Scholar Publishing Group
Zoology and Animal Physiology
https://doi.org/10.38007/ZAP.2022.030204
ISSN 2789-8369 Vol. 3, Issue 2: 43-55

Animal Application in Oral English Recognition System

Yunfeng Qiu[1], Dengfeng Yao[2*] and Xinchen Kang[2]

[1]Jingchu University of Technology, China

[2]Beijing Union University, Beijing, China

tjtdengfeng@buu.edu.cn

**corresponding author*

Keywords: Animal Language, Spoken English, Recognition System, Language Communication

***Abstract*：** Under the upsurge of artificial intelligence, the development of language recognition system has become inevitable. Now the speech recognition system has been applied to many aspects. However, there have been many difficulties in the recognition of spoken language. The purpose of this paper is to find ways to solve the problems of spoken English recognition system from animals. This paper briefly introduces animal language and spoken language recognition system through literature research and investigation. The influence of animal emotion analysis methods on the accuracy of spoken English recognition was compared through a comparative experiment. Through a questionnaire survey, the support rate of the application scenario of this technology is analyzed. The results show that the accuracy of animal emotion analysis method for male spoken English recognition in fear state has increased by 27%, and for female spoken English recognition in anger state has increased by 29%. At present, the technology corpus has a single language, and the extracted emotional features have great limitations, and there are still many places to be improved. 37% of the people hope that the spoken English recognition system based on animal emotion analysis can be applied to psychological monitoring. The current high-pressure and fast-paced life in society makes many people have psychological problems and their psychological needs will be increasing.

1. Introduction

Language communication is a natural communication mode for human beings. Since childhood, we have been learning languages spontaneously, and language communication has always run through our lives [1]. It is so natural that we have not found what a complicated phenomenon it is. Human vocal tract and vocal organs are biological organs with non-linear characteristics. They not only operate under conscious control, but also are influenced by gender and emotional state of growth factors. As a result, the sound will vary greatly due to their accent, pronunciation, clarity, volume, speed, etc. Human beings hope to further communicate with machines so as to facilitate production and life. During the transmission of voice signals, our irregular language behavior will be further distorted by background noise and echo, as well as electrical characteristics (such as microphones and other electronic devices). All these variable sound sources make speech recognition more complicated [2,3]. Voice communication with the machine to make the machine understand what you are saying has long been a dream of people. Speech recognition technology is a high technology that enables machines to convert speech signals into corresponding text or commands through recognition and understanding processes.

Speech recognition is a multi-layer pattern recognition task. After examining the sound signal, the structure is divided into a hierarchy of root units (e.g. phonemes), words, phrases and sentences [4,5]. Each layer can provide additional time constraints, for example, the pronunciation of words that have already been recognized or legal word sequences, which can make up for errors or reduce uncertainty to a lower level. The best way to limit is to use probabilistic decision-making at all lower levels, and only use discrete decision-making at the highest level. Machines are widely used in various fields such as industry, household appliances, communications, automotive electronics, medical care, home services, consumer electronics and so on [6] by recognizing and understanding high-tech technologies that convert voice signals into corresponding texts or commands. Many experts believe that speech recognition technology is one of the ten most important technology development technologies in the field of information technology from 2000 to 2010. Speech recognition technology mainly includes feature extraction technology, pattern matching criteria and model training technology. Voice recognition technology has also been fully used in vehicle networking. For example, in the wing truck networking, direct navigation of the destination can be set simply by pressing a push-to-talk spoken by customer service personnel, which is safe and convenient [7].

Polap discussed the problem of voice verification, proposed a voice verification method based on artificial intelligence, and conducted a large number of tests to verify the effectiveness of the scheme [8]. The research results are explained and discussed from the advantages and disadvantages of the solution [9]. Migowa studied the influence of speech recognition system on medication errors in pediatric outpatient department of a tertiary medical institution in Kenya, and found that errors in speech recognition system often appear on medication dosage, and incorrect dosage leads to medical accidents [10]. Jemai proposes a speech recognition learning algorithm based on fast wavelet transform (FWT). Compared with other algorithms, this algorithm has many advantages, which solves the main work of calculating connection weights in the past. It is determined by a direct solution and requires the inverse of the calculation matrix. The new algorithm is implemented by FWT iterative computation of connection weights. Secondly, a new classification method for speech recognition system is proposed [11]. Mannepalli took speech samples from native speakers with different accents for training and testing. Mel frequency cestrum coefficient (MFCC) features are extracted for each speech of the training sample and the test sample. Gaussian mixture model (GMM) is used to classify the accent of speech [12]. Nicholas research believes that the quality of speech recognition does not depend on vocabulary, and it is perhaps more important to understand language [13].

In brief, this paper discusses the application of animals in spoken English recognition system. Specifically, the main research content of this paper is roughly divided into five parts: The first part is the introduction part, which aims to make a systematic overview of the main research content of this paper from the aspects of research background, research purposes, research ideas and methods; The second part is the theoretical basis, summarizing the key technologies of spoken language recognition in detail and systematically, and introducing the working principle of spoken language recognition system. The third part is related research, which illustrates the difference between animal language and human language through data query and related experiments. The fourth part is the analysis of the data, through the specific survey data and research results, comparing the accuracy of oral English recognition system before and after using animal emotion analysis method. After animal emotion analysis, the accuracy of spoken English recognition has been greatly improved, including 27% for men in fear and 29% for women in anger. The fifth part is the summary and suggestions of this article, which is the summary of the article's achievements and the prospect of further improving the application of animal emotion analysis method in oral English system.

2. Proposed Method

2.1. Key Techniques of Spoken Language Recognition

The so-called speech recognition is to convert a speech signal into corresponding text information. The system mainly includes four parts: feature extraction, acoustic model, language model, dictionary and decoding. In order to extract features more effectively, the collected voice signals often need to be preprocessed such as filtering and framing, and the signals to be analyzed are extracted from the original signals. After that, the feature extraction converts the sound signal from the time domain to the frequency domain to provide appropriate feature vectors for the acoustic model. In the acoustic model, the score of each feature vector on the acoustic characteristics is calculated according to the acoustic characteristics. The language model calculates the probability that the sound signal corresponds to the sequence of possible phrases according to relevant theories of linguistics. Finally, according to the existing dictionary, the phrase sequence is decoded to obtain the final possible text representation.

(1) Acoustic signal preprocessing

As the premise and foundation of speech recognition, the preprocessing of speech signal is very important. In the final template matching, the feature parameters of the input speech signal are compared with the feature parameters in the template library. Therefore, only when the feature parameters that can characterize the essential features of the speech signal are obtained in the preprocessing stage, can these feature parameters be matched for speech recognition with high recognition rate. First of all, the sound signal needs to be filtered and sampled. This process is mainly to eliminate the interference between signals of frequencies other than human voice and 50Hz current frequency. This process is generally realized by using a band pass filter, setting the upper and lower ring frequencies to filter, and then quantizing the original discrete signals. After that, it is necessary to smooth the connection section between the high-frequency and low-frequency parts of the signal, so that the spectrum can be solved under the same signal-to-noise ratio condition, making the analysis more convenient and faster.

(2) Acoustic feature extraction

After the preprocessing of the signal is completed, the following is the most critical feature extraction operation in the whole process. The recognition of the original waveform cannot achieve good recognition effect. The feature parameters extracted after frequency domain transformation are used for recognition, while the feature parameters that can be used for speech recognition must meet the following requirements: the feature parameters can describe the basic features of speech as much as possible; Minimize the coupling between parameter components and compress the data. The process of calculating characteristic parameters should be made simpler and more efficient. Pitch period, resonance peak and other parameters can be used as characteristic parameters to characterize speech characteristics. At present, the most commonly used characteristic parameters in mainstream research institutions are: Linear Predictive Cestrum Coefficient (LPCC) and Mel Cestrum Coefficient (MFCC). Two kinds of characteristic parameters operate on speech signals in cestrum domain. The former uses the vocal model as the starting point and uses LPC technology to calculate cestrum coefficients. The latter simulates the auditory model, takes the output of the speech passing through the filter bank model as the acoustic feature, and then transforms it using discrete Fourier transform (DFT). The so-called pitch period refers to the vibration period of the vocal cords' vibration frequency (fundamental frequency). Because it can effectively characterize the characteristics of speech signals, pitch period detection is a crucial research point from the initial speech recognition research.

(3) Language model

The language model mainly depicts the way and habit of human language expression, and focuses on the internal relation between words and their arrangement structure. In the process of speech recognition and decoding, transfer the reference pronunciation dictionary within words and the reference language model between words. A good language model can not only improve the decoding efficiency, but also improve the recognition rate to a certain extent. Language models are divided into rule models and statistical models. Statistical language models use probability and statistics methods to describe the internal statistical laws of language units. Its design is simple, practical and has achieved good results. It has been widely used in speech recognition, machine translation, emotion recognition and other fields. The simplest but most commonly used language model is the N-element language model. The N-ray language model assumes that the probability of the current word is only related to the first N-1 words given the above environment.

(4) Decoding and dictionary

Decoder is the core component of the recognition stage. The speech is decoded by the trained model to obtain the most possible word sequence, or the recognition grid is generated according to the recognition intermediate results for subsequent components to process. The core algorithm of decoder is dynamic programming algorithm Viterbi. Due to the huge decoding space, we usually use token passing method with limited search width in practical applications. Traditional decoders generate decoded pictures completely dynamically. Such implementation takes up less memory, but considering the complexity of each component, the whole system process is complicated, which is not convenient and efficient to combine the language model with the acoustic model, and is more difficult to expand. At present, mainstream decoder implementations will use pre-generated finite state converters as pre-loaded static decoding maps to some extent. Here, we can construct four parts of language model (G), vocabulary (L), context-related information (C) and hidden Markov model (H) as standard finite state converters respectively, and then combine them through standard finite state converter operation to construct a converter from context-related phoneme substate to word.

2.2. Working Principle of Spoken Language Recognition System

Sound is actually a wave. Common mp3, wavy and other formats are compressed formats, and must be converted into uncompressed pure waveform files for processing, such as Windows PCM files, also known as wav files. In addition to a file header, the wav file stores points of the sound waveform. Before starting speech recognition, it is sometimes necessary to cut off the mute at the beginning and end to reduce the interference to the following steps. This mute cutting operation is commonly called VAD and requires some techniques of signal processing. To analyze the sound, it is necessary to frame the sound, that is, to cut the sound into small segments, each of which is called a frame. The framing operation is generally not a simple cut, but is implemented using a moving window function, which is not described in detail here. Generally, there is overlap between frames. After framing, speech becomes many small segments. However, the waveform has little ability to describe in time domain, so the waveform must be transformed. A common transformation method is to extract MFCC features. According to the physiological characteristics of human ears, each frame waveform is transformed into a multi-dimensional vector, which can be simply understood as including the content information of the frame speech. This process is called acoustic feature extraction. In practical application, there are many details in this step and MFCC is not the only acoustic feature.

In speech recognition system, the encoding of speech information in speech signals is carried out according to the time variation of amplitude spectrum. Because speech is readable, that is to say, acoustic signals can be expressed by a plurality of distinctive and discrete symbols without considering the information content conveyed by the speaker. Phonetic interaction is a cognitive process, so it must not be separated from grammar, semantics and language norms. Pre-processing, which includes sampling speech signals, overcoming aliasing filtering and removing some noise effects caused by individual pronunciation differences and environment, also takes into account the selection of basic units of speech recognition and endpoint detection. Repeated training is to make the speaker repeat the speech several times before recognition, remove redundant information from the original speech signal samples, retain key information, and then organize the data according to certain rules to form a pattern library. Thirdly, pattern matching, which is the core part of the whole speech recognition system, calculates the similarity between input features and inventory patterns according to certain rules, and then judges the meaning of input speech. In front-end processing, the original speech signal is processed first and then feature extraction is carried out to eliminate the influence brought by noise and pronunciation differences of different speakers, so that the processed signal can more completely reflect the essential feature extraction of speech, eliminate the influence brought by noise and pronunciation differences of different speakers, and enable the processed signal to more completely reflect the essential feature of speech.

2.3. Acoustic Model of Speech Recognition System

Acoustic model is a very important component in speech recognition system. The ability to distinguish different basic units is directly related to the quality of recognition results. Speech recognition is essentially a process of pattern recognition, and the core of pattern recognition is the problem of classifier and classification decision. In general, the dynamic time warping (DTW) classifier used in isolated words and small and medium vocabulary recognition has a good recognition effect, and is a very successful matching algorithm in speech recognition due to its fast recognition speed and low system overhead. However, when large vocabulary and non-specific speech recognition are used, the DTW recognition effect will drop sharply, and then the training recognition effect using Hidden Markov Model (HMM) will be significantly improved. Because the continuous Gaussian mixture model GMM is generally used to characterize the state output density function in traditional speech recognition, it is also called GMM-HMM framework. There are the following mainstream acoustic models.

(1) Gaussian mixture model

Gaussian distribution has strong ability to approximate real world data and is easy to calculate, so it is widely used in various disciplines. However, there are still many types of data that cannot be described by a Gaussian distribution. At this time, we can use a mixture of multiple Gaussian distributions to describe these data, and multiple components are responsible for different potential data sources. At this point, the random variable conforms to the density function. We call the model used to consider the data subject to Gaussian mixture distribution Gaussian mixture model. Gaussian mixture model is widely used in acoustic models of many speech recognition systems. Considering the relatively large dimension of vectors in speech recognition, we usually assume that the covariance matrix in the Gaussian mixture distribution is a diagonal matrix. This not only greatly reduces the number of parameters, but also improves the calculation efficiency. Using

Gaussian mixture model to model short-term feature vectors has the following advantages: First, Gaussian mixture model has strong modeling ability. As long as the total number of components is sufficient, Gaussian mixture model can approximate a probability distribution function with any accuracy; In addition, EM algorithm can easily make the model converge on the training data.

However, Gaussian mixture model also has a serious disadvantage: Gaussian mixture model is very poor in modeling data on a nonlinear manifold near vector space. For example, suppose some data are distributed on both sides of a sphere and very close to the sphere. If a suitable classification model is used, we may need only a few parameters to distinguish the data on both sides of the sphere. However, if we use Gaussian mixture model to describe their actual distribution, we need a lot of Gaussian distribution components to describe them accurately enough. This drives us to find a model that can use voice information more effectively for classification.

(2) Hidden Markey model

Considering a discrete random sequence, if the transition probability conforms to Markov property, i.e. the incoming state and the past state are independent, it is called a Markov chain. If the transition probability is independent of time, it is called homogeneous Mark chain. The output of Markov chain corresponds to predefined states one by one. For any given state, the output is observable and has no randomness. If we expand the output, each state output of Markov chain is a probability distribution function. In this way, the state of the Markov chain cannot be directly observed, and can only be inferred by other variables that conform to the probability distribution and are affected by state changes. We call this model, which uses hidden Markov sequence hypothesis to model data, hidden Markov model.

Corresponding to the speech recognition system, we use hidden Markov model to describe the substate changes within a phoneme to solve the problem of correspondence between feature sequences and multiple speech basic units. Using hidden Markov model in speech recognition task needs to calculate the possibility of the model on a speech segment. In training, we need to use Baum-Welch algorithm to learn the parameters of hidden Markov model and make maximum likelihood estimation. Baum-Welch algorithm is a special case of EM algorithm. The E step of calculating conditional expectation and the M step of maximizing conditional expectation are carried out iteratively and sequentially by using the probability information of the preceding and following terms.

(3) Artificial neural network

Artificial neural network was put forward in the late 1980s. Its essence is an adaptive nonlinear dynamic system based on biological neural system. It aims to fully simulate the way the neural system performs tasks. Like the human brain, a neural network is made up of interconnected neurons that influence each other's behavior. These neurons are also called nodes or processing units. Neural network imitates the activity of human neurons through a large number of nodes, and connects all nodes into an information processing system to reflect the basic characteristics of human brain function. Although ANN is accurate in simulating and abstracting human brain functions, it is, after all, an artificial neural network and is only a distributed parallel processing model that simulates biological sensing characteristics. The unique advantages of ANN and its powerful classification and input-output mapping capabilities have led to its wide application in many fields, especially in the fields of speech recognition, image processing, fingerprint recognition, computer intelligent control and expert systems. However, from the current speech recognition system, due to ANN's insufficient description of the time dynamic characteristics of speech signals, most of them use ANN combined with traditional recognition algorithms.

3. Experiments

3.1. Experimental Content

At present, speech recognition has many applications, but there are also many problems. The experiment in this paper mainly discusses the difficulties of oral English recognition system and tries to find inspiration from animals to solve these problems. To seek inspiration from animals, we must understand the connection and difference between animal language and human language. For a long time, whether there is language has been regarded as the main difference between human beings and animals. However, the distinction between language as human and animal is based on the concept of human vocal language, and this view still has some defects. With the concept of large language (language is the medium of information exchange between the two bodies), we can understand the world from a universal point of view, that is, both human beings and animals have languages, and there are striking similarities between human beings and animals in the process of using the medium to transmit information (language communication). First of all, both humans and animals use media to express meaning. Secondly, the basic form of media, that is, expression, is often an application of one's own sensory organs, or through five sensory forms such as hearing, vision, touch and smell. Thirdly, there is an imbalance in the use of media by both humans and animals. Both will choose one or both as the main means.

In a word, at the big language level, there are indeed great similarities in language between human beings and animals. It is more meaningful to make clear the difference between human and animal than to realize the connection between them. Because distinguishing the two can help us have a deeper understanding of human language. First, language is not the patent of human beings. Animals also have languages. Language is not the difference between human and animals. If we insist on finding boundary markers between human and animal languages, it is that human languages are more detailed and diverse in content than animal languages, and human beings are less meticulous and processed in form than animals. Second, audio language is only one form of media that transmits information, but it is not the only form. It is not more advanced than other media, but only the result of selection under the influence of content and the limitation of environment.

Animal expression has a direct practical purpose, while human beings also practice cognitive and aesthetic functions in the process of expression. The process of human mastering language is the process of knowing the world and acquiring knowledge, and it is also the process of improving aesthetic ability. Therefore, under the concept of large language, the difference between animals and humans is only a matter of degree, and there is no obvious boundary. However, if we must draw a clear line between the two, then we can only say that human beings are more precise and complicated in their processing of media in form than animals, and their content is more humanized and spiritual.

However, the simplification of animal language can help people to trace back the use and emotion of language. Take cats for example. Some people think that cats can only communicate with each other with purrs, meows and roars. In fact, cats have many ways of communication. They can communicate with sounds, movements, ears, mouth and tail. Besides, they also have rich vocabulary. This is especially true for the mother cat, who needs to communicate with Xiao Mao using various sounds and to teach and punish Xiao Mao. Cats purr when they are happy or satisfied, and the purrs made by different cats vary greatly. We can easily distinguish the purr from the shrill roar. The roar was made when the invading cat or dog resisted. Cats meow mainly to attract attention. If the meow is short and loud, it means the cat is looking for its owner in the house. If the meow is persistent and loud, it means the cat wants to open a door or eat. However, if a cat is locked in a travel basket or stepped on its tail, it will make a loud cry of anger.

The female cat will make a blare sound in her heat to attract the male cat. During lactation, a female cat will make a low purr to people or animals close to Xiao Mao, usually within a range of one stroke (about 0.8 ~ 1 meter), only making sound. Less than the step distance (0.5 meters), it will accompany the head to lift or rotate. When reaching the limb distance (0.3 meters), it will bare its teeth or dance until it launches the attack. Xiao Mao, who is looking for her mother, will open her mouth and make a rapid meow. Xiao Mao, who is looking for her mother, will softly make a slender meow to indicate that she needs to cuddle up. When the cat sits on the windowsill and watches the birds outside, it will make a startling noise between meow and purr.

The tone, loudness and timbre of animal language are different according to needs and emotional states. We try to use these characteristics to solve the current obstacles of spoken English recognition system.

3.2. Experimental Results

At present, the difficulties in oral English recognition are as follows: accent problem, nonstandard pronunciation and slurred speech. The speed is too fast and the pronunciation of words is stuck. Noise echo interference; Unable to understand the meaning of language. Can't understand language and emotion. We know from animals that under different conditions, the sound emitted by animals has different tones, loudness and sound quality. Integrating these factors into spoken English recognition can improve the accuracy of spoken English recognition. We have carried out relevant experiments, and the data are shown in Table 1.

Table 1. Accuracy of Oral English Recognition

Group	Analysis Mode	Gender	Happy	Sad	Anger	Fear
A	No Emotion	Male	70%	59%	51%	35%
B	No Emotion	Female	65%	55%	39%	20%
C	Contain Emotion	Male	81%	79%	72%	62%
D	Contain Emotion	Female	87%	65%	68%	38%

4. Discussion

4.1. Analysis of Influence of Animal Emotion on Recognition Accuracy

People have four basic emotions: happiness, anger, fear and sadness. Happiness is a satisfying experience when one pursues and achieves one's goal. Anger is the experience when people cannot achieve their goals due to interference. Fear is the experience when trying to get rid of or escape from a certain dangerous situation. Sadness is the experience of losing one's wish or not realizing one's ideal. On top of the above four basic emotions, many complicated emotions can be derived, such as dislike, jealousy, liking, sympathy, accusation, etc.

Under the action of these four emotions, it is bound to interfere with human language. First, we observe the accuracy of the spoken English recognition system without animal emotion analysis and draw it into a bar graph, as shown in Figure 1.

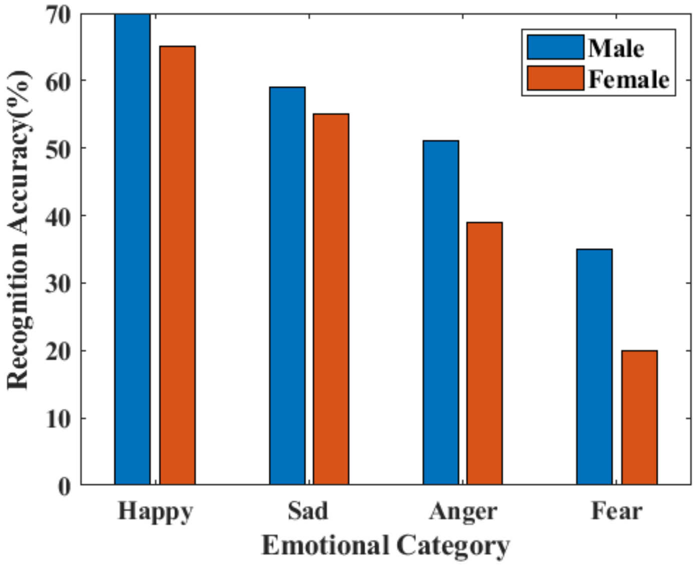

Figure 1. Accuracy of Emotional Free Oral English Recognition

As can be seen from Figure 1, the speech recognition accuracy rate is highest in happy state, followed by sad state, while the speech recognition ability is extremely weak in fear state, which may be related to discontinuous speech and weak voice. We can understand that the prosodic features of sound have changed. Prosodic features refer to a kind of phonetic features contained in speech but different from semantic content. It is embodied in the volume, the length of pronunciation, the speed of speech speed, the weight of tone and so on, which determines the cadence of speech sound and is a structural arrangement and supplement to speech expression. Whether it exists or not does not affect our listening and distinguishing of words, words and sentences, but it is closely related to the emotion contained in pronunciation. For example, when people are angry, they speak faster, with higher volume and heavier tone. When sad, the tone is low, the speed is slow and the volume is low. After animal emotion analysis, the accuracy rate is shown in Figure 2.

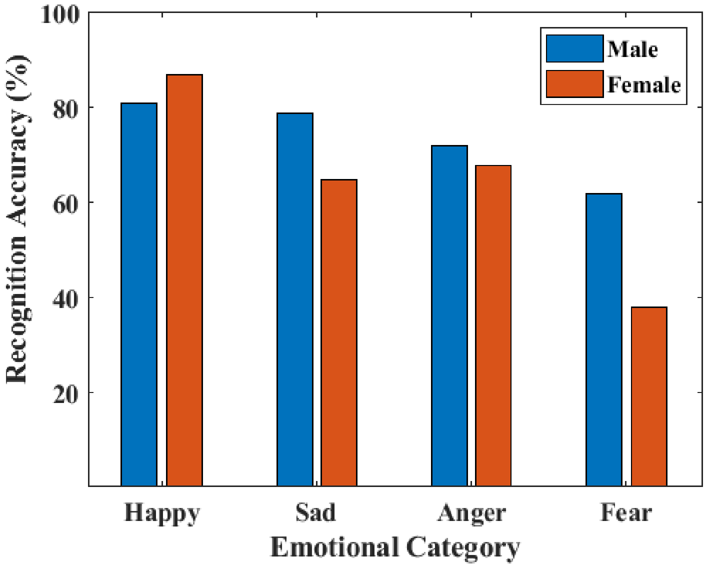

Figure 2. Animal Emotion Analysis Accuracy of Oral English Recognition

As can be seen from the data in Figure 2, after animal emotion analysis, the accuracy of spoken

English recognition has been greatly improved. We found that in a happy state, women's oral English recognition accuracy was higher than men. The difference in the accuracy of spoken English recognition between men and women may be due to the acoustic characteristics. Tone quality features are a kind of voice features used to reflect whether the speaker's voice is clear, pure and easy to recognize. Under different emotional states, people's sound quality will vary greatly. The concrete manifestation is: with emotional fluctuations, people will involuntarily produce gasps, vibrato, sobs, etc. However, under different emotional states, these acoustic performances are different. Therefore, the changes of sound quality contain rich emotional information, and the extraction of sound quality features is conducive to the recognition of voice emotions. In speech emotion recognition, the timbre features used to measure the sound quality generally include formants, respiratory laryngitis and glottis parameters. Literature research shows that there is a great correlation between speech emotion and tone quality characteristics. In order to explore the specific influence of animal emotion analysis on the accuracy of spoken English recognition. We plot the increase in accuracy into a column chart, as shown in Figure 3.

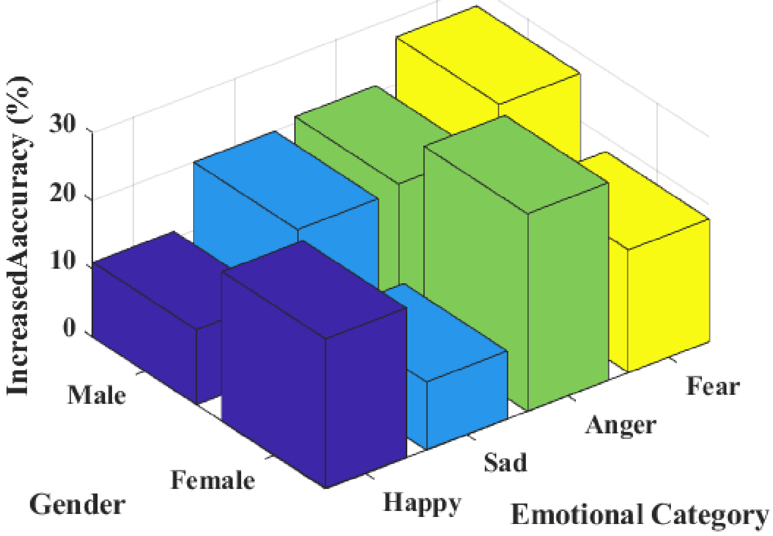

Figure 3. Animal Emotion Analysis for Accuracy Improvement

From the data in Figure 3, it can be seen that the accuracy of spoken English recognition has been greatly improved after animal emotion analysis, in which the accuracy of male recognition in fear state has been increased by 27%, and that of female spoken English recognition in anger state has been increased by 29%. It is possible that in these two states, the speech characteristics of men and women vary greatly, and the accuracy rate can be improved by animal emotion analysis.

4.2. Application Analysis of Animal Emotion Recognition System in Oral English

Speech recognition technology not only brings more functions and applications, but also, more importantly, speech, as a way of communication rich in human emotion, will project this emotion onto the human-computer relationship. Our enthusiasm and pursuit of artificial intelligence not only lies in its ability to liberate us from certain jobs, but also in our reverence for cognitive computing and emotional intelligence, as well as for speech. The spoken English recognition system of animal emotion analysis may change our communication with users.

After research, the spoken English recognition system based on animal emotion analysis is likely to be applied in service industry, smart home, smart customer service, manpower recruitment and psychological monitoring. In order to study people's needs for the application of spoken English recognition system for animal emotion analysis, a questionnaire survey was conducted. There are totally 200 questionnaires, each of which can choose 2 items. The survey results are shown in Table 2.

Table 2. Application Requirements of Oral English Emotion Recognition System

Application Scenario	Service Industry	Smart Home	Intelligent Customer Service	Manpower Recruitment	Psychological Monitoring
Number of Votes	32	16	26	52	74

Each profession has different emotional demands in communication, but it basically contains the emotion of the other party and the information revealed by his/her speech, with the emphasis on diagnosis-emotional and emotional information. We can diagnose emotions and emotions by looking at expressions and movements, listening to language and intonation, and feeling each other's feelings. Do first deal with emotions, in dealing with things. In the final analysis, users' needs are all emotional needs. In the service scenario, understanding the user's emotion can greatly improve the success rate of sales and customer satisfaction. In order to better understand the support rate of various application scenarios, we will draw the relevant data into a pie chart, as shown in Figure 4.

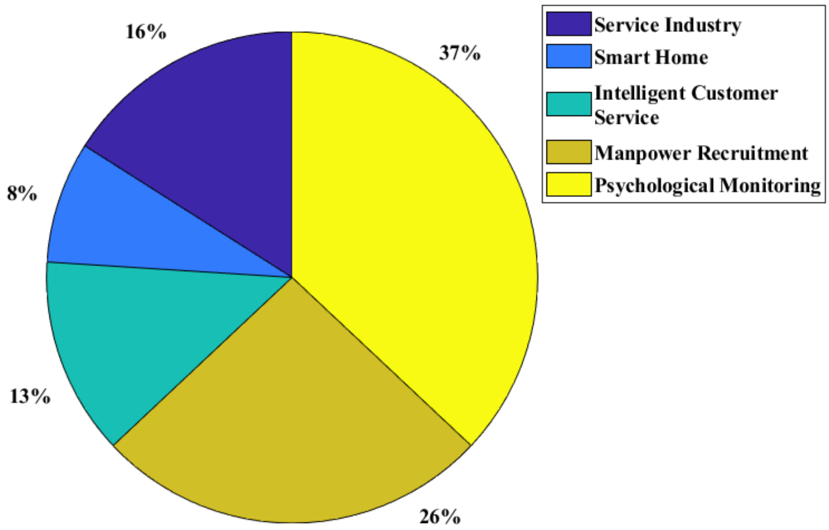

Figure 4. Application Scenario Approval Rate

As can be seen from the above figure, 37% of the people hope that the spoken English recognition system based on animal emotion analysis can be applied to psychological monitoring. At present, the high social pressure and fast pace of life make many people have psychological problems and their psychological needs will be increasing. 26% of the people want to apply this technology to human resource recruitment, and 16% want to use the spoken English recognition system based on animal emotion analysis in the service industry.

5. Conclusions

(1) The introduction of animal language and spoken language recognition system, the purpose and significance of the research and the current research situation. The so-called speech recognition

is to convert a speech signal into corresponding text information. The system mainly includes four parts: feature extraction, acoustic model, language model, dictionary and decoding. The acoustic models of spoken language recognition system mainly include: Gaussian mixture model, hidden Markov model and artificial neural network.

(2) Through literature research, the difference between animal language and human language is analyzed. Language is not the patent of human beings. Animals also have languages. Language is not the difference between human and animals. If we insist on finding boundary markers between human and animal languages, it is that human languages are more detailed and diverse in content than animal languages, and human beings are less meticulous and processed in form than animals. Voice language is only one form of media for transmitting information, but it is not the only form. It is not more advanced than other media, but only the result of selection under the influence of content and the limitation of environment.

(3) Experiments and data analysis show that the accuracy of spoken English recognition has been greatly improved after animal emotion analysis. Among them, the accuracy of male recognition in fear state has been improved by 27%, and that of female oral English recognition in anger state has been improved by 29%. At present, animal emotion analysis methods still have many limitations to be improved. 37% of the people hope that the spoken English recognition system based on animal emotion analysis can be applied to psychological monitoring. The current high-pressure and fast-paced life in society makes many people have psychological problems and their psychological needs will be increasing.

References

[1] *Yi, MoungHo; Lim, MyungJin; Ko, Hoon(2021) Method of Profanity Detection Using Word Embedding and LSTM, Mobile Information Systems,6654029 https://doi.org/10.1155/2021/6654029*

[2] *Yawen Xue, Yasuhiro Hamada, & Masato Akagi. (2016). "Emotional Voice Conversion System for Multiple Languages Based on Three-layered Model in Dimensional Space", Journal of the Acoustical Society of America, 140(4), pp.2960-2960. https://doi.org/10.1121/1.4969141*

[3] *Soumen Kanrar, & Prasenjit Kumar Mandal. (2015). "Detect Mimicry by Enhancing the Speaker Recognition System", Advances in Intelligent Systems and Computing, 339(1), pp.21-31. https://doi.org/10.1007/978-81-322-2250-7_3*

[4] *Jan Vanus, Marek Smolon, Jiri Koziorek, & Radek Martinek. (2015). "Voice Control of Technical Functions in Smart Home with Knx Technology", Lecture Notes in Electrical Engineering, 330(1), pp.455-462. https://doi.org/10.1007/978-3-662-45402-2_68*

[5] *Ibrahim El-Henawy , Marwa Abo-Elazm, (2020). Handling within-word and cross-word pronunciation variation for Arabic speech recognition (knowledge-based approach), Journal of Intelligent Systems and Internet of Things, 1(2), pp.72-79 https://doi.org/10.54216/JISIoT.010202*

[6] *Washani, N. , & Sharma, S. . (2015). "Speech Recognition System: A Review", International Journal of Computer Applications, 115(18), pp.7-10. https://doi.org/10.5120/20249-2617*

[7] *Saksamudre, S. K. , Shrishrimal, P. P. , & Deshmukh, R. R. . (2015). "A Review on Different Approaches for Speech Recognition System", International Journal of Computer Applications, 115(22), pp.23-28. https://doi.org/10.5120/20284-2839*

[8] *Cao, J., van Veen, E. M., Peek, N., Renehan, A. G., & Ananiadou, S. (2021). EPICURE: Ensemble Pretrained Models for Extracting Cancer Mutations from Literature. In 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), pp. 461-467. https://doi.org/10.1109/CBMS52027.2021.00054*

[9] Polap, Dawid, Wozniak, Marcin. (2019). “Voice Recognition by Neuro-heuristic Method”, Tsinghua Science and Technology, 24(1), pp.9-17. https://doi.org/10.26599/TST.2018.9010066

[10] Migowa, A. N. , Macharia, W. M. , Pauline, S. , John, T. , & Keter, A. K. . (2018). “Effect of a Voice Recognition System on Pediatric Outpatient Medication Errors at a Tertiary Healthcare Facility in Kenya”, Therapeutic Advances in Drug Safety, 9(9), pp.499-508. https://doi.org/10.1177/2042098618781520

[11] Jemai, O., Ejbali, R., Zaied, M., & Amar, C. B. (2015). “A Speech Recognition System Based on Hybrid Wavelet Network Including a Fuzzy Decision Support System”, International Society for Optics and Photonics, 9445(2), pp.944503. https://doi.org/10.1117/12.2180554

[12] Mannepalli, K. , Sastry, P. , & Suman, M. . (2015). “Mfcc-gmm Based Accent Recognition System for Telugu Speech Signals”, International Journal of Speech Technology, 19(1), pp.1-7. https://doi.org/10.1007/s10772-015-9328-y

[13] Nicholas R. Monto, Rachel M. Theodore, Adriel J. Orena, & Linda Polka. (2016). “The Native Language Benefit for Voice Recognition is not Contingent on Lexical Access”, Journal of the Acoustical Society of America, 140(4), pp.3227-3227. https://doi.org/10.1121/1.4970196