Evaluation of Intelligent Speech Technology in Epidemic Prevention: Take Iflytek Input Software in Chinese and Japanese Recognition as an Example

Chen Cao and Gaowu Wang*
Institute of Linguistics and Applied Linguistics, Beijing Normal University, Beijing, China
Email: wgw@bnu.edu.cn

Abstract. This study aims to investigate public’s understanding of intelligent speech technology and to evaluate the word accuracy rate of speech recognition technology in Japanese and Chinese recognition and translation, using the simulating corpus of doctor-patient communication. It is concluded from the questionnaire that currently the public has a high awareness and demand of intelligent speech products. The results of speech recognition experiments show that the word accuracy rates are 99.4% in Mandarin, 96.1% in Japanese, 94.2% in Japanese translation of Chinese, and 82.7% in Chinese translation of Japanese, respectively. In the recognition of Mandarin and Japanese, the errors mainly include missing words and word errors caused by the same or similar pronunciation. In the translation between Japanese and Chinese recognition, the whole sentence omission, sentence patterns, tenses and so on are more common.

1. Introduction
Automatic Speech Recognition (ASR) is a part of intelligent speech technology, which refers to the technology of transforming speech into text. With the development of intelligent speech technology, more and more researchers pay attention to the evaluation of ASR, which can benefit medical, education, transportation, and other industries. Recently there have been many related researches in medical field. Zhang concluded in his experiment that ASR could alleviate the subjectivity in manual detection and the shortage of doctors [1]. Renato Alejandro et al. used ASR to record cases in paediatrics and internal medicine, with an accuracy rate of 94.1% [2]. Schuster Maria et al. confirmed the accuracy of speech recognition in medical diagnosis [3]. However, most of these studies focused on speech recognition of a single language and paid less attention to the evaluation of recognition and translation of different languages. In fighting 2020 coronavirus, intelligent speech technology has to face the language diversity in international medical assistance and communication. Therefore, this paper aims to investigate the public’s understanding of speech recognition technology and the word accuracy rate of speech recognition technology in Chinese and Japanese in epidemic prevention.

Iflytek input software is a product of Iflytek company, which ranks top in speech recognition technology in China [4]. Iflytek input software has a high word accuracy rate and the largest variety of languages [5]. Besides, Iflytek-input software is widely used, whose users has exceeded 400 million and active users has reached 110 million by 2016 [6]. Therefore, this paper takes Iflytek-input as an example to evaluate speech recognition technology.
2. Materials and Methods

2.1. Questionnaire
Questionnaires were distributed online to investigate the social impression of intelligent speech technology. Totally 163 participants (53 males, 110 females, age from 17 to 60) were investigated, and 163 valid questionnaires were collected. There are 11 questions in the questionnaire, with multiple choices and fill-in-the-blank questions. The content includes age, gender, occupation, and usage of intelligent speech products, etc. After the questionnaires were collected, the answers of each question were counted and analysed.

2.2. Corpus
In order to make the spoken text close to the actual scenes, the speech experiment simulated a patient’ describing his feeling to a doctor. According to Beijing Daily [7], under the guidance of the Department of Language Information Management, Ministry of Education, the “Epidemic Prevention Language Service Team” selected 156 words and 75 phrases according to corpus statistics and medical scene investigation, which are well designed and widely used. Considering the feasibility, 21 words and 20 short sentences were chosen and modified as experimental corpus in our study.

2.3. Speech Experiment
The tools in this experiment were iTunes player in MacBook Pro, recording and memo function of mobile phone, and Iflytek-input software V9.1.9595 version.

The speakers in this experiment were four Chinese students majoring in Japanese, who had passed Japanese N1 test and had no speech or voice problems.

1) The speech materials were sent to speakers as a document, where it was stated that a quiet environment is necessary. Besides, the mobile phone should be placed flat on the table with a distance of about 30cm from the lips. Speakers should pronouncce loudly and clearly, and pause two seconds between each sentence. Speakers finished 20 sentences in Mandarin and named the recording “01”. Then they translated the sentences into Japanese and filled out the document. Then they read out the 20 Japanese sentences and named the recording “02”. At the end, these two recordings and a word file with Japanese sentences were collected by our analysers.

2) The recordings were played on the computer, and “long text” mode was set in Iflytek-input. The memo in mobile phone was to record identification results, which was placed flat at a distance of 30cm from the computer. Firstly, “Mandarin” mode was selected to recognize recording “01”. The 20 sentences were played in a row (this same step would be performed below). Secondly, “Japanese” mode was selected to recognize recording “02”. Thirdly, “Translate from Mandarin to Japanese” mode was selected to recognize recording “01”, recognizing the Mandarin speech sound to Japanese text. Fourthly, “Translate from Japanese to Mandarin” mode was selected to recognize recording “02”, recognizing Japanese speech as Chinese text.

3) The word document was filled with four recognition results of each speaker. The input sentences and output results were listed together.

4) The errors of the recognition were found out and highlighted by our analysers. Then the cause of errors was summarized.

3. Results

3.1. Public Awareness of Intelligence Speech Technology
Most of the participants have had a bachelor's degree, accounting for 69.2%. Second are people with high school diploma and master's degree, accounting for 14.5% and 12.6%, respectively. 3.1% participants have had junior middle school diploma and 0.6% have had a doctoral degree. No participants have primary school education or below. Full-time students are the main participants, accounting for 46.6%. Sales staff and management staff are the second, accounting for 9.8% and 7.4%,
respectively. Other industries such as production personnel, public relations personnel, customer service personnel, financial personnel, and so on are involved. 85.3% know intelligence speech intelligence, while 14.7% do not. Most participants have used intelligent speech products listed below. Uses of these products are in table 1.

| Options                        | Number | The percentage |
|-------------------------------|--------|----------------|
| The speech-to-text of QQ      | 108    | 66.3%          |
| The speech-to-text of Wechat  | 154    | 94.5%          |
| Siri of IOS system            | 74     | 46.4%          |
| Iflytek-input method          | 63     | 38.7%          |
| Baidu intelligent speech input method | 69 | 42.3%          |
| speech-activated navigation   | 46     | 28.2%          |
| Intelligent robots            | 30     | 18.4%          |
| Intelligent speech customer service | 48 | 29.5%          |
| Others                        | 7      | 4.3%           |

The usability is rated by a slider, from 1 representing “very impractical” to 5 “very impractical”. 85% of participants think these products are practical, indicating a positive attitude towards them. Regarding the frequency of use, 47.2% sometimes use them. 30.1% frequently use them. Only 3.7% have never used them. Besides, their convenient performance has been generally recognized. Table 2 is the positive evaluation:

| Options                        | Number | Percentage |
|-------------------------------|--------|------------|
| Saving time                   | 129    | 79.1%      |
| Freeing hands                 | 121    | 74.2%      |
| Improving the quality of life | 71     | 37.4%      |
| Improving driving safety      | 62     | 38.0%      |
| Reducing artificial costs     | 43     | 26.4%      |
| Other                         | 5      | 3.1%       |

As for disadvantages, 62.6% participants think speech products are deficient in dialect recognition. Second are low word accuracy rate and poor emotion recognition, accounting for 53.4% and 44.8%, respectively. In addition, 30.7% and 54.0% participant request speed of conversion and word accuracy rate of foreign language recognition. 11.0% participants think speech recognition products are too expensive.

3.2. Speech Recognition Experiment for Japanese Speakers
For the convenience of writing and quoting later, the speech recognition results are marked as follows: first is the name of speakers, who are represented by “Speaker1”, “Speaker2”, “Speaker3”, and “Speaker4”. Then the recognition result in Mandarin, Japanese, Japanese from Chinese and Chinese from Japanese are respectively denoted as “Task 1”, “Task 2”, “Task 3” and “Task4”. Last is the sentence number from 1 to 20. For example, “Speaker1Task1.20” refers to “Speaker1’s 20th sentence in Mandarin recognition”.

3.2.1. Word Error Rate Algorithm. Word Error Rate (WER) is used to calculate the recognition word accuracy rate, which is an important index for speech recognition evaluation. It is calculated by dividing
the total number of characters inserted, deleted and replaced by the total number of characters, and then take a percentage [8]. The formula is as follows:

\[ \text{WER}=100\% \times \text{WER} = \frac{(S + D + I)}{T} \] (1)

When calculating WER, the number of characters occupied by punctuation is not included in the calculation formula. S (substitution) stands for the number of words that substitute for the original sentence; D (deletion) stands for the number of words deleted from the original sentence, and I (insertion) stands for the additional words than the original sentence. T (total) stands for total number of words in the sentence. If the recognition result is correct, the word error rate is 0. If the number of I is large, WER may exceed 100%.

Word Accuracy Rate = \((1-\text{WER}) \times 100\%\) (2)

3.2.2. Speech Recognition in Mandarin. The main problems are tone errors and word errors.

1) Tone errors: Speaker2 Task1.8 “wo ba ba shu ye shu wan le, yao ba zhen tou” (My father is finished with the infusion and the needle needs to be pulled out). The tone of “ba” is the rising tone, which is identified as the falling-rising tone.

2) Word errors: Speaker2 Task1.17 “hu shi” (nurse) is identified as “bu shi” (not) due to identification error in initial consonant. The consonant of “hu” is [h], which belongs to glottal fricatives. The consonant of “bu” is [p], which is voiceless unaspirated [9]. Speaker3 Task1.14 “le” is identified as “de” because of similar pronunciation as well.

The word accuracy rate of Mandarin recognition of the four speakers is listed in table 3.

| Speakers    | Wrong sentence number | Word Accuracy Rate |
|-------------|-----------------------|--------------------|
| Speaker1    | none                  | 100%               |
| Speaker2    | 8, 17                 | 98.81%             |
| Speaker3    | 14                    | 99.6%              |
| Speaker4    | 7                     | 99.21%             |

3.2.3. Japanese Speech Recognition. Word errors are the main problems.

1) The confusion of unvoiced and voiced sounds. Speaker1 Task2.10 か is translated into が. Speaker2 Task2.11 “まだ” identified as “また”. Similar examples include Speaker2 Task2.12, Speaker2 Task2.14, Speaker2 Task2.16, etc.

2) The confusion of two adjacent hiragana in the fifty sound. Speaker1 Task2.13 “渦る：はかる” is identified as “上る：あがる”. は and あ are in the same paragraph. Speaker1 Task2.14 “も” is identified as “も”, for both are on the same line. Other examples include Speaker2 Task2.2, Speaker2 Task2.17, Speaker3 Task2.5, Speaker3 Task2.7, etc.

3) The confusion of words with the same hiragana but different stress. Speaker1 Task2.12 “重い：おもい” is translated as “思い：おもい”. Three hiragana of “重い” basically have the same stress, while the intensity of “思い” is on the second hiragana.

4) Adding and omitting words. Adding redundant words mainly occurs at the end of sentences. Auxiliary word “〜へ” and dial tone “ん” are usually omitted.

The word accuracy rate of Japanese speech recognition of four speakers is listed in table 4.

3.2.4. Speech Recognition from Chinese to Japanese. Meaning errors and meaningless sentences are main problems in sentence level. Meaning errors can be divided into pattern errors and tense errors.

1) Pattern errors. In Task3.6 “wo nai nai de bing yao bu yao jin” (Is my grandmother seriously ill) is translated as “私のおばあさんの病気は大丈夫です”(My grandmother's illness is not serious).
The interrogative sentence is translated into negative sentences. In Speaker 2 Task 3.5, a positive sentence is translated into a negative sentence.

(2) Tense errors. Speaker 1 Task 3.7 “wo mama de bing zhi de hao bu?” (Can my mother be cured) is translated as “wo mama de bing zhi hao le ma?” (Has my mother recovered from her illness). The recognition result indicates that the treatment has already begun, while the original sentence means the treatment has not.

(3) Meaningless sentences. Speaker 2 Task 3.7 “yi sheng, wo mama de bing zhi de hao bu? “(Doctor, can my mother be cured) is translated as “yi sheng ke yi chi wo mama de bing ma?” (Can the doctor eat my mother’s disease).

The word accuracy rate of Chinese translation of Japanese of four speakers is listed in Table 5.

Table 4. The word accuracy rate of Japanese speech recognition of four speakers.

| Speakers | Wrong sentence number | Word Accuracy Rate |
|----------|-----------------------|--------------------|
| Speaker 1 | 1, 4, 9, 10, 12, 13, 14, 15, 18, 19 | 95.64% |
| Speaker 2 | 2, 8, 11, 12, 13, 14, 15, 16, 17, 20 | 91.34% |
| Speaker 3 | 5, 7, 15 | 99.2% |
| Speaker 4 | 6, 16, 18 | 98.12% |

Table 5. The word accuracy rate of Chinese translation of Japanese of four speakers.

| Speakers | Wrong sentence number | Word Accuracy Rate |
|----------|-----------------------|--------------------|
| Speaker 1 | 3, 6, 7, 8, 11, 20 | 97.78% |
| Speaker 2 | 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 17 | 90.91% |
| Speaker 3 | 2, 6, 7, 12 | 98.73% |
| Speaker 4 | 2, 6, 8, 9, 15, 16, 17, 18, 19, 20 | 89.29% |

3.2.5. Speech Recognition from Japanese to Chinese. The problems are serious at both sentence level and word level.

(1) Meaningless sentences. Speaker 1 Task 4.7 “yi sheng, wo mama de bing zhi de hao bu?” is translated into “yi sheng, shi ni mama de bing ma? zhi hao de ke neng xing zen me yang?” (Doctor, is it your mother’s disease? Can it be cured?) The same goes for Speaker 1 Task 4.18 and Speaker 2 Task 4.2.

(2) Recognition omission. Speaker 2 Task 4.6 omits the subject, resulting in the lack of meaning of the sentence. Several sentences are wholly omitted, such as Speaker 4 Task 4.12 and Task 4.13.

(3) Word errors. Speaker 1 Task 4.15 “Hubei” is translated into “aviation performance”. Speaker 2 Task 4.7 “I” is translated into “you”. Speaker 3 Task 4.15 “Hubei” is translated into “country”.

(4) Typographical errors. The writing of “hu” in phrase “hu xi kun nan” (lose one’s breath) is translated into a variant form of “hu”, which has already been stopped using. In addition, it is found that Iflytek has a low rate of recognition and translation of medical terms.

The word accuracy rate of Japanese translation of Chinese of four speakers is listed in Table 6:

Table 6. The word accuracy rate of Japanese translation of Chinese of four speakers.

| Speakers | Wrong sentence number | Word Accuracy Rate |
|----------|-----------------------|--------------------|
| X        | 1, 3, 7, 9, 12, 14, 15, 17, 18, 19 | 70.63% |
| W        | 2, 5, 6, 7, 8, 12, 13, 14, 15, 16, 17, 20 | 83.72% |
| S        | 1, 2, 7, 9, 12, 13, 14, 15 | 91.14% |
| G        | 2, 4, 6, 11, 12, 13, 14, 16, 17, 18 | 85.47% |
4. Conclusions
It could be seen from our investigation that currently intelligence speech technology is widely used in social life. Most social groups think highly of the usability of speech intelligence products, and the frequency of using these products is high, which show the intelligence speech products have a good development prospect. However, higher requirements are put forward for the diversity and accuracy of speech recognition.

The Iflytek-input method has the highest word accuracy rate for Mandarin, with an average of 99.4%, followed by Japanese with an average of 96.1%. It can be seen that Iflytek-input method has a high word accuracy rate for a language. The word accuracy rate of Chinese translation of Japanese is 94.2%, while that of Japanese translation of Chinese is only 82.7%. Speech recognition in Chinese translation and Japanese translation increases the link of translation. By comparing the word accuracy rate of the former two, it is found that translation is a major obstacle to recognition between different languages.

The problems in Japanese and Chinese recognition are at the word level, while the problems in the latter two are mainly at the sentence level. For example, there are omission of individual words in the monolingual recognition, but whole sentence missing in translation of different languages.

This study evaluates the performance of intelligent speech technology in practical application scenarios in an independent, non-commercial and scientific way, which will be beneficial to formulating industry standards of the state for future, providing the scientific basis and reference for government to purchase intelligent speech technology products, in order to better serve for the society and people’s livelihood.

5. Future Discussion
Machine translation and artificial intelligence are not mature enough and are still in the initial stage of development [10]. Besides, the quality of machine translation is far inferior to human translation in dealing with professional terms, because “machine translation is not only a language processing problem, but also a cultural and knowledge communication processing problem” [11]. How to improve the word accuracy rate of recognition and translation of different languages remains to be solved in future studies. Sound evaluation criteria also need to be developed for speech recognition products. To solve these problems, the characteristics of language should be analysed from the perspective of linguistics, and speech recognition technology might be combined with more linguistic knowledge.

Acknowledgments
This study is supported by SSFMEC Program (16YJC740065) and NSFC Project (11974054).

References
[1]  Zhang W 2019 Summary and prospect Research on the Analysis and Evaluation of Functional Dysarthria Based on Speech Recognition chapter 5 p 44.
[2]  Renato A, Berinsky H, Daus M, et al. 2019 Design and evaluation of an automatic speech recognition model for clinical notes in Spanish in a mobile online environment Studies in Health Technology and Informatics (264) 1761-62.
[3]  Schuster M, Maier A, Bocklet T, et al. 2012 Automatically evaluated degree of intelligibility of children with different cleft type from preschool and elementary school measured by automatic speech recognition International Journal of Pediatric Otorhinolaryngology (76) 362-9.
[4]  Wan J P, Liu Z H, Wang Y, Liu W J, Zhang Q T and Xin J 2012 Patent analysis based on automatic evaluation of speech recognition technology Audio Engineering (36) 53-56.
[5]  Han Y, Jiang C H, Li S, Wang B B, Wang Y C, Qiu Q and Shi X H 2017 A preliminary study on the objective evaluation of phonological recognition technology on the phonological pronunciations of children in Mandarin Stomatology (37) 515-9.
[6]  https://c.m.163.com/news/a/C4T28CJ100097U80.html?spss=newsapp&spsw=1.
[7]  https://xw.qq.com/cmsid/20200213A00R6V00.
[8] Wu B, Lan S Y and Liu C Q 2005 An improved word error rate assessment algorithm for speech recognition Infrared and Laser Engineering (34) 106-9.
[9] Ladefoged P and Johnson K 2010 Consonantal gestures A Course in Phonetics (Wadsworth) chapter 7 p 173.
[10] Feng Z W 2018 Parallel development of machine translation and artificial intelligence Foreign Languages (Shanghai International Studies University) (41) 35-48.
[11] Chen Y 2005 Current situation and future of machine translation in Japan and China Chinese Science and Technology Translation (01) 31-33.