FORMATTING INSTRUCTIONS FOR ICLR 2023 TINY PAPER SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotion recognition from the speech is a challenging task. In this paper, we propose a modality conversion approach to improve emotion recognition performance on the MELD dataset by converting speech to text and leveraging text-based features. We propose a model called Modality-Conversion which automatic speech recognition systems (ASR) to convert speech signals to text and apply natural language processing techniques to extract text-based features. Our results show that the proposed method improves emotion recognition accuracy compared to state-of-the-art speech-based approaches on the MELD dataset. This study demonstrates the potential of modality conversion and text-based techniques for enhancing emotion recognition from speech.

1 INTRODUCTION

Emotion recognition from the speech is challenging due to audio complexity and expression variability. State-of-the-art Baseline approaches rely on a one-modality approach Chen et al. (2022). However, if there is one modality that performs better than speech Poria et al. (2019), can changing the modality improve results when using the modality that had lower performance on the same dataset? This Tiny Paper proposes a modality conversion approach to convert speech to text and use text-based features, using ASR systems and NLP techniques to extract these features. The proposed method improves emotion recognition accuracy compared to state-of-the-art speech-based approaches on the MELD dataset Chen et al. (2022). The study demonstrates the potential of modality conversion and text-based techniques for enhancing emotion recognition from speech. The paper is organized into related work, proposed method, experiments and results, and conclusion sections.

2 RELATED WORKS

Emotion recognition from speech has been studied extensively, with many approaches focusing on acoustic features extracted from speech signals Chen et al. (2022) Poria et al. (2019). In the context of emotion recognition researchers have explored using text-based features Song et al. (2022) Lee & Lee (2022), such as sentiment, emotion keywords, and linguistic patterns. Multi-modality approaches, which combine both speech and text-based features, have also been proposed and have shown promising results in emotion recognition and show that on the same dataset, text-based approaches reach better results than speech-based approaches Chudasama et al. (2022) Poria et al. (2019) Siriwardhana et al. (2020) Yoon et al. (2018) Sahu (2019). However, to the best of our knowledge, no one has explored using modality conversion to improve emotion recognition from speech on the MELD dataset.

3 METHOD

In this study, we utilized the MELD dataset, consisting of audio-visual clips from TV shows. Firstly, we extracted the audio from the clips and then utilized the Vosk API for automatic speech recognition (ASR) to convert the audio clips to text LLC (2020). We used a roBERTa-base finetuned for sentiment analysis for emotion recognition on the text transcripts obtained from the ASR system Barbieri et al. (2020). The model was fine-tuned on the MELD dataset. We evaluated the performance of

Table 1: Sample table title

Method	Model Name	WF1
Chen et al. (2022)	SpeechFormer-S	0.418
Chen et al. (2022)	SpeechFormer-B	0.419
Proposed Method	Modaliry-Conversion	0.44

our proposed modality conversion approach that is shown in Figuer1 using the weighted-F1(WF1) metric.

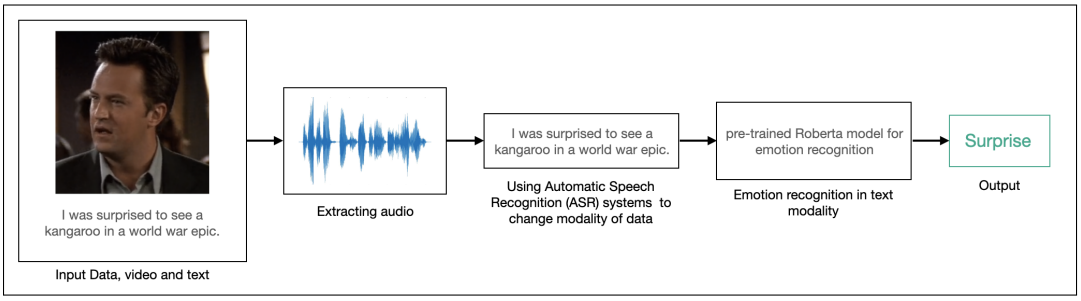

Figure 1: Our emotion recognition system pipeline consists of three main steps: speech-to-text conversion, text preprocessing, and emotion recognition using a fine-tuned Roberta model.

4 RESULTS

We compared the performance of our proposed modality conversion approach with the current state-of-the-art speech-based models on the MELD dataset. The results in Table 1 show that our approach outperformed the baseline models, achieving a weighted F1 score of 0.5. Specifically, the baseline models achieved weighted F1 scores of 0.418 and 0.419, respectively. These results indicate the effectiveness of our proposed modality conversion approach for emotion recognition from speech.

5 CONCLUSION

Based on our experiments using the MELD dataset, we proposed a modality conversion approach for speech emotion recognition. We converted speech to text using the Vosk API for ASR and then used a pre-trained Roberta model for emotion recognition on the text transcripts. Our approach resulted in a weighted F1 score of 0.428, which is a 5% improvement over the state-of-the-art speech-based models on the same dataset.

Our results demonstrate the potential of using modality conversion techniques and text-based features for enhancing emotion recognition from speech. Our approach offers an alternative to traditional speech-based approaches, which rely solely on analyzing signal features. By using text-based techniques, we can overcome some of the challenges posed by audio complexity and expression variability.

Future work can explore different ASR systems and NLP techniques for further improving the performance of the modality conversion approach. Additionally, incorporating visual cues and other modalities can be explored to further enhance emotion recognition accuracy. Overall, our study highlights the promising direction of modality conversion and multi-modal approaches for improving emotion recognition from speech.

ACKNOWLEDGEMENTS

We would like to acknowledge the support and resources provided by our institution for this research. We would also like to thank the developers of the MELD dataset, as well as the creators of the Vosk API and the pre-trained Roberta model. Lastly, we would like to express our gratitude to Soroush Gooran for his valuable insights and assistance throughout the course of this project.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://aclanthology.org/2020.findings-emnlp.148.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. SpeechFormer: A hierarchical efficient framework incorporating the characteristics of speech. In *Interspeech 2022*. ISCA, September 2022. doi: 10.21437/interspeech.2022-74. URL https://doi.org/10.21437/interspeech.2022-74.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4651–4660, 2022. doi: 10.1109/CVPRW56347.2022.00511.

Joosung Lee and Wooin Lee. CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5669–5679, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.416. URL https://aclanthology.org/2022.naacl-main.416.

MultiMedia LLC. Vosk speech recognition toolkit: Offline speech recognition api for android, ios, raspberry pi and servers with python, java, c and node, 2020. URL https://github.com/alphacep/vosk-api.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050.

Gaurav Sahu. Multimodal speech emotion recognition and ambiguity resolution, 2019. URL https://arxiv.org/abs/1904.06022.

Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. In *Proc. Interspeech 2020*, pp. 3755–3759, 2020. doi: 10.21437/Interspeech.2020-1212.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5197–5206, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.347.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, December 2018. doi: 10.1109/slt.2018.8639583. URL https://doi.org/10.1109/slt.2018.8639583.