社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－HCI－128 (5)
2008－MUS－ 75 (5)
2008／5／28

音声モーフィングによる歌声の声色強度変化の知覚特性の分析

米澤 朋子 †，鈴木 紀子 ††，安部 伸治 †，間瀬 健二 ‡，小暮 潔 ‡‡

† ATR 知能ロボティクス研究所　†† 情報通信研究機構/ATR 認知情報科学研究所
‡ 名古屋大学　‡‡ ATR 知識科学研究所

擬人的媒体等とのインタラクションに伴う音声表現に，連続的な声色表情付与を狙いとし，STRAIGHT を用いた音声モーフィングを同一歌唱者の表情無し歌声–表情付き間歌声間に適用した．合成音の声色の表情強度に関して一対比較を行った結果，第一刺激に比べて第二刺激の表情強度が増す場合と減る場合で，刺激間の差分の知覚が異なることがわかった．これに基づき声色の表情付け手法を議論する．

Transformation of Perceptual Strength for Vocal Timbre in Singing Voice using Speech Morphing

Tomoko Yonezawa †　Noriko Suzuki ††　Shinji Abe †　Kenji Mase ‡　Kiyoshi Kogure ‡‡

† ATR Intelligent Robotics and Communication Lab.
†† NICT / ATR Cognitive Information Science Lab.　‡ Nagoya University
‡‡ ATR Knowledge Science Lab.

This paper proposes and evaluate a method for synthesizing continuous expressions in vocal timbre by gradually changing spectral parameters based on STRAIGHT speech morphing algorithm among differently expressed singing voices. In order to synthesize natural change of various and continuous strengths of singing voice expressions, a singing voice without expression, "normal," is used as the base of morphing, and singing voices of the particular singer with three different expressions, "dark," "whispery" and "wet," are used as targets. The results of the paired comparison between different strengths of vocal timbre suggested that the decrease of the expression in vocal timbre is more perceived than the increase in some expressions.

1　はじめに

現代の世の中では，擬人的表現をもたらす存在の必要性が徐々に認められてきており，ペットロボットや案内ロボット，エージェントによるガイドといった様々なシーンで，人間以外の人工的で擬人的な存在が不可欠になりつつある．人らしさや生き物らしさは，あたかもそれらの内部で感じたり，考えたりしながら，我々と同様に精神生活を送っているかのように見せうるひとつの重要な表現手段として注目できる．擬人的表現の中でも，無機質で単一な音声を用いた表現では，擬人性を損なう可能性があることから，バーバル表現・ノンバーバル表現を含む音声は重要なチャンネルと言え，実際感情音声に関する研究が進められつつある[1]．

感情音声データベースを用いた感情音声合成 CHATAKO[2] なども提案されているが，この手法ではデータベースの種類に依存した不連続な感情表現の付与となってしまう．そこで，様々な強度の表情付与や，それらが変化するといった微妙な表情付与を実現する手段が必要だと考えた．他者の声からユーザの声に変換するカラオケシステム[3]などにおいて音声モーフィング手法を適用する例もあり，元音声の音質をあまり損なわずに音声補間ができる手法として注目できる．実際に F_0 や話速の異なる様々な感情音声に対する音声モーフィングの適用も試みられている[4, 5]．

そこで，我々はこれまでに STRAIGHT[6] を用いた感情音声モーフィング手法[7] を導入し，同一歌唱者の歌声に対する様々な強度における声色表情付与を実現する手法 ESVM を提案している[8]．文献[8] では，音声モーフィング手法による声色の補間が実現したことを確認し，Mean Opinion Score (**MOS**) によって計測された表情強度を逆 sigmoid

関数によるモーフィング間隔の調整により線形化することに成功した．しかし，その線形性は静的な表情強度に関する調整に過ぎず，歌声が時々刻々と徐々に変化するような状況に対して，表情強度の増加と減少に同一のパラメータ設定手法を適用しても自然な表情付与がなされるとは限らない．また，静的な声色の表情強度の知覚曲線が sigmoid 関数で近似されたことにより，カテゴリー知覚 (詳細は本節で後述) の可能性も否めない．

これまでに音声表現の強度（**表情強度**）を一対比較し被験者の正答率を調べた知覚的連続性に関する実験[9]もあるが，表情強度の比較差分の伸縮は行われておらず，表情強度の増加と減少の各状況における個別の表情付与方法についての検討は見られない．本稿では，音声におけるプロソディ情報が極力除去された歌声の声色に対し，強度変化の増減方向に応じた表情付与について検討する．

カテゴリー知覚は異なる事象に対する知覚特性として知られている．連続的に変化する刺激に対して，徐々に知覚も変化するのではなく，ある特定の変化地点において不連続に知覚が変化する．これは色彩の知覚[10]や母音の知覚[11]などにおいて議論される知覚特性である．鈴木ら[12]は，顔表情などの感情表現における知覚の不連続特性を議論している．筧ら[9]は感情音声におけるカテゴリー知覚の可能性を示唆しているが，様々な表情強度の比較順序や比較差分についての議論はない．このような知覚特性がある場合，単純に連続的なパラメータ値を設定すれば連続的に知覚されるとは限らず，知覚特性を明確化していく実験や分析が必要となる．

本稿では，歌声の声色とその変化に含まれる表現性について，知覚的な側面から評価するため，差分を多様化した一対比較を用い，声色表現にもカテゴリー知覚の可能性があるかを追求し，声色の連続的な変化における適切な表情強度について議論する．

2 歌声の声色強度変更手法 ESVM

本研究のアプローチでは，韻律や歌唱者による表情付けの差となる要素を排除し，同一の歌唱者・同一歌唱速度・同一の F_0 における，異なる声色表現を用いた歌声表情付けの，的確で滑らかな補間を狙いとする．

歌声に表情を付与する手法を検討するとき，音声パラメータを直接操作する手法も考えられるが，本研究では既存の表情付き歌声データベースを用いて自然な表情付けを実現することに焦点を当てた．そこで，声の個性と自然性の両者を保ちつつ既存のデータから音声を合成できる技法として，STRAIGHT[6]を用いた音声モーフィング[7]を適用した．様々な表情付け強度の歌声を音声モーフィングにより合成するため，まず，特定歌唱者の様々な表情付けを伴う歌声を収録した．

表 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

課題曲を童謡「ふるさと」として，20 代のアマチュア女性歌唱者の歌声をサンプリング周波数 44.1kHz で収録した．歌唱速度と F_0 を統一するため，歌唱者は，すべての歌声収録において同一の伴奏をヘッドフォンで聞きながら収録した．

様々な表情付けの中から，歌唱バリエーションに関するアマチュア歌唱者のスキルの限界を考慮し，表情付けの種類を絞った．また，表情付け同士が重複するパターンも可能になるよう，収録の際指示した表情付けは

1. 表情付けのない平坦な声色 “normal”
2. オペラ歌手の発声のような声色 “dark”
3. 子守唄のようにささやいているような声色 “whispery”
4. ポップシンガーが感情を込めたような声色 “wet”

の 4 種類とした．歌唱中は一貫した表情付けで歌うよう教示した．結果として収録された歌声は，“dark”: 後舌母音がかった太い声色，“whispery”: ホワイトノイズがかった声色，“wet”: 鼻母音がかった声色となった (表 1)．

次に，歌声の表情付けの強度や種類を多様化させるため，収録した複数の表情付き歌声に対し，STRAIGHT を用いた音声モーフィングを適用した．様々な強度で表情付けられた歌声を合成するため，表 2 の A-1 から A-3 のように表情のな

表 2　ESVM between without–with expressions

abbr.	base	target
A-1	normal	dark
A-2	normal	whispery
A-3	normal	wet

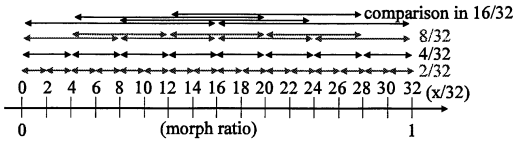

図 1　Sets of difference for paired comparisons

い歌声 “normal”(表情付け 0) から表情付き歌声 “dark” “whispery” “wet”(表情付け 1) へのモーフィングを行った.

音声モーフィング技法では F_0 や話速だけではなく，スペクトル情報についても，base から target への特徴量の補間が行われ，base を 0，target を 1 としたときの数値をモーフィング率と呼ぶ．本研究では声色の表情付与における補間 (内挿) による知覚的連続性の効果を調べるため，まずは等間隔 (0.167 間隔) でモーフィング率を設定した．また，補間だけではなく補外 (外挿) による知覚的効果も検討するため，-0.333 (-2/6) から 1.333 (8/6) まで，変化を伴わない静的な一定の表情付与を施されたモーフィング歌声を作成した.

3　声色表情付与に対する知覚強度の評価

本節では，同一歌唱者による F_0 や話速の同一の時の，声色のみの表情付与に関する知覚特性を明らかにすべく，異なる表情強度の声色を一対比較により評価することとした．これまでに STRAIGHT を用いた音声モーフィング手法による声色の連続的な補間性質は，非線形ではあるが確認されている [8]．しかしその知覚曲線は sigmoid 関数に近似され，カテゴリー知覚が起こっている可能性を否めない．これは，文献 [8] において，本稿で述べる声色表情強度の異なるモーフィング音声に対する絶対評価値の平均が，sigmoid 関数に近似される曲線を描いたことによる．つまり中間地点 (モーフィング率 0.5 付近) における比較的急激な知覚強度の変化が見られたことから,「表情付与無し」と「表情付与あり」の間にカテゴリー知覚の可能性が示されたとも考えられる．文献 [8] では，モーフィング率の補間間隔を逆 sigmoid 関数で調整することで線形的な知覚曲線を得られることが示されたが，これはカテゴリー知覚を完全に否定するものではない.

一方，文献 [9] における，モーフィング率の異なる音声刺激の一対比較では，比較間隔が規定の間隔であり，全ての組において差分に関する正答率が 0.5 近辺となっていた．この結果では，全ての比較対において差分が知覚されておらず，カテゴリー知覚のピーク値の存在を否定できない．そこで我々は，比較対のモーフィング率間の差分を 0.0625, 0.125, 0.25, 0.5, と多様化することとした．また，各差分において，表情強度の昇順の刺激提示と降順の刺激提示を準備し比較することで，1) それぞれの差分値における一対比較時の正答率にピークが存在するかどうかでカテゴリー知覚の可能性を探り，2) 降順と昇順の提示順序による差分の知覚能力に差が出るかどうかを検討する．そして，これらの結果を踏まえ，声色表情付与における適切な変化曲線を検討する.

実験仮説: I) 被験者が表情強度の強弱を正しく認識できる確率 (以後弁別率と呼ぶ) は，モーフィング率間の差分が等しい比較群の中に，明確なピークが存在する．II) 表情強度の弁別率は，刺激提示の順序 (降順・昇順) に影響されない.

実験手法: 一対比較をする際の実験刺激として，表情強度の a) 異なる差分と b) 異なる表情強度の区間において，刺激対を構成し，提示した (図 1 参照).

被験者: 聴覚に問題のない 20 代から 30 代の 20 名 (男性 10 名，女性 10 名)

実験条件および実験刺激: 実験刺激は表 2 の A-1 から A-3 を用いて，表情無し “normal” から “dark, whispery, wet” の表情付与音声の間で **32 段階**のモーフィングを施した刺激を用いた．その後，二つの刺激音声の比較対を図 1 に示すように (モーフィング率の差分が 0.0625, 0.125, 0.25, 0.5)，表情強度の昇順・降順のそれぞれにおいて作成した．比較対の音声の間に 0.5 秒の無音区間を設けた.

実験手順: 被験者は，刺激音声としてモーフィング音声の比較対を聴き，前者と後者のどちらの刺激がより表情付与が強く感じられたかを二者択一で評価するよう，実験前に教示された．判別試験

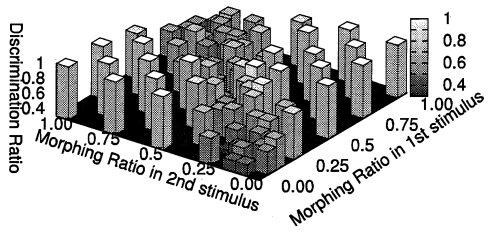

図 2　Discrimination ratios (A-1)

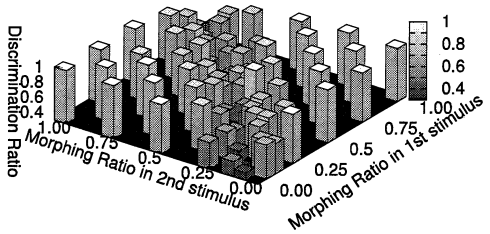

図 3　Discrimination ratios (A-2)

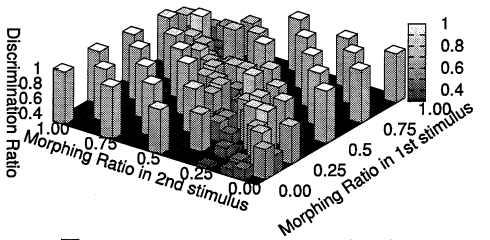

図 4　Discrimination ratios (A-3)

の確率は，理論上，被験者により刺激間に差が認識されなかった場合 0.5 となり，差分が認識された場合は 1 となる．

実験結果: 図 2〜4 に，A-1 から A-3 のモーフィング音声ごとの，各一対比較の判定結果の正答率とその標準偏差を示す．これらについて，モーフィング率の各比較差分ごと，および刺激提示における順序（昇順・降順）ごとの正答率を示したものが図 5〜7 である．

図 2〜4 を見ると，2 つの刺激間でのモーフィング率の差分が 0.5 より大きいとき，弁別率はほぼ 100 パーセントになっている．弁別率はその比較対のモーフィング率の差分が小さいほど低い値となっていることも分かる．一方，図 5〜7 を観察すると，同じ刺激対の比較において，降順・昇順の間で異なる弁別率が観察されていることも分かる．これは仮説 II を棄却する結果とも考えられる．そ

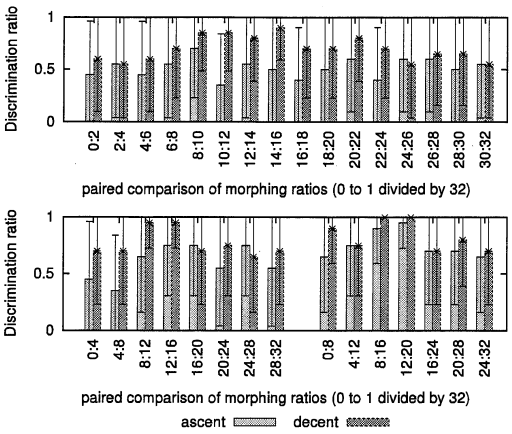

図 5　Order-separately discrimination ratio A-1

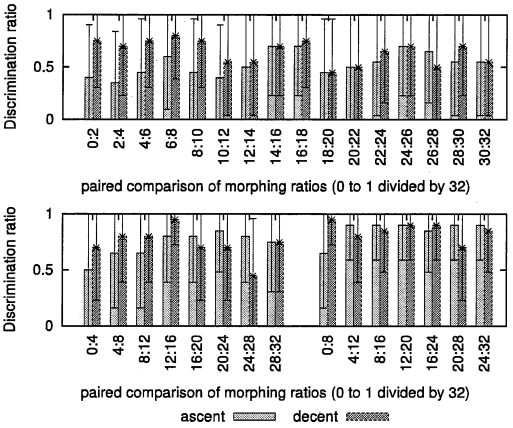

図 6　Order-separately discrimination ratio A-2

こで，刺激対のモーフィング率の中心位置，および，刺激対間のモーフィング率の差分のそれぞれにより，弁別率が存在するかどうかを調べるため，以下のとおり統計的検定を行った．

被験者の弁別率について，2 要因分散分析（$\alpha = 0.05$，要因 1 はモーフィング率の位置，要因 2 は刺激対の提示順序= 降順および昇順）を A-1 から A-3 のそれぞれで行った．結果を表 3 に示す．上記 ANOVA で有意な結果が見られた部分について Scheffé の Post-hoc 検定を施し，その結果有意差があることが確認されたものについて表 3 中の数値にマークをつけた．

以下に，有意傾向を含めた分析結果を示す．まず，i) 刺激提示において 2/32(0.0625) および

表 3　ANOVA results of discrimination tests

difference	object	A-1 "normal" ↔ "dark" $\phi^\star$	F	p	A-2 "normal" ↔ "whispery" $\phi^\star$	F	p	A-3 "normal" ↔ "wet" $\phi^\star$	F	p
	sequence orders	1	14.00	<0.01$^\diamond$	1	5.57	0.02$^\diamond$	1	0.29	=0.60
2/32 (0.0625)	morphing ratios	15	0.95	=0.51	15	1.25	=0.23	15	0.92	=0.54
	sequence orders	1	10.19	<0.01$^\diamond$	1	0.15	=0.90	1	2.53	=0.12
4/32 (0.125)	morphing ratios	7	2.37	0.02	7	1.55	=0.15	7	2.03	0.05
	sequence orders	1	2.36	=0.13	1	0.03	=0.87	1	0.29	=0.59
8/32 (0.25)	morphing ratios	6	3.75	<0.01	6	0.50	=0.81	6	5.17	<0.01 $^{\boxtimes}$
	sequence orders	1	0.21	=0.21	1	1.03	=0.31	1	1.38	=0.24
16/32 (0.125)	morphing ratios	4	3.13	0.02	4	0.90	=0.47	4	3.11	=0.07

$\star$ ϕ means DOF: degree of freedom

$^\diamond$: significant (<0.05) in post-hoc test

$^{\boxtimes}$: significant results were found by post-hoc test in pairs [0-8:8-16, 0-8:12-20, 0-8:16-24, 0-8:20-28], divided by 32.

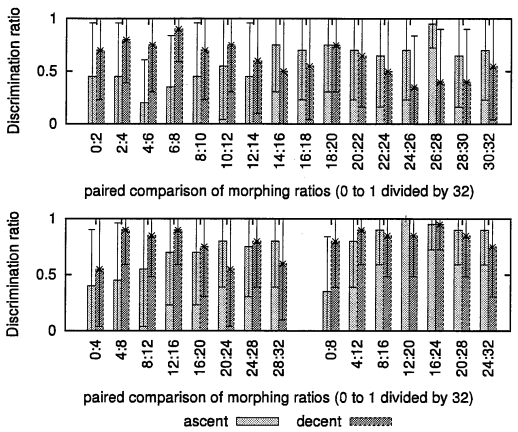

図 7　Order-separately discrimination ratio A-3

4/32(0.125) のモーフィング率の差分があるとき，表情強度の提示順序は弁別率に影響を与えることがわかった．特に降順の刺激提示において弁別率が高いことが示され，仮説 II) は棄却された．また，ii) ANOVA の結果からは，モーフィング率の比較対の中間地点の位置は，比較差分が 4/32(0.125) から 16/32(0.5) のとき，弁別率に影響する可能性が示されたが，Post-hoc 分析により，モーフィング率 0/32(0.00) と 8/32(0.25) の比較時にのみ有意傾向が見出されたに過ぎなかった．従って，差分の弁別比率に鋭いピークは見つからず，実験の仮説は棄却 I) はされたといえる．

4　考察

歌声の声色表情強度に対する知覚に関して，様々な表情強度とそれらの差分において一対比較実験を行った結果，カテゴリー知覚的な特性としての差分知覚の明瞭なピークは見出されなかった．これは，声色という表情付与の一手段が連続的に知覚されることを示しており，さらにその補間によるスムージングが，徐々に変化する表情強度を表すときに有効である可能性をも示している．カテゴリー知覚であるか否かの議論はさらに同定実験と単純な弁別実験を持って行われるべきであり，本実験では非カテゴリー知覚の可能性を示唆したに過ぎない．しかしその可能性は今後の音声モーフィングによる連続的な変化を伴う表情付与の有用性を十分に示しているといえよう．

一方，表情強度の一対比較において，昇順の刺激提示に鈍感であったかのような分析結果が得られたが，0.5 以上の刺激同士の比較の場合は必ずしも降順に対して弁別率が高かったといえない結果も見られる．例えば図 7 上段の後半部分において，むしろ昇順の刺激提示のほうが弁別率が高いようにも見える．また，図 5，6 上段においても，降順の弁別率が明確に高いのは前半のみに限られるようにも見える．よって，全般の表情強度が弱めの場合，強度を増加させるときには急激に，減少させるときには緩やかに，変化曲線を描くように調整することを提案する．また，このような特性をさらに明らかにし，滑らかな表情付与のための変化曲線を設計することが今後の課題である．

5　おわりに

本稿では，擬人的媒体とのインタラクション等に伴い必要となる自然な音声表現の可能性として，連続的な声色表情付与を狙いとし，STRAIGHT を用いた音声モーフィングを同一歌唱者の表情無し歌声–表情付き間歌声間に適用し，その知覚について調べるため，モーフィングによる合成音の声色の

表情強度に対して一対比較を行った．そして，歌声の声色表情強度の知覚がカテゴリー的でない，という仮説を支持する結果を得た．また同時に，歌声の声色の一対比較において，第一刺激に比べて第二刺激の表情強度が増す場合では（昇順），減る場合（降順）に比べて，刺激差分の知覚が比較的鋭敏になることがわかった．そこから，歌声の表情強度の付与に関する指針として，変化が昇順の場合と降順の場合に，それぞれ異なる変化曲線を与えることを提案した．今後は表情強度が時間に伴い変化する時に，本稿で示唆した昇順と降順の異なる変化手法が作用するかどうかを検証していきたい．

謝辞

本研究の一部は科研費 20700106 の助成を受け推進したものである．STRAIGHT 音声モーフィングの使用をご快諾くださった和歌山大の河原英紀氏，弁別実験について議論してくださった北陸先端科学技術大学院大学の赤木正人氏，その他，実験にご協力くださった ATR の皆様に感謝する．

参考文献

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., 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 reptitive structure in sounds," Speech Communication, 27, pp.187–207, 1999.

7) 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.

8) T. Yonezawa, N. Suzuki, S. Abe, K. Mase, and K. Kogure, *Perceptual continuity and naturalness of expressive strength in singing voices based on speech morphing*, EURASIP Journal of Audio Speech Music Process, vol.2007, no.3, ISSN:1687-4714, 2007.

9) 筧一彦，曽我部優子，河原英紀，**表情と感情音声の知覚**，信学技報 TL2005-13, pp.31–38, 2005.

10) IRL Ozgen, E; Davies. Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. Journal of Experimental Psychology-General, 131(4):477.493, 2002.

11) J. E. Flege, M. J. Munro, and R. A. Fox. Auditory and categorical effects on cross-language vowel perception. Journal of Acoustical Sciety of America, 95(6):3623.3641, 1994.

12) A. Suzuki, S. Shibui, and K. Shigemasu. Temporal characteristics of categorical perception of emotional facial expressions. Proceedings of the 26th Annual Conference of Cognitive Science Sciety, pages 1303.1308, 2004.