ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Decision fusion based multimodal hierarchical method for speech emotion recognition from audio and text

Nawal Alqurashi, Yuhua Li, Kirill Sidorov and David Marshall

Cardiff University - School of computer Science and Informatics - UK

Abstract. Expressing emotions is essential in human interaction. Often, individuals convey emotions through neutral speech, while the underlying meaning carries emotional weight. Conversely, tone can also convey emotion despite neutral words. Most Speech Emotion Recognition research overlooks this. We address this gap with a multimodal emotion recognition system using hierarchical classifiers and a novel decision fusion method. Our approach analyses emotional cues from speech and text, measuring their impact on predicted classes, considering emotional or neutral contributions for each instance. Results on the IEMOCAP dataset show our method's effectiveness: 69.45% and 65.26% weighted accuracy in speaker-dependent and speaker-independent settings, respectively.

1 Introduction

Speech emotion recognition (SER) plays a vital role in human-computer interaction. Recently, increasing attention has been directed to the study of using a variety of modalities in emotion recognition emphasizing that using more than one modality outperforms the unimodal approaches in different scenarios [1]. Utilizing information from multiple modalities leads to the use of multimodal emotion data fusion techniques. Fusion strategies typically fall into two types: feature-level (early) fusion and decision-level (late) fusion. Early fusion involves combining features from different modalities before classification, while late fusion combines decision values from individual classifiers into the final decision [2]. Traditional late fusion methods are mostly based on an ensemble of flat classifiers [3], where each example is assigned to an emotion out of a finite set of emotions at a one-level classification system and there is no hierarchical structure of emotions. However, emotion recognition is one of the real-world classification problems that are naturally cast as hierarchical classification problems [4], where emotions are classified at various levels into a predefined hierarchy of classes [5]. The differentiation between neutral and emotional speech at very early stages in the hierarchical classifier can carry considerable significance in the analysis of emotions between modalities. The intuition behind this order comes from the observation of conversations in real life, where some spoken instances can be expressed in a neutral tone yet, convey emotions through the text content rather than the tone of voice, this revealed in the experiment results from the work by Devillers et al. [6].

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Instances of such scenarios are evident in individuals diagnosed with autism spectrum disorder, social anxiety disorder and people experiencing depression, grief and loss. Conversely, some written phrases don't convey any emotional expressions and remain neutral, however, they could potentially express a clear emotion with voice tone. Thus, to analyse the relationships and intersections between neutrality and emotions, it is necessary to first differentiate between neutral and emotional occurrences within each modality. Moreover, Hierarchies effectively express generality and specificity between categories, placing broader ones at higher levels and narrower ones at lower levels [7]. However, there's no existing hierarchical structure organizing emotions from generic to specific using speech or multiple modalities. In light of these challenges, we propose a multimodal hierarchical system. We create an ensemble of hierarchical classifiers for acoustic and textual modalities independently. Our novel late fusion technique combines their predictions, offering insights into each modality's importance at each hierarchy level for predicting emotion classes.

The paper is organized as follows: Section 2 describes the proposed SER framework, Section 3 presents experimental results and discussions, and Section 4 provides the conclusion.

2 Proposed SER Methodology

2.1 Features Extraction

For the acoustic features, we focus on capturing the essential properties of the speech signal that reflect its phonetic and prosodic characteristics. We use the Librosa toolkit [8] to extract 39-dimensional Mel-frequency cepstral coefficients (MFCCs), which model the human ear's response to sound. Additionally, we use Librosa to extract handcrafted features proven useful in previous research [1]. These features, combined with the Geneva minimalistic acoustic parameter set (GeMAPS) [9], provide a comprehensive set of low-level descriptors. Using OpenSMILE [10], we extract GeMAPS, covering frequency, energy, and spectral parameters, capturing the nuances of speech. In total, we obtain 65 acoustic features. For textual features, we utilize a pretrained language model (BERT) [11] via Embedding4BERT [12] to obtain word embeddings, transforming text transcripts into a matrix that preserves semantic relationships and contextual information.

2.2 Proposed dual multimodal hierarchical approach

To perform hierarchical classification, we organize emotion categories into a two-level hierarchy. The first level distinguishes between neutral and emotional samples. Then, the second level further categorizes emotional samples into Happy, Sad, and Angry. We adjust annotations for the first level, keeping neutral samples unchanged (Neutral) and grouping emotional classes as (Emotional). For the second level, we retain the

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

original annotations for the three emotional classes. During training, Model 1(Fig. 1) is trained on the entire dataset, while Model 2 (second-level classifier) is trained only on emotional samples. During testing, Model 2 operates on results from the first level, potentially receiving misclassified non-emotional instances, providing realistic outcomes. Two hierarchical systems are used for audio and text, each providing its own predicted class.

2.3 Decision fusion based hierarchical classifiers

Inspired by Xu et al. [13], who applied Label Distribution Learning [14] to represent correlations between true labels and their siblings in hierarchical classifiers, we adapted their prediction phase method for our decision fusion. Specifically, we extend their approach by computing path scores based on label probabilities along the paths of the predicted classes from our multimodal hierarchical classifiers. The proposed decision fusion outputs the class with the highest path score as the final predicted class. To define the proposed fusion method, let h_i represent one hierarchical model, H be the set of hierarchical models we integrate in the fusion method, where i is the index of the model. We use l to represent the predicted class from a classifier in a particular level, where l_{ij} is the predicted class from the j −th level for one hierarchical model h_i. For class l, we denote its parent by $pa(l)$. We also let c_i indicate the last predicted class for the test instance x from the model h_i, thus $c_i \in C$, where C is the set of all predicted classes from different hierarchical models. We use $path(c_i)$ to express the number of classifiers from the first level leading to class c_i. In order to calculate the path score for the predicted class from each hierarchical model, we first apply Equation 1 to compute the logarithmic class probability of c_i

$$\ln\big(p(c_i|x)\big) = \sum_{j=1}^{path(c_i)} \ln\left(p\big(l_{ji}|pa(l_{ji}),x\big)\right) \quad (1)$$

Second, to avoid the impact of path length, we further divide the logarithmic class probability of the predicted class c_i by $path(c_i)$. Therefore, the path score for the predicted class is calculated by.

$$Ps(c_i|x) = \frac{\ln\big(p(c_i|x)\big)}{path(c_i)} \quad (2)$$

The final predicted class for the ensemble hierarchical models is the predicted class with the maximum path score.

$$\hat{y} = argmax_{c\in C}\, Ps(c_i|x) \quad (3)$$

By focusing on the maximum normalized path score, we aim to emphasize the most informative paths and potentially reduce the impact of less relevant information or noise, possibly leading to more accurate and reliable predictions. In the test phase (Fig.1), we demonstrate the late fusion approach. For instance, a test sample x is classified as happy in the speech model but as neutral in the text model. Using Equation 1 and 2, we calculate path scores for each predicted class. Considering path length, the

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

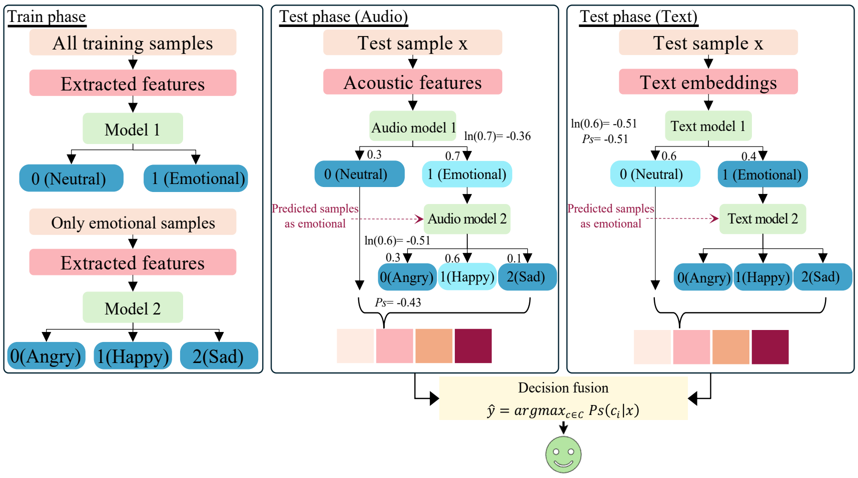

Fig. 1: Overall architecture of the proposed multimodal based hierarchical structure for SER

class with the higher score is chosen as the final prediction (e.g., happy in this example). This indicates that the speech modality strongly influences the emotional class determination for this instance. Furthermore, the hierarchy order reveals that text alone doesn't provide emotional content, evident from the first level of classification.

3 Experiments

3.1 Emotion Dataset

IEMOCAP [15] is a database of acted conversations with 10 speakers over five sessions, each with one male and one female. This study focuses on four emotional states: anger, happiness (merging excitement), sadness, and a neutral state. Using 5331 utterances with transcriptions, experiments are conducted with Speaker-Dependent (SD) and Speaker-Independent (SI) settings. In SD, data is split 80/20 for training and testing. In SI, four sessions train the model, and the last session tests it, ensuring no speaker overlap.

3.2 Results and Discussion

In this study, we use long short-term memory (LSTM) followed by two dense layers and a softmax activation to transform the LSTM output into class probabilities, which will be the classifier at each level in the hierarchical models. We set the batch size to 64 and use the Adam optimizer with cross entropy loss for training. The other hyperparameters are fine-tuned using the Optuna optimization framework [16], performing 100 iterations per model. Table 1 presents our system's performance in SD and SI settings for SER. The results demonstrate that the proposed late fusion of modalities improves recognition accuracy compared to single-modality models. Table 2 compares our method with several existing SER approaches that utilize fusion

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

techniques on the same dataset. Specifically, our system achieves 68.74% unweighted accuracy (UA) and 69.45% weighted accuracy (WA) in the SD setting, and 63.90% UA and 65.26% WA in the SI setting. Additionally, our model provides interpretability by highlighting the importance of each modality in distinguishing emotional and neutral states, offering valuable insights for the decision-making process. To illustrate and clarify this method's capabilities, we chose samples that exclusively express emotions. Thus, these samples are annotated with specific emotion classes, and the final system successfully predicts the correct emotion class for them. There are three cases of fusing the two hierarchical models: First: If the two models predict the same class (no conflict). Second: If the two models predict different classes, however they are both emotional. Third: If the two models predicted different classes one of them is a neutral class and the other is an emotional class. Table 3 shows the results of the chosen samples to illustrate the three cases of the fusion method. For example, with the “I am so sorry” instance, the fusion method takes the identical predictions from both models and produces it as the final decision. Conversely, in the case of the instance “That's so cool. Uh huh.”, both models predict it as an emotional class, but they do not agree on the specific type of emotion. The expression of this example was conveyed using a kind of screaming voice, which likely caused the speech model to predict it as angry. On the contrary, the text model easily recognized the correct emotion because it could grasp the meaning of the sentence without being influenced by the tone of voice. In such a situation, the fusion method makes its final decision as happy by selecting the highest path score between the predicted classes from the models.

4 Conclusion

We propose an ensemble of hierarchical classification models for SER, combining audio and text. Our late fusion technique, tailored for hierarchical classifiers, interprets modality importance and their relationships between categories within the hierarchy, enhancing final class prediction accuracy. Results show our framework outperforms previous multimodal fusion methods on the IEMOCAP dataset across four emotions. The proposed method can be adapted to various domains and modalities, handling multiple hierarchical models. Future work will explore these applications further.

Model		Modalities	UA	WA	F1
SD	Hierarchical models	Audio	60.16	62.01	60.71
		Text	65.31	65.75	65.28
	Late fusion	Audio + Text	68.74	69.45	68.74
SI	Hierarchical models	Audio	57.85	57.12	57.48
		Text	59.38	60.71	58.33
	Late fusion	Audio + Text	63.90	65.26	63.06

Table 1: Performance of the proposed approach on IEMOCAP

Model	UA	WA	F1
Sebastian et al. [17]	59.3	61.2	61.2
Li et al. [18]	-	63.4	-
Cho et al. [19]	64.3	63.1	-
Ours (SD)	68.7	69.4	68.7
Ours (SI)	63.9	65.2	63.0

Table 2: Performance comparison with representative methods

Sentences	SM's prediction	TM's prediction	Fusion result	Original annotation
"I am so sorry"	2-sad	2-sad	2-sad	2-sad
"That's so cool. Uh huh."	0-angry	1-happy	1-happy	1-happy
"We've got to say it to him"	0-angry	3-neutral	0-angry	0-angry
"Well, I lost them"	3-neutral	2-sad	2-sad	2-sad

Table 3: Results of chosen samples illustrating three cases of the fusion method. SM-speech model and TM- text model.

References

[1] G. Sahu, "Multimodal speech emotion recognition and ambiguity Resolution," arXiv:1904.06022, 2019.

[2] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," Multimed Tools Appl, vol. 80, no. 2, pp. 2887–2905, 2021.

[3] B. T. Atmaja and M. Akagi, "Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM," Speech Commun, vol. 126, no. June 2020, pp. 9–21, 2021.

[4] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," Knowl Based Syst, vol. 229, p. 107316, 2021.

[5] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," Data Min Knowl Discov, vol. 22, no. 1–2, pp. 31–72, 2011.

[6] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion detection in task-oriented spoken dialogues," In. Conf. on Multimedia and Expo. ICME'03, vol. 3, pp.549-552 ,2003.

[7] C. Storm and T. Storm, "A taxonomic study of the vocabulary of emotions," J. Personality Social Psychol., vol. 53, no. 4, pp. 805–816, Oct. 1987.

[8] B. McFee et al., "librosa: Audio and music signal analysis in python," in Proc. Python Sci. Conf, vol. 8, pp. 18–25, 2015.

[9] F. Eyben et al., "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," IEEE Trans Affect Comput, vol. 7, no. 2, pp. 190–202, 2016.

[10] F. Eyben, M. Wöllmer, and B. Schuller, ''Opensmile: the munich versatile and fast open-source audio feature extractor,'' in Proc. 18th ACM Int. Conf. Multimedia, pp. 1459–1462, 2010.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," arXiv:1810.04805, 2018.

[12] Y. Chai, "embedding4bert: A python library for extracting word embeddings from pre-trained language models," GitHub repository, 2020.

[13] C. Xu and X. Geng, "Hierarchical classification based on label distribution learning," AAAI Conf. Artif. Intell., pp. 5533–5540, 2019.

[14] X. Geng, "Label distribution learning," IEEE Trans Knowl Data Eng, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.

[15] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang Resour Eval, vol. 42, no. 4, pp. 335–359, 2008.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proc. of the ACM SIGKDD Int. Conf. on knowledge discovery and data Mining, Association for Computing Machinery, pp. 2623–2631, 2019.

[17] J. Sebastian, P. Pierucci, and T. L. Gmbh, "Fusion techniques for utterance-level emotion recognition combining Speech and Transcripts," Interspeech, pp. 51–55, 2019.

[18] Y. Li, P. Bell, and C. Lai, "Fusing ASR outputs in joint training for speech emotion recognition," in ICASSP, IEEE International Conf. on Acoustics, Speech and Signal Processing., pp. 7362–7366, 2022.

[19] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," Interspeech, pp. 247–251, 2018.