Building a database for likability evaluation of uttered speech

Masanori Morise^{1,*}, Fumiya Yokomori² and Kenji Ozawa²

¹*School of Interdisciplinary Mathematical Sciences, Meiji University, 4-21-1 Nakano, Nakano-ku, Tokyo, 164-8525 Japan*

²*Faculty of Engineering, University of Yamanashi, 4-3-11 Takeda, Kofu, 400-8511 Japan*

(Received 30 September 2018, Accepted for publication 18 April 2019)

Keywords: Speech database, Likability, Linguistic information, Speech analysis, Speech perception

PACS number: 43.71.Bp [doi:10.1250/ast.41.423]

1. Introduction

Speech communication is important for building relationships with others, and being able to give a good impression is especially important. To improve one's communication skill, training is generally carried out by expert trainers. On the other hand, training tools that use speech processing technology have been studied to provide the training without trainers. For example, NOCOA+ [1] has been proposed as a communication skill training tool. Such a tool requires a function to measure various kinds of impressions perceived during communication. In this paper, we focus on likability as the target impression.

The purpose of this study is to support communication skill training with a focus on likability. For this purpose, a speech database was built. Statistical analysis of uttered speech requires that the uttered sentences are unified [2]. To build a database for a likability evaluation, we first selected sentences on the basis of likability and recorded the utterances of 20 speakers reading the sentences. A basic analysis of the recorded speech was carried out, and the results showed that the tendency of the acoustic features is consistent with those in conventional studies.

2. Related works and concept of our database

Speech databases have been built for various purposes. ATR 503 sentences [3] make up one of the most popular databases built for natural language processing and automatic speech recognition studies in Japan. This database consists of phonetically balanced sentences that are arranged to enable several phonetic environments to occur at the same rate as much as possible. The Utsunomiya University spoken dialogue database for paralinguistic information studies (UUDB)[†] has been built on the basis of a concept different from that of ATR 503 sentences. This database does not use a sentence list to record natural speech. The contents are recorded on the basis of a "4-frame cartoon sorting task" [4] to obtain emotional utterances. Since speakers are not required to utter previously specified sentences, natural conversation with various kinds of emotions is recorded.

The purpose of a database for a likability evaluation is to describe what acoustic features are related to likability. To statistically analyze acoustic features, all speakers should utter

the same sentence on the basis of several levels of likability that each speaker imagines. Since there is a possibility that the sentences affect the likability perceived from uttered speech, the likability perceived from sentences should also be measured. The sentence list used for the recording is determined on the basis of the result.

3. Building of database

3.1. Determination of the sentences

We first collected sentences from four textbooks used in Japanese high schools for the evaluation. A total of 300 sentences were divided equally into three categories of likability (high, moderate, and low). Phonetic balance was ignored when collecting the sentences. In the evaluation, the sentences were randomized and presented to the subjects (15 students). Since likability may depend on the generation of the subjects, the age range of all subjects was limited to 20–24. They evaluated sentences as a score from 1 (extremely low) to 7 (extremely high).

Figure 1 shows the experimental result. The horizontal and vertical axes represent the sentence number and the average score, respectively. Sentence numbers are sorted in descending order of the score. The middle line represents the average, and the top and bottom lines represent the 95% confidence interval. The result shows that the average scores were distributed in a wide range from around 2 to 6.

We then selected 30 sentences from those in the result; ten sentences each for high, moderate, and low likability were selected. The high and low sentences were selected on the basis of the average scores, and the moderate sentences were selected from those with scores close to 4. In cases where there were sentences with the same score, the sentence with a smaller 95% confidence interval was selected as the more reliable one.

3.2. Recordings

We recorded selected utterances on the basis of the recording conditions shown in Table 1. The speakers had no experience in communication skill training. In the recordings, we instructed them to utter each sentence on the basis of the three levels of likability (high, moderate, and low). However, we provided no context regarding likability to these speakers, so they uttered all sentences on the basis of what they imagined each level to be like. The sentences were randomized and provided to each speaker, but the order was fixed to high, moderate, and low likability.

*e-mail: mmorise@meiji.ac.jp

[†]<http://research.nii.ac.jp/src/en/UUDB.html>

Fig. 1 Evaluation result of all sentences. The sentence number is sorted in descending order of score.

Table 1 Recording conditions.

Number of speakers	20 (10 males and 10 females)
Environment	Soundproof room
A-weighted SPL	18 dB
Sampling	96 kHz/24 bit
AD converter	Antelope ECLIPSE 384
Microphone	NEUMANN U87Ai

Fig. 2 Evaluation results. Error bars represent 95% confidence intervals.

A total of 1,800 utterances (20 speakers $\times$ 30 sentences $\times$ 3 levels of likability) were recorded into the database. In this paper, we analyzed a basic parameter related to likability used as described in the conventional study [5]. Since that study had revealed that speech with a higher fundamental frequency (f_0) is preferred to the original speech, we compared the f_0 values of all utterances. In the f_0 estimation, we used the Harvest frequency estimator [6] provided in the WORLD speech synthesis system [7]. The median value of the estimated f_0 contour was used for the evaluation. We measured the rate of change with moderate and another (high or low) likability. Figure 2 shows the results. These results clearly show that the speakers attempted to utter the sentences with a higher f_0 to improve likability. To decrease likability, they attempted to utter the speech with a lower f_0 .

3.3. Discussion

We first considered two types of likability in this study. Likability measured by subjective evaluation [5] depends on the perception of listeners. On the other hand, likability in recorded speech is based on the imagination of speakers. We discriminated the perceived likability from the imagined likability. The analysis result suggests that the speakers believe that a higher f_0 is preferred to a lower f_0 . Since the perceived likability depends on f_0 , imagined likability should be related to the perceived likability.

4. Conclusion

In this paper, we explained the building of a database to support the studies of likability in uttered speech. We first collected 300 sentences from four textbooks used in Japanese high schools. Recordings were then carried out by 20 speakers. A speech database with a total of 1,800 utterances was built. An analysis of the f_0 values related that the speakers attempted to increase f_0 to improve likability. This result suggests that the perceived likability shown in conventional studies is related to the imagined likability.

The next step in this study is the subjective evaluation of the uttered speech and the statistical analysis of the recorded speech. We will attempt to develop a training tool capable of generating advice during communication training on the basis of the features of utterances.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP16H05899 and JP16H01734.

References

- [1] H. Tanaka, S. Sakti, G. Neubig, T. Toda and S. Nakamura, "NOCOA+: Multimodal computer-based training for social and communication skills," *IEICE Trans. Inf. Syst.*, **E98-D**, 1536–1544 (2015).
- [2] R. Jurgens, K. Hammerschmidt and J. Fischer, "Authentic and play-acted vocal emotion expressions reveal acoustic differences," *Front. Psychol.*, **2**, 1–11 (2011).
- [3] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Commun.*, **9**, 357–363 (1990).
- [4] H. Mori, H. Kasuya, M. Nakamura and M. Amano, "Some considerations for designing spoken dialogue database from the viewpoint of paralinguistic information," *Acoust. Sci. & Tech.*, **24**, 376–378 (2003).
- [5] B. C. Jones, D. R. Feinberg, L. M. DeBruine, A. C. Little and J. Vukovic, "Integrating cues of social interest and voice pitch in men's preferences for women's voices," *Biol. Lett.*, **4**, 192–194 (2008).
- [6] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," *Proc. Interspeech 2017*, pp. 2321–2325 (2017).
- [7] M. Morise, F. Yokomori and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, **E99-D**, 1877–1884 (2016).