Dimensional Speech Emotion Recognition: a Bimodal Approach

Larissa Guder
larissa.guder@edu.pucrs.br
Pontifical Catholic University of Rio Grande do Sul
Porto Alegre, Rio Grande do Sul

João Paulo Aires
joao.souza91@edu.pucrs.br
Pontifical Catholic University of Rio Grande do Sul
Porto Alegre, Rio Grande do Sul

Dalvan Griebler
dalvan.griebler@pucrs.br
Pontifical Catholic University of Rio Grande do Sul
Porto Alegre, Rio Grande do Sul

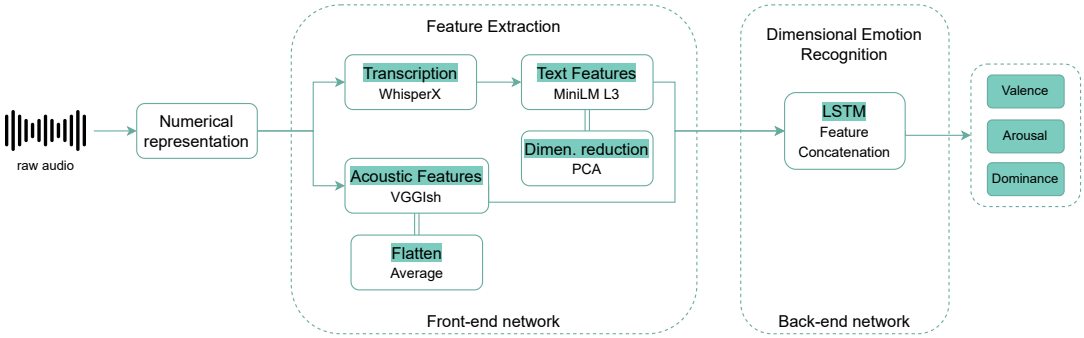

Figure 1: End-to-End Speech Emotion Recognition Architecture

ABSTRACT

Considering the human-machine relationship, affective computing aims to allow computers to recognize or express emotions. Speech Emotion Recognition is a task from affective computing that aims to recognize emotions in an audio utterance. The most common way to predict emotions from the speech is using pre-determined classes in the offline mode. In that way, emotion recognition is restricted to the number of classes. To avoid this restriction, dimensional emotion recognition uses dimensions such as valence, arousal, and dominance, which can represent emotions with higher granularity. Existing approaches propose using textual information to improve results for the valence dimension. Although recent efforts have tried to improve results on speech emotion recognition to predict emotion dimensions, they do not consider real-world scenarios, where processing the input in a short time is necessary. Considering these aspects, this work provides the first step towards creating a bimodal approach for Dimensional Speech Emotion Recognition in streaming. Our approach combines sentence and audio representations as input to a recurrent neural network that performs speech-emotion recognition. We evaluate different methods for creating audio and text representations, as well as automatic speech recognition techniques. Our best results achieve 0.5915 of CCC for arousal, 0.4165 for valence, and 0.5899 for dominance in the IEMOCAP dataset.

KEYWORDS

Affective Computing, Natural Language Processing, Streaming

In: VI Concurso de Teses e Dissertações (CTD 2024). Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (CTD'2024). Juiz de Fora/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2024.
© 2024 SBC – Sociedade Brasileira de Computação.
ISSN 2596-1683

1 INTRODUCTION

Our emotions play a subjective and controversial role, vital to our psychic survival. Understanding, to a certain extent, the emotions of other people and how they express them is fundamental to relating to each other as a society. For example, while fear is a natural protective regulator and aids decision-making, anger allows us to set limits and develop our sense of justice. An example of the importance of understanding emotions is that in autistic people, persistent deficits in emotional reciprocity and non-verbal communication, along with other factors, can lead to greater difficulty in communication and social interaction [1]. Based on this, emotion recognition is more a perspective than an exact science.

Dimensional Speech Emotion Recognition has many potential applications in the real world. Using dimensions, it is possible to map and identify anxious traces and reactions, check if a class is boring to the students, detect if a driver is tired while driving, and determine the level of customer satisfaction, among other things. However, there is a gap between the literature and the real world. While we have many approaches for SER, no one is built to support real-world scenarios by processing information as soon as it is available. Models that run on a streaming environment must be fast enough to bring results as soon as information arrives, but they also need good output accuracy.

This work introduces a dimensional speech emotion recognition approach using bimodal features focusing on a streaming scenario. This approach results from the Master's Degree in Computer Science from the School of Technology of the Pontifical Catholic University of Rio Grande do Sul (PUCRS). The dissertation defense to the evaluation board occurred on March 22, 2024. This research was also accepted for publication in the XXIV Brazilian Symposium on Computing Applied to Health (SBCAS) [5].

Our contribution was given in five main aspects: (1) the identification of the better approach for automatic speech recognition; (2)

the identification of the better way to generate the sentence embeddings for SER; (3) the identification of the better option between hand-crafted features and audio embedding for acoustic representation. (4) the identification of better options for feature fusion. (5) an architecture to execute SER on a streaming environment.

2 METHODOLOGY

To define the libraries, models, and architecture for our final approach, we conducted empirical experiments to identify the best options in each scenario. We divide the experimental process into two main steps: (1) feature selection and (2) fusion approaches. The first step is to select the best way to represent the textual and acoustic information. The second one is important to determine the best way to use both representations in our model. To evaluate all scenarios, we used the IEMOCAP (The Interactive Emotional Dyadic Motion Capture) dataset, which contains approximately 12 hours of speech in total.

Considering the application in a real-world scenario, in this case, on streaming flow, optimal libraries or models must be chosen to generate the representations of the input sources. The first step is the use of textual and acoustic information. So we define a set of experiments to select the optimal choice for (1-A) transcribing the audio, (1-B) generating a representation for the acoustic information, and (1-C) generating sentence embeddings for textual information. The objective for each one is to define the best option, considering the speed at which the data is processed and the lower error rate on evaluation. In our tests, we evaluate the ASR pre-trained models: Wav2Vec2, WhisperX, fine-tuned XLSR-53 Wav2Vec2, HuBERT, Seamless M4T v2, and Whisper v3.

Considering the different existing ways to (1-B) generate a representation for acoustic information, we selected two approaches: handcrafted features and audio embeddings. We use OpenSmile and pAA libraries for eGeMAPS, ComParE, and pAA sets to extract hand-crafted features. To generate audio embedding, we use the pre-trained VGGish and TRILL models. Since we aim to keep the sentence's meaning for recognizing emotion, we define the use of (1-C) sentence embeddings for textual information. We used models from the Sentence Transformer library to generate the embeddings: MiniLM-L12, mpnet, and MiniLM-L3. We evaluate experiments (1-B) and (1-C) using an LSTM network that predicts valence, arousal and dominance. LSTM is a learning model designed to work with sequential data, which fits the scenario of our experiments. We based our network architecture on previous work from Atmaja and Akagi [2].

Once the features used to represent the acoustic and textual data are defined, we evaluate the best way to use both types of information. To do this, we followed some of the approaches reviewed by Atmaja *et al.* [3]. We consider the (2-A) model level, (2-B) feature level, (2-C) decision-level fusion, and (2-D) average from acoustic and linguistic features.

Our final architecture is presented in Figure 1. We have a front-end and back-end block. In the front-end, we extract the features, and in the back-end, we predict the output. Given the defined architecture and the LSTM model trained, we build a streaming environment to run our pipeline. The final algorithm captures the microphone input in streaming and sends the representation to a Kafka queue every three seconds. The processing occurs in Flink, which calls a request from an external API that returns the predicted arousal, valence and dominance values for that utterance.

3 CONTRIBUTIONS

Recent reviews like Geetha *et al.* [4] and Lieskovská *et al.* [6], show a direction for future works in real-world applications that can be used in real-time. To make this possible, the processing time must be considered. However, current publications did not show the processing time necessary to execute their approach. The main focus is the feature selection for better results and the model's architecture. With the LSTM, the total prediction time for our test set was 1.2794 seconds.

Wundt and Judd [7] define that depending on the symptomatic nature of emotions, one of the forms of expressive movements is the expression of ideas. Which can be pantomimetic or descriptive. Due to genetic relationships with speech, it has a special psychological meaning. So, due to the importance of expressing ideas in emotion expression and the lack of diverse and large datasets [4], sentence representations add contextual information to predict the valence and give a modest contribution to the arousal and dominance dimension. The sentence embeddings are the best options when considering the sentence's meaning. The results on valence when using only the Mini LM L3 reflect the good results on the sentiment evaluation databases.

It is controversial to consider that speech emotion recognition can be done in real time. This is because when we consider the use of sentence embedding, the sentence must be complete to gain more context and meaning. Even if we use real-time transcription, we will deal with, in the better case, words. So, considering the average length of the annotated data chunks from IEMOCAP, we determine our windowing time to be 3 seconds of utterances.

In this research, we created a complete architecture for speech emotion recognition that can run in streaming scenarios. This is an important step because it shows that it is possible to use it to solve real-world situations.

REFERENCES

[1] American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR.* American Psychiatric Association Publishing, USA.

[2] Bagus Tris Atmaja and Masato Akagi. 2020. Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. *APSIPA Transactions on Signal and Information Processing* 9 (May 2020), e17. https://doi.org/10.1017/ATSIP.2020.14

[3] Bagus Tris Atmaja, Akira Sasou, and Masato Akagi. 2022. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication* 140 (May 2022), 11–28. https://doi.org/10.1016/j.specom.2022.03.002

[4] A.V. Geetha, T. Mala, D. Priyanka, and E. Uma. 2024. Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions. *Information Fusion* 105 (March 2024), 102–218. https://doi.org/10.1016/j.inffus.2023.102218

[5] Larissa Guder, João Aires, Felipe Meneguzzi, and Dalvan Griebler. 2024. Dimensional Speech Emotion Recognition from Bimodal Features. In *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde* (Goiânia/GO). SBC, Porto Alegre, RS, Brasil, 579–590. https://doi.org/10.5753/sbcas.2024.2779

[6] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. 2021. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* 10 (January 2021), 1163. https://doi.org/10.3390/electronics10101163

[7] W.M. Wundt and C.H. Judd. 1897. *Outlines of Psychology.* W. Engelmann.