

# VocaListener: ユーザ歌唱とその歌詞を用いた 歌声合成パラメータの自動推定システム\*

中野倫靖, 後藤真孝 (産総研)

## 1 はじめに

歌声合成システムは、個人が歌唱付き楽曲を制作するのを容易にし、歌唱の表現を自在にコントロールできる重要なツールである。また、インターネットを介した音楽の共同制作や新しいコミュニケーションを生み出している現状がある。さらに、高品質な歌声合成を目指すことは、人間の歌声知覚・生成機構の解明に繋がる取り組みでもある。歌声合成における目標の一つは、人間らしい自然な歌声を合成することにあるが、それに加えて、歌唱の表情（どう歌わせるか）を入力するインタフェースを考えることも重要である。

従来、歌声合成における歌唱の表情付けのために、表情パラメータを細かく調整できる方式 [1] があったが、ユーザによっては自分の望む歌声を作るのを困難にしていた。一方で、そのような表情付けを容易にするために、人間の歌唱音声から音高や音長などを抽出して表情パラメータとする研究があった [2]。しかし、歌声合成の条件（歌声合成システムやその音源データ）が変わると、同じパラメータを与えても合成結果の音高や音量が異なってしまう問題があった。

そこで本研究では、合成された歌唱の音高・音量を目標（ユーザ歌唱）と比較しながら、合成パラメータを反復更新していく VocaListener を提案する。これにより、歌声合成の条件の違いに依存せずに、自然な歌声を合成できる。さらに、目標自体を編集することで、ユーザ自身が歌唱できない表現（音高が声域より高い場合など）も合成できる機能も提案する。

## 2 VocaListener: ユーザ歌唱を真似る歌声合成パラメータ推定システム

本研究では、合成歌唱を目標へ近づけるコア技術を VocaListener-core、目標の音高・音量を編集する技術を VocaListener-plus と呼ぶ。また、それぞれに必要な要素技術を VocaListener-front-end と呼ぶ。

図 1 にシステム全体の流れを示す。ユーザは歌唱音声とその歌詞を入力として与え (A)、システムの分析が誤った箇所の訂正 (B, C, F) や、目標の編集 (D) を行い、最終的に目標を真似た合成歌唱を得る。以降、処理の流れに沿ってそれぞれの技術を説明する。

### 2.1 VocaListener-front-end: 要素技術

VocaListener-front-end は、歌声分析及び歌声合成に関する要素技術群である。これらの要素技術は、状況に応じて任意の手法を利用できる。これ以降、歌唱音声はサンプリング周波数 44.1kHz のモノラル音声信号を扱い、分析における処理の時間単位は 10 msec とする。また、合成パラメータとの区別を明確にするため、分析によって得られた値は観測値と呼ぶ。

歌声分析においては、音高 ( $F_0$ ) 推定 [3]、音量の計算、ビブラート区間の推定 [4] を行った。これ以降  $F_0$  ( $f_{\text{Hz}}$ ) は、次式で MIDI ノートナンバーに対応する単位の実数値 ( $f_{\text{Note\#}}$ ) へ変換して扱う。

$$f_{\text{Note\#}} = 12 \times \log_2 \frac{f_{\text{Hz}}}{440} + 69 \quad (1)$$

ここで、ノートナンバーは 1 が半音に相当する。

また、歌詞の音節毎の発音開始時刻と音長の決定（以降、歌詞アラインメントと呼ぶ）には、音声認識で用いられる Viterbi アラインメントを利用した。音響モデルには、朗読音声用の HMM [5] を、MLLR-MAP [6] によって歌唱音声に適合させて使用した。

歌声合成においては、歌声合成システムとその音源データとして、歌声合成技術 Vocaloid2 [1] の応用商品である初音ミクと鏡音リン（以下、CV01 と CV02）[7] を用いた。VSTi プラグイン (Vocaloid Playback VST Instrument) を用いて合成し、その際には合成パラメータを約 1 msec 毎に線形補間して与えた。

### 2.2 VocaListener-plus: 目標の編集

VocaListener-plus は、歌唱入力表現を広げるために目標自体を編集する機能であり、音高変更機能と歌唱スタイル変更機能の二種類がある。これらは状況に応じて利用し、使わないという選択も可能である。

#### 2.2.1 音高の変更機能

音高の変更機能として、部分的もしくは歌唱全体の音高を指定して変更する「音高トランスポーズ」と音高遷移が半音単位となるように音高を自動的にずらす「調子はずれ (off-pitch) の補正」を提案する。

調子はずれ補正では、有声音と判断された区間毎に次式から  $F_d$  を算出して補正する。

$$F_d = \underset{F}{\operatorname{argmax}} \sum_t \sum_{i=0}^{127} \exp \left\{ -\frac{(F_0(t) - F - i)^2}{2\sigma_i^2} \right\} \quad (2)$$

現在の実装では、 $\sigma = 0.17$  であり、 $F_0(t)$  には事前にローパスフィルタをかけて平滑化を行った。式 (2) は、半音間隔に大きな重みを与える関数であり、 $F_d(0 \leq F_d < 1)$  は、その算出区間の  $F_0$  軌跡がなるべく半音単位の遷移となるように音高を変更する指標となる。

#### 2.2.2 歌唱スタイルの変更機能

目標歌唱の歌唱スタイル（音高と音量の変化）を、ビブラート（音高と音量の周期的な変動）とそれ以外の区間を分けて、抑制・強調できる機能を提案する。

まず、 $F_0(t)$  にローパスフィルタをかけて、歌唱における  $F_0$  の動的変動成分 [8] を除去した  $F_{\text{LPF}}(t)$  を

\*VocaListener: An Automatic Parameter Estimation System for Singing Synthesis by Using User's Singing and Its Lyrics. by NAKANO, Tomoyasu, GOTO, Masataka (AIST)

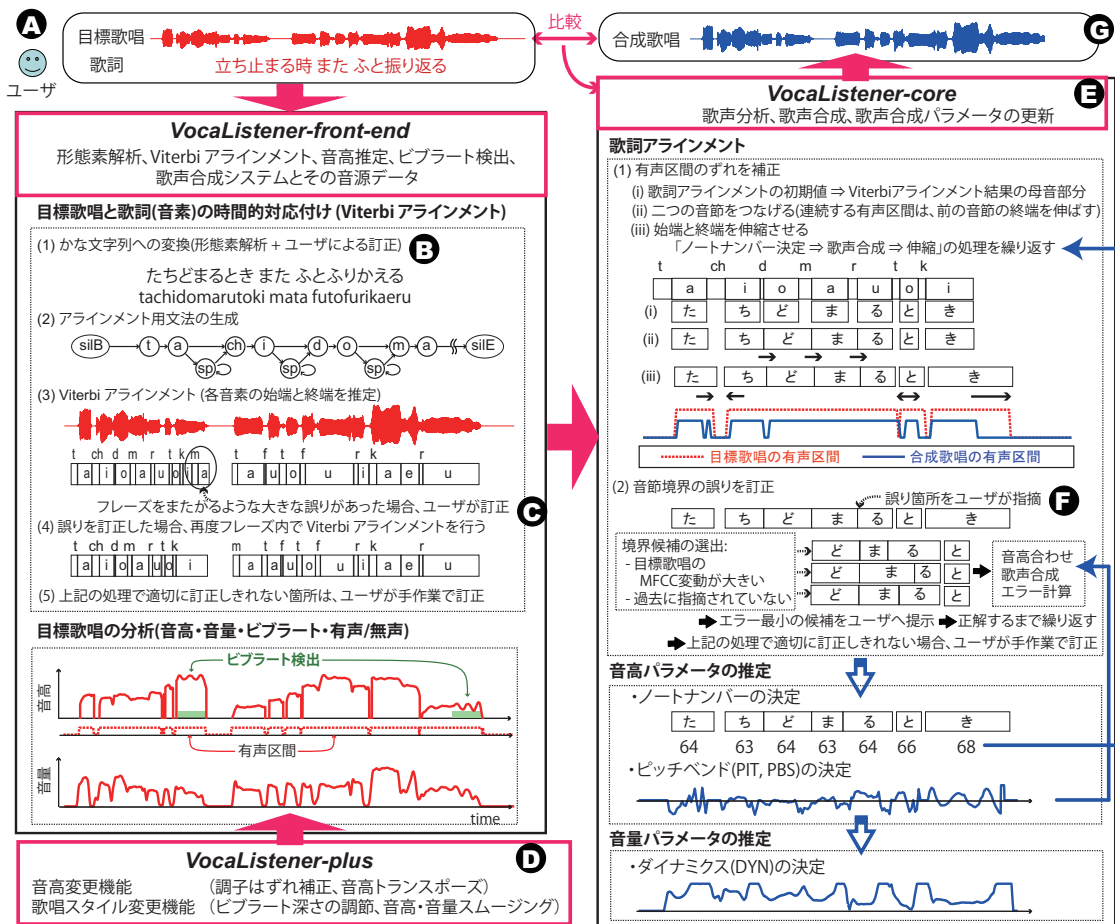


Fig. 1 VocaListener の全体像 ( VocaListener-front-end, VocaListener-plus, 及び VocaListener-core )

得た後、調節パラメータ  $r_v, r_s$  によって次式で変更の度合いを調節する (音量も同様に処理する)。

$$F_0^{(new)}(t) = r_{\{v|s\}} \times F_0(t) + (1 - r_{\{v|s\}}) \times F_{LPF}(t) \quad (3)$$

ここで  $r_v$  はビブラート区間、 $r_s$  はそれ以外の区間に適用し、 $r_v = r_s = 1$  の時に元の歌唱となる。 $r_v$  を 1 より大きくすればビブラートをより強調し、 $r_s$  を 1 より小さくすれば  $F_0$  の動的変動成分を抑制できる。

### 2.3 VocaListener-core: 合成パラメータの推定

VocaListener-core では、歌詞アラインメントと音高・音量パラメータ推定を反復計算によって行う。

#### 2.3.1 初期値の決定

まず、反復計算のための初期値を決定する。歌詞アラインメントには、Viterbi アラインメント結果における母音の開始・終了時刻を初期値として与えた。

音高に関するパラメータは、「音符の音高 (ノートナンバー)」、「ピッチベンド (PIT)」、「ピッチベンドセンシビリティ (PBS)」、音量は「ダイナミクス (DYN)」である。ここで、PIT は音符の音高を相対的に変化させる、各時刻毎に動的に設定できるパラメータであり、PBS によってその相対変化の幅を設定できる。各パラメータの設定可能な値と初期値を表 1 に示す。

PBS が 1 であれば、ノートナンバーから  $\pm 1$  半音の範囲を 16384 の分解能で表現できることになる。

Table 1 推定する歌声合成パラメータと初期値

歌声合成パラメータ		設定可能な値	初期値
音高	ノートナンバー	0 ~ 127	音節毎に設定
	PIT	-8192 ~ 8191	0 (全時刻)
	PBS	0 ~ 24	1 (全時刻)
音量	DYN	0 ~ 127	64 (全時刻)

#### 2.3.2 歌詞アラインメントの推定、及び誤り訂正

歌詞アラインメントでは、Viterbi アラインメントの性能や歌声合成システムの特徴が原因で、指定した発音開始時刻や音長と時間的にずれて合成されることがある。そこで、有声区間のずれを以下の処理によって補正する (図 1 (E), (1) 有声区間のずれを補正)。

- (i) 二つの音節が繋がっておらず、かつ、目標歌唱ではその区間が有声と判定されていた場合、前の音節の終端を次の音節の始端まで伸ばす。
- (ii) 合成歌唱の有声区間が目標歌唱とずれている音節の始端と終端を、一致するように伸縮させる。
- (iii) ノートナンバーを推定して歌声合成し、再度 (ii) を行う (ii と iii の処理を繰り返す)。

続いて、音節境界に誤りがあった場合、ユーザが指摘することで訂正する。システムがユーザに提示する新しい境界候補は、目標歌唱の MFCC の時間変化が大きい上位 3 箇所について、それぞれの候補をまず音高を後述する反復計算で合わせて合成し、目標歌唱との MFCC 距離が最小のものとした。それも誤っていたら、次の候補を提示していく。

最後に、上記の処理で適切に訂正しきれない箇所のみ、ユーザが手作業で訂正を行う。

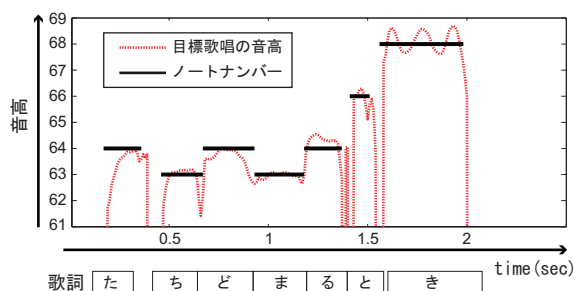


Fig. 2 目標歌唱の  $F_0$  と選択されたノートナンバー

### 2.3.3 音高パラメータの推定 (1): ノートナンバー

観測された音高 ( $F_0$ ) からノートナンバーを決定するために、歌詞の各音節 (音符) 区間に存在する音高の出現頻度から、PBS の値が小さくなるように、以下の式でノートナンバー (Note#) を推定する (図 2)。

$$\text{Note\#} = \underset{n}{\operatorname{argmax}} \left( \sum_t \exp \left\{ -\frac{(n - F_0(t))^2}{2\sigma^2} \right\} \right) \quad (4)$$

ここで、 $\sigma = 0.33$  として計算し、 $t$  は音符の始端から終端の時刻で計算した。これにより、 $F_0$  が長い時間留まっているノートナンバーが選択される。

### 2.3.4 音高パラメータの推定 (2): ピッチベンド

ノートナンバーは固定したまま、合成歌唱の音高  $F0_{\text{syn}}^{(n)}(t)$  が目標歌唱の音高  $F0_{\text{org}}(t)$  に近づくように、反復計算によって PIT と PBS を更新して推定する。

時刻  $t$ 、 $n$  回目の反復における PIT と PBS をノートナンバーに対応する値へ変換したものを  $Pb^{(n)}(t)$  とすると、更新式は以下ようになる。

$$Pb^{(n+1)}(t) = Pb^{(n)}(t) + \left( F0_{\text{org}}(t) - F0_{\text{syn}}^{(n)}(t) \right) \quad (5)$$

このようにして得られた  $Pb^{(n+1)}(t)$  から、PBS が小さくなるように、PIT と PBS を決定する。

### 2.3.5 目標音量の相対値化

目標歌唱の音量観測値は、収録条件の違い等が原因でその絶対的な値が変化するため、相対値化を行う。すなわち、音量の相対的な変化を表現するパラメータを推定するために、目標歌唱の音量を  $\alpha$  倍する。図 3 に、DYN の値を 0 ~ 127 まで変化させた合成歌唱と、目標歌唱の音量観測値をそれぞれ示す。

ここでは、図 3A のような一部の再現は断念し、全体としての再限度が高くなるよう相対値化を行う。そのために、DYN=64 の合成歌唱と目標歌唱のそれぞれの音量観測値の二乗誤差を最小とする  $\alpha$  を算出した。

### 2.3.6 音量パラメータの推定: ダイナミクス

相対値化係数  $\alpha$  は固定したまま、音量パラメータ (DYN) を反復更新する。そのためにまずは、DYN = (0, 32, 64, 96, 127) のそれぞれで実際に合成して音量観測値を算出し、その間を線形補間で求めて全ての DYN における合成歌唱の音量観測値を得る。

時刻  $t$ 、 $n$  回目の反復において、DYN から上述のように求めた音量観測値を  $Dy^{(n)}(t)$  とし、その DYN

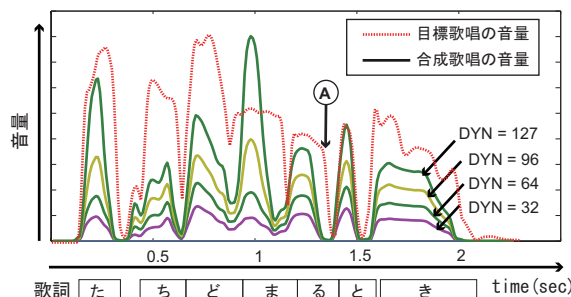


Fig. 3 目標歌唱と合成歌唱の音量観測値の違い

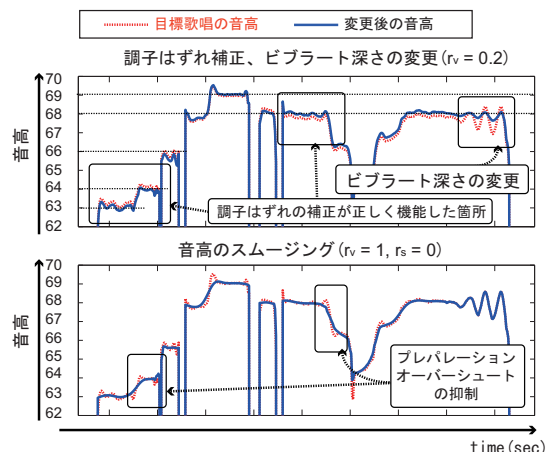


Fig. 4 VocaListener-plus の運用

で実際に合成した歌唱の音量観測値を  $Pow_{\text{syn}}^{(n)}(t)$  とすると、更新式は以下ようになる。

$$Dy^{(n+1)}(t) = Dy^{(n)}(t) + \left( \alpha Pow_{\text{org}}(t) - Pow_{\text{syn}}^{(n)}(t) \right) \quad (6)$$

このようにして得られた  $Dy^{(n+1)}(t)$  から、上述の、DYN とその音量観測値の関係を利用して、音量パラメータ DYN を決定する。

## 3 運用及び評価実験

実際の歌唱を用いた VocaListener-plus の運用結果、VocaListener-core の有効性を評価した結果を示す。

### 3.1 VocaListener-plus の運用

図 4 に、音高変更機能としての「調子はずれ補正」と、歌唱スタイル変更機能を適用した結果を示す。音高が補正されること、ビブラートの深さを変更可能なこと、音高や音量の変動を強調・抑制できることを確認した。ただし、調子はずれの補正については、有声区間毎に補正を行ったため、短い音符に相当するような箇所などが、適切に補正されない場合もあった。

### 3.2 VocaListener-core の評価実験

音源データとして CV01 及び CV02 を用い、VocaListener によって歌声合成パラメータを推定した。Vocaloid2 の合成条件は「ビブラートをつけない」、「バンドの深さを 0%」と設定した以外は全てデフォルト値を用いた。また、目標歌唱としては便宜上、RWC 研究用音楽データベース (ポピュラー音楽) RWC-MDB-P-2001 [9] の伴奏なし歌唱データを用いた。

Table 2 評価実験における目標歌唱及び音源データ

実験番号	曲番号	使用箇所	曲の長さ	目標歌唱 (歌手名)	合成用音源データ
A	No.07	1 番	103 秒	緒方智美	CV01
A	No.16	1 番	100 秒	吉井弘美	CV02
B	No.07	冒頭	2.4 秒	緒方智美	CV01,02
B	No.16	冒頭	3.5 秒	吉井弘美	CV01,02
B	No.54	冒頭	2.7 秒	凜	CV01,02
B	No.55	冒頭	2.9 秒	鍋木 朗子	CV01,02

Table 3 音節境界の誤り指摘数、及び回数 (評価 A)

曲番号	合成用音源データ	音節数	誤り指摘 $n$ 回目の誤り数			
			0 (初期値)	1	2	3
No.07	CV01	166	8	5	2	0
No.16	CV02	128	3	2	0	—

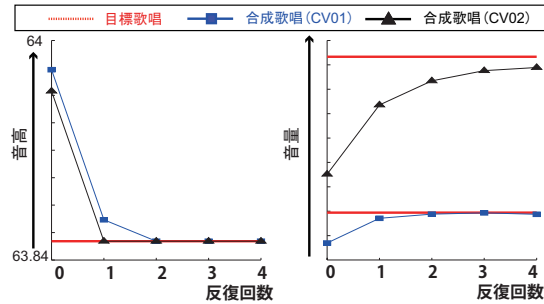


Fig. 5 反復推定による音高・音量の推移 (評価 B)。音量目標値は、CV01 と CV02 で相対値化係数  $\alpha$  が異なる。

Table 4 反復による相対エラー量 [%] (評価 B: No.07)

推定したパラメータ	合成用音源データ	反復 $n$ 回目の相対エラー量			
		1	2	3	4
音高	CV01	13.8	4.7	2.1	2.4
音高	CV02	8.1	3.7	2.3	1.7
音量	CV01	19.8	17.9	17.6	17.5
音量	CV02	16.0	14.2	13.9	13.8

太字は、エラーが増加したことを示す。

評価は、歌詞アラインメントの誤り訂正機能の有効性 (評価 A)、反復推定の必要性、及び音源データの違いに対する頑健性 (評価 B) の観点から行った。それぞれの評価で利用した楽曲を表 2 に示す。

### 3.2.1 評価 A: 歌詞アラインメントの誤り訂正回数

Viterbi アラインメント結果は、No.07 ではフレーズをまたぐ等の大きな誤りは起きず、No.16 では 2 箇所での修正が必要であった。それらを手作業で直した後の音節境界誤り数と、その指摘回数を表 3 に示す。No.07 では、計 166 個の音節について、8 箇所の境界誤りがあり、それらは最大 3 回の指摘で訂正できた。

### 3.2.2 評価 B: 反復推定における相対エラー量

歌詞アラインメントについて人手で正解を与え、以下で定義する相対エラー量 ( $\Delta err_{\{f0\}pow}^{(n)}$ ) を算出した。ここで、 $err_{\{f0\}pow}^{(n)}$  は、反復  $n$  回目における音高観測値もしくは音量観測値の目標との誤差である。

$$\Delta err_{\{f0\}pow}^{(n)} = err_{\{f0\}pow}^{(n)} / err_{\{f0\}pow}^{(n=0)} \times 100 \quad (7)$$

その結果、どの曲に対しても、反復推定によってエラーは減少した。4 回の反復における初期値からの相対エラー量は、音高に関しては 1.7~2.8%、音量に関しては 13.8~17.5%であった。No.07 についての結果を表 4 に、さらにそのうちの一箇所を図 5 に示す。

## 3.3 考察

歌詞アラインメントの結果に誤りがあった場合でも、ユーザが指摘することで訂正できた。ただし、正解の境界位置で合成した歌唱と目標歌唱との MFCC 距離は、最小とならない場合もあった。今後は、このような問題への対処も検討する必要がある。

また、図 5 及び表 4 からは、音源データに依存せず、目標歌唱の音高・音量を近似するパラメータを推定できたといえる。ただし、音高パラメータの反復推定で、エラーが増加することがあった (表 4)。これは、パラメータ推定における量子化誤差が原因と考えられる。このような誤差は音量パラメータ推定にも存在し、場合によってはエラーが若干増加した。しかし、既に高い精度で合成パラメータが得られていることが多く、合成歌唱の品質への影響は少なかった。

## 4 おわりに

本研究は、ユーザ歌唱と歌詞を入力として、それを近似する機能と入力歌唱自体を修正できる機能を持つ、歌声合成パラメータ推定システム VocaListener を提案した。本システムは、ユーザが合成パラメータを一度調整するだけで、歌声合成の条件を自由自在に切り替えながら歌声を合成できるメタ歌声合成システム実現の第一歩であると考えている。

今後は、ブレスの自動検出手法 [11] の利用によるブレス付与や、声質パラメータの推定などによる、より人間らしい歌声の合成を目指す。また、VocaListener-plus の機能拡張や、VocaListener を歌声知覚の研究に役立てる等、様々な観点から研究を行う予定である。

謝辞 本研究の一部は、科学技術振興機構 CrestMuse プロジェクトによる支援を受けました。本研究では、ヤマハ株式会社及び、クリプトン・フューチャー・メディア株式会社の「CV01」「CV02」を使用させて頂きました。本研究に対し有益なご意見を頂いた緒方 淳氏、齋藤 毅氏、藤原弘将氏 (産総研) に感謝致します。本研究では、RWC 研究用音楽データベース (ポピュラー音楽) を使用しました。

## 参考文献

- [1] 剣持, 大下. 歌声合成システム VOCALOID - 現状と課題. 情処研報 2008-MUS-74-9, pp. 51-58, 2008.
- [2] Janer, Bonada, and Blaauw. Performance-driven control for sample-based singing voice synthesis. In *Proc. (DAFx-06)*, pp. 42-44, 2006.
- [3] Camacho. Swipe: A sawtooth waveform inspired pitch estimator for speech and music, Ph.D. Thesis, University of Florida, 116p. 2007.
- [4] 中野, 後藤, 平賀. 楽譜情報を用いない歌唱力自動評価手法. 情処学論, Vol. 48, No. 1, pp. 227-236, 2007.
- [5] 河原, 住吉, 李, 坂野, 武田, 三村, 伊藤, 伊藤, 鹿野. 連続音声認識コンソーシアム 2002 年度版ソフトウェアの概要. 情処研報 2001-SLP-48-1, pp. 1-6, 2003.
- [6] Digalakis and Neumeyer. Speaker adaptation using combined transformation and Bayesian methods, *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 4, pp.294-300, 1996.
- [7] クリプトン. VOCALOID2 特集. <http://www.crypton.co.jp/mp/pages/prod/vocaloid/>.
- [8] 齋藤, 後藤, 鶴木, 赤木. 好みの歌唱様式による歌詞朗読音声からの歌唱合成. 情処研報 2008-MUS-74-6, pp. 33-38, 2008.
- [9] 後藤, 橋口, 西村, 岡. RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース. 情処学論, Vol. 45, No. 3, pp. 728-738, 2004.
- [10] 齋藤, 後藤. 歌声の個人性知覚に寄与する音響特徴の検討. 音講論集 2-Q-26, 2007.9.
- [11] 中野, 緒方, 後藤, 平賀. 無伴奏歌唱におけるブレスの音響特性と自動検出. 音講論集 1-11-12, 2008.3.