Cross-modal Coordination of Expressive Strength between Voice and Gesture for Personified Media

Tomoko Yonezawa[†]
ATR-IRC Labs. / Nagoya University
2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan
yone@atr.jp

Noriko Suzuki[‡]
ATR-MIS Labs.
2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan
noriko@atr.jp

Shinji Abe[†]
ATR-IRC Labs.
2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan
sabe@atr.jp

Kenji Mase
Nagoya University / ATR-IRC Labs.
Furo-cho, Nagoya 464-8601, Japan
mase@itc.nagoya-u.ac.jp

Kiyoshi Kogure[‡‡]
ATR-MIS Labs.
2-2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan
kogure@atr.jp

ABSTRACT

The aim of this paper is to clarify the relationship between the expressive strengths of gestures and voice for embodied and personified interfaces. We conduct perceptual tests using a puppet interface, while controlling singing-voice expressions, to empirically determine the naturalness and strength of various combinations of gesture and voice. The results show that (1) the strength of cross-modal perception is affected more by gestural expression than by the expressions of a singing voice, and (2) the appropriateness of cross-modal perception is affected by expressive combinations between singing voice and gestures in personified expressions. As a promising solution, we propose balancing a singing voice and gestural expressions by expanding and correcting the width and shape of the curve of expressive strength in the singing voice.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

General Terms

Perceptual Experimentation and Design with Cross-modality

Keywords

Cross-modality, Vocal-gestural Expression, Perceptual Experiment, Personified Puppet-interface

1. INTRODUCTION

Personified media such as humanoids, pet robots and androids are being developed with embodied interfaces for natural communication with users. Among the expressions used through these personified media, personification is an effective element for natural, life-like communication with human users [1, 5, 3]. Effective implementations of personification can be found in elderly care facilities that use therapeutic pet robots for familiar interaction with elderly people and language education for infants that employs empathetic puppets. Few reports, however, have discussed the design or effectiveness of the mutual effects among multiple modalities (**Cross-modality**), which mostly involve combinations of mutually expressive modalities such as gesture and voice.

To achieve a natural expression, it is necessary to appropriately combine expressions between modalities and to prevent contradictory or unreasonable feelings. For example, it is unnatural for a robot to make excited gestures while saying "Hello" in an emotionless voice. In that case, the expression would be comparatively natural if the voice were amusingly manipulated with prosodic and vocal sound controls corresponding to the gesture. In this paper, we call the naturalness and suitability of combinations among plural modalities **Appropriateness**. Furthermore, people tend to greatly vary their expressions, changing the strength and type of emotion over time in their everyday interpersonal communications. A natural expression for personified media needs to be attractive while the **Expressive Strength** is controlled, which depends on the content and context of the expression in question.

The purpose of our research is to build effectively personified expression in which the expressive strengths of voice and gesture are controlled in the optimal combination to achieve the desired result. In this paper, we conduct perceptual tests on the appropriateness and the expressive strength of combined voice and gesture in a personified medium. Our goal is to investigate the cross-modality arising among plural modalities in personification, especially with emotive expressions.

To measure the mutual effects of different modalities with simplified elements, the experiments consider only (i) vocal sound and (ii) pose as representative expressions of each voice and gesture. Therefore, we eliminate complicated expressions such as prosody (speech speed, F_0 and so on) and time-sequential gestures. The experiments focus on gestures of the upper body as important expressions in personified robots, as described in [7]. For video stimuli in the perceptual experiments, we have adopted the HandySinger system [18], which is an intuitive and direct controller of the expressive strength of a singing voice and gestures through a

[†]presently also with ATR-MIS
[‡]presently with NiCT / ATR-CIS
[‡‡]presently with ATR-KSL

ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.

hand-puppet interface. The experiments consist of perceptual evaluations of (1) appropriateness and (2) expressive strength of personified expressions for different strengths between expressions of voice and gestures.

2. RELATED WORKS

Personified media such as robotics, puppets and agents are becoming increasingly common in our daily lives [7, 6]. There have been several studies on personification of embodied robotics and puppets as a method of communication, story telling, and musical expression [2, 16, 13]. These expressive systems use their multi-modality nature to gain the strong advantage of involving the user's empathy through familiar personification. To fully exploit this advantage, however, it is important to analyze the perceptual effect of mutual modalities achieved by naturally and appropriately combining expressions.

The McGurk effect is well known as an auditory illusion produced by a visual experience [11]. In this regard, several studies have been conducted on audio-visual mutual effects in human expression [4] and artistic works, [8], showing that mutual expressions do not have the same effect as a singular expression. It is also possible that personified mutual expressions function differently from other types of mutual expressions, such as those described in the related works. Consequently, we attempt to confirm the cross-modal effect in personified expressions in order to produce better design guidelines.

Yamamoto et al. analyzed and applied the timing-control effect between the utterances and physical motions of a robot for natural expression [15]. This research suggests the effectiveness of applying cross-modality between physical motion and voice in personified media.

Personified expression, unfortunately, has problems not only with timing but also with expressive strength, since natural expression includes continuous or discrete changes in strength. To improve the naturalness of personified expression, in this research, we aim to clarify the effect of cross-modality with different levels of strength among expressions of multiple modalities.

3. PERSONIFIED AND MULTIMODAL EXPRESSIVE SYSTEM

3.1 Simplification of Cross-modality

To understand the effect of multiple modalities in personified and embodied media, it is useful to exclude the more complicated elements of each modality. It may be possible to establish a clearer perceptual test for the cross-modal effect by combining only the simplified expressions of each modality. This research, therefore, focuses on the strength of the auditory and visual expressions of a personified interface by using the HandySinger system, which directly manipulates a puppets's gestures through a hand-puppet interface designed for controlling singing voice expressions [18].

In our personified embodied system, we adopted the expressions of 1) upper-body gestures and 2) voice control without complicated expressions such as verbal information.

Gestures provide a non-verbal method of communication using conscious or unconscious movements of the head, hands, and other parts of the body. It is difficult and complicated to classify and interpret time-sequential gestures, so in this research we eliminate the complexity from the puppet's gestures to clarify the effect of cross-modality. Recognizing that the puppet's basic motion and poses are themselves expressions, we adopted the pose of the hand-puppet to analyze its mutual effect with voice expression.

Table 1: Varied Expressions for Personified System

Channel	Expression	Feature	abbr.
Singing Voice	normal	Flat (without Expression)	no
	dark	Closer to Back Vowel	da
	whispery	High-frequency Noise	wh
	wet	Nasal Voice	we
Pose (Gesture)	neutral	Flat (without Expression)	neu
	back	Bend Backward	bak
	droopy	Drooping Head	drp
	stretched	Stretch Forward Hands	str

neu bak drp str

Figure 1: Four Gesture-Expressions

On the other hand, expression in spoken utterances includes not only vocal sound but also prosody information such as speech speed and F_0. To observe the mutual effect of voice and gesture, it is important to remove the effects of these complicated elements to obtain the correct impression of the main element of vocal expression. In this paper, vocal expression includes only vocal sounds, eliminating the prosodic effect in order to evaluate cross-modal effects. Since spoken vocal expressions do not exist in normal speech without prosody, we adopt a singing voice that is only capable of producing expressions by vocal sounds, particularly through expressive strength.

3.2 System Structure

Simplifying the element of expressive control as explained in Section 3.1, we construct a singing-voice expression system with a hand-puppet interface to investigate the cross-modal effect. Four types of expression are prepared for each modality as a first step toward building this system's structure (Table 1 and Figure 1).

As Figure 2 shows, the system consists of the following processes:

- Employing internal sensors to detect the expressive type and strength of the hand-puppet's gestures from the user's direct controls;
- Determining the expressive strength and type of singing voice according to the detected gestures; and
- Synthesizing and outputting an expressive singing voice.

The following sections provide details of the functions listed above.

3.2.1 Sensing Gestural Expressions

To directly control the singing-voice expression via hand-puppet gestures, we adopted as the system interface a simple hand-puppet consisting of controllable parts for each of the two arms and the head. The user controls the hand-puppet with three fingers, using, for example, the thumb, forefinger, and middle finger of her/his right hand.

To capture the hand's motions as motions of the puppet itself with sufficient accuracy for the interface, we constructed

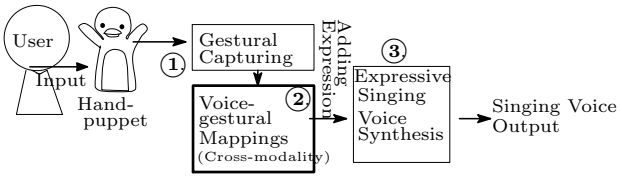

Figure 2: Synthesis of Singing Voice Expression by Hand-puppet

Figure 3: External and Internal Appearance of the Hand-puppet

a sensor-covered glove (Figure 3) as an independent capturing device and fitted it inside a stuffed penguin, which functioned as the personifying cover. The sensors' analog signals are sent to an A/D converter (Infusionsystems I-CubeX) and converted to MIDI signals. A PC (Windows XP) receives them thorough a MIDI interface (Roland UM-2).

3.2.2 Basic Mapping between Voice and Gestures

It is important to construct a simple mapping between voice and gesture expression as a prototype for studying the cross-modal effect. As a mapping of expressive types, the system is preset with configured combinations between voice and gestural expression based the simple metaphor shown in Table 2.

The mapping of expressive strength is a simple correspondence between voice and gestures. To detect the expressive strength of gestures from the sensor signals, we defined the baseline of gestures without expression ("neu") as 0 and the baseline with expression as 1. The system detects the expressive strength by comparing the differences of signals from the baselines. Accordingly, the mapping determines the expressive strength of the singing voice through the expressive strength of gestures from user inputs.

The origin and range of each sensor are calculated to normalize their values as weights of the expressions. As a result of these mappings, neutral gestures, without any power, are mapped with flat expression of the singing voice as the origin of each expression. For example, when the puppet makes an intermediately drooping gesture, the singing voice corresponds to the strength of the gesture.

3.2.3 Expressive Singing Voice Synthesis

We synthesized a singing voice with various levels of expressive strength by using a speech-morphing algorithm [10] with STRAIGHT [9]. This morphing synthesis was made using a real singing voice, with varying degrees of expressiveness from no expression to high expression.

We recorded the singing voice of a female amateur singer in her twenties at a sampling frequency of 44.1 kHz. The singer sang a Japanese nursery rhyme, "Furusato" ("Hometown"), with an accompaniment in which singing speed and F_0 were arranged in the same way. The instructions were (1) to sing flat without any expression ("no" in Table 1)

Table 2: Correspondence of Expressions between Singing Voice and Gestures

abbr.	Voice	Gesture	Metaphor
A	da	bak	Utterance like Opera Singer
B	wh	drp	Drooping Pose matching Hoarse Voice
C	we	str	Emphasis like Pop Singer
(	no	neu	Relaxed and Flat Utterance)

and with three different expressions ("da," "wh," and "we"), with the features given in Table 1, and (2) to sing with a consistent expression during each singing sample.

The singing voice expression with interpolated strength was synthesized with a speech-morphing algorithm between {no↔da, no↔wh, and no↔we}. The expressive strength of the singing voice was controlled by the morphing ratio, which is the ratio of expressive features, where "no" is 0 and "{da, wh, or we}" is 1. The singing voice was synthesized by a program using PureData [12].

4. PERCEPTUAL EXPERIMENT FOR CROSS-MODALITY

We designed and conducted perceptual experiments on cross-modality between the expressive strength of voice and gestures, using a personified puppet while assuming that there is an appropriate combination of expressive strengths. The experimental results reveal in particular: (i) the basic characteristic of expressed strength and perceived strength of mono- or multi-modal expression; and (ii) the effects of perception on (a) appropriateness and (b) expressive strength by combining different strengths of voice and gestural expressions.

The experiments evaluated perception using several video stimuli of the HandySinger system (Section 3), since it is difficult to maintain repeatability of the puppet's motions by using a robot due to the state of present control techniques.

This section describes the three experiments we conducted: to clarify the basic perceptual characteristics of expressive strength (Experiment 1, Section 4.1); to identify the effects of voice and gestural expressions (Experiment 2, Section 4.2); and to confirm the same effect of Experiment 2 in dynamical conversion (Experiment 3, Section 4.3).

4.1 Experiment 1: Perceptual Strength of Interpolated Expression

Since it has already been confirmed that variously interpolated strengths of singing-voice expression (v) are perceived continuously [14, 17], in the experiments below we attempt to clarify perceptual interpolation in both gestural expressions (g) and cross-modal expressions with the singing voice and gestures (v+g).

Hypotheses: I. Gradually changing the strength of gestural expressions makes it possible to interpolate perceptual strength; and II. By gradually strengthening vocal-gestural expressions, we can also interpolate perceptual strength just as we do with singing-voice expression.

Subjects: Twenty people aged from their twenties to low thirties (nine females and eleven males) participated in the experiments.

Environment: Subjects executed a perceptually evaluative program while wearing headphones in a soundproof room in

Table 3: R^2 Values of Linear Regression Analyses of Perceptual Interpolation

abbr.	A	B	C
g	R^2= .718	R^2=.811	R^2=.790
$v+g$	R^2=.795	R^2=.831	R^2=.818

front of a 50-inch plasma display, which showed the puppet at it's actual size, about 20 cm tall. The program described in Tcl/TK obtained evaluative input from the subjects and displayed the video of the HandySinger puppet using QuickTimeTcl3.1.

Methods: Subjects watched a video stimulus with expressions and a video stimulus without expressions to set the evaluative criteria, and then they evaluated the expressive strength of each video stimulus on a seven-level scale.

[Test 1-I]: Subjects evaluated various silent videos containing gestures made by the puppet at various strength levels (g). The test included the three types of expression listed in Table 1: {bak(Test 1-I-A), drp(I-B), and str(I-C)}.

[Test 1-II]: The subjects evaluated various videos with a singing voice and puppet gestures at various levels of expressive strength ($v+g$). The test included the three expressive combinations listed in Table 2: {da:bak(Test 1-II-a) wh:drp(II-b), we:str(II-c)}. The difference between Test 1-I and 1-II is that a singing-voice expression was or was not present.

Procedures and Conditions (Video Stimulus): The video stimulus movies were combined with a picture of the puppet's static pose expression and a singing voice at various strength levels. To control the subjects' impression when the the sound suddenly occurred, the movie began about 0.2 sec before and finished about 0.2 sec after the singing voice file, which was about 3.0 sec in Test 1-II.

We informed the subjects of the expressions with abbreviated labels {A, B, C} in Table 2 without conveying images reflecting adjectives. To confirm the continuity of perceptual strength in gestural or vocal-gestural expressions, we prepared seven levels of expressive strength in the video stimuli.

g: We presented still images for about 3.4 sec at each of the seven levels of expressive strength from a movie of the puppets' motions, ranging from the pose "neu" (without expression) as 0 up to the expressive pose "{bak, drp, or str}" as 1.

$v+g$: Still images, just as in g, were combined with the singing voice at corresponding strengths of expression. The singing voice was also synthesized to vary its strength over the seven levels.

Still images were extracted at each level of pose strength {0.00, 0.17, 0.33, 0.50, 0.67, 0.83, 1.00} detected by the sensor, and the singing-voice sounds for the video stimuli were morphed at the morphing ratio of {0.00, 0.17, 0.33, 0.50, 0.67, 0.83, 1.00}.

Instructions: The subjects were instructed to watch the video stimuli to establish the criteria for each experiment. For example, in Test 1-I-A, the instructions were to (i) view "neu" as expressive strength 0 and "bak" as expressive strength 1 in g, and (ii) judge the expressive strengths of the video stimuli on the seven-level scales.

Results: Figure 4 shows the results of MOS (Mean Opinion Score) as averages and standard deviations, and Table 3 shows R^2 values as linear approximations. Both g and $v+g$ increased in expressive strength, supporting both hypotheses I. and II.

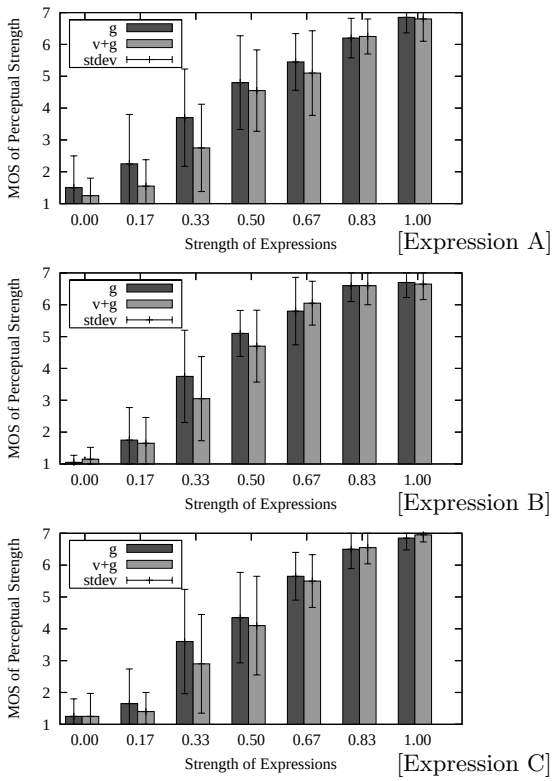

Figure 4: Perceptual Strength for Gradual Expression [Experiment 1]

Table 4: ANOVA: Perceptual Strength by Modality and Strength

item	abbr.	F-Value	p-Value
Modality	A	$F_{(1,38)}$= 3.46	$p = .07 \leftarrow$
	B	$F_{(1,38)}$= .93	$p = .93$
	C	$F_{(1,38)}$= .91	$p = .35$
Expressive Strength	A	$F_{(6,38)}$= 189.96	$p < .01$
	B	$F_{(6,38)}$= 330.82	$p < .01$
	C	$F_{(6,38)}$= 265.85	$p < .01$
Interaction (modality× strength)	A	$F_{(6,38)}$= 1.34	$p = .24$
	B	$F_{(6,38)}$= 1.61	$p = .14$
	C	$F_{(6,38)}$= .942	$p = .47$

Through the experiment with Expression A to C in a morph ratio from 0.17 to 0.50, the results of g are always higher than those of $v+g$. At low expressive strength, combining voice and gesture was always found to increase perceived expression.

On the other hand, it is possible that the difference in modalities affects perceptual evaluation since $v+g$ R^2 values were observed to be greater than g. Table 4 shows the results of analyses of variance (ANOVA) with repeated measurement of fourteen conditions and two dimensions (difference in modality: g and $v+g$ and difference in expressive strength) at $\alpha = 0.05$ and $\phi = (1, 6, 38)$ (however, α: significant level, ϕ: degree of freedom.) Except for the slight trend of significance observed for Expression A ($F = 3.46$, $p = .07$), the results indicate that the modality causes no

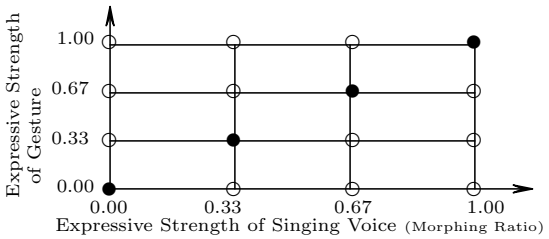

Figure 5: Conditions of Cross-modal Perception Test

significant difference, in contrast to the expressive strength, which causes a significant difference in the MOS of every expression.

In this experiment, we could confirm a basic characteristic of perceiving expressive strength. We could not, however, observe a stabilizing effect on MOS from the differences in modality and interactions (modality × expressive strength) compared with the expressive strength.

4.2 Experiment 2: Cross-modal Effect by Combining Various Expressive Strengths

In this experiment we aimed to clarify the mutual effect between the singing voice and gestures at various levels of expressive strength. We observed the perception of appropriateness and expressive strength in vocal-gestural combinations at each of the strengths of expression shown in Figure 5.

Hypotheses: I. The corresponding strength between expressions of the singing voice and poses is perceived to be highly appropriate; and II. The perceived strength is the average of voice and pose strength.

Subjects: Same subjects as in Experiment 1.

Environment: Same environment as in Experiment 1.

Methods: The sixteen video stimuli, containing four levels of strength for each singing-voice expression and pose expression, were evaluated in random order (Figure 5). The subjects evaluated the videos in relation to: (1) expressive appropriateness between the singing voice and pose (Test 2-I), and (2) the general perception of expressive strength (Test 2-II) on the seven-level scale. The subjects watched "no:neu," representing expressive strength 0, and "da:bak" or other expressions denoting expressive strength 1, with the criteria set as the same used in Experiment 1.

Procedures and Conditions (Video Stimulus): To investigate two modalities of expressive strength, we prepared each voice and pose at four levels of strength as shown in Figure 5. Singing voices and still images of poses at various levels of strength were combined in each expression {A(da:bak),B(wh:drp),C(we:str)}. Still images were extracted at expressive strengths of {0.00, 0.33, 0.67, 1.00} using sensor values in the same way as in Experiment 1, and singing voices were synthesized at the morphing ratio of {0.00, 0.33, 0.67, 1.00}.

Instructions: Subjects were instructed (i) to watch the no:neu and {da:bak, wh:drp or we:str} videos before each test to confirm the criteria, and (ii) to evaluate the videos according to the seven-level scale for: (I) appropriateness of multi-modal expressions between voice and gestures; and (II) expressive strength. We gave them a definition of "appropriateness" as "matching the sense of expression between voice and gestures."

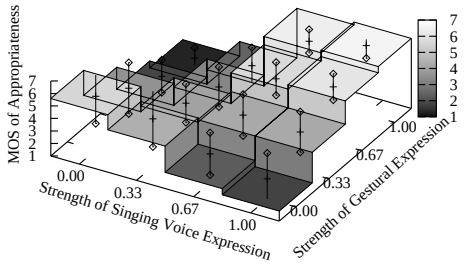

Figure 6: Appropriateness at Different Strength Levels [da:bak, Expression A]

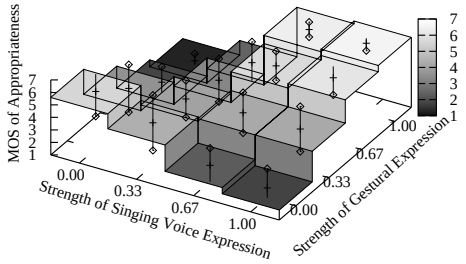

Figure 7: Appropriateness at Different Strength Levels [wh:drp, Expression B]

Results I. [Perception of Appropriateness]: Figures 6, 7, and 8 show averages and standard deviations of perception tests. Although the combinations of corresponding strengths of singing voice and gestural expression seem to provide the highest appropriateness at first glance, stronger expressions of pose lead to a higher MOS value of appropriateness at {0.67, 1.00} of the expressive strength of singing voice but to a lower MOS value at {0.00, 0.33}.

It is conceivable that cross-modality affects the perception of expressive strength, although the highest appropriateness is not always at corresponding strengths between voice and pose. For example, the MOS of appropriateness was the highest at expressive strength 1.00 of the pose for expressive strength 0.67 of the singing voice.

To clarify the cross-modal effects, we conducted an ANOVA statistical test with repeated measurement of two dimensions (voice expressions and gestural expressions), $\alpha = 0.05$ and $\phi = (3, 3, 76)$, while setting the null hypothesis as "the perception of appropriateness is not affected by differences in the expressive strengths of voice, gestures, or their mutual effect."

As Table 5 shows, the ANOVA results mostly represent significant differences. The exceptions are Expression A with the strength of pose "bak" ($F = .12$, $p = .95$) and Expression C with the strength of singing voice expression we ($F = 2.03$, $p = .12$). Therefore, the null hypothesis can essentially be rejected. The results suggest that the perception of appropriateness between singing voice and gestures is strongly affected by the expressive strength of singing voice, the expressive strength of gesture, and the interaction between them.

Results II. [Perception of Expressive Strength]: Figures 9, 10, and 11 show the averages and standard deviations from the perception tests. The figures show the strong effect of expressive strength of pose in comparison to the general perception of multi-modal expressive strength.

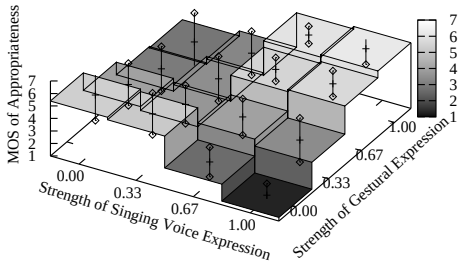

Figure 8: Appropriateness at Different Strength Levels [we:str, Expression C]

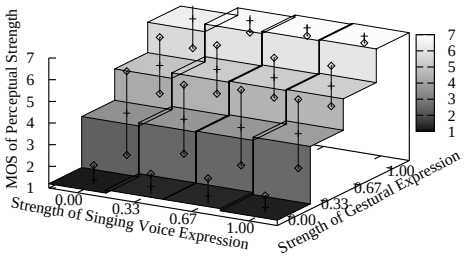

Figure 9: Perceptual Strength at Different Strength Levels [da:bak, Expression A]

Table 5: ANOVA: Test of Appropriateness with Two Factors

abbr.	item	F-Value	p-Value
A	da	$F_{(3,76)}= \underline{11.18}$	$p < \underline{.01}$
	bak	$F_{(3,76)}= .12$	$p = .95$
	Interaction	$F_{(9,76)}= \underline{23.12}$	$p < \underline{.01}$
B	wh	$F_{(3,76)}= \underline{20.84}$	$p < \underline{.01}$
	drp	$F_{(3,76)}= \underline{3.96}$	$p < \underline{.01}$
	Interaction	$F_{(9,76)}= \underline{36.99}$	$p < \underline{.01}$
C	we	$F_{(3,76)}= 2.03$	$p = .12$
	str	$F_{(3,76)}= \underline{10.51}$	$p < \underline{.01}$
	Interaction	$F_{(9,76)}= \underline{32.66}$	$p < \underline{.01}$

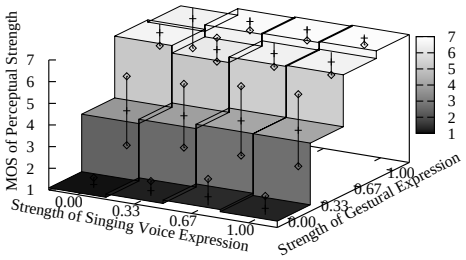

Figure 10: Perceptual Strength at Different Strength Levels [wh:drp, Expression B]

To clarify the strength of effects due to voice, pose, and mutual cross-modality, we conducted an ANOVA statistical test with the repeated measurement of two dimensions (voice expression and gestural expressions), with $\alpha = 0.05$ and $\phi = (3, 3, 76)$, setting the null hypothesis as "the perception of general expressive strength is not affected by differences in the expressive strength of voice, gestures, or their mutual effect."

Table 6 presents the ANOVA results. The test results show significant differences in the averages due to differences in expressive strength of poses for all types of expression ($F = \underline{394.38}$ in A by expressing "bak", $F = \underline{575.99}$ in B by expressing "drp", and $F = \underline{403.85}$ in C by expressing "str"). In contrast to the above results, there was no significant difference in the other results. Thus the null hypothesis concerning the pose is rejected although those of the singing voice and the mutual effect are not rejected.

These results suggest that the general perception of expressive strength is strongly affected by the expressive strength of gestures. They also indicate that the expressive strength of the singing voice and the mutual action between voice and gestures do not affect the general perception of expressive strength, which is different from what the alternate hypotheses put forward.

4.3 Experiment 3: Adequate Perception with Dynamically Expressive Change

Multi-modal expressions in real human communication momentarily change in strength. For a personified medium to express itself naturally, it is important to clarify the cross-modal effect in regard to dynamic change. In this experiment we aim to confirm the cross-modal effect on the appropriateness between voice and gestures with dynamic changes in expressive strength, just as we confirmed the cross-modal effect in the static condition in Experiment 2-I.

Hypothesis: The perceptual appropriateness between voice and gesture is higher when the expressive strength changes corresponding to the change in another factor, and the appropriateness is lower when the expressive strength values of voice and gestures do not correspond.

Subjects: Same subjects as in Experiments 1 and 2.

Environment: Same environment as in Experiments 1 and 2.

Methods: We prepared the video stimuli for multi-modal expressive strength accompanying changes in the correct (forward) sequence and in the reverse sequence between voice and gestures. The subjects evaluated the appropriateness of the video stimuli according to the seven-level scale, in random order.

Procedures and Conditions (Video Stimulus): The video stimuli include two types of combinations of expressive strength: (i) pose changes from 0 to 1 when the singing voice changes from 0 to 1; and (ii) the reverse order of (i) with respect to the singing voice. Considering the order effect, we also prepared video stimuli in which: (iii) pose changes from 1 to 0 when the singing voice changes from 1 to 0; and (iv) the reverse order of (iii) with respect to the singing voice (Table 7). These stimuli are understood to reflect changes in expressive strength along the diagonal slopes of Figs. 6 to 8.

Instructions: We defined "appropriateness" as "the matching sense of expressions between voice and gestures" as the criterion for measurement. Subjects were instructed to evaluate the appropriateness of the video stimuli on the seven-level scale based on this criterion.

Results: Figure 12 shows the results with "-" marks indicating the reverse case.

To confirm the mutual effects of voice and pose with dynamic changes in expressive strength, we conducted a T test with repeated measurement ($\alpha = 0.05$, $\phi = 19$), setting the null hypothesis as "the perception of appropriateness is not affected by differences in the combination order of expres-

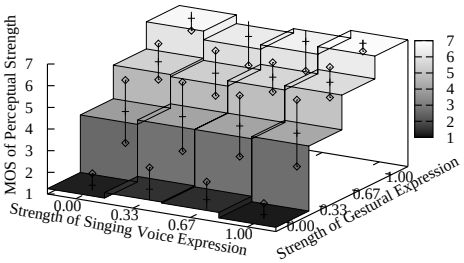

Figure 11: Perceptual Strength at Different Strength Levels [we:str, Expression C]

Table 6: ANOVA: Test of Perceptual Strength with Two Factors

abbr.	item	F-Value	p-Value
A	da	$F_{(3,76)}$= .83	p = .48
	bak	$F_{(3,76)}$= 394.38	$p <$ .01
	Interaction	$F_{(9,76)}$= .14	p = .999
B	wh	$F_{(3,76)}$= 1.12	p = .35
	drp	$F_{(3,76)}$= 575.99	$p <$ .01
	Interaction	$F_{(9,76)}$= .38	p = .94
C	we	$F_{(3,76)}$= .48	p = .70
	str	$F_{(3,76)}$= 403.85	$p <$ .01
	Interaction	$F_{(9,76)}$= .64	p = .76

sive strength between voice and pose" for each Expression from A to C.

The T values for the tests of Expressions A to C were {-5.572, -4.887, and -7.406}, and the p values were all $p <$ .01. Consequently, the null hypothesis was rejected, indicating that the difference in the combination order was significant.

These results suggest that the perceptual appropriateness is high at corresponding strengths of vocal-gestural expressions, even when expressive strength dynamically changes. We thus confirmed the hypothesis to be correct in this experiment.

5. DISCUSSION

5.1 Vocal-gestural Balance in Relation to Perceptual Strength

We first confirmed the correspondence between the strength of perception and the expressive strength of gestures in the video stimulus (Test I in Experiment 1). Test II in Experiment 1 revealed the correspondence between perception and the the strength of multiple voice and gesture expressions, just as Test I did. These results are premised on the perception of expressive strength in multi-modal or uni-modal expression, following experiments on cross-modality.

Considering the effects of voice and gestures, it is remarkable that R^2 values from the linear regression in Experiment 1 are higher in v+g than in g, although the statistical test results do not indicate any significant difference due to the difference in modalities. The curve representing the expressive strength in g appears as a raised arc on the upper side in comparison with v+g. Meanwhile, the perceptual characteristic of the expressive strength in the singing voice is approximated as a sigmoid function [17]. In comparison with g, it is possible that v+g is more strongly affected by the curve of expressive strength of the singing voice.

The results of Test II in Experiment 2 indicate a strong

Table 7: Example of Combinations in Cross-modal Expressive Strength

Combination	Strength of Singing Voice Expression	Strength of Gestural Expression
Correct Sequence	0→1	0→1
	1→0	1→0
Reverse Sequence	0→1	1→0
	1→0	0→1

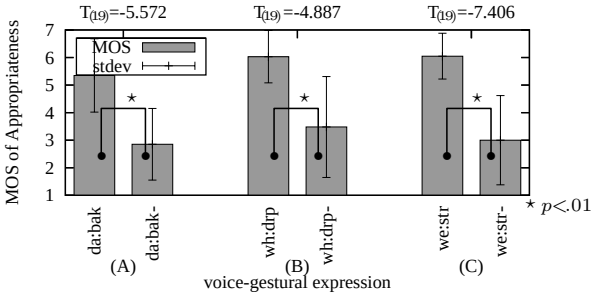

Figure 12: Appropriateness at Different Strengths with Change

effect of gestures compared to the effects of the singing voice and their mutual effect. In other words, visual expression is stronger than auditory expression, and the balance of the effect between modalities is not the same in personified expressions. Considering this perceptual phenomenon, we suggest adjusting the variational width of strength based on the results of the cross-modal effect, bearing in mind the balance between the strengths of voice and gestures.

5.2 Vocal-gestural Balance in Relation to Perceptual Appropriateness

In contrast to the perception of expressive strength strongly affected by gestures, the ANOVA of Test I in Experiment 2 shows there is a significant difference in the perceptual appropriateness due to the interaction between the vocal and gestural expressions. In addition, Experiment 3's results show that the combinations of correct/reverse sequences of expressive strength cause significant differences in appropriateness. These results mean that the combination of expressive strength in the singing voice and gestures is important for attaining the appropriateness of multi-modal expressions, although the perception of expressive strength is preferentially affected by gesture in our simplified cross-modal investigation.

When the expressive strength of the singing voice is neither 0.00 nor 1.00, the expressive strength of the gesture is appropriate for whichever of 0.00 or 1.00 is nearest (Test I of Experiment 2). That is to say, the results do not necessarily mean that all combinations of corresponding expressive strength between the voice and gestures are perceived to have the highest appropriateness. The results suggest the possibility of emphasized perception in moderate expression of the singing voice, such as 0.33 and 0.67, since the curve of expressive strength of the singing voice was approximated in a sigmoid function.

Consequently, for constant correspondence between voice and gestural expressions, it might be effective to compensate for the curve of expressive strength of the singing voice. We suggest converting the present non-linear correspondence by

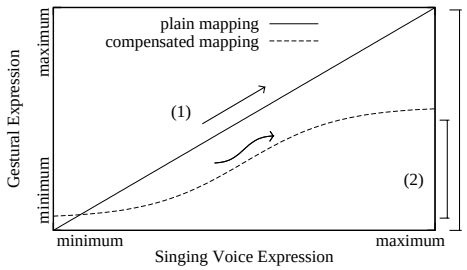

Figure 13: Suggestion for Correspondence in Expressive Strength between Voice and Gestures

setting the morphing ratio with a reciprocal sigmoid function in the same way that the expressive strength of the singing voice is linearized.

5.3 Vocal-gestural Mapping Strategy from the Perception Tests

We believe that it is suitable to use the compensation described in Section 5.2 to expand the variational width of the expressive strength of the singing voice against that of gestures, as suggested in Section 5.1, instead of simply expanding it linearly, in order to achieve a multi-modal expression that is harmonized and natural.

This suggestion is simply shown with plain and compensated mappings in Figure 13, where we (1) control the shape of the corresponding curve between expressions of voice and gestures, and (2) adjust the variational width of the expressive strength between the voice and the gesture. Detailed experiments are needed to build an accurate model of this adjusted mapping strategy.

6. CONCLUSIONS

In this paper, we described the perceptual evaluation of expressive strength in multi-modal expressions, applying a video stimulus created using the HandySinger system. This system provides a method for personified expression using a hand-puppet interface with simplified voice and gesture expressions. From the experimental results, we conclude that: (i) the expressive strength of a gesture more strongly affects the perception of its strength in multiple expressions comprising voice and gestures than just the strength of the voice; and (ii) the combination of the strength of the voice and gesture strongly affects the perception of the appropriateness in both static and dynamic multi-modal expressions. Finally, we suggested the possibility of achieving more natural expression by balancing the perceptual strength of expression in each modality, that is, properly adjusting the expressive strength of the singing voice against that of the gestures.

As future work, we plan to carry out a detailed investigation of how cross-modality should be validated for practical use in personified media by applying our elementary results as a basis. To achieve natural and life-like expressions in personified media, it is important to apply and examine the expressive characteristics of various modalities, not only voice and gesture.

7. 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, Takayuki Kanda and other ATR members for their help. This research was supported in part by the National Institute of Information and Communications Technology of Japan.

8. REFERENCES

[1] J. Bates. The role of emotion in believable agents. *Communications of the ACM*, pages 122–125, 1994.

[2] T. W. Bickmore and J. Cassell. Small talk and conversational storytelling in embodied conversational interface agent. *AAAI fall symposium on narrative intelligence*, pages 87–92, 1999.

[3] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. *Embodied Conversational Agents*. MIT Press, 2000.

[4] T. Chen and R. R. Rao. Audio-visual integration in multimodal communication. *Proc. IEEE*, 86(5):837–852, 1998.

[5] B. Duffy. Anthropomorphism and the social robot. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2002.

[6] M. Fujita and K. Kageyama. An open architecture for robot entertainment. *Proc. the First International Conference on Autonomous Agents*, pages 435–442, 1997.

[7] M. Imai, T. Ono, and T. Etani. Attractive interface for human robot interaction. *Proc. of 8th IEEE International Workshop on Robot and Human Communication (ROMAN'99)*, pages 124–129, 1999.

[8] S. Iwamiya. Multimodal communication by music and motion picture. *Proc. of 7th International Conference on Music Perception and Cognition*, pages 3–8, 2002.

[9] H. Kawahara, I. Masuda-Kasuse, and A. Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a reptitive structure in sounds. *Speech Communication*, 27:187–207, 1999.

[10] H. Kawahara and H. Matsui. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. *Proc. ICASSP'2003*, I:256–259, 2003.

[11] H. Mcgurk and M. Lewis. Space perception in early infancy: perception within a common auditory-visual space. *Science*, 186:649–650, 1974.

[12] S. M. Puckette. "pure data". *Proc. ICMC1997*, pages 224–227, 1997.

[13] D. Sekiguchi, M. Inami, and S. Tachi. Robotphone: Rui for interpersonal communication. *CHI2001 Extended Abstracts*, pages 277–278, 2001.

[14] Y. Sogabe, K. Kakehi, and H. Kawahara. Psychological evaluation of emotional speech using a new morphing method. *4th ICCS International Conference on Cognitive Science*, 2003.

[15] M. Yamamoto and T. Watanabe. Timing control effects of utterance to communicative actions on embodied interaction with a robot. *Proc. IEEE Workshop on Robot and Human Interactive Communication*, pages 467–472, 2004.

[16] T. Yonezawa and K. Mase. Musically expressive doll in face-to-face communication. *IEEE Proc. International Conference of Multimodal Interfaces*, pages 417–422, 2002.

[17] T. Yonezawa, N. Suzuki, K. Mase, and K. Kogure. Gradually changing expression of singing voice based on morphing. *Proc. Interspeech2005*, pages 541–544, 2005.

[18] T. Yonezawa, N. Suzuki, K. Mase, and K. Kogure. Handysinger: Expressive singing voice morphing using personified hand-puppet interface. *Proc. NIME2005*, pages 121–126, 2005.