Estimating Age-Dependent Degradation Using Nonverbal Feature Analysis of Daily Conversation

Natsumi Kana[1], Yumi Wakita[1(✉)], and Yoshihisa Nakatoh[2]

[1] Osaka Institute of Technology, Osaka, Japan
yumi.wakita@oit.ac.com
[2] Kyushu Institute of Technology, Fukuoka, Japan

Abstract. In this paper, we study a system that estimates the degree of decline in the driving ability of elderly people using non-verbal information from daily conversations. It is necessary for us to find the factors that would affect the calculation of the degree of decline that has reached a problematic level for functioning daily life. We focus on the cases where elderly people cannot understand their partner's speech as their hearing and concentration abilities decrease with age. We analyze the relationship between the degree of understanding of the partner's speech and the non-verbal characteristic of the response scene. Based on the results of the acoustic analysis of each utterance, the fundamental frequency (F0) and acoustic power levels of when a person can understand their partner's speech tend to be higher than those when they cannot. The analysis of the synchronization of the head motions shows that brightness value of difference image when a person can understand their partner's speech is also higher than when they cannot. These results indicate that these non-verbal factors are effective in estimating the decline in the hearing and concentration abilities of the elderly.

Keywords: Degree of decline · Fundamental frequency · Synchronism of motion · Understanding level

1 Introduction

The number of traffic accident and death for elderly people is increasing every year, the main reason being "inappropriate driving". Although there is a drop in their driving ability, some elderly people don't have consciousness of their decline. They consider that it is the same as that of their youth. Several plans for dementia prevention and health maintenance measures have been proposed. However, such tools are not actively used for elderly people who have only a slight interest in health, thereby causing the sudden occurrence of serious accidents leading to death. Therefore, it is very important for a third party to notice the decline in driving ability of the elderly and to inform them of the decrease.

We are currently working on developing a system that can estimate the degree of decline. It depends on the age of an individual and informs them when the estimated result indicates a decline. We have already determined some acoustic features that are

V. G. Duffy (Ed.): HCII 2019, LNCS 11582, pp. 222–231, 2019.
https://doi.org/10.1007/978-3-030-22219-2_17

effective for evaluating this age-dependent degradation. The voice of a speaker over the age of 75 has some special characteristics [1]. For example, their laughter often becomes a voiceless sound, the value of F0 is unstable, and its distribution of F0 increases. As compared to speakers under the age of 75, these differences are observed to be significant when using a t-test. Several other papers have illustrated that the acoustic features of the human voice are effective in detecting the degradation as it ages [2–4]. For example, Tanaka et al. [5] reported the formant frequency shift in elderly speech.

However, these reports only describe age-dependent degradation. They do not estimate whether the degree of decline has reached a problematic level in daily life. This study aims to explore a system that can be used to evaluate the degree of decline in the driving ability of elderly people based on their daily conversations and inform them when the estimated result indicates a decline. To create this system, it is necessary to find the factors that could affect the evaluation of the degree of decline before it becomes an illness.

Davies HR et al. conducted a survey on approximately 15,000 people over the age of 50 and found that the risk of dementia in people with moderate loss of hearing is approximately 1.6 times more than those with normal hearing ability [6]. It is a known fact that a person's hearing and concentration ability declines with age. An older person is more likely to have difficulty understanding their partner's speech. If a system can distinguish an elderly's response in such a case, it can inform them that their hearing or concentration abilities are problematic. It may prove to be effective in minimizing the incidences of dementia and depression.

We have established that the F0 value for a person who can understand their partner's speech tends to be higher than those who cannot [7]. In this paper, we add to analyze acoustic power level and synchronism of the gestures of the speakers by using the free conversation database of elderly people. The result of this analysis enabled us to suggest a method that can automatically estimate the comprehension level of a person.

2 Conversation Analysis

2.1 Free Conversation Recording

We recorded eight sets of 3-min dyadic conversations. Figure 1 shows the location of these recordings. We used two microphones and a video camera for this purpose. Figure 2 is an example photo extracted from the video data. The conditions of the recordings are listed in Table 1. The participating speakers were previously acquainted, but we paired those people who had no prior contact with each other.

2.2 Acoustic Analysis Method

As a method of calculating F0, an algorithm called robust algorithm for pitch tracking (RAPT) was used. RAPT realizes F0 extraction with high accuracy by postprocessing using a dynamic programming method with multiple F0 candidates obtained using a

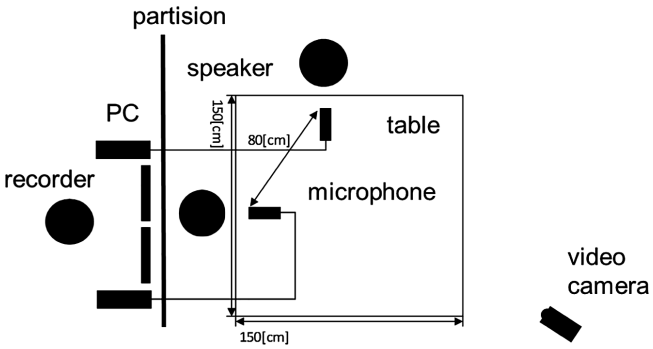

Fig. 1. Location of recording conversation

Fig. 2. Example photo extracted from the video data

Table 1. Conditions of conversation

Number of speakers	9 elderly speakers (5 males, 4 females)
Ages	62–82 years old
Number of conversations	8 conversations by 9 speakers
Conversation periods	Three minutes/conversation
Conversation condition	Free dyadic conversation

correlation method. In RAPT, the time period to the peak of the normalized cross-correlation function shown by formula (1) taken as an F0 candidate [6]. x(n) means n-th speech signal.

$$\emptyset(\mathrm{m}) = \frac{1}{\sqrt{e_1 e_m}} \sum_{\mathrm{n}=1}^{\mathrm{N}-1} \mathrm{x(n)w(n)x(n + |}m|)\mathrm{w(n + |}m|) \;\; |m| = 0, 1, \ldots, N-1 \qquad (1)$$

$$e_j = \sum_{n=j}^{j+N-1} x(n)w(n) \tag{2}$$

As a method of calculating acoustic power level "*Lp*", we used the following formula (3)

$$L_p = 10\log_{10}\left\{\frac{1}{Tp_0^2}\int x(t)^2 dt\right\} \tag{3}$$

"$p0$" means minimum audible level, the value is 20[μPa]. "T" means the utterance period.

2.3 Response Motion Analysis Method

In this paper, we focus to analyze the synchronism of response motion (Ex. nodding, head swing etc.). After AD conversion, we used a frame difference method to analyze the area around head of each speaker. After extracting image sequences from the moving data, the difference images were calculated using the following formula (4). Here, I_k indicates the value of the kth frame in the image sequences, I_{k+1} indicates the value of the (k + 1)th frame, and the absolute difference image ID is defined as follows:

$$ID_{(K,K+1)} = |I_{K+1} - I_K| \tag{4}$$

After calculating the difference values between the frames, the difference image is converted to a grayscale image. The contributions of the three colors (red, blue, and green) were 29.9%, 11.4%, and 58.7%, respectively. Binary processing was performed on the brightness value of the difference image. A brightness value equal to or larger than the threshold value was set to 1 while that less than the threshold value was set to 0. One image dataset included two speakers. We used a mask file to separate the two speakers one by one. The sum of the brightness values of all pixels for each speaker was then calculated.

To analyzing the synchronism of the motion by each speaker, we calculate the multiplied values of the sum of brightness values for each speaker. Before calculating, we used a moving average method to the sum of brightness values of each speaker to decrease the influence of a slight time lag. The calculation values multiplied after a moving average process are used to evaluate the synchronism of the motion by each speaker.

3 F0 Analysis Experiments

3.1 Response Extraction

We listened to the recorded conversations and extracted only the responses using 8 conversations database spoken by 5 male persons. Table 2 shows the number of

extracted response utterances. After extraction, we calculated the F0 of the extracted speech and excluded those whose F0 values could not be calculated.

Table 2. Number of extracted response utterances

	"Ah"	"Uh"	"Eh"	"Oh"
Number of utterance	77	96	9	5

We asked four persons to listen to the 30-second conversations, which included these response utterances, and judge the understanding level of the speakers. The judgment was performed in the following five steps: "5: He can understand the partner's speech", "4: He seems to be able to understand it", "3: I cannot say", "2: He does not seem to be able to understand it", "1: He cannot understand". We applied this 5-step evaluation to the extracted speech samples. The average value of the understanding level as the judgement results by four persons was defined as "Ave-UL".

3.2 Relationship Between F0 Value of Response Utterance and Understanding Level

We analyzed the F0 value of each response expression. Figure 3 shows the relation between the result of the 5-step evaluation and the average values of F0 of each utterance (Ave-F0). The F0 values are normalized by calculating the average of all F0 values of each speaker. The blue square points in Fig. 3 indicate the F0 values of the responses when the understanding level is under three, and the red round points indicate the F0 values when the understanding level is over three. The dotted line specifies the results of the regression analysis.

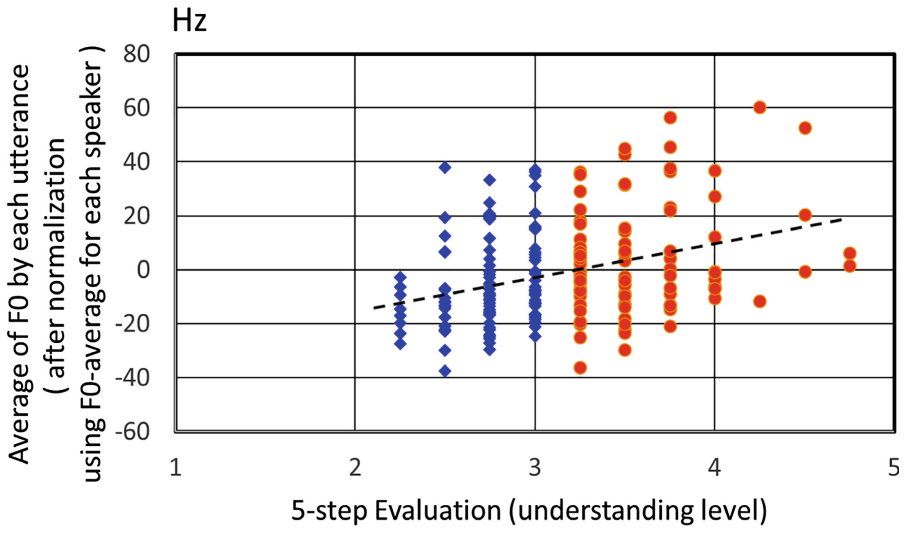

Fig. 3. Relationship between the Ave-F0 and the Ave-UL (Color figure online)

The results of the regression analysis indicate the Ave-F0 values when Ave-UL are low tend to be lower than those when the Ave-UL are high. But the tendency is modest.

3.3 Relationship Between F0 Values and Acoustic Power Levels

Figure 4 shows the relation between the "Ave-F0" and the average of acoustic power for each utterance (Ave-Power). Both of "Ave-F0" and "Ave-Power" values are already normalized by the average value for each speaker.

The response of which Ave-UL is under 2.5 was defined as **"Res-NUS".** That means the response of the speaker when they could neither understand nor listen to the partner's speech. The response of which Ave-UL is over 3.5 was defined as **"Res-US"**. That means the response when the speaker could either understand or listen to the partner's speech.

The blue triangle points in Fig. 4 indicate the data of "Res-NUS" and the red round points indicate the data of "Res-US". The two black lines show the predicted interval calculated for the data of "Res-NUS", which has a confidence level of 95%.

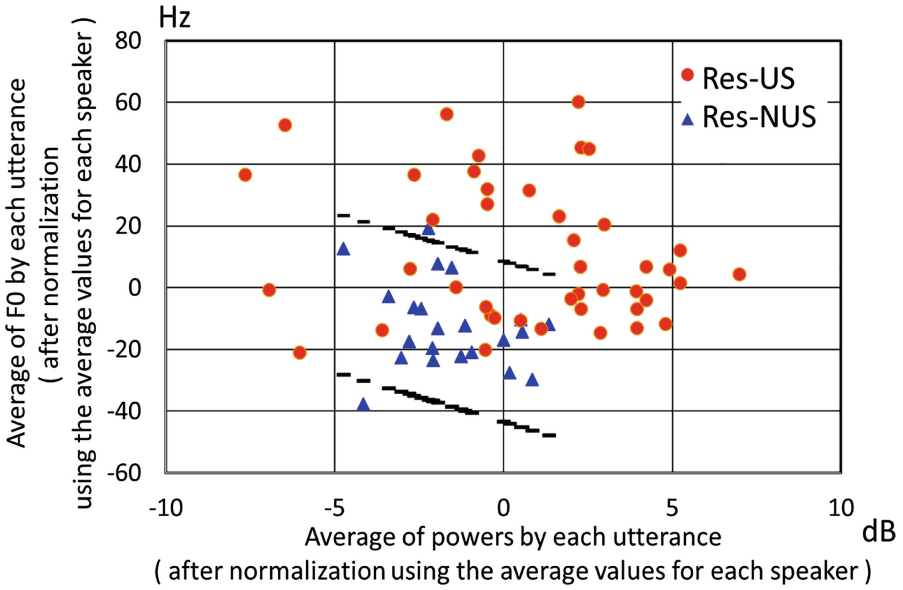

Fig. 4. Distribution of the Ave-F0 and the Ave-power of each speech utterance (Color figure online)

The results show the following:

- For almost of "Res-NUS", both of the "Ave-F0" and the "Ave-Power" values are under zero.

- In the case that both of "Ave-F0" and "Ave-Power" are over zero, all data are the case of "Res-US".
- The distribution when "Res-NUS" is narrower than that when "Res-US".
- Many data of "Res-US" were plotted outside of the prediction interval lines. These results illustrate the possibility that an elderly's understanding level of the partner's talk can estimate using F0 value and acoustic power level.

4 Brightness Analysis Experiments

4.1 Response Extraction

We extracted the 5 conversations from the database shown in Table 1. The extracted data included both response scenes of "Res-US" and "Res-NUS" in the same conversation. The total number of frames for "Res-US" and "Res-NUS" is 118 frames and 156 frames, respectively. These conversations were used to calculate the multiplied values of the sum of the brightness values between the two speakers. The sampling period of AD conversion is 0.33 s (3 frames per second). The moving averages are calculated by using 5 frames.

4.2 The Relation Between the Synchronism of the Motion of Each Speaker and Their Understanding Level

We extracted 5 conversations from database in Table 1 and calculated the multiplied values of the brightness values between the speakers. The results are platted in Fig. 5. The red round points indicate the multiplied values for "Res-US" and the blue triangle points are that for "Res-NUS".

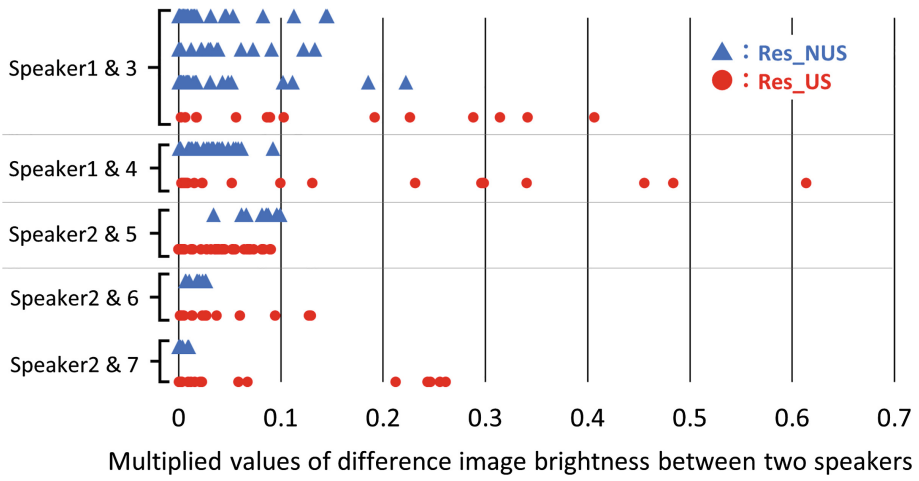

Fig. 5. Comparison of the synchronism degree between "Res-US" and "Res-NUS" (Color figure online)

The results are as follows:

- The multiplied values of "Res-US" tend to be greater than those of "Res-NUS".
- This tendency depends on the pair of speakers. In the case of only "Speaker2&5", the multiple values between for "Res-US" and for "Res-NUS" are almost same.
- The multiple values also depend on the pair of speakers.

To understand the reason behind the motion of a person who could not understand their partner's speech, we selected some response scenes in which the multiplied values are low and compared them to the brightness values by each speaker.

The Fig. 6(A) show the brightness values of two speakers in the case of "Res-NUS". The Fig. 6(B) show the multiplied values of the brightness values between the two speakers indicated in the Fig. 6(A). The two parts surrounded by dot lines in Fig. 6 (B) are examples that multiplied values are low. However, the brightness values of each speaker are not low (the parts surrounded by dot lines in Fig. 6(A)). This is due to the asynchronization of the motion between the two speakers when the duration for the moving average process is 5 frames.

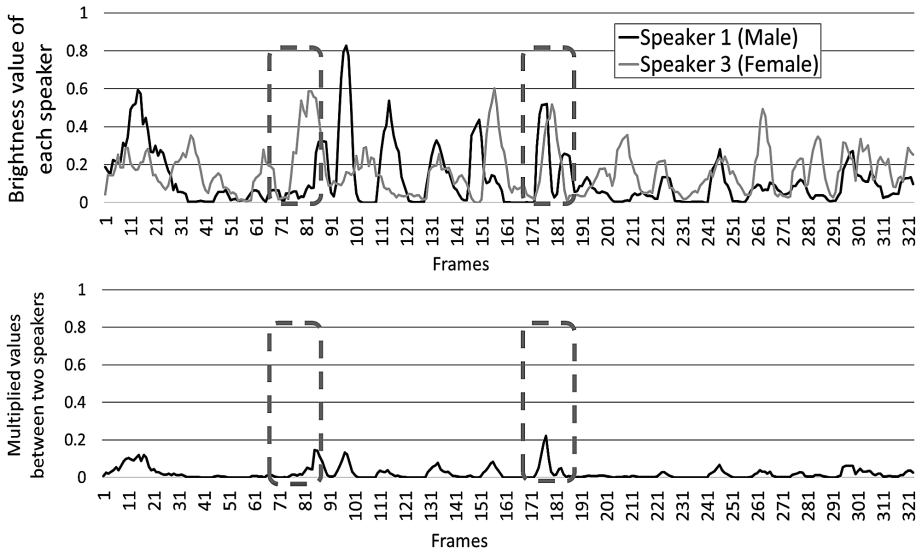

Fig. 6. (A) (Upper figure) Brightness values of two speakers in the case of "Res-NUS" (B) (Lower figure) Multiplied values between the two speakers, same as in figure (A).

The Fig. 7(A) show the brightness values of two speakers in the case of "Res-US". The Fig. 7(B) show the multiplied values of the brightness values between the two speakers indicated in the Fig. 7(A). Both of brightness values of each speaker are high and the multiplied values in Fig. 7(B) are also high. These are the cases which head motions are synchronized.

The conversation examples shown in Figs. 6 and 7 were spoken by the same speakers pair (speaker 1 and speaker 3). These figure suggest that even if the same

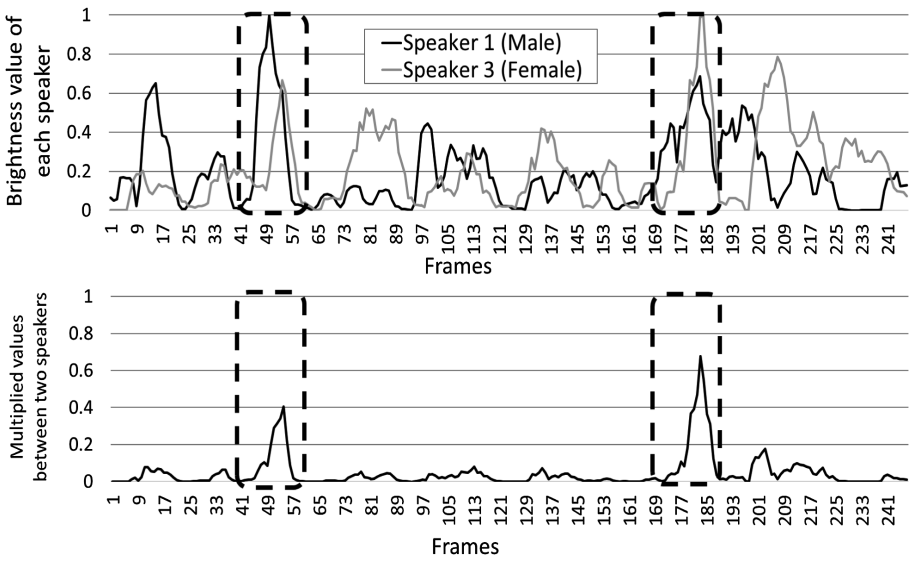

Fig. 7. (A) (Upper figure) Brightness values of two speakers in the case of "Res-US" (B) (Lower figure) Multiplied values between the two speakers, same as in figure (A)

speaker's conversation, the multiplied values change by the understanding level to the partner's talk.

5 Conclusion

We studied a system that estimates the degree of decline in the driving ability of elderly people using non-verbal information from daily conversations.

We discussed a system that estimates the degree of decline in the driving ability of elderly people using non-verbal information from daily conversations. We focused on the cases where elderly people cannot understand their partner's speech as their hearing and concentration abilities decrease with age and analyzed the F0 and acoustic power values of each utterance, and the synchronism of the head motion for each speaker using response scenes of daily conversation database. As results of analysis, it had a tendency that when understanding levels become low, both of the "Ave-F0" and "Ave-power" decreased and the synchronism of the head motion for each speaker also decrease. These results indicate that these non-verbal factors would be effective in estimating the decline in the hearing and concentration abilities of the elderly in daily life and suggest the probability of inform the degree of decline in the driving ability to elderly persons.

References

1. Wakita, Y., Matsumoto, S.: Communication smoothness estimation using F0 information. In: 2016 Proceedings of the 4th IIAE International Conference on Intelligent Systems and Image Processing, September 2016
2. Nueller, P.B., Sweeney, R.J., Barbeau, L.J.: Acoustic and morphologic study of the senescent voice. Ear Noise Throat J. **63**, 71–75 (1985)
3. Sato, K., Sakaguchi, S., Hirano, M.: Histologic investigation of bowing of the aged vocal folds. Throat J. **8**, 11–14 (1996)
4. Nishio, M., Tanaka, Y., Niimi, S.: Analysis of age-related changes in the acoustic characteristics of the voice. Jpn. Soc. Logop. Phoniatr. **50**, 6–13 (2009)
5. Tanaka, Y., Igaue, H., Mizumachi, M., Nakatoh, Y.: Study of improvement of intelligibility for the elderly speech based on formant frequency shift. Int. J. Comput. Consum. Control (IJ3C) **3**(3), 57–65 (2014)
6. Talkin, D.: A robust algorithm for pitch tracking (RAPT). In: Kleijn, W.B., Pailwal, K.K. (eds.) Speech Coding & Synthesis, pp. 495–518. Elsevier, Amsterdam (1995)
7. Natsumi, K., Wakita, Y., Nakatoh, Y.: Changes in fundamental frequency and gesture of response corresponding to the understanding level of partner's talk. In: 2018 IEEE International Conference on Artificial Intelligence in Engineering and Technology, November 2018