2019 19th International Conference on Control, Automation and Systems (ICCAS 2019)
Oct. 15~18, 2019; ICC Jeju, Jeju, Korea

Word Error Rate Comparison between Single and Double Radar Solutions for Silent Speech Recognition

Sunghwa Lee[1, 2] and Jiwon Seo[1, 2*]

[1] School of Integrated Technology, Yonsei University,
Incheon, 21983, Korea (sunghwa.lee@yonsei.ac.kr, jiwon.seo@yonsei.ac.kr)
[2] Yonsei Institute of Convergence Technology, Yonsei University,
Incheon, 21983, Korea
* Corresponding author

Abstract: Silent speech recognition (SSR) is a technology that translates human speech into text without voice information. Various sensors, such as vision, electromyography, electromagnetic articulatory, and radar sensors, can be used to build an SSR system. Because the signals of radar sensors are less intuitive, radar-based SSR research is less common and remains at a basic level compared with work on other sensors. As a basic step in this research area, in this study, we attempted to determine whether single radar or double radar shows better performance for an SSR system. To this end, we estimated the word error rate (WER) of each system. The results showed that a double-radar-based SSR system produced better WER output. This means that the number of radar sensors used in SSR can potentially affect its performance. Therefore, when we create a radar-based SSR hardware platform, how many radar sensors would be ideal for its best performance must be considered.

Keywords: silent speech recognition, radar, word error rate

1. INTRODUCTION

Silent speech recognition (SSR) is one means of communication in which human speech is communicated without voice cues [1, 2]. In place of the speaker's voice, articulatory movements (e.g., lip or tongue movement) can be utilized to recognize what the speaker has communicated. SSR is expected to supplement automatic speech recognition (ASR). For instance, in a noisy environment where the ASR tool cannot perform normally, an SSR device can provide additional speech information to the ASR system. Moreover, speakers who have lost their voice could utilize an SSR device to deliver whatever they would like to express.

SSR is an important research topic in terms of its potential to provide devices that could function as new sources of input. Recently, to improve vehicle safety and convenience, a myriad of studies have been carried out utilizing vision [3, 4], ultra-sonics [5], radar [6-10], and GPS sensors [11-13]. Specifically, an SSR device could enhance a car driver's convenience by recognizing the driver's speech or lip gestures. Furthermore, research for handicapped people is being implemented using a variety of tools. For example, haptic devices mounted on robots can guide the blind by giving them haptic feedback information about following their path [14-18]. SSR devices could also be helpful as a communication tool for people who have difficulty using their voice.

To read articulatory movement, various types of sensors can be utilized. These include vision [19], electromyography (EMG) [20, 21], electromagnetic articulatory (EMA) [22, 23], and radar [24, 25] sensors. Radar sensors, in particular, have advantages over the other sensors because they are non-invasive and contactless. Moreover, if we use impulse radio ultra-wide band (IR-UWB) radar, its pulse can penetrate human skin so that tongue motion is expected to be detected. However, so far, there have been only basic studies that utilize radar sensors for SSR. For example, in [25], the authors classified 10 numbers, from one to ten, and confirmed the potential usefulness of a single radar sensor for SSR.

To develop the concept further and make a complete radar-based SSR system, we need to consider the use of multiple radar sensors. In this study, as a first step, we compared the performance of single-radar-based and double-radar-based SSR systems based on estimation of the word error rate (WER). This paper is organized as follows. Section 2 describes how the data used in this paper was collected. Section 3 explains the way we designed the learning model, and Section 4 presents the results. Section 5 discusses ways to improve the radar-based SSR system and Section 6 summarizes the conclusions.

2. DATA PREPARATION

2.1 Experimental environment

In this study, we arranged two IR-UWB radar sensors. As seen in Fig. 1 (a) and (b), each radar sensor has one transmitting antenna and one receiving antenna, both of which were aligned vertically. The lips of the participant were about 20 cm distant from the center of the double radar set. The participant was ordered to remain unmoving except for the lips in order to eliminate other variables that might affect the recognition results.

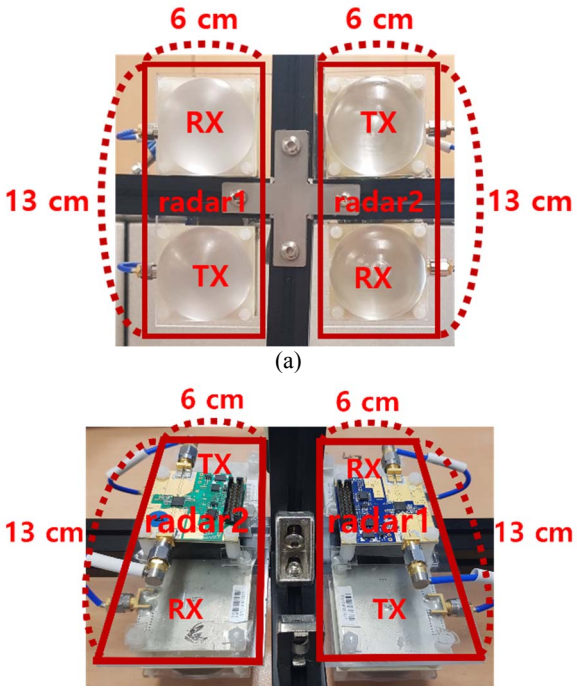

Fig. 1 Arrangement of the four antennas of two radars: (a) Front view; (b) Rear view. TX and RX indicate transmitter and receiver components, respectively.

2.2 Data constitution

In this study, all learning data and test data were from English phrases. In the learning data, the number of unique phrases, words, and phonemes was 87, 215, and 39, respectively. Most phrases in the learning data were extracted from [23], and the others were produced by us to make all phonemes appear at least 10 times in each set of phrases. The test data comprised 50 phrases, which came from telephone expressions. Moreover, all the phrases were constituted of 16 unique words and 23 unique phonemes.

One healthy, 25-year-old male speaker, whose native language is Korean but who received at least 15 years of English education, participated in the data collection. The participant recorded 10 sets of phrases to use as learning data and 15 sets of phrases to use as test data.

3. METHOD

The IR-UWB radar used in this study transmits and receives a pulse. The pulses received are averaged so that they constitute one frame with 256 range cells. In this study, the rate in frames per second (FPS) of the radar hardware was about 220, but was not fixed. Therefore, we fixed the FPS to 250 exactly, by interpolating the values of the range cells frame-by-frame using the piecewise cubic Hermite interpolating polynomial (PCHIP) method. Finally, in order to eliminate the scale difference between each radar, we normalized the range cell values to fall within the range −50 to +50. Except for interpolation and normalization, we did not apply any other pre-processing method to the radar data.

In this study, all learning and tests were implemented using the hidden Markov toolkit (HTK) [26]. We produced three types of SSR systems using Gaussian mixture model-hidden Markov model (GMM-HMM), of which two were based on a single radar method and the other one is based on a double radar method. This does not mean that we collected data separately for each system, instead we recorded data once based on a double radar and used the data selectively. Specifically, as shown in Fig. 2, to learn a GMM-HMM model, two single radar systems each utilized a single frame from each radar and the double radar system utilized both frames from the two single radars. In other words, because one frame has 256 range cells, a single radar system has a 256-feature vector size and a double radar system has a 512-feature vector size.

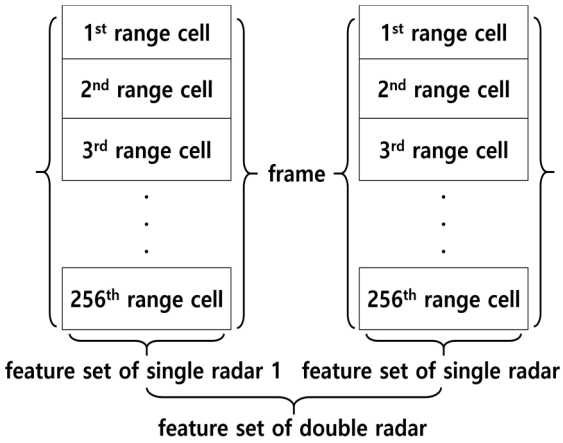

Fig. 2 Feature sets of single radars and double radar to implement GMM-HMM

4. RESULTS

Each GMM-HMM-based SSR system produced the expected words according to the test data. Table 1 shows the recognition results for the test data from each radar SSR system. In Table 1, "H is the number of correct labels, D is the number of deletions, S is the number of substitutions, I is the number of insertions, and N is the total number of labels in the transcription files" [26]. In this quote, "labels" corresponds to words. Table 2 shows the word accuracy (WAcc) and WER of each radar SSR system derived using Eqs. (1) and (2) [27, 28].

$$\text{WAcc} = \frac{H-I}{N} \times 100 \quad (1)$$

$$\text{WER} = \frac{S+D+I}{N} \times 100 \quad (2)$$

A higher WAcc or a lower WER means that the estimated recognition results reflect the actual participant's mouth movement better. As shown in Table 2, the double radar system provides better performance than do the two single radar systems, in terms of both of WAcc and WER. These results indicate that increasing the number of radar sensors adequately could potentially improve the performance of a radar-based SSR system

Table 1 Results of each radar SSR system.

	H	D	S	I	N
Single radar1	389	751	1152	188	2292
Single radar2	319	1078	895	7	2292
Double radar	399	1055	838	54	2292

Note: "*H* is the number of correct labels, *D* is the number of deletions, *S* is the number of substitutions, *I* is the number of insertions, and *N* is the total number of labels in the transcription files" [26].

Table 2 WAcc and WER of each radar SSR system.

	WAcc	WER
Single radar1	8.77 %	83.03 %
Single radar2	13.61 %	87.08 %
Double radar	15.05 %	82.59 %

Note: WAcc and WER stand for word accuracy and word error rate, respectively.

5. DISCUSSION

The WER of the radar-based SSR system could be improved in three ways. First, we could modify the features of the radar data. In this paper, except for interpolation and normalization, there was no other pre-processing step. Therefore, if a feature that better reflects the property of articulatory movement is developed, a lower WER could be achieved. Second, instead of using GMM-HMM, the use of a deep learning-based model, such as a deep neural network (DNN) or long short-term memory (LSTM), could lead to better results. Actually, as seen in [23], an LSTM model performed better than GMM-HMM in an EMA-sensor based SSR system. Third, as we confirmed in this paper, we could increase the number of radar sensors in the SSR system. However, there is no guarantee that more sensors would equal better results. Therefore, it will be important to figure out the optimal number of radar sensors for the SSR system.

6. CONCLUSIONS

Effective SSR has widespread applicability. According to circumstances, it could assist or replace existing ASR devices. Moreover, the SSR device could become a communication tool for voice-impaired people. In this study, we utilized radar sensors for SSR technology. To manufacture a radar-based hardware platform for SSR, we need to decide how many radar sensors to use to detect a speaker's articulatory movements. In this paper, on the basis of WER, we showed that double radar performs better than single radar in SSR. It shows that more radar sensors can result better performance, potentially.

ACKNOWLEDGEMENTS

This research was supported by the Basic Science Research Program thorough the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, and Future planning (NRF-2016R1C1B1010910). This research was also supported by the Ministry of Science and ICT (MSIT), Korea, under the "ICT Consilience Creative Program" (IITP-2019-2017-0-01015) supervised by the Institute of Information & Communications Technology Planning & Evaluation (IITP). The authors gratefully acknowledge Dr. Myungjong Kim for technical advice.

REFERENCES

[1] T. Hueber, G. Bailly, and B. Denby, "Continuous articulatory-to-acoustic mapping using phone-based trajectory HMM for a silent speech interface," *Proc. of the 13th Annual Conference of the International Speech Communication Association*, pp. 723-726, 2012.

[2] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, Vol. 52, No. 4, pp. 270-287, 2010.

[3] W. Song, Y. Yang, M. Fu, Y. Li, and M. Wang, "Lane detection and classification for forward collision warning system based on stereo vision," *IEEE Sensors Journal*, Vol. 18, no. 12, pp. 5151-5163, 2018.

[4] L. Zhang, X. Li, J. Huang, Y. Shen, and D. Wang, "Vision-based parking-slot detection: a benchmark and a learning-based approach," *Symmetry*, Vol. 10, No.3, p. 64, 2018.

[5] J. H. Rhee and J. Seo, "Low-cost curb detection and localization system using multiple ultrasonic sensors," *Sensors*, Vol. 19, No. 6, p. 1389, 2019.

[6] K. A. Smith, C. Csech, D. Murdoch, and G. Shaker, "Gesture recognition using mm-wave sensor for human-car interface," *IEEE Sensors Letters*, Vol. 2, No. 2, pp. 1-4, 2018.

[7] S. Lee and J. Seo, "IR-UWB radar-based near-field head rotation movement sensing under fixed body motion," *Proc. of the International Conference on Electronics, Information, and Communication*, 2018.

[8] S. Lee and J. Seo, "IR-UWB radar based near-field intentional eyelid movement sensing under fixed head and body motions," *Proc. of the International Conference on Control, Automation and Systems*, pp. 1959-1962, 2017.

[9] Y. H. Shin, S. Lee, and J. Seo, "Autonomous safe landing-area determination for rotorcraft UAVs using multiple IR-UWB radars," *Aerospace Science and Technology*, Vol. 69, pp. 617-624, 2017.

[10] S. K. Leem, F. Khan, and S. H. Cho, "Vital sign monitoring and mobile phone usage detection using IR-UWB Radar for Intended use in car crash

prevention," *Sensors*, Vol. 17, No. 6, p. 1240, 2017.
[11] J. Seo, T. Walter, and P. Enge, "Availability impact on GPS aviation due to strong ionospheric scintillation," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 47, No. 3, pp. 1963-1973, 2011.
[12] J. Seo, T. Walter, "Future dual-frequency GPS navigation system for intelligent air transportation under strong ionospheric scintillation," *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 5, pp. 2224-2236, 2014.
[13] J. Lee, Y. T. Morton, J. Lee, H.-S. Moon, and J. Seo, "Monitoring and mitigation of ionospheric anomalies for GNSS-based safety critical systems," *IEEE Signals Processing Magazine*, Vol. 34, No. 5, pp. 96-110, 2017.
[14] H. S. Moon, J. Baek, and J. Seo, "Effect of redundant haptic information on task performance during visuo-tactile task interruption and recovery," *Frontiers in Psychology*, Vol. 7, Art. No. 1924, 2016.
[15] H. S. Moon and J. Seo, "Observation of human response to a robotic guide using a variational autoencoder," *Proc. of the IEEE International Conference on Robotic Computing*, pp. 258-261, 2019.
[16] H. S. Moon and J. Seo, "Prediction of human trajectory following a haptic robotic guide using recurrent neural networks," *arXiv preprint arXiv:1903.01027*, 2019.
[17] A. Ghosh, L. Alboul, J. Penders, and H. Reed, "Following a robot using a haptic interface without visual feedback," *proc. of the 7th Int. Conf. Advances in Computer-Human Interaction (ACHI)*, pp. 147-153, 2014.
[18] Y. H. Hsieh, Y. C. Huang, K. Y. Young, C. H. Ko, and S. K. Agrawal, "Motion guidance for a passive robot walking helper via user's applied hand forces," *IEEE Trans. Human–Mach. Syst.*, Vol. 46, No. 6, pp. 869-881, 2016.
[19] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 455-462, 2017.
[20] G. S. Meltzner, G. Colby, Y. Deng, and J. T. Heaton, "Signal acquisition and processing techniques for sEMG based silent speech recognition," *Proc. of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4848-4851, 2011.
[21] M. Wand, C. Schulte, M. Janke, and M. Schultz, "Array-based Electromyographic Silent Speech Interface," *Proc. of the 6th International Conference on Bio-Inspired Systems and Signal Processing*, pp. 89-96, 2013.
[22] M. Kim, B. Cao, T. Mau, and J. Wang, "Multiview representation learning via deep CCA for silent speech recognition," *Proc. of the INTERSPEECH 2017*, pp. 2769-2773, 2017.
[23] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 12, pp. 2323-2366, 2017.
[24] A. M. Eid and J. W. Wallace, "Ultrawideband speech sensing," *IEEE Antennas Wireless Propagation Letters*, Vol. 8, pp.1414-1417, 2009.
[25] Y. H. Shin and J. Seo, "Towards contactless silent speech recognition based on detection of active and visible articulators using an IR-UWB Radar," *Sensors*, Vol. 16, No. 11, p. 1812, 2016.
[26] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason D. Povey, V. Valtchev, and P. Woodland, "The HTK book – version 3.2," Cambridge University Engineering Department, U.K., 2002.
[27] M. Morchid, R. Dufour, and G. Linaré, "Impact of word error rate on theme identification task of highly imperfect human–human conversations," *Computer Speech and Language*, Vol. 38, pp. 68-85, 2016.
[28] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with gabor feature extraction," *Proc. of ICSLP*, pp. 25-28, 2002.