Models development for audio evaluation in affective computing

Mário F. C. Silva[1], Ilídio C. Oliveira[2], Susana Brás[2]

[1]Universidade de Aveiro, Aveiro, Portugal

[2]Instituto de Engenharia Eletrónica e Informática de Aveiro, Departamento de Eletrónica, Telecomunicações e Informática, Universidade de Aveiro, Aveiro, Portugal

Corresponding author: Mário F. C. Silva (e-mail: mariosilva@ua.pt).

ABSTRACT This paper presents the developed work in the dissertation in the field of Speech Emotion Recognition (SER) using traditional machine learning and deep learning approaches. The main objectives of the work were to develop and evaluate models on emotional datasets and explore feature engineering, validation methodologies, audio preprocessing techniques, and data stratification. The experimental results highlight the effectiveness of the proposed models and provide valuable insights into the SER area.

INDEX TERMS Affective Computing, Machine Learning, Speech Emotion Recognition, Voice Processing

I. INTRODUCTION

Our dissertation, " Models development for audio evaluation in affective computing", explores the field of Speech Emotion Recognition (SER), that uses technology for understanding human emotions through speech analysis. The main goal was to research and develop accurate SER models capable of identifying emotions from audio data. This paper provides a concise overview of the work, highlighting the results and major conclusions derived from the dissertation.

II. METHODS

Our methodology consisted of several steps to ensure a comprehensive study of SER. We began with an extensive literature review to explore the latest advancements in emotions, emotional datasets, and SER approaches. For our experiments, we selected the widely recognized and validated IEMOCAP dataset, comprising 5531 audio files with gender-balanced recordings and dimensional and discrete labels. To enhance data quality, we employed audio preprocessing techniques, effectively reducing noise and eliminating silence frames. Our investigation encompassed two distinct approaches for SER: a traditional feature-based approach that involved an elaborate audio feature engineering process, and a deep learning-based approach that leveraged transfer learning techniques. To ensure a comprehensive evaluation of our proposed models, we conducted cross-dataset evaluations. Additionally, through in-depth data analysis and stratification strategies, we identified specific conditions within the IEMOCAP dataset to enhance the models' performance. Finally, we developed a robust SER pipeline capable of efficiently processing and classifying emotions in both real-time and offline scenarios.

III. RESULTS

A. SER DEVELOPMENT

The traditional feature-based approach, which involved audio feature extraction, analysis, and selection, achieved notable performance with an accuracy of 60.69% on the IEMOCAP dataset. This approach showcased its computational efficiency by utilizing a 1-dimensional 33 audio feature vector as input for an Extreme Gradient Boosting (XGBoost) model. On the other hand, the deep learning approach, employing transfer learning techniques, achieved an accuracy of 58.24% on the same dataset, with a fine-tuned ResNet50 model using spectrogram 3D images. This model's success can be attributed to its enhanced feature extraction and generalization capabilities.

B. MODELS CROSS-DATASET VALIDATION

Our proposed models trained on the IEMOCAP dataset were then tested on multiple datasets, including eNTERFACE'05, EMO-DB, and CREMA-D. These datasets provide a diverse set of challenges for our models, including variations in speech content, recording conditions, emotional expressions, and spoken language, making them suitable for evaluating the robustness of our models across different scenarios. The results obtained are displayed in Table I below.

TABLE I
CROSS-DATASET VALIDATION OF THE MODELS TRAINED ON IEMOCAP

Dataset	Model	Accuracy	Predictions Time
eNTERFACE'05	Traditional	32.22%	0.17
	Deep Learning	36.67%	0.25
EMO-DB	Traditional	38.35%	0.10
	Deep Learning	38.35%	0.18
CREMA-D	Traditional	45.22%	0.10
	Deep Learning	54.14%	0.30

C. DATA STRATIFICATION

Furthermore, our comprehensive analysis of the data using various stratification strategies enabled us to uncover the shortcomings of the IEMOCAP dataset and the proposed models. To achieve this, we applied a set of conditions between the dimensional annotations of valence, arousal, and dominance (VAD) and the corresponding categorical emotion. In addition, we also applied a minimum duration condition of 1 second, as it was necessary to ensure the presence of sufficient emotional data for the models to learn effectively. As a result of this process, we obtained a total of 4200 audio files, with a nearly gender balanced distribution, including 52.9% male and 47.1% female speakers.

D. SER PIPELINE

Our developed SER pipeline is designed to effectively analyze audio streams, supporting both real-time and offline processing. It comprises multiple stages, each playing a vital role in accurately identifying emotional content from the audio signal. The pipeline begins by consuming audio data with a minimum duration of 1 second. To ensure compatibility with our models, the audio is normalized. Next, voice activity detection is performed, identifying segments that contain speech. Finally, SER classification is conducted on these speech segments. To validate the pipeline's effectiveness, the entire IEMOCAP dataset, comprising 10,039 annotated segments, was ingested. The pipeline accurately detected 7,698 speech segments, with similar predictions to the dataset annotations.

IV. DISCUSSION

In the field of speech emotion recognition (SER), inconsistent labeling processes in available datasets pose a major challenge. To address this, we utilized the widely used and validated IEMOCAP dataset, enabling more accurate model performance comparisons. Our developed audio preprocessing operations effectively reduced noise and eliminated silence frames at the beginning and end of the audio. Our proposed traditional model, XGBoost, achieved the second-highest accuracy among state-of-the-art articles employing the traditional approach on the same dataset. The ResNet50 model yielded comparable results, indicating a potential for improvement through techniques like fine-tuning, displayed in Table II.

TABLE II
STATE-OF-THE-ART SER MODELS PERFORMANCE ON IEMOCAP

Model	Input	Evaluation Strategy	Accuracy
Traditional Feature-Based SER Approaches			
Ensemble of Random Forest, XGBoost and Multilayer Perceptron [1]	8-dimensional features vector	80:20 split (train-test)	56.00%
Multi-level binary decision trees [2]	384 features vector	10-fold CV	58.46%
XGBoost [Ours]	33 features vector	5-fold CV	60.69%
CNN [3]	193 features vector	5-fold CV	64.30%
Deep Learning-Based SER Approaches			
ResNet50 [Ours]	3-D Spectrogram Image	5-fold CV	58.24%
CNN and RNN [4]	Log-Spectrogram	5-fold CV	64.22%
3-D attention-based convolutional RNN [5]	3-D Mel-Spectrogram Image	10-fold CV	64.70%
CNN and LSTM with attention [6]	Mel-Spectrogram	5-fold CV	67.00%
Quaternion CNN [7]	Mel spectrogram encoded in an RGB quaternion domain	5-fold CV	70.46%

Cross-dataset validation findings supported the superiority of the deep learning model for SER tasks on diverse datasets, owing to its enhanced feature extraction and generalization capabilities across different factors. In contrast, the traditional approach's reliance on a feature input vector derived solely from the training dataset limits its applicability to other datasets. However, the traditional model's faster prediction speed renders it more suitable for real-time applications. Furthermore, our conducted data stratification process identified specific conditions that significantly improved performance within the IEMOCAP dataset and cross-dataset evaluation. This analysis effectively addressed biases and accounted for the subjective nature of emotions in labeled data, enhancing model robustness and applicability. The developed emotion recognition pipeline demonstrated versatility and scalability, enabling the creation of voiced audio segments through a voice activity detection (VAD) tool. Validation of the pipeline using the IEMOCAP raw dataset aligned with expectations, considering its requirement of at least 1 second of audio data while accommodating shorter audio segments within the dataset. As anticipated, the distribution of predicted emotions reflected the training data, given its training on utterances from the same dataset.

V. CONCLUSION

In conclusion, our study significantly contributes to the field of SER through a thorough investigation and evaluation of both traditional and deep learning approaches. We propose an XGBoost and a ResNet50 with competitive performances and potential for real-world applications. Through the data stratification study of the dataset, we successfully identified high-quality data that is applicable in diverse environments. Additionally, the development of our pipeline enables the efficient detection and construction of speech segments for accurate emotion recognition. Our research contributes to the advancement of SER understanding and implementation. Future work could concentrate on refining the proposed models, exploring additional datasets, or using techniques to enhance SER accuracies, such as multimodal approaches that combine speech, facial expressions, or text, which holds promise for deepening the recognition of emotions.

REFERENCES

[1] G. Sahu, Multimodal speech emotion recognition and ambiguity resolution, 2019. doi: 10.48550/ARXIV.1904. 06022.

[2] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, «Emotion recognition using a hierarchical binary decision tree approach», Speech Communication, vol. 53, no. 9-10, pp. 1162–1171, Nov. 2011. doi: 10.1016/j. specom.2011.06.004.

[3] D. Issa, M. F. Demirci, and A. Yazici, «Speech emotion recognition with deep convolutional neural networks», Biomedical Signal Processing and Control, vol. 59, p. 101 894, May 2020. doi: 10.1016/j.bspc.2020.101894.

[4] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, «Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms», in Proc. Interspeech 2018, 2018, pp. 3683–3687. doi: 10. 21437/Interspeech.2018-2228.

[5] M. Chen, X. He, J. Yang, and H. Zhang, «3-d convolutional recurrent neural networks with attention model for speech emotion recognition», IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440–1444, 2018. doi: 10.1109/LSP.2018.2860246.

[6] Z. Zhao, Z. Bao, Y. Zhao, et al., «Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition», IEEE Access, vol. 7, pp. 97 515–97 525, 2019. doi: 10.1109/access.2019.2928625.

[7] A. Muppidi and M. Radfar, «Speech emotion recognition using quaternion convolutional neural networks», in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Jun. 2021. doi: 10.1109/icassp39728.2021.9414248.