AUTOMATIC VOICE ASSIGNMENT TOOL FOR INSTANT CASTING MOVIE SYSTEM

Yoshihiro Adachi† ‡*,Shinichi Kawamoto*‡*,Tatsuo Yotsukura*‡*,Shigeo Morishima*†*, and Satoshi Nakamura*‡

†Science and Engineering, Waseda University, 3-4-1 Okubo Shinjuku-ku Tokyo, 169-8555 Japan
‡ATR Spoken Language Communication Research Laboratories,
2-2-2 Keihanna, Science City, Kyoto, 619-0288 Japan
xyadachi@toki.waseda.jp, shinichi.kawamoto@atr.jp, tatsuo.yotsukura@atr.jp,
shigeo@waseda.jp, satoshi.nakamura@atr.jp

ABSTRACT

In Instant Casting movie System, a personal CG character is automatically generated. The character resembles a participant in a face geometry and texture. However, the voice of a character was an alternative voice determined by the gender of the participant. Therefore sometimes it's not enough to identify the personality of a CG character.

In this paper, an automatic pre-scored voice assignment tool for a personal CG character is presented. Voice is essential to identify a personal character as well as a face feature. Our proposed system selects the most similar voice to the participants from voice database, and assigns it as a voice of CG character. Voice similarity criterion is presented by combination of eight acoustic features. After assigning voice data to a personal character, the voice track is played back in synchronization with the movement of the CG character. 60 voice variations have been prepared to our voice database. Validity of the assigned voice has been evaluated by MOS value. The proposed method has achieved 68% of the theoretical figure that is calculated by preliminary experiments.

Index Terms— Speech analysis, Speaker recognition, Computer applications

1. INTRODUCTION

Instant Casting movie System (ICS) is a visual entertainment system that anyone can appear in a movie as a CG character [1]. The CG character closely resembles each participant, and role-plays animatedly. In ICS, all processes are performed automatically (Figure 1): scanning the facial shape and the face image, reconstructing the 3D face model, and generating the movement and appearance in a screen. However in terms of voice, a voice actor's voice is assigned to each character depending on the gender of the participant. Therefore, the voice of the character is not enough to identify oneself from others.

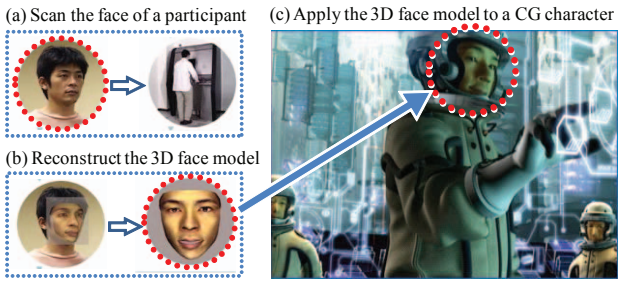

Fig. 1. Instant Casting movie System.

To this problem, we decided to select a similar voice actor to the participant from a voice actor database (DB), and assign the selected voice to the character. This will require the system that can select the similar voice actor to the participant from a voice actor DB. Furthermore, the system needs to apply the selected voice to the scene of a movie.

In speaker recognition which deals with similarity of speech data, the similarity between speakers is calculated based on likelihood of Gaussian Mixture Model (GMM) [2]. The aim of speaker recognition is to perceive oneself, not to search a perceptual similar speaker to a target. Basically, we focus on a perceptual similarity rather than similarity of speaker models. Meanwhile, as for the relation between perceptual similarity and acoustic similarity, Amino proved the strong correlation between perceptual similarity and cepstral distance [3]. Nagashima proved the strong correlation between perceptual similarity and spectrum distance at 2 - 10 kHz with speech data in which utterance speed and intonation were controlled by speakers [4]. We estimate the perceptual similarity using a sentence of speech, because the personality of a speaker appears not only in voice quality but also in utterance speed or prosodic intonation. Therefore, we need to consider multiple acoustic features for various voice characteristics.

In this paper, we describe a method to combine the multiple acoustic features for calculating the perceptual similarity, and our implementation system to realize the ICS considering the participant's voice character.

2. SELECTING SIMILAR SPEAKER

2.1. Estimation method

We estimate the perceptual similar speaker to participants using a combination of multiple acoustic features for more accurate estimation. 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 a number of acoustic features, x_i is a distance of ith acoustic features between speech data, α is the weighting coefficient for each distance of acoustic feature.

We use 8 acoustic features related with the voice personality. These are Mel Frequency Cepstral Coefficient (Static: 12 + Dynamic: 13 = 25 dimension) [2], STRAIGHT Cepstrum of over 35 dimensions and that of 1 dimension [5], Spectrum of over 2.6 kHz, STRAIGHT-Ap under 2 kHz [6] that is a parameter of STRAIGHT [7], fundamental frequency, formants (F1 - F4), and spectrum slope between 0 kHz - 3 kHz [8]. To extract these acoustic features, we use the 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. It can estimate the perceptual similarity accurately, because it represents the temporal structure of acoustic features.

2.2. Optimization of weighting coefficients

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

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

In this equation, α is the ranking of perceptual similarity by the subject. β is that of acoustic similarity using our method. N is the number of speech data. For this optimization, we used speech data uttered by 36 speakers.

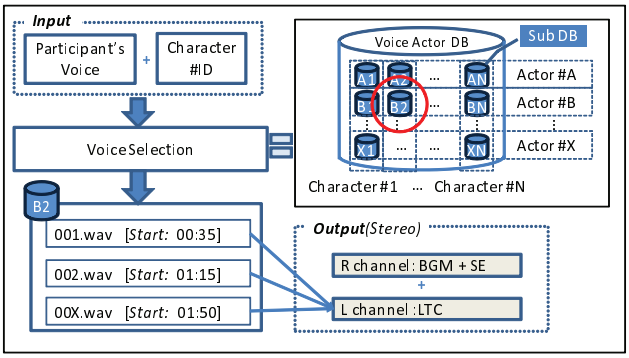

Fig. 2. Voice assignment system.

3. SYSTEM IMPLEMENTATION

3.1. Voice assignment system

Figure 2 shows a voice assignment system. In our proposed system, inputs are participant's voice and character's ID. The voice assignment system selects a sub DB according to the similar voice actor's ID and character's ID from a voice actor DB. Each sub DB has speech data of phrases and information of its playback time. These speech data played based on a Longitudinal Time Code (LTC). With LTC, we are able to deal with the speech data per phrase and easily apply the voice conversion with less quality deteriorating in the future.

3.2. Constructing voice actor DB

Because the DB of the proposed method affects the quality of work, the construction of the DB is important. In general, it is difficult for beginners to read scripts along to the movement of a character in a movie. Therefore, the DB is recorded by professional voice actors. Recorded contents are the speech data of scripts "Grand Odyssey" that has been exhibited at the 2005 World Exposition in Aichi Japan. We have recorded 60 kinds of voice for our proposed system. (Total 5,340 sentences)

3.3. Prototype system

Figure 3 shows an implementation of our proposed method for ICS. Participants record own voice for selecting the similar voice actor to own using the recording PC + headset. The recording PCs select the similar voice actor from the prepared DB. The most similar voice actor's ID and participant's ID are sent to the audio composite PC. On the other hand, the system scans the participant's facial shape with the 3D face scanner and then constructs a 3D facial model for the CG character of the participant. Sound PC outputs stereo audio consisting of LTC at L channel and sound (Mixture of BGM and SE (Sound Effects)) at R channel. This stereo audio is input to the sound distributor. The sound distributor distributes LTC to the audio composite PC and the image composite PC, BGM and

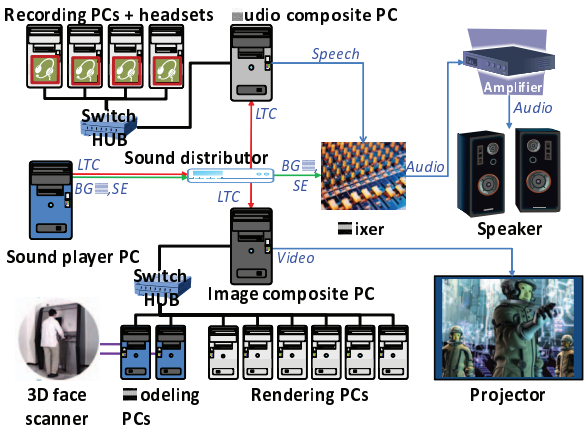

Fig. 3. System implementation for ICS.

SE to the mixer. The audio composite PC sends speech data to the mixer based on LTC. The mixture of speech data and sound is sent to the speaker through the amplifier. The rendering PC composes pictures with pre-rendered background pictures and the 3D facial model of a participant. The image composite PC loads the composite pictures and outputs to the projector.

4. EVALUATION

We made subjects experience the advanced ICS with our method, and asked them following questions.
(1) Could you recognize your character?
(2) Have you felt that a selected voice matches to your character?
(3) Do you agree that ICS with our method is more interesting and exciting than the previous system that assigns the character's voice only according to gender?
Subjects answer the question 1 with "Yes" or "No", the question 2, 3 on a scale 1 (absolutely no) to 5 (absolutely yes).

The number of subjects is 172. We limit the number of participants of each time to 15, because of the fairness in appearance times for characters. Sentence for the selection of the voice actor is a Japanese containing all Japanese vowel sounds, "amembo akaina aiuoe". We record the voice from each participant. Then, the system selects the similar voice actor.

The number of valid responses is 144, and that of person who answered "Yes" for the question 1 is 113. Figure 4 shows results of the question 2 by the 113 subjects. The rate of results at over 4 is 35%. There are two reasons why we couldn't achieve the theoretical figure. First, voice actors perform the script professionally. Therefore, the speech data are quite different from the way that participants usually speak. Second, our voice we always hear and recognize as own is a mixture of air-transmitted and bone-transmitted. However, playing speech data is just air-transmitted one.

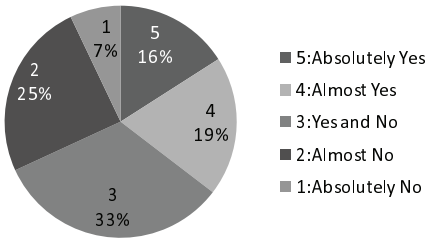

Fig. 4. Rate of subjects who have felt that a selected voice matches own character.

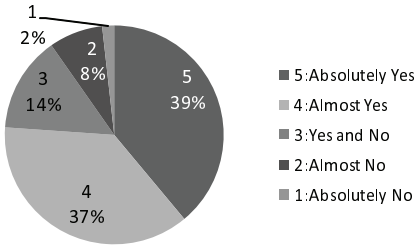

Fig. 5. Rate of subjects who satisfied ICS with our voice selecting system.

Figure 5 shows the result of the question 3. This figure shows 76% subjects are interested in ICS with our similar voice assignment system. Therefore, our system is effective for an entertainment such as ICS.

5. DISCUSSION

5.1. Voice actor DB

We investigated the ideal DB size for the voice assignment system. We have experimented speech data uttered by 28 voice actors. Sentence of the speech is a Japanese, "amembo akaina aiuoe". The number of subjects are 20 (Male:8, Female:12, Ages:19-24 years old). At first, we present pairs of speech data which are uttered by two speakers (regard as Speaker A and Speaker B) to the subjects. Then we ask the subject that "If we change the Speaker A to the Speaker B, how much do you feel mismatch?" The subjects answer with score 0 - 100. (0: The subject feels the mismatch very much, 70: the mismatch is permissible, 100: The subject doesn't feel the mismatch at all.) The pairs of speech data for evaluation is considered counterbalance. Furthermore, we remove the pair of same speech because the pair will be scored 100 by the subjects. We regard an average score by 20 subjects as a score for the pair of speech data, and assess the necessary DB size using these scores.

We divided the 28 voice actors into two groups: one with one speaker for input, and the other with 27 speakers for the DB based on leave-one-out cross validation. Then we nar-

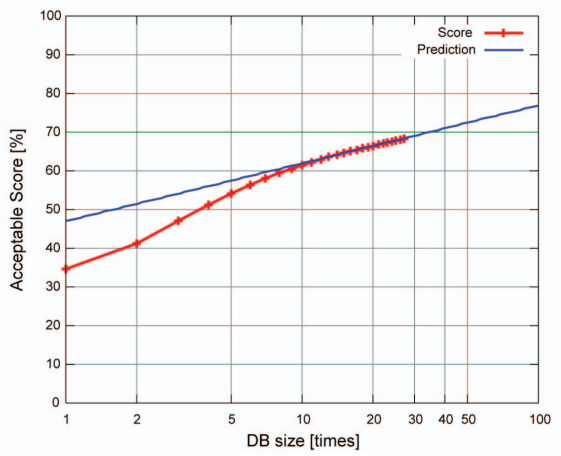

Fig. 6. Ideal DB size for ICS.

rowed 27 speakers in the DB down to N speaker at random, and select the most match speaker to the input speaker from the DB of N speaker. We regard the average score of 28 kinds of input speakers by the subjects as a prospective score when we select the similar speaker from a DB with N times size against to the number of input speakers. Figure 6 shows the results of the DB size N = 1 - 27 and its line of tangency. Gradient of this line of tangency is calculated with N = 20. From this linearly approximated result, the acceptance rate of a selected voice becomes 70 when the N is 30 - 40. We have experimented using speech data uttered by youthful speakers. If we categorize speakers into 3 stages such as children, youth and elder according to age, we need the DB with 90 - 120 time size for the number of input speakers. Since the number of input speakers in ICS is 20, the DB for the voice assignment system needs 1,800 - 2,400 voice actors.

5.2. System evaluation

Since our DB includes 60 kinds of voice and the number of participants is 15, the DB size in Figure 6 is 4 [times]. Now therefore, "Acceptable Score" in Figure 6 is 51.27% . This "Acceptable Score" represents the rate to be tolerated the mismatch by participants and it is regarded as an expectation value for the DB size. Because the evaluation result is 35% and "Acceptable Score" is 51.27% , our implementation method has achieved around 68% (35/51.27) of the theoretical figure.

6. CONCLUSIONS

We have proposed the automatic voice assignment tool considering a participant's voice character. Our system selects a perceptual similar voice actor from a voice actor DB. Additionally speech data uttered by the similar voice actor are played in synchronism with movements of participant's CG character using LTC. Herewith, we improve the mismatch between the face and voice for participant's character. We have evaluated our proposed system with 60 kinds of voices as database. As a result, we conclude that our system has achieved 68% of the theoretical figure for experimental condition and been preferred compared with the previous system by 76% of experiment participants.

In future work, we increase the size of the DB. To select the more similar voice actor, we need to record speech data uttered by amateurs. However, a problem is that speech data uttered by amateurs would degrade a quality of a movie.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] Akinobu Maejima, Shuhei Wemler, Tamotsu Machida, Masao Takebayahasi, and Shigeo Morishima, "Instant casting movie theater: The future cast system," *The IEICE Transaction on Information and System*, vol. E91-D, no. 4, pp. 1135–1148, 2008.

[2] D.A. Reynolds, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. On Acoust. Speech and Audio Processing*, vol. 3, no. 1, 1995.

[3] K. Amino and T. Arai, "Speech similarity in perceptual speaker identification," *Proc. of Acoustical Society of Japan 2006 Autumn Meeting*, pp. 273–274, 2006.

[4] I. Nagashima, M. Takagiwa, Y. Saito, Y. Nagano, H. Murakami, M. Fukushima, and H. Yanagawa, "An investigation of speech similarity for speaker discrimination," *Proc. of Acoustical Society of Japan 2003 Spring Meeting*, pp. 737–738, March 2003.

[5] T. Kitamura and T. Saitou, "Contribution of acoustic features of sustained vowels on perception of speaker characteristic," *Proc. of Acoustical Society of Japan 2007 Spring Meeting*, pp. 443–444, March 2007.

[6] T. Saitou and T. Kitamura, "Factors in /vvv/ concatenated vowels affecting perception of speaker individuality," *Proc. of Acoustical Society of Japan 2007 Spring Meeting*, pp. 441–442, March 2007.

[7] H. Kawahara, "Straight: An extremely high-quality vocoder for auditory and speech perception research," *Computational Models of Auditory Function (Eds. Greenberg and Slaney)*, pp. 343–354, 2001.

[8] N. Higuchi and M. Hashimoto, "Analysis of acoustic features affecting speaker identification," *Proc. of EUROSPEECH '95*, pp. 435–438, 1995.