Speech Sentiment Analysis Based on Basic Characteristics of Speech Signal

Zijun Yang^{a,*}, Lifeng Zhang^a, Seiichi Serikawa^b

^aKyushu Institute of Technology, Kitakyushu, 8048550, Japan

^aKyushu Institute of Technology, Kitakyushu, 8048550, Japan

*Corresponding Author: zijun.yang529@mail.kyutech.jp

Abstract

As the pace of life continues to accelerate, people's life pressure is increasing. With the accumulation of time, people's mental and psychological conditions have been affected to a certain extent. These mental illnesses will not cause any impact under normal circumstances, but once they break out, they will cause trauma that cannot be ignored in life or even in society. Therefore, we hope to design a system program that can chat with humans in daily life, and it can feel the human's emotional changes in daily conversations. When humans have negative emotions, it can comfort us in time and even warn humans when our negative emotions reach a certain limit. When humans have positive emotions, it can give humans affirmative approval and encouragement. Based on this concept, we must first analyze the different emotions that humans design in daily conversations. This article is mainly based on the basic characteristics of audio signals to judge the user's emotional changes. The database we use is six different emotional voices recorded by four voice actors, and each voice contains 50 single sentences for emotional recognition analysis.

keywords: Voice emotions analysis, Voice feature value, Voice speed

1 Introduction

In our lives, there are many confirmed mental illness patients and hidden potential mental illness patients. With the rapid increase in life pressure, everyone in our lives may have certain spiritual risks. Perhaps we may live routinely as usual, but there seems to be a bomb in the spirit. Maybe when or being overwhelmed by the last straw and breaking out, they may hurt themselves due to excessive stress, or they may retaliate against society due to excessive stress, causing social panic. Take

depression as an example: According to statistics from the WHO, there are approximately 264 million depression patients worldwide, and up to 1 million depression patients choose to commit suicide under the torture of illness each year. Although there are effective treatments for depression, most patients cannot receive treatment for various reasons. In addition to depression accidents, mental illnesses caused by excessive stress also include anxiety, split personality, and other diseases. When human beings cannot avoid the harm caused by the stress of life, we need to consider how to relieve this stress. We hope to use modern technology to help people and use artificial intelligence to realize voice sentiment analysis of the positive or negative emotions contained in people's daily dialogue. Through the sentiment analysis of the user, we can judge the user's emotional change at all times, and give positive encouragement and comfort when the user's emotion is negative; when the user's emotion is positive, express recognition and appreciation for their success. Especially for patients, it is possible to analyze whether they are onset or not, and if they cannot be comforted by talking, they can promptly send an alarm to community medical staff or emergency contacts to avoid unnecessary loss of life and social property. At the same time, in daily conversations, you can also capture dangerous signals and give users a vigilant response. This article mainly focuses on the analysis of voice emotions and analyzes the specific manifestations of the feature values of different emotions according to the different feature values of the voice. In the future, voice emotion recognition will be performed on the basis of the analysis results.^{(1),(2)}

2 Basic voice feature value analysis

First, we analyze the different performances of some basic voice features under different emotions. These ba-

Table 1: Voice duration of different emotions(s)

Emotions	Angry	Fear	Happy	Neutral	Sad	Surprise
Times	1.783	2.041	1.622	1.550	2.317	1.519

sic characteristics include voice duration, speech pressure, speech time-domain analysis: short-term energy, short-term average amplitude, short-term zero-crossing rate, and short-term average amplitude difference. In terms of the length of speech, it can also be clearly felt in our daily conversations that when we have positive emotions, our speaking speed is usually faster, and when we have negative emotions, our speaking speed it will be slower. The response to the speech duration means that we use a shorter time for positive emotions and a longer time for negative emotions for the same sentence. Therefore, we get all the voices in the database and get their duration. The specific performance of the duration of different emotions as shown in Table. 1. We can see that the sequence of time duration for different emotions is: Sad > Fear > Angry > Happy > Neutral > Surprise.

For sound pressure, as the most basic physical quantity to quantitatively describe sound waves, sound pressure is the excess pressure generated by sound disturbance and is a function of spatial position and time. Since sound pressure measurement is relatively easy to implement, and other acoustic parameters such as particle vibration velocity can also be obtained indirectly through sound pressure measurement, sound pressure has become the most commonly used physical quantity to quantitatively describe the properties of sound waves. What we usually talk about is the effective sound pressure, which is obtained by calculating the root mean square value of the instantaneous sound pressure against time in a certain time interval. Supposing the length of the voice is T and the number of discrete points is N , the effective sound pressure calculation formula as shown in Eq. (1):

$$P_e = \sqrt{\frac{1}{T} \sum_{n=1}^N x^2 \Delta t} \quad (1)$$

Where x represents the sampling point. We count the sound pressure of all voices in the database, and the result as shown in Fig. 1. After further analysis, the sound pressure of different emotions is obtained, and the result as shown in Fig. 2. We can see that the sound pressure of emotional “surprise” is too large, and the pressure of emotional “sad” is the smallest.

Then comes the time-domain analysis, which is to extract the time domain parameters of the speech signal. The time-domain analysis is usually used for the most basic parameter analysis and applications, such as speech segmentation, preprocessing, classification, etc. In the

Fig. 1: The result of sound pressure statistics

Fig. 2: Different emotions with sound pressure

time domain, the voice signal has the characteristics of “short-term”, that is, on the whole, the characteristics of the voice signal change with time, but in a short time interval, the voice signal remains stable. Commonly used time-domain parameters include short-term energy, short-term average amplitude, short-term zero-crossing rate, and short-term average amplitude difference function.^{(3), (4)} The first is short-term energy. Regarding the energy of the speech signal, it will vary considerably over time. Suppose the short-term energy of the n th frame of speech signal $x_n(m)$ is represented by E_n as shown in Eq. (2) :

$$E_n = \sum_{m=1}^N x_n^2(m) \quad (2)$$

The final result as shown in Fig. 3. We can see that emotions “angry” and “surprise” have greater short-term energy. When our emotions are “angry” and “surprise”, our emotions have greater fluctuations, and emotion “sad” has the smallest short-term energy. When we speak with “sad” emotion, we have a more stable mood.

Followed by short-term average amplitude. short-term average amplitude is a function that measures the change in the amplitude of the speech signal, which is very sensitive to high levels (calculated by the square). Short-term average amplitude function can be used M_n and as shown in Eq. (3):

$$M_n = \sum_{m=1}^N |x_n(m)| \quad (3)$$

Fig. 3: The result of short-term energy

Fig. 4: The result of short-term average amplitude

Fig. 5: The result of short-term zero-crossing rate

M_n is also a representation of the energy level of a frame of speech signal, and should not be squared to cause a big difference. The result as shown in Fig. 4.

Then there is the short-term zero-crossing rate. The short-time zero-crossing rate indicates the number of times a frame of speech signal waveform crosses the horizontal axis. That is, the number of times the symbol of one sample point changes before and after. We usually use Z_n to express as shown in Eq.(4). The result as shown in Fig. 5.

$$Z_n = \frac{1}{2} \sum_{m=1}^N |sgn[x_n(m)] - sgn[x_n(m-1)]| \quad (4)$$

Finally, is the average amplitude difference function. The short-term average amplitude difference function of the speech signal is the same as the short-term autocorrelation function, which represents the same periodic characteristic of the speech cycle, but different short-term

Fig. 6: The result of short-term average amplitude difference

average amplitude differences have valley characteristics rather than peaks at each integral multiple of the period. Therefore, the pitch period can also be determined by calculating the short-term average amplitude difference. Its advantage is: because the function only needs to add, subtract and take the absolute value calculation, the algorithm is very simple and easy to implement by hardware so that the short-term average amplitude difference function method is quite common in pitch detection. The result as shown in Fig. 6.

Based on the above analysis, we can get a rough conclusion based on different parameters, as shown in Table. 2. The manifestations of different emotions under different parameters. According to the table, we can see that different parameters have different manifestations under different emotions. We will further analyze these manifestations later.

3 Voice speed analysis

In the above, we learned that different emotions have different manifestations on different feature values. For example, the voice of emotion “Sad” has the characteristics of smoother ups and downs, longer overall time, and less energy. But we also found that the contained energy is easily interfered with by other factors, such as the distance of the sound, the noise of the environment, etc. So we quickly thought that we can analyze a person’s emotions by speaking speed. Similar to the duration of the voice, when we have positive emotions, our speech speed will become faster, and when we have negative emotions, our speech speed will become slower. So how can we get the parameter of speech rate. After analyzing the language, we found that the pronunciation of Chinese, Japanese and other texts is “one character, one sound”, that is, a character will have one pronunciation, and this pronunciation is basically composed of consonants with vowels or a single vowels to complete. In phonetics, we call the sounds that vibrate the vocal cords voiced sounds, and those that do not vibrate are called

Table 2: Different parameters correspond to the performance of different emotions

Eigenvalues	Voice duration	Speech pressure	Energy	Average amplitude	Zero-crossing rate	Average amplitude difference
MAX	Sad	Surprise	Angry	Surprise	Happy	Neutral
MIN	Neutral	Sad	Sad	Sad	Sad	Angry

unvoiced sounds. Consonants are clear and voiced, and vowels in most languages are voiced, and nasal, side, and semi-vowels are also voiced. Therefore, as long as we can distinguish unvoiced and voiced sounds, we can roughly distinguish between vowels and consonants. After that, we only need to find a way to mark the position of all the vowels in a sentence, and then we can get the value of the speech rate of a sentence through simple calculations. In the process of short-term time-domain analysis, we know that the energy of the speech signal will change considerably over time, especially since the energy of the unvoiced segment is generally much smaller than that of the voiced segment. For the short-term zero-crossing rate, the zero-crossing rate of unvoiced sounds is higher, and the zero-crossing rate of voiced sounds is lower. For the short-term average amplitude difference function, for voiced sounds, the short-term energy change of the average amplitude difference function is higher, while the short-term energy change of the unvoiced average amplitude difference function is lower. Based on these characteristics, we can basically get the relevant parameters from the speech signal to the speech rate.

Taking short-term energy as an example, we get its basic short-term energy image. According to the energy of voiced sound is much greater than that of unvoiced sound, we know that several peaks are voiced sounds, and then we further filter the waveform to get a smooth curve. The filter first uses the average filter, and then uses the Savitzky-Golay filter to perform secondary filtering on the result. The result as shown in Fig. 7 and a relatively complete and smooth curve is obtained. Integrity refers to the peak value in the original image. In the case of retention, because the peak corresponds to the voiced point, it should be retained and the vibration caused by other factors such as noise should be smoothed. After filtering, we need to extract peak points. We obtain the maximum value by deriving the image, and mark it on the image, as shown in Fig. 8. We found some unnecessary peaks and removed them. The final result as shown in Fig. 9.

At this point, we have obtained the coordinates of the basic voiced point. The abscissa represents the frame number of the point in the sentence, and the ordinate represents its speech energy value. We alternate the ab-

Fig. 7: Comparison chart before and after filtering

Fig. 8: Filtered extreme points

Fig. 9: Extreme point after partial rounding

scissas and divide it by the total number of frames of

Table 3: Different parameter speech rate

Eigenvalues	Energy	Average amplitude	Zero-crossing rate	Average amplitude difference
MAX	Fear	Fear	Fear	Surprise
MIN	Angry	Angry	Happy	Neutral

the sentence to indicate the interval between two voiced sounds, and then calculate the average of all intervals. The obtained value is used as a reference parameter for speech rate and compared. The result as show in Table. 3. According to the results of our calculations, the “fear” in negative emotions is basically the slowest speaking rate of all emotions, and the “angry” in negative emotions speaks faster.

4 Conclusions

We first analyzed the specific expression of speech feature values in different emotions, and based on the performance, the difference in unvoiced and voiced sounds caused by different parameters was determined to determine the vowels and consonants and used this as a reference to determine the speech rate of different emotions. Finally got a rough conclusion. In the later stage, we will expand and optimize the database based on this idea to get more accurate conclusions, and build a new tag database based on the conclusions, so as to lay the foundation for the later deep learning.

References

- (1) K. Wang, N. An, B. N. Li, Y. Zhang and L. Li: “Speech Emotion Recognition Using Fourier Parameters”, in IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 69-75, 1 Jan.-March 2015
- (2) C. Busso, S. Lee and S. Narayanan: ”Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection,” in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 4, pp. 582-596, May 2009
- (3) X. M. Cheng, P. Y. Cheng and L. Zhao: “A Study on Emotional Feature Analysis and Recognition in Speech Signal”, 2009 International Conference on Measuring Technology and Mechatronics Automation, Zhangjiajie, Hunan, 2009, pp. 418-420
- (4) R. Nawaz, H. Nisar and V. V. Yap: “Recognition of Useful Music for Emotion Enhancement Based on