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

2008－MUS－74 (6)
2008－SLP－70 (6)
2008／2／8

好みの歌唱様式による歌詞朗読音声からの歌唱合成

森山 剛† 小沢 慎治††

† 東京工芸大学工学部メディア画像学科 〒243–0297 神奈川県厚木市飯山1583
†† 愛知工科大学工学部情報メディア学科 〒443–0047 愛知県蒲郡市西迫町馬乗50–2
E-mail: †moriyama@mega.t-kougei.ac.jp, ††ozawa@aut.ac.jp

あらまし 自分の声で好みの歌唱様式を有した歌声を作るには，歌唱を訓練し，歌唱様式を学び，さらに歌声を発する恥ずかしさを克服しなければならず，非常に困難である．そこで，歌詞を朗読するだけで，その音声と楽譜を入力とし，ユーザの声の話者性を損なわずに，自由な歌唱様式を有した歌唱を合成する手法を提案する．本手法により，楽譜さえあれば，どんな曲でも，いつでもどこでも何度でも，自分の声で歌わせることができ，さらにリズムや旋律を工夫してジャズや演歌といったジャンルを演出したり，他人の歌い方を真似したり，音痴を修正したりできる．楽譜を編集する過程で，歌唱様式をどう実現すれば良いか学習でき，また自分が歌う前に，自分が歌った場合のイメージを掴むこともできる．聴取実験により，朗読音声の話者性や歌詞の音韻性を損なわず，歌唱様式の基本となる演奏記号を合成できることが示された．

Transformation of Reading to Singing with Favorite Style

Tsuyoshi MORIYAMA† and Shinji OZAWA††

† Dept. of Media and Image Technology, Tokyo Polytechnic University Iiyama 1583, Atsugi-shi, Kanagawa, 243–0297 Japan
†† Dept. of Information Media, Aichi University of Technology 50–2, Manori, Nishihazama-cho, Gamagori-shi, 443–0047 Japan
E-mail: †moriyama@mega.t-kougei.ac.jp, ††ozawa@aut.ac.jp

Abstract We propose a method of transforming reading speech of lyrics to singing voice. It is capable of realizing favorite style in the transformation, i.e., a specific genre and an expression. Generating singing voice by one's own voice requires the person to train singing, to learn how to realize singing styles, and to overcome hesitation in singing out. The proposed method only requires the user to read the lyrics. It then allows the user to generate singing voice of any music, anytime, anywhere, and any number of times. The user can edit the music for generating a specific singing style, mimicing other's style of singing, and correcting the problematic portion of his or her singing. The user can also learn how to realize a specific style in singing and hear how it sounds when he or she sings on his or her own. Experimental results demonstrated that the proposed method was able to synthesize a comprehensive set of basic indications such as *crescendo* in the synthesized singing holding the voice quality of the speaker.

1. はじめに

1980年代初頭にMIDIが策定され，また同時期にパソコンが急速に低価格化及び高性能化したことで，DTM (DeskTop Music) といった言葉に象徴されるように，個人が音楽を作曲編集する環境が急速に整えられた．これは主に，楽器の音源を用意して，五線譜に記述した楽曲を演奏させる方式であり，YAMAHAから2004年に発

売された歌声合成システム VOCALOID [1] は，それを歌声に拡張したものの一つである [2]〜[4]．また，インターネットの普及により，自分の音楽作品を自分のウェブサイトを通して配布したり，自己紹介の一部としたりすることで，個人固有の情報を発信することが一般的になった．

しかし現状では，音楽編集ソフトウェアを開いても，五線を表示し入力を待つだけで，経験や才能に恵まれないユーザにとっては，何をすべきか全くわからない．また，音楽は絵画に比べて，楽器や歌をただ奏するだけでも，他人の鑑賞に堪えるクオリティを達成することは難しく，表現の面白さや独自性を発現するレベルへは，到底至らないのが現状である．殊更に歌は，自分の歌声を作ろうと思っても，歌を練習し，好みの歌唱様式（ジャンルや表現方法）を練習し，さらに歌声を発する恥ずかしさを克服しなければならず，非常にハードルが高い．

著者らはこれまで，合成したい歌唱の，楽譜と歌詞の朗読音声とを入力とし，朗読音声の話者性と歌詞の音韻性を損なわずに歌唱を合成する研究を行ってきた [5], [6]．歌唱では，五線譜による音高や音長の指示だけでなく，その楽曲の作曲された時代やジャンル，人間の声の特性や演奏上の効率，さらに演奏者の個性や音楽的解釈（以下，歌唱様式と呼ぶ）が加わることによって，豊かな音楽表現が可能となる．従って，自然な歌唱を合成するには，このような様々な歌唱様式を制御することのできる自由度を有する合成手法が必要である．

2. 歌い方を決定する要因

歌唱は単旋律であるから，音楽の三要素（リズム，旋律，和声）のうち，リズムと旋律が主たる要素である．すなわち歌唱の歌い方とは，対象楽曲を歌う際に，目標とする歌唱様式を実現するために，リズムと旋律を適切に操作する方法である．

本研究では，歌唱の歌い方を決定する要因として，譜面通り「普通に」歌うための基本技術による要因と，さらにジャンルや音楽表現といった歌唱様式，言わば「色付け」を行うための応用技術による要因，の二つに分類する．また歌い方以前に，ピッチ（声の高さ）の滑らかな遷移や自然な揺らぎといった，人の歌声が自然に備える要因がある．

2.1 人の歌声の自然性による要因

人の歌声は，声道の弾性による調音結合によって，異なる音高でかつ隣り合う音符におけるピッチの遷移が滑らかになる（階段状に遷移するピッチは機械的に聴こえる）．

また，同じ音韻及び音高が持続する場合，ビブラートと呼ばれる，ピッチの自然な揺れ（周波数変調）がかかる（この現象はベルカント唱法に特徴的である）．

このような，人の歌声の自然性によって発生する歌い方の要因を，自然性要因と呼ぶこととする．

2.2 歌唱の基本技術による要因

歌詞には，母音と有声子音，無声子音に加えて，促音に見られるような空白が含まれる．歌唱の旋律は，このうち声帯の振動を伴う有声音によって，楽譜で決められたリズム通りに紡がれる必要がある．しかし，無声子音から始まる場合（例：「あした」の「し」），音符の開始時刻から無声子音（この例では/sh/）を始め，それに母音（この例では/i/）を続けると，拍頭より遅れて有声音が始まるため，本来のリズムから遅れて聴こえてしまう．これを避けるために，実際の歌唱では，無声子音を拍頭よりも早く始め，母音が拍頭で始まるようにする「子音の先取り」と呼ばれるテクニックが使われる．

また，一つの呼吸で一つのフレーズ（音楽的なまとまり）を，自然な呼吸によって歌う場合，p（弱い声）から始め，フレーズの中心に向かって自然な crescendo（次第に大きくする） と accelerando（次第に速くする） を伴って，滑らかに f（強い声）に向かって膨らませた後，decrescendo（次第に小さくする） と ritardando（次第に遅くする） を伴って再び p に戻す，メッサ・ディ・ヴォーチェと呼ばれるテクニックが用いられる [7]．

このような，歌唱の基本技術で決定される歌い方の要因を，基本要因と呼ぶこととする．

2.3 歌唱様式のための応用技術による要因

ジャンルや音楽表現といった歌唱様式において，リズムや旋律につけられる変化は，演奏記号によって記述される．演奏記号とは，1) 演奏の速度を与えるもの，2) 演奏の速度の揺らぎを与えるもの，3) 演奏の強弱を与えるもの，4) 曲想を示すもの，5) 音と音の切り方や結び方を与えるもの，6) （楽器固有の）奏法を与えるものを含む．特定のジャンルや音楽表現を演奏する際には，これらの演奏記号を時系列に組み合わせて，リズムや旋律を適切に操作する．

このような，歌唱様式のために演奏記号によって決定される歌い方の要因を，色付け要因と呼ぶこととする．

3. 提案する歌唱合成法

図 1 に，提案する歌唱合成法の概要を示す．まず，ユーザが歌詞を朗読した音声から，ピッチ及びパワー（声の大きさ）の時間軌跡を，さらに音声認識器により音素位置をそれぞれ計算する．音声認識誤りは，手動で補正する．また，対象楽曲の楽譜から，音高，音量，音長の時系列を決定し，以下に述べる自然性要因，基本要因，色付け要因を考慮して，合成するピッチ及びパワーの軌跡を得る．次に，朗読音声を音節ごとにセグメントし，ピッチ

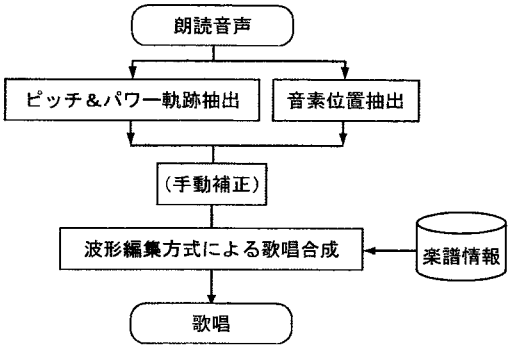


図 1　波形編集による歌詞朗読音声からの歌唱合成

表 1　楽譜情報の例

1	2	3	4	5	6	7	8	9
こ	8	H2	54				cresc. p-f	
の	8	E3						
み	8	Fi3					f	
ち	4	Gi3						
は	4	Gi3				tie	decresc. f-mf	
	8	Gi3						
い	8	Ci4				ten.	mf	
つ	8	H3						
か	8	Gi3						
き	8	Fi3						
た	8	Gi3						
み	4	Fi3						
ち	4	E3						

同期波形重畳法（PSOLA）[8] を用いて，合成するピッチ及びパワーの軌跡を用いて歌唱を合成する．PSOLA では，無声音（空白及び無声子音）は，朗読音声から切り出されたまま用い，有声子音は，ピッチとパワーのみ変更し，母音は，ピッチとパワーに加えて長さも変更する．

3.1　楽譜情報のコーディング

五線譜から，表 1 に示す例のような楽譜情報に変換する．1 列目から歌詞，音符長，音高，速度（全体），速度（音符），装飾（全体），装飾（音符），強弱（全体），強弱（音符）を示す．

3.2　自然性要因のパラメータ化

2.1 節で述べた自然性要因を，ピッチ軌跡に関してパラメータ化する．

3.2.1　ピッチの遷移のさせ方

異なる音高で隣り合う音符間で生ずる，ピッチの滑らかな遷移を，式 (1) のように表現する．

$$\tau(t) = \tau_0 + (\tau_1 - \tau_0)\frac{1-\cos\frac{\pi t}{T_t}}{2} \qquad (1)$$

ここで，T_t は遷移時間，τ は合成音声の対数ピッチ周期，τ_0 及び τ_1 は遷移前，遷移後の対数ピッチ周期である．

3.2.2　ビブラートのさせ方

同じ音韻及び音高が続く際に生ずる，ビブラートの揺れの周波数を 6 回/秒とし，式 (2) を合成するピッチの対数軸に掛けることによって表現する．

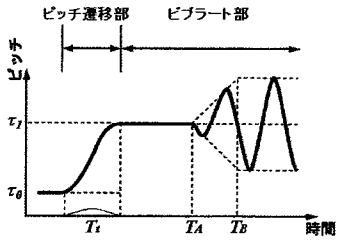


図 2　自然性要因のパラメータ化による合成ピッチ軌跡

$$\tau(t) = \tau_0 + A\sin 12\pi t \qquad (2)$$

$$A = \begin{cases} 0 & 0 \leqq t < T_A \\ \frac{t}{T_B - T_A} A_{\max} & T_A \leqq t < T_B \\ A_{\max} & T_B < T \end{cases}$$

A はビブラートの振幅，T_A は揺れ開始時間，T_B は振幅が最大になる時刻である．

ピッチの自然な遷移とビブラートを考慮することにより，図 2 に示すようなピッチ軌跡となる．

3.3　基本要因のパラメータ化

2.2 節で述べた基本要因を，ピッチとパワーの軌跡に関してパラメータ化する．

3.3.1　歌詞の乗せ方

「子音の先取り」を実現するために，無声子音と母音から構成される音節を音符に乗せる際，母音を音符の開始時刻から始まるように配置する．

3.3.2　メッサ・ディ・ヴォーチェのさせ方

メッサ・ディ・ヴォーチェを実現するために，図 3 に示すように，フレーズの歌い始めと歌い終わりにパワーの変化をつける．音の立ち上がる時刻を T_0，立ち下がる時刻を T_1 とし，式 (3) のように表現する．

$$p(t) = \begin{cases} 0.5 - 0.5\cdot\cos\frac{\pi(t-T_0)}{T_u} & T_0 \leqq t \leqq T_0 + T_u \\ 0.5 + 0.5\cdot\cos\frac{\pi(t-T_1)}{T_d} & T_1 \leqq t \leqq T_1 + T_d \end{cases} \qquad (3)$$

ここで，T_u は立ち上がりに，T_d は立ち下がりに，それぞれかかる時間である．

3.4　色付け要因のパラメータ化

2.3 節で述べた色付け要因に関して，音高，音量，音長に関して演奏記号を分類したものを表 2 に示す．表の一つのセルに属する演奏記号のうち，同じパラメータで表現できるものをクラスとしてまとめ，クラスに対して共通のパラメータ化を行う．

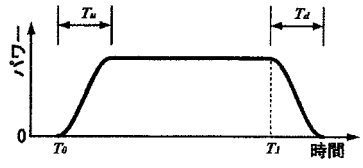
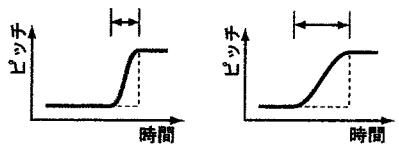


図 3　メッサ・ディ・ヴォーチェの模式図

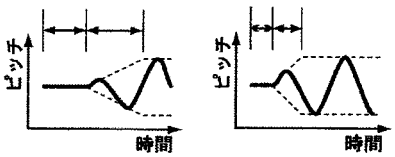
表 2　色付け要因の構成要素となる演奏記号

		対音高	対音量	対音長
対範囲の指示	一様		**forte クラス** - fortississimo - fortissimo - forte - mezzo forte - mezzo piano - piano - pianissimo - pianississimo	**largo クラス** - ♪=n - largo - adagio - lento - andante - moderato - allegro - vivace - presto
	推移		**cresc. クラス** - crescendo - decrescendo	**accel. クラス** - accelerando - ritardando
対音符の指示		portamento trill turn tenuto	marcato **staccato クラス** - staccatissimo - staccato - mezzo staccato **accento クラス** - accento - rinforzando - forzato - sforzato - sforzando	fermata



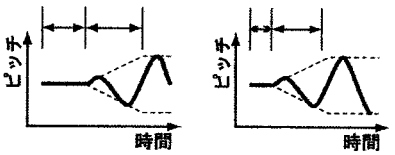
(a) portamento 無し　　(b) portamento 有り

図 4　portamento の表現



(a) trill 無し　　(b) trill 有り

図 5　trill の表現



(a) tenuto 無し　　(b) tenuto 有り

図 6　tenuto の表現

(a) 置換前の楽譜　　(b) 置換後の楽譜

図 7　turn における楽譜情報置換

3.4.1　音高に関するもの

a）portamento のパラメータ化

音高の遷移を極端にゆっくりと行うことにより,「音をずり上げる」効果が出る．通常の音高移動では，3.2.1 節で述べたように，図 4(a) に示すような短時間のピッチ遷移を伴うが，これを図 4(b) のように，長時間かけてピッチを遷移させることでこれを表現する．

b）trill のパラメータ化

拍頭において，音高を一つ上の音高と連続的に繰り返すことで，旋律を修飾する．これはビブラートが急速に立ち上がる様子と似ているため，本手法では図 5(b) のように，ビブラートが安定するまでの時間（$T_B - T_A$）を短くすることで表現する．

c）tenuto のパラメータ化

音符の長さいっぱいに音を延ばす歌い方を，図 6 に示すように，ビブラート開始までの時間を短くすることで表現する．

d）turn のパラメータ化

trill の一種であるが，図 7(b) のように楽譜情報を置換することで表現する．

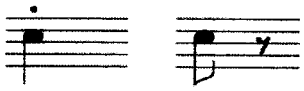
3.4.2　音量に関するもの

a）forte クラスのパラメータ化

fortississimo, fortissimo, forte, mezzo forte, mezzo piano, piano, pianissimo, pianississimo のそれぞれについて，パワーの絶対値を与えることで表現する．

b）cresc. クラスのパラメータ化

目標の音量に向かって，次第に音量を増大（crescendo）もしくは減少（decrescendo）させる歌い方を，次の式 (4) で与えられるゲイン $g(t)$ を元のパワーに掛け合わせることで表現する．



(a) 置換前の楽譜　(b) 置換後の楽譜

図 8　staccato における楽譜情報置換

(a) 置換前の楽譜　(b) 置換後の楽譜

図 9　mezzo staccato における楽譜情報置換

(a) 置換前の楽譜　(b) 置換後の楽譜

図 10　staccatissimo における楽譜情報置換

$$g(t) = 1 + (r_c - 1) \times \frac{t}{T_c} \qquad (4)$$

r_c は目標の元のパワーに対する比，T_c は音量の増大もしくは減少を行う区間の長さである．

c） staccato クラスのパラメータ化

音符を音符の長さよりも短く歌う歌い方を，staccato, mezzo staccato, staccatissimo のそれぞれについて，図 8，図 9，図 10 のように楽譜情報を置換することで表現する．

d） accento クラスのパラメータ化

前後の音に比べて強調して歌う歌い方を，次の式 (5) で与えられるゲイン $g(t)$ を元のパワーに掛け合わせることで表現する．

$$g(t) = 1 + r_a \cos \frac{\pi t}{T_a} \qquad (5)$$

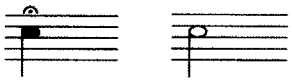
r_a は，accento, rinforzando, forzato, sforzato, sforzando のそれぞれで決定される強度，T_a は音符の長さである．

e） marcato のパラメータ化

一音一音の輪郭をはっきりとごつごつ歌う歌い方を，図 11(b) のように楽譜情報を置換し，accento クラスとしてパラメータ化する．

(a) 置換前の楽譜　(b) 置換後の楽譜

図 11　marcato における楽譜情報置換



(a) 置換前の楽譜　(b) 置換後の楽譜

図 12　fermata における楽譜情報置換

3.4.3　音長に関するもの

a） largo クラスのパラメータ化

♪=n（1 分間に四分音符が n 個入る速さの意味）は指定された速度を用い，largo, adagio, lento, andante, moderato, allegro, vivace, presto のそれぞれについては，速度の絶対値を与えることで表現する．

b） accel. クラスのパラメータ化

目標の速度に向けて次第に速くしたり（accelerando）遅くしたり（ritardando）する歌い方を，式 (6) のように音符の長さを変化させて表現する．ここでは，拍をブロックの集合と考え，拍を構成するブロックの長さを伸縮させることで，拍全体の長さを伸縮させる．

$$T = \sum (\Delta T_0 + (\Delta T_1 - \Delta T_0) \frac{n}{N}) \qquad (6)$$

T は音符の長さ，ΔT は 1 ブロックの長さ，ΔT_0 は 1 ブロックの元の長さ，ΔT_1 は 1 ブロックの目標の長さ，n は音符を構成するブロック数，N は速度を速くしたり遅くしたりする区間のブロック総数である．

c） fermata のパラメータ化

図 12(a) のような指示によって，拍子の運動を停止し，音符を長く伸ばす歌い方であるため，図 12(b) のように楽譜情報を置換することで表現する．

3.5　演奏記号間の関係

演奏記号に，次の 3 つの条件が成立すると仮定する．

- 音高，音量，音長は，それぞれ独立に変更される
- 音符に関する演奏記号による変更は，全体に関する演奏記号による変更に加算する
- 表 2 中の一つのクラスから，二つ以上の演奏記号が同時に現れることはない．

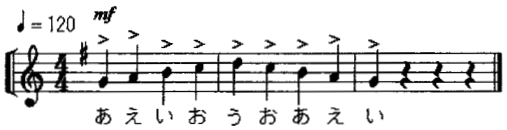


図 13　accento の譜例

表 3　各演奏記号の合成歌唱に対する評定値（平均 ± 標準偏差）

演奏記号	評定値	演奏記号	評定値
accento	4.8±1.8	tenuto	3.8±2.8
sforzato	4.9±2.1	fermata	4.0±2.8
staccatissimo	3.6±2.1	turn	3.5±2.0
staccato	3.7±2.8	portamento	4.0±2.3
mezzo staccato	4.6±2.1	trill	1.7±1.8
cresc. → decresc.	3.6±2.4	decresc. → cresc.	5.4±1.8

4. 歌唱の合成及び聴取実験

声楽科学生の歌唱を分析して決定した各パラメータの値を付録に示す．本手法を用いて朗読音声から歌唱を合成し，音楽大学学生 11 名を被験者とする聴取実験を行った．

4.1 対象楽曲と譜例

図 13 の譜例に示すように，母音の歌詞を用い，表 3 の各演奏記号を有する楽曲をそれぞれ用意した．

4.2 朗読音声と分析方法

演奏記号の表現性能に関しては，女性話者 1 名による母音列の朗読音声，話者性及び音韻性の保存性能に関しては，男性話者 3 名による各 3 曲分の計 9 朗読音声を収録した．44.1KHz サンプリング，16bit 量子化した朗読音声に対して，自己相関法によってピッチ周期を求めた．PSOLA で用いるピッチ波形は，ピッチ周期内のローカルピークを探索し，ローカルピークを中心とする 2 ピッチ分の波形をハニング窓を乗じて切り出した．パワーは，切り出された波形の短時間平均を計算した．音素位置の抽出には，Julius ディクテーションキット [9] を用いた．

4.3 聴取実験結果

各演奏記号に対応する歌唱を被験者に聴かせ，「記号を表現していない」を左端，「記号を十分表現している」を右端に配置した 7 段階評定尺度で評価させた．各演奏記号について，被験者間の平均と標準偏差を表 3 に示す．trill を除いて良好な結果が得られた．話者性及び音韻性については，合成歌唱においていずれも保存されていると評価された．

5. ま と め

歌唱の歌い方を決定する要因を，人の歌声の自然性，歌唱の基本技術，歌唱様式のための応用技術の三つに分類した．それぞれの要因について，歌唱の音高及び音量の時間軌跡と音長を決定する規則を作成し，各規則のパラメータを実際の歌唱から決定した．本規則を用いて合成した歌唱に対して聴取実験を行ったところ，歌詞を朗読した音声の話者性や歌詞の音韻性を損なわずに，種々の演奏記号を表現する自由度を有することが示された．

今後，特定の歌唱様式が，どのような演奏記号の組み合わせで成り立つかを与える規則を体系化し，歌唱様式を指定するだけで，楽譜情報が自動的に編集される仕組みを検討する．

付　　録

3 章の各パラメータの設定値は，以下の通りである．式 (1) の $T_t = 0.1s$，式 (2) の $T_A = 0.25s, T_B = 0.4s$，式 (3) の $T_u = T_d = 0.1s$，portamento における $T_t = 0.5s$，trill における $T_A = T_B = 0.0s$，tenuto における $T_A = 0.0s, T_B = 0.15s$，forte クラスにおけるパワーの絶対値は，fortississimo で 30000，fortissimo で 20000，forte で 10000，mezzo forte で 5000，mezzo piano で 2500，piano で 1000，pianissimo で 500，pianississimo で 250，式 (4) の r_c は，パワーの目標値が設定されていない場合，crescendo で $r_c = 2.0$，decrescendo で $r_c = 0.5$，式 (5) の r_a は，accento で $r_a = 1.0$，rinforzando で $r_a = 1.3$，forzato で $r_a = 1.7$，sforzando で $r_a = 2.0$，largo クラスにおける速度の絶対値は，largo で 60，adagio で 70，lento で 75，andante で 83，moderato で 120，allegro で 146，vivace で 213，presto で 362，accel. クラスでは，4 分音符を 240 ブロックとし，速度の目標値が設定されていない場合，accelerando で $\Delta T_1 = 2.0\Delta T_0$，ritardando で $\Delta T_1 = 0.5\Delta T_0$，fermata では元の音符の 2 倍に伸ばすこととした．

文　　献

[1] YAMAHA: “VOCALOID”, http://www.vocaloid.com/ (2004).
[2] 赤木, 清水：“STRAIGHT を用いた話声からの歌声合成”, 電子情報通信学会技術研究報告, **103**, 154, pp. 13–18 (2003).
[3] 吉田, 中嶌：“歌声合成システム”, 情報処理学会研究報告, **98**, 14, pp. 45–46 (1998).
[4] 酒向, 宮島, 徳田, 北村：“隠れマルコフモデルに基づいた歌声合成システム”, 情報処理学会論文誌, **45**, 3, pp. 719–727 (2004).
[5] 新田, 森山, 小沢：“歌詞朗読音声から歌唱への変換”, 日本音響学会講演論文集, 2-5-11, pp. 645–646 (1997).
[6] 新田：“演奏記号を考慮した合成規則に基づく歌唱合成システム”, 修士論文, 慶應義塾大学大学院理工学研究科電気工学専攻 (1998).
[7] 大賀：“美しい日本語を歌う”, 第 4 章, pp. 59–85, カワイ出版 (2003).
[8] 阪本, 斉藤, 鈴木, 橋本, 小林：“波形重畳法を用いた日本語テキスト音声合成システムについて”, 信学技報, SP95-6 (1995).
[9] 河原, 李：“連続音声認識ソフトウエア Julius”, 人工知能学会誌, **20**, 1, pp. 41–49 (2005).