Speech recognition for companion robots

Zhuoyan Zheng

College of William and Mary, Williamsburg, United States of America, 23185.

alyssa510@126.com

Abstract. This paper is based on the goal of solving a global problem — the aging problem. Part of the recognition process is completed by using the Hidden Markov Model and the N-gram model as two main models. This paper clarifies the mechanism of speech recognition and develops a project to achieve simple speech recognition using an interdisciplinary research approach. The results show that the basic recognition of speech content can be achieved, especially when it comes to recognizing male voices. The results of this study suggest that further research and development is needed to identify female voices or higher-pitched voices.

Keywords: speech recognition, HMM, N-gram model, python

1. Introduction

Nowadays, the aging problem in the population is becoming more and more serious around the world. The author hopes that people can solve this problem by using artificial intelligence. In China, people can often see news like this: A ninety-year-old man fell down from his bed and died because nobody else was at home and he missed the best time for treatment. The reason for the frequent hearing of this kind of news is that China has the largest number of elderly people. In 2020, the year of the seventh census, there were 264.02 million people in China who were 60 or older, making up 18.70% of the total population. There were also 190.64 million people who were 65 or older, making up 13.50% of the total population. The proportion of people over 60 grew by 5.44% in 2011 compared to 2010[1]. To deal with this problem, people not only need to implement a birth encouragement plan but also have to find ways of taking care of the elderly. One of the solutions is to use modern techniques. By applying artificial intelligence in our daily lives, the old may gain a more convenient life[2]. The passage mainly uses literature, systematic research methods, and interdisciplinary research methods to solve the use of speech recognition in our daily lives. This paper explains the mechanism of speech recognition and develops a project to realize simple speech recognition using interdisciplinary research methods.

2. Project

2.1. Speech recognition

To make a machine understand what people are talking about, a typical interdisciplinary task is needed[3]. Firstly, by using the signal model, scientists can acquire the basis of a theoretical description for a signal processing system. For example, people can remove the noise when they are transmitting their phonetic sources. Secondly, the signal model can assist scientists in better understanding the audio source that they supplied to the computer. Last but not the least, the reason that the signal model is so important is that it can work perfectly well in practice and help people realize practical systems effectively[4, 5, 6].

2.2. Hidden Markov Model(HMM)

Since the 1980s, statistical model-based methods represented by hidden Markov model(HMM) methods(which can be also called Markov sources or functions of Markov chains) have gradually take over a dominant position in speech recognition research. While using HMM, scientists exploit short-term stationary properties of speech signals. To recognize speech requires people to split it into several different parts. First of all, scientists collect the wave of a sound and then turn it into a silent utterance. They try to match them with real words used in our daily lives. The concept of features is quite important in this process[7, 8]. An input speech stream is transformed into a series of feature vectors after feature extraction. Sequence labeling is a problem that voice recognition systems must solve in order to determine the appropriate word sequence based on the feature vector sequence[9]. For ease of understanding, the author first explains what HMM is through a general description.

HMM consists of the following components:

1. Q =$q_1q_2...q_N$(HMM hidden state set)
2. A=$a_{01}a_{02}...a_{n1}...a_{nn}$(state transition probability matrix)
3. O=$o_1o_2...o_N$(observations, each from the dictionary V= $v_1,v_2,...,v_V$)
4. B=bi(ot)

The set of likelihoods of observations, is also known as emission probabilities. Each represents the probability that the observations ot are generated by state i at time t.

5.q_0, q_{end}(start state and end state)

In speech recognition, the hidden state Q of HMM is phone or sub-phone word. To simplify the problem, the author assumes that the feature is only divided into two values. So the problem is: Given the observation sequence O, guess the most “probable” state sequence. The passage models this feature by using Hidden Markov Mode(HMM). HMM is a generative model, assumes the model has a sequence of hidden states, but cannot see it. What people see is a sequence of observations. HMM will have two probabilities to use the two important information mentioned above: the jump probability of the hidden state and the emission probability. The transition probability of a state refers to the probability of jumping from one hidden state to another hidden state[10, 11].

The purpose of the HMM model is to estimate the “most likely” model(that is, these two probabilities) based on the sequence of observations. So what is the most likely model? There are infinite(or just many) possibilities for both of these probabilities. But once the model is determined, people can calculate the probability of seeing a sequence under this model(the calculation method will be introduced later). If there is one(or more) model that maximizes this probability, then the result can be considered “optimal”. In theory, people can “traverse” all possible models and find the “optimal” model, but in reality, the possible model parameters are infinite, so a smarter way is needed to find this optimal model. The output probability of the HMM model is calculated by the Viterbi algorithm, which is the commonly used algorithm for finding the optimal path. The viterbi algorithm is a frame synchronization algorithm[12]. The path calculated for each frame is kept in the corresponding register (that is, the tack structure in this program), and each path must continue to be extended in the next frame. The path of the model exit state will be expanded into two, and the expansion of the path in the exit state will be more complicated, depending on the upper-level grammar. It depends on the number of candidate transfer paths provided. Therefore, the number of paths corresponding to each frame increases exponentially, which is an important reason for the excessive computational burden. In fact, most of the paths have extremely low scores and are not competitive in subsequent calculations, so they can be completely excluded. If each frame is processed by removing most of the lower scoring paths, a significant amount of computation can be saved. Because the probability value is generally much smaller than 1, so the probability after taking the logarithm is used as the output value[13, 14].

2.3. N-gram model

During the decoding process, it may reach hundreds of thousands of words, which will be very complicated, and the recognition results will not be ideal due to too many combinations of recognition results. Therefore, only the acoustic model is not enough, and the language model needs to be introduced to constrain the recognition results. The language model is used for calculating a sentence appearing probability, and it is also used for judging whether a sentence is reasonable. The Markov hypothesis is used to address the issue of having too many free parameters; it states that the likelihood of a random word arriving depends only on the n-1 words that come before it[15]. N-gram language models are statistical language models built on the aforementioned premises. In the statistical language model, the probability P(word sequence) is used to gauge how closely a word sequence adheres to linguistic conventions. The word sequence search path should be abandoned more during the decoding process the lower the P(word sequence) is.

Here is the P (word sequence) calculation formula:

Unigram:

$$p(s) = p(w_1)p(w_2)...p(w_n)$$
$$=\prod_{i=1}^{n} p(w_i \backslash w_i - 1) \tag{1}$$

Bigram:

$$p(s) = p(w_1)\ p\ (w_2|w_1)\ p\ (w_3|w_2)\ ...\ p\ (w_n|w_{n-1})$$
$$=p\ (w_1)\ \prod_{1=2}^{n} p(w_i \backslash w_{i-1}) \tag{2}$$

Trigram:

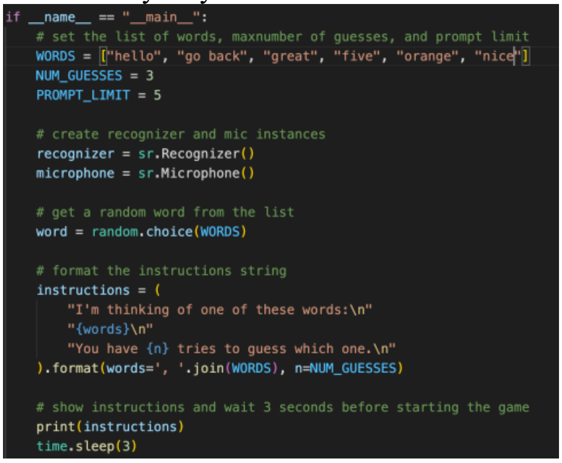
$$p(s)=p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)\ ...p(w_n|w_{n-2},w_{n-1})$$
$$=p(w_1)\ p\ (w_2|w_1)\prod_{i=3}^{n} p(w_i \backslash w_{i-2}, w_{i-1}) \tag{3}$$

Unigram is used when we consider that words are independent, or that each word depends only on itself, then the objective function can be rewritten as this. Bigram means each word depends on its previous word. The function can be written as this one. Similarly, the trigram can be written as this function. Usually, when N is bigger than 2, we call them N-gram model. From the effect of the model, theoretically, if the value of n is larger, the effect would be better. The degree of the effect improvement reduces as n increases in value, though, in the mean time. It also has a distinguishability and dependability issue. The distinguishability would be better if there are more parameters, but meanwhile,if a single parameter has fewer instances, which means the reduce of the reliability[16].

The maximum likelihood algorithm is used to estimate the probability of the language model using a large amount of text corpus. However, the number of statistical expectations is limited, so there will be sparse data, which will lead to zero probability or inaccurate estimation. For word sequences that do not appear or appear in a small amount as expected, use smoothing techniques for indirect prediction. There are three main types of smoothing techniques, including the discount method, interpolation method, and the back-off method. The discount method is to reduce the probability of non-zero items from the probability of the existing observed value to allocate some probability to the unobserved value for increasing the probability of the 0 items. But this does not consider the relationship between the low-order model and the high-order model, so it can not be used alone. The interpolation method is to linearly combine the high-order model and the low-order model, making full use of the high-order and low-order language models, and assign the high-order probability information to the low-order model. The fallback method is to estimate the unobserved high-order model based on the low-order model[17].

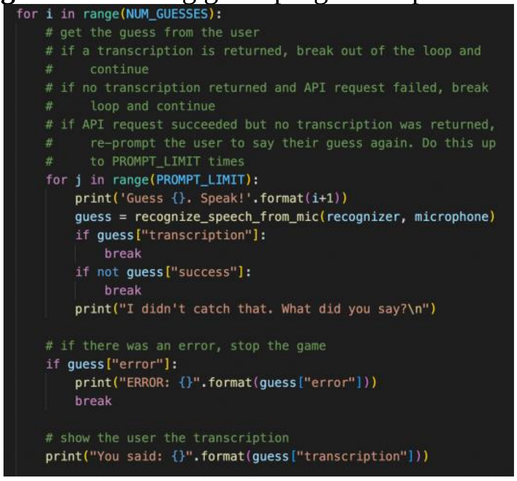
2.4. Guessing game

Pocketsphnix is a lightweight recognizer library written in C, developed by CMUSphnix. By installing this toolkit, we can simply realize some speech recognition. By using PyAudio, people can use Python to play and record audio on many different platforms with ease. Pyaudio helps the computer get people’s speech and give feedback together with the help of pocketsphnix. Using sphnix and pyaudio, a guessing game is designed. Figure 1 shows the introduction word of the guessing game program, and Figure 2 shows the loop design of the guessing game program. By giving six words to the computer and then gaining three times for guessing the word that the computer chose, people can know how recognition works and the problems they may encounter in the future.



```
if __name__ == "__main__":
    # set the list of words, maxnumber of guesses, and prompt limit
    WORDS = ["hello", "go back", "great", "five", "orange", "nice"]
    NUM_GUESSES = 3
    PROMPT_LIMIT = 5

    # create recognizer and mic instances
    recognizer = sr.Recognizer()
    microphone = sr.Microphone()

    # get a random word from the list
    word = random.choice(WORDS)

    # format the instructions string
    instructions = (
        "I'm thinking of one of these words:\n"
        "{words}\n"
        "You have {n} tries to guess which one.\n"
    ).format(words=', '.join(WORDS), n=NUM_GUESSES)

    # show instructions and wait 3 seconds before starting the game
    print(instructions)
    time.sleep(3)
```

Figure1. Guessing game program import words.



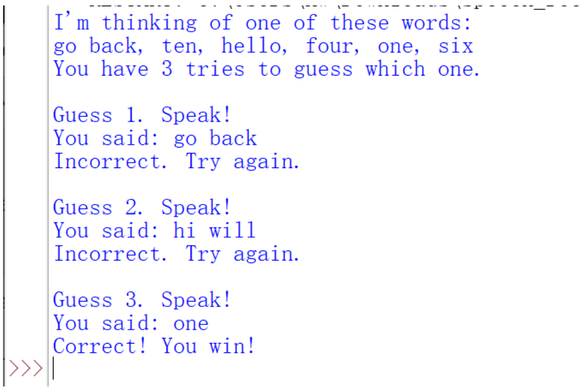
```
for i in range(NUM_GUESSES):
    # get the guess from the user
    # if a transcription is returned, break out of the loop and
    #     continue
    # if no transcription returned and API request failed, break
    #     loop and continue
    # if API request succeeded but no transcription was returned,
    #     re-prompt the user to say their guess again. Do this up
    #     to PROMPT_LIMIT times
    for j in range(PROMPT_LIMIT):
        print('Guess {}. Speak!'.format(i+1))
        guess = recognize_speech_from_mic(recognizer, microphone)
        if guess["transcription"]:
            break
        if not guess["success"]:
            break
        print("I didn't catch that. What did you say?\n")

    # if there was an error, stop the game
    if guess["error"]:
        print("ERROR: {}".format(guess["error"]))
        break

    # show the user the transcription
    print("You said: {}".format(guess["transcription"]))
```

Figure 2. Guessing game program loop design.

3. Result

The result is shown below.

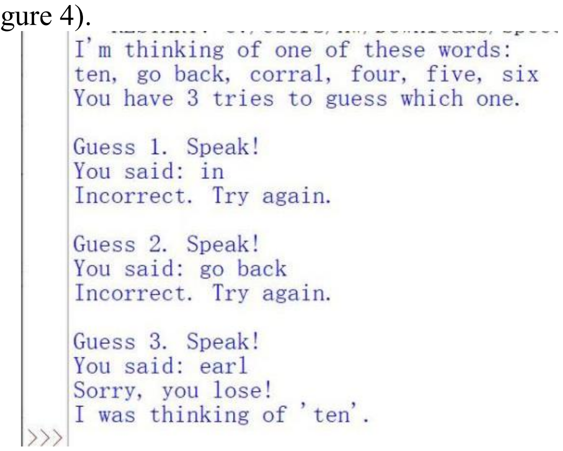
Condition 1 is the computer successfully recognized the word the author said and the author won the game(as is shown in Figure 3.)



```
I'm thinking of one of these words:
go back, ten, hello, four, one, six
You have 3 tries to guess which one.

Guess 1. Speak!
You said: go back
Incorrect. Try again.

Guess 2. Speak!
You said: hi will
Incorrect. Try again.

Guess 3. Speak!
You said: one
Correct! You win!
>>>
```

Figure 3. The computer recognizing successfully.

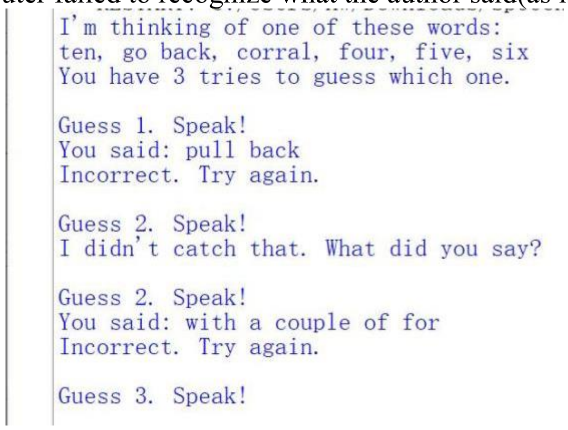
Condition 2 is the computer successfully recognized the word the author said and the author lost the game(as is shown in Figure 4).



```
I'm thinking of one of these words:
ten, go back, corral, four, five, six
You have 3 tries to guess which one.

Guess 1. Speak!
You said: in
Incorrect. Try again.

Guess 2. Speak!
You said: go back
Incorrect. Try again.

Guess 3. Speak!
You said: earl
Sorry, you lose!
I was thinking of 'ten'.
>>>
```

Figure4. Losing game.

Condition 3 is the computer failed to recognize what the author said(as is shown in Figure 5).



```
I'm thinking of one of these words:
ten, go back, corral, four, five, six
You have 3 tries to guess which one.

Guess 1. Speak!
You said: pull back
Incorrect. Try again.

Guess 2. Speak!
I didn't catch that. What did you say?

Guess 2. Speak!
You said: with a couple of for
Incorrect. Try again.

Guess 3. Speak!
```

Figure 5. Recognition fails.

4. Discussion

During the test for the guessing game, the author found that the feedback from the computer differs from one's anticipation. When the author says the same words with different tones, the computer recognizes them as different words. Men and women's voices also get different results. Next breakthrough in speech recognition may be emotional recognition. Just return to the beginning of the report, designing a companion nursing robot is a good entry point and research direction. Short talk and conversation are the basic and most important skills of nursing robots. Considering that the thinking logic and articulation abilities of the elderly are not as clear and coherent as those of the young, in such a complex situation, it is necessary to capture more accurate meanings. Traditional speech recognition and dialogue generation should be optimized and improved, and some additional personalization functions should be added, such as step counting, wake-up and alarm functions, etc., to assist the elderly[18].

5. Conclusion

This paper clarifies the mechanism of speech recognition and develops a project to achieve simple speech recognition using an interdisciplinary research approach. The results show that the basic recognition of speech content can be achieved, especially when comes to recognizing male voices. The results of this study suggest that further research and development is needed to identify female voices or higher-pitched voices.

Acknowledgment

I have received a lot of help and support while writing this dissertation. I want to start by expressing my gratitude to my supervisor, whose knowledge was extremely helpful to me in developing my research topics and methodologies. Your astute criticism encouraged me to sharpen my reasoning and raised the caliber of my work. I also want to express my gratitude to Yuhang Ming, my tutor, for his invaluable advice during my academic career. You gave me the resources I required to pick the appropriate study topic and specifics and successfully complete my report. I also want to thank my parents for making sensible choices and listening to me with compassion. You are by my side at all times.

References

[1] National Bureau of Statistics http://www.stats.gov.cn/ztjc/zdtjgz/zgrkpc/dqcrkpc/ggl/202105/t20210519_1817702.html

[2] An Qi. 2022. Research on the Design of Home Elderly Care Robot Based on Demand Hierarchy[J]. Industrial design, (5):119-121.

[3] Wang Haikun, Pan jia, Liu Cong. Research on design of household elderly escrot robots based on demand level, iFLYTEK Co., Ltd. Artificial Intelligence Research Institute, Anhui, Hefei, 230088, TP393, A , doi: 10.11959/j.issn.1000-0801.20

[4] Hu, H. (2010). Nonlinear discriminant analysis based feature dimensionality reduction for automatic speech recognition. State University of New York at Binghamton.

[5] Lin Shengye. 2019. Speech recognition technology[J]. Peak data science, (4):182-183.

[6] Yu, W., Zeiler, S., & Kolossa, D. (2022). Reliability-Based Large-Vocabulary Audio-Visual Speech Recognition. Sensors, 22(15), 5501.

[7] Zhao Li, Zou Cairong, Wu Zhenyang. 2002. A HMM Speech Recognition Method Introducing Inter-frame Correlation Information[J]. Journal of electronics and information technology, 23(4):327-331.

[8] Xia Jun, Zhou Xiangzhen, Sui dong. 2022. Research on Machine Translation Method Based on Cyclic Generation Countermeasure Network[J]. Journal of Nanjing Normal University(Natural Science Edition), 45(1):104-109.

[9] Duan Qingqing. 2021. Visual analysis of GRAPHEME-PHONEME correspondence in English based on HMM algorithm[J]. Computer Applications and Software, 38(8):44-50.

[10] Zhang Jingxuan. 2021. Research on sequence-to-sequence acoustic modeling for speech generation[D]. Anhui:University of Science and Technology of China.

[11] Huang Xiaoqi, Fan Sheng, Chen Wenguang, et.al,. 2021. Research on intelligent speech interaction algorithm based on Viterbi decoding technology[J]. Electronic Design Engineering, 29(10):37-41.

[12] MUHIB ULLAH. 2020. Automatic Speech Recognition[D]. Guangzhou: South China University of Technology.

[13] Zhu Xiang. 2020. Hybrid English Speech Recognition Algorithm Based on HMM and Cluster[J]. Computer Measurement and Control, 28(5):175-179.

[14] Yu Xiaoming, Bai Song. 2009. Research on speech recognition system based on forward and backward HMM[J]. Computer Engineering and Design, 30(18):4339-4341.

[15] Yin Chen, Wu Min. 2018. Survey on N-gram Model[J]. Computer Systems & Applications, 27(10):33-38.

[16] Xu Hongkui, Lu Jiangkun, Zhang Zifeng, et.al, 2022. Chinese Speech Recognition Based on Conformer and N-gram[J]. Computer Systems & Applications, 31(7):194-202.

[17] Lai Dedi, Luo Zhihui, Ma Yinglong. 2021. Label order optimization method of classifier chains based onCO—occurrence analysis[J]. Systems Engineering and Electronics, 43(9), 9.

[18] Yang Jiabing. 2021. Research on Old People's Chatting Robot Based on Deep Learning[D]. Guangdong University of Technology.