Computer Speech and Language 77 (2023) 101424

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

# A new speech corpus of super-elderly Japanese for acoustic modeling

Meiko Fukuda [a,*], Ryota Nishimura [a], Hiromitsu Nishizaki [b], Koharu Horii [c], Yurie Iribe [d], Kazumasa Yamamoto [e], Norihide Kitaoka [c]

[a] *Tokushima University, Department of Computer Science, Tokushima, Japan*
[b] *University of Yamanashi, The Graduate School of Interdisciplinary Research, Kofu, Japan*
[c] *Toyohashi University of Technology, Department of Computer Science and Engineering, Toyohashi, Japan*
[d] *Aichi Prefectural University, School of Information Science and Technology, Nagoya, Japan*
[e] *Chubu University, Department of Computer Science, Japan*

ARTICLE INFO

*MSC:*
00-01
99-00

*Keywords:*
Speech corpus
Elderly
Speech recognition
Acoustic feature

ABSTRACT

The development of accessible speech recognition technology will allow the elderly to more easily access electronically stored information. However, the necessary level of recognition accuracy for elderly speech has not yet been achieved using conventional speech recognition systems, due to the unique features of the speech of elderly people. To address this problem, we have created a new speech corpus named EARS (Elderly Adults Read Speech), consisting of the recorded read speech of 123 super-elderly Japanese people (average age: 83.1), as a resource for training automated speech recognition models for the elderly. In this study, we investigated the acoustic features of super-elderly Japanese speech using our new speech corpus. In comparison to the speech of less elderly Japanese speakers, we observed a slower speech rate and extended vowel duration for both genders, a slight increase in fundamental frequency for males, and a slight decrease in fundamental frequency for females. To demonstrate the efficacy of our corpus, we also conducted speech recognition experiments using two different acoustic models (DNN-HMM and transformer-based), trained with a combination of data from our corpus and speech data from three conventional Japanese speech corpora. When using the DNN-HMM trained with EARS and speech data from existing corpora, the character error rate (CER) was reduced by 7.8% (to just over 9%), compared to a CER of 16.9% when using only the baseline training corpora. We also investigated the effect of training the models with various amounts of EARS data, using a simple data expansion method. The acoustic models were also trained for various numbers of epochs without any modifications. When using the Transformer-based end-to-end speech recognizer, the character error rate was reduced by 3.0% (to 11.4%) by using a doubled EARS corpus with the baseline data for training, compared to a CER of 13.4% when only data from the baseline training corpora were used.

## 1. Introduction

Digital technologies have improved the quality of our lives by making many tasks and activities more convenient, especially under the various socio-economic constraints imposed by the coronavirus pandemic, by allowing online shopping and interaction with our friends, family and colleagues, using video conferencing, for example. In recent years in Japan, the use of the internet by

[*] Corresponding author.
*E-mail address:* fukuda.meiko@tokushima-u.ac.jp (M. Fukuda).

https://doi.org/10.1016/j.csl.2022.101424
Received 13 April 2022; Accepted 17 June 2022
Available online 24 June 2022

the elderly has been increasing. A survey conducted by the Cabinet Office of the Japanese government (Cabinet Office, 2021) found that the percentage of elderly people over 70 who use a smartphone or tablet was 40.8%. The survey asked those who were not using these tools why they did not use them. The second most common reason was, "I do not know how to use them" (42.4%), while the third most common reason was, "I think I could ask my family members to use them for me if necessary". Regarding the results of this survey, the Ministry of Internal Affairs and Communications of Japan said that the lack of progress in utilization of digital information by the elderly could prevent them from enjoying the benefits of connecting to the internet, and that they could be left behind in a digital society, hence it is essential to support their utilization of these resources (Ministry of Internal Affairs and Communication, Japan, 2021).

Another hurdle encountered by the elderly is that the visual acuity and hand-motor function of those who are already familiar with PCs and smartphones decline as they age, which can make it challenging for them to continue to operate PCs and smartphones. Speech recognition technology can be used to address this problem, since this mode of operation is more convenient for elderly users with impaired vision and motor skills. This promising modality can allow the elderly to maintain their access to the internet through their electronic devices, enabling them to enjoy its benefits. Conventional speech recognition technologies have not yet demonstrated sufficient recognition accuracy for elderly speech, however. One significant factor may be the acoustic models used to recognize elderly speech, which are generally trained using the voices of younger adults (Wilpon and Jacobsen, 1996; Anderson et al., 1999; Vipperla et al., 2008; Pellegrini et al., 2012). Therefore, in order to create acoustic models suitable for recognizing elderly speech, a large-scale corpus of elderly speech 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 (Baba et al., 2001). The average age of the S-JNAS speakers is 67.6, but life expectancy in Japan has risen to 84.2, resulting in a wide age gap. In addition, it has been observed that the older the speaker, the lower the speech recognition accuracy (Anderson et al., 1999; Pellegrini et al., 2012). Therefore, we decided to build a super-elderly speech corpus for acoustic modeling, following the same methodology used for the S-JNAS corpus. We named our new corpus Elderly Adults Read Speech (EARS). So far, we have recorded very elderly speakers in four regions of Japan for our corpus, resulting in a database of 123 speakers containing 5,617 utterances. We continue to record additional participants, and plan to make this corpus available to the public through Japan's National Institute of Informatics. We hope that many other investigators will use our new corpus for research on speech recognition for the elderly.

Many studies have reported on the various characteristics of elderly voices which differ from those of younger people, such as changes in fundamental frequency (*Fo*) and formant frequencies, slower speech rate, increased number of pauses, extended pauses between words or inserted in a word, unclear pronunciation and hoarseness (Torre and Barlow, 2009; Miyazaki et al., 2010; Winkler et al., 2003; Pellegrini et al., 2013; Makiyama and Hirano, 2017). The acoustic features of the speech of the Japanese elderly are generally the same as those of elderly speakers of other languages (Miyazaki et al., 2010; Nishio et al., 2011). However, there have been few comprehensive studies on the acoustic features of the speech of the Japanese super elderly, so in this paper we compare the *Fo*, speaking rate and vowel duration of elderly and super elderly speakers using S-JNAS and our EARS corpus. Based on our analysis, super-elderly speech differs from the speech of elderly in some ways, thus acoustic modeling using the speech of adults, in general, is likely insufficient for the recognition of elderly speech.

In this paper, we demonstrate the effectiveness of using the EARS corpus when training acoustic models for elderly speech, by conducting speech recognition experiments using DNN-HMM and transformer-based acoustic models trained using a combination of JNAS corpus, CSJ and S-JNAS conventional Japanese speech corpora and EARS speech data.

We use these additional speech corpora because EARS currently contains an insufficient amount of speech data to create original acoustic models. Therefore, we constructed our acoustic models using a combination of speech data from three existing speech corpora in addition to the super-elderly speech data in our corpus (Fukuda et al., 2020). These corpora are the Japanese Newspaper Article Sentences Read Speech Corpus (JNAS) (Itou et al., 1999), S-JNAS (Baba et al., 2001), and the Corpus of Spontaneous Japanese (CSJ) (Furui et al., 2000). Additional information about the first two corpora is provided in the Related Work section of this paper, while the third corpus is discussed in Section 5.1. In order to demonstrate a simple method of using our corpus, the training data of our corpora was expanded through duplication, from one to ten times, and was also used to train the acoustic models for various numbers of epochs, without any modifications.

The rest of this paper is structured as follows. In Section 2, related work is introduced, especially regarding previously developed speech corpora. In Section 3, we provide details of the development and content of our new corpus. In Section 4, we explore the acoustic characteristics of Japanese elderly and super-elderly speech using the S-JNAS and EARS corpora. In Section 5, we report the results of our speech recognition experiments, which show the effectiveness of using EARS when modeling the speech of the super elderly. In Section 6, we discuss the findings of this study, and finally, in Section 7, we summarize the conclusions of this paper.

## 2. Related work

The design of our corpus is based on the JNAS and S-JNAS speech corpora, so in this section we will describe these two corpora. The JNAS corpus is a famous Japanese speech corpus widely used for constructing the acoustic models in automatic speech recognizers (ASRs) for standard, Japanese, adult speech (Itou et al., 1999). It was created by the Acoustical Society of Japan using the recorded speech of 306 participants (153 males and 153 females) who were mostly 20 to 49 years old. Each participant read aloud one set of newspaper article sentences (100 sentences) and one set of phoneme balanced sentences (about 50 sentences), for use as training data. These sentences are described in Section 3.3 of this paper. We estimate that the total recording time of the

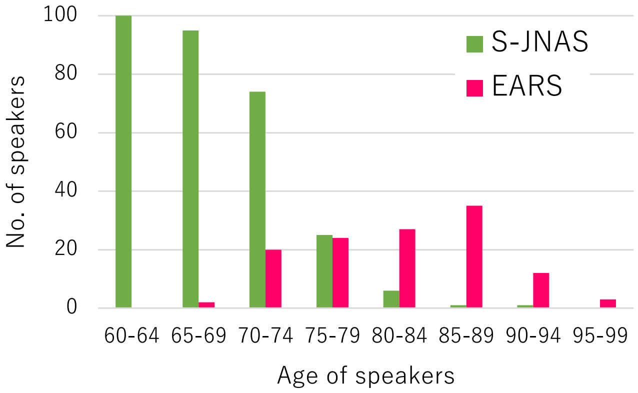

**Fig. 1.** Comparison of the age distributions of speakers in S-JNAS and EARS corpora.

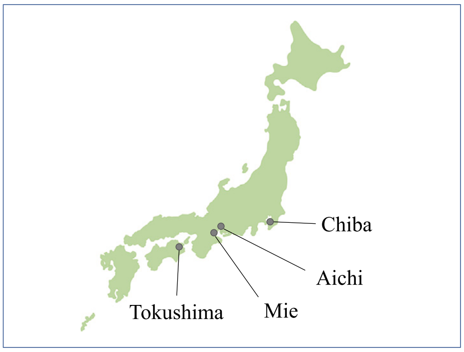

**Fig. 2.** Areas where recording sessions were conducted.

JNAS corpus is 72.4 h. These utterances were recorded using both a headset microphone and a desktop microphone. The WAV format was used for the speech files (16 kHz, 16-bit, Mono).

The S-JNAS corpus, a version of JNAS created using elderly speech, was compiled as a New Energy and Industrial Technology Development Organization (NEDO) project (Baba et al., 2001). The training set consists of the speech of 301 participants (151 males and 150 females) recorded as they read aloud one set of newspaper article sentences and two sets of phonetically balanced sentences per speaker. The test set also includes 301 participants (151 males and 150 females), who read aloud one set of newspaper article sentences and two sets of task sentences per speaker. The newspaper article sentences and phonetically balanced sentences are the same as those used for the JNAS corpus. We estimate that the total recording time is 133.4 h. The ages of the S-JNAS participants ranged 60-90, with an average age of 67.6. These utterances were also recorded using both a headset microphone and a desktop microphone, and the speech files are also in the WAV format (16 kHz, 16-bit, Mono).

## 3. Construction of a new super-elderly speech corpus

In this section, we describe the specifications of our super-elderly speech corpus, EARS. Our research project was approved by the ethics committees of Nagoya University, Tokushima University and Toyohashi University of Technology in Japan. All of the participants provided written, informed consent for the recording of their speech.

### *3.1. Selection criteria for participants and their age distribution*

There are only two criteria for selecting participants, and one of those is age. Since the goal in creating our new corpus is to construct a super-elderly version of S-JNAS, as many people over eighty years of age as possible were selected. As a result, the mean age of our participants, whose ages range from 70 to 99 years old, is currently 83.1, which is two decades older than those in S-JNAS corpus, who have a mean age of 63.4. Fig. 1 shows a comparison of the age distribution of speakers in the S-JNAS corpus and in the EARS corpus. Tables 1 and 2 show the age distributions of the subjects whose speech is included in the training and test data sets of our corpus, respectively. We recorded the speech of elderly people in four regions of Japan, as shown in Fig. 2, and there are no significant differences in the average ages of the participants from these four areas.

Our second selection criterion was the health of the participants, but this was not strictly applied since another goal when creating this corpus is to allow the creation of acoustic models for people with various health conditions, enabling people who are

**Table 1**
Number of speakers in each age group and their gender in EARS training data.

| Age | Male | Female | Subtotal |
|---|---|---|---|
| 70–74 | 2 | 8 | 10 |
| 75–79 | 8 | 14 | 22 |
| 80–84 | 8 | 19 | 27 |
| 85–89 | 5 | 27 | 32 |
| 90–94 | 5 | 9 | 14 |
| 95–99 | 1 | 3 | 4 |
| Total | 29 | 80 | 109 |

**Table 2**
Number of speakers in each age group and their gender in EARS test data.

| Age | Male | Female | Subtotal |
|---|---|---|---|
| 70–74 | 0 | 1 | 1 |
| 75–79 | 3 | 1 | 4 |
| 80–84 | 1 | 3 | 4 |
| 85–89 | 0 | 3 | 3 |
| 90–94 | 0 | 1 | 1 |
| 95–99 | 1 | 0 | 1 |
| Total | 5 | 9 | 14 |

**Table 3**
Number of subjects from each area whose speech was used for training data.

| Recording area | Male | Female | Subtotal |
|---|---|---|---|
| Tokushima | 11 | 22 | 33 |
| Mie | 1 | 5 | 6 |
| Aichi | 16 | 50 | 66 |
| Chiba | 1 | 3 | 4 |
| Total | 29 | 80 | 109 |

not in good health to also use speech recognition technology. However, it was essential that the recording sessions not be a burden to the participants mentally or physically, so only those in good enough condition to take part in the recording process were asked to participate. Hence, our speakers included people with various health conditions, such as a tendency to dementia, impaired vision, impaired hearing and use of dentures. Participants' living arrangements were not considered as a selection criterion, thus there are speakers in our corpus who live in nursing homes, who utilize care services only during the day and who did not use care facilities at all. Note that there are more women than men in our corpus, because the gender ratio of the participants was not considered when selecting participants either. Of course it would be better if the gender distribution of the participants was more balanced, but because of the demographics of the elderly population in Japan this has been difficult so far. Balancing the gender ratio of the speakers is one of our future goals.

### *3.2. Recording area and number of subjects*

Recording of super-elderly speech was conducted at elderly nursing facilities in four prefectures of Japan (Tokushima, Mie, Aichi and Chiba)[1] (Fukuda et al., 2020), as shown in Fig. 2. It is generally said that there are 16 dialects of Japanese. One study found that these regional variations in spoken Japanese have a 2% to 4% influence on speech recognition accuracy (Kudo et al., 1996). Hence, we plan to balance the distribution of speaker dialects within the EARS corpus by conducting recording sessions in additional regions in the future.

Currently, a total of 109 Japanese native speakers (29 men and 80 women) have been selected as training data speakers for our corpus. The number of participants in each of the four regions vary however, as shown in Table 3, owing to the difficulty of recruiting subjects. The total recording time of our EARS training data is 11.52 h. The participants in our experiment, i.e., the speakers who speech was included in the test set, were totally different from the speakers whose speech data was use in the training set. The total number of study participants was 14, who were selected from among the EARS corpus participants in three of the four prefectures (excluding Aichi) where data was collected, as shown in Table 4. The total recording time of test data is 0.33 h. In the future, additional recording sessions for test data will be held in Aichi prefecture.

[1] The speech data recorded in Nagasaki and Yamagata prefectures, which was reported in our previous paper, was excluded from the data used in this study, due to a joint research contract with a private company.

**Table 4**
Number of subjects from each area whose speech was used for test data.

| Recording area | Male | Female | Subtotal |
|---|---|---|---|
| Tokushima | 2 | 3 | 5 |
| Mie | 2 | 3 | 5 |
| Aichi | 0 | 0 | 0 |
| Chiba | 1 | 3 | 4 |
| Total | 5 | 19 | 14 |

**Table 5**
Number of utterances in each set of EARS training data sentences.

| Set name<br>(# sentences) | Male | Female<br>(# utterances) | Subtotal |
|---|---|---|---|
| Set A (50) | 1 (50) | 10 (500) | 11 (550) |
| Set B (50) | 3 (150) | 7 (350) | 10 (500) |
| Set C (50) | 5 (250) | 5 (250) | 10 (500) |
| Set D (50) | 1 (50) | 9 (450) | 10 (500) |
| Set E (50) | 6 (300) | 4 (200) | 10 (500) |
| Set F (50) | 2 (100) | 12 (600) | 14 (700) |
| Set G (50) | 3 (150) | 10 (500) | 13 (650) |
| Set H (50) | 3 (150) | 7 (350) | 10 (500) |
| Set I (50) | 3 (150) | 9 (450) | 12 (600) |
| Set J (53) | 2 (106) | 7 (371) | 9 (477) |
| Total # utterances<br>[# speakers ] | 1456<br>[29] | 4021<br>[80] | 5477<br>[109] |

**Table 6**
Number of utterances in each set of EARS test data sentences.

| Set name<br>(# sentences) | Male | Female<br>(# utterances) | Subtotal |
|---|---|---|---|
| T1 (10) | 2 (20) | 1 (10) | 3 (30) |
| T2 (10) | 0 (0) | 3 (30) | 3 (30) |
| T3 (10) | 0 (0) | 2 (20) | 2 (20) |
| T4 (10) | 2 (20) | 2 (20) | 4 (40) |
| T5 (10) | 1 (10) | 1 (10) | 2 (20) |
| Total # utterances<br>[# speakers ] | 50<br>[5] | 90<br>[9] | 140<br>[14] |

### 3.3. Selection of Japanese sentences

Since the design of the EARS corpus was based on that of the S-JNAS corpus, we also used the ATR 503 sentences (Kurematsu et al., 1990) and JNAS newspaper article sentences as scripts for our speakers. However, unlike the S-JNAS corpus, the ATR 503 sentences were used only when recording training data, while the newspaper article sentences were only used when recording test data. Our reason for doing this will be explained below.

The ATR 503 sentences corpus contains 503 sentences. These include 402 two-phoneme sequences and 223 three-phoneme sequences (625 items in total). The sentences are extracted from newspapers, textbooks, journals, letters, novels, and the like, and reproduce the phonetic balance that appears in Japanese as much as possible. The sentences are split into ten text sets (Set A-J) (Kurematsu et al., 1990).

The JNAS newspaper article sentences were taken from Japan's Mainichi daily newspaper. These sentences contains 155 test sets consisting of 1,000 total sentences, but only 50 of these sentences were used when recording the EARS test set. They were divided into five sets of 10 sentences each (Sets T1-T5). When the training set of the S-JNAS corpus was created, each speaker read two sets of ATR 503 sentences (about 100 sentences) and one set of newspaper article sentences (about 100 sentences). However, since our speakers are super-elderly, and many of them have difficulty reading a large number of sentences, such as dementia tendency or poor eyesight, the number of sentences read aloud by each participant was lessened to reduce the burden. Each participant read only one set of the ATR 503 sentences (containing either 50 or 53 sentences) and 50 JNAS newspaper article sentences. Since we wanted the corpus to include as many Japanese phonemes as possible, and since the appearance probabilities of the phonemes in Japanese were vital for constructing the super-elderly acoustic models, we decided to use the ATR 503 sentences as the EARS training data. And since it is better to use natural Japanese as test data, we used the JNAS news article sentences as the EARS test data.

The current total number of utterances of all of the speakers in the EARS training data is 5,477, as shown in Table 5, while the total number of newspaper article utterances contained in the EARS test data is 140, as shown in Table 6.

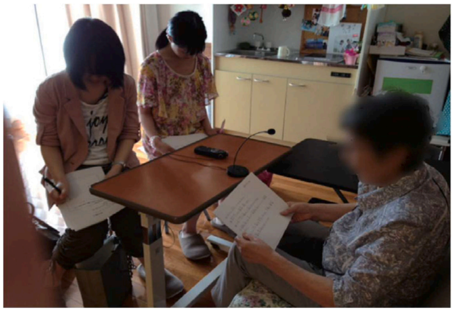

**Fig. 3.** Typical recording scene in nursing home.

### 3.4. Data collection procedure

Most of the read speech recording sessions were held at elderly care facilities, but some were recorded at a university. Fig. 3 shows a typical recording scene at an elderly care facility. Participants were informed in advance about the recording procedure and printouts of the ATR 503 sentences and JNAS newspaper article sentences were distributed to the speakers, except in Aichi, where only the ATR 503 sentences were used. All of the sentences were printed in both standard Japanese kanji and hiragana characters. Participants were allowed to practice reading the sentences aloud before the start of the recording session. If they made a mistake when reading a sentence aloud, the recording staff from our research group encouraged them to read it aloud again. In consideration of the physical condition of the participants, if the speakers wanted to rest or end the recording session, they could stop at any time. The likelihood that participants were suffering from dementia was evaluated by the recording staff using the HDS-R (Hasegawa's Dementia Scale-Revised) (Imai and Hasegawa, 1994). Finally, each participant's mood was evaluated by staff members of the elderly care facilities, using a survey. The results of the dementia evaluations and mood surveys will not be made public, however.

### 3.5. Devices used for recording

The devices used to record the read speech of the participants were a lapel microphone (Sony ECM-88B), a desktop microphone (Audio-Technica AT9930) and an 8-track field recorder (Tascam DR-680MKII). However, in Aichi, a linear PCM recorder (Tascam DR-05 ver.2) and the same Audio-Technica desktop microphone were used.

### 3.6. Database details

The duration of the recorded speech in our corpus is approximately 13.4 h in total. The speech waves were digitized at a sampling frequency of 16 kHz using 16-bit linear PCM. The speech recordings were divided into sentence units manually, and silent pauses of about 150msec were added at the beginning and end of almost every sentence.

The recorded speech was manually transcribed into text data by trained employees since the speakers often shuttered, uttered fillers or misread the sentences. The phonemes and words of the sentences were rewritten to correspond to the speakers' actual phonation. Demographic and assessment information for the speaker (age, gender, recording location, dementia score and mood assessment), as well as the Set ID (A-J or T1-T5) of the sentences being read, were also included in the database.

## 4. Comparison of acoustic features of elderly and super-elderly speech

To understand the differences between the acoustic features of elderly and super-elderly speech, we examined fundamental frequency (*Fo*), speaking rate and vowel duration.

### 4.1. Speech material

The recorded voices of elderly and super-elderly speakers obtained from the S-JNAS and EARS corpora, respectively, reading aloud the ATR 503 sentences, were the data used for our comparison. The ages of the speakers range from 60 to 90 in the S-JNAS corpus, and from 70 to 98 in the EARS corpus. Japanese has five vowels (/a/, /e/, /i/, /o/, /u/), all of which were examined in this comparison. To avoid the effect of co-articulation as much as possible, measurement targets were limited to consonant-vowel combinations. All such combinations which appeared in the vocabulary of the ATR 503 sentences were examined. The number of measurement targets for each vowel ranged from 169 to 425, depending on which set of the ATR 503 sentences was being evaluated and which vowels happened to appear in that set.

There are a different number of speakers in these two corpora. In order to accurately approximate the regression lines of the scatter plots, the number of subjects in both corpora was randomly reduced for each age, in order to balance the number of speakers in each age group.

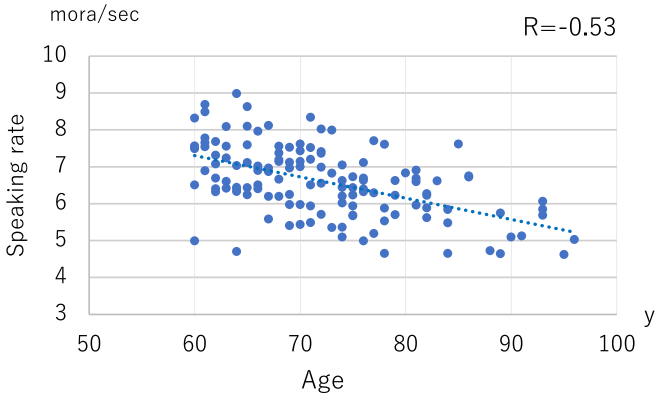

**Fig. 4.** Scatter plot and regression line for speaking rates (mora/sec) of males. Each dot corresponds to one speaker.

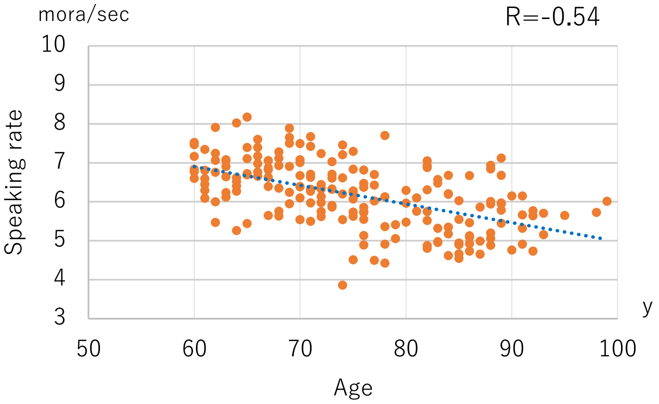

**Fig. 5.** Scatter plot and regression line for speaking rates (mora/sec) of females of various ages. Each dot represents one speaker.

### 4.2. Measurement of acoustic features

To compute the acoustic values, forced-alignment made of Julius Japanese speech recognizer (Lee and Kawahara, 2009) was used to perform forced alignments between the audio signal and the word and phone sequences automatically. Vowel durations and speaking rates (mora/second) were calculated from the alignment results with reference to the beginning and the ending point of each vowels. Median *Fo* values of the vowels were estimated with the cross-correlation algorithm.

### 4.3. Result of acoustic measurements

The result of acoustic measurements is depicted using scatter plots and regression lines. Figs. 4 and 5 shows prolongation of speaking rates (mora/second) with aging for both genders. Related to these prolongations, the mean duration of vowels with age for both genders are shown in Figs. 6 and 7. We can see that *Fo* slight increases of with age for the male speakers, but slight decreases with age for the female speakers in Figs. 8 and 9.

Previous studies have revealed that the speech characteristics remain fairly consistent from 20s to 50s, and that speech recognition performance is also similar for speakers belonging to these generations. However, as speakers reach their 60s, speech recognition performance begins decreasing with age. The tendencies observed here, in the changes in speakers' voices which occur with age, are likely due to degradation of the body, although there may be various other changes in our voices which also occur besides those examined here.

## 5. Speech recognition experiments

This study aims to demonstrate the effectiveness of EARS as a training corpus for the recognition of elderly speech using a simple training method. We created an expanded version of EARS by duplicating our corpus of super-elder speech many times without any modifications. We then combined this data with speech data from three other corpora of Japanese speech. This was done because the amount of speech data in EARS corpus alone was insufficient for creating original acoustic models. We controlled the ratio of EARS super-elderly speech in the training data by duplicating EARS and by training the model repeatedly.

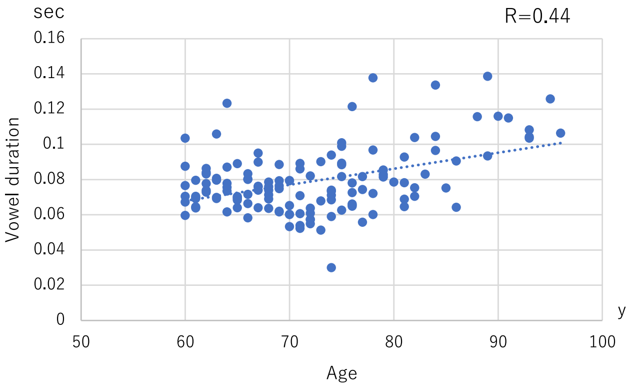

**Fig. 6.** Scatter plot and regression line for vowel duration of males of various ages. Each dot represents one speaker.

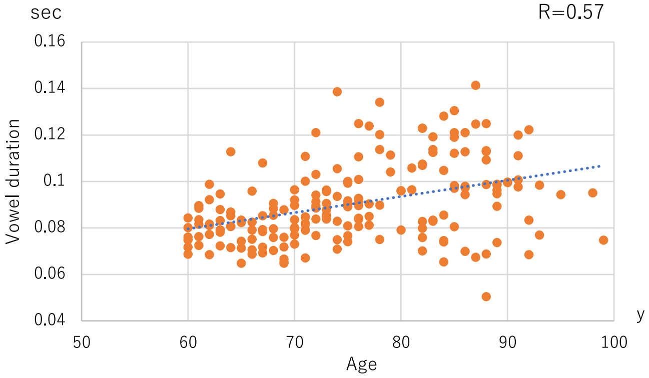

**Fig. 7.** Scatter plot and regression lines for vowel duration of females of various ages. Each dot represents one speaker.

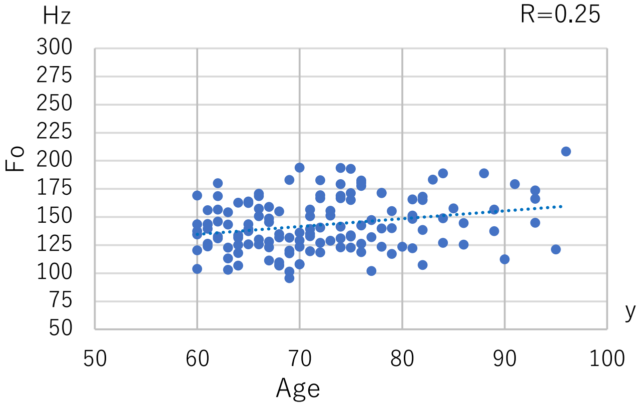

**Fig. 8.** Scatter plot and regression line for *Fo* of males. Each dot corresponds to one speaker.

### 5.1. Experimental setup for DNN-HMM-based speech recognizer

We used the Kaldi DNN-HMM speech recognition toolkit in our experiment. The training and test procedures followed the CSJ recipe, thus the "nnet3" neural network in the toolkit[2] was used, therefore the neural network in the model was a time-delay neural network (TDNN) (Peddinti et al., 2015). The recognizer used word-based units defined by a phoneme-based pronunciation

[2] https://github.com/kaldi-asr/kaldi/blob/master/egs/csj/s5/local/nnet3/run_tdnn.sh

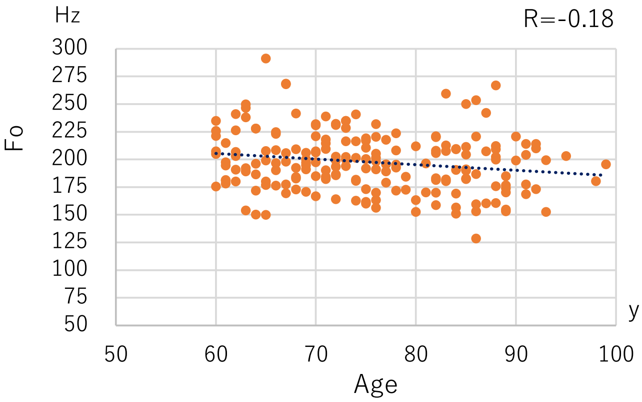

**Fig. 9.** Scatter plot and regression lines for *Fo* of females. Each symbol corresponds to one speaker.

dictionary, and the word models were created using concatenation of triphone models. Each triphone model consisted of three states, and the topology was left-to-right. The speech data was expanded using speed perturbation with factors of 1.1 and 0.9. The speech features derived from the expanded speech data consisted of 40-dimensional MFCCs and a 100-dimensional i-vector for each frame.

In our previous study, we used three Japanese speech corpora (JNAS, S-JNAS and CSJ) to create the training data for our baseline acoustic models and obtained good results (Fukuda et al., 2020). Based on the results of that study, the same three corpora were also used for model training in this study. The number of duplication of EARS training data when training the acoustic models were once, twice, three, five or ten times. The models were trained for 1, 2 or 4 epochs. The character error rate (CER) and word error rate (WER) and sentence error rate (SER) were calculated at each iteration.

Both the JNAS and S-JNAS corpora contain speech data from speakers reading the ATR 503 sentences aloud, as does EARS. While the speech data in the former two corpora is very fluent, many of the utterances in EARS corpus are disfluent, since they include non-standard, personal speaking habits, such as a large number of fillers, hesitation, rephrasing, shuttering and many pauses, because of the speakers' more advanced ages and the inclusion of speakers with dementia symptoms. Thus, the speaking styles of the EARS speakers are more similar to spontaneous speech than to the read speech in the JNAS and S-JNAS corpora. Therefore, the CSJ corpus (Furui et al., 2000) was used to bridge the gap between EARS and the two read speech corpora by adding it to the training set used to train the acoustic models. Although the CSJ does not include any super-elderly speakers, the speaking styles of the speakers often include many fillers, shuttering and rephrasing. The CSJ is the most extensive corpus of spontaneous Japanese speech in the world, with 1,417 speakers and a total training data recording time of about 520.8 h. It mainly consists of speeches from academic conferences and simulated public speaking.

For our DNN-HMM framework, we also need a language model which has been trained separately with a large amount of text data. In this study, our language model was constructed using the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014), using the same method as in our previous study (Fukuda et al., 2020). The BCCWJ corpus is the most extensive corpus of contemporary written Japanese sentences, and was collected from various publications such as books, magazines, newspapers and websites, thus the sentences include a wide variety of vocabulary and syntax. The EARS test set consists of newspaper sentence utterances, and thus the written Japanese BCCWJ data is suitable for training the language model. We could have used another newspaper text corpus to train our language model, in order to obtain better recognition performance, but we mainly wanted to evaluate the acoustic models in this study, thus we used a relatively ”loose” linguistic constraint. Using the CSJ and BCCWJ corpora, we trained a trigram model with a vocabulary size of 164k for the search in the decoding stage.

Our test data consisted of the utterances of our 14 test speakers, as previously mentioned in Sections 2.2 and 2.3 of this paper (see Tables 2, 4 and 6). All of the test speakers are not the same speakers whose speech was used for the training data.

### 5.2. Experimental setup for transformer-based speech recognizer

In our experiment, we also examined the effect on elderly speech recognition performance of using an end-to-end speech recognizer. Various end-to-end speech recognizers are available within the ESPnet2 framework (Watanabe et al., 2021) The setup of the ESPnet2 followed the original CSJ recipe of ESPnet2. Note that end-to-end models learn acoustic and linguistic information in an integrative manner.

There are some *ad hoc* methods which use an additional language model, such as shallow fusion, but in this study we wanted to investigate the relationship between speech recognition performance and super-elderly acoustic information, thus we did not use any additional language models. The S-JNAS speech data used for model development was manually divided into training data (497,680 utterances) and validation data (7,312 utterances). Model training was repeated for 20 epochs. For our evaluation, we saved the model parameters of the ten best performing epochs, according to the validation sets, and averaged them at the end of training. The hyper-parameters used for training and recognition were the same as the default values of the CSJ recipe, except that we used

**Table 7**
Results of speech recognition experiments when using the DNN-HMM acoustic model.

| Training data | Epoch | WER (%) | CER (%) | SER (%) |
|---|---|---|---|---|
| BL[a] | 1 | 17.56 | 15.80 | 67.86 |
| BL | 2 | 19.00 | 16.91 | 70.71 |
| BL | 4 | 18.78 | 16.94 | 70.00 |
| BL + EARS×1 | 1 | 11.42 | 9.33 | 56.43 |
| BL + EARS×1 | 2 | **10.73** | **9.08** | **55.71** |
| BL + EARS×1 | 4 | 11.51 | 9.79 | 56.43 |
| BL + EARS×2 | 1 | 11.68 | 9.71 | 56.43 |
| BL + EARS×2 | 2 | 11.86 | 10.47 | 59.29 |
| BL + EARS×2 | 4 | 12.12 | 10.17 | 59.29 |
| BL + EARS×3 | 1 | 11.60 | 9.74 | 57.41 |
| BL + EARS×3 | 2 | 12.07 | 10.31 | 57.86 |
| BL + EARS×3 | 4 | 12.03 | 10.52 | 56.43 |

[a] Baseline model, trained using only JNAS, S-JNAS and CSJ speech data.

**Table 8**
Results of speech recognition experiments when using the transformer acoustic model.

| Training data | CER (%) | SER (%) |
|---|---|---|
| BL[a] | 13.4 | 80.0 |
| BL + EARS×1 | 12.2 | **74.3** |
| BL + EARS×2 | **11.4** | 76.4 |
| BL + EARS×3 | 13.2 | 75.7 |

[a] Baseline model, trained using only JNAS, S-JNAS and CSJ data.

Transformer instead of Conformer, because our preliminary experiment showed that Transformer performed slightly better than Conformer. The speech data was augmented using speed perturbation with factors of 0.9 and 1.1. SpecAug (Park et al., 2019) was also used. The features consisted of 40-dimensional Mel filter bank outputs. Recognition was performed using character units. The number of characters was 3,316. We chose a hybrid method called Joint CTC-attention, which primarily uses a transformer-based encoder–decoder, but a CTC decoder is also used as a supplementary process. Validation data consisted of the same utterances used for our DNN-HMM experiments. Next, the EARS speech data was duplicated from one to ten times and added to the baseline training data (JNAS, S-JNAS, CSJ) without any modifications.

The character error rate (CER) and sentence error rate (SER) were calculated at each iteration. The speech data from all four corpora was manually divided into training data (502,757 utterances) and validation data (7,624 utterances).

### 5.3. Experimental results

#### 5.3.1. DNN-HMM experiments

As shown in Table 7, the CER of the baseline model (epoch 2) was 16.9%. During our speech recognition experiments, the lowest CER achieved was 9.08%, when the EARS speech data was used once during training in epoch 2. Simply training the DNN-HMM acoustic model repeatedly with the EARS data was not shown to be effective. This might be the results of a trade-off between the ratio of super-elderly and adult speech used for training, and variation within the training corpus.

#### 5.3.2. Transformer experiment

When using the transformer acoustic model trained with the baseline training set, the recognition CER was 13.4% and the SER was 80.0%, as shown in Table 8. The CER fell significantly, by 2.0%, when the EARS data was doubled using data augmentation and included in the training set. The SER decreased significantly, by 5.7%, when the EARS data was used once. The 2% lower CER when using the EARS data indicates that including super-elderly speech data when training the acoustic model, without modification, improved speech recognition accuracy compared to the baseline training method. However, when the volume of EARS speech was expanded three to ten times, no additional improvement in speech recognition accuracy was observed (data not shown). Thus, there is a limitation to the effectiveness of simply using additional copies of the same EARS data during training, as was also observed in the DNN-HMM experiments.

## 6. Discussion

Our new EARS corpus of super-elderly speech, featuring 123 speakers, is the first of its kind in Japan. Our preliminary comparison of elderly speech from the S-JNAS corpus and super-elderly Japanese speech from our EARS corpus revealed a decrease in speaking rate and increase in the duration of vowels in super-elderly Japanese speech. Both of these phenomena, as well as changes observed in the fundamental frequency of male and female speech, showed a tendency to increase with the age of the speaker, and thus are likely associated with aging. Whether these changes cause the deterioration of speech recognition accuracy is an issue for future research. Since our EARS corpus is not very large, we are not yet able to create original acoustic models for the elderly using only our corpus, but we plan to collect further recordings of super-elderly speakers in the future to expand our corpus. In this study, we have leveraged a limited amount of data to demonstrate the promise of this approach.

### 6.1. Acoustic comparison of elderly and super-elderly speech

We first investigated speaking rates, vowel duration and *Fo* in the speech of the elderly and super-elderly. As in previous studies, we confirmed that prolongation of speaking rates and increased vowel duration were associated with aging. These changes may be due to the physical consequences of aging, e.g., neuromuscular degeneration, in conjunction with slower processing times, as well as reduced auditory feedback (Harnsberger et al., 2008; Smith et al., 1987; Linville, 1996). By slowing their speaking rates, the elderly may also be trying to make their vague articulation as clear as possible (Fletcher et al., 2015).

Scatter plots and regression results showing the relationship between fundamental frequency of elderly speech in relation to the age of the speaker revealed that *Fo* tends to increase with age for males but to decrease with age for females. Previous studies have also reported changes in *Fo* in old age, with some researchers observing increases in *Fo* (not always to a significant degree) (Nishio and Niimi, 2008; Torre and Barlow, 2009; Sebastian et al., 2012), some observing no change (Eichhorn et al., 2018), and some observing decreases (Albuquerque et al., 2019; Tykalova et al., 2020). The increase in *Fo* in males after middle age has been attributed to a drop in vocal fold mass and stiffness in the vocal fold tissues, as well as thyroarytenoid muscle atrophy (Nishio and Niimi, 2008; Albuquerque et al., 2019). In females, many studies have reported that *Fo* decreases with age (Nishio and Niimi, 2008; Decoster, 1999; Albuquerque et al., 2019; Tykalova et al., 2020; Eichhorn et al., 2018). It has been suggested that the mechanism behind this decrease is the change in the balance of sexual hormones related to menopause, which usually occurs between 45 and 55 years of age (Honjo and Isshiki, 1980; Linville, 1996). However, it has also been observed that a woman's hormonal balance changes significantly from pre-menopause through menopause, but that no marked decrease in *Fo* is generally observed during this period (Lã and Ardura, 2020). Other researchers have found that a decrease in the *Fo* of some female speakers can already be observed in their 20s, and that this process continues even up to the age of 60 (Nishio and Niimi, 2008; Guimarães and Abberton, 2005), which suggests that the decrease in *Fo* in female speech begins long before menopause. Lã and Ardura suggest that the hormonal changes associated with menopause are not the only factor causing the decrease in female *Fo*, and that aging is also an important factor in changes in the acoustic features of speech (Lã and Ardura, 2020).

### 6.2. Speech recognition experiments

Although the size of the EARS corpus is not large enough to be used independently for acoustic model training, we were able to improve speech recognition accuracy for the elderly speech by using it in combination with three existing Japanese speech corpora (JNAS, S-JNAS, and CSJ), to successfully compensate for the acoustic changes in speech caused by aging (Fukuda et al., 2020).

In this study, we duplicated the EARS data repeatedly, without any modifications, and the recognition results exceeded the level of accuracy achieved in our previous study. When using Transformer, the best character error rate (CER) achieved was 11.4%, when the EARS data was duplicated, reducing the CER 2.0% (reducing it relatively by 15.0%) from the baseline-only training method. However, when the DNN-HMM model was trained using three or more times the amount of elderly speech, there was no further improvement in recognition accuracy over using EARS only once.

In light of these results, EARS has been shown to be a useful resource when creating acoustic models for elderly speech. Still, the effectiveness of simply expanding the EARS data is limited. When using the DNN-HMM, where the degree of freedom of the model is small, the properties of the model were not affected as much by the size of the EARS component of the training data. On the other hand, the amount of EARS training data had some influence in the case of the End-to-End approach with a high degree of freedom. However, increasing the ratio of only one part of the data reduces generalization performance, so a moderate increase in the amount of additional data is desirable.

In the future, we plan to enlarge the EARS corpus. We will also further investigate changes in speech features across age groups, and develop more efficient and practical techniques for using our corpus.

## 7. Conclusion

In this study, we have introduced EARS, the first corpus of super-elderly Japanese speech, with the aim of improving the accuracy of automatic speech recognition for elderly Japanese users of this senior-friendly technology. By comparing speech in the S-JNAS and EARS corpora, we were able to identify some of the characteristics of elderly versus super-elderly Japanese speech, which are a reduction in speaking rate, an increase in the mean duration of vowels, an increase in *Fo* in males and a decrease in *Fo* in females. Speech recognition performance is said to decrease with age, beginning in a person's 60s. The changes in speech related to aging observed in this study, through our comparison of elderly and super-elderly speech, although there may be other types of changes in the acoustic features of speech which also occur with aging, in addition to those observed in this investigation.

We conducted speech recognition experiments using elderly speech to demonstrate the effectiveness of the EARS corpus for the acoustic modeling of elderly speech. Our experimental results using Transformer showed that speech recognition accuracy was improved when multiple copies of the EARS speech were used for training, without any modifications, in conjunction with the baseline corpora. This method of training acoustic models is simple, but it increased speech recognition accuracy significantly when using a Transformer-based speech recognizer. In future work, we want to investigate changes in other speech features, such as formant frequencies, vowel articulation (using detailed indexes of formant frequencies such as Vowel Space Area), and the number and duration of pauses inserted by elderly speakers. We also plan to enlarge our speech corpus and develop more efficient and effective techniques for using our corpus.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Funding

This study was partially supported by the JSPS KAKENHI Grant-in-Aid for Scientific Research, Grant Numbers, 19H01125 and 19K12022, Japan and by ROIS NII Open Collaborative Research 2019, Grant Numbers 19S0403, Japan.

## References

Albuquerque, L., Oliveira, C., Teixeira, A.J., Sa-Couto, P., Figueiredo, D., 2019. Age-related changes in European Portuguese vowel acoustics. In: INTERSPEECH. pp. 3965–3969.

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: 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258), Vol. 1. IEEE, pp. 145–148.

Baba, A., Yoshizawa, S., Yamada, M., Lee, A., Shikano, K., 2001. Elderly acoustic model for large vocabulary continuous speech recognition.

Cabinet Office, J., 2021. Annual report on the Ageing Society 2021. https://www8.cao.go.jp/kourei/whitepaper/w-2021/html/zenbun/s1_3_1_4.html. Last (Accessed 27 October 2021).

Decoster, W., 1999. Acoustic differences between sustained vowels perceived as young or old. Logopedics Phoniatrics Vocology 24 (1), 1–5.

Eichhorn, J.T., Kent, R.D., Austin, D., Vorperian, H.K., 2018. Effects of aging on vocal fundamental frequency and vowel formants in men and women. J. Voice 32 (5), 644–e1.

Fletcher, A.R., McAuliffe, M.J., Lansford, K.L., Liss, J.M., 2015. The relationship between speech segment duration and vowel centralization in a group of older speakers. J. Acoust. Soc. Am. 138 (4), 2132–2139.

Fukuda, M., Nishizaki, H., Iribe, Y., Nishimura, R., Kitaoka, N., 2020. Improving speech recognition for the elderly: A new corpus of elderly Japanese speech and investigation of acoustic modeling for speech recognition. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 6578–6585.

Furui, S., Maekawa, K., Isahara, H., 2000. A Japanese national project on spontaneous speech corpus and processing technology. In: ASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop. ITRW.

Guimarães, I., Abberton, E., 2005. Fundamental frequency in speakers of portuguese for different voice samples. J. Voice 19 (4), 592–606.

Harnsberger, J.D., Shrivastav, R., Brown, Jr., W., Rothman, H., Hollien, H., 2008. Speaking rate and fundamental frequency as speech cues to perceived age. J. Voice 22 (1), 58–69.

Honjo, I., Isshiki, N., 1980. Laryngoscopic and voice characteristics of aged persons. Arch. Otolaryngol. 106 (3), 149–150.

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 J. Psychiatry 4 (2), 20.

Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., Itahashi, S., 1999. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. J. Acoust. Soc. Japan (E) 20 (3), 199–206.

Kudo, I., Nakama, T., Watanabe, T., Kameyama, R., 1996. Data collection of Japanese dialects and its influence into speech recognition. In: Proceeding of Fourth International Conference on Spoken Language Processing, Vol. 4. ICSLP'96, IEEE, pp. 2021–2024.

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 Commun. 9 (4), 357–363.

Lã, F.M., Ardura, D., 2020. What voice-related metrics change with menopause? A systematic review and meta-analysis study. J. Voice.

Lee, A., Kawahara, T., 2009. Recent development of open-source speech recognition engine julius. In: Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference. Asia-Pacific Signal and Information Processing Association, 2009 Annual, pp. 131–137.

Linville, S.E., 1996. The sound of senescence. J. Voice 10 (2), 190–200.

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. Lang. Resour. Eval. 48 (2), 345–371.

Makiyama, K., Hirano, S., 2017. Aging Voice. Springer.

Ministry of Internal Affairs and Communication, Japan, 2021. Information and communication in Japan, white paper 2021. https://www.soumu.go.jp/johotsusintokei/whitepaper/index.html. Last (Accessed 24 October 2021).

Miyazaki, T., Mizumachi, M., Niyada, K., 2010. Acoustic analysis of breathy and rough voice characterizing elderly speech.. J. Adv. Comput. Intell. Intell. Informatics 14 (2), 135–141.

Nishio, M., Niimi, S., 2008. Changes in speaking fundamental frequency characteristics with aging. Folia Phoniatrica Logopaedica 60 (3), 120–127.

Nishio, M., Tanaka, Y., Niimi, S., 2011. Analysis of age-related changes in the acoustic characteristics of voices. J. Commun. Res. 2 (1).

Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.

Peddinti, V., Povey, D., Khudanpur, S., 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proceedings of INTERSPEECH 2015. pp. 3214–3218. http://dx.doi.org/10.21437/Interspeech.2015-647.

Pellegrini, T., Hämäläinen, A., De Mareüil, P.B., Tjalve, M., Trancoso, I., Candeias, S., Dias, M.S., Braga, D., 2013. A corpus-based study of elderly and young speakers of European portuguese: Acoustic correlates and their impact on speech recognition performance. In: INTERSPEECH. pp. 852–856.

Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M.S., Braga, D., 2012. Impact of age in ASR for the elderly: Preliminary experiments in European portuguese. In: Advances in Speech and Language Technologies for Iberian Languages. Springer, pp. 139–147.

Sebastian, S., Babu, S., Oommen, N.E., Ballraj, A., et al., 2012. Acoustic measurements of geriatric voice. J. Laryngol. Voice 2 (2), 81.

Smith, B.L., Wasowicz, J., Preston, J., 1987. Temporal characteristics of the speech of normal elderly adults. J. Speech Lang. Hear. Res. 30 (4), 522–529.

Torre, P., Barlow, J.A., 2009. Age-related changes in acoustic characteristics of adult speech. J. Commun. Disord. 42 (5), 324–333.

Tykalova, T., Skrabal, D., Boril, T., Cmejla, R., Volin, J., Rusz, J., 2020. Effect of ageing on acoustic characteristics of voice pitch and formants in Czech vowels. J. Voice.

Vipperla, R., Renals, S., Frankel, J., 2008. Longitudinal study of ASR performance on ageing voices.

Watanabe, S., Boyer, F., Chang, X., Guo, P., Hayashi, T., Higuchi, Y., Hori, T., Huang, W.-C., Inaguma, H., Kamo, N., et al., 2021. The 2020 espnet update: New features, broadened applications, performance improvements, and future plans. In: 2021 IEEE Data Science and Learning Workshop. DSLW, IEEE, pp. 1–6.

Wilpon, J.G., Jacobsen, C.N., 1996. A study of speech recognition for children and the elderly. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Vol. 1. IEEE, 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, Vol. 3. pp. 2869–2872.