2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

VOTE400(Voide Of The Elderly 400 Hours): A Speech Dataset to Study Voice Interface for Elderly-Care

Minsu Jang[1], Sangwon Seo[2], Dohyung Kim[1], Jaeyeon Lee[1] and Jaehong Kim[1] and Jun-Hwan Ahn[2]

***Abstract*— This paper introduces a large-scale Korean speech dataset, called VOTE400, that can be used for analyzing and recognizing voices of the elderly people. The dataset includes about 300 hours of continuous dialog speech and 100 hours of read speech, both recorded by the elderly people aged 65 years or over. A preliminary experiment showed that speech recognition system trained with VOTE400 can outperform conventional systems in speech recognition of elderly people's voice. This work is a multi-organizational effort led by ETRI and MINDs Lab Inc. for the purpose of advancing the speech recognition performance of the elderly-care robots.**

I. INTRODUCTION

Voice interface is the most intuitive, comfortable and universal interface for interacting with service robots. Recent advancement of commercial cloud-based speech-to-text (STT) services allowed devising a voice interface for service robots a very simple process of integrating a service API for STT into the robot SW system.

While these commercial systems work very well with adults in the ages of between 20 and 60, it easily fails with voices from older adults aged 65 years or over. It is known that speech signals from older adults bring about difficulties for automated speech recognition as they tend to be imprecise in consonant pronunciation, include tremors, and have slower articulations [1].

In the need to develop a speech recognition system that are specialized to the speech signals from older adults, we built a speech dataset by collecting large-scale dialogue and read speech from older adults. The result of our effort is 400 hours of Korean speech data which we named as 'VOTE400 (**V**oice **O**f **T**he **E**lderly **400** Hours) and open-sourced for any non-commercial research projects (`https://ai4robot.github.io/mindslab-etri-vote400`).

II. DATASET DESCRIPTION

A. Dataset Collection

For recruiting and collecting voice data from older adults, we got assistance from a Korean governmental office, called Dok-Geo-No-In-Jong-Haap-Ji-Won Center (Ji-Won Center: `http://www.1661-2129.or.kr/index.html`) devoted to the support of older adults living alone. With the support from the Ji-Won Center, we could collect a large-scale dialog speech and read speech from a number of older adults across various regions of South Korea.

[1]Minsu Jang, Dohyung Kim, Jaeyeon Lee and Jaehong Kim are with Electronics and Telecommunications Research Institute, Daejeon-si, South Korea `minsu at etri.re.kr`

[2]Sangwon Seo and Jun-Hwan Ahn is with MINDs Lab Inc., Kyungki-do, South Korea `asdn9353 at mindslab.ai`

1) Dialog Speech: To collect spontaneous speech data from older adults, we could utilize a support program of Ji-Won Center called Saa-Raang-It-Gi where social workers regularly visit elderly people's homes for consulting on health-related issues and relieving loneliness. After explaining about the data collection experiment and getting consent of participation from the elderly, conversations between a social worker and an elderly were recorded using a smartphone.

The recordings from these program sessions were sent to Ji-Won Center and a screening process was performed to remove every dialogue involving sensitive personal information. Then, a quality assurance process was followed to filter out speech segments incomprehensible by human listener due to imprecise pronunciation or significant noise.

2) Read Speech: To amend the relatively low-quality of the dialog speech dataset, we launched another data collection process to acquire read speech from older adults. We built and utilized in the process a dedicated speech collection system, where a tablet-based client program presents a sentence to read to an elderly user; makes a recording and sends it to a server; where the recording is inspected to be accepted or not. In total, the number of unique sentences chosen to be read by participants was 2,250. These sentences were selected by considering how often these could be casually uttered by older adults in daily lives.

TABLE I
RAW DATA COLLECTION OF DIALOG SPEECH

Region(R)	No. Participants	Len. (hrs)
Seoul-si(SE)	620	122
Busan-si(PS)	242	90
Daegu-si(DG)	202	33
Gwangju-si(GJ)	179	63
Daejeon-si(DJ)	275	66
Ulsan-si(WS)	80	28
Goyang-si(GG)	335	69
Gangwon-do(GW)	178	45
Chungcheongbuk-do(CB)	252	92
Chungcheongnam-do(CN)	317	46
Jeollanam-do(JN)	323	103
Gyeongsangbuk-do(GB)	378	116
Total	3,381	873

B. Dialog Speech Data

The total number of elderly participants is 3,381 and the total length of recordings is 873 hours. This is the result of collective efforts by regional senior citizens welfare institutes

collaborating with the Ji-Won Center. Table I shows the regional distributions of all the participants and the length of recordings per region.

After the screening and the QA process mentioned in the previous subsection, we finalized 300 hours of dialog speech to be included in the VOTE400 dataset. Transcription for every final speech data was done by human annotators. In VOTE400, we provide for a recording session in a WAV file. The audio format of the WAV file is as shown in table II. Every WAV file is accompanied by a transcription text file encoded in ISO-8859. The transcription does not include audio-text alignment information.

TABLE II
VOTE400 DIALOG SPEECH AUDIO FORMAT

Property	Value
Format.	PCM
Format Settings	Little/Signed
Codec ID	1
Bit Rate Mode	Constant
Bit Rate.	256
Channel(s)	1
Sampling Rate	16 kHz
Bit Depth	16 bits

The file name of each recording follows the pattern of `<P-ID>_<G>_<A>_<R>_<DT>`, where `P-ID` is a unique participant ID; `G` is a gender value (`F` for female, `M` for male), `A` is a age value, `R` is a regional code, and `DT` is the data-time of the recording session.

Participants and speech audio statistics for VOTE400 dialog dataset are shown in table III and table IV.

TABLE III
DEMOGRAPHICS OF VOTE400 DIALOG SPEECH

Region(R)	No. Participants	Age (μ/σ)
Seoul-si(SE)	251(F:210,M:41)	78.98/5.13
Daegu-si(DG)	108(F:95,M:13)	80.33/6.08
Gyoungki-do(GG)	110(F:83,M:27)	80.17/5.41
Chungcheongnam-do(CN)	6(F:6,M:0)	77.00/3.69
Jeollanam-do(JN)	70(F:56,M:14)	80.76/4.90
Busan-si(PS)	160(F:137,M:23)	78.70/5.51
Daejeon-si(DJ)	96(F:72,M:24)	78.81/5.24
Gangwon-do(GW)	109(F:94,M:15)	80.07/5.50
Gyeongsangbuk-do(GB)	98(F:95,M:3)	80.87/4.48
Gwangju-si(GJ)	87(F:70,M:17)	79.39/5.77
Chungcheongbuk-do(CB)	17(F:17,M:0)	80.47/5.51
Ulsan-si(WS)	58(F:49,M:9)	76.97/4.48
Total	1,170(F:984,M:186)	79.47/5.37

C. Read Speech Data

The total number of elderly participants is 104 and the total length of recordings is 100 hours. Table VI shows the statistics of VOTE400 read speech data.

Audio format of the VOTE400 read speech data is as shown in V, which is slightly different from the format of the dialog speech data.

TABLE IV
SPEECH AUDIO STATISTICS FOR VOTE400 DIALOG SPEECH

Region(R)	Len.(secs)	Len.(μ/σ)
Seoul-si(SE)	151,010	601.63/239.83
Daegu-si(DG)	60,740	562.42/228.14
Gyoungki-do(GG)	107,935	981.23/357.19
Chungcheongnam-do(CN)	5,193	865.62/293.98
Jeollanam-do(JN)	81,767	1,168.10/294.85
Busan-si(PS)	200,207	1,251.30/255.85
Gangwon-do(GW)	95,420	875.42/158.18
Daejeon-si(DJ)	123,138	1,282.70/293.83
Gyeongsangbuk-do(GB)	71,175	726.28/308.80
Gwangju-si(GJ)	92,699	1,065.52/276.53
Chungcheongbuk-do(CB)	20,135	1,184.41/309.54
Ulsan-si(WS)	70,754	1,219.90/254.43
Total	1,080,179	923.23/380.17

TABLE V
VOTE400 READ SPEECH AUDIO FORMAT

Property	Value
Format.	PCM
Format Settings	Little/Signed
Codec ID	1
Bit Rate Mode	Constant
Bit Rate.	705.6 kb/s
Channel(s)	1
Sampling Rate	44.1 kHz
Bit Depth	16 bits

The file name of read speech data follows the pattern of `PID_<P-ID>_<DATE>_<SENTENCE-NO>_<R>`, where `P-ID` is a unique participant ID, `DATA` is the date of recording, `SENTENCE-NO` is a serial number put to each of the recorded sentences, and `R` is the region code as shown in table VI. Each WAV file contains a single sentence, accompanied by a transcription text file encoded in `EUC-KR`.

Though the number of sentences chosen and presented to the participants was originally 2,250, the final total number of unique sentences in VOTE400 read speech data is 7,832, due to mistakes and slight variations in real utterances by older adults.

III. PRELIMINARY EXPERIMENT

We conducted a preliminary experiment by training a MINDs Lab Inc.'s proprietary baseline speech recognizer(`M`), which is based on LSTM architecture, and estimating the STT accuracy using VOTE400. After fine-tuning the baseline with 50 hours each of dialog speech data and read speech data of VOTE400, a simple test with 100 sentences from different regions was performed and the results are as shown in table VII, along with the results when the sentences were tested on a commercial cloud-based STT engine(`C`).

IV. SUMMARY

We described a Korean speech dataset VOTE400 which is collected entirely from older adults of more than 75 years old. VOTE400 contains 300 hours of dialogue speech data and 100 hours of read speech data, with proficient varieties

TABLE VI

REGIONAL DISTRIBUTIONS OF VOTE400 READ DATASET

Region(G)	No. Persons	No. Sent.	Len.(μ/σ)
Gyeongsangnam-do(GB)	20	22,575	3.18/1.38
Seoul-si(SE)	18	19,220	3.31/1.49
Jeollanam-do(JN)	21	21,393	3.36/1.52
Daegu-si(DG)	25	26,950	3.60/1.87
Gangwon-do(GW)	20	21,676	2.73/1.12
Total	104	111,814	3.25/1.54

TABLE VII

STT PERFORMANCE TEST RESULTS WITH VOTE400

Region(G)	Gender	Acc. M (%)	Acc. C (%)
SE	M	90	90
SE	F	90	80
GW	M	80	90
GW	F	90	80
DG	M	70	80
DG	F	90	80
GN	M	90	80
GN	F	80	80
JN	M	70	50
JN	F	80	60

in gender and regions. To our knowledge, VOTE400 is by far one of the largest voice datasets that is oriented to voices of the elderly. We hope that this dataset will be useful to study older adult's voice features and realize voice technologies that work sufficiently well in elderly-care robotics.

ACKNOWLEDGMENT

This work was supported by the Institute of Information communi-cations Technology Planning Evaluation(IITP) grant funded by the Koreagovernment(MSIT) (No. 2017-0-00162, Development of Human-care RobotTechnology for Aging Society)

REFERENCES

[1] Vacher, M., Aman, F., Rossato, S. and Portet, F., 2015, August. Development of automatic speech recognition techniques for elderly home support: Applications and challenges. In International Conference on Human Aspects of IT for the Aged Population (pp. 341-353). Springer, Cham.