Dimensional Speech Emotion Recognition from Bimodal Features

Larissa Guder[1], João Paulo Aires[1], Felipe Meneguzzi[12], Dalvan Griebler[1]

[1]School of Technology – Pontifical Catholic University of Rio Grande do Sul (PUCRS)
Porto Alegre – 90619–900 – Brazil

[2]University of Aberdeen – Aberdeen – Scotland

{larissa.guder,joao.souza91}@edu.pucrs.br

felipe.meneguzzi@abdn.ac.uk, dalvan.griebler@pucrs.br

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 to 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 quickly is necessary. Considering these aspects, we take 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. Our final architecture achieves a Concordance Correlation Coefficient of 0.5915 for arousal, 0.1431 for valence, and 0.5899 for dominance in the IEMOCAP dataset.*

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 [Association 2022]. Based on this, emotion recognition is more a perspective than an exact science.

Besides the ways used to determine emotions in psychology, two approaches have been used to recognize emotions using deep learning: discrete classes and dimensional [Lieskovská et al. 2021]. In discrete classes, the six emotions considered essential by [Ekman 1999]: anger, disgust, fear, happiness, sadness, and neutral are used, where

the model must classify the input according to the most correlated class. On the other hand, [Russell 1980] defines a dimensional approach through the circumplex model of affect. The circumplex model considers two dimensions: arousal and valence. Each dimension has a value that ranges from -1 to 1. Arousal is related to calming or exciting the tonality of speech, while valence represents how pleasant or not it is. With the score of each dimension, it is possible to correlate to a specific emotion. For example, fear and anger can be defined with low valence and high arousal. [Mehrabian 1996] adds the dominance dimension, representing how emotion influences a person's behavior. It is important that models recognize emotions and respect each person's idiosyncratic diversity.

Leaving the direct application in psychology, different sectors benefit from recognizing emotions daily. The review by [Geetha et al. 2024] identifies sectors like education, healthcare, marketing and advertising, human-robot interaction, security and surveillance, customer service, sports, entertainment, gaming, and the automotive industry. Conversely, the preoccupation with privacy and the possible emotional state exploration to induce the user to buy some services or products is discussed by [Testa et al. 2023].

The lack of data for training and testing deep learning models makes it difficult for the field of SER to grow [de Lope and Graña 2023]. Existing datasets have a small amount of available data, are less diverse than necessary, or are too different from real-world data. Even when focusing only on speech emotion recognition, it is necessary to consider that human emotion perception involves multiple senses, being multimodal [Geetha et al. 2024]. So, to overcome, and extract more information from only spoken data, the use of textual information can improve the precision of the predictors. Some authors have already shown that using text features, such as word embeddings, improves valence prediction [Triantafyllopoulos et al. 2022, Srinivasan et al. 2022, Ghriss et al. 2022, Atmaja and Akagi 2020, Atmaja and Akagi 2021, Sogancioglu et al. 2020, Julião et al. 2020]. However, including new features in the processing usually increases the time necessary to generate an output. For instance, the inclusion of text features requires first transcribing the audio to use it as input.

The main problem is that the existing approaches for speech emotion recognition do not evaluate the processing time necessary to extract the information from audio and predict the emotion dimensions. Also, there are no dimensional applications in the streaming environment. In this way, the main aim of this work is to create an architecture for speech emotion recognition that is useable in a streaming environment. We combine audio and sentence embeddings for speech emotion recognition to make this possible. In addition, we empirically show the effectiveness by evaluating the time necessary to extract and process the features, the Mean Squared Error (MSE) metric for emotion recognition, and the Word Error Rate (WER) for automatic speech recognition.

The article is divided into three sections: first, we discuss some related works and the main differences between them and our proposal in Section 2. Our main findings are presented in Section 3, where we present our final architecture and compare the results with state-of-the-art approaches. Finally, Section 4 discusses our contributions and future directions.

2. Related Work

In this section, we introduce some related work found in the literature. We divide the analysis of the related works in two ways: (1) approaches that use dimensional emotion recognition and text features; and (2) approaches that apply their models in a streaming scenario. This division was necessary because we did not identify any work that used a bimodal model with dimensional data in a streaming scenario. For the specific scenario, we have only a few models that use classes and audio-only data.

Using text features, more precisely word embeddings, demonstrably improves results on the valence dimension. While the dominance and arousal are affected only by the acoustic features [Triantafyllopoulos et al. 2022, Srinivasan et al. 2022, Ghriss et al. 2022, Atmaja and Akagi 2020, Atmaja and Akagi 2021, Sogancioglu et al. 2020, Julião et al. 2020]. We notice the use of GloVe by [Atmaja and Akagi 2020, Atmaja and Akagi 2021] and more recent approaches, such as BERT [Srinivasan et al. 2022, Julião et al. 2020, Sun et al. 2020] and a derivation of it called camemBERT [MacAry et al. 2021], and DeBERTaV3 [Ispas et al. 2023].

All of them use word representation level. We evaluate the use of sentence-level representations. This is because we will infer the emotion based on a sentence, not for each pronounced word. Keeping on that way, the meaning and the context of the words in the sentence.

Using dimensional approaches, we found a focus on the acoustic features used. For example, we used eGEMAPS and ComParE feature sets, which improved SER results. Considering emotions, audio embeddings were explored in the music emotion recognition task. [Koh and Dubnov 2021] evaluate L3-Net [Cramer et al. 2019] and VGGish models. For SER, [Wang et al. 2022] explored VGGish, but for categorical evaluation, while for dimensional [Julião et al. 2020] explored the use of x-vectors [Snyder et al. 2018] embedding, [Sun et al. 2020] even evaluated the use of VGGIsh, but not for the bimodal approach. More recent approaches consider the use of w2v2 [Triantafyllopoulos et al. 2022] and HuBERT [Ispas et al. 2023, Srinivasan et al. 2022] to generate the representations. Independent of the method to extract the features from the audio, even using pre-trained models or hand-crafted options, none had the time to process this information. Our approach compares the ComParE, eGeMAPS, pAA feature sets, and TRILL and VGGISH models for audio embeddings.

Focusing on approaches developed for streaming scenarios, we only found works that use classes. Also, our focus is on bimodal features, while [Stolar et al. 2017], [Bertero et al. 2016], and [Lech et al. 2020] use only acoustic features. Another point is that these papers are from before 2020, and after that, we do not have publications that focus on SER that run on a streaming environment, different from the ASR task, where we have some new approaches over the years, such as [Dominguez-Morales et al. 2018, Singh et al. 2019, Leow et al. 2020] and [Saeki et al. 2021]. It is important to notice that only [Lech et al. 2020] provides metrics for evaluating streaming scenarios. [Stolar et al. 2017] and [Bertero et al. 2016] only mentioned that their approaches are in real-time but do not show the result.

Dimensional Speech Emotion Recognition has many potential applications in the real world. Using dimensions, it is possible to map and identify anxious traces and reac-

tions, check if a class is boring to the students, detect if a driver is tired while driving, and determine customer satisfaction, among other things. However, there is a gap between the literature and the real world, in which we have many approaches for SER, but no one is built to support real-world scenarios with processing information as soon as they are 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. Because of this, this work aims to combine SER, deep learning, and streaming to build a robust approach that can be applied to the real world.

3. Proposed Architecture

This section presents our proposed approach to recognizing emotions in speech in a streaming environment. We combine sentence embeddings with audio embeddings to generate representations from text and audio information. Then, we predict the arousal, valence, and dominance values through an LSTM network. The following sub-sections present the dataset, proposed architecture, evaluation results, and streaming application.

3.1. Dataset

To evaluate our experiments, we use the IEMOCAP (The Interactive Emotional Dyadic Motion Capture) dataset [Busso et al. 2008]. IEMOCAP contains multimodal information, combining video, speech, motion capture of face, and text transcriptions. Using different sources of information can lead to more robust predictions. However, in some cases, visual information is not available. From the features on the dataset, we use speech and text transcriptions in our approach. In total, the dataset contains approximately 12 hours of speech. IEMOCAP provides an AVD score and an emotion class annotation for each utterance. VAD scores range from 1 to 5. The dataset contains approximately 12 hours of speech. Since IEMOCAP does not contain information about the split ratio, we divided it into 60/20/20 ratios for training, testing, and validation. The validation set was used to compute the results of all experiments. In total, the 1992 utterances from the dataset have 8909 seconds of duration.

We normalized to a -1 to 1 scale with the Equation 1. This normalization is since the original Russel approach uses the -1 to 1 scale, which is the pattern we use in our final architecture.

$$\frac{x - \left(\frac{max-min}{2} + 1\right)}{\frac{max-min}{2}} \tag{1}$$

3.2. End-to-End Speech Emotion Recognition Architecture

Information extraction from the speaker's speech is a crucial step in the process of recognizing emotions in speech. We can consider two different kinds of information: the acoustic, which involves how the speech is pronounced, and the textual, which contains the meaning of the speech. The inclusion of this information requires a pre-processing step. For this, our end-to-end architecture consists of two blocks:

the front-end and the back-end. The front-end is responsible for extracting features from the input signal, while the back-end is responsible for processing the information

from the front-end and predicting the output. We detail the architecture in Figure 1. Given raw audio, we transform it into a mono waveform and resample it into a 16 kHz sample rate. Due to the VGGIsh input limitation, we limit the audio length to 10 seconds. We extract two types of features from the waveform: textual and acoustic.

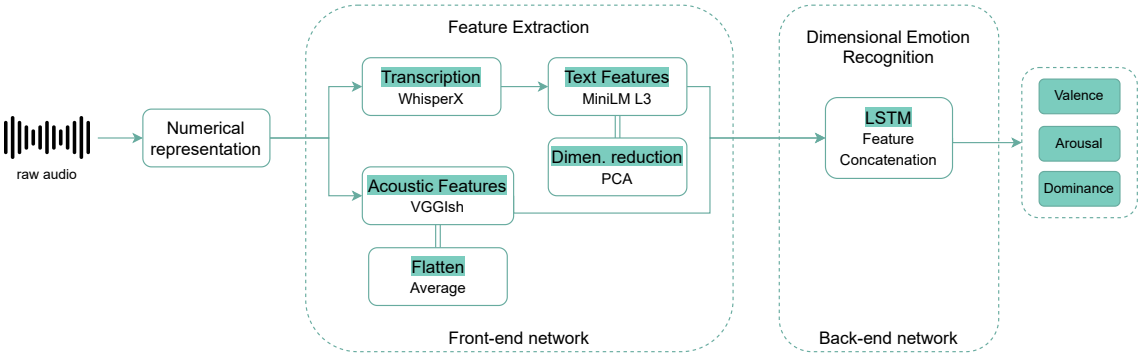

Figure 1. End-to-End Speech Emotion Recognition Architecture

The objective of the front-end is to extract and pre-process the textual and acoustic features, providing the correct shape to the back-end network so that it can concatenate and process it. The expected output is two vectors with 128 dimensions each. For acoustic features, we generate audio embedding using the pre-treined VGGIsh model. VGGish [Hershey et al. 2017] is a modification of the VGG16 architecture [Simonyan and Zisserman 2015], a popular convolutional neural network. The authors trained the VGGish model on a large YouTube dataset. To process each second of audio, VGGish needs an average of 2.97ms. The model generates a vector with 128 dimensions for each second of audio. We calculate the average from all rows in the matrix as a flattened function, generating a unique vector with 128 dimensions for the whole audio for our back-end network.

The text features require an extra processing stage. We use the WhisperX model to convert the input waveform into text, thus allowing sentence embedding to be generated for textual representation. The average time to transcribe each second of audio was 2.803 ms. To generate the sentence embedding, we use the MiniLM L3 pre-trained model. The MiniLM [Wang et al. 2020] is a task-agnostic and distilled approach focusing on a lightweight version of Transformer-based models. Using the teacher-student architecture, the authors propose a distilled version of the self-attention heads of the teacher to make this possible. MiniLM produces a representation with 384 dimensions for sentences and needs, on average, 0.47ms to process each second of transcribed audio. To match the size of the audio features, we use the Principal Component Analysis (PCA) algorithm to reduce the dimension to 128.

The back-end network uses an LSTM network to process the incoming data. The first layer concatenates both feature sets. We use the order $audio, text$. After the input layer, we use a batch normalization layer to standardize the features. We use only two LSTM layers, the first with 128 units and the second with 256 units, followed by a dense layer with 64. We apply a dropout with a 0.25 probability after the dense layer. The output is a dense layer with three values corresponding to valence, arousal, and dominance dimensions. We use $tanh$ as the activation function and Adam optimizer with a 0.001

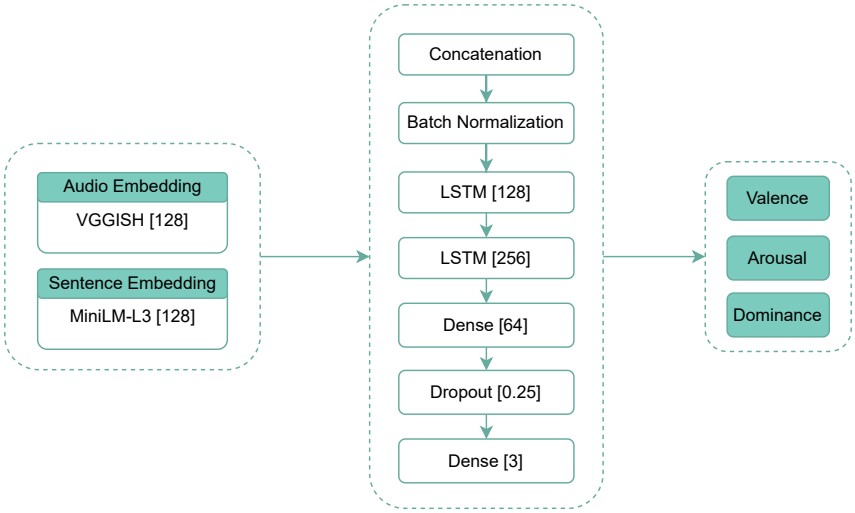

Figure 2. Back-end Architecture

learning rate.

3.3. Evaluation Results

We train and evaluate our model on the IEMOCAP dataset. On IEMOCAP, we used the solution provided by [Atmaja and Akagi 2020] as a baseline to compare our approach. [Atmaja and Akagi 2020] also uses an LSTM model with GloVe for textual features combined with pAA HSF for acoustic features.

The main point in defining our architecture is the time necessary to process the incoming data. While [Atmaja and Akagi 2020] focuses on word embedding, with GloVe, we focus on capturing the sentence's meaning through the sentence embedding from MiniLM L3. The MiniLM L3 was tested on the Sentiment Analysis task and performed well on Stanford Sentiment Treebank (SST) [Socher et al. 2013]. The textual embedding focuses on improving the valence dimension; the task is close to sentiment analysis, going from negative to positive perspectives.

Table 1. IEMOCAP evaluation results

Mode	CCC/MSE			AVG
	Valence	Arousal	Dominance	
Baseline				
Bimodal LSTM (GloVe + HSF from pAA) [Atmaja and Akagi 2020]	**0.418**	0.571	0.500	0.496
Our approach				
VAD MiniLM-L3 VAD	0.4165	0.2989	0.2989	0.3381
LSTM Concat (VGGISH + MiniLM-L3 PCA) VAD	0.1431	**0.5915**	**0.5899**	0.4415

On the acoustic side, the use of VGGIsh to recognize emotions has been explored by [Pham et al. 2023] in bimodal categorical speech emotion recognition and by [Koh and Dubnov 2021] in music emotion recognition. [Pham et al. 2023] uses the concatenation of VGGIsh and BERT to recognize emotions. In addition to the mode to recognize emotion, the main difference in our approach is in the architecture used and the

textual representation. Originally, VGGIsh was trained to focus on audio classification tasks and achieved better results than hand-crafted features on the Audio Set Acoustic Event Detection (AED) classification task. Using GPU, the processing time of VGGIsh took 2.97ms per second of audio, while the approach of [Atmaja and Akagi 2020] uses pAA with 9.13ms per second. Analyzing the best scenario for each dimension, on valence, we have a loss of 0,359% of CCC in relation to baseline, while for arousal, we have a gain of 3.59%, and for dominance, 17.98%.

3.4. Streaming

The streaming implementation took place in two ways: one for evaluation and the other for real-world application. This is necessary since there are no datasets available for streaming scenarios. So, to make the evaluation possible, we iterate over the data, preserving the duration of each file annotated. In the real-world scenario, we used a window time-based to split the incoming signal. We present the architecture in Figure 3.

To generate the audio input streaming, we use the pyAudio streaming function to capture the signal from the microphone as mono. We specify the params used to capture the audio in the Table 2. The number of chunks is calculated by multiplying the chunk length and the sample rate. The chunk represents the number of frames into a mel spectrogram input, calculated over the number of samples divided by the hop length. We use a mono channel.

Table 2. pyAudio parameters for audio capturing

Parameter	Value
Sample Rate	16000
N FFT	400
N MELS	80
Hop Length	160
Chunk Length	30
Number of Samples	30 * 16000
Chunk	480000 / 160
Format	pyaudio.paInt16
Channels	1

After the windowing process, we convert the input signal into a numerical representation. We use the Whisper function, which uses FFmpeg to convert the signal into a waveform. After that, we use the Kafka producer to send the waveform to the queue, which Flink will process. To predict the values for valence, arousal, and dominance, we created an API using Flask to receive the requests from Flink. We use an API because Tensorflow models cannot be used in a streaming environment.

Our API has four different endpoints; in that way, we can use different producers in Flink. First, we transcribe and generate the audio embedding. After that, using the transcription, we generate the sentence embedding and apply the PCA to reduce dimensionality. With both embeddings, we predict the three dimensions using our LSTM model. After getting the prediction, we remove the waveform from the Kafka queue.

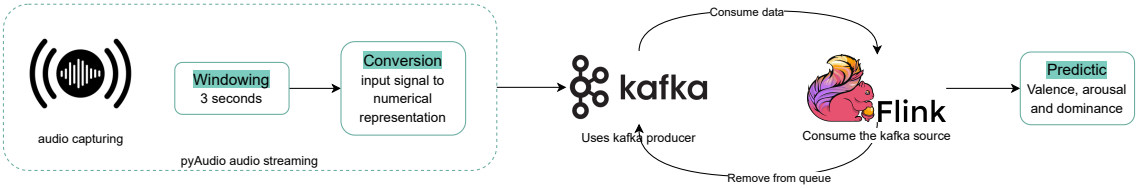

Figure 3. Architecture used for streaming speech emotion recognition

4. Discussion

Different to the literature, our highest gain using a bimodal approach was on the dominance dimension instead of valence as presented in the related works. We achieved a 17.98% gain in CCC in comparison to the [Atmaja and Akagi 2020] approach. This is more correlated to the way used to represent the audio withVGGIsh. The incorporation of sentence embeddings adds only 4.12% of CCC.

When we compare the results for valence using only the Mini LM L3 model, they are similar to the bimodal approach of [Atmaja and Akagi 2020] (0.418 vs 0.4165). Our main issue is the dimensionality reduction for using sentence embedding in the concatenation layer in the Keras model. This is necessary to obtain a sentence embedding with the same number of dimensions as the audio embedding. In the case where we only use PCA as input, CCC is reduced to 0.1055. The concatenation provided a better result, with 0.1431 CCC. But it is still worse than the original size one. This occurs because we apply the PCA after generating the embedding. A possible solution is adding a new dense layer to the Mini LM model and producing the embedding directly with 128 dimensions.

From the most recent machine learning approaches to extract information from audio, we evaluate the VGGIsh and TRILL models, regarding that they are used to feed our LSTM network. Another possible option is to use a CNN network with features from Wav2Vec2, Wav2Vec2-BERT 2.0, Hubert, and another model that generates more complex representations. Wav2Vec2-BERT 2.0, for example, creates a representation with 1024 dimensions for each x ms. To be able to use only one dimension, we apply an average function to the VGGIsh matrix embedding. They produce an array for each second of audio input.

Recent reviews like [Geetha et al. 2024] and [Lieskovská et al. 2021], 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 1897] 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 [Geetha et al. 2024], 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 if we consider the use of sentence embedding, the sentence must be complete to get more context and meaning from it. 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 and MSP-PODCAST, we determine our windowing time to be 3 seconds of utterances.

5. Conclusion

This work introduces a dimensional speech emotion recognition approach using bimodal features, that can be applied in a streaming environment. As a result, we achieve 0.5915 of CCC for arousal, 0.1431 for valence, and 0.5899 for dominance. 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 AVD values for that utterance.

In future work, we plan to use a pretrained version of the Mini LM L3 model to directly produce vectors with 128 dimensions as the output. With this, we aim to increase the CCC for the valence dimension as well. By consolidating the best features, we also aim to test with new architectures, such as Transformer, and use different datasets to train and evaluate our approach. Finally, considering the streaming scenario, we aim to add a sink operation and use a visual approach to understand the model prediction output.

References

Association, A. P. (2022). *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*. American Psychiatric Association Publishing.

Atmaja, B. and Akagi, M. (2020). Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. *APSIPA Transactions on Signal and Information Processing*, 9.

Atmaja, B. and Akagi, M. (2021). Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm. *Speech Communication*, 126:9–21.

Bertero, D., Siddique, F. B., Wu, C.-S., Wan, Y., Chan, R. H. Y., and Fung, P. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1042–1047, Austin, Texas. Association for Computational Linguistics.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Cramer, A. L., Wu, H.-H., Salamon, J., and Bello, J. P. (2019). Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019 - 2019 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3852–3856.

de Lope, J. and Graña, M. (2023). An ongoing review of speech emotion recognition. *Neurocomputing*, 528:1–11.

Dominguez-Morales, J. P., Liu, Q., James, R., Gutierrez-Galan, D., Jimenez-Fernandez, A., Davidson, S., and Furber, S. (2018). Deep spiking neural network model for time-variant signals classification: a real-time speech recognition approach. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Ekman, P. (1999). Basic emotions. In Dalgleish, T. and Powers, M. J., editors, *Handbook of Cognition and Emotion*, pages 4–5. Wiley.

Geetha, A., Mala, T., Priyanka, D., and Uma, E. (2024). Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, 105.

Ghriss, A., Yang, B., Rozgic, V., Shriberg, E., and Wang, C. (2022). Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2022-May, pages 7347–7351.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (2017). Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 131–135. IEEE Press.

Ispas, A.-R., Deschamps-Berger, T., and Devillers, L. (2023). A multi-task, multi-modal approach for predicting categorical and dimensional emotions. In *ACM International Conference Proceeding Series*, page 311 – 317.

Julião, M., Abad, A., and Moniz, H. (2020). Exploring text and audio embeddings for multi-dimension elderly emotion recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2020-October, pages 2067–2071.

Koh, E. S. and Dubnov, S. (2021). Comparison and analysis of deep audio embeddings for music emotion recognition. *CoRR*, abs/2104.06517.

Lech, M., Stolar, M., Best, C., and Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers in Computer Science*, 2.

Leow, C. S., Hayakawa, T., Nishizaki, H., and Kitaoka, N. (2020). Development of a low-latency and real-time automatic speech recognition system. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pages 925–928.

Lieskovská, E., Jakubec, M., Jarina, R., and Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10.

MacAry, M., Tahon, M., Esteve, Y., and Rousseau, A. (2021). On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion

recognition. In *2021 IEEE Spoken Language Technology Workshop, SLT 2021 - Proceedings*, pages 373–380.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology*, 14:261–292.

Pham, N. T., Dang, D. N. M., Pham, B. N. H., and Nguyen, S. D. (2023). Server: Multimodal speech emotion recognition using transformer-based and vision-based embeddings. In *Proceedings of the 2023 8th International Conference on Intelligent Information Technology*, ICIIT '23, page 234–238, New York, NY, USA. Association for Computing Machinery.

Russell, J. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39:1161–1178.

Saeki, T., Takamichi, S., and Saruwatari, H. (2021). Low-latency incremental text-to-speech synthesis with distilled context prediction network. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 749–756.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Singh, R., Yadav, H., Sharma, M., Gosain, S., and Shah, R. R. (2019). Automatic speech recognition for real-time systems. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 189–198.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2018-April, page 5329 – 5333.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Sogancioglu, G., Verkholyak, O., Kaya, H., Fedotov, D., Cadée, T., Salah, A., and Karpov, A. (2020). Is everything fine, grandma? acoustic and linguistic modeling for robust elderly speech emotion recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2020-October, pages 2097–2101.

Srinivasan, S., Huang, Z., and Kirchhoff, K. (2022). Representation learning through cross-modal conditional teacher-student training for speech emotion recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2022-May, pages 4298–4302. cited By 0.

Stolar, M. N., Lech, M., Bolia, R. S., and Skinner, M. (2017). Real time speech emotion recognition using rgb image classification and transfer learning. In *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–8.

Sun, L., Lian, Z., Tao, J., Liu, B., and Niu, M. (2020). Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*, MuSe'20, page 27–34, New York, NY, USA. Association for Computing Machinery.

Testa, B., Xiao, Y., Sharma, H., Gump, A., and Salekin, A. (2023). Privacy against real-time speech emotion detection via acoustic adversarial evasion of machine learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7.

Triantafyllopoulos, A., Wagner, J., Wierstorf, H., Schmitt, M., Reichel, U., Eyben, F., Burkhardt, F., and Schuller, B. (2022). Probing speech emotion recognition transformers for linguistic knowledge. In *Proc. Interspeech 2022*, volume 2022-September, pages 146–150.

Wang, C., Ren, Y., Zhang, N., Cui, F., and Luo, S. (2022). Speech emotion recognition based on multi-feature and multi-lingual fusion. *Multimedia Tools and Applications*, 81:4897–4907.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Wundt, W. and Judd, C. (1897). *Outlines of Psychology*. W. Engelmann.