Machine Learning Algorithms for Speech Emotion Classification

Chuqiao Yao

School of Electrical and Energy Power Engineering, Yangzhou University, China

Abstract. Artificial intelligence products work by imitating the way humans think. 'Communication' is one of the most important links in the sustainable development of human society. Meanwhile, allowing artificial intelligence to communicate with humans is also one of the main aspects of scientific research. In recent years, the level of development of artificial intelligence has already made some smart products that can chat with humans when they are bored. It can be predicted that in the future, we can have robot pets to accompany us. In order to make companion-type products to be more similar to humans, it is necessary to implant algorithms that could recognize human's expressions of emotion. this paper is introduced two methods that can help AI to recognize emotion in human speech. One is to recognize words that express emotion, and the other is to recognize the tone of speech to judge emotion. Finally, it discusses the feasibility of this algorithm in reality.

Keywords: Artificial Intelligence; Speech Emotional; Classification.

1. Introduction

1.1 Background

It has been nearly seven decades since the concept of artificial intelligence was proposed. It is playing an increasingly important role in our lives. AI can be seen everywhere in our daily life, such as fingerprint recognition and face recognition installed on mobile phones or doors for locking and unlocking, robots that can play chess with us, and smart search function which can conjecture our preferred based on our browsing history etc. What they all have in common is that they learn the way humans behave in order to provide more convenience for our lives. Almost all of us have smart voice assistants installed on our mobile phones. It appears when we call out its name and completes the task according to the instructions we give. Like calling someone, checking the weather, playing a song, navigating to a destination, etc. Having a mobile assistant can save many things that need to be done manually. Sometimes they can chat with us as friends and tell us jokes when we are bored. But in the process of communicating with it, we will find that the voice assistant can only complete the work according to the commands we issue, and cannot communicate with us for a long time like a real human. Because humans are sentient beings, it is difficult for machines that are run by code to understand human emotions. Considering that in a few years from now we may have robotic assistants in our homes, communicating with them is essential.

It is easy for us to understand that when speaking, the same words may have different meanings if they are spoken in different tones. For artificial intelligence, if it only recognizes words, it is likely to make mistakes in judgement, and make strange reactions. Therefore, it may destroy the experience of human-computer communication. In order for artificial intelligence and humans to communicate better, an important part of it is trying to make artificial intelligence understand human emotions. When human speak, we can express our emotions in two ways, one is to say words that are emotionalized, the other is to reflect the mood at that time through the tone. So, if artificial intelligence can recognize these two contexts, it can get along with humans better.

1.2 Literature Review

Zhang et.[1] studied that: Speech recognition technology converts speech data into digital information through artificial intelligence processing technology, and extracts the required characteristic parameters from it. The machine cannot directly recognize a piece of speech into text, but through decomposition, encoding, decoding, and several steps, the sound signal is divided into

tiny parts and then pieced together to form the final recognition result. The process has five steps from capturing sound signals to recognizing speech.Figure1 shows that in the process design, the first step is to collect data to establish a data set, and the second step is to extract the characteristics of the data set. The third step is to achieve model training through data sets, and the fourth step is to conduct application testing. The last step is to identify and output the final result.

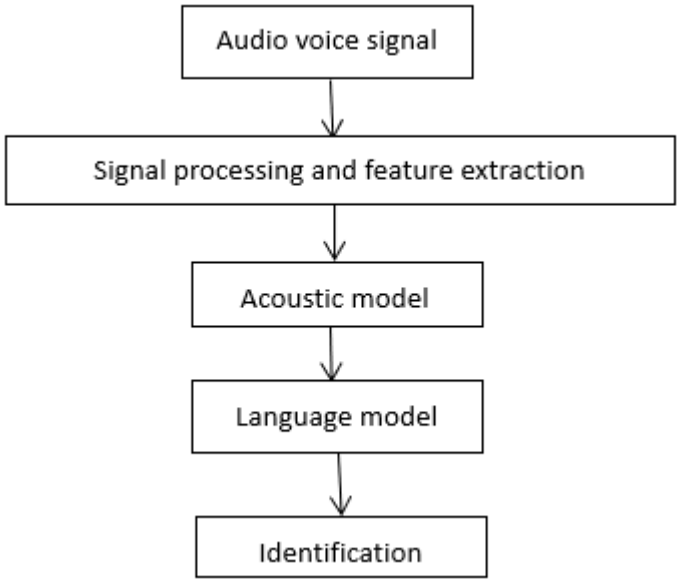

Figure 1. The calculation process of identification results

Zhang et.[2] Studied that in the flow design, after the audio signal is input, the signal is enhanced. The speech signal is converted from the time domain to the frequency domain. Then the special name extraction is given to the acoustic model to output high-quality signal data. The decoder can calculate the data in the acoustic mode and the language mode, and output a priority word sequence in the recognition result.

2. Methodology

2.1 Introduction of the First Method: Recognize Emotional Words

The first method is an algorithm that recognizes emotional words. For example, ‘happy’, ‘joyful’...They can mean excitement. And ‘sad’, ‘disappointed’ ... can mean depression. In addition, ‘good’ and ‘poor’ are also opinion words that can reflect emotions.

Recognizing emotional words requires inserting the recognition algorithm in the third step described in Figure 1. A machine goes through a total of three steps when recognizing speech. The first step is voice input, a voice signal is received. The second step is Automatic Speech Recognition (ASR), which is the process of converting sound into text, which is equivalent to the role of the ear. This step generally uses an "end-to-end" approach based on Markov Models (HMM) of deep neural networks. It needs to go through four processes of "input, encoding, decoding, and output". The computer cannot directly recognize the input sound signal, so the input signal must be cut into small segments, and then each segment is represented by a vector according to certain rules. Then put the compiled vector into the acoustic model, you can decode the letters corresponding to each small segment, and connect them to translate into a complete sentence. The third step is Natural Language Processing (NLP), which is the process of understanding and processing, equivalent to the brain. Recognizing emotional words requires inserting the recognition algorithm in the third step. Once the keywords are identified in the translated text, it is possible to guess what the sentiment expressed by the sentence is.We can insert as many emotional words into the program as possible. Once the

keywords are identified in the translated text, it is possible to guess what the sentiment expressed by the sentence is. Sentiment tendency analysis is mainly divided into two parts: sentiment tendency category and opinion extraction. Sentiment tendency is to identify whether words in a sentence are positive or negative. This type of analysis has a good effect on recommending content and services to users and evaluating restaurants. Opinion extraction is the extraction of opinion vocabulary, which is of great help in establishing a service or content evaluation system.

2.2 Introduction of the Second Method: Recognizing the Tone of Speech

The second is an algorithm that recognizes the tone of voice when humans speak. Sound is a sound wave produced by the vibration of an object. It is a wave phenomenon that propagates through a medium and can be perceived by the human or animal auditory organs. The object that initially vibrates is called the source of the sound. Sound travels in vibrations in the form of waves. Sound is the motion of sound waves propagating through any medium. Frequency is the number of sound waves passing a given point per second. A syllable is the smallest unit of speech that can be naturally detected by hearing, and a syllable consists of three parts: initials, finals, and tones. A word may consist of one or more syllables, and according to the syllables, it can be divided into different categories. A phoneme is the smallest unit of speech that is analyzed from a syllable, and speech can be decomposed into the smallest unit that is a phoneme.

We can draw up some communication scenes in advance, like in a movie, let two people talk and record the scenes. Put it into the computer to analyze the audio data. We can develop systems that understand 50 human emotional signals. Speech signals have time-domain characteristics and short-term energy.

The speech signal has time-varying characteristics and is a non-stationary random process. But its properties remain largely unchanged over a short period of time. In the time domain, the speech signal can be directly represented by its time waveform. Among them, the unvoiced segment is similar to white noise, with high frequency, but small amplitude and no obvious periodicity; while voiced sound has obvious periodicity, and has large amplitude and relatively low frequency. These time-domain features of speech signals can be analyzed by methods such as short-term energy and short-term zero-crossing rate. Since the energy of speech signals varies over time, the difference in energy between unvoiced and voiced sounds is quite significant. Therefore, the analysis of short-term energy and short-term average amplitude can describe this characteristic change of speech.

The so-called speech recognition is to convert a piece of speech signal into corresponding text information. The system mainly includes four parts: feature extraction, acoustic model, language model, dictionary and decoding. The audio data which is processed ahead of time work such as filtering and framing is performed on the sound signal, and the audio signal that needs to be analyzed is properly extracted from the original signal; the feature extraction work converts the sound signal from the time domain to the frequency domain to provide suitable acoustic models for the acoustic model. feature vector; in the acoustic model, the score of each feature vector on the acoustic feature is calculated according to the acoustic characteristics; while the language model calculates the probability that the sound signal corresponds to the possible sequence of phrases according to the theory related to linguistics; finally, according to the existing dictionary , decode the sequence of phrases to get the final possible text representation.

This system can measure and track changes in speed, volume, pitch, timbre and prolonged speech pauses. Categorize different contexts that can express the same emotion, analyze the common points of these voices in terms of frequency, pitch, etc., and summarize the frequency and band of each emotion. Real-time monitoring and analysis of audio data during speech recognition. When the frequency matches the frequency range representing a certain emotion, it can be inferred that this emotion is expressed.

Processing ahead of time: 1. The mute removal at the head and tail ends reduces the interference to the subsequent steps. The operation of mute removal is generally called VAD. 2. Framing the sound, that is, cutting the sound into small segments, each segment is called a frame,

Feature extraction: The purpose is to convert each frame of waveform into a multidimensional vector containing sound information.

Acoustic model (AM): obtained by training speech data, the input is a feature vector, and the output is phoneme information.

Dictionary: The correspondence between words or words and phonemes. In simple terms, Chinese is the correspondence between pinyin and Chinese characters, and English is the correspondence between phonetic symbols and words.

Language Model (LM): Obtain the probability that a single word or words are related to each other by training a large amount of text information.

Decoding: It is to output the audio data after feature extraction through acoustic model, dictionary and language model.

3. Discussion

For the first algorithm for recognizing words, because there are countless words that can represent emotions in daily expressions, we cannot guarantee that all emotional words can be entered into the program. And we don't need to express anger by saying emotional words like "I'm already mad!". Emotions are expressed most of the time through tone of voice and expressions, body movements, etc. In addition, everyone speaks with a different accent, and there are various dialects in the world. Machines cannot avoid the problem of identifying the wrong words when recognizing and refining speech. In the era of networking, new network terms will be generated from time to time. At this time, developers need to update the program in a timely manner. Therefore, it is not enough to recognize words alone. This shows that artificial intelligence cannot really communicate with humans as smoothly as humans in the short term. When recognizing our emotions, we also need our deliberate expressions to help them understand emotions. Or install a camera on the machine, combined with the recognition of emotional vocabulary and tone, plus the use of the camera to capture facial expressions, it may also be necessary to capture body movements, and comprehensive analysis in multiple aspects can improve the accuracy of judging emotions. In addition, it takes a heavy workload to finish this work, but the functionality of the resulting product may not be worth that much money.

4. Conclusion

To sum up, the first algorithm (recognizing emotional words) is simpler and more accurate in recognizing words. But it can't cover all the emotional words, and developers need to constantly update the thesaurus. The advantage of the second algorithm (Recognizing the tone of speech) is that it only needs to simulate the context in advance. When there is no specific word, it can also rely on the tone to identify emotions, which has a wider scope of application. The disadvantage is that it needs complex programs to implement, and may need the help of cameras to improve accuracy. AI emotion recognition algorithm can help elderly people living alone and some autistic patients in the medical industry to a certain extent, communicate with them as friends, pets, and even doctors, and relieve them from the distress caused by loneliness. In the face of mental problems, AI can intervene to triage patients before the doctor's consultation, gain a preliminary understanding of the disease, and divide patients into preliminary departments, aim at lightening the load for busy doctors.

5. Future Work

In the future, speech recognition human emotion technology still needs a lot of investment research. Improve the accuracy of recognition and try to reach the level equivalent to that of humans. This can be commercialized and put into production. Make a product like a home assistant robot or an office assistant robot and put it into production. As the recognition accuracy of speech recognition technology has been greatly improved, and the application scenarios have become more and more abundant, we believe that the development of artificial intelligence speech recognition capabilities is

not yet its peak, and research and market applications in this field have yet to be explored. We predict that intelligent speech recognition can achieve the goal of multilingual, large-scale, human-machine collaboration in the future.

References

[1] Zhang Yanning, Chang Ying, Chen Haiyan, Zhang Jingfeng, research on AI based speech recognition technology. [TN912.34; TP18] information technology, 2022.8.

[2] Zhang Huie, Li Caihong, Wang Jin, et al. Design and implementation of multimedia memo based on Android [J]. Computer knowledge and technology, 2019, 15 (17): 102-103.

[3] Sp A, Mrt A, Ebe A, et al. Speech emotion recognition based on machine learning tactics and algorithms – ScienceDirect [J]. Materials Today: Proceedings, 2021.