Construction of a corpus of elderly Japanese speech for analysis and recognition

Construction of a Corpus of Elderly Japanese Speech for Analysis and Recognition

Norihide Kitaoka[1], Yurie Iribe[2] and Hiromitsu Nishizaki[3]
[1]Department of Computer Science, Tokushima University,
2-1 Minamijohsanjima, Tokushima, Japan
[2]School of Information Science and Technology, Aichi Prefectural University,
1522-3 Ibaragabasama, Nagakute-shi, Aichi, Japan
[3]Graduate School of Interdisciplinary Research, Faculty of Engineering, University of Yamanashi,
4-3-11 Takeda, Kofu-shi, Yamanashi, Japan
kitaoka@is.tokushima-u.ac.jp, iribe@ist.aichi-pu.ac.jp, hnishi@yamanashi.ac.jp

Abstract

We have constructed a new speech data corpus using the utterances of 100 elderly Japanese people, in order to improve the accuracy of automatic recognition of the speech of older people. Humanoid robots are being developed for use in elder care nursing facilities because interaction with such robots is expected to help clients maintain their cognitive abilities, as well as provide them with companionship. In order for these robots to interact with the elderly through spoken dialogue, a high performance speech recognition system for the speech of elderly people is needed. To develop such a system, we recorded speech uttered by 100 elderly Japanese who had an average age of 77.2, most of them living in nursing homes. Another corpus of elderly Japanese speech called S-JNAS (Seniors-Japanese Newspaper Article Sentences) has been developed previously, but the average age of the participants was 67.6. Since the target age for nursing home care is around 75, much higher than that of most of the S-JNAS samples, we felt a more representative corpus was needed. In this study we compare the performance of our new corpus with both the Japanese read speech corpus JNAS (Japanese Newspaper Article Speech), which consists of adult speech, and with the S-JNAS, the senior version of JNAS, by conducting speech recognition experiments. Data from the JNAS, S-JNAS and CSJ (Corpus of Spontaneous Japanese) was used as training data for the acoustic models, respectively. We then used our new corpus to adapt the acoustic models to elderly speech, but we were unable to achieve sufficient performance when attempting to recognize elderly speech. Based on our experimental results, we believe that development of a corpus of spontaneous elderly speech and/or special acoustic adaptation methods will likely be necessary to improve the recognition performance of dialog systems for the elderly.

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

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, lack of familiarity with ICT on the part of many elderly and the limited set of available interaction modalities, since this technology is mainly designed with younger users in mind. Hence, 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. 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 for 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 such as 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, and may also ease loneliness. 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, the speech recognizers which would need to be used in interfaces such as spoken dialogue systems do not currently 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. Some studies have found differences in the acoustic characteristics of the speech of the elderly and that of younger people (Winkler, 2003). In particular, elderly speech frequently contains inarticulate speech, which occurs when the speaker does not fully open the mouth.

Additionally, acoustic models are often constructed using the speech of adults, excluding the aged. As a result, it has been reported that deterioration in speech recognition accuracy with elderly users is caused by mismatches 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 designed for the elderly.

To address this problem, we have constructed a new speech data corpus using the utterances of 100 elderly Japanese people in three age categories; the young-old, old-old and oldest-old, in order to improve recognition of the speech of the elderly. In this study we compare the characteristics of our new corpus with those of two other speech databases which have been used to construct acoustic models in Japan

Figure 1: Recording speech in a Japanese nursing home for the elderly.

(JNAS and S-JNAS), and then experimentally evaluate our corpus by comparing speech recognition performance when using each of the three corpora. In order to evaluate the effect of using spontaneous speech data, we also used the Corpus of Spontaneous Japanese (CSJ), which consists of speech from presentations given at Japanese acoustics conferences, in our experiment. In addition, we compared using elderly read speech versus dialog speech as our test data. Since our end goal is a speech recognition system for the elderly that can recognize dialog speech, we performed several pilot experiments using these various sources of test data.

2. Data Collection

Between May 2014 and February 2015, we collected 9.2 hours (5,030 sentences) of read speech from 100 elderly Japanese subjects. During data collection we recorded read speech from elderly subjects at four nursing homes for the elderly and at one university. Figure 1 shows a typical recording scene at a nursing home. 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). Table 1 shows the number of speakers recorded at each location. In this section, 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, it has not been conclusively determined whether or not there are 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, psychological stress and tension. Furthermore, features which are often considered to be typical of elderly speech can be related to situational circumstances, such as lexical and grammatical factors which are associated with 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. The age and sex distribution of our speakers are shown in Table 2. All of the speakers lived in Aichi prefecture. 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 distributions 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. 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, after each recording session a subjective mood evaluation survey was conducted with each subject by facility staff, and their likelihood of suffering from dementia was assessed using the HDS-R (Hasegawa's Dementia Scale-Revised).

Table 1: Number of speakers and their average age at each recording location.

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: Age and sex distributions of speakers in our corpus.

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

2.3 Selection of Japanese Sentences

When choosing sentences for the speakers to read, the goal was to create a corpus of read speech suitable for training acoustic models, thus sentence selection and structure were based on the existing JNAS speech corpora, a typical corpus used for constructing Japanese acoustic models (Iso, 1988; Kurematsu, 1990). The JNAS database consists of sentences from newspaper articles and is 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 (625 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 JNAS ATR phonetically-balanced sentences as phrases for our speech corpus. The number of utterances of each sentence set in each corpus is shown in Table 3.

2.4 Database

The total duration of the recorded speech in our corpus is approximately 9.2 hours. Each speaker was recorded using a desktop microphone, and their speech was stored with a 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 were changed to correspond to what the speakers actually said. The database includes information about the speakers (age, gender, subjective assessment of mood and likelihood of dementia), recording location, text transcription (in Japanese), set (A~J) and sentence number as attribute information.

Table 3: Number of utterances in each sentence set.

Set Name (Number of 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: Age and sex distribution of JNAS speakers.

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
Age unknown	1	1	2
Total	153	153	306

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

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

2.5 Emotion labels

Analysis of elderly speech is indispensable for the development of dialogue systems and dialogue interfaces for elderly people. In addition to the content being conveyed by speech, the speaker's emotions are also part of the messages being conveyed, particularly during dialogues, so information about speaker emotion is increasingly being incorporated into dialogue control in dialogue systems. However, studies examining the emotional content of speech generally focus on the adult speech of subjects of relatively young ages, and little research has been done on the emotional content of the speech of the elderly. Therefore, our corpus includes emotion labels for the speech of our elderly speakers.

How to apply emotion labels to speech is an important task, and descriptive methods have largely been divided into models based on basic emotional perspectives and models based on dimensional perspectives of emotion. For our corpus, we adopted the eight basic emotions identified by Plutchik (Plutchik, 2001) which include the continuums of "joy - sadness", "acceptance - disgust", "fear - anger" and "surprise - anticipation", which we translated into Japanese. After making each speech recording, staffs with experience working with the elderly elicited information from the speakers about their emotional state. The pairs of adjectives were then used to label the emotional states of the speakers. Then, a five-point evaluation was performed on the speaker's adjective pairs. For example: emotion label of Subject 1 is a very joy, a little acceptance, a weak fear and a little surprise.

2.6 Dementia testing using HDS-R

As Japanese society rapidly ages, the number of people with dementia in the country increases each year. In July 2008, the Ministry of Health, Labor and Welfare launched an "emergency project to raise the quality of medical care and improve daily life for citizens with dementia." Early diagnosis of dementia is listed as one of measures to be taken, so it has been studied through the analysis of physical movement, brain wave measurement, cognitive assessments, etc. However, there have been few studies on the relationship between speech and dementia. For this reason, we conducted a simple diagnostic test used to measure dementia risk with each speaker after recording their speech so that we could analyze possible characteristics of elderly speech which might be linked with dementia.

Two widely used cognitive assessments used to detect dementia are the revised Hasegawa's Dementia Scale (HDS-R) (Imai, 1994) and the Mini-Mental State Examination (MMSE) (Fostein, 1975). The HDS-R consists of nine questions on age, time, place, memory, calculation, etc., and has an evaluation scale of 0 to 30 points. The MMSE examination consists of eleven questions such as time, place, memory, calculation, writing, drawing figures, etc., and also scores subjects on a scale of 0 to 30. Both tests indicate a tendency towards dementia if a subject's score falls below a given threshold. The HDS-R only detects the presence or absence of a tendency towards dementia, while the MMSE ranks the severity of a subject's dementia based on their score. For this study, we adopted the simpler HDS-R. Each subject who provided speech for our corpus was scored using the HDS-R at a later date. As a result, 11 out of the 100 original participants were judged to have a tendency towards dementia.

3. Japanese Speech Corpora

3.1 Comparison of Speech Corpora

Sets of phrases are selected for speech corpora so that the result is a collection of read speech suitable for training acoustic models for a wide variety of speech-driven applications, including dictation. Although elderly speech corpora for various languages exist (Cucchiarini, 2006; Hamalainen, 2012), in this paper we focus on Japanese corpora. The JNAS corpus is typically used to construct Japanese acoustic models for standard adult speech data. Here we compare our new, elderly speech database with the JNAS database and the S-JNAS elderly speech database. 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 of ATR phoneme-balanced sentences (around 100 sentences). The age distributions of the speakers in the JNAS and S-JNAS corpora are shown in Tables 4 and 5, respectively, and the number of utterances of each sentence in each of the three corpora is shown in Table 3.

Table 6: Recording equipment.

Database	Microphone	Recorder
JNAS	Desktop microphone: Sony ECM530, etc. Headset microphone: Senhhizer HMD410 & HMD25-1	Not given
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

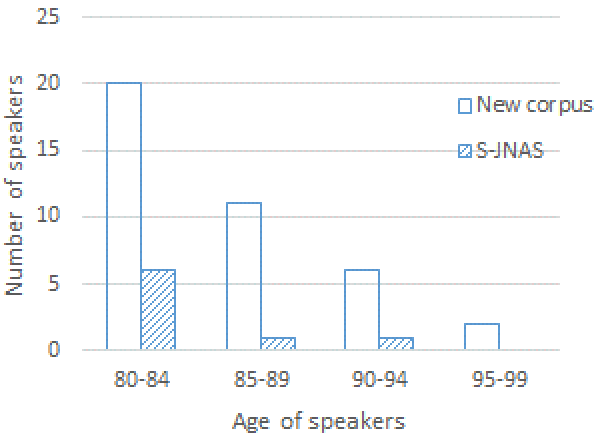

Figure 2: Age distributions of new corpus and S-JNAS.

The majority of the speakers in the JNAS corpus were from 20 to 39 years old, while in the S-JNAS corpus 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.

The JNAS speech data was recorded at 39 facilities, mostly universities and research institutes. The S-JNAS speech data 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 the JNAS and S-JNAS varied depending on the recording facility. 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 volume level of all of the speech data in each corpus, and then calculated the average speaking rate for each corpus by dividing the total duration of the speech in each corpus by the total number of morae in each corpus. The results, presented in Table 7, reveal that the average speaking rate in the JNAS corpus corresponds to the average rate of speech of typical Japanese utterances (Han,

Table 7: Average speaking rate for each corpus.

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

Table 8: Average speaking rates in new corpus for each age group.

Age	**Average speaking rate [mora/sec.]**	**Standard Deviation**
60-69	6.21	0.74
70-79	5.89	1.14
80-89	5.65	1.15
90-99	4.88	0.83

1994). Regarding the 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, which represents about 40% of the speakers. Figure 2 compares the number of subjects in each age group 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.

Table 8 shows the rate of speech for each age group in our corpus. We can see how the rate of change in speaking rates increases when we compare the speech of speakers 60 to 79 years old with the speech of speakers older than 80. Although differences in age-related changes between individuals are large, we can clearly see from these results that aging causes a decline in speaking rates, and that this change becomes especially noticeable when the speakers are over 90 years old.

In this paper we compared the average speaking rates of each corpus by calculating total speech time. However, although duration of one mora is almost constant in the Japanese language, the duration of speech changes slightly depending on the position of the phonemes (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. Speech Recognition Experiments

By examining the spectral features of speech, it becomes clear that there are various differences between the speech of older and younger adults besides speech duration, and that these differences are likely to affect speech recognition performance, the improvement of which is the goal of our research. First, we examine speech recognition performance by conducting speech recognition experiments using JNAS and S-JNAS utterances with acoustic models constructed using JNAS and S-JNAS data. We then use 30 utterances from our new corpus, consisting of acoustically balanced sentences read by elderly persons, as our test data. As a comparison, we also used 200 utterances from JNAS and 10,313 utterances from S-JNAS, which are newspaper article sentences, as test data. To train our acoustic models, we used training scripts based on the CSJ recipe in the Kaldi speech recognition toolkit (Povey, 2011). For our language model, we only used the sentences included in the training data, thus there was a very strong linguistic constraint.

Table 9: WERs (%) when using various data sources during speech recognition testing. Note that the language model constraints were very strong, thus the WERs are much lower (better) than for conventional approaches.

Train	JNAS	JNAS	S-JNAS	JNAS	S-JNAS
Eval.	JNAS	S-JNAS	S-JNAS	Ours	Ours
GMM(1)	6.00	14.35	8.27	37.67	26.28
GMM(2)	2.79	6.79	4.05	20.47	15.35
DNN(1)	2.38	4.74	3.60	19.77	13.02
DNN(2)	2.34	4.92	3.41	23.72	13.49

GMM (1): GMM-HMM (LDA+MLLT)
GMM (2): GMM-HMM (LDA+MLLT+SAT+MMI+fMMI)
DNN (1): DNN-HMM (Cross-entropy optimization)
DNN (2): DNN-HMM (state-level Minimum Bayes Risk (sMBR) w/ lattice regeneration)

Table 10: WERs (%) for speech recognition experiments using utterances of elderly speakers and BCCWJ language model (w/o acoustic adaptation).

Acoustic models		**JNAS**	**S-JNAS**	**CSJ**
Test data	Read	55.50	39.45	59.17
	Dialog	64.64	67.52	45.28

Table 11: WERs (%) for speech recognition experiments using utterances of elderly speakers and BCCWJ language model (w/ acoustic adaptation).

Original acoustic models		**JNAS**	**S-JNAS**	**CSJ**
Test data	Read	44.61	38.30	34.06
	Dialog	67.98	70.67	47.83

Recognition results are shown in Table 9. By comparing the various approaches, we can see that state-of-the-art Deep Neural Network - Hidden Markov Model (DNN-HMM) acoustic models clearly do improve recognition performance. When using JNAS and S-JNAS data for both training and evaluation, we can see from our results that matched acoustic models are very effective for obtaining good performance. By comparing the effect of using the JNAS versus the S-JNAS corpus for training, we can also see that use of the S-JNAS data improves recognition performance for our elderly target speakers. However, even when we use the S-JNAS corpus, recognition performance for the speech of our target speakers are still insufficient. We may be able to improve recognition performance by using our new corpus as training data, but the quantity of data is currently insufficient to train DNN-HMM acoustic models.

Next, we investigated recognition performance under more realistic conditions using a more general language model with DNN-HMM acoustic models. We constructed a language model using the Balanced Corpus of Contemporary Written Japanese (BCCWJ), and in addition to JNAS and S-JNAS we also used the CSJ to train the acoustic models, in order to investigate the effect on performance when recognizing spontaneous speech.

Note that we also collected additional samples of speech from 39 elderly people (13 males and 26 females) in

Tokushima, under recording conditions similar to those used during our collection of data in Nagoya. We then included this data in the training data. As for test data, we used newly recorded utterances of five elderly persons reading newspaper articles aloud, and utterances extracted from dialogs between each of eight elderly persons and an interviewer.

Experimental results are shown in Table 10. From these results we can see that recognition of speech generated by elderly speakers is a very difficult task. Results when using the S-JNAS read speech data for training the acoustic model were much better than when using the JNAS read speech data, which is understandable, but even the use of S-JNAS data was not effective for the recognition of elderly dialog speech. In contrast, the CSJ trained acoustic models were more effective for dialog speech recognition, indicating that matching the speaking styles of training and test data is very important, even if there is a “generation gap” between the speakers.

We also used acoustic models adapted with transfer training, using JNAS, S-JNAS, and CSJ models, respectively, as the original models and then applying additional back-propagation to the models using the read speech of elderly people from our new corpus. Results are shown in Table 11. In the case of read speech, performance of all of the original acoustic models were improved, however this was not the case for dialog speech. The adaptation data was read speech, and thus it was very effective for improving performance with read speech, by “bridging the gap” of the various generational differences discussed in the previous sections. As for dialog speech, we could not achieve any improvements, even when the original acoustic models were trained using the CSJ data. These results indicate that mismatches between test data and training/adaptation data are highly detrimental to recognition accuracy.

Although we have not achieved acceptable performance in our attempt to develop a dialog recognition system for the elderly, we have learned valuable lessons, and we think that there may be two other ways to tackle this problem:

Collection of spontaneous elderly speech for model training.

Development of a new adaptation method which does not corrupt important characteristics of the original acoustic models, for example, corruption of the spontaneity of the CSJ trained models.

In our future work, we will use these two techniques to improve our method, but the first technique is very costly, thus we will mainly concentrate on the second technique.

5. Conclusions

In this study we attempted to improve recognition of the speech of elderly Japanese by spoken dialog systems through the creation a superior corpus of elderly Japanese speech. Experimental evaluation of our new corpus, using a variety of ASR systems and techniques, showed that it was not effective for improving elderly speech recognition performance. However, we did confirm that using elderly speech when developing speech recognition systems for the elderly is effective, and that matching the speaking styles of the training and test data for the language model (read speech vs. spontaneous speech vs. dialog speech) is also very important. In addition, we discovered that it is important not to corrupt the acoustic model during adaptation. We also confirmed that state-of-the-art DNN-HMM acoustic models improve speech recognition performance. Finally, we consider our process for developing a corpus of elderly speech to have been largely successful, as we were able to collect a sizeable corpus of high-quality read speech at a low cost, from a subset of the population that is relatively difficult to engage.

6. Acknowledgements

This work was partially supported by the Center of Innovation (COI) Program of the Japan Science and Technology Agency (JST), the Strategic Information and Communications R&D Promotion Program (SCOPE) of the Ministry of Internal Affairs and Communications of Japan, and JSPS KAKENHI Grant-in-Aid for Scientific Research (B) Grant Number 17H01977.

7. References

Acartürk, C., Freitas, J., Fal, M., 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 Technologies, Volume 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

Fostein, M.F., Fostein, S.F., McHugh, P.R., (1975). MiniMental State: A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res, vol. 12, pp. 189-198.

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.

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.

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.

Maekawa, K., Koiso, H., Furui, S., Isahara, H., (2000). Spontaneous speech corpus of Japanese, In Proc. of LREC2000, pp. 2013-2018.

Ota, M., Ladd, D.R., 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.

Plutchik, R., (2001). The nature of emotions, American Scientist, 89(4), pp. 344-350.

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

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

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.