Review of Research on Speech Emotion Recognition

Yali Yang(✉) and Fangyuan Xu

School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China
17803428680@163.com

Abstract. Language is an effective way to express human emotions, but emotions are difficult to describe and judge with computers, so it is an important task to analyze their emotions through speech. We summarized the current situation of speech emotion recognition from five aspects: the development of speech emotion recognition, emotion description model, emotion speech database, feature extraction, and emotion recognition algorithm. By summarizing and analyzing these five aspects, we can predict the future development trend of speech emotion recognition and improve recognition accuracy combined with other information for joint analysis. In this paper, we summarized the commonly used emotion database, compared the popular attention mechanism in speech emotion recognition, and finally, proposed the prospect of speech emotion recognition.

Keywords: Speech emotion recognition · Recognition algorithm · Attention mechanism

1 Introduction

The expression of human emotions is carried out in various ways. The speaker's emotional state can be effectively understood based on the speaker's reaction, behavior, body posture, and physiological signals. However, it is very difficult to interpret the speaker's emotional state from the speech in a non-contact way [1]. Speech emotion recognition, one of an important direction in the speech research, is to judge the emotional state that the speaker wants to express based on the person's speech. As an important branch of artificial intelligence, speech emotion recognition has been widely used in the diagnosis and evaluation of depression or some mental diseases, and in analyzing the emotions of students so as to make course contents adjustment in distance learning and online learning [2]. The task of speech emotion recognition is to find the features related to emotion from the speaker's speech, and then judge the speaker's emotion through these features [3].

There are currently two models describing speech emotions: discrete and dimensional emotion models. The former describes emotions as discrete emotions, such as happiness and anger, which are widely used in the research of speech emotion recognition. The latter describes the emotional state as points in a multi-dimensional emotional

X. Jiang (Ed.): MLICOM 2021, LNICST 438, pp. 315–326, 2022.
https://doi.org/10.1007/978-3-031-04409-0_29

space, which is actually a Cartesian space with each dimension. Each dimension represent an emotional feature. The dimensional emotional database is relatively less than the discrete database. The current ones are mainly VAM, Semaine, etc. [4]. With the research on speech emotion recognition, it has made significant progress in two aspects: one is the combination of speech emotion recognition and deep learning algorithms for feature extraction, emotion classification and regression that make the recognition efficiency higher. The other is a variety of methods are integrated and the more and more complex networks are used in the process of emotion recognition.

This paper will describe the current technology and progress from five aspects: the development of speech emotion recognition, emotion description models, emotion speech databases, feature extraction, and emotion recognition algorithms. In this paper we mainly discussed the research progress of feature extraction, recognition and classification algorithms in the process of speech emotion recognition, summarized the limitations of speech emotion recognition in the context of deep learning algorithms, and finally proposed the future prospects of speech emotion recognition.

2 The Development of Speech Emotion Recognition

The earliest speech emotion recognition originated in the mid-1980s. Acoustic statistical features was used for the first time in emotion analysis and obvious speech differences was found under different emotions. In 1987, Professor Minsky proposed the idea of "making computers have emotional capabilities" [5], and consequently more researchers are engaged in the research of speech emotion recognition. In the early 1990s, the MIT laboratory constructed an "emotion editor" [6] that used the collection and analysis of various emotional signals from the outside world to identify various emotions, and then proposed the viewpoint of affective computing in 1995. At the end of the ninth decade in the twentieth century, Moriyama applied the voice interface of a graphics acquisition system to an e-commerce system [7], making speech emotion recognition commercially available. In this period, speech emotion recognition is still in its infancy stage, mainly analyze acoustic features. At this time, emotion recognition infrastructure was incomplete, the sample data was relatively simple, and no generalized method was established [8].

Speech emotion recognition has been developed at a high speed in the 21st century under the background of the rapid development of computers and deep learning algorithms. In 2000, the "Speech and Emotion" symposium was held in Ireland. For the first time, scholars of speech emotion recognition research were gathered together to summarize the previous problems of speech emotion recognition, and proposed the future development direction and goals of speech emotion recognition. And subsequently some important speech emotion recognition special issues were opened so that it attracted worldwide attention and entered a golden period of development. At the same time, many research institutions and universities started this research, such as the multimedia laboratory led by Professor Pricard of Massachusetts Institute of Technology, and the Institute of Human-Computer Communication at the Technical University of Munich in Germany. Speech emotion recognition in China started at the beginning of this century. In 2001, Professor Zhao Li of Southeast University was the first person to study on

speech emotion information. Chinese Academy of Sciences host the first Chinese Affective Computing and Intelligent Interaction Academic Conference in 2003, and the first International Affective Computing and Intelligent Interaction Academic Conference was held in 2005. Southeast University, Tsinghua University, Zhejiang University laboratories have made more contributions to speech emotion recognition in the past ten years. The Institute of Automation, Chinese Academy of Sciences completed the recording of the CASIA Chinese emotion database. In the development of speech emotion recognition, Support Vector Machine (SVM) in traditional machine learning methods is relatively basic for emotion recognition. For example, Zhang [9] compared several SVMs' recognition efficiency based on different kernel functions, used genetic algorithm and particle swarm algorithm to optimize the parameters and achieved better results. With the development of deep learning algorithms and the rise of attention mechanisms, Wang improved the efficiency and accuracy of recognition through speech emotion recognition models based on CNN, BLSTM, attention mechanisms and the Adam optimizer. Bao [10] combined CNN with BLSTM to extract the spatiotemporal features of speech emotion, and then improved the recognition performance by using the attention mechanism. At this stage, speech emotion recognition has been largely developed, the efficiency of recognition significantly improved, and the accuracy of discrete speech emotion recognition obviously increased, and the speech database greatly enriched. However, there are some emotions such as irony and sarcasm still do not have a standard emotional set method and publicly recognized data, and some complex emotions cannot be identified.

3 Emotional Description

3.1 Model

Emotion recognition description models are divided into two types: discrete and dimensional. Discrete emotion recognition models, which is widely used, include basic and relatively independent emotions from all human emotions, such as anger, fear, happiness, etc. Table 1 list the division of basic emotions by different scholars.

Table 1. Basic emotions by different scholars [11].

Scholars	Sentiment classification
Ekman, Friesen [12]	Joy, sadness, anger, fear, sorrow, disgust
Arnold	Anger, aversion, courage, dejection, despair, dear, hate, hope, love, sadness
Gray	Desire, happiness, interest, surprise, wonder, sorrow
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Carroll Izard	Interest, Joy, Surprise, Sadness, Anger, Disgust, Contempt, Self-Hostility, Fear, Shame, Shyness, Guilt

The dimensional emotion recognition model believes that emotions are depend of each other. There is a smooth transition from one emotion to another. Each emotion corresponds to a point in the dimensional space. An emotional state can be expressed in multiple dimensions [4]. Compared with discrete emotions, the dimensional emotion recognition model is more detail, and can more precisely represent people's emotions in real life. Evaluation and activation are the two main dimensions that describe the main aspects of emotions. For some dimensional emotions [13], we even can't recognize what emotions they are. Nowadays, common dimensional models include two-dimensional, three-dimensional [14] and four-dimensional emotional models. The figure below is a two-dimensional dimension emotion recognition model (Fig. 1).

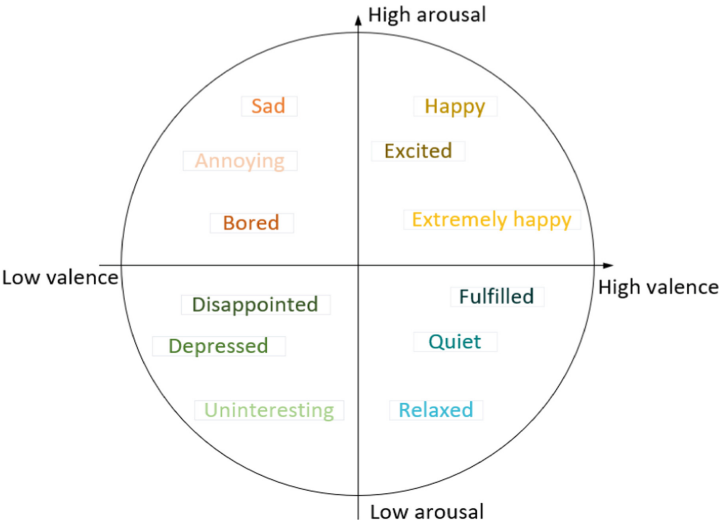

Fig. 1. Two-dimensional emotion recognition model.

3.2 Emotional Speech Database

As the performance of emotion recognition is also closely related to the quality of the database, emotional speech database is the key to speech emotion recognition. Emotional speech database is divided into discrete and dimensional, and then divided into natural type, performance type and guide type according to the classification of emotion generation. Speech recording is generally carried out in the laboratory in order to ensure the quality of speech. Due to differences in language and expression of emotions, there is no standard database that has been accepted by all researchers. However, speech emotion databases have become more and more perfect.

(1) Belfast data set [15]
This emotion database is an English database [16] which was recorded by 40 recorders, consist of are 20 men and 20 women, interpreted 35–40 sentences. This

database contains 5 basic discrete emotions, including anger, sadness, happiness, fear and neutrality.

(2) EMO-DB data set [17]
The EMO-DB data set was recorded by the Technical University of Berlin. This data set is interpreted in German. It consists of 5 long and 5 short sentences expressed by five men and five women with 7 different emotions, a total of 800 sentences. The database contains 7 emotions: neutral, angry, scared, happy, sad, disgusted and bored. This data set was recorded in a recording studio, with a high degree of emotional freedom, a strong sense of reality, and no specific emotional preference.

(3) IEMOCAP data set
The IEMOCAP data set is collected by the Sail Lab of the University of Southern California. It is an interactive emotional binary motion capture data set. It contains data of ten actors and actresses in improvisation or scripted scenes. There are videos, voice, facial motion capture and text transcription. The dialogue of this data set is divided into two parts, one is scripted dialogue, and the other is free to play in specific situations. There are 7433 sentences in total and six emotions are labeled, including neutrality, happiness, sadness, anger, surprise and excitement as well as dimension labels, such as activation, dominance and so on. This data set is often used for multi-modal emotional research due to its high quality and multi-modal information.

(4) CASIA data set
The CASIA data set is a Chinese data set recorded by the Institute of Automation, Chinese Academy of Sciences. It contains six different emotions: angry, happy, scared, sad, surprised and neutral. A database of 9600 sentences was recorded by four professional speakers. It contains 300 sentences with the same text and different sentiment, and 100 sentences with different text that can better express the sentiment features.

(5) VAM data set [18]
The VAM data set is a typical open dimensional data set. This data set is a purely natural and unrestricted voice communication database. It is obtained from a German interview program and contains three parts: a speech warehouse, a video warehouse and an expression warehouse. There are a total of 1018 sentences, expressing the emotions Valence, Activation and Dominance in three dimensions, and the label value is between -1 and 1. This data set expresses more negative emotions.

(6) SEMAINE data set [19]
The SEMAINE data set is a multi-modal dialogue material. It is a typical dimensional data set that consists of a dialogue between four robots with fixed emotions and 20 users. This data set is recorded in a professional recording studio and consists of four emotional dimensions, namely Valence, Arousal, Expectancy, and Power. The first three dimensions are continuous values between -1 and 1, and Power is a continuous value greater than or equal to 0. Part of the data in this dataset is used in AVEC2012 competition.

4 Feature Extraction

4.1 Prosodic Features

In the traditional way of prosody expression [20], it is generally defined by linguists. It has stress [21], pitch, rhythm, etc., which can help listeners to better understand speech. Among the most commonly used prosodic features [22] are duration, baseband [23], energy, etc. Fundamental frequency features include a large number of features that characterize speech emotions, which are very important for speech emotion recognition. The pitch period is the reciprocal of the vocal cord vibration frequency, and the period of vocal cord vibration is actually the pitch period. The fundamental frequency has a large range of changes, so it is difficult for researchers to detect. The commonly used methods for extracting fundamental frequency features include autocorrelation function method [24], average amplitude difference method and wavelet method [25]. The first two methods are aimed at the time domain part of speech, and the wavelet method is mainly frequency domain part of speech. The amplitude energy of sound signals of different emotions is different. The emotional amplitude energy of surprise and happiness are increased, while the emotional energy of sadness is decreased. Prosody does not affect the content of the speech [26], but it affects the true meaning of the content. The same sentence has different meanings for different prosodic structures. Nowadays, the research of speech emotion recognition uses more prosodic features as an auxiliary reference.

4.2 Based on Spectral Correlation Features

With the popularity and application of deep learning, spectrum widely used in speech emotion recognition include spectrograms and other variants [27]. Spectrograms can reflect different emotions that usually the abscissa represents time and the ordinate represents frequency. The coordinate point value represents the energy of the speech data, and the size of the energy value is generally expressed by color. The darker the color, the stronger the speech energy of the point. Commonly used features are the Mel frequency cepstral coefficient, which is displayed in the Mel label [28]. For the cepstrum parameters extracted from the frequency domain, the Mel scale describes the non-linear characteristics [29] of the human ear frequency. Some people used Gabor wavelet to combine gray-level co-occurrence matrix, Tamura method with LBP method to extract features of spectrogram to improve the accuracy of feature extraction. With the development of deep learning algorithms, traditional feature extraction methods sometimes have no superiority. Other built a feature extraction method for the context of the spectrum sequence. After the cepstral coefficients were obtained by the discrete cosine transform, the context processing was carried out. And the cepstrum coefficients [30] are combined with the feature coefficients obtained from the context processing to obtain the spectral series context features. This extraction method significantly improves the performance of feature extraction and reduces the influence of many external factors in feature extraction.

5 Recognition Algorithm

5.1 Classic Machine Learning Algorithms

The Hidden Markov Model (HMM) [31] algorithm was proposed in the 1970s and has become a common method in signal processing. It is a statistical analysis signal model in speech emotion recognition, mainly regarding the cumulative probability of speech. In the course of training and recognition using the HMM algorithm, each emotion corresponds to an HMM model [32] whose parameter is obtained from the emotion samples. Although the algorithm has good system scalability, its shortcomings involved in unqualified classification and decision-making ability, inaccurate recognition in similar emotions, and poor robustness. In view of the above-mentioned shortcomings of the HMM algorithm, the researchers combined the HMM with strong time modeling ability and the ANN algorithm with strong classification learning ability to improve the robustness. Some Combine RBF and HMM algorithms to form an emotion recognition model to improve recognition efficiency.

Gaussian Mixed Model (GMM) [33] is one of the common speech emotion recognition models, which is used to represent the acoustic characteristics of sound units. It describes the emotional feature parameters through the linear weighted superposition of Gaussian probability density functions. The calculation [34] amount of the GMM algorithm is much smaller than that of the HMM, so the robustness of the GMM algorithm is better than that of the HMM [35]. In the improved GMM speech emotion recognition model, an optimal parameter set was found and then a final GMM model was obtained by iteration, which verifies that the improved GMM recognition model has better performance and higher recognition rate than the traditional GMM model.

Support vector machine (SVM) [36] is a classifier developed from the generalized portrait algorithm for pattern recognition in 1964. The traditional SVM is mainly used to solve binary classification problems. SVM seeks an optimal classification hyperplane as much as possible in solving linear classification problems [37]. In order to convert nonlinear problems into linear problems, a kernel function was introduced in solving nonlinear classification problems. For the nonlinear SVM, it is very important to select the kernel function, due to the choice of kernel function will affect the classification effect in dealing with nonlinear problems. On the speech emotion recognition technology based on feature selection and decision tree combined with SVM, recognition efficiency of genetic algorithm and particle swarm are compared in parameters selection. It is concluded that the performance of the genetic algorithm is better than particle swarm algorithm, and the recognition efficiency of the decision tree combined with SVM in feature selection strategy is higher than that of the traditional SVM algorithm.

5.2 Deep Learning Algorithm

Convolutional Neural Network (CNN) [38] is a type of feedforward neural network that includes convolutional calculations. The algorithm, began in the 1980s and 1990s, includes an input layer, a hidden layer, and an output layer [39]. The hidden layer includes a convolutional layer, a pool Convolution layer and fully connected layer. The convolution layer, the core of CNN, is mainly used to perform convolution operations on

data through convolution kernels. The other improved CNN by using multi-level residual convolution. The neural network model, which contains multiple convolutional pooling layers and a multi-level residual structure, further reduces the amount of calculation, thereby improving the efficiency of recognition.

The long and short-term memory network is the most commonly used network for speech emotion recognition. The LSTM [40] network is a special network structure with three gate mechanisms: forget gate, input gate and output gate. This network can effectively determine information that should be forgotten or retained, which solves the learning of long-distance information. The structure in the hidden layer in the recurrent neural network is replaced with a long and short-term memory module. The idea of self-loop is introduced in LSTM. The weight value of self-loop is determined according to the context. And the gates control the weight of self-loop that change dynamically in accumulated time. Since LSTM can only achieve one-way transmission, when the sentence sequence is changed, the key part sentence appears in the back of network and thus make LSTM inapplicable. So a bidirectional long-short term memory network (BiLSTM) [40] was proposed, which can be regarded as an acyclic graph composed of two unidirectional LSTM networks [41], and output results are obtained under consideration of the relevant influencing factors in the front and back parts of network. The recognition rate of BiLSTM is much higher than the LSTM, whereas the latency of BiLSTM is relatively longer [42]. For example, researchers compared the unidirectional to the bidirectional long-short term, and drew a conclusion the BiLSTM can learn the front and back information of the current speech frame.

5.3 Attention Mechanism

The attention mechanism [43] is originated from human vision. When processing information, humans focus on a part of all information while ignoring other unimportant visible parts. It was originally used for machine translation and has become an important concept in neural network. The common attention mechanisms [44] are Soft Attention, Local Attention and Hard Attention according to the calculation area of attention.

Soft Attention is widely used in natural language processing and speech recognition. This mechanism calculates the weight probability of all keys, and then input them into the next layer. Soft attention covers the whole range of the network because each key has its own weight. This mechanism can be embedded in a network for training and is so rational that can be used as a comparison standard for other attention mechanisms to measure the recognition efficiency.

Hard Attention is a mechanism relative to Soft Attention. It attaches importance to part of the information, while the Soft Attention calculates all keys. The mechanism sample is a random process that calculate part of hidden layers of the encoder according to the weight probability of keys and accurately locate a certain key, which means the probability of the selected part is 1 and all other parts are 0. This method has high requirements for precise positioning, and thus rarely used in research.

Local Attention is a compromise and combined mechanism of the first two mechanisms. The mechanism focuses on a part of the decoder hidden states and selectively calculates a small window area, which can avoid the same computational overhead as Soft Attention and train easier than Hard Attention. The accuracy of the mechanism depends

mainly on the accuracy of the position area. Overall, the Soft Attention mechanism has a wider range of applications than other two attentions.

The attention mechanism can be divided into General Attention and Self Attention according to the category of the information used. The mechanism of General Attention uses external information, which is suitable for the construction of two-paragraph text relationship, while Self Attention only uses internal information, key equals value and value equals query. Self Attention, proposed by Google in the transformer model, occurred between the internal elements of the input or the output. Self Attention mechanism can be better applied to long-distance dependent learning, because it easily capture the long-distance interdependence features in sentences. Researchers combined the local attention advantages of CNN, RNN and Self Attention mechanism in the ACRNN recognition model, and enabled the model to learn the most emotional features. As a result, the efficiency of ACRNN is about 0.5% higher than CRNN and then the effectiveness of the attention mechanism was proved in speech emotion recognition.

According to hierarchical relationships in the structure, the attention mechanism can also be defined single-layer attention mechanism, multi-layer attention and multi-head attention. Multi-head-attention performs linear transformations on query, key and value, and then performs and splices many times of scaling dot product attention, and finally perform a linear transformation. At this time the value obtained is the result of the multi-head attention mechanism. Some people used the multi-head attention mechanism to construct a module to combine the audio module with the video module. And finally the emotional prediction value was obtained through linear transformation.

With the wide application of attention mechanism, more and more types of attention mechanism emerged. Researchers proposed Component Attention and Monotonic Attention in speech emotion recognition research, compared the recognition efficiency of different attention mechanism, showed that Monotonic Attention is better than other attention mechanisms on the IEMOCAP data set, and thus furtherly proved that the attention mechanism play an indispensable role in the current speech emotion recognition.

6 Conclusion

From the application of classic machine learning algorithms to the improved various deep learning speech emotion recognition models, the current research on speech emotion recognition has made great achievements through decades of development. Furthermore, the addition of attention mechanism into speech emotion recognition made its the efficiency furtherly improved. This article describes five aspects of speech emotion recognition: the development of speech emotion recognition, emotion description models, emotion speech databases, feature extraction, and emotion recognition algorithms. Due to the high complexity of human emotion, the research of speech emotion still faces many problems. Firstly, speech emotion recognition lacks a widely recognized database. It is difficult to collect the data and organize the database as the result of complexity of emotions. Besides, most of the existing databases are recorded in the laboratory, so it is very difficult to label emotions in dimensional databases. However, the speech database is the basis of research and an inevitable requirement; Secondly, the development of emotion recognition tend to the combination of multimodality, such as speech,

expression and text are combined to recognize the emotion, which makes the recognition efficiency higher than each one of them; Finally, different languages and cultures lead to different ways of emotional expression. Therefore, different models were built due to different speakers and language differences in the training data set in the training process of the speech emotion recognition model. In addition, there is untargeted research on some complex human emotions. Above all, there are broad prospects in speech emotion recognition.

References

1. Swain, M., Routray, A., Kabisatpathy, P.: Databases, features and classifiers for speech emotion recognition: a review. Int. J. Speech Technol. **21**(1), 93–120 (2018). https://doi.org/10.1007/s10772-018-9491-z
2. Weninger, F., Wöllmer, M., Schuller, B.: Emotion Recognition in Naturalistic Speech and Language—a Survey. In Emotion Recognition, pp. 237–267. Wiley (2015). https://doi.org/10.1002/9781118910566.ch10
3. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011). https://doi.org/10.1016/j.patcog.2010.09.020
4. Lugger, M., Yang, B..: Psychological Motivated Multi-Stage Emotion Classification Exploiting Voice Quality Features. In tech Open (2008). https://doi.org/10.5772/6383
5. Minsky, M.: The society of mind. Person. Forum **3**(1), 19–32 (1987)
6. Cahn, J.: Generation of affect in synthesized speech. J. Am. Voice I/O Soc. **8** (2000)
7. Moriyama, T., Ozawa, S.: Emotion recognition and synthesis system on speech. In: Proceedings IEEE International Conference on Multimedia Computing and Systems, vol. 1, pp. 840–844 (1999). https://doi.org/10.1109/MMCS.1999.779310
8. Williams, C.E., Stevens, K.N.: Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. **52**(4B), 1238–1250 (1972). https://doi.org/10.1121/1.1913238
9. Zhang, L.: Speech emotion recognition algorithm based on modified SVM–Journal of Computer Applications. Accessed 03 Sept 2021. https://en.cnki.com.cn/Article_en/CJFDTotal-JSJY201307039.htm
10. Zhao, Z., Bao, Z., Zhang, Z., Cummins, N., Wang, H., Schuller, B.: Attention-Enhanced Connectionist Temporal Classification for Discrete Speech Emotion Recognition, pp. 206–210 (2019). https://doi.org/10.21437/Interspeech.2019-1649
11. Ortony, A., Turner, T.J.: What's basic about basic emotions?. Psychol. Rev. **97**(3), 315–331 (1990). https://doi.org/10.1037/0033-295X.97.3.315
12. Varghese, A.A., Cherian, J.P., Kizhakkethottam, J.J.: Overview on emotion recognition system. In: 2015 International Conference on Soft-Computing and Networks Security (ICSNS), pp. 1–5 (2015). https://doi.org/10.1109/ICSNS.2015.7292443
13. Borchert, M., Dusterhoft, A.: Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In: 2005 International Conference on Natural Language Processing and Knowledge Engineering, pp. 147–151 (2005)
14. Cowie, R., et al.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. **18**(1), 32–80 (2001)
15. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E.: FEELTRACE: an instrument for recording perceived emotion in real time, undefined (2000). Accessed 05 Sept 2021. https://www.semanticscholar.org/paper/FEELTRACE%3A-an-instrument-for-recording-perceived-in-Cowie-Douglas-Cowie/5b35fe6950db00ad32f6af8ad0162028e15c2f27

16. McGilloway, S., Cowie, R., ED, C., Gielen, S., Westerdijk, M., Stroeve, S.: Approaching automatic recognition of emotion from voice: a rough benchmark (2000)
17. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech, p. 4 (2005)
18. Grimm, M., Kroschel, K., Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database, pp. 865–868 (2008)
19. Mckeown, G., Valstar, M., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions, pp. 1079–1084 (2010)
20. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog (2002)
21. Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E.: desperately seeking emotions: actors, wizards, and human beings. In: Proceedings of the ISCA Workshop on Speech and Emotion, pp. 195–200 (2000)
22. Beeke, S., Wilkinson, R., Maxim, J.: Prosody as a compensatory strategy in the conversations of people with agrammatism. Clin. Linguist Phon. **23**(2), 133–155 (2009). https://doi.org/10.1080/02699200802602985
23. Busso, C., Lee, S., Narayanan, S.S.: Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Audio Speech Lang. Process. **17**, 582–596 (2009). https://doi.org/10.1109/TASL.2008.2009578
24. Lee, C.M., Narayanan, S.S., Pieraccini, R.: Combining acoustic and language information for emotion recognition (2002)
25. Lee, C., et al.: Emotion Recognition based on Phoneme Classes, presented at the Proceedings of ICSLP (2004). https://doi.org/10.21437/Interspeech.2004-322
26. Leinonen, L., Hiltunen, T., Linnankoski, I., Laakso, M.-L.: Expression of emotional–motivational connotations with a one-word utterance. J. Acoust. Soc. Am. **102**(3), 1853–1863 (1997). https://doi.org/10.1121/1.420109
27. Ramdinmawii, E., Mohanta, A., Mittal, V.K.: Emotion recognition from speech signal. In: TENCON 2017 - 2017 IEEE Region 10 Conference, pp. 1562–1567 (2017). https://doi.org/10.1109/TENCON.2017.8228105
28. Bou-Ghazale, S.E., Hansen, J.H.L.: A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Trans. Speech Audio Process. **8**(4), 429–442 (2000)
29. Bitouk, D., Verma, R., Nenkova, A.: Class-level spectral features for emotion recognition. Speech Commun. **52**(7–8), 613–625 (2010). https://doi.org/10.1016/j.specom.2010.02.010
30. Chauhan, R., Yadav, J., Koolagudi, S.G., Rao, K.S.: Text independent emotion recognition using spectral features. In: Aluru, S., et al. (eds.) IC3 2011. CCIS, vol. 168, pp. 359–370. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22606-9_37
31. Cairns, D.A., Hansen, J.H.L.: Nonlinear analysis and classification of speech under stressed conditions. J. Acoust. Soc. Am. **96**(6), 3392–3400 (1994). https://doi.org/10.1121/1.410601
32. Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., Cox, C.: ASR for emotional speech: clarifying the issues and enhancing performance. Neural Networks **18**(4), 437–444 (2005). https://doi.org/10.1016/j.neunet.2005.03.008
33. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974). https://doi.org/10.1109/TAC.1974.1100705
34. Yuan, G., Lim, T.S., Juan, W.K., Ringo, H.M.-H., Li, Q.: A GMM based 2-stage architecture for multi-subject emotion recognition using physiological responses. In: Proceedings of the 1st Augmented Human International Conference, pp. 1–6. New York, NY, USA (2010). https://doi.org/10.1145/1785455.1785458
35. Tang, H., Chu, S.M., Hasegawa-Johnson, M., Huang, T.S.: Emotion recognition from speech VIA boosted Gaussian mixture models. In: 2009 IEEE International Conference on Multimedia and Expo, pp. 294–297 (2009). https://doi.org/10.1109/ICME.2009.5202493

36. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. **70**(3), 614–636 (1996). https://doi.org/10.1037/0022-3514.70.3.614
37. Burges, C.J.C.: a tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 121–167 (1998). https://doi.org/10.1023/A:1009715923555
38. Aldeneh, Z., Provost, E.M.: Using regional saliency for speech emotion recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2741–2745 (2017)
39. Fayek, H.M., Lech, M., Cavedon, L.: Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks **92**, 60–68 (2017). https://doi.org/10.1016/j.neunet.2017.02.013
40. Tzirakis, P., Zhang, J., Schuller, B.W.: End-to-End speech emotion recognition using deep neural networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5089–5093 (2018)
41. Tzinis, E., Potamianos, A.: Segment-based speech emotion recognition using recurrent neural networks. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 190–195 (2017)
42. Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231 (2017)
43. Bahdanau, D., Cho, K., Bengio, Y.: eural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs, stat] (2016). Accessed 02 Sept 2021. http://arxiv.org/abs/1409.0473
44. Luong, T., Pham, H., Manning, C.D.: Fective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421. Lisbon, Portugal (2015)