Real-Time Talking Avatar on the Internet Using Kinect and Voice Conversion

Takashi Nose
Graduate School of Engineering
Tohoku University
Sendai City, Japan

Yuki Igarashi
Graduate School of Engineering
Tohoku University
Sendai City, Japan

***Abstract*—We have more chances to communicate via the internet. We often use text/video chat, but there are some problems, such as a lack of communication and anonymity. In this paper, we propose and implement a real-time talking avatar, where we can communicate with each other by synchronizing character's voice and motion from ours while keeping anonymity by using a voice conversion technique. For the voice conversion, we improve accuracy of the voice conversion by specializing to the target character's voice. Finally, we conduct subjective experiments and show the possibility of a new style of communication on the internet.**

***Index Terms*—Talking avatar; Voice conversion; Kinect; Internet; Real-time communication**

I. Introduction

Typical examples for human communication via internet are text, voice, and video chatting. Users of the text chat input text with interfaces such as a keyboard and easily communicate to each other. However, the text-based communication has difficulty in expressing emotions and intentions correctly, which sometimes leads to misunderstanding of the user's internal state. In addition, the lack of the real-time communication is often stressful. On the other hand, the video chat has an advantage in communicating both linguistic and para-linguistic information through the facial expression and the speaking style [1], [2], [3]. Although the video chat is the most advanced and rich communication tool of the chat systems, one of the biggest problems in the use of the audio and video information is the lack of anonymity, and we must choose an appropriate tool depending on the situation.

In this paper, we present a real-time talking avatar system in which the user's motion and speech are reflected to the avatar in real-time. The system enables us to communicate to each other via the network without directly conveying the user's personal information. The system of the real-time talking avatar consists of the two technologies as follows:

- Voice conversion: Converting the speaker characteristics of the user to that of the avatar using neural network in real-time [4]
- Synchronization skeleton: Capturing the user's motion and reflecting the motion to the avatar in real-time using Kinect [5]

In the voice conversion, we focus on the character of the avatar, virtual singer *Hatsune Miku*. We conduct an experiment where the speech parameters extracted from the input speech is averaged using moving average to improve the reproducibility of the robotic voice of the character.

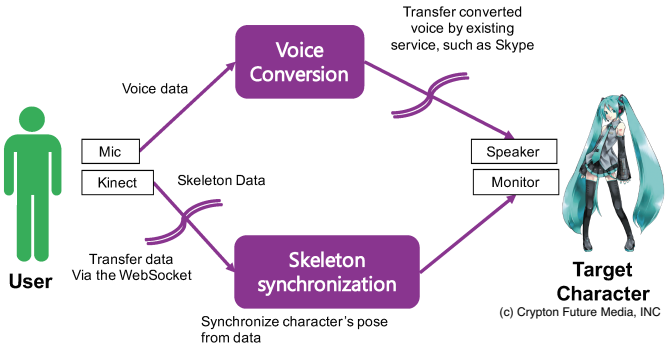

Fig. 1: Overview of the proposed real-time talking avatar system via internet.

We develop the real-time talking avatar system using the voice conversion and synchronization skeleton techniques, which enables the users to anonymize their voice and facial information. We conduct subjective evaluations and compare the proposed system to the conventional text and video chat systems in terms of the anonymity, entertainment, and communicability.

The rest of this paper is organized as follows: Section II overviews the real-time takling avatar system proposed in this paper. Section III introduces the voice conversion depending on the target character to improve the conversion performance based on neural networks, and the objective experimental evaluations for the voice conversion part is shown in Section IV. Section V explains how to control the character's motion using Kinect. The total subjective evalution is conducted in Section VI and the result is discussed. Finally, Section VII summarizes this study and shortly discusses the remaining issues as conclusions.

II. System overview

We use two sets of Kinect for Windows and microphone, and a PC to control them and network environment to realize the real-time communication using the proposed talking avatar.

The synchronization of the user's motion including hand gestures is achieved by acquiring the skeleton data of the user and by reflecting the data to the character model of the avatar. In this study, we use Kinect v2 and Skeletal Tracking of Kinect SDK v2 [6] to acquire the user's motion data. The Kinect is able to obtain the position data of twenty-five joints per one user.

To reflect the user's motion, only the rotation parameters of each joint and the position of center part of pelvis, which is given as SpineBase in Kinect SDK, are extracted and transmitted to the client user using a communication protocol, e.g., WebSocket. The client system receives the transmitted data and maps them to each joint and position of the character model. Finally, the avatar image having similar pose to the user is outputted to the display.

In the voice conversion, the speech of the user is recorded using a microphone, and the speech parameters, i.e., spectral and excitation parameters, are extracted and converted to those of the target character using a neural network that is one of the nonlinear mapping techniques. The speaker-adapted speech is obtained by synthesizing speech from the converted parameters. The speech data after the voice conversion is outputted through a virtual sound device using several windows APIs. By introducing the virtual device, we can use the converted voice as a source of existent voice conference applications such as Skype. Figure 1 shows the overview of the proposed system.

III. Character-dependent voice conversion

Voice conversion is a technique for changing the input speaker's voice characteristics to that of another speaker (target speaker) while keeping the linguistic information. In this study, the target speaker is not a person but an avatar, i.e., virtual singer Hatsune Miku as shown in Figure 2 to increase the entertainment factor in the communication.

A. Character's voice for avatar

We use a voice of a Japanese famous virtual character, Hatsune Miku of singing voice synthesis software *VOCALOID* [7] developed by YAMAHA [8]. Figure 2 shows an illustration of Hatsune Miku. Recently, the singing voice of Hatsune Miku is used for many users and is becoming much popular in the community cite such as Youtube and Niconico [9] in Japan [10]. By using VOCALOID, users can easily synthesize singing voice by inputting melody and lyrics. We prepare the parallel speech of human and Hatsune Miku using VOCALOID and use the synthesized speech for the training of the voice conversion. In this study, we chose the Miku as the target speaker to take the friendly feeling into account. Figure 3 shows an example of controlling character model of of Hatsune Miku using Kinect v2.

B. Voice conversion based on neural networks

Figure 4 shows an overview of the voice conversion part, In the voice conversion, we use neural networks to map the spectral features of the input speech of a certain user to those of the target character [4]. By using neural network, the multi-dimensional feature vectors consisting of the spectral parameters are mapped in a non-linear form. Since the relation of corresponding frames between source and target speakers is highly non-linear, the neural network is known to be effective compared to the traditional mapping technique based on Gaussian mixture model (GMM) [11], [12], [13] that has been widely used in the study of voice conversion [14]. The process of the voice conversion from the user's voice to the Hatsune Miku's voice is as follow:

- Prepare parallel speech data of two speakers uttering the same sentences.
- Extract mel-cepstral coefficients and fundamental frequency (F0) data using Speech Signal Processing Toolkit (SPTK) [15]
- Perform dynamic time warping [16] to align the frames of the spectral features of the two speakers.
- Train neural networks that maps source speaker's features to those of the target speaker and obtain a set of weight

Fig. 2: Virtual character Hatsune Miku of a singing voice synthesizer VOCALOID.

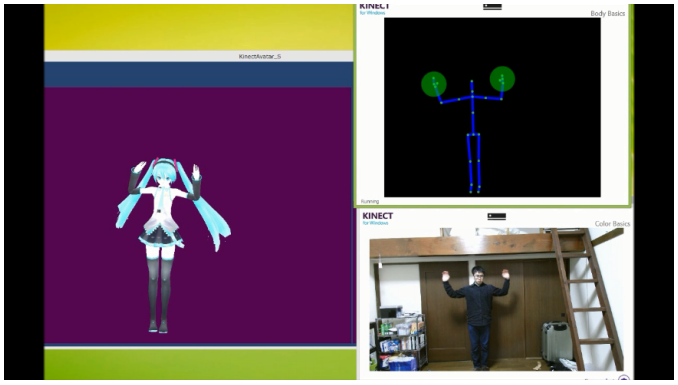

Fig. 3: Example of controlling character model of Hatsune Miku using Kinect v2.

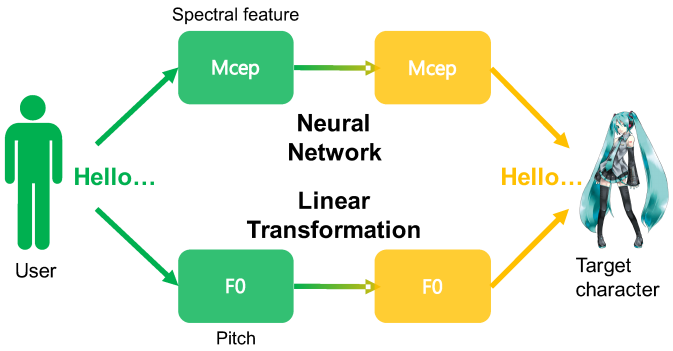

Fig. 4: Overview of the voice conversion part using neural networks for spectral mapping and linear transformation for pitch conversion.

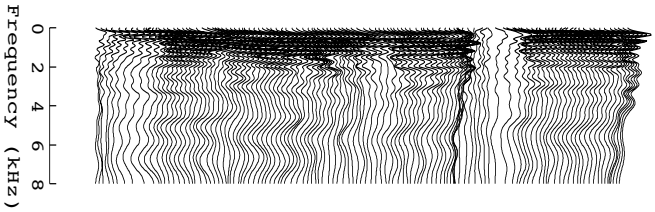

Fig. 5: Running spectrum extracted from the speech /yobou/ uttered by Hatsune Miku in Japanese.

parameters.

- Convert the spectral features of the input speech of the user using the trained neural networks.
- Convert the log F0 parameter using affine transformation so that the mean and variance parameters become the same between the two speakers.
- Synthesize speech from spectral and F0 parameters.

For the F0 conversion, we use affine transformation defined by

$$y = (x - \mu_x)/\sigma_x * \sigma_y + \mu_y \quad (1)$$

where μ_x and σ_x are global mean and variance of the source speaker, respectively, and μ_y and σ_y are global mean and variance of the target speaker, respectively.

Figures 5 and 6 show examples of the running spectrum and the F0 sequence of synthetic speech of Hatsune Miku. As is shown in the figure, the trajectory of the spectral envelope and log F0 is smooth, which is different from those of the human speech. To utilize this property, we apply smoothing filter by moving average for the spectral and log F0 parameter sequences after the voice conversion process.

IV. Voice Conversion Experiments

A. Experimental conditions

For the experiments, we used a hundred sentences, i.e., subsets A and B of the ATR phonetically-balanced Japanese sentences [17]. A male non-professional speaker uttered the sentences. We used fifty sentences of the subset A for the

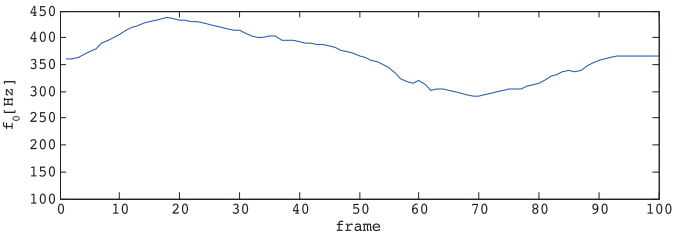

Fig. 6: Log F0 sequence extracted from the speech /yobou/ uttered by Hatsune Miku in Japanese.

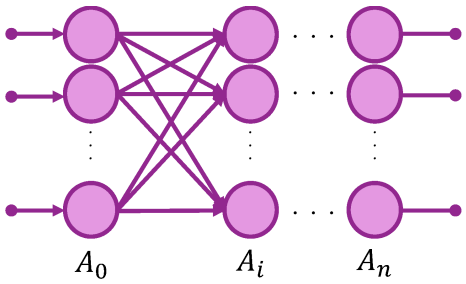

Fig. 7: Example of neural network.

training of the neural networks in the voice conversion and used fifty sentences of the subset B for the testing. As for the structure of the neural network, we fixed the number of the hidden layers to one to achieve the real-time processing of voice conversion though we might improve the performance by increasing the number of the layers under the condition where we have a sufficient amount of training data. From a preliminary experiment, we set the number of units of the hidden layer to fifty. The numbers of units for input, hidden, and output layers were 25, 50, 25, respectively.

The 0th to 24th mel-cepstral coefficients were extracted using mcep command of SPTK where we set the window length to 25 ms and the frame shift to 5 ms. The log F0 was extracted using pitch command with the RAPT algorithm [18] with the same frame shift. To create the parallel speech of Hatsune Miku, we generated log F0 sequences using Open JTalk [19]. The log F0 sequences were then quantized into several levels with a semitone interval for each mora and the Miku's voice corresponding to the training text was generated using VOCALOID. As a result, we created the Miku's speech data of subset A of the ATR sentences.

B. Results and analysis

We used mel-cepstral distance as the objective measure of spectral reproducibility. The mel-cepstral distance of the n-th frame is defined as

$$d(n) = \sum_{k=1}^{M} (c_n^{(\mathrm{t})}(k) - c_n^{(\mathrm{s})}(k) \quad (2)$$

where $c_n^{(\mathrm{t})}(k)$ and $c_n^{(\mathrm{s})}(k)$ is the k-th mel-cepstral coefficient of n-th frame of source and target speakers, respectively. We calculated the mel-cepstral distance between the original

TABLE I: Average mel-cepstral distance between source and target speakers before and after the voice conversion.

	Average (dB)	Min. (dB)	Max. (dB)
Before	16.025	14.312	17.319
After	9.356	8.251	10.368

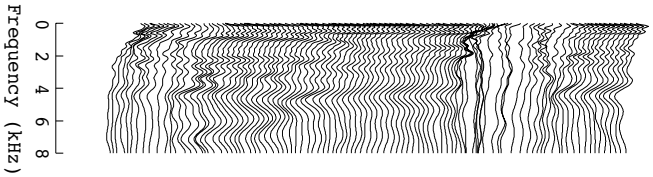

Fig. 8: Running spectrum before voice conversion.

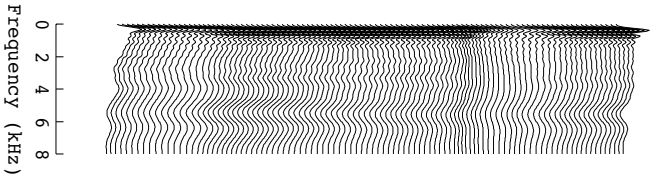

Fig. 9: Running spectrum after voice conversion.

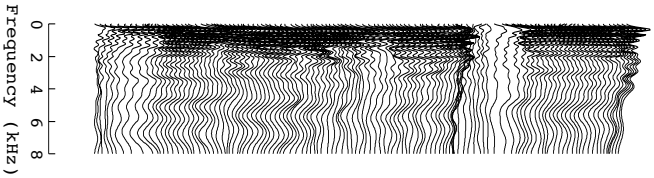

Fig. 10: Running spectrum of original speech of the target speaker (Hatsune Miku).

speech of the target speaker and the converted speech. For comparison, we also calculated the mel-cepstral distance of speech between the source (before conversion) and the target speakers. Table I shows the average, minimum, and maximum values of the mel-cepstral distance for the test data.

From the table, it is seen that there was large spectral difference between the source and the target speaker but the difference became smaller after the voice conversion with spectral mapping based on neural networks. Figures 8 and 9 show examples of running spectra before and after the voice conversion. For the reference, we also show the running spectrum of the original speech in Figure 10.

To evaluate the effect of smoothing after the voice conversion, we also calculated the mel-cepstral distance between converted speech and the target speaker's speech by changing the width for the moving average. Table II show the result. From the table, we found that the mel-cepstral distance decreases by introducing moving average is smallest when the number of frames is six. This result indicates that the smoothing operation by moving average well captures the voice property of the target character.

Next, we evaluated the effect of the smoothing by moving average for the log F0 sequence with subjective evaluation. We conducted a listening test where subjects listened to the converted speech samples and rated the speaker similarity using five-point scale: 5 (similar), 4 (a little similar), 3 (undecided), 2 (a little dissimilar), and 1 (dissimilar). Table III shows the result.

TABLE II: Width (# of frames) in moving average and mel-cepstral distance.

# of frames	Average (dB)	Min. (dB)	Max. (dB)
0	9.356	8.251	10.368
1	9.288	8.225	10.303
2	9.316	8.273	10.357
3	9.228	**8.182**	10.332
4	9.199	8.201	10.215
5	9.202	8.271	10.216
6	**9.179**	8.316	**10.168**
7	9.189	8.295	10.345
8	9.211	8.382	10.347
9	9.274	8.412	10.406
10	9.271	8.506	10.359

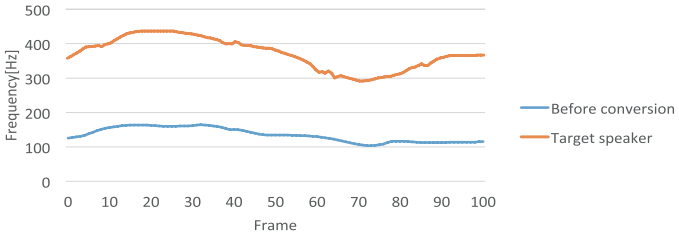

Fig. 11: F0 contours of the source speaker (before conversion) and the target speaker.

From the table, we found that the speaker similarity substantially degrades when the F0 conversion is not applied. Figures 11 and 12 show examples of F0 contours before and after the F0 conversion, respectively. The original F0 contour of the target speaker is also shown in the figures. From the figures, it is seen that the F0 contour of the source speaker became closer to that of the target speaker even when the target speaker is a virtual character, Hatsune Miku.

In contrast to the case of spectral features, the smoothing was not effective in the case of F0, and over-smoothing also degrades the speaker similarity. A possible reason of the degradation is that the F0 smoothing by moving average does not take the mora units and their average pitch into account. Since the speech of Hatsune Miku was synthesized using VOCALOID, the F0 became flat within each mora. To improve the conversion performance, we explicitly use the mora information for the input speech by forced alignment technique. However, it is not easy to use the alignment information in the real-time application. Hence, this is a remaining problem in this study.

V. Synchronization to the Character's Motion

In this study, we need to acquire skeleton information of the user to synchronize the motion of the user to that of the character model. We use Kinect for Windows v2 that is a

TABLE III: Subjective evaluation results comparing converted speech with moving average of log F0 to the converted speech without moving average.

# of frames	Similar	A little similar	Undecided	A little dissimilar	Dissimilar
No conversion	0	0	0	0	**5**
0	1	**3**	1	0	0
5	0	**4**	1	0	0
10	0	0	**4**	1	0
15	0	0	0	2	**3**
20	0	0	0	1	**4**

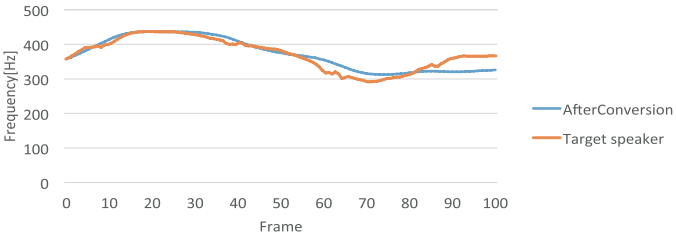

Fig. 12: F0 contours of the source speaker (after conversion) and the target speaker.

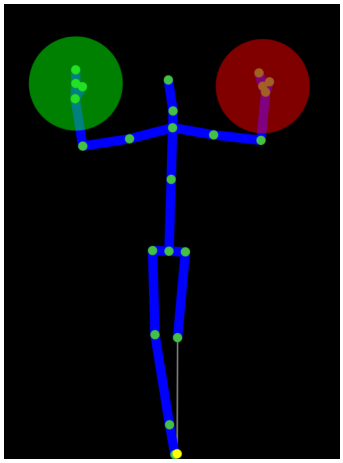

Fig. 13: Example of skeleton detection by Kinect v2.

motion sensor device developed by Microsoft corporation. By using this sensor device, we can obtain the information of twenty-five joints per one person, and we can also obtain the position and rotation information of each joint. Kinect for Windows v2 has better performance than the conventional Kinect sensor in terms of the accuracy and resolution, and the face recognition and facial expression analysis are supported. However, we only utilize the function of the synchronization skeleton in this study.

The synchronization skeleton is realized by acquiring rotation information of each joint of the user, calculating the rotation between the joints, and synchronizing with the rotation of each joint of the character model. By applying these processing, even when there is a large difference of body size and body characteristics between the user and the target character,

Fig. 14: Example of controlling MMD model of Hatsune Miku by Kinect v2.

mapping is not sensitive to the difference, and natural mapping is achieved. In addition, to deal with the longitudinal and lateral movement, first we set the position of the calibration to the origin and calculate the relative position of the pelvis of the center, and apply the position to the character model. This enables the system to display the character as if the character was walking to the direction corresponding to the user's walking motion to the same direction.

We used Unity [20], which is an environment for the programming easily handling the character model, for the implementation of the above functions. For the character model, we used Lat-style Miku Ver.2.31 [21] that is a 3D CG model of Hatsune Miku for MikuMikuDance (MMD) [22]. Figures 13 and 14 show an example of the mapping from the user's skeleton information to the character's motion.

VI. Demonstration Experiments of Real-Time Talking Avatar

On the basis of the experimental results in the previous sections, we implemented a real-time talking avatar system. To demonstrate the effectiveness of the system, we conducted an experiment where five subjects used the system for the communication to a reference person.

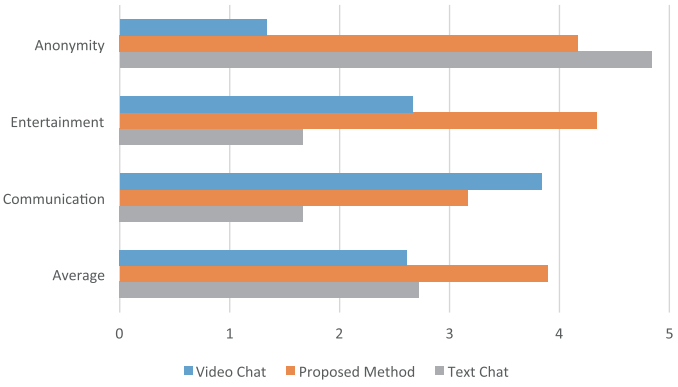

Fig. 15: Comparison of the effectiveness of the proposed real-time talking avatar to the other communication systems.

Fig. 16: Example of the conversation using the real-time talking avatar system.

A. Experimental procedure

Users were naive and did not know the conversation partner, and the communication through the system was performed in the following order.

1) Text chat
2) Real-time talking avatar (the proposed system)
3) Video chat

Users evaluated the three systems in terms of the following three criteria using five-point scale: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), 1 (Bad).

- Anonymity: Whether the user directly recognizes the partner through the communication
- Entertainment: How the user enjoys the communication
- Communication: How the conversation with the system continues smoothly

The real-time talking avatar is proved to be effective if the proposed system outperforms the video chat in terms of Anonymity and outperforms the text chat in terms of Entertainment and Communication criteria.

B. Experimental results

Figure 15 shows the results. From the aspect of Anonymity, the proposed system is slightly worse than text chat but is substantially better than the video chat. In the experiment, we found that the user sometimes obtained the information about the partner's personality and gender through the gesture and habit of the user even though the motion and the voice were conveyed through the character model. However, it is difficult to completely separate the information related to the motion and personality, and hence the slight degradation of anonymity would be acceptable in the real communication of the most of users. As for the degree of the entertainment, the proposed system gave higher score than both the text and video chat systems. One of the reasons is that the attractive target character instead of a real person enhances the pleasantness in the conversation, which is the main purpose of this work. The advantage of our system is that the user can choose the target character so as to be fun for him/her. In the factor of Communication, the real-time conversation of the proposed system made the score much higher than the text chat and the score was close to that of the video chat. The communication performance would be improved by the advance of Kinect and motion and face tracking SDKs.

The above results are summarized as follows. The proposed real-time talking avatar system has the intermediate property between the text chat and the video chat, and the average score is highest of three systems, which indicates that our system is the more balanced and attractive system than the conventional text and video chat systems.

VII. Conclusions

In this paper, we presented a novel communication tool via internet, where the communication style is different from the conventional text and video chat systems. The most attractive point of our system is that both of the anonymity and entertainment factors are achieved at a sufficient level while keeping the smoothness of the communication in real-time. The system utilizes Kinect-based motion capturing and processing and a voice conversion technique, and the user can choose the favorite character and voice to anonymize him/herself. In the voice conversion part, the property of the target speaker was taken into account, and we showed that the smoothing operation using moving average increase the the spectral reproducibility when the width was appropriately set.

In the future work, the performance improvement in the voice conversion is important. Especially, the noise robustness is highly required in the real environment in our daily life. Also the use of facial information including emotional expressions is beneficial for more advanced human communication with high anonymity and security.

Acknowledgment

Part of this work was supported by JSPS Grant-in-Aid for Scientific Research 15H02720 and Step-QI School of Tohoku University.

REFERENCES

[1] C. Neustaedter and S. Greenberg, "Intimacy in long-distance relationships over video chat," in *Proc. the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 753–762.
[2] O. Boyaci, A. G. Forte, and H. Schulzrinne, "Performance of video-chat applications under congestion," in *Proc. 11th IEEE international symposium on multimedia*, 2009, pp. 213–218.
[3] J. Scholl, P. Parnes, J. D. McCarthy, and A. Sasse, "Designing a large-scale video chat application," in *Proc. the 13th annual ACM international conference on Multimedia*, 2005, pp. 71–80.
[4] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. ICASSP*. IEEE, 2009, pp. 3893–3896.
[5] Microsoft Kinect for Windows, http://www.microsoft.com/en-us/kinectforwindows/.
[6] Tracking Users with Kinect Skeletal Tracking, https://msdn.microsoft.com/ja-jp/library/jj131025.aspx.
[7] H. Kenmochi and H. Ohshita, "Vocaloid–commercial singing synthesizer based on sample concatenation," pp. 4011–4010, 2007.
[8] YAMAHA Corporation, http://www.yamaha.com/.
[9] Niconico, http://www.nicovideo.jp/.
[10] M. Hamasaki, H. Takeda, and T. Nishimura, "Network analysis of massively collaborative creation of multimedia contents: case study of hatsune miku videos on nico nico douga," in *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*, 2008, pp. 165–168.
[11] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.
[12] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, vol. 1, 1998, pp. 285–288.
[13] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.
[14] Y. Stylianou, "Voice transformation: a survey," pp. 3585–3588, 2009.
[15] The SPTK working group, "Speech Signal Processing Toolkit (SPTK)," http://sp-tk.sourceforge.net/ (2015.9.24).
[16] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.
[17] 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 Communication*, vol. 9, no. 4, pp. 357–363, 1990.
[18] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, pp. 495–518, 1995.
[19] Open JTalk, http://open-jtalk.sourceforge.net/.
[20] Unity, http://unity3d.com/.
[21] Lat-style Miku Ver.2.31, https://bowlroll.net/file/30199.
[22] T. Yoshikawa, "Miku miku dance starter pack," 2010.