Scenario-based Speech Assignment Techniques for Instant Casting Movie System

Shin-ichi Kawamoto[12], Yoshihiro Adachi[23], Yamato Ohtani[2],
Tatsuo Yotsukura[2], Shigeo Morishima[3], and Satoshi Nakamura[12]

[1] National Institute of Information and Communications Technology,
Knowledge Creating Communication Research Center,
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0288 Japan
{shinichi.kawamoto, satoshi.nakamura}@nict.go.jp
[2] Advanced Telecommunication Research Institute International,
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288 Japan
{yoshihiro.adachi, yamato.ohtani, tatsuo.yotsukura}@atr.jp
[3] Waseda University,
3-4-1 Okubo Shinjuku-ku Tokyo, 169-8555 Japan
shigeo@waseda.jp

Abstract. In this paper, we propose an improved Future Cast System (FCS) that enables anyone to be a movie star while retaining their individuality in terms of how they look and how they sound. The proposed system produces voices that are significantly matched to their targets by integrating the results of multiple methods: similar speaker selection and voice morphing. After assigning one CG character to the audience, the system produces voices in synchronization with the CG character's movement. We constructed the speech synchronization system using a voice actor database with 60 different kinds of voices. Our system achieved higher voice similarity than conventional system; the preference score of our system was 56.5% over other conventional system.

1 Introduction

Instant Casting movie System (ICS) is a visual entertainment system that makes it possible for anyone to appear in a movie as a CG character [1]. The CG character closely resembles the participant, and role-plays animatedly. In ICS, all processes are performed automatically (Figure 1): the scanning of the facial shape and image, the reconstruction of the 3D face model, and the generation of on-screen appearance and movement. In terms of the voice, however, a voice actor's voice is assigned to each character depending on the gender of the participant. Our goal is to develop a voice assigning function for the ICS. The participant's CG character should be assigned a voice that is matched and similar to the participant's own voice. Furthermore, the voice quality should be better matched with the movie quality. Degradation of the voice quality has a significant effect on the overall quality of ICS. To improve the process of assigning matched voices to the CG characters, two types of voice generation approaches are proposed:

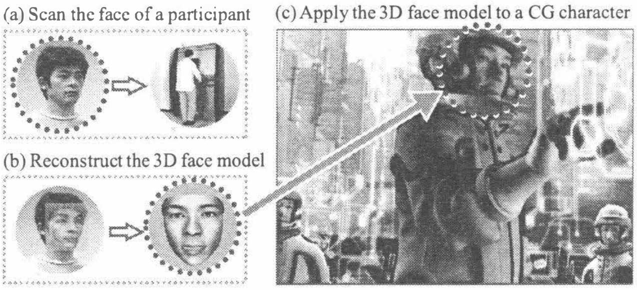
(a) Scan the face of a participant
(b) Reconstruct the 3D face model
(c) Apply the 3D face model to a CG character

Fig. 1. Instant Casting movie System

similar speaker selection and voice morphing. In order to develop a prototype system and to evaluate voice similarity, a voice actor database (DB) is designed with 60 different kinds of voices.

The similar speaker selection process selects from the voice actor DB an actor with a similar voice 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. In speaker recognition, which deals with elements of similarity in speech data, the similarity between speakers is calculated based on the Gaussian Mixture Model (GMM) [2]. The aim of speaker recognition is to perceive oneself, and not to search for a speaker who is perceptually similar to a target. Basically, we focus on perceptual similarity rather than the similarity of speaker models. As for the relationship between perceptual similarity and acoustic similarity, Amino proved that the existence of a strong correlation between perceptual similarity and cepstral distance [3]. Further, Nagashima found evidence of a strong correlation between perceptual similarity and spectrum distance, at 2 - 10 kHz, using speech data in which utterance speed and intonation were controlled by speakers [4]. We use spoken sentences to estimate perceptual similarity because the personality of a speaker is evinced not only in the voice quality but also in the utterance speed and prosodic intonation. Therefore, we need to consider multiple acoustic features for various voice characteristics. The key technology of our method is to combine the multiple acoustic features that are used to calculate perceptual similarity with our implementation system so as to realize within the ICS, a closely matched voice for the participant's character.

The Voice morphing approach is based on blending a few voices to generate a similar voice to that of the participant. Our approach is based on STRAIGHT [5] voice morphing, which is an extra high quality voice morphing technique [6]. 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.

2 Constructing Voice Actor DB

Since the nature and depth of the DB, in relation to our proposed method, affects the quality of our work, careful construction of the DB is imperative. In general, it is difficult for beginners to read scripts to go along with the movement of a character in a movie. Therefore, the DB is recorded by professional voice actors. The recorded content is speech data taken from the scripts of "Grand Odyssey" that was exhibited at the 2005 World Exposition in Aichi, Japan. We have recorded 60 different kinds of voices for our proposed system. (Total: 5,340 sentences)

2.1 Statistics of DB

We investigated the subjective age and gender of each voice actor in the DB. We have experimented with 60 kinds of speech data gleaned from various voice actors. The specific Japanese sentence used in our experiment is "amembo akaina aiueo." The number of subjects is 20.

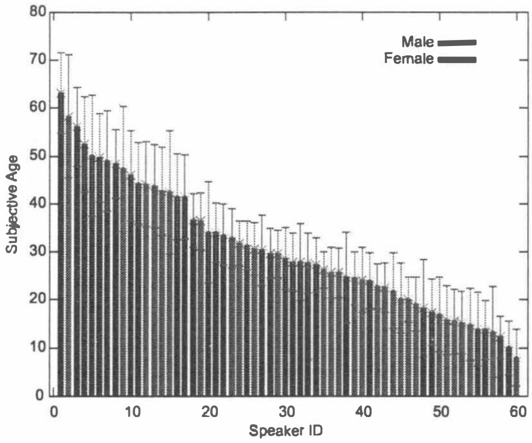

Fig. 2. Statistics of Voice Actor DB (Average and Standard Deviation of Subjective Age, and Subjective Gender)

Figure 2 shows the results of the subjective age and gender for each voice. The y-axis indicates the average subjective age. The error bars denote the standard deviations of the subjective ages. The two colors indicate the subjective genders which are decided by majorities among the voted results. This histogram shows that our voice actor DB covers a wide range of participants, in terms of their age. In addition, our DB gets a balanced male-female ratio.

2.2 Ideal DB Size for Selecting Similar Speaker

We investigated the ideal DB size for the voice assignment system. We have experimented speech data uttered by 28 voice actors. The Japanese sentence "amembo akaina aiueo" is used. The number of subjects is 20. At first, we present to the subjects pairs of speech data which are uttered by two speakers (regarded as Speaker A and Speaker B). Then we ask the subject "If we change Speaker A to Speaker B, how much of a mismatch do you feel?" The subjects were instructed to answer by checking a score in the range 0 - 100. (0: the subject perceives a mismatch very strongly; 70: the mismatch is permissible; 100: The subject doesn't feel a mismatch at all.) The pairs of speech data for evaluation are considered counterbalanced. Furthermore, we remove the pairs of same speech because such pairs will obviously be scored as 100 by the subjects. We consider the average of the scores given by the 20 subjects as for each pair of speech data, and assess the necessary DB size using these scores.

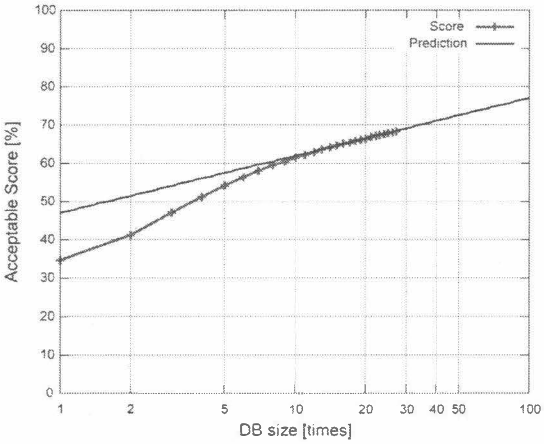

Fig. 3. Ideal DB size for ICS

We divided the 28 voice actors into two groups: one, with a single speaker for input, and the other, with 27 speakers for the DB, based on the principle of leave-one-out cross validation. Then we narrowed the 27 speakers in the DB down to N speakers at random, and selected the speaker most matched to the input speaker from the DB of N speakers. We regarded the average scores given by the subjects on the 28 kinds of input speakers as the prospective scores when we conducted similar speaker selection from the DB with N times size against to the number of input speakers. Figure 3 shows the results of the DB size $N =$ 1 - 27 and its line of tangency. The 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 the speakers into 3 stages such as children, youth and elderly according to their age, we need a 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.

3 Selecting Similar Speaker

3.1 Estimation method

We estimate the perceptual similar speakers to 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 features between the speech data, and α is the weighting coefficient for each such distance between acoustic features.

We use 8 acoustic features related with the voice personality. These are the Mel Frequency Cepstral Coefficient (Static: 12 + Dynamic: 13 = 25 dimension) [2], the STRAIGHT Cepstrum of over 35 dimensions and that of 1 dimension [7], the Spectrum of over 2.6 kHz, the STRAIGHT-Ap under 2 kHz [8] that is a parameter of STRAIGHT [5], 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 perceptual similarity represented 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 perceptual 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 perceptual similarities 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 perceptual similarity and that of acoustic similarity. Acoustic similarity is calculated using the 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 with respect to perceptual similarity ascribed by the subject. β 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

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 [10]. STRAIGHT-based morphing [6] handles the feature vectors of time-frequency representation derived by STRAIGHT [5], STRAIGHT spectrogram, Aperiodicity Map and F_0. 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 [11]. The procedure is almost 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 $\boldsymbol{r}$.

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

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

$$\begin{aligned}\hat{\boldsymbol{r}} &= \arg\min_{r} \|\boldsymbol{y} - \hat{\boldsymbol{x}}_{mrp}\|^2 \\ &= \arg\min_{r} \left\| \boldsymbol{y} - \sum_{s=1}^{S} r_s \boldsymbol{x}_s \right\|^2 \qquad (4)\\ \boldsymbol{r} &= [r_1, r_2, \cdots, r_S]^\top \qquad (5)\end{aligned}$$

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

$$\epsilon(\boldsymbol{r}) = \sum_{\tilde{f}=1}^{F} \sum_{\tilde{t}=1}^{\tilde{T}(\boldsymbol{r})} \left(y_{\tilde{t}}(f) - \boldsymbol{x}_{\tilde{t}}(\tilde{f})\boldsymbol{r} \right)^2 \tag{6}$$

where $y_{\tilde{t}}(f)$ is the feature vector of a target speaker with a regularized time domain, $\boldsymbol{x}_{\tilde{t}}(\tilde{f}) = \left[x_{\tilde{t}}^{(1)}(\tilde{f}), x_{\tilde{t}}^{(2)}(\tilde{f}), \cdots, x_{\tilde{t}}^{(S)}(\tilde{f}) \right]$ 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 $\hat{\boldsymbol{r}}$. $\tilde{T}(\hat{\boldsymbol{r}})$ is the regularized speech duration in the time domain as shown below.

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

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 an equation 6 analytically. Therefore we use an iterative approach to solve for a blending ratio vector using the following formulae.

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

$$\begin{aligned} E_n &= \left(\frac{\partial^2 \acute{\epsilon}(\hat{\boldsymbol{r}}_n)}{\partial \hat{\boldsymbol{r}_n}^2} \right)^{-1} \frac{\partial \acute{\epsilon}(\hat{\boldsymbol{r}}_n)}{\partial \hat{\boldsymbol{r}_n}} \\ &= \left(\overline{\boldsymbol{X}}_n^{\top} \overline{\boldsymbol{X}}_n \right)^{-1} \overline{\boldsymbol{X}}_n^{\top} \left(\overline{\boldsymbol{Y}}_n - \overline{\boldsymbol{X}}_n \hat{\boldsymbol{r}}_n \right) \end{aligned} \tag{9}$$

where $\acute{\epsilon}(\hat{\boldsymbol{r}})$ is an approximation formula of equation 6 that assumes the blending ratio $\hat{\boldsymbol{r}}_n$ is constant. $\overline{\boldsymbol{X}} = \left[\boldsymbol{X}_1^{\top}, \boldsymbol{X}_2^{\top}, \cdots, \boldsymbol{X}_{\tilde{T}(\hat{\boldsymbol{r}}_n)}^{\top} \right]^{\top}$ are regularized feature vectors in the time-frequency domain that are updated n times by the blending ratio vector $\hat{\boldsymbol{r}}_n$. $\overline{\boldsymbol{Y}} = \left[\boldsymbol{y}_1^{\top}, \boldsymbol{y}_2^{\top}, \cdots, \boldsymbol{y}_{\tilde{T}(\hat{\boldsymbol{r}}_n)}^{\top} \right]^{\top}$ are regularized feature vectors in the time domain that are updated by a blending ratio vector $\hat{\boldsymbol{r}}_n$. $\boldsymbol{X}_{\tilde{t}_n}$, $\boldsymbol{x}_{\tilde{t}_n}^{(s)}$, and $\boldsymbol{y}_{\tilde{t}_n}$ are defined as follows.

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

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

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

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 the Prototype System

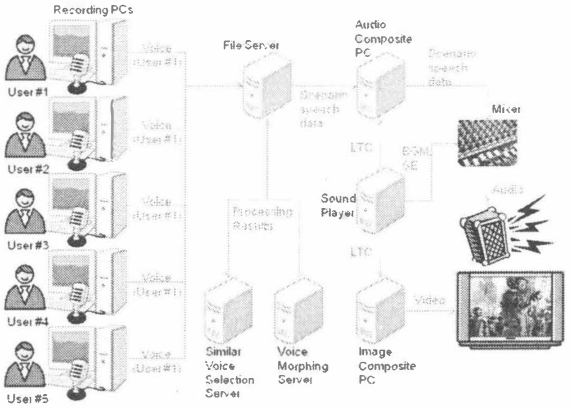

Fig. 4. Overview of Prototype System

Figure 4 shows our prototype system for ICS. Participants record their own speech using recording PCs. Similar Speaker Selection Servers and Voice Morphing Servers calculate the results of scenario-based speeches that are intended to be similar to the participants' voices. These scenario-based 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 consist 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 with over 100 participants, including children and elderly people, and was found to work well.

6 Evaluation

In this subjective evaluation, we ascertain speaker similarity in the result voices obtained by two approaches: selecting similar speakers and voice morphing. At first, we present three voices (target speaker's voice X, and the result voices obtained by the two approaches A and B) to the subjects. Then we ask the subject "Which voice is similar to voice X?" The subjects answer saying either A or B. The number of subjects is 40. The number of target speakers is 6. The pairs of voices, A and B, entered for evaluation are considered counterbalanced. In this evaluation, we prepare two groups: 1) read speech that includes 16 utterances (SENT), and 2) FCS scenario speech with BGM and sound effects that includes 39 utterances (FCS). Both approaches use the same voice actor DB that is described in section 2.

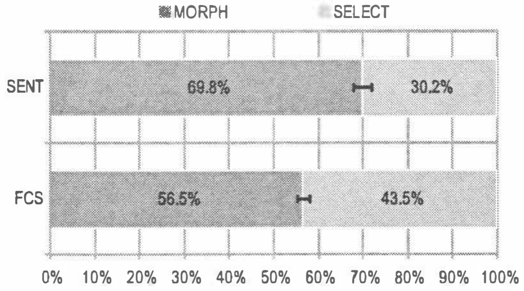

Fig. 5. Subjective evaluation result of speaker similarity

Figure 5 shows preference scores and their confidence intervals in the comparison between the two approaches. SELECT refers to voices that are a result of selecting similar speakers, and MORPH implies voices that are a result of voice morphing. Fig. 5 indicates that voice morphing is slightly better than similar voice selection in this particular situation. Since the voice used in the conventional system [1] was selected from 2 speakers based on gender information (male or female), we are in a position to assert that the approach of selecting similar speakers based on our voice actor DB works better than the conventional system [1]. In addition, combining the two approaches will provide the system with an output that is more similar to the target participant's voice.

7 Conclusion

In this paper, we described a novel scenario with regard to a speech assignment system for extending the casting functions of an instant casting movie system. In this system, two types of scenario-based speech assignment technologies are combined; similar speaker selection and voice morphing. We constructed a voice actor DB that includes 60 kinds of voices, and implemented a prototyping system with speech synchronization mechanism using the LTC. Our system worked well for various participants. Speaker similarity was found to improve with the use of our techniques, based on a conventional FCS.

One important course of future work is to develop an automatic speech quality evaluation technique. 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. This technology will reduce operational costs of our system.

It is also important to establish an archetype for the designing strategy for the construction of a voice actor DB. System performance depends on the quality of the voice actor DB. Constructing a voice actor DB is a time- and money-consuming 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.

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., Takebayahasi, M., Morishima, S.: Instant casting movie theater: The future cast system. The IEICE Transaction on Information and System **E91-D**(4) (2008) 1135–1148
2. Reynolds, D.: Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. On Acoust. Speech and Audio Processing **3**(1) (1995)
3. Amino, K., Arai, T.: Speech similarity in perceptual speaker identification. In: Proc. of Acoustical Society of Japan 2006 Autumn Meeting. (2006) 273–274
4. Nagashima, I., Takagiwa, M., Saito, Y., Nagano, Y., Murakami, H., Fukushima, M., Yanagawa, H.: An investigation of speech similarity for speaker discrimination. In: Proc. of Acoustical Society of Japan 2003 Spring Meeting. (March 2003) 737–738
5. Kawahara, H.: Straight: An extremely high-quality vocoder for auditory and speech perception research. Computational Models of Auditory Function (Eds. Greenberg and Slaney) (2001) 343–354
6. Kawahara, H., Matsui, H.: Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In: Proc. of ICASSP. Volume 1. (2003) 256–259
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. (March 2007) 443–444
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. (March 2007) 441–442
9. Higuchi, N., Hashimoto, M.: Analysis of acoustic features affecting speaker identification. In: Proc. of EUROSPEECH. (1995) 435–438
10. Slaney, M., Covell, M., Lassiter, B.: Automatic audio morphing. In: Proc. of ICASSP. (1995) 1001–1004
11. 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. (March 2006) 229–230