Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 4 (2011) 1336–1342

Procedia
Computer Science

International Conference on Computational Science, ICCS 2011

Personalized Translation Listening System for Age Groups

Do-won Jang[a,*], Dong-Ju Kim[a], Kwang-Seok Hong[a]

[a]*School of Information and Communication Engineering, Sungkyunkwan University, Republic of korea*

Abstract

Present translation systems play back their results with specific speed and volume regardless of the type of user. In this paper, we describe a personalized translation listening system by identifying user information. User groups are classified by age group after inputting speech. This suggested translation system gives the translated result back to the user, speaking at an emphasized speech based on the user's age. The personalized translation listening system performs age recognition, speech recognition, language recognition, Google machine translation, and TTS web services. We showed the comparison of normal speech and emphasized speech by realizing a user-oriented translation system in a mobile device. A Mean Opinion Score was used for the experiment. Average score of the Mean Opinion Score experiment was 3.21.

"Keywords: Personalized translation; Listening system; Age recognition; Age Groups; Mobile device"

1. Introduction

Many studies about effective translation are under way in order to give people smoother communication [1-4]. In Japan, the US, Italy and Germany, studies about translation are very popular. India is also very involved in translating between the country's 22 official languages [5, 6].

As the diffusion rate of mobile device and wireless internet grows higher, more people have a mobile device and can use it everywhere. Thus, this paper suggests a translation system that utilizes two speech recognizers and machine translation based on Google's free web service. In order to cover the weakened hearing ability of old age groups [7], this thesis also emphasizes different speaking patterns for each age group. This aims to increase the accessibility and marketability of the translation system. This paper compares normal speech and emphasized speech by realizing a user-oriented translation system in a mobile device.

This paper is organized as follows: Section 2 introduces the related research of speech translation and speech recognition. Section 3 shows the structure of the translation system and the detailed information of each process. Section 4 is about the experiments about the translation system. Finally, the conclusions of this paper are presented in Section 5.

* Corresponding author. Tel.: +82-31-290-7196.
E-mail address: dowonjj@skku.edu.

Selection and/or peer-review under responsibility of Prof. Mitsuhisa Sato and Prof. Satoshi Matsuoka
doi:10.1016/j.procs.2011.04.144

2. Relative Works

2.1. Speech Translation

The Consortium for Speech Translation Advanced Research (C-STAR) was organized in order to develop an effective translation tool [8]. ATR R&D lab (in Japan), Carnegie Mellon University (in the US), ITCIRST R&D lab (in Italy), Karlsruhe University (in Germany), the CLIPS group (in France) and NLPR (in China) all take part in this research. IBM set the commercialization of auto translation technology as their top priority in their five year business plan. Google declared that it will provide a mobile translation service in the years to come.

2.2. Speech Recognition

The National Institute of Standards and Technology (NIST) started the Advanced Research Projects Agency (ARPA) project with Ministry of National Defense aid in the early 1970's, and the project name has been changed. Now called the Defense Advanced Research Projects Agency (DARPA), it does research every year collaborating with global speech recognition research labs. IBM released the Via Voice program. Microsoft developed the Speech API (SAPI, the standard of voice recognition interfaces) and the Whisper (a search engine based on voice recognition).

In Europe, the European Union (EU) took the lead in this research. The CORETEX project under control of the UK, Germany, Italy, and France aims to set the standard of their respective countries' languages.

In Japan, the focus is on Japanese language translation technology. Japan started the development of an auto translation system in 1986. Toshiba, Fujitsu, Toyohashi University of Technology, and Tokushima University are all doing exclusive research work about auto translation technology.

3. Translation System Structure

3.1. Structure

The overall Personalized Translation Listening System is shown in Fig. 1. Once user 1 chooses the language and inputs the speech, the speech data will be transferred to the server. When the age recognition and translation work is done on the server, the translated data will be sent to the mobile device again so that user 2 can listen to the translated speech. Then, user 2 inputs the speech in another language and user 1 listens to the translated speech.

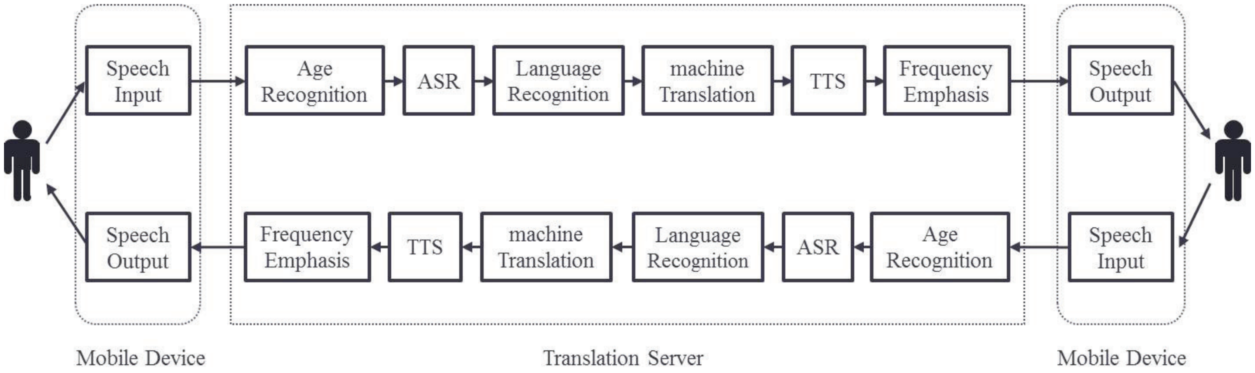

Fig. 1. Personalized Translation Listening System procedures. Users on the right and left are using different languages.

The system structure is based on a client-server system. Thus, the large number of users can access the main server and use the updated system for just one server.

3.2. Speech Input

The user has to select two languages before speech input. The first combo box is user 1's language; the second combo box is user 2's language. By following this procedure, the program can condense the possible languages down to two so that rate of language recognition can be increased.

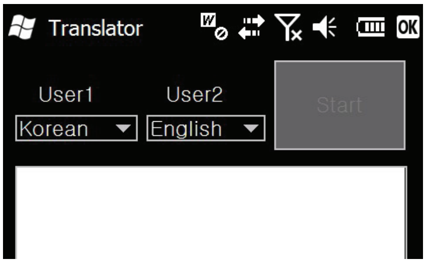

Fig. 2. When a user selects two different kinds of languages, initialization of the translation system is started. Then, when initialization is complete, the start button is enabled and users can input the speech.

When press the start button, the user can input the speech data. Then the system finds the start / end point of input speech data by calculating each frame of energy. The detected speech data will be saved in the PCM file format and transferred to the server. The program shows the wavelength graph on the screen so the user can check whether or not it has been cut. This progress repeats until the system finds out the accessible speech data.

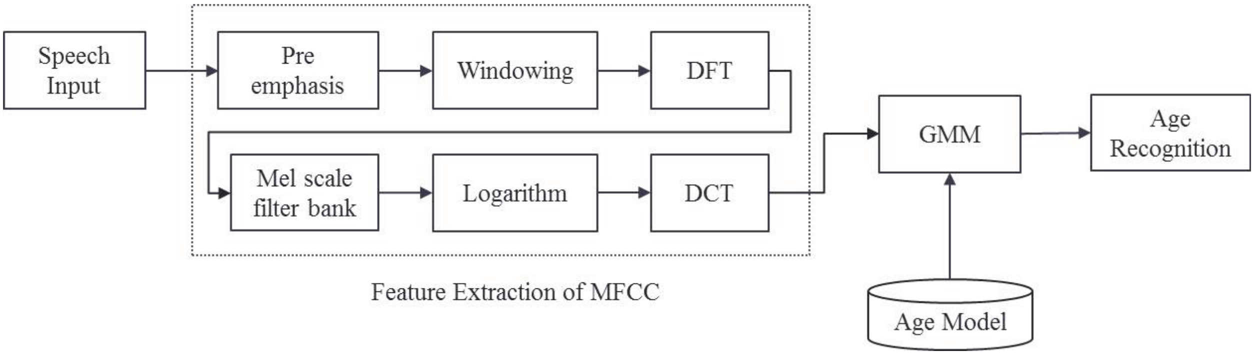

Fig. 3. Age group recognition. This process extracts MFCC features, in the form of 39 dimensional vectors containing the 12 MFCCs and the normalized frame energy, along with their first and second derivatives.

3.3. Age Group Recognition

The progress of age recognition follows Fig. 3. MFCC coefficients are extracted from the input speech data. Similarity is figured out using each age group based on the GMM algorithm. The maximum value of each age group’s result among similarity is chosen as the input age group's recognition result. There are four recognized age groups: children, young people, adults, and senior citizens [9].

- Children: ≤ 13 years
- Young people: 14-19 years
- Adults: 20-64 years
- senior citizens: ≥ 65 years

3.4. Translation System Procedure

Progress on the server is made regardless of the user. After inputting speech data, the user has to wait until the result comes back from the server. Let's take a closer look at the translation on the server. In the process of speech recognition, the selected user 1 and user 2 languages from Fig. 2 are input in the PCM file format (16 kHz, 16 bit) into the speech recognizer. The speech recognizers we used for this research are the HMM Tool Kit (HTK) for Korean language recognition and PowerEASR for English recognition of HCI Lab Co., Ltd. If speech recognition fails, a failure message will be saved and sent to the user.

In the process of language recognition, the similarity between two language recognizers will be summed up and compared in order to identify the language being spoken. The higher value will be chosen as the user's language and the lower value will be set as the language that needs to be translated. If the difference between the two values is less than 20, the server will send a fail message to the user's mobile device.

In the process of machine translation [10], the server requests machine translation work from Google by connecting to the wireless internet. In order to utilize Google translation, the input language, output language, and the sentence that needs to be translated should be included. The input language and the output language at this process will be set as what the user chooses at the language recognition process. And the sentence that needs to be translated will be translated from what the user recorded during the speech recognition process.

Text to Speech (TTS) transfer means making a speech file from text utilizing Google's TTS web service [11, 12]. The server sends a call to Google's TTS service and receives a result back in an MP3 file format from Google. After TTS transfer work is done, the MP3 file will be sent to the user's mobile device from the server.

3.5. Emphasize Frequency

The frequency band of normal conversation ranges from 100 Hz to 8 kHz. Referring to Table 1, you can see that voice energy of 100 Hz to 2 kHz takes 98% and the voice articulation of 500 Hz to 4 kHz takes 83% of whole frequency band [13].

Table 1. Energy and Articulation of speech frequency band.

Frequency band(Hz)	Energy (%)	Articulation (%)
100~500	60	5
500~1000	35	35
1000~2000	3	35
2000~4000	1	13
4000~8000	1	12

Given Fig. 4 showing that hearing loss depends on age, a different frequency band will be emphasized for each age group.

The emphasis of frequency band is applicable only when the age recognition result is adults or senior citizens. In case of adults, it follows Formula (1). In case of senior citizens, it follows the formula (2) [13].

$$D_a = \begin{cases} 0.0016f + 1.9, & if\ 0.5kHz < f \leq 1kHz \\ 0.003f + 0.5, & if\ 1kHz < f \leq 2kHz \\ -0.0004f + 7.3, & if\ 2kHz < f \leq 4kHz \end{cases} \qquad (1)$$

$$D_a = \begin{cases} 0.002f + 16, & if\ 0.5kHz < f \leq 1kHz \\ 0.007f + 11, & if\ 1kHz < f \leq 2kHz \\ 0.006f + 13, & if\ 2kHz < f \leq 4kHz \end{cases} \qquad (2)$$

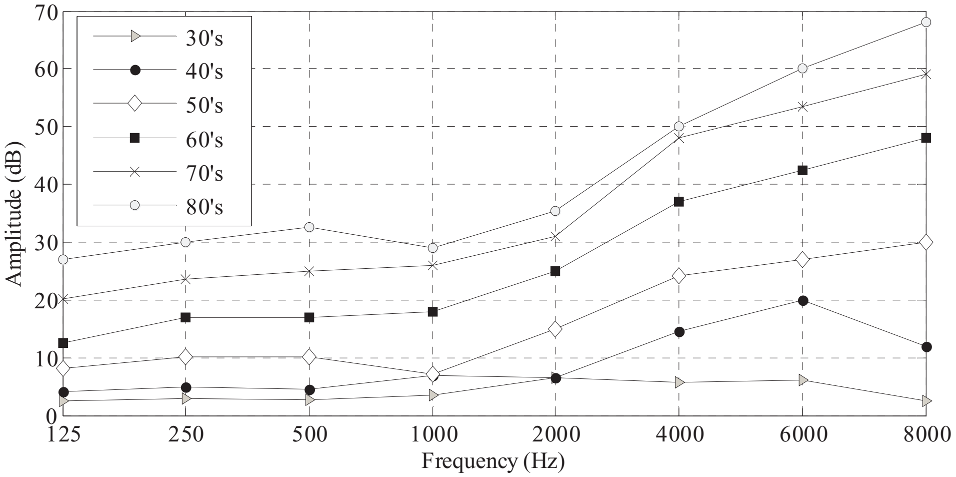

Fig. 4. Hearing loss depending on age groups.

3.6. Playing TTS

The mobile device maintains a listen state until the translation work finishes. After downloading the translated speech file from the server, the mobile device plays the result file automatically.

4. Experiment of Translation System

4.1. Environment

In this paper we used a Show Omnia (SPH-M8400) with Windows Mobile 6.5 Professional OS as the mobile device. The server computer was equipped with an Intel i5 750 CPU, 4GB RAM with Windows 7 OS. Wireless Internet was used for the wireless AP in the lab. Ten people, aged 48 – 66 years old, participated in the experiment. The Mean Opinion Score (MOS) was used as the experimental method. First, ten participants listened to the original translated speech. Then, the participants listened to the emphasized speech and chose a score from 1 to 5 to evaluate the quality of the speech. Table 2 shows the standard MOS form.

Table 2. Representation of the standard form of MOS. The MOS is the arithmetic mean of all the individual scores, and can range from 1 (worst) to 5 (best).

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Bad	Very annoying and objectionable

4.2. Result

An example of frequency emphasis is shown in Fig. 5. We show the frequency emphasis from 500 Hz to 4 kHz. Table 3 shows the individual score of the experiment and average score per sentence and user. The average score of the MOS experiment is 3.21. Because the score of MOS test were higher than 3, we confirm that the speech of emphasized frequency heard a little better than not emphasized.

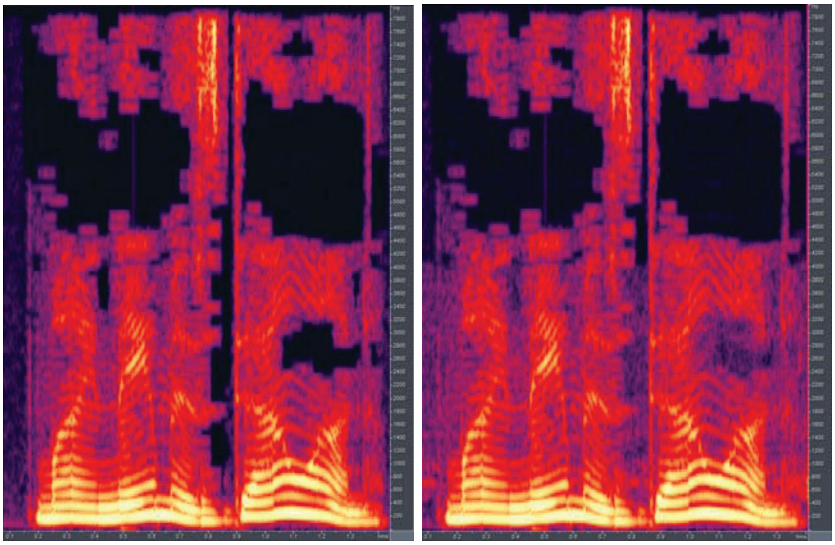

Fig. 5. Speech spectrogram (a) before frequency emphasis; (b) after frequency emphasis.

Table 3. MOS Experiment Score.

Sentence	User1	User2	User3	User4	User5	User6	User7	User8	User9	User10	Average
1	2	1	3	3	3	3	3	3	3	3	2.7
2	1	3	3	3	3	3	3	3	3	2	2.7
3	2	4	3	3	3	3	3	3	3	3	3
4	4	4	3	4	3	3	3	3	3	3	3.3
5	4	4	4	4	3	4	4	3	3	3	3.6
6	5	4	3	3	3	3	3	3	3	4	3.4
7	4	5	3	4	3	3	3	3	3	4	3.5
8	3	5	3	3	3	3	4	3	3	3	3.3
9	4	3	3	4	3	3	3	3	3	3	3.2
10	5	4	3	3	3	3	3	3	3	4	3.4
Average	3.4	3.7	3.1	3.4	3	3.1	3.2	3	3	3.2	**3.21**

5. Conclusion and Future Work

In this paper, we showed the comparison of normal speech and emphasized speech by creating a personalized translation listening system in a mobile device. The suggested system is composed of age recognition, speech recognition, language recognition, Google machine translation, and TTS web service. The system also emphasizes the specific speech of each age group. The mobile device only inputs a user's speech which was then transmitted to the server to handle all the processes like recognition and translation. The processing speed was improved. Average score of the MOS experiment was 3.21. We believe that our proposed methods will the groundwork of development in personalized speech translation.

Future work for this research will be about adding gender recognition to this process in order to improve the age recognition rate. In addition, In order to increase the rate of speech recognition, we will detect the distance of speaker.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2010-0027789) and Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0020210).

References

1. Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Genichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, Jin-Song Zhang, Hirofumi Yamamoto, Eiichiro Sumita, Seiichi Yamamoto, The ATR Multilingual Speech-to-Speech Translation System, Audio, Speech, and Language Processing, pp.365-376, 2006.
2. F. Casacuberta, H. Ney, F.J. Och, E. Vidal, J.M. Vilar, S. Barrachina, I. Garcia-Varea, D. Llorens, C. Martinez, S. Molau, F. Nevado, M. Pastor, D. Pico, A. Sanchis, C. Tillmann, Some approaches to statistical and finite-state speech-to-speech translation, Computer Speech and Language, volume 18, pp.25-47, 2004.
3. Schultz T., Black A.W., Challenges with Rapid Adaptation of Speech Translation Systems to New Language Pairs, In Proceedings of the ICASSP, 2006.
4. JongHo Shin, Panayiotis G. Georgiou, Shrikanth Narayanan, Towards modeling user behavior in interactions mediated through an automated bidirectional speech translation system, Computer Speech and Language 24, pp.232-256, 2010.
5. Sakriani Sakti, Noriyuki Kimura, Michael Paul, Chiori Hori, Eiichiro Sumita, Satoshi Nakamura, Jun Park, Chai Wutiwiwatchai, Bo Xu, Hammam Riza, Karunesh Arora, Chi Mai Luong, Haizhou Li, The Asian network-based speech-to-speech translation system, Automatic Speech Recognition & Understanding, pp.507-512, 2009.
6. Gary Anthes, Automated translation of Indian languages, Communications of the ACM, Volume 53, Issue 1, pp.24-26, 2010.
7. Korea Industrial Safety Assiciation, http://www.safety.or.kr
8. Boitet C., GETA's MT methodology and its current development towards personal networking communication and speech translation in the context of the UNL and C-STAR projects, In Proceedings of the PACLING-97, pp.23-57, 1997.
9. F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J.G. Bauer, B. Littel, Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications, in Proceeding of ICASSP, pp.1089–1092, 2007.
10. Adam Lopez, Statistical machine translation, ACM Computing Surveys, Volume 40, Issue 3, 2008.
11. M. Chu, C. Li, H. Peng, E. Chang, Domain adaptation for TTS systems, In Proceedings of the ICASSP, 2002.
12. M. Beutnagel, A. Syrdal, P. Brown, Preselection of candidate units in a unit selection-based text-to-speech synthesis system, In Proceedings of the ICSLP, pp.314-317, 2000.
13. Kelly L. Tremblay, Michael Piskosz, Pamela Souza, Effects of age and age-related hearing loss on the neural representation of speech Cues, Clin Neurophysiol 114, pp.1332-1343, 2003.