important and what is not. It consists
of a gating mechanism and there are three types of gates: input, forget and output. Their equations are mentioned below:

\[
\begin{align*}
    f_t &= \sigma_g \left( W_f x_t + U_f h_{t-1} + b_f + C_{t-1} \right) + X_f \\
    i_t &= \sigma_g \left( W_i x_t + U_i h_{t-1} + b_i \right) \\
    o_t &= \sigma_g \left( W_o x_t + U_o h_{t-1} + b_o \right) \\
    c_t &= f_t c_{t-1} + i_t \sigma_c \left( W_c x_t + U_c h_{t-1} + b_c \right) + f_t c_t \\
    h_t &= o_t \sigma_h (c_t)
\end{align*}
\]

Where initial values are \(c_0=0\) and denote the element-wise product the time step, \(x_t\) refers to the LSTM unit’s input vector, \(f_t\) is the activation vector of forgetting gate, \(i_t\) refers to the activation vector of input gate, \(o_t\) refers to the activation vector output gate and finally, \(h_t\) is the hidden state vector, \(c_t\) is the cell state vector and \(W, U, b\) are weight, bias matrices which need to learn during training.

![Basic LSTM Cell](image)

Fig3: Visualization of an LSTM cell

We feed the feature vectors into the network as input and then run the output of the LSTM network through a soft-max layer to get probability scores for each of the six emotion classes. We don’t need another decoder network to transform it back from hidden to output space because we’re using feature vectors as input, which reduces network size.

5. Experimental Results

In this block, will go over all implementation details used in this work.

(a) To process the audio files and extract features from them, we used Librosa, a Python library.

(b) To implement the LSTM classifiers described earlier, use PyTorch.

(c) To regularize the hidden space of the LSTM classifiers, we employ a shut-off mechanism known as dropout [22], in which a subset of neurons is not used for final prediction. This is shown to improve network robustness and prevent over-fitting.

Here, randomly divided our dataset into a set of test (20%) and a train (80%). To achieve better performance, all experiments use the same split. Different batch sizes were used for different models. Hyper-parameters for all the models under the three experiment settings could be found in the released repository.

5.1 Metric used in this analysis

In this, begin by describing the various evaluation metrics that were used before reporting results for the three experiment settings.
(a) Accuracy: This refers to the percentage of test samples that are classified correctly.

(b) Precision: This measure tells us out of all predictions, how many are present in the ground truth. It is calculated using the formula:

\[ \text{Precision} = \frac{t_p}{t_p + f_p} \]  

(c) Recall: This measure counts the number of correct labels in the predicted output. It is calculated using the formula:

\[ \text{Recall} = \frac{tp}{tp + f_N} \]  

Here, \( t_p \), \( f_p \), and \( f_N \) stand for true positive, false positive, and false negative respectively. We can compute these values from the confusion matrix.

(d) F-score: It is also referred to as the harmonic mean of recall and precision. This calculation was included as accuracy is not a complete measure of a model’s predictive power but F-score is since it is more normalized.

5.2 Multimodal Evaluation

Table 2 represents the average recognition accuracies for RAVDESS and SAVEE datasets. Here, compared both audio and text features to recognition emotions with different deep learning models.

| Dataset   | Audio  | Text  | Audio+Text (multi-modal) |
|-----------|--------|-------|--------------------------|
| RAVDESS   | 67.9   | 75.1  | 85.34                    |
| SAVEE     | 65.74  | 77.56 | 83.7                     |

Fig 4: F1 score for RAVDESS and SAVEE dataset – Audio, Text, and Multimodal (Audio Text)

So, finally, the Confusion matrix for prediction corrects emotion using text and audio features.

![Normalized confusion matrix](image)

Fig 5: Confusion Matrix for Text and Speech Emotion

6. Future Scope and Conclusion
Finally, this paper presented a joint multimodal i.e. using audio and text-based speech emotion recognition. For the dataset i.e. RAVDESS and SAVEE, perform the feature engineering and send it to the LSTM cell to analyze the emotions. The results presented here, comparing the F1 score for both data sets given text, audio, and both (multimodal). By applying some advanced fusion-based deep learning framework to enhance performance and accuracy.

References
[1] A. Kolakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wrobel, “Emotion recognition and its applications,” in Human-Computer Systems Interaction: Backgrounds and Applications. Springer, 2014, pp. 51–62.
[2] G. Castellano, L. Kessous, and G. Caridakis, “Emotirecognition through multiple modalities: face, body gesture, speech,” in Affect and emotion in human-computer interaction. Springer, 2008, pp. 92–103.
[3] S. Yoon, S. Dey, H. Lee, and K. Jung, “Attentive modality hopping mechanism for speech emotion recognition,” in ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2020, pp. 3362–3366.
[4] Carlos Busso, Murtaza Bulut, and Shrikanth Narayanan, “Toward effective automatic recognition systems of emotion in speech,” Social Emotions in Nature and Artifact, p. 110, 2014.
[5] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu, “Emotional chatting machine: Emotional conversation generation with internal and external memory,” 2018.
[6] Dong Yu and Li Deng, AUTOMATIC SPEECH RECOGNITION., Springer, 2016.
[7] Google, “Cloud speech-to-text,” http://cloud.google.com/speech-to-text/, 2018.
[8] P. Meyer and T. Fingscheidt, “A New Evaluation Methodology for Speech Emotion Recognition With Confidence Output”, Proc. of ITG Conference on Speech Communication, pp. 81-84, Sept. 2014.
[9] K. Lu and Y. D. Jia, “Audio-visual emotion recognition with boosted coupled HMM,” in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 1148–1151, Tsukuba, Japan, November 2012.
[10] S. S. Narayanan, S. Lee, and A. Metallinou, “Audio-visual emotion recognition using Gaussian mixture models for face and voice,” in Proceedings of the 2008 Tenth IEEE International Symposium on Multimedia, pp. 250–257, Berkeley, CA, USA, December 2008.
[11] D. Li and J. Qian, “Text sentiment analysis based on long short-term memory,” in Proceedings of the 2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI), pp. 471–475, Wuhan, China, October 2016.
[12] L. L. Chao, J. H. Tao, M. H. Yang, Y. Li, and Z. Wen, “Long short-term memory recurrent neural network based on encoding method for emotion recognition in video,” in Proceedings of the 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2752–2756, Shanghai, China, March 2016.
[13] Bjorn Schuller, Gerhard Rigoll, and Manfred Lang, “Hidden Markov model-based speech emotion recognition,” in Multimedia and Expo, 2003. ICME’03. Proceedings. 2003 International Conference on. IEEE, 2003, vol. 1, pp. I–401.
[14] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” Speech Communication, vol. 53, no. 9-10, pp. 1162–1171, 2011.
[15] Dario Bertero and Pascale Fung, “The first look into a convolutional neural network for speech emotion detection,” in Acoustics, Speech, and Signal Processing (ICASSP). 2017 IEEE International Conference on. IEEE, 2017, pp. 5115–5119.
[16] Huang, Z., Dong, M., Mao, Q., and Zhan, Y. (2014). “Speech emotion recognition using CNN,” in ACM (Orlando, FL), 801–804.
[17] Fayek, H., Lech, M., and Cavedon, L. (2015). “Towards real-time speech emotion recognition using deep neural networks,” in ICSPCS (Cairns, QLD), 1–6.
[18] Lim, W., Jang, D., and Lee, T. (2016). “Speech emotion recognition using convolutional and recurrent neural networks,” in Proceedings of the Signal and Information Processing Association Annual Conference and Conference (Jeju), 1–4.

[19] Mao, Q., Dong, M., Huang, Z., and Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks, IEEE Trans. Multimedia 16, 2203–2213. DOI: 10.1109/TMM.2014.2360798

[20] John Gideon, Soheil Khorrham, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost, “Progressive neural networks for transfer learning in emotion recognition,” Proc. Interspeech 2017, pp. 1098–1102, 2017.

[21] 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

[22] Livingstone SR, Russo FA (2018) 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 13(5): e0196391.

[23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[25] P. V. Rouast, M. Adam, and R. Chiong, “Deep learning for human affect recognition: Insights and new developments,” IEEE Transactions on Affective Computing, 2019.

[26] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in Proc. of the 24th Int. Conf. on Machine learning, 2007, pp. 209–216.

[27] F. Eyben, M. Wollmer, and B. Schuller, “Opensmile: the Munich versatile and fast open-source audio feature extractor,” in Proc. of the 18th ACM Int. Conf. on Multimedia. ACM, 2010, pp. 1459–1462.

[28] O’Shaghnessy, D. (1987). Speech Communication: Human and Machine (Boston, MA: Addison-Wesley Longman Publishing Co., Inc.), 120–520.

[29] Schuller, B., Steidl, S. A, and Batliner, A. (2009a). “The interspeech 2009 emotion challenge,” in Proceedings INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association (Brighton, UK), 312–315.

[30] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” in Platform Technology and Service (PlatCon), 2017 International Conference on. IEEE, 2017, pp. 1–5.

[31] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” Language resources and evaluation, vol. 42, no. 4, pp. 335, 2008.