Emotion Recognition in Human Voice Speech Based on Machine Learning

Xiaorui Wang(✉)

City University of Hong Kong, Kowloon, Hong Kong
xwang2456-c@my.cityu.edu.hk

Abstract. Emotion recognition of speech is the basis of human-computer interaction interface, and has also made remarkable development in the past decade. This paper summarizes the preprocessing and feature extraction methods and different existing emotion models and database. The speech is collected and converted into signals, and then the extracted features are identified through the speech recognition model. It also explains the still existing shortcomings at the present stage, such as emotional complexity and high-quality emotional corpus is difficult to obtain. In addition, this paper also introduces the future development and optimization direction of speech emotion recognition. It's important for researchers to accurately describe the association between emotion and acoustic characteristics, and construct a recognition model that is reasonable and as close to the emotion processing mechanism of human brain.

Keywords: Emotion recognition · Machine learning · Emotion feature extraction · Classification algorithms

1 Introduction

The classification and recognition of emotion in speech is a link that cannot be ignored in the development of artificial intelligence, which help us to better realize the human-computer interaction and improve the emotional intelligence of computers. The expression and reception of emotions are an inseparable part of the communication between people. These emotional changes come from the reception of subtle information, including the tone of speech, cadence, specific tone words, and some fixed words. Speech emotion recognition (SER) uses a computer to automatically identify the emotional state of the input speech. It has broad application prospects in the fields of mental health monitoring, educational assistance, personalized content recommendation, customer service quality monitoring and other fields. At present, speech and emotion recognition mainly consists of the following processes: preprocessing, feature extraction and emotion classification. In the past 30 years, scholars all over the world have focused on the field of emotional speech recognition, which has made remarkable achievements and progress, but there are still many defects and challenges to be improved and supplemented by future generations. Based on the existing achievements in the research field of speech and emotion recognition, this paper will summarize the research progress in this field and prospect the future technology development trend.

Y. Chen et al. (Eds.): ICMETSS 2022, ASSEHR 693, pp. 149–157, 2023.
https://doi.org/10.2991/978-2-494069-45-9_19

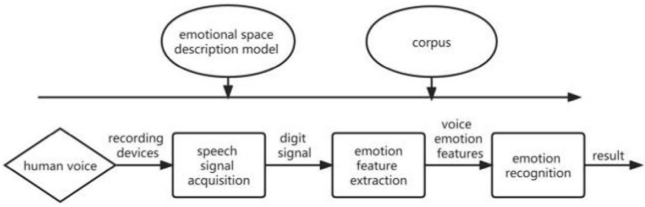

Fig. 1. Framework of a standard speech emotion recognition system

2 Analysis

2.1 Composition of the Speech and Emotion Recognition System

The whole SER process is mainly divided into three parts: speech signal acquisition, emotion feature extraction and emotion recognition.

The voice signal acquisition process is to collect human voice through recording devices such as microphone and to convert the real voice into signals for subsequent processing. The feature extraction module extracts the obtained speech signals for emotion-related parameters. Finally, then the identification operation is performed. In addition, the speech recognition system also includes the emotional space description model and the corpus. Different emotion description models have different standards, and the results will also have a certain impact (Fig. 1).

2.2 Research Status

In the late 1980s to the early 1990s, an "emotion editor" has been constructed by the MIT Multimedia Lab to collect various emotional signals in the outside world. Facial expression signals, human physiological signals, voice signals were used to jointly process and identify emotions, and let the machine to make an appropriate simple response to various emotions [1]; In 1999, Moriyama proposed a model of the linear association between speech and emotion, Accordingly, the image acquisition system voice interface that can recognize the user emotion is built in the e-commerce system, The initial application of voice emotion in e-commerce, is realized. Overall, speech emotion recognition research in this period is still in the primary stage [2], speech emotion recognition research mainly focuses on emotional acoustic characteristics analysis on the one hand, as the study object of emotional speech sample is more small scale, low nature, simple semantic characteristics, although there are a considerable number of valuable research results, but did not form a widely recognized, system theory and research methods.

Since the beginning of the 21st century, with the emergence of computer multimedia information processing technology and other research fields and the rapid development of artificial intelligence, the research of speech and emotion recognition has been given more urgent requirements, and the pace of development has been gradually accelerated. In 2000, the first ISCA Workshop on Speech and Emotion International Conference in Ireland brought together scholars dedicated to emotion and speech studies. in recent years, Several conferences and journals focusing on emotion computing,

including speech and emotion computing, have been created, And has received worldwide attention, One of the more famous is: the Affective Computing and Intelligent Interaction biennial meeting, which began in 2005, Started in 2009, Founded in the IEEE Transactions on Affective Computing journal in 2010. At the same time, the top universities and research offices in various countries have also started their research on speech and emotion recognition, including Media Research Laboratory led by MIT Picard; Emotion Research Laboratory led by Soberer at University of Geneva; Emotional Voice Group led by Cowie and Douglas-Cowie at Queen's University in Belfast; The Institute of Human-Computer Interaction and Media Integration, Tsinghua University and Human-computer voice interaction team at the Schuller of the Technical University of Munich.

2.2.1 Emotion Feature Extraction

At present, the acoustic features used for speech emotion recognition can be roughly summarized into three types: prosody features, spectrum-based related features and sound quality features. These features are often extracted in frames, but they participate in emotion recognition in the form of global feature statistics. The unit of global statistics is generally auditory independent statements or words, and the commonly used statistical indicators include extreme value, extreme value range, variance, and so on.

2.2.1.1 Spectra-Based Correlation Features

Spectrum-based related features are considered to reflect the correlation between the shape change of vocal tract and articulator movement [6], which have been successfully used in the field of voice signal processing, including voice recognition, voice recognition and so on. By studying the related spectral characteristics of emotional speech, the team of Nwe found that the emotional content in speech has a significant impact on the distribution of spectral energy in each spectrum interval [7]. For example, speech sounds that express pleasure emotions show high energy in the high frequency band, while those that express sadness show significantly different low energy in the same frequency band. In recent years, more and more researchers have applied spectral related characteristics to the recognition of speech emotion to [7], and played a role in improving the recognition performance of the system. The emotion discrimination ability of related spectral characteristics cannot be ignored. Linear spectral features (linear-based spectral feature) used in speech and emotion recognition tasks generally include: LPC (linear predictor coefficient) [8], OSALPC (one-sided autocorrelation linear predictor coefficient) [9], LFPC (log-frequency power coefficient) [5]; Inverse spectrum features (cepstral-based spectral feature) generally include: LPCC (linear predictor cepstral coefficient), OSALPCC (cepstral-based OSALPC), MFCC (mel-frequency cepstral coefficient) and so on. At present, it seems to be inconclusive on the emotional differentiation ability of linear spectrum characteristics and backward spectrum characteristics. Bou-Ghazale studied the performance of inverted and linear spectral features in the pressure voice detection (detecting speech under stress) task, and found that the discrimination ability of inverted spectral features OSALPCC, LPCC and MFCC was significantly better than the linear spectral features LPC and OSALPC. However, Nwe reached the opposite conclusion.

Specifically, HMM was used as a classifier for speaker-related identification of six emotions including anger, disgust, fear, pleasure, sadness and surprise, showing that LFPC achieved 77.1%, compared to 56.1% and 59.0% for LPCC and MFCC, respectively.

2.2.1.2 Sound Quality Features

Sound quality is a subjective evaluation index given by people, which is used to measure whether speech is pure, clear and easy to identify. Acoustic manifestations that affect the sound quality include wheezing, trebrato, SOB, and often appear in the speaker's emotional and uncontrollable. In the hearing discrimination experiment of voice emotion, the change of sound quality was unanimously identified by the listeners as closely related to the expression of speech emotion. In the study of speech and emotion recognition, the acoustic characteristics used to measure sound quality generally include: resonance peak frequency and its format frequency and bandwidth, frequency perturbation and shimmer, glottal parameter, etc.

In a series of works, Lugger extracted the first and fourth resonant frequencies and the corresponding bandwidth as sound quality features, together with prosodic features such as base frequency, for speaker-irrelevant speech emotion recognition. Li extracted frequency and amplitude perturbations as sound quality parameters for the corpus data in the SUSAS database, and HMM (hidden Markov model) was used as an identifier. Compared to the baseline performance of 65.5%, the feature combination of MFCC and frequency perturbation is 68.1%, the MFCC and amplitude perturbation achieves 68.5%, and the best performance is 69.1% from a common combination of MFCC, frequency perturbation and amplitude perturbation.

2.2.1.3 Rhythm Features

Rhythm refers to the changes of pitch, sound length, speed and weight above the semantic symbols, which is a structural arrangement of the expression mode of speech flow. Its existence or not does not affect our listening to words, words, sentences, but determines whether a sentence sounds natural to the ear and cadence. Prosody features are also known as "supersegment features" or "hyperlinguistic features". Their emotion discrimination ability has been widely recognized by researchers in the field of speech and emotion recognition, and is widely used, among which the most commonly used rhythmic features are duration, pitch, energy and so on.

Origlia using base frequency and energy-dependent maximum, minimum, mean, and standard deviation to form a 31-dimensional promedic feature set, achieving a recognition rate of nearly 60% on a multilingual emotional corpus with Italian, French, English, and German. Seppanen used 43-dimensional global prosodic features of base frequency, energy, and time correlation for emotion recognition in Finnish, achieving a 60% recognition rate when the speaker was not related. Iliou and Wang used the base frequency, energy and time length for irrelevant emotion recognition of German speakers and emotion recognition of Chinese Mandarin emotion, with recognition rates of 51% and 88%, respectively.

2.2.2 Three Machine Learning Classification Algorithms

Random Forest, Support vector machine and K-nearest neighbor are the three most commonly used classification algorithms.

2.2.2.1 Random Forest

Random forest is an integrated algorithm based on the principle of "a set of weak estimates will be combined together to form a strong estimate". The random forest classifier designed in this paper is a set of 100 decision trees. Let X and Y represent the total number of emotions and the total number of samples, respectively. A set of bootstrap samples were selected for each decision tree. The decision tree is then constructed by assigning an $y < Y$ variable on each node until all the variables are exhausted. At each node of the tree, the classification model was fitted with their variables, and the cutoffs were found. After training, species predictions of unknown samples can determine [3, 4] by voting the majority of all prediction ranks across all individual trees.

2.2.2.2 Support Vector Machine (SVM)

Support vector machine is a new machine learning method proposed by [7] in the 1990s. It is a generalized linear classification algorithm for the binary classification of data by supervised learning. The basic idea of the SVM is to map its input vector to another high-dimensional space by performing nonlinear transformations of the non-linear separable samples. To achieve linear separation in this new space, finding the optimal classification hyperplane maximizes the distance between the hyperplane and the sample sets of different classes, thus achieving the maximum generalization power.

It has evolved into four types: 1) linear separable types. When trying to split 2 data types, there is at least one split plane that can completely separate 2 types, which have no boundaries, and therefore cannot control the error well, called linear separable types. 2) The linear type. Cannot completely divide 2 data types, but can make the vast majority of data segmentation, this type is called a linear type. 3) Nonlinear type. When a type is not linearly separable, the SVM handles data classification by providing a "soft boundary" that allows some elements of a set of data to fall on the other side, but allows them to pass through this hyperplane without producing any significant anomalies, allowing the system to become less rigorous and more robust.

2.2.2.3 KNN Algorithm

Traditional KNN algorithm compares with every sample vector in sample space in order to find k neighbors of classification of the sample. From the perspective of classification process, KNN most directly makes use of the relationship between samples and samples, reduces the adverse effects of improper category feature selection on classification, and can minimize the error items in the classification process. In order to find out the k neighbors of the samples to be classified, it is necessary to compare with each sample vector in the sample space. When the training samples are large, the classification speed decreases. In order to solve the problem of excessive similarity calculation, the KNN classification method based on inverted index is proposed.

Table 1. Various definitions of emotion from different researchers [10]

Researcher	Emotion
Arnold	Anger, aversion, courage, dejection, desire despair, dear, hate, hope, love, sadness
Ekman, Friesen, Ellsworth	Anger, disgust, fear, joy, sadness, surprise
Gray	Desire, happiness, interest, surprise, wonder, sorrow
Fridja	Desire, happiness, interest, surprise, wonder, sorrow
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James	Fear, grief, love, rage
McDougall	Fear, disgust, elation, fear, subjection, tender-emotion, wonder Pain, pleasure
Mower	Pain, pleasure
Oatley, Johnson-Laird	Anger, disgust, anxiety, happiness, sadness
Panksepp	Anger, disgust, anxiety, happiness, sadness
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Wats0On	Fear, love rage
Weiner, Graham	Happiness, sadness

2.2.3 Two Types of Mainstream Emotion Description Models

Emotional description can be divided roughly into two forms: discrete and dimension.

2.2.3.1 Discrete Emotion Description Model

The former describes emotion as discrete, adjective labels, such as joy, anger, etc., which are widely used in people's daily communication process, while also being widely used in early emotion-related research. Rich language labels describe a large number of emotional states, so which of these emotional states are more universal? This question boils down to the determination of the basic emotion categories. It is generally believed that those emotional categories that can span different human cultures and can even be shared by humans and social mammals are basic emotions. Table 1 [10] lists the definitions and divisions of basic emotions by different scholars. Among them, the six basic emotions (also known as big six) proposed by American psychologist Ekman are widely used in the field of emotion-related research today.

2.2.3.2 Dimension Emotion Description Model

The latter describes emotional states as points in a multidimensional emotional space. The emotional space here is actually a Cartesian space, and each dimension of the space corresponds to a psychological attribute of the emotion. In theory, the space's ability to describe emotions is able to cover all emotional states. In other words, any and real emotional state can find the corresponding mapping point in the emotional space, and

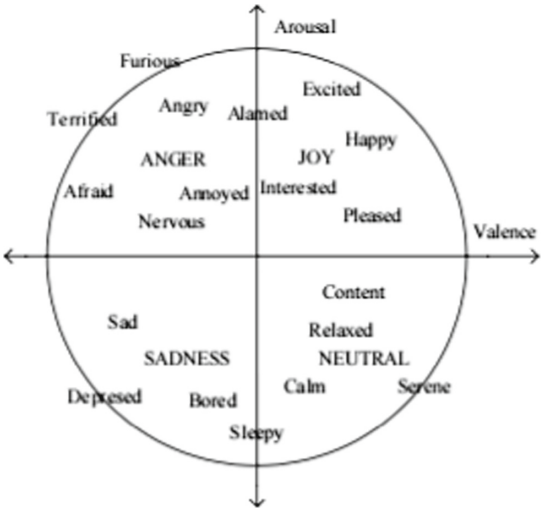

Fig. 2. Arousal-Valence emotional space

the value of the coordinate value of each dimension reflects the strength of the emotional state in the corresponding dimension (Fig. 2).

2.2.3.3 Advantages

Two expression models have different strengths: in terms of model complexity, The discrete description model is relatively concise and easy to understand, It is conducive to the initiation and development of relevant research work, However, the dimensional model faces the problem of how to transform between qualitative emotional state and quantitative spatial coordinates; In terms of the ability to describe emotions, The emotion description ability of the discrete emotion models shows great limitations, In the most cases, It can only depict a single, finite variety of type of emotion, Yet the emotions that people experience in their daily life are subtle and varied, Even the complex and vague (for example, the emotions people show when startled, It often contains fear and even fear.), in a manner of speaking, There are still large barriers between discrete descriptions and descriptions of spontaneous emotions, However, the dimensional emotion model describes the emotion from a multi-faceted, continuous perspective, Well addressing the description of spontaneous emotions, it also largely avoids the ambiguity problem of discrete emotion labels with accurate numerical values.

2.2.4 Current Emotional Speech Database

There is no unified standard for the emotion database in the current domain, and different classification results can be obtained according to different classification criteria. For example, according to different kinds of language it can be divided into English, German, Chinese, etc.

2.2.4.1 FAU AIBO Children's German Emotional Speech Database

FAU AIBO [11] recorded 51 children (aged 10 to 13, 21 men and 30 women) during

the electronic pet AIBO game with SONY A total duration of 9.2 h (excluding pauses), including 48,401 words. Voice through the A high-quality wireless headset was collected and recorded by DAT-recorder, sampled at 48 kHz (then compressed to 16 kHz), and quantified at 16bit. For the purposes of the record in a real emotional voice, the staff makes the children believe that AIBO can respond to and execute their verbal commands, but in fact, AIBO is Secretly manipulated by the staff members. The labeling work was completed by five college students majoring in linguistics, and the final labeling method was decided by voting Fruit, 11 emotion tags including joyful, irritated, angry, neutral, etc.

2.2.4.2 VAM Database

The VAM Database [12] is an unpaid database for scientific research purposes through a German-language television talk show called "Vera am Mittag" Live recording results, and voice and video are saved at the same time, so the database contains three parts: corpus, video library, and expression library. The conversation was about nothing Script-restricted, pure natural communication without emotional guidance. Take the VAM-audio library, which contains 947 sentences of recorded data from 47 program guests, In the wav format, 16 kHz sampling, 16bit quantification. All data were saved in sentences (1018 sentences), annotated in Valence, Activation and Dominance was performed on these three emotional dimensions, with annotation values ranging between 1 and 1. The annotation work is done by multiple annotators together, and the final emotional value is the phase Average of the ators.

2.2.4.3 Belfast English Emotion Database

The Belfast emotion database [14] was recorded by Cowie of Queen University, and by 40 recorders (18–69, 20 men and 20 women) to 5 The three paragraphs are deduced. Each paragraph contains 7 to 8 sentences, and has a certain emotional, such as tendency, anger, fear, neutral, sadness, happiness.

3 Conclusion

There is still the problem of many features and many irrelevant features in speech and emotion recognition data. Data enhancement can be employed in preprocessing in subsequent studies to solve the problems such as overfitting. Studying data enhancement in speech samples considers the characteristics of human sounds. The frequency range of human voices is mainly divided into male and female 2 parts. The male voice is divided into Tenor, Baritone, and Bass. Similarly, female voices can be divided into Soprano, Mezzo-soprano and Alto. The study shows that the bass and baritone have the same frequency range as the alto and mezzo-soprano; the soprano is slightly higher than the tenor; the average frequency of the female voice is about 1 to 1.75 times that of the male voice.

References

1. Cahn JE. The generation of affect in synthesized speech. Journal of the American Voice Input/Output Society, 1990, 8:1−19.

2. Moriyama T, Ozawa S. Emotion recognition and synthesis system on speech. In: Proc. of the 1999 IEEE Int'l Conf. on Multimedia
3. Computing and Systems (ICMCS). Florence: IEEE Computer Society, 1999. 840−844. doi: https://doi.org/10.1109/MMCS.1999.779310
4. PAUL A, MUKHERJEE D P, DAS P, et al. Improved random forest for classification [J]. IEEE transactions on image processing, 2018, 27(8): 4012–4024
5. BUTT A M, BHATTI Y K, HUSSAIN F. Emotional speech recognition using SMILE features and random forest tree [M]. BI Yaxin, BHATIA R, KAPOOR S. Intelligent Systems and Applications. Cham: Springer, 2020.
6. BOSER B E, GUYON I M, VAPNIK V N. A training algorithm for optimal margin classifiers [C]// Proceedings of the 5th Annual Workshop on Computational Learning Theory. Pittsburgh, PA, USA, 1992: 144–152.
7. Benesty J, Sondhi MM, Huang Y. Springer Handbook of Speech Processing. Berlin: Springer-Verlag, 2008. doi: https://doi.org/10.1007/978-3-540-49127-9
8. Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden Markov models. Speech Communication, 2003,41(4): 603−623. doi: https://doi.org/10.1016/S0167-6393(03)00099-2
9. Rabiner LR, Schafer RW. Digital Processing of Speech Signal. London: Prentice Hall, 1978.
10. Hernando J, Nadeu C. Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. IEEE Trans. on Speech and Audio Processing, 1997, 5(1):80−84. doi: https://doi.org/10.1109/89.554273
11. Ortony A, Turner TJ. What's basic about basic emotions. Psychological Review, 1990,97(3):315−331. [doi: https://doi.org/10.1037/0033-295X.97.3.315.
12. Steidl S. Automatic classification of emotion-related user states in spontaneous children's speech [Ph.D. Thesis]. Erlangen: University at Erlangen Nurberg, 2009.
13. Grimm M, Kroschel K, Narayanan S. The vera am mittag german audiovisual emotional speech database. In: Proc. of the 2008 IEEE Int'l Conf. on Multimedia and Expo (ICME). Hannover: IEEE Computer Society, 2008. 865−868. doi: https://doi.org/10.1109/ICME.2008.4607572
14. McGilloway S, Cowie R, Douglas-Cowie E, Gielen S, Westerdijk M, Stroeve S. Approaching automatic recognition of emotion from voice: A rough benchmark. In: Proc. of the 2000 ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. Belfast: ISCA, 2000. 207−212.