HandySinger: Expressive Singing Voice Morphing using Personified Hand-puppet Interface

Tomoko Yonezawa
ATR IRC Labs.
Hikari-dai, Seika-cho
Kyoto 619-0288, Japan
yone@atr.jp

Noriko Suzuki
ATR MIS Labs.
Hikari-dai, Seika-cho
Kyoto 619-0288, Japan
noriko@atr.jp

Kenji Mase
Nagoya University
Furoh, Chikusa,
Nagoya 464-8601, Japan
mase@itc.nagoya-u.ac.jp

Kiyoshi Kogure
ATR IRC Labs.
Hikari-dai, Seika-cho
Kyoto 619-0288, Japan
kogure@atr.jp

ABSTRACT

The HandySinger system is a personified tool developed to naturally express a singing voice controlled by the gestures of a hand puppet. Assuming that a singing voice is a kind of musical expression, natural expressions of the singing voice are important for personification. We adopt a singing voice morphing algorithm that effectively smoothes out the strength of expressions delivered with a singing voice. The system's hand puppet consists of a glove with seven bend sensors and two pressure sensors. It sensitively captures the user's motion as a personified puppet's gesture. To synthesize the different expressional strengths of a singing voice, the "normal" (without expression) voice of a particular singer is used as the base of morphing, and three different expressions, "dark," "whisper" and "wet," are used as the target. This configuration provides musically expressed controls that are intuitive to users. In the experiment, we evaluate whether 1) the morphing algorithm interpolates expressional strength in a perceptual sense, 2) the hand-puppet interface provides gesture data at sufficient resolution, and 3) the gestural mapping of the current system works as planned.

Keywords

Personified Expression, Singing Voice Morphing, Voice Expressivity, Hand-puppet Interface

1. INTRODUCTION

Personification enriches the expressions of communications and emotional performances. A singing voice can be considered a kind of personified musical expression that pretends to evoke someone else by using the voice. Moreover, musical expressions are enriched by the verbal, nonverbal and emotional expressions of a singing voice. Therefore, it is important to control the expressions of a singing voice in real time for use as an effective medium for real-time performances and communications. However, there has been little research on the utility of a singing voice as a medium with a personified control system; in addition, few synthesis methods for a singing voice have been developed for actual real-time performance with a sufficient range of expressions.

NIME'05, May 26-28, 2005, Vancouver, BC, Canada.

Our approach employs Expressive Singing Voice Morphing (**ESVM**) for real-time musical expressions with personification. ESVM synthesizes a singing voice with an indiscrete and smooth expression that is suitable for natural real-time sound control.

Although personification has been embodied in virtual agents[1] and robots[2], they still do not have a sufficient variety of expressions and natural movements. We employed a hand-puppet interface for personification to achieve the feeling of direct human control in the real world. The hand puppet covers a fundamental tool of the user's bodily expressions, i.e., the hand, with a personified surface. Therefore, it is considered effective for intuitively personified expressions. A user can make gestures with her/his hand with actual feeling and a sense of touch.

In this paper, we first explain our motivation by referring to related research. Next, we show our system configuration in detail: 1) ESVM synthesis method, 2) hand-puppet interface, and 3) control mappings on expressional elements. We then evaluate and analyze our system configuration based on these three aspects. Furthermore, we discuss the effect of personified complex media. Finally, we conclude that a personified interface is suitable for personified musical expression.

2. MOTIVATION

Naturally personified human-like expression involves empathy for the performer of the expression in human-human communication. We believe that the personified hand-puppet interface gives an actual feeling of the performance to the performer. The performer can gesture intuitively by controlling the puppet with her/his hand from the inside. The sense of touch creates an illusion that lets the performer feel as if her/his hand were the puppet's body. In recognizing the puppet's gestures by its personified shape, the audience can empathize with the expression conveyed in the performance without difficulty.

To involve the performer and the audience more deeply, personified sound expression is important. We found that musical expressions are more effective than voice-like sounds, which have continuous F_0, in face-to-face communication[3]. A singing voice includes musical information and the emotional nonverbal expressions of a human's voice. HandySinger generates a singing voice for familiar and intuitive personification of another person's voice. For natural expressions that change constantly, we propose a method of interpolating the strengths of expression of a particular singer's singing voice.

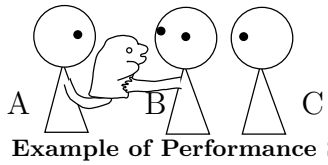


Figure 1: Example of Performance Situation

For instance, this configuration allows the performer to teach children how a singing voice is expressed based on bodily gestures. A performance of pretending to be others, as in a puppet show, takes advantage of this configuration. We assume the situation illustrated in Figure 1: i) *A* performs a singing voice by using the hand-puppet interface. She/he can feel the sound expression changing with the gestures of her/his hand. ii) *B* can not only listen to but also touch and feel the sound's feedback from within the puppet. iii) *C* can listen to the naturally interpolated expressions and see the gesture of the puppet at the same time. iv) Based on iii), *A* feels both a) the sense of touch by *B* occurring within the hand-puppet and b) a change in the singing voice expression corresponding to a).

3. RELATED WORKS

Stuffed puppets have been used for personification in a variety of expressive media, such as pet robots and the covers of cellar phones. We have conducted research on a musically expressive doll[3] that controls the sound parameters of several musical instruments. In this system, the user feels a sense of touch from outside of the puppet and musical feedback at the same time. This configuration leads the users to concentrate on only either the musical control or the affective interaction with the puppet. To solve this problem, in this paper we propose using a hand puppet that contains a part of the user's body, i.e., the hand, and enables easier control from inside.

Mulder et al.[4] introduced a musical tone controller using two gloves. Fels et al.[5] used a glove interface as a phoneme controller. "Squeezevox"[6] is a phoneme and pitch controller with an accordion. Their works differ from our approach, which transmits the meaning of a hand movement through the gestures of the puppet. In this research, it is important to make the feeling of the sound feedback coincide with the gesture's meaning.

Cano et al.[7] proposed a karaoke system for singing voice morphing between different singers, from the user's voice to the voice of a professional singer. Sogabe et al.[8] and Matsui et al.[9] investigated the sound morphing of emotional speech by a particular speaker. The former research used different singers, and the latter includes different values of speech speed and F_0. Thus the previous research efforts in sound morphing provided new synthesized sound; in contrast, our research aims to vary and smooth out the expression in the voice of a particular singer using the same singer, speed, and F_0.

4. SYSTEM CONFIGURATION

To express various singing voices with a hand puppet, the system needs 1) variously expressed singing voices, 2) an input device that measures the motion of the hand puppet, and 3) the ability to adapt the motion of the hand puppet to an adequate voice expression. We constructed these three main axes of the "HandySinger" system configuration. Details of the system configuration are described as follows.

Table 1: Expression Types in Recorded Voices

expression	singing instruction
"normal"	flat, without expressions
"dark"	entirely like interior tongue vowel
"whisper"	including more white noises
"wet"	entirely nasal voice

Table 2: ESVM Synthesis

	base	target		base	target
A-1	normal	dark	B-1	whisper	dark
A-2	normal	whisper	B-2	wet	dark
A-3	normal	wet	B-3	wet	whisper

4.1 ESVM Synthesis

We collected variously expressed singing voices by a particular singer for use in singing voice synthesis based on varying the expression's strength with morphing technology. It is possible to synthesize the voice parameters, but we focused on vocal synthesis from the existing data for more natural expression. Voice morphing is an appropriate synthesis technique for maintaining individuality and naturalness at the same time.

We recorded the singing voice of a female amateur singer in her twenties at a sampling frequency of 44.1 kHz. The singer was instructed to sing in four types of expressions: "normal," "dark," "whisper," and "wet" voice (Table 1) while keeping each expression consistent in her singing. Among various expressions, we selected the above four from the viewpoint of the technical skill involved in the song types. Here, "dark" emphasizes expressiveness like that produced by an *opera* singer, "whisper" is a hoarse voice like a lullaby sung as interlude expressions in certain songs, and the "wet" expression is used in *pop music* for temporally emotional emphasis.

The amateur singer sang a Japanese nursery rhyme, "Furusato" ("Hometown"), with an accompaniment that arranges speech speed and F_0 in the same way. We synthesized the variously expressed morphed singing voices by applying STRAIGHT Morphing [10].

As shown from A-1 to A-3 in Table 2, we first synthesized morphed singing sounds expressed at various strengths by using "normal" as the base and the three types of singing voice as the targets. Then, as shown from B-1 to B-3 in Table 2, we synthesized morphed singing sounds with two kinds of expressions by using each pair of expressed voices as both the base and the target. To adopt not only interpolation but also extrapolation for the emphasized expressions, the morphing rate was set to 0 or less and to one or more. As sufficient steps for tracing the interpolation, we set the morphing rates from -0.333 (-2/6) to 1.333 (8/6) over eleven steps with equal intervals of 0.167 (1/6).

4.2 Hand-Puppet Interface

As an input device for singing voice expression, a hand puppet must consist of personified parts of the body that can be adequately controlled for gestural expression. It is noteworthy that the gesturing interface needs at least the upper half of the puppet's body, as in robot construction[2]. Accordingly, this system employs a hand puppet consisting of two arms and a head controlled by using three fingers. In the current implementation, we did not incorporate mouth control for singing timing in order to concentrate on expression by whole-body gestures and the motion of the puppet

itself.

It is important to capture the motion of the hand as the motion of the puppet itself for building a suitable input interface in terms of appearance. For capturing sufficiently accurate data as an expressional control for the singing voice, the hand puppet consists of a stuffed penguin as a personifying cover and a sensor-covered glove (Figure 2) as an independent capturing device. As shown in Figure 3, the thumb of the right hand controls the left arm of the puppet, the middle finger the right arm of the puppet, and the forefinger the head of the puppet. For sensing the motion of the puppet's gestures, this system has seven bend sensors and two pressure (touch) sensors (Table 3). Each finger of the glove has bend sensors at two axes. As shown in the right image of Figure 2, the glove has two pressure (touch) sensors at the tips of the thumb and middle finger corresponding to the hands of the puppet. Not only the bend-forward but also the bend-back movement of the forefinger can be detected because the forefinger slightly bends forward in advance when the head of the puppet looks straight up.

The sensors' analog signals are sent to an A/D converter (Infusionsystems I-CubeX) and changed into MIDI signals. A PC (Windows XP) receives them thorough a MIDI interface (Roland UM-2). A sound control program (Section 4.3) produces a singing voice, and the user listens to the singing voice from speakers connected to the PC. Figure 4 shows the sensors capturing typical gestures: 1) hand-waves using the thumb, 2) "nodding" with the forefinger, 3) hand-waves using both the thumb and the middle finger, 4) bend-back of both the thumb and the middle finger, 5) "clap" or "clasp" with the thumb and the middle finger.

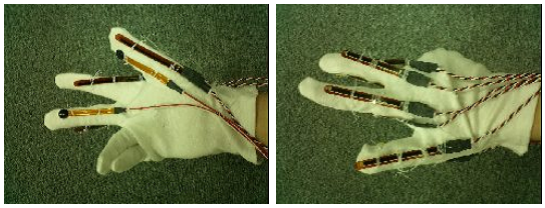
4.3 Sound Control Mappings

For assignment of singing voice expression to appropriate motion, it is important to consider the user's physical experience, such as the difficulty of the mapping rules. For instance, we need to consider whether a physically difficult hand motion matches the difficulty of emphasized expressions. We thus defined the physical situation of the hand in terms of the physical and emotional situation of the voice as explained below.

To control singing voice expression at various kinds and strengths, we defined controllable parameters: 1) singing voice volume, 2) type of voice expression, "dark," "whisper," or "wet," and 3) strength of each expression type. In the current implementation, "dark" strength, A-1, is mapped to expansive *opera*-like gestures for emphasis in a song, "whisper" strength, A-2, is mapped to a drooping gesture of the head, and "wet" strength, A-3, is mapped to a stretching gesture of the arm as done by *pop* singers.

The gestures can be separated into time-sequential gestures and temporal gestures, but sound controllers are not appropriate for time-sequential gestures because feedback must be intuitive in the experience provided by HandySinger. Table 4 shows an outline of the mapping used in this configuration. The *neutral* shape without any power is mapped to "normal" as the origin of each expression. The origin and range of each sensor are calculated to normalize their values for weights of the other expressions. We adopted the bend of the wrist as a tilt of the puppet body, and this is independent of the other hand gestures.

The sound control mapping software is built with a PureData[11] program. The program selects expressional cate-



A) the back side B) the palm side

Figure 2: Sensors on the Glove

A) appearance B) inside of the puppet

Figure 3: Glove installed in the Puppet

Table 3: Sensor Values installed in the Glove

sensing	sensor	max	min	destination
thumb	bend1	110	60	the palm side
thumb	bend2	100	0	forefinger side
forefinger	bend3	90	35	the palm side
forefinger	bend4	110	35	middle finger side
middle finger	bend5	110	35	the palm side
middle finger	bend6	110	35	forefinger side
wrist	bend7	50-60	20-30	bend backward
thumb	touch1	127	<127	fingertip pressure
middle finger	touch2	127	<127	fingertip pressure

⋆ maximum motion: minimum in sensor value

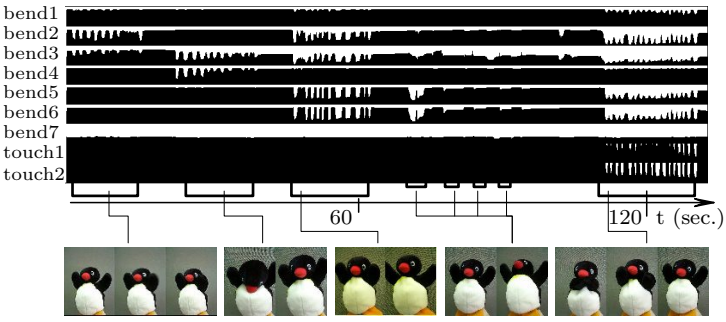


Figure 4: Sensor Signals from the Hand Puppet

Table 4: Parameter Mapping to Gestures

expression	parts of puppet body	gesture
"dark"	head & hands hands	bend-back clasp
"whisper"	head	droop
"wet"	hands	stretch forward
volume	tilt of the body	bend-back

gories of singing voice and playback and controls the smoothing of the volume with expression at the desired strength. When the strength of each expression increases suddenly over 0.5, smoothing works with a time delay of 50 milliseconds.

5. SYSTEM EVALUATIONS

In this section we evaluate 1) the perceptual effect and the naturalness of the interpolated expression of a singing voice, 2) whether the interface design is capable of acquiring motion data at adequate resolution, and 3) whether the system can translate sensor signals into singing voice controls.

5.1 Perception Test of Singing Voice Morphing

To examine the effect of the morphed singing voice, we conducted a perception experiment. In this test, we aimed to verify that the morphing method enables perceptual interpolation even when the base and target have the same singer, the same speed, and the same F_0.

Hypotheses: We proposed three hypotheses: ⋆1) that expressions of the singing voice are different from each other, ⋆2) that the expression level is changed by morphing, and ⋆3) that the morphed voice made from the two expressed voices is different from "normal."

Method: The experiment's subjects listened to stimulating sounds through headphones attached to a Windows 2000 PC and gave subjective evaluations in seven grades: "completely suitable, very suitable, somewhat suitable, indeterminate, somewhat unsuitable, very unsuitable, completely unsuitable" according to the instructed criteria on the GUI interface of the Tcl/Tk program.

Subjects: Thirteen people aged from twenties to lower-thirties (six females and seven males).

Stimulating Sound: We adopted the synthesized morphed sound shown in Table 2 and four original sounds: "normal," "dark," "whisper" and "wet." As a sample of the morphed voice, we selected six morae, "Ko Bu Na Tsu Ri Shi," in the morphed song from the synthesized song data described in Section 4.1. Speech speed is about 2.0 morae/second, and F_0 range is approximately 300 Hz to 450 Hz on average in each musical interval. Each sound length is about 3.0 seconds.

Procedure: The experiment was conducted through each of the tests described below.

⋆1) Subjects evaluated sounds using the seven grades listed above in pairwise comparison between normal and (normal, dark, whisper or wet) while each pair was continuously played back.

⋆2) Subjects listened to the morphed sound of A-1 in Table 2 and judged two evaluation items, I) expression of "dark" and II) naturalness, according to the seven grades. They did the same experiments for A-2 as "whisper" and A-3 as "wet." In preparation for evaluating an item, before this experiment subjects were instructed to listen to a control "dark" sound to confirm what is defined as "dark." They were also instructed to base the criterion of naturalness on how much they felt the sound resembles a human voice.

⋆3) Subjects identified each member of the pair [normal, (morphed voice from B-1 to 3)] continuously played back at the seven grades.

Results of Perception Test: The MOS averages of the identification results of perception test⋆1) compared with "normal" are shown in Figure 5. To verify the difference between "normal" and the other sounds, Table 5 shows T-test results between the identification of ["normal", "dark," "whisper" or "wet."] These results indicate that the expressed singing voice is accurately perceived as different from "normal" in perceptual feeling.

The results of perception test⋆2) (Figure 6) show that MOS averages of the expressional strength correspond to the morphing ratio. Figure 7 shows the naturalness of the morphed sound. Although we estimated MOS values to be higher at the morphing ratio of 0 and 1, a deeper expression was not recognized as natural in the cases of A-1 and A-3. It is possible that a morphed voice with a continuous hard expression was recognized as an artificial voice.

Perception test⋆3) shows that expressional morphing between two expressed voices gives a new expression that is different from "normal." Figure 8 shows that the morphed voices from B-1 to B-3 at a morphing ratio around 0.5 are comparatively more similar to "normal" than are the voices at other ratios. Therefore, we used morphing ratios of 0.33 to 0.67 for B-1 to B-3. To verify that "normal" is similar to B-1, B-2 and B-3, the T-test results are shown in Table 6. These show that a synthesized voice is not significantly different from "normal" even at a morphing ratio around 0.5.

Thus we confirmed that our hypotheses were correct in this experiment. These results show that the morphing of singing voices can supply rich expression by varying the kinds and strengths of expressions in the perceptual measure.

5.2 Data Capturing Test

Sensor data are used as sample points of the continuous signal and as a trigger of the change in its value. The system thus needs data over 10-Hz. To control ESVM in eleven steps, it is sufficient to capture the controller's resolution over 11 stages, and there is no remarkable difference in perceptual sense between a pair of the synthesized ESVM from one morph ratio to another. That is to say, the system needs a motion range that is over 10 percent of the sensor resolution.

Our system configuration meets the demand for resolution of 127 stages and a frequency of 20 Hz in nine channels, although we consider the resolution within the motion range. To confirm the range of each sensor, we captured sensor signals by the gestures of the subjects.

Hypotheses: 1) The users can perform gestures using the hand puppet at sufficient resolution within the range of each motion. 2) The sensors can exceed the desired resolution over eleven stages. 3) The sensor signals are not different among subjects with different palm sizes.

Method: To investigate whether the input signals are significant, subjects were assigned to perform gestures: 1) bend left hand toward its belly twenty times, 2) nod twenty times, 3) bend right hand toward its belly twenty times, 4) clap twenty times, 5) bend its head and hands backward two times, and 6) bend the body backward while using the wrist of the subject two times. The experimenter instructed the subjects on which gestures to make by performing with another similar puppet as a model, and the sensor signals of the gestures were recorded at the same time. The experimenter also measured the length of each subject's palm.

Subjects: Eight females, 24–36 years old.

Results: Every subject could gesture according to assign-

Table 5: T-test between the Identification of ["normal", "dark," "whisper" or "wet"] ($\alpha = 0.01$)

"dark"	"whisper"	"wet"
$t_{(12)} = 15.0$ $p < .01$	$t_{(12)} = 10.2$ $p < .01$	$t_{(12)} = 24.0$ $p < .01$

Table 6: T-test of Identification between ["normal", from B-1 to 3] ($\alpha = 0.01$)

morph ratio	B-1	B-2	B-3
0.33(2/6)	$t_{(12)} = 9.70$ $p < .01$	$t_{(12)} = 9.96$ $p < .01$	$t_{(12)} = 9.06$ $p < .01$
0.50(3/6)	$t_{(12)} = 12.8$ $p < .01$	$t_{(12)} = 7.25$ $p < .01$	$t_{(12)} = 9.80$ $p < .01$
0.67(4/6)	$t_{(12)} = 7.23$ $p < .01$	$t_{(12)} = 5.33$ $p < .01$	$t_{(12)} = 9.68$ $p < .01$

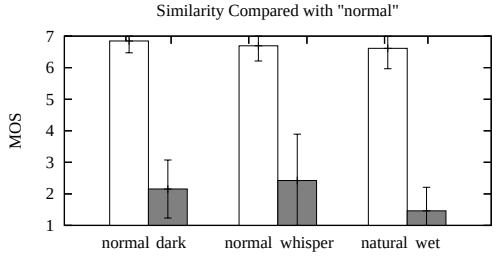


Figure 5: Similarity of Expression

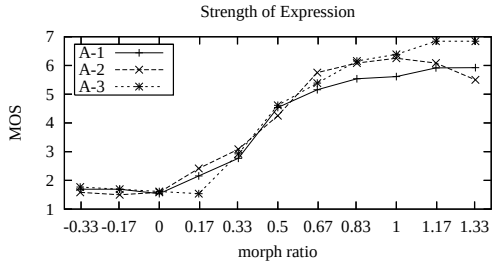


Figure 6: Evaluation of Expressional Strength

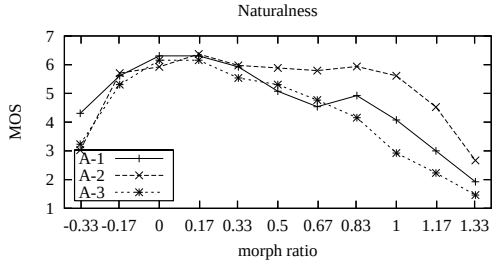


Figure 7: Evaluation of Naturalness

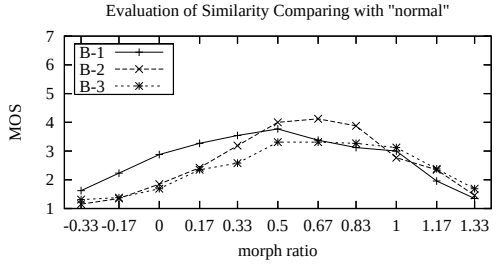


Figure 8: Similarity of Morphing Sound between Expressed Voices

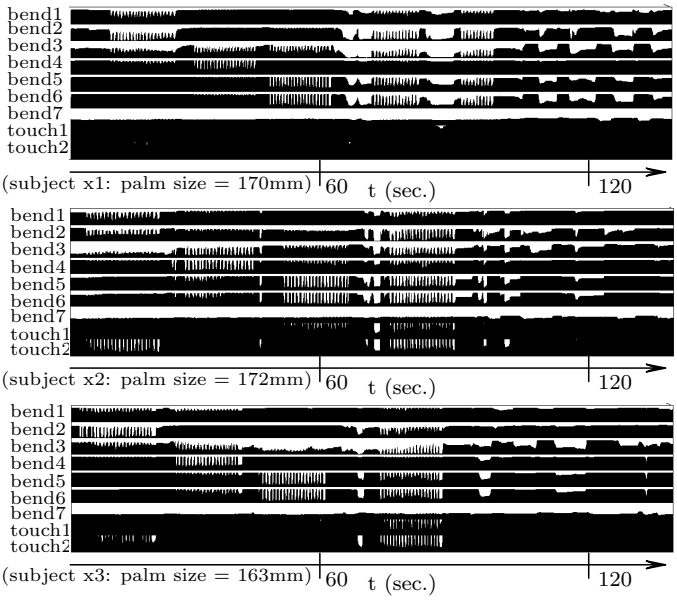


Figure 9: Example of Data Capturing Test

ments 1) to 3), 5) and 6). They enjoyed controlling the puppet from inside. The subject who had the smallest palm could not gesture assignment 4) because her finger length was too short to reach to the glove's fingertip. Figure 9 shows examples of the sensor signals. Figure 10 shows the results of the data range and the palm size for each subject. The larger average and the smaller standard deviation are preferable as higher resolution of the motion. The results show that the sensor signals exceeded the desired resolution, although the smallest palm produced a slightly lower resolution.

5.3 Confirmation of System Operation

We tried to confirm the mapping operation explained in Section 4.3 by showing the recorded signals and the control parameters. For around 45 seconds, the user first performs "bend back" with the body tilted backward. Then, she bends the hands of the puppet backward with the body tilted. Next, she stretches the hands of the puppet ahead, and finally she gestures by clasping the puppet's hands.

Figure 11-A is an example of several sensors' inputs. Figure 11-B shows the extracted information of gesture from the signals of Figure 11-A. Figure 11-C shows the controlled parameter of each expressional strength. Thus we confirmed the system's successful operation: how the user listened to the smooth singing voice's expression corresponded to the gestures of the user's right hand.

6. DISCUSSION

Expressivity of ESVM: ESVM has been employed for interpolation between "normal" without any expression and with three types of expressions that were determined beforehand. The number of expressions was appropriate for control by one hand. Now let us discuss the types of expressions used for the singing voice. An emotional speech database is constructed from several emotional categories. Moreover, in contrast with speech, musical scores limit the speed and the intonation (F_0) of a singing voice. To collect more effective expressions, we would additionally need to select the most effective item of the existing singing voice

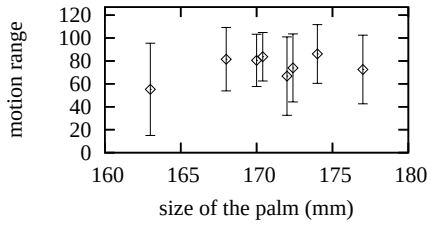


Figure 10: Data Range and Palm Size

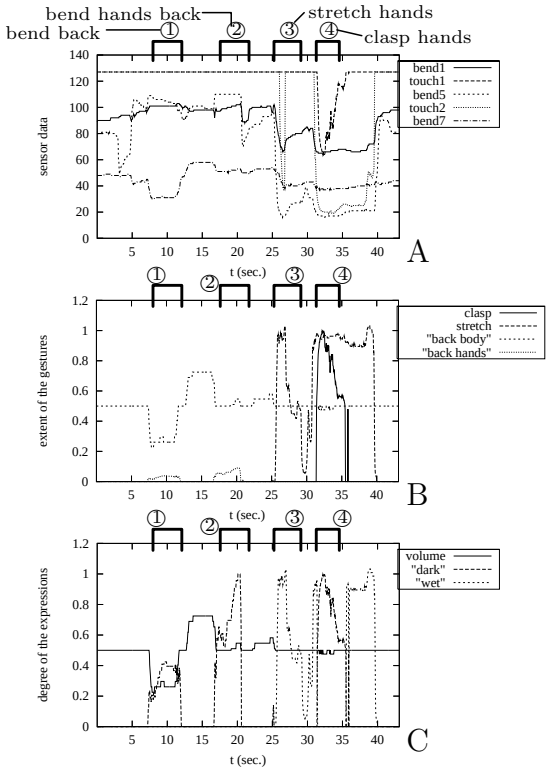


Figure 11: Relations among Sensors, Gestures, and Expressions

from hundreds of evaluative adjective pairs.

Effectiveness of Hand-puppet Input: We confirmed that it is very easy to gesture with a hand-puppet interface, which every subject could use in performance. Although the sensing range slightly changes depending on the size of the palm, this configuration could measure the existence and the extent of movement. We consider the sensor signals to be sufficiently accurate because the size of the palm reflects the movement of the stuffed animal. It would also be possible to improve the sensing method by incorporating a calibration function. If only one system were applied for several persons, it would need a device that is adjustable to the size and shape of the palm.

Mappings: As confirmed in the system operation, the mapping strategy reflects the user experience based on the situation of the singing voice expressions; accordingly, a stressed pose of bending all finger back makes a "dark" and high-volume sound, and a lighthearted pose of holding something makes a "wet" sound with "whispering," as we intended. We should adopt not only fixed mappings but also new mapping designs developed by the user as multimodal expressions formed by gestures and singing voice. In the future, we aim to find a method to measure the corresponding feeling and satisfaction of both the performer and the audience.

7. CONCLUSIONS

In this paper, we introduced expressive singing voice morphing by using a hand-puppet interface for natural and personified expression. Our system, HandySinger, has intuitive controls and appearance with the sense of touch experienced within a hand puppet. From the results of the perceptual test, we confirmed that ESVM significantly enriched the expressions of a singing voice through the interpolation of expressional strength. Effective expressions of the singing voice and the appearance gave the user an intuitive experience of pretending with a cute puppet.

As future work of the system implementation, we will examine automatic clustering of both the singing voice and the hand puppet's gestures. For more intuitive musical expressions, it would be useful to control the timing of the utterance of the singing voice and the lyrics by incorporating manipulation of the puppet's mouth.

Acknowledgments

The authors would like to thank Prof. Hideki Kawahara for permission to use the STRAIGHT morphing system. We also thank Dr. Norihiro Hagita, Yoshinori Sakane and other ATR members for their help and discussions on this work. This research was supported in part by the National Institute of Information and Communications Technology of Japan.

8. REFERENCES

[1] Bickmore, Timothy W. and Cassell, J.: "Small talk and conversational storytelling in embodied conversational interface agent," AAAI fall symposium on narrative intelligence, pp. 87–92, 1999.

[2] Imai, M., Ono, T., and Etani, T., "Attractive Interface for Human Robot Interaction," Proceedings of 8th IEEE International Workshop on Robot and Human Communication (ROMAN'99), pp. 124–129, 1999.

[3] Yonezawa, T., Clarkson, B., Yasumura, M., and Mase, K., "Context-aware Sensor-Doll as a Music Expression Device," CHI2001 Extended Abstracts, pp. 307–308, 2001.

[4] Mulder, A., Fels, S., and Mase, K., "Design of Virtual 3D Instruments for Musical Interaction," Graphics Interface, pp. 76–83, June, 1999.

[5] Fels, S. and Hinton, G. E., "Glove-TalkII: A neural network interface which maps gestures to parallel formant speech synthesizer controls," IEEE Transactions on Neural Networks, vol. 8, num. 5, pp. 977–984, 1997.

[6] Cook, P. R., Leider C., "Squeeze Vox: A New Controller for Vocal Synthesis Models," Proc. ICMC2000, pp. 519–522, 2000.

[7] Cano, P., Loscos, A., Bonada, J., Boer, M., and Serra, X., "Voice Morphing System for Impersonating in Karaoke Applications," Proc. ICMC2000, pp. 109–112, 2000.

[8] Sogabe, Y., Kakehi, K., and Kawahara, H., "Psychological evaluation of emotional speech using a new morphing method," CD-ROM Proc. International Conference on Cognitive Science, 114, 2003.

[9] Matsui, H. and Kawahara, H., "Investigation of Emotionally Morphed Speech Perception and its Structure Using a High Quality Speech Manipulation System," Proc. Eurospeech'03, pp. 2113–2116, 2003.

[10] Kawahara, H. and Matsui, H., "Auditory Morphing Based on an Elastic Perceptual Distance Metric in an Interference-free Time-frequency Representation," Proc. ICASSP'2003, vol. I, pp. 256–259, 2003.

[11] Puckette, S. M., "Pure data." Proceedings of the International Computer Music Conference, pp. 224–227, 1997.