Speech Corpus Spoken by Young-old, Old-old and Oldest-old Japanese

Yurie Iribe[1], Norihide Kitaoka[2] and Shuhei Segawa[3]
[1]School of Information Science and Technology, Aichi Prefectural University
1522-3 Ibaragabasama, Nagakute, Japan
[2]Department of Information Science and Intelligent Systems, Tokushima University
2-1, Minamizyousanzima, Tokushima, Japan
[3]Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Japan
iribe@ist.aichi-pu.ac.jp, kitaoka@is.tokushima-u.ac.jp, segawa.shuhei@g.sp.m.is.nagoya-u.ac.jp

Abstract

We have constructed a new speech data corpus, using the utterances of 100 elderly Japanese people, to improve speech recognition accuracy of the speech of older people. Humanoid robots are being developed for use in elder care nursing homes. Interaction with such robots is expected to help maintain the cognitive abilities of nursing home residents, as well as providing them with companionship. In order for these robots to interact with elderly people through spoken dialogue, a high performance speech recognition system for speech of elderly people is needed. To develop such a system, we recorded speech uttered by 100 elderly Japanese, most of them are living in nursing homes, with an average age of 77.2. Previously, a seniors' speech corpus named S-JNAS was developed, but the average age of the participants was 67.6 years, but the target age for nursing home care is around 75 years old, much higher than that of the S-JNAS samples. In this paper we compare our new corpus with an existing Japanese read speech corpus, JNAS, which consists of adult speech, and with the above mentioned S-JNAS, the senior version of JNAS.

Keywords: elderly speech corpus, nursing home care, speech corpus construction, speech recognition

1. Introduction

Previous research suggests that elderly people have more difficulty using information and communication technology (ICT) than younger adults (Júdice, 2010). The main reasons for this are the complexity of existing user interfaces and the limited set of available interaction modalities, since this technology is mainly being designed with younger users in mind. Hence, from the point of view of ICT, adapting the technology to better suit the needs of the elderly, for instance by increasing the choice of available interaction modalities, will help ensure that the elderly have access to these technologies. Some previous research suggests that speech is the easiest and most natural modality for human-computer interaction (HCI) (Acartürk, 2015). Speech is also the preferred modality when interacting with mobile devices when users have permanent impairments such as arthritis, or when temporary limitations such as driving or carrying objects make it difficult to use other modalities like touch.

It is hoped that ICT can be used to help maintain the health of the elderly. Daily verbal interaction helps them maintain their cognitive ability, reducing the risk of dementia. In a super-aging society such as Japan, where we face an acute shortage of care workers, spoken dialogue systems could play an important role. However, speech recognizers, which would need to be used in such interfaces and spoken dialogue systems, do not yet work well for elderly users. A mismatch between the acoustic model and the acoustic characteristics of user speech is one factor which reduces speech recognition accuracy. In particular, elderly speech frequently contains inarticulate speech, which occurs when the speaker does not fully open their mouth. Therefore, elderly speech is likely to have different acoustic characteristics than the speech of younger people. Additionally, acoustic models are often constructed using the speech of adults but excluding aged persons. As a result, it has been reported that deterioration in speech recognition accuracy with elderly users is caused by the mismatch between acoustic models and the acoustic characteristics of elderly speech (Anderson, 1999; Baba, 2001; Vipperla, 2008). Therefore it is important to construct an acoustic model which takes into account the characteristics of elderly speech in order to improve the speech recognition accuracy of speech applications for the elderly.

We have constructed a new speech data corpus, using the utterances of 100 young-old, old-old and oldest-old Japanese people, to improve speech recognition accuracy of the speech of elderly people. In addition, we have compared our corpus with other speech databases which have been used to construct acoustic models in Japan

Figure 1: Recording scene in an elderly nursing home

2. Data Collection

Between May 2014 and February 2015, we collected 9.2hours (5,030 sentences) of read speech from 100 elderly, Japanese subjects. In this chapter, we describe the collected speech in detail.

2.1 Speaker Selection

Although it is possible to observe differences between teenage speech, young adult speech and elderly speech at the acoustic/phonetic level, there do not seem to be any clear-cut, age-related acoustic/phonetic differences in human speech. This is partly because the aging of the speech organs is influenced by factors such as the abuse or overuse of the vocal folds, smoking, alcohol consumption, and psychological stress and tension. Furthermore, what is usually considered to be typical of elderly speech is more often related to situational circumstances, such as lexical and grammatical factors which are commonly used to identify different sociolinguistic registers. While it might be impossible to precisely determine an exact age at which an individual's speech should be considered to be elderly, researchers usually regard 60-70 years of age as the minimum age range for elderly speech. Therefore, for our corpus we decided to collect speech from subjects aged 60 and over. Apart from age, literacy and basic technical comprehension requirements, we had no other criteria for selecting speakers. We did not, for example, aim at a specific ratio of female to male speakers or screen speakers for pronunciation, etc. During data collection, we recorded the read speech of 100 elderly subjects over 60 years of age. The age and sex distribution of the speakers is shown in Table 1. We collected speech from elderly subjects at four nursing homes for the elderly and at one university. Fig. 1 shows a recording scene at a nursing home. Table 2 shows the number of speakers recorded at each location. The number of elderly subjects recorded at nursing homes was 56, and their ages ranged from 66 to 98 (average age: 82). The number of elderly subjects recorded at the university was 44, and their ages ranged from 60 to 78 (average age: 71). Some of the subjects were suffering from dementia and more than half of the subjects were living in nursing homes. In the popular S-JNAS elderly Japanese speech database, there are only eight speakers over 80 years of age, and the overall age distribution is also biased, therefore we chose as many subjects over 80 as possible. As a result, we were able to include recorded speech from 39 individuals more than 80 years of age in our corpus. This speech data should prove valuable for acoustic modeling and elderly speech analysis.

2.2 Data Collection Procedure

Each speaker uttered about 50 ATR phoneme-balanced sentences, for a total of about 9.2 hours of recorded speech for all of the subjects combined. All of the speakers lived in Aichi prefecture. The utterances were recorded using a desktop microphone. We explained the recording procedure and provided the sentences to be read to each subject, printed in kana characters. The subjects then practiced reading the sentences. Rest breaks were provided during the recording process in consideration of the physical condition of the subjects. In addition, we conducted a post-recording, subjective evaluation survey regarding the speaker's mood (which was completed by nursing home staff instead of by the subjects) and tested the level of dementia using HDS-R (Hasegawa's Dementia Scale for Revised) the subjects after recording.

Age	Male	Female	Total
60-64	1	3	4
65-69	3	10	13
70-74	3	22	25
75-79	6	13	19
80-84	4	16	20
85-89	1	10	11
90-94	3	3	6
95-99	1	1	2
Total	22	78	100

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

Recording location	Number of speakers	Average age
Nursing home A	10	85.5
Nursing home B	10	82.8
Nursing home C	17	80.6
Nursing home D	19	81.4
Nagoya University	44	70.8

Table 2: Average age and number of speakers at each recording location.

2.3 Selection of Japanese Sentences

When choosing sentences for speakers, the goal was to create a corpus of read speech suitable for training acoustic models. Sentence selection and structure were based on the existing JNAS (Japanese Newspaper Article Speech) Japanese speech corpora, a typical corpus used for constructing Japanese acoustic models (Iso, 1988; Kurematsu, 1990). The JNAS database consists of sentences from newspaper articles divided into 155 text sets of about 100 sentences per set, with 16,176 sentences in total. In addition, it contains ATR phonetically balanced sentences divided into 10 text sets with about 50 sentences per set and 503 sentences in total. The ATR phonetically balanced sentences included 402 two-phoneme sequences and 223 three-phoneme sequences, with 625 of these items in total. The phonetically balanced sentences were extracted from newspapers, journals, novels, letters, and textbooks, etc., so that different phonetic environments occur at the same rate as much as possible. The sentences consist of Set A ～ Set I (50 sentences each) and Set J (53 sentences). We selected these same ATR phonetically balanced sentences as phrases for our speech corpus. The number of speakers uttering each phrase set in each corpus is shown in Table 5.

2.4 Database

Age	Male	Female	Total
10-19	1	0	1
20-29	90	81	171
30-39	40	47	87
40-49	11	16	27
50-59	5	5	10
60-	5	3	8
unknown	1	1	2
Total	153	153	306

Table 3: Age and sex distribution of JNAS speakers.

Set name (sentences)	JNAS	S-JNAS	New corpus
Set A (50)	1,600	2,950	500
Set B (50)	1,600	2,950	500
Set C (50)	1,600	2,950	500
Set D (50)	1,400	2,950	500
Set E (50)	1,600	3,000	500
Set F (50)	1,700	3,000	500
Set G (50)	1,600	3,100	500
Set H (50)	1,600	3,100	500
Set I (50)	1,400	3,050	500
Set J (53)	1,272	3,233	530
Total [Number of speakers]	15,372 [306]	30,283 [301]	5,030 [100]

Table 4: Number of sentences for each set.

Age	Male	Female	Total
60-64	47	52	99
65-69	49	46	95
70-74	39	35	74
75-79	11	14	25
80-84	4	2	6
85-89	1	0	1
90-94	0	1	1
95-99	0	0	0
Total	151	150	301

Table 5: Age and sex distribution of S-JNAS speakers.

The total duration of speech in our corpus recorded is approximately 9.2 hours (Table 7). Each speaker was recorded using a desktop microphone, and their speech was stored with wav header. The speech waves were digitized at a sampling frequency of 16 kHz using 16 bit audio. The recorded speech data was then divided into sentence units, and a pause of about 300 ms was inserted before and after each sentence. The new corpus was transcribed and the transcription of the speech data was verified and edited manually by trained employees who listened to the recorded speech data. When necessary, the phonemes and words of the sentences given to the speakers were changed to correspond to what the speakers actually said. The database includes information about the speakers (age, gender, subjective assessment of mood and level of dementia), recording facilities, Japanese text transcription, set (A~J) and sentence number as attribute information.

Database	Microphone	Recorder
JNAS	Desktop microphone: Sony ECM530, etc. Headset microphone: Senhhizer HMD410 & HMD25-1	No mention of the reference
S-JNAS	Desktop microphone: Sony ECM530 Headset microphone: Senhhizer HMD25-1	DAT PCM-R500
New corpus	Desktop microphone: Audio-Technica AT9930	TASCAM DR-05 VERSION2

Table 6: Recording equipment.

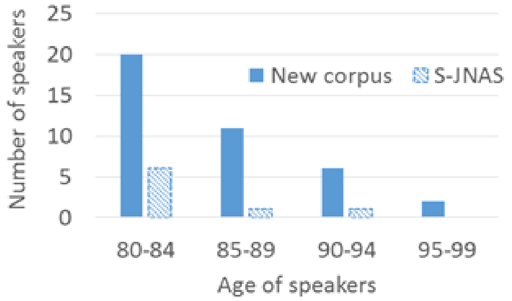

Figure 2: Age distribution of new corpus and S-JNAS

3. Japanese Speech Corpora

3.1 Comparison of Speech Corpora

Sets of phrases were selected for each corpus which would result in a collection of read speech suitable for training acoustic models for a wide variety of speech-driven applications, including dictation. Although the elderly corpus for various language exist (Cucchiarini, 2006; Hamalainen, 2012), in this paper I refer to Japanese corpus. JNAS is typically used to construct Japanese acoustic models for standard adult speech data. Here we compare our new, elderly speech database with the JNAS and S-JNAS elderly speech databases. The JNAS database includes 306 speakers, with each speaker uttering about 50 ATR phoneme-balanced sentences, while the S-JNAS includes 301 speakers, with each speaker uttering two sets (around 100 sentences) of ATR phoneme-balanced sentences. The age distributions of the speakers in the JNAS and S-JNAS corpora are shown in Tables 3 and 4, respectively, and the totals for the number of sentences uttered by speakers in each corpus are shown in Table 5.

The majority of speakers in the JNAS corpus were from 20 to 39 years old. In the S-JNAS corpus, on the other hand, most of the speakers ranged from 60 to 69 years old, with an average age of 67.6 and equal numbers of male and female speakers.

JNAS speech data was recorded at 39 facilities, mostly at universities and research institutes. S-JNAS speech data

Database	Total speech duration [s]	Average speaking rate [mora/s]
JNAS	64,403.0	7.66
S-JNAS	176,824.4	5.44
New corpus	33,008.0	4.98

Table 7: Average speaking rate for each corpus.

was recorded at two facilities in Nara Prefecture, which were not elderly facilities. Table 6 shows the type of recording equipment used to record each corpus. The microphones and recorders used to record both JNAS and S-JNAS varied among the recording facilities used. For our corpus, we used the same microphone and recorder with all of our subjects.

3.2 Comparison of Speech Duration

We conducted Voice Activity Detection (VAD) based on the speech power of all of the speech data in each corpus, and calculated average speaking rate by dividing the total duration of speech by the total number of mora in each corpus respectively. The results are presented in Table 7. These results reveal that the average speaking rate in the JNAS corpus corresponds to the average rate of speech of typical Japanese utterances (Han, 1994). As for S-JNAS and our new corpus, it is clear that these utterances are spoken more slowly than the average Japanese utterance, providing an example of how the acoustic characteristics of elderly speech are different from the speech of younger people.

Regarding age distribution, speakers aged over 80 accounted for only 3% (8 persons) of the speakers in the S-JNAS corpus, while our new corpus includes 39 subjects over 80, about 40% of the speakers. Fig. 2 compares the number of subjects over 80 in the S-JNAS corpus and in our new corpus. The average age of speakers in the new corpus is approximately 10 years older than in the S-JNAS corpus, and the average speaking rate in the new corpus is also slower than in the S-JNAS corpus. If the cause of this difference in speaking rates is due to the difference in the age distribution of the speakers, the difference in speaking rates may further increase when the speech of speakers 60 to 79 years of age is compared to the speech of speakers older than 80.

In this paper we compared the average speaking rate of each corpus by calculating total speech time. However, in the case of the Japanese language, although duration of one mora is almost constant, the duration of speech changes slightly depending on the position of the phoneme (Ota, 2003). Therefore, in order to accurately calculate speech duration, it will be necessary to precisely calculate the duration of each mora. Moreover, there is a possibility that noise in the recorded speech affected the VAD process, so it may also be necessary to improve our VAD technique.

4. Conclusion

Between June 2014 and February 2015, we collected almost 9.2 hours (5,030 sentences) of read speech from 100 elderly Japanese subjects. We recorded 56 elderly subjects whose ages ranged from 66 to 98 (average age: 82) at four nursing homes, and recorded 44 elderly subjects whose ages ranged from 60 to 78 (average age: 71) at a university. We consider our effort to be a successful example of how to collect a large corpus of high-quality read speech at a low cost from a subset of the population that is relatively difficult to engage.

Because most of the recorded subjects lived in Aichi prefecture, the collected speech may contain a distinctive Aichi accent. In the future, we will need to record the speech of subjects living in various parts of Japan. We also plan to increase the amount of speech data collected from elderly persons over the age of 80.

5. Acknowledgements

This work was partially supported by Center of Innovation (COI) Program of the Japan Science and Technology Agency (JST).

6. References

Acartürk, C., Freitas, J., Fal, M., and Dias, M.S, (2015) Elderly Speech-Gaze Interaction: State of the Art and Challenges for Interaction Design, *Universal Access in Human-Computer Interaction. Access to Today's TechnologiesVolume 9175 of the series Lecture Notes in Computer Science*, pp 3-12.

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*.

Cucchiarini C., Van hamme, H., van Herwijnen, O., Smits, F, (2006) JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality, IN *Proc. of International Conference on Language Resources and Evaluation*, pp. 135-138

Hamalainen, A., Pinto, F.M., Dias, M.S., Júdice, A., Freitas, J., Pires, C.G., Teixeira, V.D., Calado, A., Braga, D, (2012) The First European Portuguese Elderly Speech Corpus, IN *Proc. of IberSPEECH 2012*.

Han, M.S, (1994) Acoustic manifestations of mora timing in Japanese, *Journal of the Acoustical Society of America*, vol.96, no.1, pp.73-82.

Iso, K., Watanabe, T., Kuwabara, H, (1988) Design of a Japanese Sentence List for a Speech Database, *Preprints, Spring Meeting of Acoustic Society of Japan*, Paper 2-2-19, pp. 89-90 (in Japanese).

Júdice, A., Freitas, J., Braga, D., Calado, A., Dias, M., Teixeira, A., Oliveira, C, (2010) "Elderly Speech Collection for Speech Recognition Based on Crowd Sourcing, IN *Proc. of DSAI2010*, pp.103-110.

Kurematsu, A., Takeda, K., Sagisaka, S., 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.

Ota, M., Ladd, D.R., and Tsuchiya, M, (2003) Effects of foot structure on mora duration in Japanese?," IN *Proc. of 15th International Congress of Phonetic Science (ICPhS-15)*, pp.459-462.

Vipperla, R., Renals, S., Frankel, J, (2008) Longitudinal study of ASR performance on ageing voices, IN *Proc. of Interspeech 2008*.