Emotion Recognition in Noisy Environments: Performance Analysis using Convolutional Neural Networks

Maria Grazia Borzì[1], Ludovica Beritelli[1], Valerio Francesco Puglisi[2], Roberta Avanzato[1] and Francesco Beritelli[1]

[1]Department of Electrical, Electronic and Computer Engineering University of Catania, Catania, Italy
[2]Department of Computer Science, University of Catania, Catania, Italy

Abstract

Recognition of emotional state through the sound of the voice is a crucial element in human interactions. This process, known as emotional prosody, allows an individual's emotions to be interpreted without the need to see his or her face or body language. The ability to identify emotions through voice is critical in areas such as psychology, health care, marketing and human-computer interaction technology. Moreover, in artificial intelligence systems and voice interfaces, the analysis of emotional prosody can improve the user experience and make interactions more personalized and intuitive. This study explores the robustness of a neural network in recognizing emotions within speech signals, considering the presence of environmental noise at different SNR levels. The study involves using public datasets containing different classes of emotions, training a CNN network to classify these emotional states, and then testing to evaluate the obtained performance. The results show that emotion recognition in speech is more accurate when the recording is clean and free of other noises. However, misclassification depends on the type of noise present and the type of emotion: indeed, it turns out that some emotions are more frequently misclassified than others. In fact, it is observed that accuracy is higher when using RAW data rather than MFCCs, with an accuracy of 75% compared to 60%. In addition, adding noise to the recordings results in a decrease in model performance.

Keywords

Speech Emotion Recognition (SER), Emotion classification, Audio signal analysis, Model robustness to environmental noise, Convolutional neural networks (CNN)

1. Introduction

Speech signal is the fastest and most natural method of communication among humans, prompting researchers to focus on making this mode of communication equally efficient in human-machine communication [1]. In addition to the challenges in speech recognition of words, it is essential to attribute context and nuance of meaning to words, including recognition of carried emotions.

Emotion Recognition (ER) is the process of identifying human emotions, and the use of technology for this process represents a relatively young area of research. So far, much of the research in this field has involved recognition of facial expressions through video and images, vocal expressions through audio, expressions written in text, and physiology measured by wearable devices. Emotion recognition can be used in various ways depending on the context, such as in speech or face recognition, referred to as "Speech Emotion Recognition" (SER) [2] and "Face Emotion Recognition" (FER), respectively.

Despite significant attention to this field, the task of SER systems remains challenging to implement for several reasons. First, it is not always clear which linguistic features are most relevant in distinguishing an emotion. In addition, there is considerable acoustic variability introduced by factors such as speed, accent, and voice volume, which can affect the waveform of audio files but still express the same emotion, making it complex for machines to understand the emotion and locate useful information.

In [3], a study is presented in which decision trees and Convolutional Neural Network (CNN) are used as emotion classifiers from English and Canadian audio data. In [4], a model using a CNN and combined input data from audio and text data is proposed. The study [5] proposes a semi-supervised learning (SSL) method for translingual emotion recognition when only a few labeled examples are available in the target domain.

The potential applications of this technology are numerous: in [6] the use of SER for detecting dangerous situations in public transportation to improve public safety is proposed. While [7] proposes a system for recognizing the emotional state of children while playing video games.

In this paper, the authors propose a study that aims to explore the robustness of a neural network in recognizing emotions within speech signals, considering the presence of environmental noise at different levels of signal-to-noise ratio (SNR). Using public datasets containing a wide range of emotion classes, such as the RAVDESS [8] and the USC-IEMOCAP [9], the goal is to evaluate the performance of the neural network under realistic conditions.

The comparison of using raw audio data (RAW) and Mel's frequency cepstral coefficients (MFCC) as inputs to the model aims to determine which type of input is more robust than noise and produces better results in emotion recognition. In addition, the study includes analysis of emotion misclassification trends to identify any patterns in the confounds between emotion classes. By adding noise of different types to voice recordings and evaluating the performance of the model under these conditions, the study seeks to understand how noise affects the emotion recognition process and what challenges may arise in realistic contexts. The ultimate goal is to contribute to the creation of more robust and reliable models for voice emotion analysis, with potential applications in areas such as automotive and telephone services.

The paper is organized as follows: In the 2 section, the methodology used in this paper for emotional recognition using speech signals is reviewed. In fact, this section discusses the dataset used for training and testing the network, the pre-processing phase of audio signals, and the addition

ICYRIME 2023: 8th International Conference of Yearly Reports on Informatics, Mathematics, and Engineering. Naples, July 28-31, 2023
borzi.m@studium.unict.it (M. G. Borzì); beritelli.ludovica@gmail.com (L. Beritelli); valerio.puglisi@phd.unict.it (V. F. Puglisi); roberta.avanzato@unict.it (R. Avanzato); francesco.beritelli@unict.it (F. Beritelli)

of noise to validate the robustness of the model. In the 3 section, the results obtained in the case of audio sequences without and with added ambient noise are discussed.

2. Proposed Work

The proposed method is to use neural networks to recognize emotions from audio signals. A public dataset is employed to train the models and evaluate them realistically. Next, audio data are preprocessed and two input approaches are explored: raw audio data and MFCC parameters. The robustness of the model to the addition of environmental noise is also evaluated. Finally, the neural network is trained on both inputs to classify the emotional state through the speech signal. The above is described in detail in the subsections to follow.

2.1. Datasets

The use of public datasets is a crucial element in the development of neural network-based emotion recognition models. The availability of well-annotated and diverse datasets allows researchers to train and evaluate their models under realistic conditions, ensuring greater generalization and reliability of the results obtained [10, 11, 12, 13, 14, 15, 16, 17, 18].

Prominent among the many datasets in the literature are RAVDESS (Ryerson Audio Database for Emotional Speech and Song) and USC-IEMOCAP (University of Southern California-Emotional Motion Capture). These datasets offer a wide range of voice recordings of individuals expressing a variety of emotions, allowing scholars to explore and analyze in detail the acoustic features associated with each emotion.

There are also other datasets available in different languages, such as German, Chinese, Mandarin, and others, which are useful for the development of SER systems in those languages.

The RAVDESS dataset, which is a large set of audiovisual files commonly used in Speech Emotion Recognition (SER) systems, was selected for this study. The dataset includes recordings of 24 actors from North America, including 12 female and 12 male actors. The recordings include both speech and singing and are categorized by loudness (normal or loud) and by emotion type, with 8 emotions labeled: calm, neutral, happiness, surprise, sadness, fear, anger, and disgust.

The choice of this dataset is also due to the fact that it avoids false positives regarding gender and volume classification. In addition, for this study, only simple audio data related to recordings were analyzed, excluding video and sung recordings.

2.2. Pre-processing Phase

Before providing the input data to the neural network, a pre-processing and feature extraction phase was performed on the audio data.

The first operation was to equalize the length of the recordings to 4 seconds.

Then, two parallel studies were conducted: the first study is related to using the RAW data of the audios as input to the neural network; the second study is related to extracting MFCC parameters from the audio signal.

In both cases, a moving window with an offset of 200 milliseconds was applied in the input generation process, which starting from the four-second audio file generated about ten files of two seconds each.

Finally, the dataset was divided into learning (70%) and testing (30%) sets.

2.3. Robustness to Environmental Noise

On the data belonging exclusively to the test set, further manipulations were carried out: first, noise was added to the recordings. The types of noise used in the study include:

- Babble-type noise [19]: typical noise of a group of people talking in the background (cocktail effect)
- Office-type noise [20]: typical background noise found in workplace settings.

Then, the powers were normalized in order to obtain a determined signal-to-noise ratio (SNR) with respect to the useful signal. The SNR values used in the study were 10 dB, 15 dB and 20 dB.

2.4. Neural network training

The neural network used in this study is of the 1D type and has also been used in [21, 22]. This network consists of 5 layers, of which the first four are convolutions (1D convolution layers, Batch normalization layers, ReLU layers and Pooling layers) and the last layer is the output (Softmax).

In addition, two different neural network trainings were conducted: one using the raw 2-second data of the audio signal as input, and the other using a vector of 50 MFCC parameters extracted from voice recordings of the same duration. Both trainings aimed to evaluate the model's ability in classifying emotional state through the speech signal.

3. Experimental Results

Analysis of the experimental data provides insight into the performance of the emotion recognition model under varying input conditions. Through the examination of both clean and noise-added data, relevant observations emerge regarding the overall performance of the system. This study focuses on three main scenarios: clean datasets, datasets with babble-type noise, and datasets with office-type noise. In each of these scenarios, a performance comparison is conducted between the use of MFCC parameters and RAW raw data.

3.1. Dataset clean

Analyzing the results obtained from the 2-second-length noise-free data, it can be seen that the accuracy is higher when RAW data (63%) is used compared to MFCC data (55%).

By applying a recurrence filter with observation window equal to 2 seconds and offset of 200 milliseconds to these data, the accuracy increases to 75% in the case of RAW data and 60% in the case of MFCC data. These results are displayed in Figure 1.

Next, looking at the confusion matrix in Figure 2, we see that using MFCC data there is a misclassification rate of 16% between Neutral and Happy, 19% between Fearful and Happy, and 16% between Fearful and Sad.

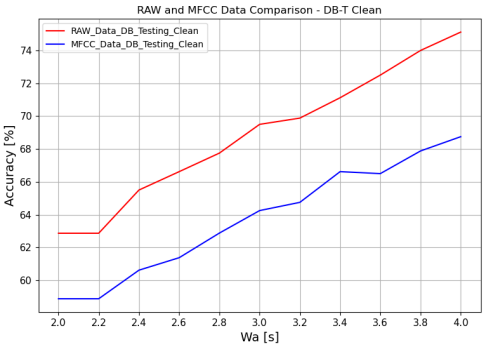

Figure 1: Model accuracy tables with clean dataset

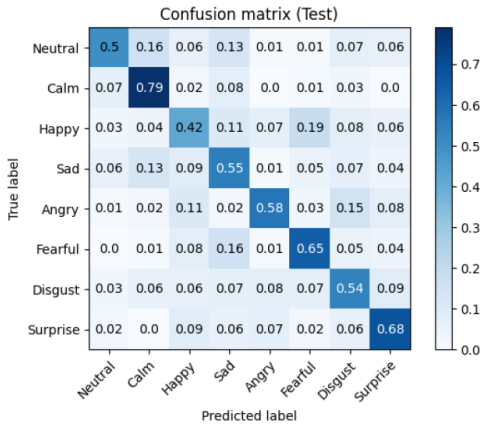

Figure 2: Model confusion matrix with clean MFCC dataset.

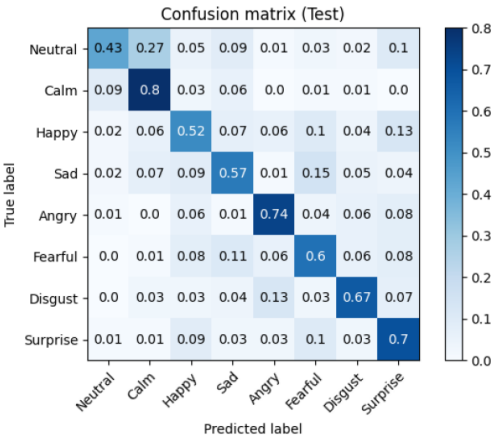

Figure 3: Model confusion matrix with clean RAW dataset.

Finally, by analyzing the confusion matrix related to the cleaned RAW files in Figure 3, it can be seen that generally the classes are recognized more accurately than in tests using MFCC data.

3.2. Dataset with babble-type noise

Testing data to which babble-type noise has been added, a decrease in model performance is observed. In addition, for

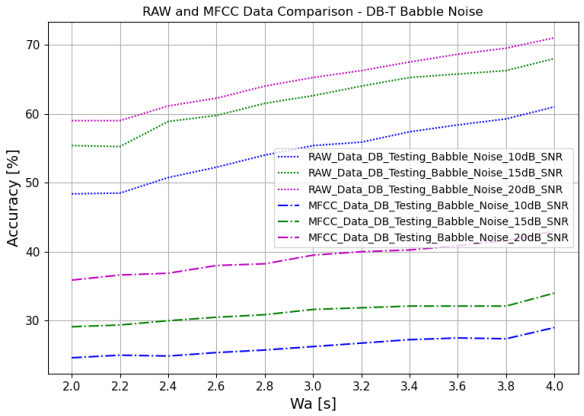

Figure 4: Model accuracy tables with noisy babble type dataset.

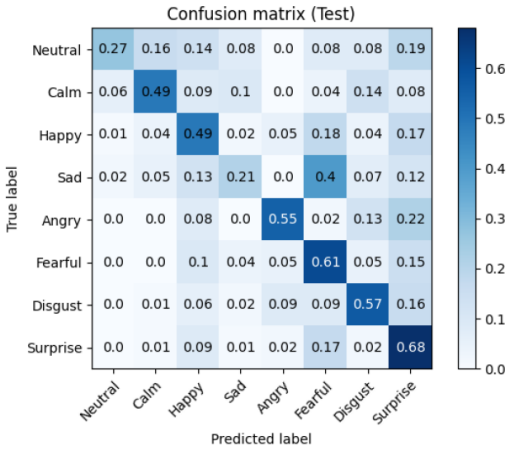

Figure 5: Model confusion matrix with noisy RAW dataset of babble type at 10 dB SNR.

the same SNR, the test detects better results using RAW files rather than MFCCs, as shown in Figure 4. Also, as expected, as SNR increases, accuracy decreases.

In tests with RAW recordings to which babble-type noise was added with a signal-to-noise ratio of 10 dB, it can be seen that the “Neutral" and “Sad" emotions are the most misclassified. In particular, it is observed that Neutral is often interchanged with Calm, Happy and Surprise, while Sad is strongly misclassified with Fearful, as shown in Figure 5.

By raising the SNR value to 20 dB, a general increase in network performance is evident in Figure 6, with the Sad emotion no longer being strongly misclassified. However, Neutral is still mistaken for Calm and Sad, while Surprise (which goes from an accuracy of 68% to 59%) is mistaken for Fearful.

3.3. Dataset with office-type noise

Similarly to the noisy babble-type recordings, the office-type recordings also showed lower accuracy results than the test with clean data, as could be expected. Also, as in the case of babble noise, for the same SNR, a decrease in accuracy is observed if MFCC instead of RAW data is tested. It is also noted that as the signal-to-noise ratio (SNR) increases, the accuracy decreases, as evidenced in Figure 7.

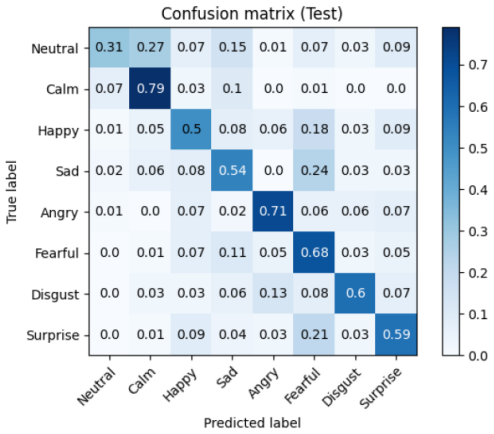

Figure 6: Model confusion matrix with noisy RAW dataset of babble type at 20 dB SNR.

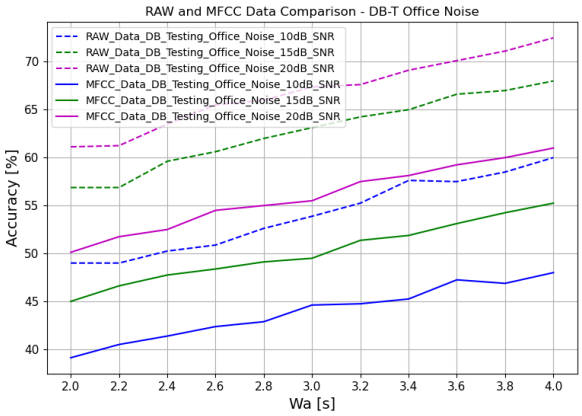

Figure 7: Model accuracy tables with noisy office-type dataset.

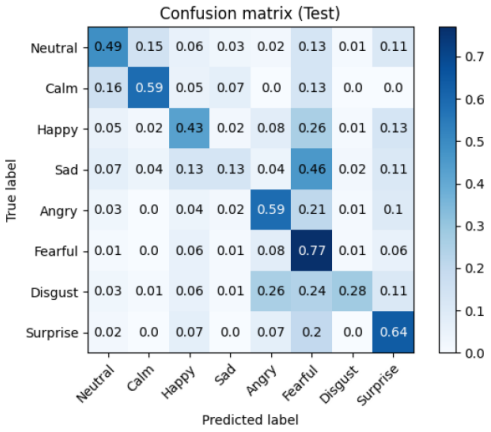

Figure 8: Model confusion matrix with noisy RAW office-type dataset at 10 dB SNR.

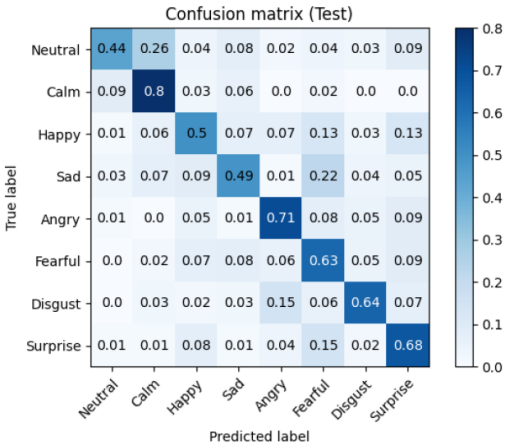

Figure 9: Model confusion matrix with noisy RAW dataset of office type 20 dB SNR.

By analyzing the confusion matrices related to RAW files with office-type noise and signal-to-noise ratio of 10 dB, it is possible to observe the results of the model in emotion class recognition. Under these test conditions, it is observed that the network strongly misclassifies the Sad emotion as Fearful 46% of the time, while Disgust is frequently misclassified as Angry and Fearful. Such a confusion matrix is shown in Figure 8.

By increasing the SNR value to 20 dB, a general increase in network performance is observed, with Sad and Disgust emotions no longer being strongly misclassified. Fearful goes from an accuracy of 77% to 69%, while Neutral goes from 49% to 44% accuracy. The confusion matrix of RAW files with office-type noise and SNR 20 dB is shown in Figure 9.

4. Conclusion

In this study, the robustness of a neural network in the face of noise variation in audio data in the context of Speech Emotion Recognition was examined.

From the tests conducted and related results, it appears that emotion recognition in speech is more accurate using RAW data (75%) than using MFCCs as input (60%).

In addition, worse performance was observed in the presence of noise in recordings. In particular, in crowded contexts where the acoustic cocktail party phenomenon occurs, SER-type systems experience difficulties in performing their task (accuracy decreases up to 30% in the worst case, with an SNR of 10 dB), while in contexts where noise differs significantly from human speech, better results are obtained (accuracy decreases up to 10% in the worst case, with an SNR of 10 dB).

In general, the most frequently misclassified classes turn out to be “neutral" and “sad."

These results outline interesting prospects for future developments, especially in real-world and potentially real-time applications. For example, in automotive or telephone services, where the recorded and analyzed voice usually belongs only to the involved interlocutors.

References

[1] F. Beritelli, A. Gallotta, C. Rametta, A dual streaming approach for speech quality enhancement of voip service over 3g networks, in: 2013 18th International Conference on Digital Signal Processing (DSP), IEEE, 2013, pp. 1–5.

[2] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern recognition 44 (2011) 572–587.
[3] N. Damodar, H. Vani, M. Anusuya, Voice emotion recognition using cnn and decision tree, in: International Journal of Innovative Technology and Exploring Engineering (IJITEE), volume 8, 2019.
[4] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, P. Yenigalla, Deep learning based emotion recognition system using speech features and transcriptions, arXiv preprint arXiv:1906.05681 (2019).
[5] M. Agarla, S. Bianco, L. Celona, P. Napoletano, A. Petrovsky, F. Piccoli, R. Schettini, I. Shanin, Semi-supervised cross-lingual speech emotion recognition, Expert Systems with Applications 237 (2024) 121368.
[6] E. Mancini, A. Galassi, F. Ruggeri, P. Torroni, Disruptive situation detection on public transport through speech emotion recognition, Intelligent Systems with Applications 21 (2024) 200305.
[7] P. Kozlov, A. Akram, P. Shamoi, Fuzzy approach for audio-video emotion recognition in computer games for children, Procedia Computer Science 231 (2024) 771–778.
[8] S. R. Livingstone, 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 13 (2018) e0196391.
[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (2008) 335–359.
[10] N. Brandizzi, S. Russo, G. Galati, C. Napoli, Addressing vehicle sharing through behavioral analysis: A solution to user clustering using recency-frequency-monetary and vehicle relocation based on neighborhood splits, Information (Switzerland) 13 (2022). doi:10.3390/info13110511.
[11] G. Capizzi, G. L. Sciuto, C. Napoli, M. Woźniak, G. Susi, A spiking neural network-based long-term prediction system for biogas production, Neural Networks 129 (2020) 271–279.
[12] V. Ponzi, S. Russo, V. Bianco, C. Napoli, A. Wajda, Psychoeducative social robots for an healthier lifestyle using artificial intelligence: a case-study, volume 3118, 2021, pp. 26 – 33.
[13] G. Capizzi, G. Lo Sciuto, M. Woźniak, R. Damaševicius, A clustering based system for automated oil spill detection by satellite remote sensing, in: Artificial Intelligence and Soft Computing: 15th International Conference, ICAISC 2016, Zakopane, Poland, June 12-16, 2016, Proceedings, Part II 15, Springer, 2016, pp. 613–623.
[14] G. De Magistris, R. Caprari, G. Castro, S. Russo, L. Iocchi, D. Nardi, C. Napoli, Vision-based holistic scene understanding for context-aware human-robot interaction 13196 LNAI (2022) 310 – 325. doi:10.1007/978-3-031-08421-8_21.
[15] G. Capizzi, G. Lo Sciuto, C. Napoli, R. Shikler, M. Woźniak, Optimizing the organic solar cell manufacturing process by means of afm measurements and neural networks, Energies 11 (2018) 1221.
[16] A. Alfarano, G. De Magistris, L. Mongelli, S. Russo, J. Starczewski, C. Napoli, A novel convmixer transformer based architecture for violent behavior detection 14126 LNAI (2023) 3 – 16. doi:10.1007/978-3-031-42508-0_1.
[17] G. De Magistris, M. Romano, J. Starczewski, C. Napoli, A novel dwt-based encoder for human pose estimation, volume 3360, 2022, pp. 33 – 40.
[18] F. Bonanno, G. Capizzi, A. Gagliano, C. Napoli, Optimal management of various renewable energy sources by a new forecasting method, 2012, pp. 934 – 940. doi:10.1109/SPEEDAM.2012.6264603.
[19] A. Varga, H. J. Steeneken, Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech communication 12 (1993) 247–251.
[20] J. Thiemann, N. Ito, E. Vincent, DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments, in: Proc. Meetings Acoust, 2013, pp. 1–6.
[21] W. Dai, C. Dai, S. Qu, J. Li, S. Das, Very deep convolutional neural networks for raw waveforms, in: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017, pp. 421–425.
[22] R. Avanzato, F. Beritelli, Heart sound multiclass analysis based on raw data and convolutional neural network, IEEE Sensors Letters 4 (2020) 1–4.