Gradually Changing Expression of Singing Voice based on Morphing

Tomoko Yonezawa^{1),2)}, Noriko Suzuki³⁾, Kenji Mase^{2),1)}, Kiyoshi Kogure¹⁾

1) ATR Intelligent Robotics and Communication Laboratories

2) Nagoya University

3) ATR Media Information Science Laboratories

{yone, noriko, mase, kogure}@atr.jp

Abstract

We have developed a method for synthesizing a singing voice by gradually changing the musical expression based on speech morphing. This paper shows the advantages of this method in comparison with the approach of binary discrete transformation between two expressions, confirmed by statistical analyses of perception tests. In order to synthesize different expressional strengths of a singing voice, a “normal” (without expression) voice of a particular singer is used as the base of morphing, and three different expressions, “dark,” “whispery” and “wet,” are used as the target. Through our experiments, we confirmed i) the proposed morphing algorithm effectively interpolates the expressional strength of a singing voice, ii) an approximate equation of the perceptual sense can be used to calculate the morph ratio at a perceptually linear interval, and iii) our gradual transformation method can generate a natural singing voice from the interpolation of two different expressions.

1. Introduction

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 has variously transformed expressions. Recent approaches to speech synthesis with emotional expressions include rule-based methods such as those mentioned by Schröder[1] and corpus-based methods such as CHATAKO[2]. While rule-based methods mainly address prosody controls, corpus-based methods cover vocal sound. However, there are gaps between the different expressions or emotions achieved by each corpus.

To solve this problem, we have developed a method for gradually changing expression by using speech morphing. By focusing on the effectiveness of vocal sound without prosodic features, our approach thus achieves the purpose of its name: Expressive Singing Voice Morphing (ESVM).

Cano et al.[3] 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.[4] and Matsui et al.[5] investigated the sound morphing of emotional speech by a particular speaker. The former research used different singers, and the latter included 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 by using the same singer, speed, and F_0 .

Figure 1 shows an example of the anticipated advantages: while the strength of the original vocal expression gradually transforms, corpus-based voice synthesis provides binary dis-

Figure 1: Example of Advantage achieved with Morphing Expression

crete transformation depending on categories of vocal expression. In contrast to the discontinuity of the existing methods, our approach can accurately approximate natural expression.

In this paper, we explain the synthesizing method of ESVM by a speech morphing algorithm[6] using STRAIGHT[7] and confirm the effectiveness of interpolation and extrapolation of the singing voice. We then approximate MOS (mean opinion score) values 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.

2. ESVM Synthesis

We collected variously expressed singing voices by a particular singer for use in singing voice synthesis based on varying strength with morphing technology. It is possible to synthesize the voice parameters, but we focused on vocal synthesis from the existing data for a 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,” “whispery,” and “wet” (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, “whispery” 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.

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

Table 1: Expression Types in Recorded Voices

expression	singing instruction
“normal”	flat, without expressions
“dark”	entirely like interior tongue vowel
“whispery”	including more white noise
“wet”	entirely nasal voice

Table 2: ESVM Synthesis

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

the emphasized expressions, the morphing rate was set to 0 or less and to 1 or more. As sufficient steps for tracing the shape of 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).

3. Perception Test of ESVM

In this section we evaluate 1) the perceptual effect and the naturalness of the linearly interpolated expression of a singing voice, 2) whether an approximate expression enables perceptually linear interpolation, and 3) whether the interpolation of the expressional strength enables natural transformation between different expressions of a singing voice.

3.1. Consistent ESVM: Experiment #1 and #2

3.1.1. Experiment #1: Linear (Original) Interpolation

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 and extrapolation 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 morphed voice made from the two expressed voices is different from “normal,” and ★3) that the expression level is changed by morphing.

Method: The subjects listened to stimulating sounds in randomized order 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” for the instructed criteria on the GUI of the Tcl/Tk program.

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

Stimulating Sounds: We adopted the synthesized morphed sound shown in Table 2 and four original sounds: “normal,” “dark,” “whispery” 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 2. 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 sound was 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, whispery or wet] while each pair was continuously played back.

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

Figure 2: Pairwise Comparison between “normal” and Expressions

Table 3: T-test between the Identification of “normal” and [“dark,” “whispery” or “wet”] ($\alpha = 0.01$)

“dark”	“whispery”	“wet”
$t_{(12)} = 15.0$	$t_{(12)} = 10.2$	$t_{(12)} = 24.0$
$p < .01$	$p < .01$	$p < .01$

seven grades.

★3) 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 “whispery” and A-3 as “wet.” In preparation for evaluating an item, subjects were instructed to listen to a control “dark” sound before this experiment to confirm what is defined as “dark.” They were also instructed to base the criterion of naturalness on how much they feel the sound resembles a human voice.

Results in Simple Linear Interpolation: The MOS averages of the identification results of perception test ★1) compared with “normal” are shown in Figure 2. To verify the difference between “normal” and the other sounds, Table 3 shows T-test results between the identification of [“normal”, “dark,” “whispery” or “wet.”] These results indicate that the expressed singing voice is accurately perceived as different from “normal” in perceptual feeling.

Perception test ★2) shows that expressional morphing between two expressed voices gives a new expression that is different from “normal.” Figure 3 shows that the morphed voices from B-1 to B-3 at a morphing ratio of around 0.5 are comparatively more similar to “normal” than are the voices at other ratios. Therefore, we focused on 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 by T-test (Table 4.) The results show that a synthesized voice is significantly different from “normal” even at a morphing ratio of around 0.5.

The results of perception test ★3) (Figure 4) show that MOS averages of the expressional strength correspond to the morphing ratio. Figure 5 displays the naturalness of the morphed sound. Although we estimated MOS values 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. It is possible that a morphed voice with a continuous hard expression was recognized as an artificial voice.

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.

3.1.2. Experiment #2: Reciprocal Sigmoid Interpolation

In Figure 4, MOS values change steeply around a morph ratio of 0.5. We assume that the MOS values of ESVM strength by linear interpolation are approximated into a *sigmoid* expression (Equation (1), where x is the morph ratio), especially for the morph ratios between 0.0 and 1.0.

Figure 3: Similarity of Morphing Sound Synthesized between Expressed Voices

Table 4: 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)	$t_{(12)} = 7.23$ $p < .01$	$t_{(12)} = 5.33$ $p < .01$	$t_{(12)} = 9.68$ $p < .01$

$$MOS \approx \frac{5}{1 + e^{-6x+3}} + 1.5 \quad (1)$$

To synthesize ESVM in perceptually linear MOS values, we propose using a morph ratio computed by Equation (2), in which a is the original morph ratio at a regular interval, from the reciprocal function of Equation (1).

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

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

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

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

Stimulating Sounds: 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 A-3-sig, which are the same as A-1-sig.

Method and Procedure: Same as test *3) in Experiment #1.

Results of Reciprocal Sigmoid Interpolation: The MOS averages of the expressional strength shown in Figure 6 seem to be linear compared with the MOS values in Figure 4. To verify the adequacy of the MOS value's approximation by Equation (1), we calculated R-square values (Table 5) and compared them. As a result, each R-square value for A-1 to 3 shows that Equation (1) has improved the adequacy.

We next verified that the morph ratio computed by Equation (2) leads to perceptual linearity. In comparing A-1-sig to 3 with A-1 to 3 in the linear approximation, we can see that each R-square value has improved, as we intended in this experiment. These results demonstrate that our hypothesis is correct.

3.2. Gradual Expression using ESVM: Experiment #3

To verify the effectiveness of ESVM interpolation as the strength of the expression changes with time, we focused on the naturalness of the singing voices with/without interpolation. By adopting "whispery," which was natural in the results of test

Figure 4: Evaluation of Expressional Strength

Figure 5: Evaluation of Naturalness

*3), 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 of the expressional 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 linear interpolation.

Method: Subjects evaluated the naturalness of the expressional change in the stimulating sound by using the seven grades given above.

Subjects: Same as Experiment #2.

Stimulating Sounds: To verify the hypotheses, we set up three conditions of the method of expressional change: Switch is sudden binary change, Linear is linear change, and rec-Sigmoid is the reciprocal sigmoid change of the morph ratio. The first one is a currently used conventional method, while the latter two are proposed here. We focused on the "whispery" singing voice, which is basically natural as mentioned in Section 3.1.1.

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 the order by averaging the results of a) "normal" to "whispery" and b) "whispery" to "normal."

Procedure: Subjects listened to the morphed sound of A-1 in Table 2 and evaluated the *naturalness of expressional change*, according to the seven grades based on the criterion of how much they feel the sound resembles a human voice.

Results of Perception Test with Expressive Change: Figure 7 shows the averages of MOS values of naturalness. We intended to show in a standardized way the results among the three methods by using analyses of variance (ANOVA) with repeated measurement ($\alpha = 0.01, \phi = 18, 2$). There were significant differences among Switch, Linear and rec-Sigmoid (ANOVA, $F=35.249, p<.0001$). The results of the multiple comparisons are ({Switch, Linear}, 2.026, Scheffé,

Figure 6: Evaluation of Expressional Strength in Compensated Ratio

Table 5: R-square Values of Approximation

morph type	linear			rec-sigmoid		
	A-1	A-2	A-3	A-1-sig	A-2-sig	A-3-sig
linear	0.91	0.86	0.91	0.99	0.97	0.96
sigmoid	0.95	0.97	0.97			

$p < .0001$ {Switch, rec-Sigmoid}, 2.158, $p < .0001$ {Linear, rec-Sigmoid}, 0.132, $p = 0.9013$, which show that the interpolation methods (Linear and rec-Sigmoid) are more natural than the Switch method.

4. Discussion

We now discuss the effectiveness of our proposed ESVM. The perception tests demonstrate that ESVM between “normal” (without expression) and three types of expression interpolates and extrapolates the strength of the expression of the singing voice. This finding suggests that the vocal expression’s morphing is practical even under the condition of same singer, same F_0 , and same speed.

The results also make it reasonable to conjecture that the reciprocal sigmoid equation calculates morph ratios that provide an appropriate interval for linear perception in the expressional strength.

We finally confirmed that both types of gradual transformation, Linear and rec-Sigmoid, have an advantage over the binary discrete transformation (Switch) from the aspect of naturalness, which is what we aimed for. These results demonstrate that ESVM 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 with unit concatenation based on a corpus. ESVM can also be adopted in real-time voice synthesis with natural expressions for application to robotics and personified interfaces[8].

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

5. Conclusions

This paper proposed a method for singing voice synthesis, called ESVM, by gradually changing expression based on a speech morphing algorithm. From the results of perception tests, we concluded that this method has advantages over the approach of binary discrete transformation between two expressions.

As future works expand the capabilities of ESVM, its extrapolation should be developed to improve the synthesized

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
Reciprocal 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 = $\{0 < x < 1\}$

Figure 7: Naturalness of Change

voices by incorporating emphasized expressions, which would enrich the current implementation.

6. Acknowledgements

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.

7. References

- [1] Schröder, M., “Emotional Speech Synthesis: A Review,” Proc. Eurospeech, volume 1, pp. 561–564, 2001.
- [2] Iida, A., Iga, S., Higuchi, F., Campbell, N., Yasumura, M., “A Speech Synthesis System with Emotion for Assisting Communication”, Proc. ISCA Workshop on Speech and Emotion, pp. 167–172, 2000.
- [3] Cano, P., Loscos, A., Bonada, J., Boer, M., and Serra, X., “Voice Morphing System for Impersonating in Karaoke Applications,” Proc. ICMC’2000, pp. 109–112, 2000.
- [4] Sogabe, Y., Kakehi, K., and Kawahara, H., “Psychological evaluation of emotional speech using a new morphing method,” 4th ICCS International Conference on Cognitive Science, 2003.
- [5] 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.
- [6] 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.
- [7] Kawahara, H., Masuda-Kasuse, I., and Cheveigne, A., “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, 27, pp.187–207, 1999.
- [8] Yonezawa, T., Suzuki, N., Mase, K., and Kogure, K., “HandySinger: Expressive Singing Voice Morphing using Personified Hand-puppet Interface,” Proc. NIME2005, pp. 121–126, 2005.