Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 23807, 9 pages
doi:10.1155/2007/23807

Research Article

Perceptual Continuity and Naturalness of Expressive Strength in Singing Voices Based on Speech Morphing

Tomoko Yonezawa,[1,2,3] Noriko Suzuki,[4] Shinji Abe,[1,3] Kenji Mase,[2,1] and Kiyoshi Kogure[5]

[1] *ATR Intelligent Robotics and Communication Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan*
[2] *Nagoya University, Furo-cho, Chikusa, Nagoya 464-8601, Japan*
[3] *ATR Media Information Science Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan*
[4] *National Institute of Information and Communication Technology/ATR Cognitive Information Science Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan*
[5] *ATR Knowledge Science Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan*

Received 30 November 2006; Revised 23 April 2007; Accepted 17 August 2007

Recommended by S. Voran

This paper experimentally shows the importance of perceptual continuity of the expressive strength in vocal timbre for natural change in vocal expression. In order to synthesize various and continuous expressive strengths with vocal timbre, we investigated gradually changing expressions by applying the STRAIGHT speech morphing algorithm to singing voices. Here, a singing voice without expression is used as the base of morphing, and singing voices with three different expressions are used as the target. Through statistical analyses of perceptual evaluations, we confirmed that the proposed morphing algorithm provides perceptual continuity of vocal timbre. Our results showed the following: (i) gradual strengths in absolute evaluations, and (ii) a perceptually linear strength provided by the calculation of corrected intervals of the morph ratio by the inverse (reciprocal) function of an equation that approximates the perceptual strength. Finally, we concluded that applying continuity was highly effective for achieving perceptual naturalness, judging from the results showing that (iii) our gradual transformation method can perform well for perceived naturalness.

1. INTRODUCTION

The familiar impression of personified media is well received by humans in social interaction, as discovered in animal and puppet therapy in caring for the elderly and traumatized children. To maintain the illusion of personification [1], it is absolutely imperative to develop a natural and human-like expressive method for personified robots and agents. In the same way that android research [2, 3] has led to the development of highly human-like skin and faces for natural appearances, there have also been various approaches to achieving natural personification in vocal, gestural, and facial expressions. We focused on the need for a realistic and natural voice synthesis method at the same time that other projects pursued the personification of robots and agents. For example, it is desirable to make human-like voices for human-like agents having various expressions in other multimodal channels. One of the critical expectations placed on personified robotics and agents is the ability to produce an expressive voice, so naturalness of voice is a key design challenge that must be met. A naturally expressed voice should be able to gradually transform expressions in a variety of ways.

Speech expressions have various parameters, such as speech speed, tone pitch, intensity, and tone timbre, which have perceptual characteristics consisting of spectrum features. To control these expressions, we should investigate each parameter through its independent evaluation. In particular, we stress the importance of vocal timbre for such applications. Although a singing-voice expression has various parameters, singers sometimes need to represent their expressions by vocal timbre alone, since the musical tone is set by the musical piece's score. Focusing on the expressive strength of vocal timbre in a singing voice, this paper investigates the continuity and naturalness of the voice's perceptual strength with the aim of achieving the capability to synthesize human-like vocal timbre. We developed a method for synthesizing a singing voice by gradually changing the musical expression based on speech morphing. To show the advantages

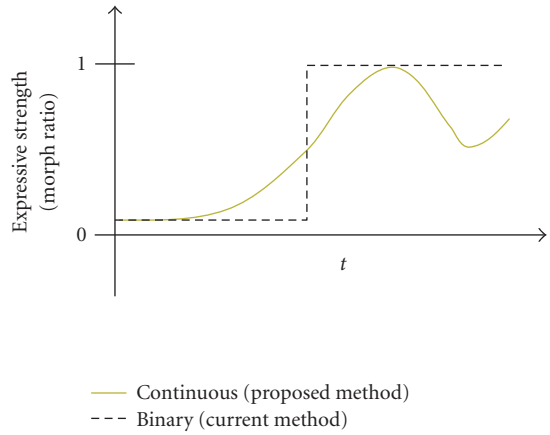


Figure 1: Advantage achieved with morphing expression.

of this method in comparison with the approach of binary discrete transformation between two different expressions, we present statistical analyses of various perceptual evaluation studies on the expressive strength of tone timbre in a singing voice.

Approaches were used to develop speech synthesis capable of emotional expressions with each particular database of various sizes and types. For example, rule-based methods have smaller databases, in contrast to corpus-based methods. The former mainly involve signal processing, including prosodic approaches, while the latter combine speech materials aiming at natural vocal timbre, although the quality of the generated sound depends on the corpus. In order to produce various and continuous expressive strengths of a singing voice with corpus-based methods, we employed a speech morphing algorithm for a singing voice without expression as the base of morphing and singing voices with three different expressions as the target.

Figure 1 shows an example of the anticipated advantages of our approach. In attempting to generate a continuous expression corresponding to human perception of expressive strength, the current corpus-based voice synthesis can provide only a binary discrete expression, which depends on the categories of vocal expression in existing databases. In contrast to the discontinuous transformation by the current method, our proposed approach can provide naturally continuous expression by controlling the morphing ratio.

In this paper, we explain a method for synthesizing various expressive strengths of vocal timbre by adopting a speech morphing algorithm [4], which uses STRAIGHT [5], for singing-voice expressions. Furthermore, we confirm the effectiveness of both interpolation and extrapolation of the singing voice. We then approximate the curve of the expressive strength through subjective evaluations for perceptually linear synthesis of the singing-voice expressions. Finally, we discuss the advantage of the gradual transformation's naturalness of expression through a comparison with binary discrete transformation.

Section 2 describes related works and the context of this research, and Section 3 explains the interpolation method for singing-voice expression using a speech morphing algorithm with a simple model of continuous perception of expressive strength. Section 4 verifies the perceptual naturalness and the differences among synthesized materials. These findings provide the premise upon which we base the subsequent experiments. Section 5 presents a perceptual evaluation of our proposed method of handling the continuity of the expressive strength in a singing voice. Section 6 evaluates the naturalness of the gradual expression of morphed sound based on verified perceptual continuity. In Section 7, we discuss the perceptual model of expressive strength in a singing voice through perceptual evaluations of continuity, a method of linearization, and perceived naturalness. Finally, we conclude our paper in Section 8.

2. RELATED WORKS

2.1. Database size and synthesis processes

Recent approaches to speech synthesis with emotional expressions consisting of vocal timbre and other parameters use particular databases with various sizes and types. Each polar system can be roughly categorized into two methods: rule-based methods such as those explored by Schröder [6] and Erickson [7], and corpus-based methods such as CHATAKO [8] incorporating various data-collection methods such as [9]. While rule-based methods mainly address prosody controls, corpus-based methods can cover expression of vocal timbre with various databases. However, there are gaps between the different expressions or emotions achieved by each corpus when the database has multiple and discrete expressions as in the case of CHATAKO. For natural vocal timbre synthesis, both intermediate expressions and dynamic change of the expression by time should be obtained.

To solve this problem, we have developed a method for gradually changing an expression by using speech morphing. By focusing on the effectiveness of using vocal timbre without prosodic features, our approach employs speech morphing in a corpus-based method. Prosodic approaches such as the F_0 control technique proposed by Saitou et al. for natural singing voice synthesis [10], are expected to be strong tools in combination with vocal timbre control based on perceptual models.

2.2. Emotional speech morphing

Saitou et al. [11] introduced a method for converting speech to a singing voice. Although it is an efficient method in the corpus-based approach, reforming F_0 in the singing voice with a model poses the risk of synthesizing a nonexistent voice of the singer since the "singing-ness" includes personality. To use a natural and realistic singing voice to the extent possible, we perform singing-voice morphing among real singing voices, which proves to be ideal for evaluating the naturalness and continuity of expressive strength. Cano et al. [12] proposed a karaoke system for singing-voice morphing between different singers, from the user's voice to the voice of a professional singer. Although they employ singing-voice

Table 1: Expression types in recorded voices.

Singing instruction	Label of expression
Expressionless	"normal"
Operatic voice by uttering like back and rounded vowels	"dark"
Including more breathy and aspirated voice	"whispery"
Entirely nasal voice	"nasal"

morphing similar to our approach, these research efforts in sound morphing produced a new synthesized sound of a nonexistent singer. In contrast, our research aims to vary and smooth out the expression in the voice of a particular singer by using only that singer in the morphing process.

Sogabe et al. [13] and Matsui and Kawahara [14] investigated the sound morphing of Japanese emotional speech by a particular speaker. Mareüil et al. [15] also employed speech morphing for emotional speech in various languages. Accordingly, emotional speech morphing is now one of the primary methods of synthesizing middle-emotive expressions. These approaches involve using different values of speech speed and F_0 along with the emotional expressions. In contrast, our research aims to vary and smooth out the expression in a voice by using a singing voice at the same speed and F_0.

3. MORPHING METHOD FOR EXPRESSIVE SINGING VOICE

To control the expressive strength of tone timbre in voice, we propose a method of morphing speech from a flat singing voice to an expressive singing voice [16]. It is possible to synthesize the voice parameters, but we focused on vocal synthesis from the existing data for a more natural expression. Furthermore, speech morphing is an appropriate synthesis technique for maintaining individuality and naturalness at the same time.

As materials of the proposed method, variously expressed singing voices by a particular singer must be collected for use in singing voice synthesis, which is based on varying strength by using morphing technology. We recorded the singing voice of a female amateur singer in her twenties at a sampling frequency of 44.1 kHz. The amateur singer sang a Japanese nursery rhyme, "Furusato" (Hometown), with a recorded piano accompaniment that is audible to the singer to set the same speech speed and F_0. The singer was instructed to sing in the four types of expressions listed in Table 1 while keeping each expression consistent in her singing. For the sake of convenience, these expressions are labeled as "normal," "dark," "whispery," and "nasal" in this paper. Among various expressions, we selected the above four from the viewpoint of variation in the technical skill involved in the song types. Here, for example, "dark" emphasizes expressiveness like that produced by an *opera* singer, "whispery" is a hoarse voice like a lullaby sung as interlude expressions in certain songs, and the "nasal" expression is used in *pop music* for temporally emotional emphasis.

Table 2: Coordination of expressive singing voice synthesis.

Abbr.	Base	Target
A-1	"normal"	"dark"
A-2	"normal"	"whispery"
A-3	"normal"	"nasal"

As the basis for evaluating naturalness, it is important to produce a sound that is as natural as possible. Therefore, we synthesized the variously expressed morphed singing voices by applying a speech morphing algorithm [4] based on STRAIGHT [5]. STRAIGHT is a speech analysis-synthesis system based on channel-vocoder architecture that is designed to eliminate interference by periodicity. It does this by applying a pitch-adaptive and time-frequency smoothing method that separately extracts the spectrum envelope and source components. These analyzed components are appropriate for synthesizing a smooth change in the perceptual attributes between different voices.

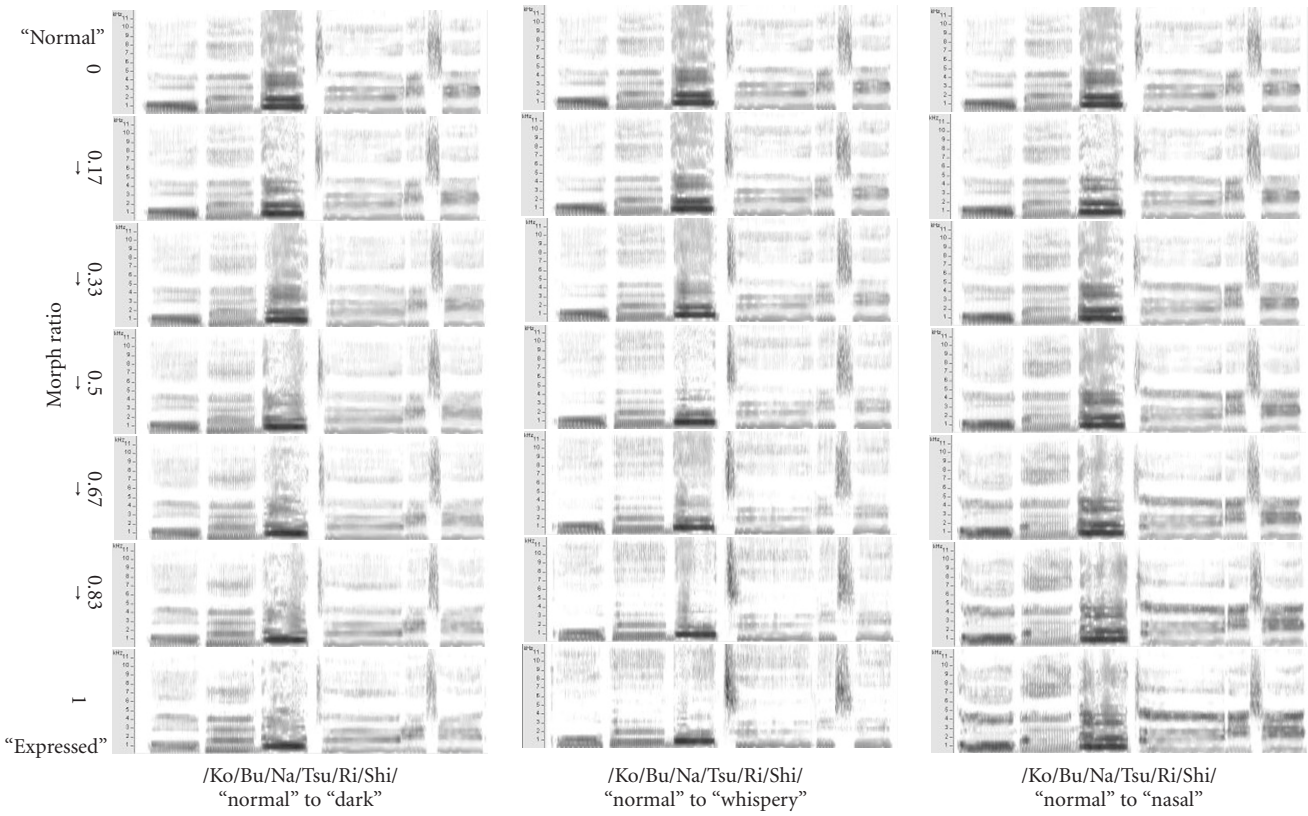
As shown from A-1 to A-3 in Table 2, we first synthesized morphed singing sounds expressed at various strengths by using "normal" as the base and the three types of the singing voice as the targets. Figure 2 shows the spectrogram for each expression when the voices are synthesized at morph ratios from 0 (base) to 1 (target). We adopted not only interpolation but also extrapolation for the emphasized or opposite expressions. As sufficient steps for tracing the shape of interpolation, we temporary 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. PERCEPTUAL TESTS FOR VERIFICATION OF PREMISES

In this section, as an essential preliminary step, we use the subjective evaluations for expressive strength to verify the premises upon which we base the perceptual tests described below. This is done by confirming (a) the perception of naturalness in the synthesized singing voice at various morphing ratios and (b) the difference in perception between expressive singing voices and a monotone singing voice without expression.

4.1. Naturalness of expressive eorphing sound

To examine the effect of the morphing synthesis on naturalness, we conducted a perceptual evaluation to verify the levels of naturalness at the base and target positions for different morphing ratios. The morphing algorithm converts the spectral information from the base and target, so we expected the existing sounds at morph ratios 0 and 1 to perform at high naturalness, the interpolated sounds to perform at sufficient naturalness, and the extrapolated sounds to perform at low naturalness.



∗ Window size of these graphs: 128 points (≒ 2.9 ms); window type: hamming

Figure 2: Spectrograms of synthesized voices with morph ratios from 0 to 1.

Hypothesis

The perception curve of naturalness in singing voice morphing is highest at morphing ratios 0 and 1, slightly lower at a morphing ratio around 0.5, and lowest under 0 or over 1.

Method

The subjects listened to audio stimuli in randomized order through headphones attached to a Windows PC and gave subjective evaluations based on a seven-point rating scale: "completely suitable, very suitable, somewhat suitable, indeterminate, somewhat unsuitable, very unsuitable, and completely unsuitable" for the instructed criterion of "naturalness," using the GUI of the Tcl/Tk program.

Subjects

Thirteen people aged in their twenties to lower-thirties (six females and seven males).

Stimuli

We adopted the synthesized morphed sounds shown in A-1 to A-3 of Table 2 while using six morae, "Ko Bu Na Tsu Ri Shi."

Procedure

Subjects listened to the morphed sounds of A-1 to A-3 in Table 2 and judged the evaluation item of "naturalness" according to the seven-point rating scale. They were also instructed to base the criterion of naturalness on how much they felt the sound resembled a human voice.

Results

Figure 3 displays the averages of evaluated naturalness by the morphing ratios. Although we expected the subjective evaluations to be higher at the morphing ratios of 0 and 1, a deeper expression was not recognized as natural in the cases of A-1 and A-3. Two-factor ANOVAs show significant differences among morphing ratios ($F = 41.692$, $p < .01$) and interactions between morphing ratios by expressive types ($F = 2.320$, $p < .01$), differently from the factor of expressive types ($F = 2.398$, $p = .105$). The results of post-hoc tests (Fisher), except extrapolations, show significant differences between A-1/A-3 ($p = .039$) and between A-2/A-3 ($p < .01$) only at morph ratio 1. Here, it is possible that a morphed voice with a consistently strong expression was perceived as an artificial voice.

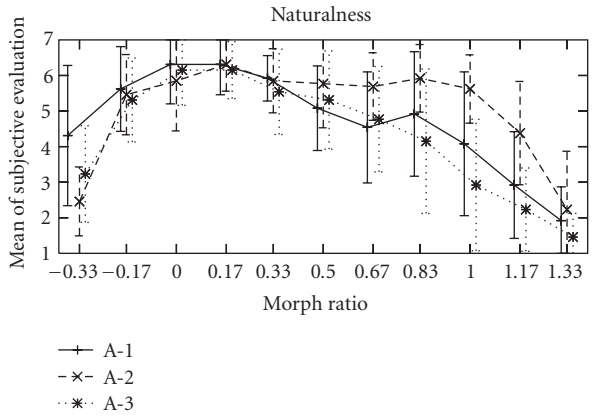


Figure 3: Evaluation of naturalness at each morphing ratio.

4.2. Perceptual discrimination

As another basic premise of the perceptual experiments in this paper, it is also necessary to confirm perceptual identification with discriminant evaluations between a singing voice without expression ("normal") and expressive singing voices ("dark," "whispery," and "nasal"). The following experiments feature discrimination tests using perceptual evaluations.

Hypotheses

The expressions of the singing voices are different from each other.

Method

The subjects listened to audio stimuli in randomized order through headphones attached to a Windows PC and gave subjective evaluations on a seven-point rating scale: "completely suitable, very suitable, somewhat suitable, indeterminate, somewhat unsuitable, very unsuitable, completely unsuitable" for the instructed criterion of "similarity to each other," using the GUI of the Tcl/Tk program.

Subjects

Thirteen people aged in their twenties to lower-thirties (six females and seven males).

Stimuli

We adopted the four voices with different expressions: "normal," "dark," "whispery," and "nasal," converted by the STRAIGHT analysis-synthesis process from each original sound in order to assimilate the stimuli with the morphed voice. As a sample of the singing voice, we selected six morae, "Ko Bu Na Tsu Ri Shi," in the morphed song from the synthesized song data described in Section 3. Speech speed was about 2.0 morae/second, and the F_0 range was approximately 300 Hz to 450 Hz on average in each musical interval. The length of each tune was about 3.0 seconds.

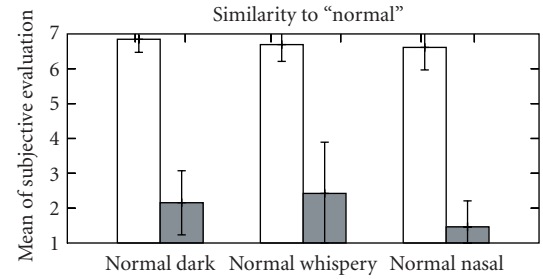


Figure 4: Pairwise comparison between "normal" and expressions.

Table 3: T-test of the identification between "normal" and "dark," "whispery," or "nasal" ($\alpha = 0.01$).

"dark"	"whispery"	"nasal"
$t_{(12)} = 15.0$	$t_{(12)} = 10.2$	$t_{(12)} = 24.0$
$P < .01$	$P < .01$	$P < .01$

Procedure

Subjects evaluated stimuli using the seven-point rating scale described above in pairwise comparison between "normal" and ("normal," "dark," "whispery," or "nasal") while each pair was continuously played back.

Results

The means of subjective evaluations of the identification results of the perceptual evaluations compared with "normal" are shown in Figure 4. To verify the difference between "normal" and the other voices, Table 3 shows the T-test results of identification between ("normal" and "dark," "whispery," or "nasal.") These results indicate that the expressed singing voices are accurately perceived as different from "normal" in perceptual feeling.

5. PERCEPTUAL TESTS OF THE CONTINUITY OF EXPRESSIVE STRENGTH

The perceptual continuity in expressive strength must be verified in order to design natural change in this parameter. In this section, we evaluate and verify the continuity of expressive strength in a singing voice using synthesized sound.

5.1. Perceptual continuity of expressive strength

To confirm the perceptual continuity of expressive strength in a singing voice, the following evaluation shows the perceptual strength of a singing-voice expression synthesized at various morphing ratios.

Hypotheses

The perceived expressive strength of a singing voice is continuously changed by the morphing ratio.

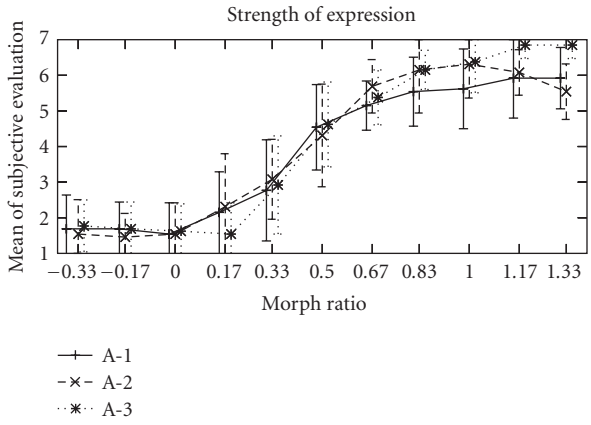


Figure 5: Evaluation of expressive strength.

Method

Same as Section 4.2.

Subjects

Same as Section 4.2.

Stimuli

We adopted the synthesized morphed sound shown in Table 2. As a sample of the morphed voice, we selected six morae, "Ko Bu Na Tsu Ri Shi," the same used in Section 4.2, in the morphed singing voice from the synthesized song data described in Section 3.

Procedure

Subjects listened to the morphed sound of A-1 in Table 2 and judged the evaluation item of expressive strength of "dark" according to the seven-point rating scale. They did the same experiments for A-2 ("whispery") and A-3 ("nasal"). In preparation for evaluating an item, subjects were instructed to listen to a control "dark" sound before this experiment to confirm their understanding of what we defined as "dark."

Results

The results of perceptual evaluations are shown in Figure 5. The subjective evaluations of the expressive strength correspond to the morphing ratio in the figure. Thus we confirmed that our hypotheses were correct in this experiment. These results indicate that the morphing of a singing voice can supply rich expression by varying the kind and strength of expressions in the perceptual measure.

5.2. Approximation and compensation of perceptual curve

We assume that the curves of subjective evaluations by linear interpolation are approximated by a sigmoid expression (Equation (1), where x is the morph ratio), especially for the morph ratios between 0.0 and 1.0,

$$\text{evaluated values} \approx \frac{5}{1+e^{-6x+3}} + 1.5. \tag{1}$$

To synthesize the expressive strength of a singing voice in perceptually linear evaluated values, we propose using a morph ratio computed by (2), in which a is the original morph ratio at a regular interval, from the reciprocal function of (1),

$$\text{compensated morph ratio} = 0.5 - \frac{1}{6}\log\left(\frac{1}{a} - 1\right) \quad (0 < a < 1). \tag{2}$$

To verify linear perception with a compensated morph ratio as proposed, we conducted the perception experiment described below.

Table 4: R-square values of approximation in linear interpolation.

Approximation	A-1	A-2	A-3
Linear	0.91	0.86	0.91
Sigmoid	0.95	0.97	0.97

Hypothesis

Expressions of a singing voice are interpolated to be perceptually linear when the morph ratio is set by (2).

Subjects

Nineteen people aged in their twenties to thirties (seven females and twelve males).

Stimuli

We synthesized the singing voices in seven steps from 0 to 1 with compensated intervals of morph ratio: 0, 0.39, 0.45, 0.5, 0.55, 0.61, 1. The prepared singing voices, called A-*1-sig*, were synthesized from the base and target by a method similar to A-1 in Table 2, with A-*2-sig* and *3-sig*, which are the same as A-1-sig.

Method and procedure

Same as tests in Section 5.1.

Results of reciprocal sigmoid interpolation

The evaluated values of the expressive strength shown in Figure 6 seem to be linear compared with the evaluated values' curves in Figure 5. To verify the adequacy of the approximation of the evaluated values' curves by (1), we calculated R-square values of the approximation and compared these values with linear approximation (Table 4). As a result, each R-square value for A-1 to 3 shows that (1) improved the approximation.

We next verified that the morph ratio computed by (2) leads to perceptual linearity. In comparing A-1-sig to 3 with

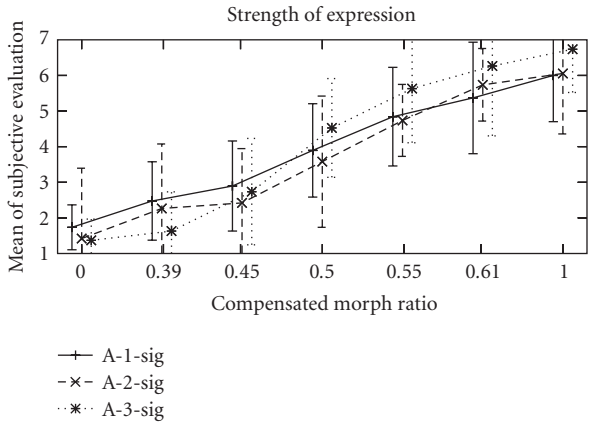


Figure 6: Evaluation of expressive strength in compensated ratio.

Table 5: R-square values with sigmoid compensated strengths.

Approximation	A-1-sig	A-2-sig	A-3-sig
Linear	0.99	0.97	0.96

the linear approximation shown in Table 4, we can see that each R-square value was improved with compensated interpolation (Table 5), as we intended in this experiment. These results demonstrate that our hypothesis is correct.

6. PERCEPTUAL TEST ON THE NATURALNESS OF TEMPORALLY CHANGING EXPRESSIVE STRENGTH

In this section, we evaluate the naturalness of a singing voice with continuous change in expressive strength; clarifying this value is the main target of this research.

To verify the effectiveness of the interpolation method for expressive strength with change over time, we focused on the naturalness of the singing voices with/without interpolation of expressive strength from 0 to 1. By adopting "whispery," which was stably evaluated as natural in the results of Section 4.1, we compared (i) the singing voices smoothly morphed between "normal" and "whispery" to (ii) the singing voices suddenly switched between them. In this experiment we also evaluated (iii) the effectiveness of the reciprocal sigmoid interpolation of the morph ratio.

Hypotheses

(1) The gradual change in the expressive strength between "normal" and "whispery" is more natural than the sudden binary switch between them. (2) The reciprocal sigmoid interpolation more effectively achieves naturalness of the singing voice than does the linear interpolation.

Method

Subjects evaluated the naturalness of the change in expressive strength of the stimulus by using the seven-point rating scale given above.

Table 6: Type of morph ratio in gradual expression.

Method of change		Morph ratio					
		Ko	Bu	Na	Tsu	Ri	Shi
Switch	(a)	0	0	0	1	1	1
	(b)	1	1	1	0	0	0
Linear	(a)	0	0.2	0.4	0.6	0.8	1
	(b)	1	0.8	0.6	0.4	0.2	0
Rec-Sigmoid	(a)	0	0.40	0.47	0.53	0.60	1
	(b)	1	0.60	0.53	0.47	0.40	0

morph ratio: "normal" = 0; "whispery" = 1; morphing voice = $\{0 < x < 1\}$.

Subjects

Nineteen people aged in their twenties to thirties (seven females and twelve males).

Stimuli

To verify the hypotheses, we set up three conditions of the method of expressive change: *Switch* is the sudden binary change, *Linear* is the linear change, and *rec-Sigmoid* is the reciprocal sigmoid change of the morph ratio. The first one is currently used in conventional methods, while the latter two are proposed here. We focused on the "whispery" singing voice, which is basically natural as mentioned in Section 5.1.

Corresponding to each condition, we applied the morph ratio shown in Table 6 to each mora, "Ko Bu Na Tsu Ri Shi." These were connected at each end of the morae at zero cross points cut from A-2 and A-2-sig, which were synthesized in six steps of morph ratio.

We eliminated the effect of order by averaging the results of (a) "normal" to "whispery" and (b) "whispery" to "normal."

Procedure

Subjects listened to the morphed sound of A-2-sig, synthesized in Section 5.2, and evaluated the *naturalness of expressive change* according to the seven-point rating scale based on the criterion of how much they felt the sound resembled a human voice.

Results of perception test with expressive change

Figure 7 shows the means of subjective evaluations for naturalness. We intended to show in a standardized way the results among the three methods by using ANOVA with repeated measurement ($\alpha = 0.01$, $\phi = 18, 2$). There were significant differences among Switch, Linear, and rec-Sigmoid (ANOVA, $F = 35.25$, $p < .01$). The results of the multiple comparisons are ({Switch, Linear}: 2.03, Scheffé, $p < .01$; {Switch, rec-Sigmoid}: 2.16, $p < .01$; {Linear, rec-Sigmoid}: 0.13, $p = 0.90$). These findings show that the interpolation methods (Linear and rec-Sigmoid) provide more natural results than does the Switch method.

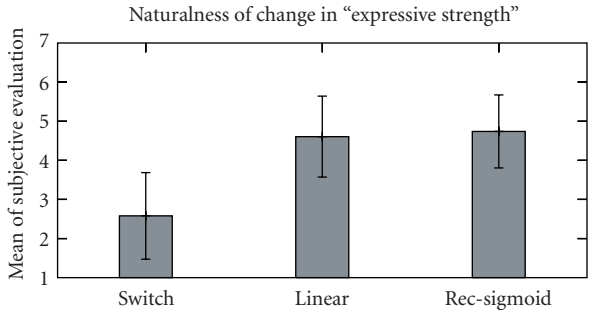


Figure 7: Naturalness of change.

7. DISCUSSIONS

7.1. Continuity of expressive strength

The evaluation in Section 4.2 confirms a fundamental characteristic of perception: The expressive singing voices "dark," "whispery," and "nasal" were discriminated from "normal," that is, a flat singing voice without expression.

The perceptually continuous tendency of interpolated and extrapolated expressions was found in Section 5.1 although it had a nonlinear curve as with the sigmoid functions. Next, we discuss the effectiveness of compensating expressive strength in a singing voice. Section 5.2 proposed a method using a compensated morphing ratio to linearize the perceptual curve. The results also indicate that it is reasonable to conjecture that the reciprocal sigmoid equation calculates morph ratios that provide an appropriate interval for linear perception of expressive strength. It is also assumed that the nonlinear curve was well controlled for linearization.

7.2. Effectiveness of smoothing expressive strength

We now discuss the effectiveness of controlling expressive strength in a singing voice. First, the evaluation results of Section 4.1 show that the interpolated sounds are as natural as the original stimuli by comparing the evaluated values of naturalness around the middle morphing ratios with the average of that at morphing ratios 0 (base) and 1 (target). From these results, it is assumed that the speech analysis-synthesis conversion by STRAIGHT has no serious effect on perceptual naturalness.

The results in Section 6 confirm that both types of gradual transformation, Linear and rec-Sigmoid, have an advantage over the binary discrete transformation (Switch) from the viewpoint of naturalness, which is our goal. Consequently, this work has demonstrated that singing voice synthesis achieved by morphing vocal timbre can make a strong contribution toward the production of natural expressions in a singing voice. Such positive effects will be applied to speech and singing voice synthesis along with unit concatenation based on a corpus. This method can also be adopted in real-time voice synthesis with natural expressions for application to robotics and personified interfaces [17].

There was no significant difference between Linear and rec-Sigmoid although we could improve the perceptual linearity using the latter. This result poses a new question as to how perceptual linearity is related to the naturalness of singing-voice expression. Consequently, the proposed approach also needs additional verification of this difference by using longer and shorter voice samples as well as stronger and weaker expressions in the transformation.

8. CONCLUSION

This paper experimentally investigated the perceptual continuity of expressive strength in vocal timbre. This was carried out using the gradually changing expression of a singing voice based on a speech morphing algorithm.

From the results of the evaluations, we concluded that continuous strength in expression of vocal timbre has advantages over the approach of binary discrete transformation between two expressions. Furthermore, we confirmed that the curve of the change in expressive strength is not linear in a perceptual sense, and that compensated expressive strength causes perceptual linearization.

The clarified effectiveness of perceptually natural and continuous expression by gradual change in vocal timbre can be combined with other prosodic approaches such as F_0 in corpus-based synthesis. Integrating the elements of expression, which have been sufficiently evaluated for their effect on perception, will ensure a high degree of expressiveness. The synthesis of highly natural continuity in an expressive voice is important not only for advancing speech synthesis but also for creating attractive personification.

ACKNOWLEDGMENTS

The authors would like to thank Professor Hideki Kawahara for permission to use the STRAIGHT morphing system. We also thank Dr. Norihiro Hagita, Mr. Yoshinori Sakane, Dr. Kenji Susami, and other ATR personnel for their help. This research was supported in part by the National Institute of Information and Communications Technology of Japan.

REFERENCES

[1] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 177–190, 2003.

[2] T. Minato, K. F. MacDorman, M. Shimada, S. Itakura, K. Lee, and H. Ishiguro, "Evaluating humanlikeness by comparing responses elicited by an android and a person," in *Proceedings of the 2nd International Workshop on Man-Machine Symbiotic Systems*, pp. 373–383, Kyoto, Japan, November 2004.

[3] D. Hanson, "Exploring the aesthetic range for humanoid robots," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society in Cooperation with the 5th International Conference on Cognitive Science (CogSci/ICCS '06)*, pp. 16–20, Vancouver, BC, Canada, July 2006.

[4] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 256–259, Hong Kong, April 2003.

[5] 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*, vol. 27, no. 3-4, pp. 187–207, 1999.

[6] M. Schröder, "Emotional speech synthesis: a review," in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech '01)*, vol. 1, pp. 561–564, Aalborg, Denmark, September 2001.

[7] D. Erickson, "Expressive speech: production, perception and application to speech synthesis," *Acoustical Science and Technology*, vol. 26, no. 4, pp. 317–325, 2005.

[8] A. Iida, S. Iga, F. Higuchi, N. Campbell, and M. Yasumura, "A speech synthesis system with emotion for assisting communication," in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 167–172, Belfast, Northern Ireland, UK, September 2000.

[9] N. Campbell, "Developments in corpus-based speech synthesis: approaching natural conversational speech," *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, pp. 376–383, 2005.

[10] T. Saitou, M. Unoki, and M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, vol. 46, no. 3-4, pp. 405–417, 2005.

[11] T. Saitou, N. Tsuji, M. Unoki, and M. Akagi, "Analysis of acoustic features affecting "singing-ness" and its application to siging-voice synthesis from speaking-voice," in *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP '04)*, vol. 3, pp. 1929–1932, Jeju, Korea, October 2004.

[12] P. Cano, A. Loscos, J. Bonada, M. Boer, and X. Serra, "Voice morphing system for impersonating in karaoke applications," in *Proceedings of the International Computer Music Conference (ICMC '00)*, pp. 109–112, Berlin, Germany, August 2000.

[13] Y. Sogabe, K. Kakehi, and H. Kawahara, "Psychological evaluation of emotional speech using a new morphing method," in *Proceedings of the 4th Joint International Conference on Cognitive Science (ICCS/ASCS '03)*, Sydney, Australia, July 2003.

[14] H. Matsui and H. Kawahara, "Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech '03)*, pp. 2113–2116, Geneva, Switzerland, September 2003.

[15] P. B. Mareüil, P. Célérier, and J. Toen, "Generation of emotions by a morphing technique in English, French and Spanish," in *Proceedings of Speech Prosody*, pp. 187–190, Aix-en-Provence, France, April 2002.

[16] T. Yonezawa, N. Suzuki, K. Mase, and K. Kogure, "Gradually changing expression of singing voice based on morphing," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 541–544, Lisbon, Portugal, September 2005.

[17] T. Yonezawa, N. Suzuki, K. Mase, and K. Kogure, "Handysinger: expressive singing voice morphing using personified hand-puppet interface," in *Proceedings of the 5th International Conference on New Interfaces for Musical Expression (NIME '05)*, pp. 121–126, Vancouver, Canada, May 2005.