Personalized Voice Assignment Techniques for Synchronized Scenario Speech Output in Entertainment Systems

Shin-ichi Kawamoto[1,3], Tatsuo Yotsukura[2], Satoshi Nakamura[3], and Shigeo Morishima[4]

[1] Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa, 923-1211, Japan
kawamoto@jaist.ac.jp
[2] OLM Digital Inc.,
Mikami Bldg. 2F, 1-18-10 Wakabayashi, Setagaya-ku, Tokyo, 154-0023, Japan
yotsu@olm.co.jp
[3] National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
satoshi.nakamura@nict.go.jp
[4] Waseda University,
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan
shigeo@waseda.jp

Abstract. The paper describes voice assignment techniques for synchronized scenario speech output in an instant casting movie system that enables anyone to be a movie star using his or her own voice and face. Two prototype systems were implemented, and both systems worked well for various participants, ranging from children to the elderly.

Keywords: Instant casting movie system, post-recording, speaker similarity, voice morphing, synchronized speech output.

1 Introduction

Instant casting movie system (ICS) is an entertainment system that allows participants to instantly appear in a movie as a CG character [1]. In ICS, all processes are performed automatically (Fig. 1): scanning of the face shape and image, reconstruction of the 3D face model, and generation of onscreen appearance and movement. However, assigned voices of the participants' CG characters were only switched pre-recorded male or female voices, depending on the gender of the participant. Our goal is to extend the functionality of voice assignment for the ICS. The participant's CG character should be assigned a voice that is matched and similar to the participant's own voice. In addition, the voice quality should be better matched to the CG movie quality. Since the CG character has various facial expressions, the voice should communicate various expressions of emotions. Various speech synthesis techniques have been proposed [2-5]. However, it is difficult to generate high quality voices that match the various scenes.

R. Shumaker (Ed.): Virtual and Mixed Reality, Part II, HCII 2011, LNCS 6774, pp. 177–186, 2011.

Excessive deformation of the speech waveform is especially difficult without degrading voice quality.

In this paper, we introduce three approaches to extend the voice assignment function: quick post-recording, similar speaker selection and voice morphing. For short scenarios, participants can record scenario speeches directly using our post-recording tool. For long scenarios, participants can participate in the movie easily using the combined approaches of similar speaker selection and voice morphing. We have developed prototype speech synchronization systems. Scenario speeches can be output synchronously along with the CG character using our systems.

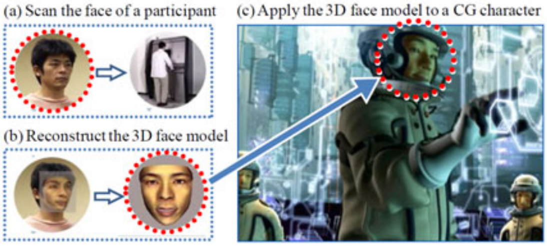

(a) Scan the face of a participant
(b) Reconstruct the 3D face model
(c) Apply the 3D face model to a CG character

Fig. 1. Instant casting movie system (ICS)

2 Quick Post-recording Tool

If the most significant feature of voice casting is that the CG character acts as the participant with the participant's voice, one of the best solutions is to directly record the scenario voice from the participants. However, it is difficult for participants to record their own voices for synchronization with the movie. This recording task, called post-recording, is widely used in movie and animation production. The accuracy of synchronization between voice and acting depends on the skill of voice actors. Adjusting the timing of the scenario speech takes a long time for non-professional voice actors. In addition, these post-recording environments for professional use are expensive and complicated, and must be manually operated by skillful sound engineers. Considering the cooperative work of ICS, the following system requirements of the post-recording tool must be provided: 1) various timings of post-recording information; 2) automatic post-processing function of recorded voices.

2.1 Designing the Timing Information of Post-recordings

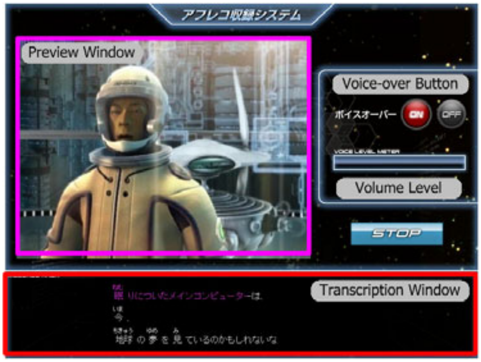
An important system requirement of a post-recording system is to produce scenario speech synchronously with the character's action in the movie. To fulfill this requirement, our tool provides users six types of timing information: movie, BGM/SE, time code, reference voice for voice-over acting, colored text like karaoke, and rhythmic information. Rhythmic information is generated by repetitively displaying the target scene over a constant period. Our tool can also be selected with/without a reference voice for voice-over acting, since the reference voice sometimes influences the user's acting.

2.2 Automatic Post-processing

To use our post-recording tool in ICS, participant's voices should be inserted into a movie soon after they are recorded. However the post-processing of recorded voices is time-consuming, especially for sound engineers who manually extract the essential parts of recorded voices to be synchronized with the corresponding movie shots. Thus an automatic post-processing mechanism for recorded voices should be implemented to eliminate such time-consuming work.

2.3 Prototype of Post-recording Tool

We developed a post-recording tool that fulfills the requirements. Fig. 2 shows a screenshot of our tool. Participants operate this tool with a touch-screen. The user speaks the script, which is synchronized with the video images. This tool has the following functions for supporting voice and lip synchronization: displaying the time code and the transcript text, and voice-over enable/disable buttons. The voice volume of user utterances can be checked using a meter. The video images for the post recording are repeatedly played by pushing the "OK" button to send a signal that the user felt a sentence recorded correctly. In our tool for ICS, manual post-processing is not needed, since our tool automatically extracts and saves the user voices that fit the length of the target scenes.



眠りについたメインコンピューターは、
今、
地球の夢を見ているのかもしれないな

Fig. 2. Screenshot of post-recording tool

2.4 Evaluation of Post-recording Tool

In this section, we evaluate our prototype system in terms of three points: stability of real world use, ease of operation, and shortening of voice acquisition time.

Stability. We exhibited our system from February 10th to 11th, 2008 at the Miraikan Museum in Tokyo. The system worked well with 162 participants (including 79 children) for post-recording. While the children needed the support of a parent or an attendant to use our system, many participants could easily understand and operate it.

Ease of operation. We ran a subjective test to evaluate our system. 99 subjects between the ages of 18 and 63 tested it and answered some questions. To evaluate ease of operation, we asked the following question: "*Was the operation of this post-recording system easy?*" As a result. 64% of the subjects responded that our system's interface could be operated easily (Very Easy=13%, Easy=51%, Neither=11%, Difficult=11%, Very Difficult=23%).

Acquisition time. To evaluate the efficiency of data acquisition, we measured operation time using our system. For comparison, we measured the manual recording operation time using popular sound recording equipment with an operator. In manual recording, the subject speaks the lines of the script while watching the movie and the scenario text, and listening to the BGM/SE and the reference voice. After this manual recording, the operator must manually split the recorded voice using sound editing software. In manual recording, the average recording and splitting operation times of the eleven subjects were 64 minutes and 69 minutes, respectively. On the other hand, using our system, the recording operation time averaged 77 minutes with 91 subjects. In our system, the voice splitting operation time includes the recording time. Our system reduced the total recording time by 42% in comparison with manual recording. In addition, the number of professional operators can also be reduced using our system, since the subjects themselves operated our system in this evaluation of operating time.

3 Selecting Similar Speakers

The similar speaker selection process selects from the voice actor DB an actor with a voice similar to that of the participant, and assigns the selected voice to the character. In this approach, the voice quality does not undergo any degradation. However, the possibilities with regard to the extent of voice similarity that can be achieved are limited to the size of the voice actor DB. Various acoustic features related to the perception of speaker similarity were reported independently [6-9]. However the personality of a speaker is evinced not only in the voice quality but also in the prosodic intonation. Therefore, we need to consider multiple acoustic features for various voice characteristics. The key technology of our method is to combine multiple acoustic features that are used to calculate perceptual similarity into our system implementation so as to realize within the ICS, a closely-matched voice for each participant's character.

3.1 Estimation Method

We estimate the perceived similarity of speakers by participants using a combination of multiple acoustic features for greater accuracy. The perceptual similarity estimate s is calculated using Equation 1.

$$s = -\sum_{i=1}^{n} \alpha_i x_i \tag{1}$$

In this equation, n is the number of acoustic features, x_i is the distance of the ith acoustic feature between the voices, and α is the weighting coefficient for each such distance between acoustic features.

We use 8 acoustic features related to voice personality. These are the Mel Frequency Cepstral Coefficient (Static: 12 + Dynamic: 13 = 25 dimensions)[6], the STRAIGHT Cepstrum of over 35 dimensions and 1st dimension [7], the Spectrum of over 2.6 kHz, the STRAIGHT-Ap under 2 kHz [8] that is a parameter of STRAIGHT [10], the fundamental frequency, the formants (F1 - F4), and the spectrum slope between 0 kHz - 3 kHz [9]. To extract these acoustic features, we use a window length of 25 ms and the shift rate of 10ms.

We calculate the distance between the acoustic features with Dynamic Time Warping (DTW). DTW distance is commonly used in a wide range of pattern recognition systems. It can estimate perceptual similarities accurately because it represents the temporal structure of acoustic features.

3.2 Optimization of Weighting Coefficients

To increase the correlation between the perceived similarity as reported by the subjects and that estimated using our method, we optimize the weighting coefficient α in Equation 1. To select a target speaker from a speaker DB, we represent the perceived similarities of the other speakers to the target by ranking the speakers in a permutation. The ranking is determined by quick sort based on subjective judgment. A subject judges similarity considering various speech features. Then the weighting coefficients α are optimized using the steepest descent method to increase the Spearman's rank correlation coefficient between the ranking of perceived similarity and that of acoustic similarity. Acoustic similarity is calculated using Equation 1. Spearman's rank correlation is shown in Equation 2.

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(\beta_i - \gamma_i)^2}{N^3 - N} \tag{2}$$

In this equation, β is the ranking with respect to perceived similarity ascribed by the subject. γ is the acoustic similarity ranking derived by using our method. N is the number of units of speech data. For this optimization, we used speech data uttered by 36 speakers.

4 Voice Morphing

The Voice morphing approach is based on blending a few voices to generate a voice similar to that of the participant. Our approach is based on STRAIGHT [10] voice morphing, which is an extra-high-quality voice morphing technique [11]. The key technology in our approach enables the automatic estimation of the optimal blending weights required to generate a voice similar to that of the participant.

4.1 Two Speakers' Voice Morphing

The basic idea of voice morphing is to generate an intermediate voice from two source voices by using an arbitrary blending ratio [12]. STRAIGHT-based morphing [11] handles the feature vectors of time-frequency representation derived by STRAIGHT [10], STRAIGHT spectrogram, Aperiodicity Map and F0. Time-frequency transformation of each feature vector is represented as a simple piecewise bilinear transformation with the same blending ratio.

4.2 Multiple Speakers' Voice Morphing

Takahashi et al. extended a conventional STRAIGHT-based morphing system to a multiple-speaker morphing mechanism [13]. The procedure is almost the same as that of conventional STRAIGHT-based morphing; it involves 1) anchor points, characteristic corresponding points in the time-frequency domain that are manually assigned on each reference spectrogram; 2) time-frequency transformation, which is derived from target and reference feature vectors based on the anchor points; and 3) reference feature vectors, which are morphed to mapped target feature vectors $\boldsymbol{x}_{mrp}$ based on Equation 3 with a blending ratio vector $\boldsymbol{r}$.

$$x_{mrp} = \sum_{s=1}^{S} r_s \boldsymbol{x}_s \tag{3}$$

where r_s is the blending ratio for speaker s, and $\boldsymbol{x}_s$ are the feature vectors for speaker s.

4.3 Voice Morphing for Generating Specific Speakers

We introduce our technique which can estimate a blending ratio vector to generate a specific speaker's voice based on multiple-speaker morphing. In this paper, we estimate a blending ratio vector that satisfies the following formula.

$$\hat{\mathbf{r}} = \arg\min_{r} \left\| \mathbf{y} - \hat{\mathbf{x}}_{mrp} \right\|^2 = \arg\min_{r} \left\| \mathbf{y} - \sum_{s=1}^{S} r_s \mathbf{x}_s \right\|^2 \tag{4}$$

$$\mathbf{r} = \left[r_1, r_2, \ldots, r_S \right]^{\mathrm{T}} \tag{5}$$

where $\mathbf{y}$ is the feature vector of target speaker, and $\hat{\mathbf{r}}$ is the estimated blending ratio vector. In this method, the blending ratio vector is minimized to yield the following formula.

$$e(\mathbf{r}) = \sum_{\tilde{f}=1}^{F} \sum_{t=1}^{\tilde{T}(\mathbf{r})} \left(y_{\tilde{t}}(f) - x_{\tilde{t}}(\tilde{f})\mathbf{r} \right)^2 \tag{6}$$

where $y_{\tilde{t}}(f)$ is the feature vector of a target speaker with a regularized time domain, $x_{\tilde{t}}(\tilde{f}) = \left[x_{\tilde{t}}^{(1)}(\tilde{f}), x_{\tilde{t}}^{(2)}(\tilde{f}), \ldots, x_{\tilde{t}}^{(S)}(\tilde{f})\right]$ are feature vectors for reference speakers, and S is the number of reference speakers. $\tilde{f}$ and $\tilde{t}$ refer to time and frequency domain elements, respectively, regularized by anchor points and a blending ratio vector $\hat{\mathbf{r}}$. $\tilde{T}(\hat{\mathbf{r}})$ is the regularized speech duration in the time domain, as shown below.

$$\tilde{T}(\hat{\mathbf{r}}) = \sum_{s=1}^{S} \hat{\mathbf{r}}_s T_s \tag{7}$$

Since the STRAIGHT-based voice morphing process controls various features by the same blending ratio vector in the time-frequency domain, it is difficult to solve Equation 6 analytically. Therefore, we use an iterative approach to solve for a blending ratio vector using the following formulae.

$$\hat{\mathbf{r}}_{n+1} = \hat{\mathbf{r}}_n - \alpha E_n \tag{8}$$

$$E_n = \left(\frac{\partial^2 e'(\hat{\mathbf{r}}_n)}{\partial \hat{\mathbf{r}}_n{}^2} \right)^{-1} \frac{\partial e'(\hat{\mathbf{r}}_n)}{\partial \hat{\mathbf{r}}_n} = \left(\overline{\mathbf{X}}_n{}^T \overline{\mathbf{X}}_n\right)^{-1} \overline{\mathbf{X}}_n{}^T \left(\overline{\mathbf{Y}}_n - \overline{\mathbf{X}}_n \hat{\mathbf{r}}_n\right) \tag{9}$$

where $e'(\hat{\mathbf{r}})$ is an approximation formula of Equation 6 that assumes the blending ratio $\hat{\mathbf{r}}_n$ is constant. $\overline{\mathbf{X}} = \left[\overline{\mathbf{X}}_1{}^T, \overline{\mathbf{X}}_2{}^T, \ldots, \overline{\mathbf{X}}_{\tilde{T}(\hat{\mathbf{r}}_n)}^T\right]^T$ are regularized feature vectors in the time-frequency domain that are updated n times by the blending ratio vector $\hat{\mathbf{r}}_n$. $\overline{\mathbf{Y}} = \left[\mathbf{y}_1{}^T, \mathbf{y}_2{}^T, \ldots, \mathbf{y}_{\tilde{T}(\hat{\mathbf{r}}_n)}^T\right]^T$ are regularized feature vectors in the time domain that are updated by a blending ratio vector $\hat{\mathbf{r}}_n$. $X_{\tilde{t}_n}$, $\mathbf{x}_{\tilde{t}_n}^{(s)}$, and $\mathbf{y}_{\tilde{t}_n}$ are defined as follows.

$$\mathbf{X}_{\tilde{t}_n} = \left[\mathbf{x}_{\tilde{t}_n}^{(1)}, \mathbf{x}_{\tilde{t}_n}^{(2)}, \ldots, \mathbf{x}_{\tilde{t}_n}^{(S)}\right] \tag{10}$$

$$\mathbf{x}_{\tilde{t}_n}^{(s)} = \left[x_{\tilde{t}_n}^{(s)}(1), x_{\tilde{t}_n}^{(s)}(2), \ldots, x_{\tilde{t}_n}^{(s)}(\tilde{f}_n), \ldots, x_{\tilde{t}_n}^{(s)}(F)\right]^T \tag{11}$$

$$\mathbf{y}_{\tilde{t}_n} = \left[y_{\tilde{t}_n}(1), y_{\tilde{t}_n}(2), \ldots, y_{\tilde{t}_n}(\tilde{f}_n), \ldots, y_{\tilde{t}_n}(F)\right]^T \tag{12}$$

In this paper, we adopt α as 1, the number of iterations as 20, and the number of reference speakers as 8.

5 Implementation of Synchronized Speech Output

Fig. 3 shows our prototype system for ICS. Participants record their own speech using recording PCs. Input is a voice speaking a sentence which was recorded by each participant. Similar Voice Selection Servers and Voice Morphing Servers calculate the results of scenario speeches that are intended to be similar to the participants' voices. We have recorded 60 different kinds of voices to construct the voice DB of this system. This DB covers a wide range of participants, in terms of age (subjective ages: 6-63). In addition, this DB has a balanced male-female ratio (subjective gender: male=29, female=31). These scenario speeches are played based on a Longitudinal Time Code (LTC) that represents time synchronization with video images. The Display PC outputs video images and stereo audio, which consists of the LTC and the recorded sound (a mixture of BGM and SE). The audio composite PC sends speech data to the mixer based on the LTC. The mixture of speech data and sound is sent to the audio speakers. The image composite PC also sends video data to the display based on the LTC. As for the prototype system's movie content, we used “Grand Odyssey,” which was exhibited at the 2005 World Exposition in Aichi, Japan. Our system can be used easily and quickly by various participants-ranging from children to the elderly-because it is based on participants having to record one specific sentence only. We exhibited our system on March 20-22, 2009, at the Miraikan Museum in Tokyo. The system was tested by over 100 participants, including children and elderly people, and was found to work well.

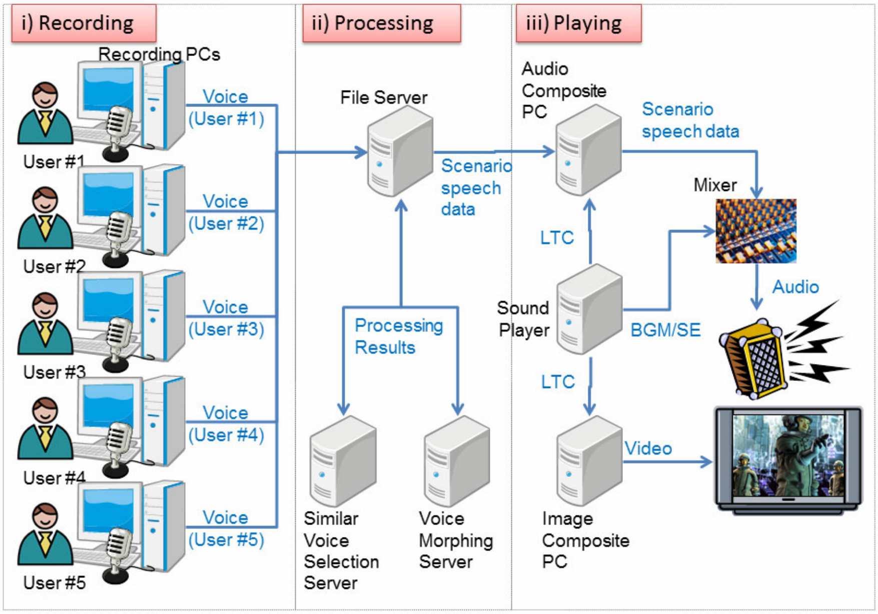


Fig. 3. Overview of prototype system

6 Discussion and Conclusion

In this paper, we described two types of scenario speech assignment systems to extend the casting functions of ICS. Our system worked well for various participants. Since synchronization of these scenario speeches was based on LTC, all three approaches can be combined into the same system.

The voice post-recording tool can be operated intuitively with various timing information. In this system, participants feel that they participate directly in the movie production, since participants uttered their scenario speech directly. However, the post-recording task is time-consuming, if the scenario speech is too long. In our post-recording strategy, scenario speech should be kept short to avoid overload of participants' tasks. In addition, it is difficult for most people to utter exaggerated expressions matched to the target scenes. Thus, the quality of movie output depends on participants. One of our future target projects is to modify the expression style of recorded voices automatically to match the target scenes.

The combined approach of similar voice selection and voice morphing also worked well by recording an input sentence for each participant. In this approach, workload of participants is less than in the post-recording approach, since participants only recorded a read speech, without exaggerated expression. Quality of output movie is stable, since scenario speech is based on voice DB, which was recorded by professional voice actors preliminarily. However, similarity of output voice depends on the size of voice DB. It is important to establish an archetype for the design strategy for the construction of a voice actor DB. Constructing a voice actor DB is a time- and money-intensive task. In voice morphing, assigning anchor points is also time-consuming. In order to improve the efficiency of these tasks, our system needs to incorporate other features. In addition, the use of the voice morphing technique leads to a slight degeneration in the speech quality of output voices. At present, we check the voice quality of outputs manually. One important course of future work is to develop an automatic speech quality evaluation technique. This technology will reduce operational costs of our system.

Acknowledgments. This research is supported by the Special Coordination Funds for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology.

References

1. Maejima, A., Wemler, S., Machida, T., Takebayahashi, M., Morishima, S.: Instant Casting Movie Theater: The Future Cast System. The IEICE Transactions on Information and Systems E91-D(4), 1135–1148 (2008)
2. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: The HMM-based speech synthesis system version 2.0. In: Proc. of ISCA SSW6, Bonn, Germany (2007)
3. Kawai, H., Toda, T., Yamagishi, J., Hirai, T., Ni, J., Nishizawa, N., Tsuzaki, M., Tokuda, K.: XIMERA: A Concatenative Speech Synthesis System with Large Scale Corpora. IEICE Trans. J89-D-II(12), 2688–2698 (2006)

4. Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. of ICASSP, pp. 373–376 (1996)
5. Clark, R.A.K., Richmond, K., King, S.: Multisyn: Open-domain unit selection for the Festival speech synthesis system. Speech Communication 49(4), 317–330 (2007)
6. Reynolds, D.: Robust text-independent speaker identication using gaussian mixture speaker models. IEEE Trans. On Acoust. Speech and Audio Processing 3(1) (1995)
7. Kitamura, T., Saitou, T.: Contribution of acoustic features of sustained vowels on perception of speaker characteristic. In: Proc. of Acoustical Society of Japan 2007 Spring Meeting, pp. 443–444 (2007)
8. Saitou, T., Kitamura, T.: Factors in /vvv/ concatenated vowels affecting perception of speaker individuality. In: Proc. of Acoustical Society of Japan 2007 Spring Meeting, pp. 441–442 (2007)
9. Higuchi, N., Hashimoto, M.: Analysis of acoustic features affecting speaker identification. In: Proc. of EUROSPEECH, pp. 435–438 (1995)
10. Kawahara, H.: Straight: An extremely high-quality vocoder for auditory and speech perception research. In: Greenberg, Slaney (eds.) Computational Models of Auditory Function, pp. 343–354 (2001)
11. Kawahara, H., Matsui, H.: Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In: Proc. of ICASSP, vol. 1, pp. 256–259 (2003)
12. Slaney, M., Covell, M., Lassiter, B.: Automatic audio morphing. In: Proc. of ICASSP, pp. 1001–1004 (1995)
13. Takahashi, T., Nishi, M., Irino, T., Kawahara, H.: Average voice synthesis using multiple speech morphing. In: Proc. of Acoustical Society of Japan 2006 Spring Meeting, pp. 229–230 (2006)