Literature Review

Introduction

With the development of human-computer interaction technology, more and more researchers pay attention to the emotional factors in speech signals. With the purpose of understanding the framework of the speech emotion recognition, I reviewed some research in this field in recent years. Although different literatures have used different algorithms to realize speech emotion recognition or focus on taking certain measures to improve the accuracy of speech emotion recognition, I hold the idea that they will have common and necessary components in processing logic or processing flow, which is beneficial for us to have a better understanding of the components needed in the framework of the speech emotion recognition.

The speech emotion recognition

Emotion is the basic psychological attribute of human beings and higher organisms. It is one of the most important external communication channels for emotional expression. It is thought that speech signal of human can carry a lot of emotional information (Xinzhou, 2017). In speech emotion analysis, the basic emotional states include calm, fear, anger, pleasure, irritability, surprise, disgust and so on. Speech emotion recognition refers to the feature extraction of emotion signals in frames through computer processing. Based on the extracted features, computer can simulate human perception and understanding of human emotion to infer the type of speech emotion (Akçay et al., 2020). According to Xinzhou (2017) and El Ayadi (2011), we can use human-computer interaction or existing speech database as the input of speech emotion recognition system. Based on the requirements of the project, The component to get the sound emitted by users in our project are indispensable, meaning we can the use machine to obtain the voice signal of the communication object through the external interface. A complete speech emotion system must include two parts. The first part is the sound signal processing and extracting effective features. The second part is a good speech emotion classification algorithm (Zhiyan et al., 2019). For the first part, the collected speech signals generally need to be pre-processed, such as framing, pre-emphasised, simple denoising and so on. In El Ayadi (2011), they also convert user voice into text as input to the language model, whose output will be as the factor to help the system recognize the word sequence and emotions. And feature parameter extraction is the basis of speech emotion recognition that plays a crucial role in the performance of emotion recognition. At present, these extracted features can be divided into three categories: prosody related features, sound quality related features and spectrum based related features (Xinzhou, 2017), as shown in the following table.

The categories of extracted features	**Description**
Prosody related features	Prosody is the structural arrangement of speech signal

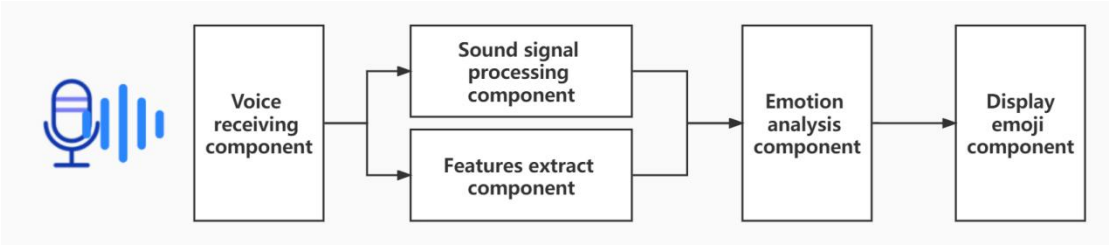
	expression, which mainly includes the changes of length, pitch, weight and speed in speech signal.
sound quality related features	Sound quality is often measured from the purity, legibility and clarity of speech. When people are very excited and difficult to control in the face of some special situations, they will show choking, vibrato and wheezing.
Spectrum based related features	Speech emotion is closely related to the level of spectrum energy on the spectrum. For example, happiness shows high energy in the high-frequency spectrum.

For the second part, we need to use the component that integrates emotion analysis algorithms to analyze the features aiming to classify the emotion of the user's input speech. A variety of algorithms can be used, such as HMM, KNN, SVM and so on.

Algorithm	**Mechanism type**	**Advantage**	**Disadvantage**
DTW	Calculate the distance of time series data	Reliable time alignment between reference and test patterns is obtained	The heavy computational burden required to find the optimal time alignment path
HMM	Statistical model of Markov process	The modeling ability of time series is strong and the expansibility of the system is good	The model has high complexity, general fitting function and poor robustness
KNN	Supervised simple machine learning algorithm	The algorithm is simple and the theory is mature	Large amount of calculation, weak interpretability and large amount of memory
SVM	Machine learning algorithm based on statistical learning theory	Good robustness and global optimization	The recognition efficiency of large-scale samples is low, and it is difficult to solve the multi classification problem
CNN	Depth neural network in space	Shared convolution kernel, stronger generalization ability and good feature	Gradient dissipation is easy to occur

		classification effect	
RNN	Temporal depth neural network	Strong ability to model sequence content	It is prone to gradient dissipation or gradient explosion

Once the user's voice emotion is analyzed. The component needs to display the corresponding emoji to the user. Based on above discussion, the components needed in the framework of the speech emotion recognition are shown in the following figure.



The processing flow of these components can be summarized: the machine obtains the voice signal input by the user through the external interface, and then the component pre-processes the voice signal and extracts the characteristics of the voice signal samples and uses the relevant algorithms to analyze and get the result of the emotional state. Finally, the system makes a decision to select the corresponding emoji to display to the user.

References

Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, *59*, 101894.

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56-76.

Zhiyan, H. A. N., & Jian, W. A. N. G. (2019, June). Speech emotion recognition based on deep learning and kernel nonlinear PSVM. In *2019 Chinese control and decision conference (CCDC)* (pp. 1426-1430). IEEE.

Xu Xinzhou (2017). *Research on Speech Emotion Recognition based on Emotion Feature Information Enhancement* (doctoral dissertation, Southeast University).https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CDFDLAST2018&filename=1018002870.nh

Pang huan.(2012). *The Research on Feature Extraction and Recognition of Emotional Speech*(Master's thesis, Changsha University of

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition,* 44(3), 572-587.

Technology).https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2012&filename=1012348443.nh