Multitask-based Shared Feature Learning

Subjects: Computer Science, Artificial Intelligence

Contributor: Yiping Ma , Wei Wang

Speech emotion recognition (SER), a rapidly evolving task that aims to recognize the emotion of speakers, has become a key research area in affective computing. Various languages in multilingual natural scenarios extremely challenge the generalization ability of SER, causing the model performance to decrease quickly, and driving researchers to ask how to improve the performance of multilingual SER. To solve this problem, an explainable Multitask-based Shared Feature Learning (MSFL) model is proposed for multilingual SER. The introduction of multi-task learning (MTL) can provide related task information of language recognition for MSFL, improve its generalization in multilingual situations, and further lay the foundation for learning MSFs.

speech emotion recognition multi-task learning feature learning

1. Introduction

Speech emotion recognition (SER), which aims to recognize the emotion of speakers via extracting the acoustic features from the speech signal, started in the 1990s [1] and is a key research area in affective computing. Although modalities such as facial expressions, text, and physiological signals are known to have important and prominent results in affective computing [2][3][4], research reveals that speech modality, as the most convenient and natural medium of communication, has many advantages over the modalities described above [5]: compared with facial expressions, speech is a stronger temporal sequence, and it is easy to identify emotional changes from the entire sequence; compared with text, speech is more expressive in its intonation; compared with physiological signals such as those taken from an electroencephalogram (EEG), it is easy to collect speech data through the lightweight devices. In view of these advantages, SER has made breakthroughs in theoretical methods and key technologies after nearly three decades of development [6][7][8] and is widely used in intelligent vehicles [9], distance education [10], medical treatment [11], media retrieval systems [12] and other fields.

However, emotion implies huge intra-class and inter-class differences, and objective factors such as gender [13], age [14], language [15], and speaker [16] reduce the performance of the existing methods. Since most studies in SER focus on a single language, multilingual SER has been addressed only in a few studies, focusing on the effect of diverse languages on SER. Feraru et al. [17] found that the effect of SER in the same language or within the same language family is higher than that across languages or language families, which means that the model will be limited by the trained corpus due to the single language category and the small sample size of the corpus. This is also the case in real life. Taking SER in the teaching classroom as an example, language courses have lower emotion recognition performance than non-language courses due to the poor generalization of the model in multilingual scenarios. In other words, existing models mainly focus on feature learning within the corpus to improve the performance of SER, without considering the influence of different languages in multiple corpora, far from the SER in real complex situations. A natural question arises: how can researchers improve SER in multilingual scenarios? To answer this question, the connection between different languages and SER is investigated in this research, and the key challenges are summarized as follows: (1) What are the similarities in the feature representation of speech emotion expression in different languages? (2) How can researchers build a generalization model that can improve the performance of SER in multiple languages simultaneously?

To address the above challenges, scholars have attempted to find breakthroughs from the perspective of features and models. For the first issue, researchers have explored fusing multiple features to find the features that are more conducive to multilingual emotional expression, and lead to an improvement. Currently, feature fusion has included the fusion between traditional handcrafted features [18][19][20][21][22][23], the fusion between traditional handcrafted features and deep features [24][25][26][27], and the fusion between deep features [28][29]. In addition, feature selection is also an effective way to obtain optimal acoustic features. Li et al. [30] used a three-layer model consisting of acoustic features, semantic primitives, and emotion dimensions to map multilingual acoustic features to emotion dimensions by using a model inspired by human emotion perception ability, where acoustic features were selected using Fisher discriminant ratio and sequential forward selection methods were used to select features and develop a shared standard acoustic parameter set. For the second issue, the main point is how to control the influence of language as an objective factor on SER. Most of the existing studies have been conducted by training language recognition classifiers for model selection [31][32] or improving the model to enhance the generalization [33][34][35][36][37][38][39]. Although these approaches show good results, there are still some limitations in these efforts. Researchers can summarize as follows: (1) existing methods emphasize controlling the influence of the languages and ignore the intrinsic connection between languages. (2) Most studies either focus on the exploration between multilingual features or on the improvement of models, with little consideration given to studying models and features together.

A more efficient approach alternative to address the above limitations is multi-task learning (MTL), which is inspired by the fact that humans can learn multiple tasks simultaneously and use the knowledge learned in one task to help the learning of the remaining tasks [40]. MTL can learn robust and generalized feature representations from multiple tasks to better enable knowledge sharing between tasks, with the core idea of reducing the risk of overfitting for each task by weighing the training information between tasks. From this, it can be assumed that if the machine learns both language recognition and emotion recognition, will the multilingual shared features learned during the MTL training contribute to the improvement of multilingual SER? The literature proves that this is a feasible method. Lee [41] investigated multilingual SER across English and French using MTL trained with gender recognition and language recognition. Through comparative experiments, he confirmed that the MTL strategy can lead to further improvements under all conditions and is effective for multilingual SER. Zhang et al. [42] proposed a multi-task deep neural network with shared hidden layers and jointly trained several SER tasks from different corpora. This method achieved large-scale data aggregation and obtained feature transformation of all corpora through shared hidden layers. Sharma [43] combined 25 open-source datasets to create a relatively larger multilingual corpus, which shows good performance on a multilingual and multi-task learning SER system based on the multilingual pre-trained wav2vec 2.0 model. In his experiments, several auxiliary tasks were performed, including gender prediction, language prediction, and three regression tasks related to acoustic features. Gerczuk et al. [44] created a novel framework with the concept of residual adapters for multi-corpus SER in a deep transfer learning perspective, where the multi-task transfer experiment of the model trained a shared network for all datasets while only the adapter modules and final classification layers were specific to each dataset. Experiments showed that multi-task transfer experiment led to increased results for 21 of the 26 databases and achieved the best performance. From these studies, it is clear that MTL applied to SER is beneficial for aggregating data, sharing features, and establishing emotional representations. However, previous studies have only applied MTL to improve the generalization ability of models, and have not fully taken into account the interpretability of the model and its generated shared features. In other words, MTL should not only be a method to improve model generalization but is also an effective way to analyze and explain shared features.

To this end, considering the variability of emotional expressions in different languages, researchers propose an explainable Multitask-based Shared Feature Learning (MSFL) model for multilingual SER, which can improve the SER performance of each language and effectively analyze multilingual shared features (MSFs). Based on the basic idea of MTL, the module can be divided into a task-sharing module and a task-specific module, where the task-sharing module is the key component of MSFL, as it undertakes the feature selection and transformation to uncover the generalized high-level discriminative representations. The task-specific module is for the classification of emotion and language tasks. Specifically, the task-sharing module utilizes a long short-term memory network (LSTM) and attention mechanism from a new perspective, where LSTM uses the global feature dimensions as time steps to obtain long-term dependencies of features and the attention mechanism layer enables the model to better understand the important contribution of each feature in MSFs by assigning different weights. The weights of MSFs generated from the attention mechanism are essential to explain the reason for the improved validity and generalizability of the MSFL model and its MSF features.

2. Deep Learning for Speech Emotion Recognition

Early SER techniques relied on extensive feature engineering and performed emotion recognition by traditional machine learning models, such as Hidden Markov Model (HMM), Support Vector Machine (SVM), and Gaussian Mixture Model (GMM) [45]. The flourishing development of deep learning has broadened the representation of acoustic features, and feature extraction is no longer limited to traditional feature engineering. The method of extracting deep representation by using the powerful feature learning ability of deep neural networks has gradually become mainstream, which has also laid the foundation for end-to-end models. SER formally steps into the era of relying on deep learning technology and achieving good performance. Convolution neural networks (CNNs) [46] and recurrent neural networks (RNNs) [47] have become the common deep neural networks in SER. CNNs are designed to process data with a grid-like topology, such as time series and image data, and generally contain convolutional layers, pooling layers, and fully connected layers. Since these can overcome the scalability problem of standard neural networks by allowing the multiple regions of the input to share the same weights [48], they have been widely used in SER to learn the frequency and time domain representations from spectrum images [49]. However, to enhance the interpretability of features, researchers use the traditional handcrafted features, which are always applied in deep neural networks (DNNs) and RNNs. Since DNNs are basic for deep learning, researchers will thus introduce RNNs in detail.

The self-connection property of an RNN has a great advantage in dealing with the temporal sequence problem, but with continuous training, the gradient will disappear and it is difficult to deal with the long temporal sequence, which promotes the proposal of the long short-term memory network [50]. Differing from the RNN, the LSTM adds one cell unit that holds the data for a common time and enables it to call the last calculated value. To protect and control information in the cell state, the LSTM sets up three gate structures including input gate, forgetting gate and output gate. Since the LSTM can both learn the long-term dependence in the data and effectively alleviate the gradient disappearance problem during the training process, frame-level and spectral features generally input LSTM to learn long-term contextual relationships in speech [51]. On this basis, bidirectional long short-term memory (BiLSTM) is proposed to obtain the present and future information in an utterance [52]. To strengthen the capability of capturing the long-time dependency in sequential data, Wang et al. [53] combined BiLSTM with a multi-residual mechanism, where the multi-residual mechanism targets the pattern of the relationship between the current time step and further distant time steps instead of only one previous time step. Additionally, the attention mechanism, which is borrowed from human visual selective attention and was first introduced into SER by Mirsamadi [54], is often combined with LSTM to select the importance of a sentence or some frame segments in the whole sentence on the time series [55].

3. Multi-Task Learning for Speech Emotion Recognition

Multi-task learning (MTL), also known as joint learning, learning to learn, and learning with auxiliary tasks, was proposed by Caruana in 1997 [56]. Its successful applications in natural language processing, computer vision, speech recognition, and other fields demonstrate the irreplaceable advantages of this learning paradigm. It is more conducive to alleviating the data sparsity problem by exploiting shared low-dimensional representations in multiple related tasks. In this way, representations learnt in the MTL scenario become more generalized, which helps improve the performance [57]. However, the premise of applying this method is that all tasks need to correlate. Otherwise, it will generate a negative transfer phenomenon and reduce the inter-task learning effect, so selecting tasks with a strong correlation is crucial for multi-task SER. Based on previous research, the related auxiliary tasks are summarized into four categories. The first is about different emotion representations such as dimensional emotion [58], the second is about the different objective factors related to speech emotion such as gender [59], speaker [60], and language [41], the third is about the different feature representations [61], and the final category is the related task from different databases [42]. Through these tasks, the multi-task SER model can share common feature representations to improve the generalization ability and performance of the model. To establish the link between languages and emotions for multilingual SER, the language recognition is regarded as an auxiliary task in this study.

Thung et al. [62] divide MTL models into single-input multi-output models (SIMO), multi-input multi-output models (MIMO), and multi-input single-output models (MISO). According to the existing literature and the basic situation of SER, multi-task SER can be classified into SIMO and MIMO. SIMO usually takes the traditional handcrafted features [63] or spectrograms [64] as model input, and outputs multiple task targets. MIMO trains the model with multiple sources of data such as multimodal [65], multi-corpus [42], and multi-domain [66] data as inputs, and the probability of one input source predicted as one target is defined as the task. The framework of the two models is shown in **Figure 1**. Generally, during MTL model optimization, the loss function of the model is combined with the sum of weighted loss functions of tasks. Previous studies on multi-task SER have applied an experience-based weight-adjusting method to assign different weights for each task. However, the loss magnitudes of different tasks during training may not be consistent, and this will lead to being dominated by a certain task at a certain stage and being more inclined to fit a certain kind of task. Therefore, scholars have started to balance the task gradients adaptively during the training process to improve the performance of all tasks. Inspired by this, an adaptative loss balancing method called gradient normalization is introduced to improve the performance of the two tasks in the proposed model [67].

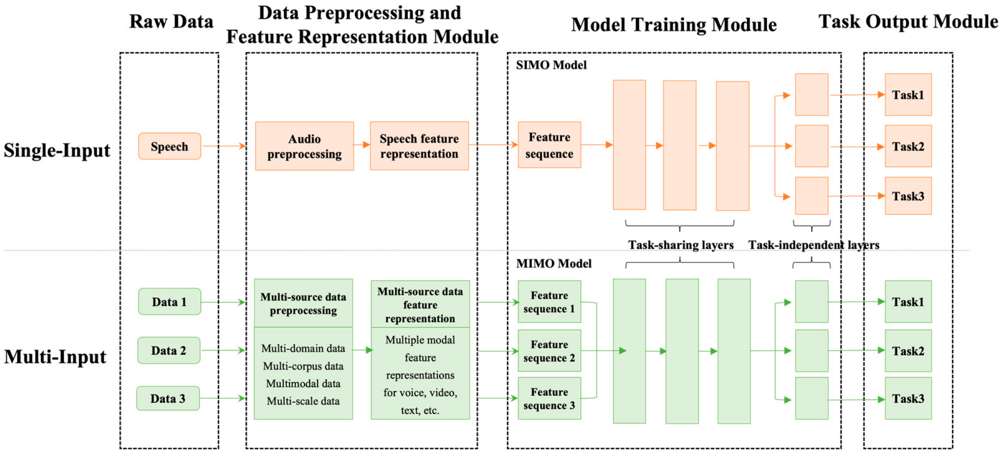

Figure 1. The framework of multi-task SER based on SIMO and MIMO.

References

1. Dellaert, F.; Polzin, T.; Waibel, A. Recognizing Emotion in Speech. In Proceedings of the Fourth International Conference on Spoken Language Processing, ICSLP ’96, Philadelphia, PA, USA, 3–6 October 1996; Volume 3, pp. 1970–1973.

2. Savchenko, A.V.; Savchenko, L.V.; Makarov, I. Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network. IEEE Trans. Affect. Comput. 2022, 13, 2132–2143.

3. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. J. Mach. Learn. Research. 2022, 21, 5485–5551.

4. Zhong, P.; Wang, D.; Miao, C. EEG-Based Emotion Recognition Using Regularized Graph Neural Networks. IEEE Trans. Affect. Comput. 2022, 13, 1290–1301.

5. Li, H.F.; Chen, J.; Ma, L.; Bo, H.J.; Xu, C.; Li, H.W. Dimensional Speech Emotion Recognition Review. Ruan Jian Xue Bao/J. Softw. 2020, 31, 2465–2491. (In Chinese)
6. Kakuba, S.; Poulose, A.; Han, D.S. Attention-Based Multi-Learning Approach for Speech Emotion Recognition with Dilated Convolution. IEEE Access 2022, 10, 122302–122313.
7. Jiang, P.; Xu, X.; Tao, H.; Zhao, L.; Zou, C. Convolutional-Recurrent Neural Networks with Multiple Attention Mechanisms for Speech Emotion Recognition. IEEE Trans. Cogn. Dev. Syst. 2022, 14, 1564–1573.
8. Guo, L.; Wang, L.; Dang, J.; Chng, E.S.; Nakagawa, S. Learning Affective Representations Based on Magnitude and Dynamic Relative Phase Information for Speech Emotion Recognition. Speech Commun. 2022, 136, 118–127.
9. Vögel, H.-J.; Süß, C.; Hubregtsen, T.; Ghaderi, V.; Chadowitz, R.; André, E.; Cummins, N.; Schuller, B.; Härri, J.; Troncy, R.; et al. Emotion-Awareness for Intelligent Vehicle Assistants: A Research Agenda. In Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems, Gothenburg, Sweden, 28 May 2018; pp. 11–15.
10. Tanko, D.; Dogan, S.; Burak Demir, F.; Baygin, M.; Engin Sahin, S.; Tuncer, T. Shoelace Pattern-Based Speech Emotion Recognition of the Lecturers in Distance Education: ShoePat23. Appl. Acoust. 2022, 190, 108637.
11. Huang, K.-Y.; Wu, C.-H.; Su, M.-H.; Kuo, Y.-T. Detecting Unipolar and Bipolar Depressive Disorders from Elicited Speech Responses Using Latent Affective Structure Model. IEEE Trans. Affect. Comput. 2020, 11, 393–404.
12. Merler, M.; Mac, K.-N.C.; Joshi, D.; Nguyen, Q.-B.; Hammer, S.; Kent, J.; Xiong, J.; Do, M.N.; Smith, J.R.; Feris, R.S. Automatic Curation of Sports Highlights Using Multimodal Excitement Features. IEEE Trans. Multimed. 2019, 21, 1147–1160.
13. Vogt, T.; André, E. Improving Automatic Emotion Recognition from Speech via Gender Differentiation. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06); European Language Resources Association (ELRA): Genoa, Italy, 2006.
14. Mill, A.; Allik, J.; Realo, A.; Valk, R. Age-Related Differences in Emotion Recognition Ability: A Cross-Sectional Study. Emotion 2009, 9, 619–630.
15. Latif, S.; Qayyum, A.; Usman, M.; Qadir, J. Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages. In Proceedings of the 2018 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 19 December 2018; pp. 88–93.
16. Ding, N.; Sethu, V.; Epps, J.; Ambikairajah, E. Speaker Variability in Emotion Recognition—An Adaptation Based Approach. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 5101–5104.

17. Feraru, S.M.; Schuller, D.; Schuller, B. Cross-Language Acoustic Emotion Recognition: An Overview and Some Tendencies. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 125–131.
18. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans. Affect. Comput. 2016, 7, 190–202.
19. Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; Narayanan, S.S. The INTERSPEECH 2010 Paralinguistic Challenge. In Proceedings of the Interspeech 2010, ISCA, Chiba, Japan, 26–30 September 2010; pp. 2794–2797.
20. Qadri, S.A.A.; Gunawan, T.S.; Kartiwi, M.; Mansor, H.; Wani, T.M. Speech Emotion Recognition Using Feature Fusion of TEO and MFCC on Multilingual Databases. In Proceedings of the Recent Trends in Mechatronics Towards Industry 4.0; Ab. Nasir, A.F., Ibrahim, A.N., Ishak, I., Mat Yahya, N., Zakaria, M.A., Abdul Majeed, A.P.P., Eds.; Springer: Singapore, 2022; pp. 681–691.
21. Origlia, A.; Galatà, V.; Ludusan, B. Automatic Classification of Emotions via Global and Local Prosodic Features on a Multilingual Emotional Database. In Proceedings of the Fifth International Conference Speech Prosody 2010, Chicago, IL, USA, 10–14 May 2010.
22. Bandela, S.R.; Kumar, T.K. Stressed Speech Emotion Recognition Using Feature Fusion of Teager Energy Operator and MFCC. In Proceedings of the 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 3–5 July 2017; pp. 1–5.
23. Rao, K.S.; Koolagudi, S.G. Robust Emotion Recognition Using Sentence, Word and Syllable Level Prosodic Features. In Robust Emotion Recognition Using Spectral and Prosodic Features; Rao, K.S., Koolagudi, S.G., Eds.; SpringerBriefs in Electrical and Computer Engineering; Springer: New York, NY, USA, 2013; pp. 47–69. ISBN 978-1-4614-6360-3.
24. Araño, K.A.; Gloor, P.; Orsenigo, C.; Vercellis, C. When Old Meets New: Emotion Recognition from Speech Signals. Cogn Comput 2021, 13, 771–783.
25. Wang, C.; Ren, Y.; Zhang, N.; Cui, F.; Luo, S. Speech Emotion Recognition Based on Multi-feature and Multi-lingual Fusion. Multimed. Tools Appl. 2022, 81, 4897–4907.
26. Sun, L.; Chen, J.; Xie, K.; Gu, T. Deep and Shallow Features Fusion Based on Deep Convolutional Neural Network for Speech Emotion Recognition. Int. J. Speech Technol. 2018, 21, 931–940.
27. Yao, Z.; Wang, Z.; Liu, W.; Liu, Y.; Pan, J. Speech Emotion Recognition Using Fusion of Three Multi-Task Learning-Based Classifiers: HSF-DNN, MS-CNN and LLD-RNN. Speech Commun.

2020, 120, 11–19.

28. Al-onazi, B.B.; Nauman, M.A.; Jahangir, R.; Malik, M.M.; Alkhammash, E.H.; Elshewey, A.M. Transformer-Based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion. Appl. Sci. 2022, 12, 9188.
29. Issa, D.; Fatih Demirci, M.; Yazici, A. Speech Emotion Recognition with Deep Convolutional Neural Networks. Biomed. Signal Process. Control. 2020, 59, 101894.
30. Li, X.; Akagi, M. Improving Multilingual Speech Emotion Recognition by Combining Acoustic Features in a Three-Layer Model. Speech Commun. 2019, 110, 1–12.
31. Heracleous, P.; Yoneyama, A. A Comprehensive Study on Bilingual and Multilingual Speech Emotion Recognition Using a Two-Pass Classification Scheme. PLoS ONE 2019, 14, e0220386.
32. Sagha, H.; Matějka, P.; Gavryukova, M.; Povolny, F.; Marchi, E.; Schuller, B. Enhancing Multilingual Recognition of Emotion in Speech by Language Identification. In Proceedings of the Interspeech 2016, ISCA, San Francisco, CA, USA, 8 September 2016; pp. 2949–2953.
33. Bertero, D.; Kampman, O.; Fung, P. Towards Universal End-to-End Affect Recognition from Multilingual Speech by ConvNets. arXiv 2019, arXiv:1901.06486.
34. Neumann, M.; Thang Vu, N. goc Cross-Lingual and Multilingual Speech Emotion Recognition on English and French. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5769–5773.
35. Zehra, W.; Javed, A.R.; Jalil, Z.; Khan, H.U.; Gadekallu, T.R. Cross Corpus Multi-Lingual Speech Emotion Recognition Using Ensemble Learning. Complex Intell. Syst. 2021, 7, 1845–1854.
36. Sultana, S.; Iqbal, M.Z.; Selim, M.R.; Rashid, M.M.; Rahman, M.S. Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks. IEEE Access 2022, 10, 564–578.
37. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Schuller, B.W. Self Supervised Adversarial Domain Adaptation for Cross-Corpus and Cross-Language Speech Emotion Recognition. IEEE Trans. Affect. Comput. 2022.
38. Tamulevičius, G.; Korvel, G.; Yayak, A.B.; Treigys, P.; Bernatavičienė, J.; Kostek, B. A Study of Cross-Linguistic Speech Emotion Recognition Based on 2D Feature Spaces. Electronics 2020, 9, 1725.
39. Fu, C.; Dissanayake, T.; Hosoda, K.; Maekawa, T.; Ishiguro, H. Similarity of Speech Emotion in Different Languages Revealed by a Neural Network with Attention. In Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 3–5 February 2020; pp. 381–386.

40. Caruana, R. Multitask Learning. Mach. Learn. 1997, 28, 41–75.
41. Lee, S. The Generalization Effect for Multilingual Speech Emotion Recognition across Heterogeneous Languages. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5881–5885.
42. Zhang, Y.; Liu, Y.; Weninger, F.; Schuller, B. Multi-Task Deep Neural Network with Shared Hidden Layers: Breaking down the Wall between Emotion Representations. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4990–4994.
43. Sharma, M. Multi-Lingual Multi-Task Speech Emotion Recognition Using Wav2vec 2.0. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6907–6911.
44. Gerczuk, M.; Amiriparian, S.; Ottl, S.; Schuller, B.W. EmoNet: A Transfer Learning Framework for Multi-Corpus Speech Emotion Recognition. IEEE Trans. Affect. Comput. 2021.
45. Akçay, M.B.; Oğuz, K. Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers. Speech Commun. 2020, 116, 56–76.
46. Wang, W.; Cao, X.; Li, H.; Shen, L.; Feng, Y.; Watters, P. Improving Speech Emotion Recognition Based on Acoustic Words Emotion Dictionary. Nat. Lang. Eng. 2020, 27, 747–761.
47. Hsu, J.-H.; Su, M.-H.; Wu, C.-H.; Chen, Y.-H. Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations. IEEE/ACM Trans. Audio Speech Lang. Process. 2021, 29, 1675–1686.
48. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Epps, J. Direct Modelling of Speech Emotion from Raw Speech. In Proceedings of the Interspeech 2019, ISCA, Graz, Austria, 15 September 2019; pp. 3920–3924.
49. Wu, X.; Cao, Y.; Lu, H.; Liu, S.; Wang, D.; Wu, Z.; Liu, X.; Meng, H. Speech Emotion Recognition Using Sequential Capsule Networks. IEEE/ACM Trans. Audio Speech Lang. Process. 2021, 29, 3280–3291.
50. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. Neural Comput. 1997, 9, 1735–1780.
51. Wang, J.; Xue, M.; Culhane, R.; Diao, E.; Ding, J.; Tarokh, V. Speech Emotion Recognition with Dual-Sequence LSTM Architecture. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6474–6478.

52. Graves, A.; Jaitly, N.; Mohamed, A. Hybrid Speech Recognition with Deep Bidirectional LSTM. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013; pp. 273–278.

53. Wang, Y.; Zhang, X.; Lu, M.; Wang, H.; Choe, Y. Attention Augmentation with Multi-Residual in Bidirectional LSTM. Neurocomputing 2020, 385, 340–347.

54. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.

55. Hu, D.; Wei, L.; Huai, X. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7042–7052.

56. Zhang, Y.; Yang, Q. An Overview of Multi-Task Learning. Natl. Sci. Rev. 2018, 5, 30–43.

57. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. Survey of Deep Representation Learning for Speech Emotion Recognition. IEEE Trans. Affect. Comput. 2021.

58. Zhang, Z.; Wu, B.; Schuller, B. Attention-Augmented End-to-End Multi-Task Learning for Emotion Prediction from Speech. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6705–6709.

59. Li, Y.; Zhao, T.; Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In Proceedings of the Interspeech 2019, ISCA, Graz, Austria, 15 September 2019; pp. 2803–2807.

60. Fu, C.; Liu, C.; Ishi, C.T.; Ishiguro, H. An End-to-End Multitask Learning Model to Improve Speech Emotion Recognition. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Virtual, 18–21 January 2021; pp. 1–5.

61. Li, X.; Lu, G.; Yan, J.; Zhang, Z. A Multi-Scale Multi-Task Learning Model for Continuous Dimensional Emotion Recognition from Audio. Electronics 2022, 11, 417.

62. Thung, K.-H.; Wee, C.-Y. A Brief Review on Multi-Task Learning. Multimed Tools Appl 2018, 77, 29705–29725.

63. Xia, R.; Liu, Y. A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space. IEEE Trans. Affect. Comput. 2017, 8, 3–14.

64. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Epps, J.; Schuller, B.W. Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition. IEEE Trans. Affect. Comput. 2022, 13, 992–1004.

65. Atmaja, B.T.; Akagi, M. Dimensional Speech Emotion Recognition from Speech Features and Word Embeddings by Using Multitask Learning. APSIPA Trans. Signal Inf. Process. 2020, 9, e17.

66. Kim, J.-W.; Park, H. Multi-Task Learning for Improved Recognition of Multiple Types of Acoustic Information. IEICE Trans. Inf. Syst. 2021, E104.D, 1762–1765.

67. Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; Rabinovich, A. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 3 July 2018; pp. 794–803.

Retrieved from https://encyclopedia.pub/entry/history/show/89865