Voice Conversion for Improving Perceived Likability of Uttered Speech

Shinya HORIIKE\textsuperscript{†a}, Nonmember and Masanori MORISE\textsuperscript{††,†††, Member}

SUMMARY To improve the likability of speech, we propose a voice conversion algorithm by controlling the fundamental frequency (F0) and the spectral envelope and carry out a subjective evaluation. The subjects can manipulate these two speech parameters. From the result, the subjects preferred speech with a parameter related to higher brightness.

\textit{key words:} speech analysis/synthesis, voice conversion, spectral envelope, likability

1. Introduction

Speech is an important communication for building relationships with others; uttered speech leaves an impression on the other person. Leaving a good impression on another person is important for smoothly building a good personal relationship. Speech training for improving communication skills has been carried out by expert trainers. Training tools that use speech processing technology are being studied to provide training without trainers. For example, NOCOA+\textsuperscript{1} has been proposed as a social skill training tool. Such tools would support not only social skills training but also speech training for foreigners.

There are many kinds of impressions perceived during communication (e.g., linguistic, para-linguistic, and non-linguistic information). Conventional systems have focused on linguistic information by using an automatic speech recognition (ASR) system. Uttering exact phoneme sequences is important, but humans can also receive a good impression from speech that includes incorrect phonemes. To show the relationship between speech and likability, acoustic features are important. We therefore focus on likability as the important factor in speech communication. Several speech parameters have been examined in relation to the perceived impression from speech. For example, uttered speech was used as judgment material for the speaker\textsuperscript{2}. Emotion recognition from speech waveforms\textsuperscript{3} has been examined by analyzing acoustic and linguistic cues. Pitch information is an important feature\textsuperscript{4} in human mate choice.

We propose a voice conversion algorithm to improve the likability of uttered speech. This algorithm attempts to improve likability while maintaining speaker individuality. Since one simple way is to shift the pitch, we propose an algorithm to convert the timbre information by controlling the spectral envelope. To control the speech parameters, we use a high-quality vocoder.

2. Related Work

A study\textsuperscript{5} examined the relationship between the fundamental frequency (F0) and personality impression in the Big-Five factor model\textsuperscript{6}. It used English conversational teaching materials as the original speech. Five kinds of speech stimuli we synthesized in which the F0 contour was uniformly converted. These stimuli were listened to by a total of 132 people including 78 men and 54 women. As a result of the experiment, there was a tendency in which the subjects judged that the speaker leaves a higher diplomatic impression when the F0 of their speech is higher.

Another study\textsuperscript{7} examined the influence of the F0 of female speech on likability. The authors of that study used utterances of “I really like you” and “I don’t really like you” by four women. In addition, they also used pitch-shifted speech. These speech stimuli were listened to by a total of 60 people consisting of 30 men and 30 women. The results showed that the male subjects preferred the speech with higher pitch. It also showed that measurement of perceived likability is difficult because of the dependency of gender.

The purpose of this study is to show the acoustic features related to perceived likability in each gender. Since the conventional studies showed that pitch information is important, we focus on not only pitch but also timbre. We propose a voice conversion algorithm by converting F0 and the spectral envelope.

3. Proposed Algorithm

The algorithm converts the speech parameters obtained by a vocoder. Below, we explain the vocoder used in the experiment and the algorithm used in the F0 conversion and spectral envelope conversion.

3.1 Vocoder Used for Voice Conversion

The proposed algorithm requires a high-quality vocoder that can manipulate the pitch and timbre. The conventional
method [7] manipulates F0 by using the Pitch Synchronous Overlap Add (PSOLA) technique, whereas our algorithm converts not only F0 but also the spectral envelope. In this research, we used WORLD [8] (D4C edition [9]) as the high-quality vocoder that fulfills the requirements for manipulating the speech parameters related to pitch and timbre. It decomposes the speech waveform into F0, spectral envelope, and aperiodicity. The algorithm does not use aperiodic.

3.2 F0 Conversion

Here, we explain how to manipulate F0 for improving likability. In the conventional algorithms [5], [7], F0 is controlled by multiplying it by a coefficient. To fix the dynamic range of intonation, F0 should be converted to the mel scale [10]. In the proposed algorithm, F0 is converted to the mel scale by using the following equation.

\[
m = 1127.01048 \log \left( \frac{f}{700} + 1 \right),
\]

where \( f \) represents the input frequency (Hz). When F0 is shifted by adding a coefficient on the mel scale, we can control the pitch without changing the dynamic range of intonation.

3.3 Spectral Envelope Conversion

In voice conversion, not only F0 but also the spectral envelope is converted to improve the naturalness of the synthesized speech. Spectral stretching is generally carried out as the conversion. However, spectral stretching changes speaker individuality. To maintain speaker individuality, we propose another approach instead of spectral stretching.

Our conversion algorithm linearly shifts the spectrum envelope. The spectral envelopes in all frames are shifted on the basis of a parameter related to the amount of shift. Equation (2) is used when the speech conversion is performed by changing the value of \( v \).

\[
\hat{H}(n, f) = \begin{cases} 
H(n, f - v) & (f_{\min} \leq f - v \leq f_{\max}) \\
H(n, f_{\max}) & (f - v > f_{\max}) \\
H(n, f_{\min}) & (f - v < f_{\min}) 
\end{cases}
\]

where \( H(n, f) \) is the spectrogram, and \( f \) and \( n \) represent the frequency and the frame index, respectively. \( f_{\max} \) and \( f_{\min} \) represent the maximum and minimum frequencies used as a safeguard. In cases where \( v \) is larger than \( f \), the frequency in the spectrogram is negative. In this case, the algorithm refers to \( H(n, f_{\min}) \). The usage of \( f_{\max} \) is the safeguard for not exceeding the Nyquist frequency. We therefore set the parameters \( f_{\max} \) and \( f_{\min} \) to the Nyquist frequency and 0 Hz, respectively.

Figure 1 shows an example of spectral envelopes of the source and the converted spectral envelope (\( v = 150 \text{ Hz} \)) in a frame. The vertical and horizontal axes represent logarithmic amplitude and frequency, respectively. From this figure, the spectral envelope is shifted in the positive direction on the frequency axis.

Since the proposed method also changes the spectral centroid [12], it is defined as the way to control the perceived brightness. In cases where we convert the spectral envelope with a large value of \( v \), the speaker individuality of the synthesized speech changes. We confirmed in advance of the experiments that the speaker individuality is maintained by using a parameter in the range used in the following evaluation.

4. Subjective Evaluation

Regarding likability, there is a dependence on speakers in the variation in F0 and the spectral envelope. The best parameters would depend on the speaker. Therefore, it is not appropriate to fix the variations in F0 and the spectral envelope. We designed a GUI for experiments in which the subject manipulates these variations and reproduces the converted speech. The effectiveness of this GUI was verified in an experiment in which each subject could freely control parameters to obtain speech with the maximum likability.

4.1 Experimental Conditions

Before the evaluation, we aurally checked the sound quality of synthesized speech and confirmed that there was no fatal error that would cause degradation of the synthesized speech. This check was important to avoid degradation when using the vocoder.

Table 1 represents the conditions in the evaluation, and the subjects evaluated each speech using the GUI shown in

![Figure 1: Examples of spectral envelope (\( v = 150 \text{ Hz} \)).](image)
Fig. 2 GUI used in subjective evaluation.

Fig. 2. The subject could manipulate F0 and the spectral envelope by using sliders set in the horizontal and vertical axes. The subject evaluated one speech stimulus in three steps.

1. The subject scores the likability of speech stimuli on a scale of 1 to 7 (full marks).
2. The subject manipulates F0 and the spectral envelope to obtain the maximum likability. They can perform these manipulations until they accept the result.
3. The subject scores the likability of the converted speech as in the first step.

Since we aim to affect the likability of speech in daily life, we conducted the experiment in an ordinary room with an A-weighted SPL of 35 dB. Twenty persons with normal hearing ability participated in the evaluation. We used a set of headphones (SENNHEISER HD650) for the evaluation.

The speech stimuli used for the evaluations were 18 sentences spoken by three men and three women in a database for likability research[13]. The sentences consisted of Japanese text of around eight seconds, categorized in three likability levels (low, middle, and high). The sampling frequency and quantization were 96 kHz and 24 bit, respectively.

4.2 Results

Figures 3 and 4 show the experimental results. In Fig. 3, the vertical axis represents the baseline score, and the horizontal axis represents the average improvement compared with the baseline. The error bar represents the 95% confidence interval. Since a score of 7 is the highest in this evaluation, the improvement score always indicates zero. The results of the analysis indicated that there was no speech whose likability decreased as a result of the parameter conversion.

In Fig. 4, the vertical and horizontal axes represent the variation in F0 and the variation in the spectral envelope for each speech, respectively. Regarding the variation in F0, the likability scores of the female speech increased by converting it to high. This result supports the findings of previous studies[7]. However, the results for the male speakers were different; the optimal value of the F0 shift scaling depended on the speaker.

Regarding the spectral envelope, the likabilities of all speakers increased as a result of controlling the spectral centroid as high. This result did not depend on gender, and it showed that the perceived brightness affected the perceived likability. The optimal values of all speakers were positive, whereas there was a difference between genders.

5. Discussion

The results showed that the proposed algorithm can improve the likability of uttered speech. In this section, we discuss how the results of the subjective evaluation can be used to improve the likability of uttered speech by adjusting it parameters.

5.1 Discussion on the Variation in F0

The variation in obtaining the highest likability was around 35 mel for female speech. This result was similar to that of the previous research in which likability improved when F0 was increased. The previous research [13] showed that the speakers attempted to utter sentences with a higher pitch to improve its likability. The result of the current evaluation showed that the subject raised the pitch of the speech to improve its likability. This shows that female speech with good likability may be considered to be the same by speakers and
The variation in F0 in the male speech was different from that of the female speech. The research [13] showed that the speakers attempted to utter the sentences with a higher pitch to improve likability similar to what was done for female speech. However, the results of the current evaluation indicated that the subject did not necessarily convert the uttered speech to a higher pitch to improve likability. This shows that male speech considered to be likable by speakers is different from male speech perceived as by listeners.

5.2 Discussion on the Variation in the Spectral Envelope

Regarding the spectral envelope, the variation in female speech was around 50 Hz, and the variation in male speech was around 20 Hz. This result showed that the variations in the spectral envelope are converted to high regardless of gender; it also suggests that the brightness is related to perceived likability. On the other hand, the variation in female speech is higher than that of male speech. We hypothesized that this difference was due to the difference in the variation in F0 and examined the relationship between the variations in F0 and the spectral envelope.

Figure 5 shows the result. The vertical and horizontal axes represent the variation in F0 and the variation in the spectral envelope for each speech, respectively. The parameter $r$ represents the correlation coefficient between the variations in F0 and the spectral envelope. This result shows a significantly positive correlation between them ($p < 0.01$). In turn, it suggests that the variation in the spectral envelope was related to the variation in F0, but the optimum value was always positive.

6. Conclusion

We proposed a voice conversion algorithm for improving the likability of uttered speech. The algorithm can convert the spectral envelope by using one parameter related to the spectral centroid. A subjective evaluation was carried out to verify the effectiveness of this proposed algorithm. As a result, all subjects were able to appropriately set speech parameters to improve the likability of the speech. The result also showed that the perceived brightness was related to the perceived likability.

The next step of this research is to develop a likability measurement model from the waveform. This model would be used in speech training as an indicator of objective likability.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP19K127346 and JP16H05899.

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., vol.E98-D, no.8, pp.1536–1544, 2015.
[2] S.A. Collins, “Men’s voices and women’s choices,” Animal Behaviour, vol.60, no.6, pp.773–780, 2000.
[3] T. Holzheim, A. Schmitt, F. Metze, and M. Wagner, “Anger recognition in speech using acoustic and linguistic cue,” Speech Communication, vol.53, no.9-10, pp.1198–1209, 2011.
[4] K. Pisanski, A. Oleszkiewicz, J. Plachetka, M. Gmiterek, and D. Redy, “Voice pitch modulation in human mate choice,” Proc. RSB2018, vol.285, no.1893, pp.1–8, 2018.
[5] T. Uchida and N. Nakaune, “Effects of fundamental frequency and speech rate on impression formation,” The Japanese Journal of Psychology, vol.75, no.5, pp.397–406, 2004 (in Japanese).
[6] L.R. Goldberg, “An alternative “description of personality”: The big-five factor structure,” Journal of Personality and Social Psychology, vol.59, no.6, pp.1216–1229, 1990.
[7] 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,” Biology Letters, vol.4, no.2, pp.192–194, 2008.
[8] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Inf. & Syst., vol.E99-D, no.7, pp.1877–1884, July 2016.
[9] M. Morise, “D4C, A band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, vol.84, pp.57–65, 2016.
[10] S.S. Stevens, J. Volkman, and E.B. Newman, “A scale for measurement of the psychological magnitude pitch,” The Journal of the Acoustical Society of America, vol.8, no.3, pp.185–190, 1937.
[11] H. Ye and S. Young, “High quality voice morphing,” Proc. ICASSP, vol.1, pp.9–12, 2004.
[12] E. Schubert and J. Wolfe, “Does timbral brightness scale with frequency and spectral centroid?” Acta Acustica United with Acustica, vol.92, pp.820–825, 2006.
[13] M. Morise, F. Yokomori, and K. Ozawa, “Building of a database for likability evaluation of uttered speech,” Proc. UAC2018, 2 pages, 2018.