Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6578–6585
Marseille, 11–16 May 2020

Improving Speech Recognition for the Elderly: A New Corpus of Elderly Japanese Speech and Investigation of Acoustic Modeling for Speech Recognition

Meiko Fukuda[1], Hiromitsu Nishizaki[2], Yurie Iribe[3], Ryota Nishimura[1], Norihide Kitaoka[4]
[1]Tokushima University, Tokushima, Japan
[2]University of Yamanashi, Kofu, Japan
[3]Aichi Prefectural University, Nagoya, Japan
[4]Toyohashi University of Technology, Toyohashi, Japan
fukuda.meiko@tokushima-u.ac.jp, hnishi@yamanashi.ac.jp, iribe@ist.aichi-pu.ac.jp, nishimura@is.tokushima-u.ac.jp, kitaoka@tut.jp

Abstract

In an aging society like Japan, a highly accurate speech recognition system is needed for use in electronic devices for the elderly, but this level of accuracy cannot be obtained using conventional speech recognition systems due to the unique features of the speech of elderly people. S-JNAS, a corpus of elderly Japanese speech, is widely used for acoustic modeling in Japan, but the average age of its speakers is 67.6 years old. Since average life expectancy in Japan is now 84.2 years, we are constructing a new speech corpus, which currently consists of the utterances of 221 speakers with an average age of 79.2, collected from four regions of Japan. In addition, we expand on our previous study (Fukuda, 2019) by further investigating the construction of acoustic models suitable for elderly speech. We create new acoustic models and train them using a combination of existing Japanese speech corpora (JNAS, S-JNAS, CSJ), with and without our 'super-elderly' speech data, and conduct speech recognition experiments. Our new acoustic models achieve word error rates (WER) as low as 13.38%, exceeding the results of our previous study in which we used the CSJ acoustic model adapted for elderly speech (17.4% WER).

Keywords: Japanese speech corpus, elderly, acoustic modeling

1. Introduction

In recent years in Japan, the use of the internet has increased among the elderly. A survey conducted by the Ministry of Internal Affairs and Communications found that the percentage of elderly people who searched for information on the internet within the past year was 76.6% for those aged 60-69, 51.0% for those aged 70-79, and 21.5% for those aged 80 and older (Ministry of Internal Affairs and Communications, 2019). However, as their visual acuity and hand motor function declines, it becomes increasingly difficult for the elderly to operate PCs and smartphones, which could result in a digital divide between the elderly and the younger generation. As governmental and corporate services are increasingly provided via the internet, this access problem will become more severe. We believe that speech recognition technology is a promising modality for addressing this problem, allowing the elderly to maintain access to information through their electronic devices.

Furthermore, as society ages labor shortages are becoming apparent in Japan. As one countermeasure, a joint project between the Ministry of Health, Labor and Welfare and the Ministry of Economy, Trade and Industry is underway to promote the development and deployment of robots for assisting the elderly with their daily activities. In 2017, assisted living devices using robot technology for communication with the elderly was added to the list of priorities for this project (Ministry of Health, Labor and Welfare, 2017). Since speech recognition technology is essential for these robots to function effectively, improving the accuracy of recognition of elderly speech has become an urgent issue since conventional speech recognition technology has not demonstrated sufficient accuracy when processing elderly speech.

One significant factor in the recognition of elderly speech may be the acoustic models used for speech recognition, which are created using the voices of younger adults (Wilpon, 1996; Anderson, 1999; Vipperla. 2008; Pellegrini, 2012; Fukuda, 2019). Many studies have reported various characteristics of elderly voices which differ from those of younger people, such as hoarseness, slower speech, extended pauses between words, unclear pronunciation and phonological changes due to dementia, etc. (Toth, 2018; Miyazaki, 2010; Winkler, 2003). Therefore, in order to create acoustic models suitable for the recognition of elderly speech, a large-scale elderly speech corpus is required. In Japan, the Japanese Newspaper Article Sentences Read Speech Corpus of the Aged (S-JNAS) has been widely used in studies for more than ten years. The average age of S-JNAS participants is 67.6 years old, but Japan's life expectancy has risen to 84.2. There is also a wide range of ages among S-JNAS speakers. Therefore, we decided to build a super-elderly speech corpus for acoustic modeling, following the methodology of the S-JNAS corpus.

We have already collected recordings from four regions of Japan, and have created a database of 221 elderly speakers containing 11,107 utterances. However, the amount of speech data we have collected, as well as the number of regions sampled, are not yet sufficient for accurate acoustic modeling, therefore we are continuing to reach out to nursing care facilities for additional participants. We plan to make this corpus publicly available through the National Institute of Informatics in Japan. We hope that many other investigators will use our new corpus for research on speech recognition for the elderly.

In this study, we also conducted speech recognition experiments using our corpus. Since we still do not have enough speech data to create original acoustic models for the elderly, we constructed acoustic models using the combined speech data of three existing speech corpora (JNAS, S-JNAS, CSJ) in addition to the elderly speech data in our corpus, achieving a word error rate (WER) of 13.21%. This result exceeded the recognition accuracy we reported in our previous study, in which CSJ-based

acoustic models were adapted to elderly speech, achieving a WER of 17.4%. Scatter plots were created using the WER for each speaker, and the mild relationship between age and WER was examined. Although the number of speakers was insufficient, a mild positive correlation was found between age and WER (correlation coefficient: 0.59). In other words, the accuracy of speech recognition tended to decrease as the age of the speaker increased.

In the future, we would like to do further research on acoustic modeling and speaker adaptation, while continuing to record elderly speech in various regions of Japan in order to increase the scale of our corpus.

2. Construction of Our Corpus

2.1 Recording Area Selection

There are 16 dialects in Japan, and these dialects have been found to effect the accuracy of speech recognition (Kudo, 1996). Therefore, when selecting recording areas for our corpus, we sought to include a balanced distribution of Japanese dialects. So far, we have recorded the utterances of elderly participants in Nagoya, Tokushima, Yamagata, and Nagasaki, and we are currently recording in Kisarazu and Suzuka (Fig. 1). We plan to record more speech in other parts of Japan as well.

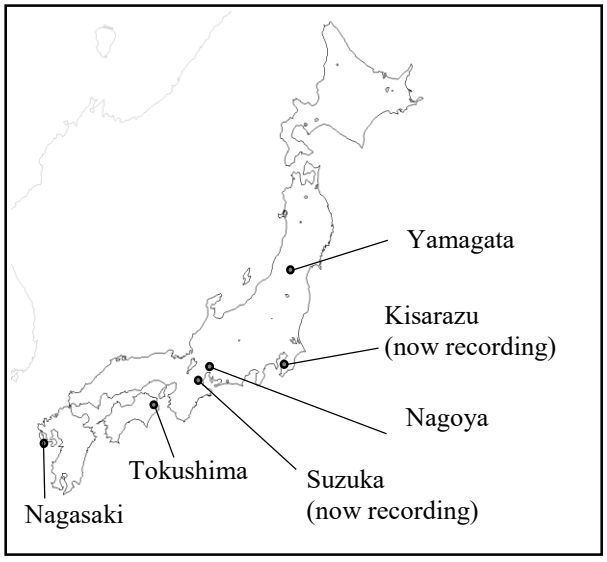

Figure 1: Areas where elderly speech data was recorded

2.2 Speaker Selection

We have only two criteria for selecting participants, their age and health status. In the currently available database of elderly Japanese speech, S-JNAS, the average age of the speakers is 67.6 years old, and there are only 8 speakers over 80 among a total of 301 speakers (Baba, 2001). In order to include more speech from older age groups in our corpus than S-JNAS, we tried to gather speech from as many speakers over 80 years old as possible. At present, the average age of the speakers in our corpus is 79.2 years old, which is significantly higher than S-JNAS. Figure 2 compares the age distribution of participants in S-JNAS and in our corpus. Table 1 shows the age and sex distribution of all of the speakers in our corpus. Table 2 shows the average age of the speakers in each recording area. Tables 3 to 6 show the age and sex distributions for each recording area.

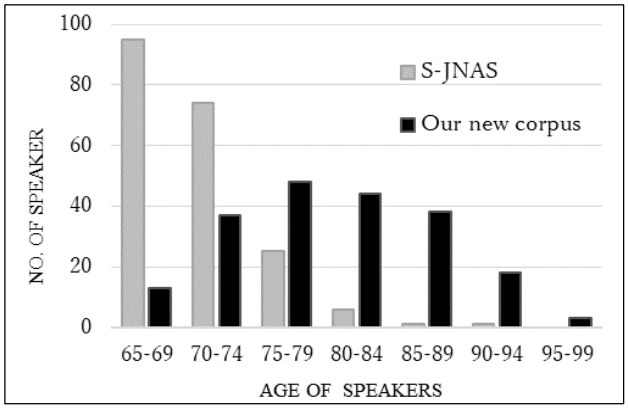

Figure 2: Comparison of the age distributions of speakers in S-JNAS and in our new corpus

Age	Male	Female	Total
65-69	6	7	13
70-74	23	34	57
75-79	17	31	48
80-84	10	34	44
85-89	7	31	38
90-94	6	12	18
95-99	1	2	3
Total	**70**	**151**	**221**

Table 1: Age and sex distribution of speakers in our corpus

Recording Area	Male	Female	Total Avg. of Each Area
Nagoya	82.9	81.6	81.7
Tokushima	83.2	86.6	85.5
Yamagata	73.4	NA*	73.4
Nagasaki	75.3	76.2	76.0
Total	**78.7**	**81.5**	**79.2**

** not available*

Table 2: Average age of male and female speakers in each recording area, and total average age for each region

Age	Male	Female	Total
65-69	2	1	3
70-74	0	7	7
75-79	3	10	13
80-84	5	16	21
85-89	1	10	11
90-94	3	4	7
95-99	1	1	2
Total	**15**	**49**	**64**

Table 3: Age and sex distribution of speakers in Nagoya

Because we think it important that people with various health conditions are able to use speech recognition technology, our speakers include people with various health conditions, such as a tendency to dementia, poor vision and use of dentures. We did not use the living arrangements of potential participants as a selection criterion, so there are speakers in our corpus who live in nursing homes, who use daycare services only during the day, and who do not use care facilities at all. We also did not consider the gender ratio of our participants. As a result, currently the total number male participants in our corpus is 70, and the total number of female participants is 151 (Table 1).

Age	Male	Female	Total
65-69	0	1	1
70-74	1	0	1
75-79	4	1	5
80-84	3	4	7
85-89	3	13	16
90-94	2	6	8
95-99	0	1	1
Total	**13**	**26**	**39**

Table 4: Age and sex distribution of speakers in Tokushima

Age	Male	Female	Total
65-69	4	5	9
70-74	13	27	40
75-79	10	20	30
80-84	1	14	15
85-89	3	8	11
90-94	1	2	3
95-99	0	0	0
Total	**32**	**76**	**108**

Table 5: Age and sex distribution of speakers in Nagasaki

Age	Male	Female	Total
65-69	0	0	0
70-74	9	0	9
75-79	0	0	0
80-84	1	0	1
85-89	0	0	0
90-94	0	0	0
95-99	0	0	0
Total	**10**	**0**	**10**

Table 6: Age and sex distribution of speakers in Yamagata

Since the number of people who volunteered to participate in our data collection project was limited, in all regions our test speakers were the same participants as our training speakers. We were only able to record five speakers in Tokushima, and only ten in Yamagata. In contrast, we had a sufficient number of participants in Nagasaki, but to match the conditions in Tokushima and Yamagata we also recorded the Nagasaki test and training data using the same speakers. Moreover, we were unable to record any test data in Nagoya, therefore we did not do any speech recognition experiments with the Nagoya data. However, we did use the recorded training data from Nagoya, in addition to the data from the other regions, when building our acoustic models.

2.3 Selection of Japanese Sentences

The design of our corpus is based on S-JNAS, so we also used the ATR 503 sentences and JNAS newspaper article sentences as the script for our participants. However, unlike S-JNAS, we used the ATR 503 sentences as training data and the newspaper article sentences as test data.

The ATR 503 sentences are phonetically balanced Japanese sentences extracted from newspapers, textbooks, journals, letters, novels, and so on. They include 402 two-phoneme sequences and 223 three-phoneme sequences (625 items in total), and reproduce the phonetic balance that appears in Japanese as much as possible. The ATR 503 sentences are divided into 10 text sets (Sets A – J) (Kurematsu, 1994; NII-SRC, 2019).

For the S-JNAS corpus, each of the training data speakers read aloud two sets of ATR 503 sentences (about 100 sentences) and one set of the newspaper article sentences (about 100 sentences). However, for our corpus, since many of our speakers are very elderly, and some have limited vision or a tendency towards dementia, we limited the number of sentenced we asked each participant to read in order to reduce the burden. But since we wanted to record as many Japanese phonemes as possible, we selected the ATR 503 sentences as the training data and assigned each participant only one set of sentences (50 or 53 sentences). The total number of utterances of all of the speakers is 11,107. Table 7 shows the number of utterances for each set.

Set Name (Number of sentences)	Male	Female	Total
Set A (50)	7 (350)	15 (750)	22 (1,100)
Set B (50)	8 (400)	14 (700)	22 (1,100)
Set C (50)	9 (450)	15 (750)	24 (1,200)
Set D (50)	5 (250)	18 (900)	23 (1,150)
Set E (50)	10 (500)	12 (600)	22 (1,100)
Set F (50)	5 (250)	18 (900)	23 (1,150)
Set G (50)	8 (400)	15 (750)	23 (1,150)
Set H (50)	8 (400)	14 (700)	22 (1,100)
Set I (50)	6 (300)	15 (750)	21 (1,050)
Set J (53)	4 (212)	15 (735)	19 (1,007)
Total Utterances [Number of speakers]	**3,512 [70]**	**7,595 [151]**	**11,107 [221]**

Table 7: Number of utterances in each sentence set

The JNAS newspaper article sentences were extracted from the Japanese *Mainichi* daily newspaper. There are 155 text sets consisting of 1,000 total sentences in the JNAS, but we extracted 50 of these sentences and divided them into 5 sets of 10 sentences each (Sets T1 - T5). For the same reason mentioned above, we reduced the number of sentences to be read by each participant to only 1 set (10 sentences).

Recording Areas	Nagoya, Tokushima, Yamagata, Nagasaki
Subjects	221 subjects Average age: 79.2
Subject selection	– as old as possible – healthy enough to participate without difficulty – those with a tendency for dementia or with other health issues were included
Training data	ATR phonetically balanced sentences 50 or 53 sentences per person
Testing data	JNAS newspaper sentences 10 sentences per person*
Sampling frequency	16kHz
Contents	a. recorded speech data divided into sentence units b. three text transcriptions per person: – transcribed faithfully to pronunciation – printed in hiragana, katakana or kanji characters with hiragana transcription c. demographic information about the speaker (age, gender, recording location,)

** except in Nagoya*

Table 8: Overview of our corpus

2.4 Data Collection Procedure

We recorded most of the read speech of our elderly participants at elderly care facilities, but some were recorded at town halls or a university. A typical recording scene at a nursing home is shown in Figure 3.

We explained the purpose of our research to each volunteer before recording, and only those who agreed participated.

Figure 3: Typical recording scene at a nursing home

At the beginning of each session, we explained the recording procedure and distributed printouts of the ATR 503 sentences and JNAS newspaper article sentences (except in Nagoya, where only the ATR 503 sentences were used). All of the sentences were printed in both standard Japanese and in hiragana characters. We began recording after a few minutes of practice reading aloud.

In consideration of the physical condition of the participants, speakers were allowed to rest, or to terminate the recording session, at any time. Additionally, after each recording session a subjective mood evaluation survey was conducted with each participant by members of the facility staff, and their likelihoods of suffering from dementia were assessed using the HDS-R (Hasegawa's Dementia Scale-Revised) (Imai, 1994), except in Yamagata.

2.5 Devices for Recording

The devices that we used to collect utterances were a desktop microphone (Audio-Technica, AT9930) and a linear PCM recorder (TASCAM, DR-05 VERSION2) in Nagoya, or a lapel microphone (SONY, ECM-88B), a desktop microphone and an 8-track field recorder (TASCAM, DR-680MKII) in the other areas.

2.6 Database

The total duration of the recorded speech in our corpus is approximately 21.7 hours. We digitized the speech waves at a sampling frequency of 16 kHz using 16-bit audio, and divided the speech data into sentence units. At the beginning and end of each sentence, we included a silent pause of about 150 msec as often as possible. Trained employees transcribed the recorded speech data manually into text data because the speakers often shuttered, uttered fillers or misread the sentences. The phonemes and words of the sentences were rewritten to correspond to actual phonation. The database also includes demographic and assessment information about the speakers (age, gender, recording location). It also includes the Set ID (A-J or T1-T5) of the sentences. Table 8 shows an overview of our corpus.

Eval. Data	JNAS	S-JNAS	CSJ
Baseline	25.53	21.85	27.25
Adaptation with our corpus	21.57	20.24	17.42

Table 9: WERs (%) for speech recognition experiments using JNAS, S-JNAS and CSJ trained acoustic models and BCCWJ language model, with and without acoustic adaptation

3. Speech recognition experiments

In our previous study (Fukuda, 2019), we used existing Japanese speech corpora to create the baselines for acoustic models, since the amount of speech data in our corpus is not yet sufficient for creating original acoustic data. The three speech corpora we tested included different types of data, to see whether the use of different acoustic models influenced speech recognition accuracy for our corpus test data: adult speech (JNAS), elderly speech (S-JNAS) and spontaneous speech (CSJ). JNAS is a famous speech corpus widely used for constructing acoustic models for standard Japanese adult speech. The 306 participants were mostly 20-49 years old (Itou, 1999). Each participant read aloud one set of newspaper article sentences (100 sentences) and one set of phoneme balanced sentences (about 50 sentences) for training data. The total recording time is 72.4 hours. The S-JNAS corpus was described previously in Section 2.2 of this paper. There were 301 participants and the total recording time is 133.4 hours. The Corpus of Spontaneous Japanese (CSJ), is a corpus mostly consisting of speech from academic conferences and simulated public speaking (Furui, 2000). There are 1,417

speakers, with a total training data recording time of about 520.8 hours. The CSJ corpus does not include super-elderly speakers, but the speaking style is spontaneous so fillers, shuttering and rephrasing are often observed. In contrast, our speakers include the super-elderly (Table 1 and Figure 2), as well as people with a tendency for dementia, poor vision, dentures, weak exhalation and non-standard personal speaking habits, such as uttering many fillers. As a result, many of our speakers are not very fluent. There was also a wider variety of speaking styles in our corpus than in the JNAS and S-JNAS corpora, so we thought that the spontaneous speaking style contained in the CSJ corpus might be more similar to utterances of our speakers.

As in our previous study, we created our language model using the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa, 2014), which is the most extensive corpus of contemporary written Japanese. These sentences are collected from various publications such as books, magazines, newspapers and web sites, and they include a wide variety of vocabulary and syntax. To construct our speech recognizer, we used training scripts based on the CSJ recipe in the Kaldi toolkit (Povey, 2011). DNN-HMMs were used for the acoustic models. The DNN used for the DNN-HMM was a simple feed-forward network (nnet1).

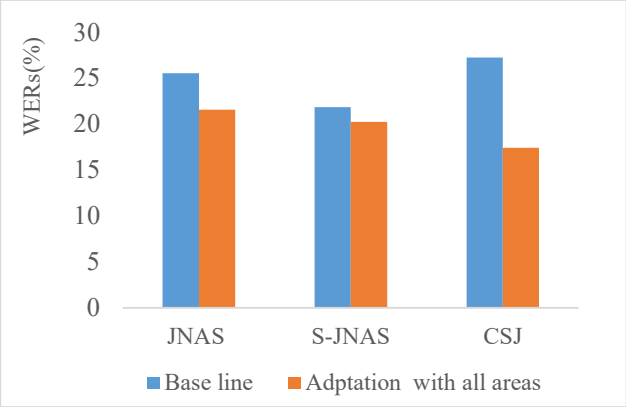

Figure 4: WERs (%) for speech recognition experiments using JNAS, S-JNAS and CSJ trained acoustic models and BCCWJ language model with or without acoustic adaptation

The speech recognition experiment was performed in three stages. The first step was to create baseline acoustic models for JNAS, S-JNAS and CSJ. In the second stage, each set of acoustic models was adapted to elderly speech using our corpus, based on transfer training using back propagation. As a third step, we re-adapted the acoustic models created in the second stage using speech data from each region except Nagoya, to adapt the models to each dialect. (We did not record any test data in Nagoya.) Word error rates (WERs) at each stage were calculated. S-JNAS achieved the lowest WER of 21.85% among the baseline acoustic models in the first stage. However, the CSJ acoustic model achieved the lowest WER value after adaptation for elderly voices (Fig. 4 and Table 9), and the same tendency was observed in the third stage. Table 10 shows the re-adaptation results using data from the CSJ models, which were the best results among the three acoustic models , which were the best results among the three acoustic models, which were the best results among the three acoustic models.

Eval. Data	Baseline (CSJ)	Adaptation with our corpus	Re-adaptation for each area
Nagasaki	20.12	15.03	12.80
Tokushima	59.31	34.14	32.76
Yamagata	12.73	7.96	6.37
Total	**27.25**	**17.42**	**15.59**

Table 10: WERs (%) for speech recognition experiments using CSJ trained acoustic models and BCCWJ language model with two types of acoustic adaptation

In order to develop more effective methods of training acoustic models for elderly speech, thereby achieving better speech recognition results, in this study we created the baseline acoustic models using large amounts of data. We used the combined JNAS, S-JNAS and CSJ corpora (with total recording times of 72.4, 133.4 and 520.8 hours, respectively), as well as our own corpus of speech data (21.7 hours), to create acoustic models we called “mixed” models. To examine the effect of including our own, more elderly speech data, we also created other acoustic models using the combined JNAS, S-JNAS, and CSJ corpora which did not include our speech data. We call these acoustic models “mixed w/o our corpus” models and compared them with the “mixed” models. The results of our speech recognition experiments are shown in Table 11 and Figure 5.

Eval. Data	**Mixed w/o our corpus**	**Mixed**
Nagasaki	11.61	10.50
Tokushima	31.03	27.82
Yamagata	5.57	5.84
Total	**14.40**	**13.21**

Table 11: WERs (%) for speech recognition experiments using acoustic models trained with “mixed w/o our corpus” data and “mixed” (all corpora) data, using a BCCWJ-based language model

The WER of the CSJ-trained acoustic model evaluated in our previous study was 27.25%, and when adapted to elderly speech the WER was 17.4%. In comparison, the WER for the acoustic model proposed in this study, which was trained using the “mixed” data from three large speech corpora and our own super-elderly data was 13.21%, which was much better than the method proposed in our previous study. Based on these results, acoustic models trained with large amounts of speech data from multiple corpora may be more effective for elderly speech recognition, even without speaker adaptation in our experimental procedure. The WER for our “mixed without our data” model (trained with

data from multiple corpora, but without our super-elderly speech data) was 14.40%. Although the total audio time of

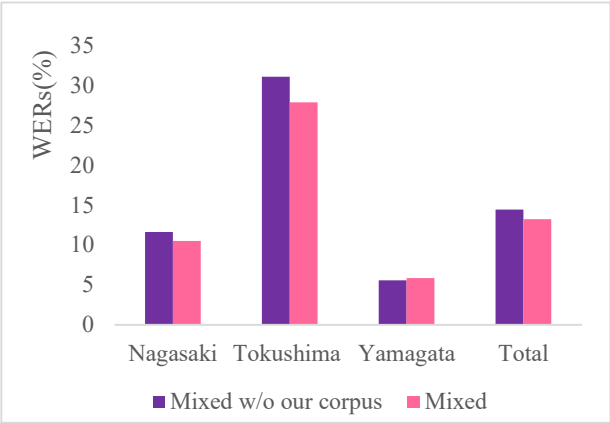

Figure 5: WERs (%) for speech recognition experiments using acoustic models trained with "mixed w/o our corpus" data and "mixed" (all corpora) data, using a BCCWJ-based language model

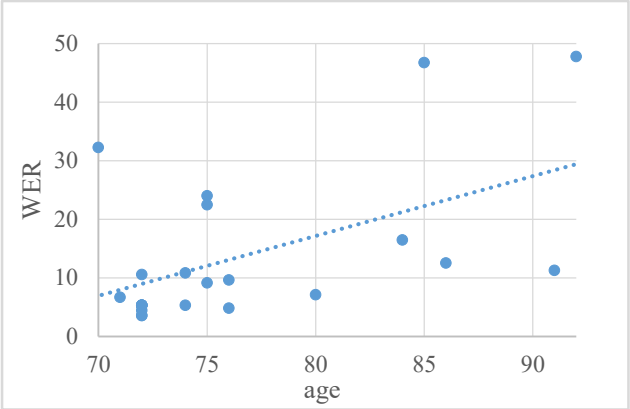

Figure 6: Correlation between age and WER in the mixed model without our corpus (corr. = 0.53)

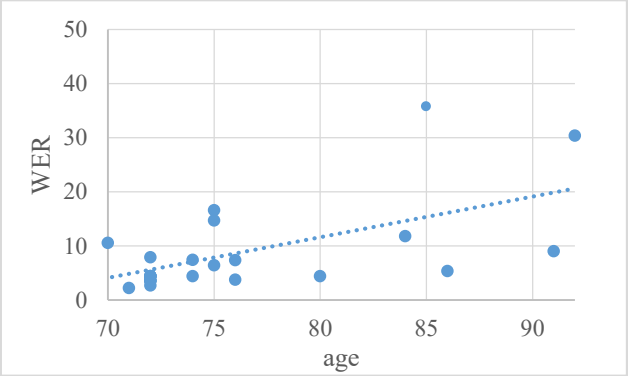

Figure 7: Correlation between age and WER in the mixed model (corr. = 0.59)

our corpus is only 21.7hr, much smaller than the total audio time of the combined JNAS, S-JNAS and CSJ corpora, there was a slight improvement in performance when our super-elderly speech data was added to the training data of the acoustic models.

Considering the experimental results by region, all of our recognition results for the speakers from Tokushima were worse than for speakers from the other areas. The first likely reason for this outcome is the high average age of the Tokushima participants (85.5 years). Because of their ages, many of the participants from Tokushima were not fluent when reading aloud. A second likely reason for the lower recognition results is also partially related to the participants higher ages. Unlike in the other areas, there was insufficient coaching of the Tokushima participants by the recording staff, such as prompting them to read the text more carefully when they made mistakes, or having them re-read the text aloud when they made serious errors.

In contrast, speech from the Yamagata speakers obtained the best recognition results, and this may have been because the average age of the participants was the lowest, at 73.4 years, which may have helped them to read aloud more fluently. Moreover, there were only 10 participants from Yamagata, so the training and test speakers were the same people. For this reason, in the future we will need to conduct an additional experiment involving a larger number of Yamagata participants.

Eval. Data	Mixed w/o our corpus	Nagasaki adapted	Tokushima adapted	Yamagata adapted
Nagasaki	11.61	**10.99**	13.66	11.61
Tokushima	31.03	32.18	**27.97**	36.21
Yamagata	5.57	6.37	7.87	**4.86**
Total	**14.40**	**12.39***		

* *The average of the values (**bold** numbers) obtained using adaptation to the dialects of each region.*

Table 12: WERs (%) for speech recognition experiments using acoustic models trained with mixed w/o our corpus corpora and adaptation to regional speech, using the BCCWJ language model

We created scatter plots of age and word error rate for each speaker and calculated the correlation coefficients for the mixed model without our corpus (Fig. 6), and for the mixed model (Fig. 7). Even though the number of speakers included in our corpus is insufficient to draw a firm conclusion, the word error rate tended to increase with the age of the speaker. The correlation coefficient (Pearson product-moment correlation coefficient) was 0.53 for the mixed model without our corpus and 0.59 for the mixed model, both showing mild correlations between age and the word recognition error rate. As for the slope of the regression line, the slope of the mixed model is slightly more gradual than that of the mixed model without our corpus, suggesting that the inclusion of our super-elderly speech data resulted in only a small reduction in the tendency of WERs to increase with increasing speaker age. We want to observe the behavior of these correlations in more detail in a future study by increasing the number of elderly speakers included in our corpus.

Furthermore, we adapted the mixed w/o our corpus model using data from speakers with regional dialects, the WER decreased only slightly for the Nagasaki and Yamagata dialect. In Tokushima, the WER decreased from 31.03% to

27.97%, so an obvious adaptation effect was observed despite the poor level of recognition accuracy. Even in comparison with the mixed model, the WER in the Tokushima dialect after adaptation was significantly lower than the mixed model, but there was no significant difference between the other two dialects results. Further study on dialect adaptation is needed in order to improve recognition accuracy more consistently, as observed in these two regions.

4. Conclusion

In this paper we have described our on-going effort to construct a new corpus of Japanese super-elderly speech for acoustic modeling. So far, we have collected 11,107 utterances from 221 people living in the Nagoya, Tokushima, Yamagata and Nagasaki regions of Japan, with a total recording time of about 21.7 hours, organized into a database. The average age of our speakers is 79.2 years old, much older than the average age of speakers in S-JNAS, the existing Japanese corpus of elderly speech, whose participants have an average age of 67.6 years. We believe our new database will be useful for creating acoustic models suitable for the super-aged. We are currently recording in another region, and plan to record in more areas in order to achieve a diverse, large-scale corpus.

In our previous paper, we performed speech recognition experiments using existing Japanese speech corpora (JNAS, S-JNAS and CSJ), applying speaker adaptation to elderly speech, using our data, to the acoustic models created using each corpus. Among these three corpora, the CSJ-trained acoustic models achieved the best results (a WER of 17.4%).

In this study, we tried to create more effective acoustic models for elderly speech recognition without using speaker adaptation, by combining existing Japanese speech corpora with our data for model training. We hypothesized that recognition accuracy would increase if the volume of speech data used to create the acoustic models was increased, so we combined the existing Japanese speech corpora (JNAS, S-JNAS, CSJ) and our new corpus of super-elderly speech for training new acoustic models. Using this approach, we were able to achieve a WER of 13.21%, which was better than the CSJ-trained acoustic models adapted to elderly voices proposed in our previous study. To determine the effect of using elderly speech when training the acoustic model, we also experimented with acoustic models created using data only from the three existing corpora, and achieved a WER of 14.40%. Although the total audio time of our corpus was much smaller than those of the JNAS, S-JNAS and CSJ, a positive effect was observed when our elderly speech data was also used to train the acoustic model. In the future, we will study methods of adapting this acoustic model, with the aim of improving the accuracy of speech recognition for elderly people.

Although the number of speakers was insufficient to be conclusive, a mild positive correlation between age and WER was observed (correlation coefficient: 0.59), i.e., the accuracy of speech recognition tended to decrease as the age of the speaker increased. We want to further investigate this correlation in the future, after increasing the number of speakers.

Finally, our corpus will be made public by the National Institute of Informatics in Japan. We anticipate that it will be used by other investigators for research on speech recognition for the elderly in Japan.

5. Acknowledgements

This study was partially supported by the Strategic Information and Communications R&D Promotion Program (SCOPE) of the Ministry of Internal Affairs and Communications of Japan, by the JSPS KAKENHI Grant-in-Aid for Scientific Research, Grant Numbers 17H01977, 19H01125 and 19K12022, and by ROIS NII Open Collaborative Research 2019, Grant Numbers 19S0403.

6. Bibliographical References

Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., Hudson, R., (1999). Recognition of elderly speech and voice-driven document retrieval. In Proc. of ICASSP.

Baba, A., Yoshizawa, S., Yamada, M., Lee, A., Shikano, K., (2001). Elderly Acoustic Model for Large Vocabulary Continuous Speech Recognition. In Proc. of EUROSPEECH 2001.

Fukuda, M., Nishimura, R., Nishizaki, H., Iribe, Y., Kitaoka, N., (2019). A new corpus of elderly Japanese speech for acoustic modeling, and a preliminary investigation of dialect-dependent speech recognition, In Proc. of COCOSDA 2019.

Furui, S., Isahara, H., Maekawa, K., (2000). A Japanese national project on spontaneous speech corpus and processing technology, ISCA Workshop on Automatic Speech Recognition.

Imai, Y., Hasegawa, K., (1994). The Revised Hasegawa's Dementia Scale (HDS-R): Evaluation of its usefulness as a screening test for dementia. Hong Kong Coll Psychiatr, vol. 4, pp. 20-24.

Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., (1999). JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, J.Acoust. Soc, Jpn, vol. 20, pp. 1999-2006.

Kitaoaka, N., Iribe, Y., Nishizaki, H., (2018). Construction of a corpus of elderly Japanese speech for analysis and recognition, LREC May 2018.

Kudo, I., Nakama, T., Watanabe, T., Kameyama, R., (1996), Data collection of Japanese dialects and its influence into speech recognition, Proc. ICSLP 1996.

Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., Shikano, K., (1990). ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis, Speech Communication, vol. 9, pp. 357-363.

Maekawa, K., Yamazaki, M., Ogiso, T. , Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., Den, Y., (2014). Balanced corpus of

contemporary written Japanese, Language Resources and Eval., Vol. 48, No. 2, pp. 345–371.

Ministry of Internal Affairs and Communications, (2019), Information and communications in Japan, http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r01/pdf/n3200000.pd

Ministry of Health, Labor and Welfare, (2017), https://www.mhlw.go.jp/stf/houdou/0000180168.html

Miyazaki, T., Mizumachi, M., Niyada, K., (2010). Acoustic analysis of breathy and rough voice characterizing elderly speech, J. of Advanced Computational Intelligence and Intelligent Informatics, pp. 135-141.

NII-SRI, (2019). Speech Resources Consortium, http://research.nii.ac.jp/src/en/ATR503.html

Pellegrini, T., Trancoso I., Hamalainen A., Calado A., Dias M. S., Braga D., (2012). Impact of age in ASR for the elderly preliminary Experiments in european portuguese, IberSPEECH 2012, CCIS 328, pp139-147.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., (2011). The Kaldi Speech Recognition Toolkit, Proc. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Toth L., Hoffmann I., Gosztolya G., Vincze V., Szatloczki G., Banreti Zo., Papaski M., Kalman J., (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech, Current Alzheimer research, 15, pp130-138.

Vipperla, R., Renals, S., Frankel, J., (2008). A longitudinal study of ASR performance on aging voices. In Proc. of Interspeech 2008.

Wilpon, J., Jacobsen, C., (1996). A study of speech recognition for children and the elderly, IEEE Press, pp. 349-352.

Winkler, R., Brückl, M., Sendlmeier, W., (2003). The aging voice: an acoustic, electroglottographic and perceptive analysis of male and female voices. In: Proc. of ICPhS 03, Barcelona, pp. 2869-2872.