楽譜情報を用いない歌唱力自動評価手法

中野倫靖[†] 後藤真孝[††] 平賀譲[†]

本論文では，歌唱力を自動的に評価するシステム開発の第 1 段階として，ポピュラー音楽における歌唱力の「うまい」「へた」を，楽譜情報を用いずに自動的に識別する手法を提案する．従来，訓練された歌唱者の歌唱音声に関する音響学的な考察は行われてきたが，それらの研究は歌唱力の自動評価に直接適用されたり，人間による評価と結び付けて検討されたりすることはなかった．本論文では，聴取者の歌唱力評価の安定性を聴取実験によって確認し，そこで得られた結果から歌唱音声に「うまい」「へた」をラベル付けして自動識別実験を行った．そのための特徴量として，歌唱者や曲に依存しない特徴であることを条件に，相対音高とビブラートの 2 つを提案する．聴取実験では，22 人の聴取者を被験者とし，聴取者間の評価に相関があった組の割合は 88.9%（$p < .05$）であった．また，600 フレーズのラベル付けされた歌唱音声に対して識別実験を行った結果，83.5%の識別率を得た．

An Automatic Singing Skill Evaluation Method for Unknown Melodies

Tomoyasu Nakano,[†] Masataka Goto[††] and Yuzuru Hiraga[†]

As a first step towards developing an automatic singing skill evaluation system, this paper presents a method of classifying singing skills (*good/poor*) that does not require score information of the sung melody. Previous research on singing evaluation has focused on analyzing the characteristics of singing voice, but were not directly applied to automatic evaluation or studied in comparison with the evaluation by human subjects. In order to achieve our goal, two preliminary experiments, verifying whether the subjective judgments of human subjects are stable, and automatic evaluation of performance by a 2-class classification (*good/poor*), were conducted. The approach presented in the classification experiment uses pitch interval accuracy and vibrato as acoustic features which are independent from specific characteristics of the singer or melody. In the subjective experiment with 22 subjects, 88.9% of the correlation between the subjects' evaluations were significant at the 5% level. In the classification experiment with 600 song sequences, our method achieved a classification rate of 83.5%.

1. はじめに

本研究では，人間の歌唱理解能力を備えたシステム開発の一環として，楽譜情報を用いない歌唱力自動評価手法を実現することを目的とする．歌唱力の自動評価は様々なアプリケーションで有用であり，たとえば，歌唱の評価が低い原因を提示することで歌唱力向上を支援するシステムが実現できる．さらに，どのようなメロディにも有効な音響特徴量を明らかにできれば，それを音楽情報検索の検索キーとして活用できる可能性がある．

従来，歌唱音声の特性を明らかにする研究や，人間の歌唱理解に関する研究はあったが，それを歌唱力の自動評価につなげた研究事例はなかった．歌唱音声の特性としては，Singer's Formant ☆が存在すること[1),2)]，基本周波数（以下，F_0 と呼ぶ）には歌唱音声特有の変動があること[3),4)] が明らかとなっている．また，人間の歌唱理解に関しては，歌声知覚における心理的特徴の分析[5),6)] と音響特徴量との関連付け[6)]，歌声らしさを特徴付ける F_0 軌跡に関する考察[7)]，朗読音声と歌唱音声の人間の識別能力に関する調査と自動識別[8)]，歌唱音声の音響解析に基づく歌唱力評価の考察[9)~12)]，などの研究事例がある．ほかにも，歌唱の生成における議論として，歌唱の F_0 制御モデルの提案[7),13)]，歌声生成の総合的かつ詳細な検討[14)] などがある．

本論文では，歌唱力を自動的に評価するシステム開発の第 1 段階として，ポピュラー音楽における人間の

† 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba
†† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

☆ 男性声楽歌手の母音（有声音）において 2.5 kHz から 3 kHz の範囲に発生するフォルマント[1)]．女性歌手についても，4 kHz 付近に顕著なピークが観察される場合がある[2)]．

歌唱力評価を調査し，その結果をもとに「うまい」「へた」を自動識別する手法を提案する．特に，原曲を知らない聴取者の歌唱力評価と，楽譜情報が未知であっても有効な音響特徴量を対象とする．人間の歌唱力評価を調査するのは，人間の評価に安定性があることの確認と，歌唱力評価において着目すべき特徴の検討のためである．自動識別では，聴取実験で明らかになった特徴のうち，原曲や歌唱者の違いに左右されない相対音高とビブラートに関する音響特徴量を用いる．

以下，2 章で人間による歌唱力評価実験について述べ，3 章で自動評価のための特徴を提案して識別実験を行う．最後に，4 章でまとめと今後の研究について述べる．

2. 歌唱力評価の聴取実験

歌唱力の自動評価手法を実現することを目的として，人間による歌唱力の評価が安定していることを示し，人間が着目している歌唱力の特徴を検討する．実験では，複数の歌唱者の歌唱音声（伴奏なし）について，聴取者が原曲を知らない条件で歌唱力を評価させた．聴取者間の評価に高い相関があれば，人間による歌唱力評価は安定しているといえる．人間が着目している歌唱力の特徴は，聴取者の内省報告をもとに検討する．また，聴取者による評価結果を自動評価に利用するために，それぞれの歌唱音声に対して「うまい」「へた」のラベル付けを行う．

2.1 歌唱力評価のための尺度

声を評価する従来の取り組みとしては，英語発音の自動評価があり，自動的に算出した発音の評価値と，聴取者による n 段階評価（$n = 5$, 7 など）の評価値の相関を求める方法がとられていた[15),16)]．しかし，歌唱力評価の場合，聴取者の音楽経験の違いなどが原因で，評価値の基準が聴取者によって異なる可能性，基準が同一でもその距離感（ある 2 つの歌唱の定量的な評価値の差）が聴取者によって異なる可能性がある．

そこで，そのような問題を回避するために，順位法による評価を行う．順位付けによって評価をすれば，その順序関係さえ一定であれば対処でき，距離尺度の違いを吸収した評価値が得られることが期待できる．

本研究では，2 つの順序の類似性を測るために Spearman の順位相関係数 ρ を用いた[17)]．要素数が N の（同順位のない）順序を $\boldsymbol{a} = (a_1, a_2, \ldots, a_N)$ のような順位ベクトルとして表現すると，2 つの順位ベクトル $\boldsymbol{a}$ と $\boldsymbol{b}$ の順位相関係数 ρ は，次式によって定義される．

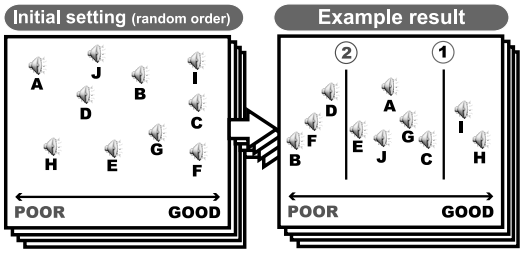

図 1 実験画面と順序付けの例（スピーカのマークは，ダブルクリックすると歌声が再生され，ドラッグすると移動する）
Fig. 1 Example subjective evaluation session using the interface screen.

$$\rho = 1 - \frac{6}{N^3 - N} \sum_{i=1}^{N} (a_i - b_i)^2 \qquad (1)$$

ここで，a_i と b_i は順位ベクトル $\boldsymbol{a}$ と $\boldsymbol{b}$ における要素 i の順位を表す．ρ は，2 つの順序が一致するときには 1 に，互いに逆順であれば -1 になる．N が 10 の場合は $\rho \geq 0.7333$，$\rho \geq 0.5636$ であればそれぞれ，危険率 1%水準・5%水準で有意な相関がある[17)]．

2.2 歌唱力評価用インタフェース

各刺激を何度も比較聴取しながら，それぞれの順位を変更しやすいように，**図 1** に示すようなインタフェースを用意した．図におけるスピーカのマークが各曲刺激を表し，マークにコメントを付けることもできる．順位付けにはインタフェースの横軸のみを用い，縦軸には特に意味を持たせていない．

被験者は，ランダムに配置された 10 個の曲刺激（図 1 左：A, B, . . . , J）を，右ほどうまく，左ほどへたであるようにマウス操作で並べ替える（図 1 右では，H が最もうまく，B が最もへたと評価されている）．

2.3 被 験 者

実験で使用する曲刺激を初めて聴取する 22 人の大学生の男女（19 歳〜29 歳）を被験者とし，11 人ずつの 2 つのグループ（A，B）に分けて実験を行った．

被験者 22 人中，16 人が楽器演奏経験者，2 人がボーカル・合唱経験者，4 人が絶対音感があると申告をした．

2.4 刺激および装置

曲刺激としての歌唱音声は，AIST ハミングデータベース（AIST-HDB）[18)] と RWC 研究用音楽データベース（RWC-MDB）[19)] から抜粋して使用した．本実験で AIST-HDB から用いたデータは，ポピュラー音楽データベース（RWC-MDB-P-2001）から 4 曲 8 カ所を抜粋して，その曲を初めて聴く被験者が 5 回聴いた後に，思い出しながら歌う音声を収録したものである．歌唱時には歌詞のみを提示し，楽譜は提示していない．したがって，本来は「うまい」歌唱者でも，思

表 1 聴取実験に用いた 40 人による計 80 個の曲刺激
Table 1 80 stimuli (by 40 singers) used in the subjective experiment.

グループ	曲番号	抜粋箇所	言語	性別	刺激数
A	No.27	出だし	日本語	男	10 人分
	No.28	出だし	日本語	女	10 人分
	No.90	出だし	英語	男	10 人分
	No.97	サビ	英語	女	10 人分
B	No.27	サビ	日本語	男	10 人分
	No.28	サビ	日本語	女	10 人分
	No.90	サビ	英語	男	10 人分
	No.97	出だし	英語	女	10 人分

※曲番号は RWC-MDB-P-2001

い出しながら歌っているという点で，収録されたデータは「うまくない」可能性もある．

被験者はそれぞれ，歌詞の言語と歌唱者の性別，曲の種別が異なる 4 曲分それぞれ 10 歌唱ずつの曲刺激（40 個）を聴取する（**表 1**）．各曲刺激は 10 人の歌唱者が同一曲を歌ったものであり，AIST-HDB から 9 個，RWC-MDB-P-2001 から 1 個の歌唱（ただし，伴奏のない歌声）を使用した．グループ A，B それぞれの刺激セット（4 曲分）は，同じ歌唱者群（40 人）が同じ曲の異なる部分（出だし/サビ）を歌ったものであり，被験者は A，B どちらか 1 つの刺激セットを評価する．曲刺激（80 個）の平均長は 12.5 sec であった．

呈示する曲刺激は 16 kHz/16 bit サンプリングのモノラル音声信号である．音量の違いによる聴取印象の変化を抑えるために振幅最大値を統一し，十分聴きやすい一定の音量でヘッドフォン聴取させた．

2.5 実験手順

被験者は実験に関する教示を受けた後，4 曲それぞれについて，10 人分の歌唱を比較聴取しながら並べ替える．曲の呈示順は，被験者ごとにランダムとした．ここで，被験者には，以下のような教示を行った．

- 同順位がないように並べ替えること
- 横軸に沿って歌唱力の順位付けを行うこと
- 歌唱力の差を間隔で表現すること
- 縦軸には意味がなく，自由に使ってよいこと
- 何度も聴取し，比較しながら並べ替えること

すべてを並べ替えた後，4 曲それぞれに対し図 1 右の ①② のような 2 本の線を引いてもらった．これは，① より右が「うまい」，② より左が「へた」となるように引かせたものである（図 1 右では，うまい = H, I, へた = B, F, D）．

最後に，歌唱力の評価基準に関する内省報告をとった．

2.6 結　　果

被験者間の順位付けに対して順位相関係数 ρ を計算すると，相関があった組の割合は**表 2** のようになった．これは，各グループの 4 曲それぞれに 11 人の評価があるため，55 組（$= 11 \times 10/2$）の相関を計算した結果である．一例として「グループ A/英語歌詞/女性歌唱者」を評価した場合の ρ を濃淡で示す（**図 2**）．ただし，最小の ρ が 0.42 であったため，0.4（白）$\leq \rho \leq 1$（黒）として表示している．**表 3** には，55 組の ρ の統計量（平均，標準偏差，最大，最小）をそれぞれ示す．

また，得られた結果を自動評価に利用するために，以下のような基準で，40 人の歌唱者による計 80 個の

表 2 有意に相関があった組の割合
Table 2 Percentage of significant pairs.

グループ	言語	性別	p < .01	p < .05
A	日本語	男	96.4% (53)	100.0% (55)
	日本語	女	74.6% (41)	90.9% (50)
	英語	男	61.8% (34)	89.1% (49)
	英語	女	41.8% (23)	80.0% (44)
	overall (220)		68.6% (151)	90.0% (198)
B	日本語	男	45.5% (25)	72.7% (40)
	日本語	女	72.7% (40)	98.2% (54)
	英語	男	52.7% (29)	89.1% (49)
	英語	女	74.6% (41)	90.9% (50)
	overall (220)		61.4% (135)	87.8% (193)
overall (440)			65.0% (260)	88.9% (391)

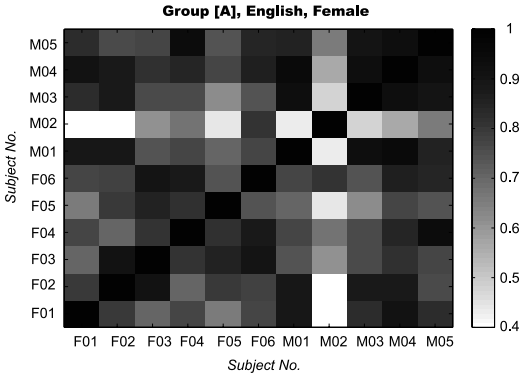

図 2 ρ 行列の濃淡表示
Fig. 2 Graphical display of the ρ-matrix.

表 3 順位相関係数 ρ の統計量
Table 3 Statistics values of Spearman's ρ.

グループ	言語	性別	平均	標準偏差	最大	最小
A	日本語	男	0.87	0.07	0.99	0.71
	日本語	女	0.77	0.14	0.95	0.38
	英語	男	0.75	0.14	0.96	0.28
	英語	女	0.69	0.14	0.98	0.42
B	日本語	男	0.64	0.22	0.98	0.03
	日本語	女	0.81	0.13	0.99	0.39
	英語	男	0.73	0.14	0.98	0.36
	英語	女	0.76	0.14	0.96	0.36

※それぞれ 55 組から計算

表 4 うまい/へたのラベル付け結果
Table 4 Results of labeling (good/poor).

グループ	言語	性別	うまい	へた	それ以外
A	日本語	男	3/10	2/10	5/10
	日本語	女	3/10	3/10	4/10
	英語	男	4/10	2/10	4/10
	英語	女	3/10	2/10	5/10
B	日本語	男	1/10	3/10	7/10
	日本語	女	3/10	3/10	4/10
	英語	男	2/10	2/10	6/10
	英語	女	3/10	4/10	3/10

曲刺激に「うまい」「へた」のラベル付けを行った．

うまい

被験者による最多得票が「うまい」であり，「へた」への得票がなかった曲刺激．

へた

被験者による最多得票が「へた」であり，「うまい」への得票がなかった曲刺激．

ラベル付けの結果を**表 4** に示す．

最後に，人間が着目する歌唱力の特徴を明らかにするために，内省調査によって得られたコメントを，その内容に応じて分類しながらまとめた結果を**表 5** に示す．ただし，音高（ピッチ）の正確さに関するコメントは，いずれの被験者も音程という用語で表現していた．これは，一般的にヴォーカリストが音をはずしたときに “音程が悪い” といういい方をするためと考えられる．

2.7 考　　察

歌詞の言語と歌唱者の性別，曲の種別一定の条件においては，表 2，表 3 に示すように被験者間の評価に高い相関があり，人間による歌唱力評価は安定していることが分かる．これに対し，評価基準の重要性は被験者で異なり，好みが評価に影響するというコメントがあった（表 5）．しかし，それにもかかわらず被験者間で評価は一定であったことから，歌唱力は個々人の評価基準の違いや好みに依存せずに，客観的に評価できるといえる．すなわち，歌唱力の自動評価は可能であり，「うまい」「へた」のラベル付けの結果は，自動評価手法を評価するための正解として利用できる．

また被験者は「声質」，「発音」，「音程」，「リズム」，「テクニック」に関係する特徴に着目していることが分かった．これらは，（非専門家である）被験者の内省報告であり，厳密な定義に基づいて得られた指標ではないが，歌唱力を評価するうえで重要な知見となる．興味深いコメントとして「うまい/へたは，聴いてすぐ分かる（3〜5 秒）」，「まず，うまい/へたの 2 グループに分割した」があった．すなわち，人間による評価では，まず（最初の数秒で）「うまい」か「へた」かを判断してからより詳細な評価を行っている可能性があり，自動評価において，「うまい」「へた」を識別する手法の開発は重要だと考えられる．

表 5 歌唱力評価に関する内省報告
Table 5 Introspective reports about singing skill evaluation.

分類	コメント例
評価基準の重要性	声質（声量，声の伸び・張り）＞ 音程 ＞ リズム 音程 ＞ リズム ＞ フレーズ感 ＞ 声量・声質 リズム ＞ 音程 ＞ 声質 （“＞” の左側ほど重視）
声質・音色	楽しそうな曲は楽しそうに歌ってほしい． 声が明るい方が良い．暗いと評価が下がる． 無理に高音を出そうとしていると評価が下がる． 苦しそうに歌うと評価が低い． 声量が足りないと評価が下がる． 声に張りや艶があると評価が高い．
発音	歌はメッセージを伝えるので，良い発音が必要．
音程	歌いたい音程を歌えているかどうか． 1 つの音符を同じ高さで歌えているかどうか． 強調して歌っている音がずれると減点．
リズム	一定のリズムを崩していないか．
キー	キーの違いは評価に反映しない． 声（キー）が高い人の方が評価が高くなる．
テクニックスキル	ビブラートがあると評価が高くなる． 音が急に変わるところで滑らかに歌えるか． 単語が変なところで伸びていると評価が低い． 歌らしくない歌（朗読，棒読み）は評価が低い． 感情移入していると評価が高い（抑揚・声質）． 声が伸びるときに音がフラットだと評価が下がる． 音の終わり方（伸ばし方）が良いと評価が高い． 節回しが不自然な人の評価を低くした．
評価方法	うまい/へたは，聴いてすぐ分かる（3〜5 秒）． まず，うまい/へたの 2 グループに分割した． 評価のために正解楽譜が欲しい．
好み	掠れながら歌う歌が好き． 好みが評価に影響する．

3. 歌唱力の自動評価手法

本章では，12 人の歌唱者による 100 種類のメロディの歌唱全 600 サンプルに対して，前章で得られた結果に基づいて「うまい」「へた」のラベル付けをしたデータを対象に，歌唱力評価手法の評価実験を行う．使用する歌唱サンプルは，16 kHz/16 bit サンプリングのモノラル音声信号である．

3.1 音響特徴量の抽出

表 5 からは歌唱力を評価するうえで様々な音響特徴量が考えられるが，その中でも，歌唱力評価において重要であり，歌唱者の個人性や曲の種別への依存が少ないと考えられる，相対音高とビブラートに着目する．

本章で提案する特徴量は，$\boxed{n}$のように，特徴量の番号を示す n を四角で囲んで示す．また，周波数は

すべて対数スケールで示し，cent 単位で表す．西洋平均律では，半音が 100 cent にあたる．中央ハ音の周波数 f_c（$= 440 \times 2^{\frac{3}{12}-1} = 261.62\ldots$ Hz）の cent 値を 4800 cent とすると，周波数 f_{Hz} の音の cent 値 f_{cent} は

$$f_{\mathrm{cent}} = 1200 \log_2 \left(\frac{f_{\mathrm{Hz}}}{f_c} \right) + 4800 \qquad (2)$$

で表される．

3.1.1 相対音高

音高の正確さを評価する場合，楽譜情報が既知であり，伴奏などの基準音高が明確なら，歌唱音声の基本周波数（F_0）を楽譜上の各音に対応する音高と照合すればよい．従来，カラオケにおける自動採点システムでは，このような方法がとられていた[20]．しかし，本論文の問題設定では，楽譜情報が未知であり，かつ，歌唱者がどういう絶対音高で歌っているかは固定されていないため，このような照合を行うことができない．そこで，相対音高に着目して評価を行う．西洋平均律を前提にすれば，相対音高が半音（100 cent）単位であるかを調べることで，歌唱対象の楽譜情報に依存せずに音高の正確さを評価できる．

歌唱音声の F_0 が半音単位で遷移できているかを評価するには，100 cent 間隔のグリッドを配置し，歌唱音声の F_0 がどれだけそのグリッド上に存在するかを見ればよい．すなわち，ある時刻の F_0（x cent）が，オフセット F（$0 \le F < 100$）からのグリッド周波数（$F, F+100, F+200, \ldots$）にどれだけ適合しているかを評価する．

そのために，コムフィルタの考え方に基づいたフィルタ $p(x;F)$ を用いる．$p(x;F)$ はオフセット F から 100 cent ごとに大きな重みを与える関数であり，次式のような混合ガウス分布で定義する．

$$p(x;F) = \sum_{i=0}^{\infty} \frac{\omega_i}{\sqrt{2\pi}\sigma_i} \exp\left\{ -\frac{(x-F-100i)^2}{2\sigma_i^2} \right\} \qquad (3)$$

現在の実装では，すべての i について，重み $\omega_i = 1$，標準偏差 $\sigma_i = 16$ cent としている．

このようなフィルタ $p(x;F)$ を用いて，時刻 t においてオフセット F からのグリッド上に F_0 が存在する可能性 $P_g(F,t)$ を式 (4) のように定義し，時刻 t を終端とする窓幅 T の矩形窓をシフトさせながら算出する．式 (4) で，$F_0(t)$ は時刻 t における F_0，$P_{F0}(t)$ は時刻 t において $F_0(t)$ が F_0 周波数である可能性（高調波構造が相対的にどれだけ優勢かを高調波構造上のパワーから算出した値）であり，後藤らの手法[21]を用いて 10 msec ごとに推定した．

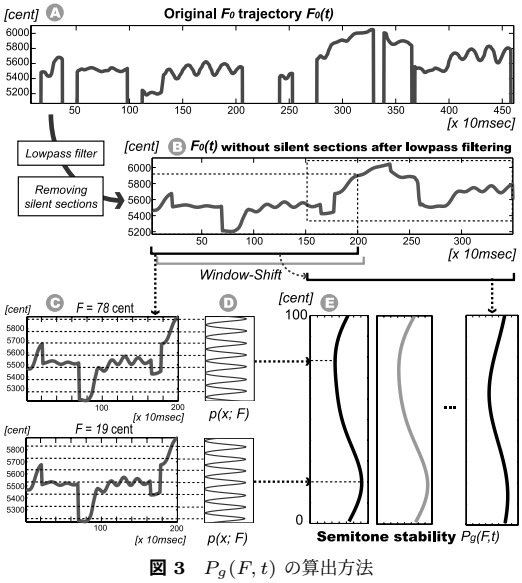

図 3 $P_g(F,t)$ の算出方法
Fig. 3 Overview of calculation method of $P_g(F,t)$.

$$P_g(F,t) = \int_{t-T}^{t} p(F_0(\tau);F) P_{F0}(\tau) d\tau \qquad (4)$$

$P_g(F,t)$ の計算手順を**図 3** に示す．図 3Ⓐ には $F_0(t)$ を，図 3Ⓑ には $F_0(t)$ をカットオフ周波数 5 Hz のローパスフィルタによって平滑化☆した後に無音区間を切り詰めたものを表示している．平滑化を行うのは，歌唱音声の F_0 特有の変動（ビブラートやオーバシュートなど）[7] による影響を除去するためである．$P_g(F,t)$ は，200 点（2 sec $= T$）の矩形窓を 5 点（50 msec）ずつシフトさせて算出した．図 3ⒸⒹⒺ に，$F = 19$ と $F = 78$ の場合のグリッド，$p(x;F)$，$P_g(F,t)$ の計算結果をそれぞれ示す．この例では，$F = 19$ が $F = 78$ よりもよく適合している．

仮に F_0 が 100 cent の整数倍で遷移していれば，$P_g(F,t)$ は，それに応じたオフセットに鋭いピークを 1 つ持つ．しかし，100 cent 以外の遷移が増えるにつれて，ピークが鋭くなくなったり 2 つ以上出現したりする．F_0 が半音単位で遷移できている場合には，その長時間平均 $g(F)$ も同様のピークを持つことが期待できる（**図 4**）．すなわち，$g(F)$ のピークの鋭さを評価することで，音高の正確さを評価できる．そこで，鋭いピークであれば値が小さくなるような，次式に示す 2 次モーメント M を特徴量 1 とする．

☆ FIR フィルタを使用し，F_0 推定のオクターブエラーや無音区間による不自然な平滑化を避けるために，無音や閾値（300 cent）以上の周波数変化がない区間のみを用いて平滑化を行う．

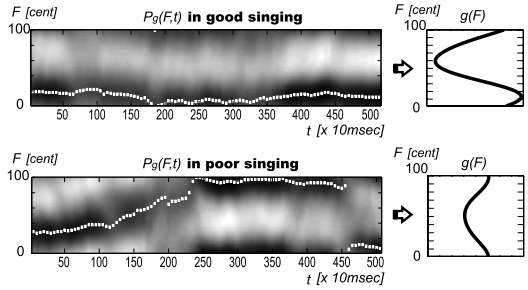

図 4 うまい歌唱とへたな歌唱の $P_g(F,t)$ とその長時間平均 $g(F)$
Fig. 4 Examples of $P_g(F,t)$ and its long-term average $g(F)$ in good/poor singing.

$$M = \int_{F_g-50}^{F_g+50} (F_g - F)^2 \, g(F) dF \tag{5}$$

ここで，F_g は $g(F)$ を最大にする周波数であり，$g(F)$ は $[F_g - 50, F_g + 50]$ の範囲の和が 1 となるように正規化する．ただし，$g(F) = g(F \pm 100)$ である．

また，$g(F)$ の傾斜を直線近似した傾き b_g も特徴量 2 として追加する．そこで，$g(F)$ を F_g で折り返した以下の関数 $G(F)$：

$$G(F) = \frac{g(F_g + F) + g(F_g - F)}{2}, \quad (0 \le F \le 50) \tag{6}$$

を最小自乗法で直線近似する．すなわち，a_g と b_g をパラメータとして次式を最小化することで b_g を得る．

$$err_g^2 = \int_0^{50} (G(F) - (a_g + b_g F))^2 \, dF \tag{7}$$

3.1.2 ビブラート

ビブラートとは，主に音を伸ばすときに周期的に音高を変化させる（揺らす）歌唱テクニックであり，熟達した歌唱者が頻繁に用いる重要なテクニックである．音高の周期的な変化の有無さえ適切に検出できれば，歌唱対象の楽譜情報に依存せずに用いることができるため，歌唱力評価に効果的な音響特徴量といえる．

再現率よりも精度を重視して検出を行うために，速さ（rate：毎秒に生じる揺らぎの回数）と，変調幅（extent：ビブラート区間の平均音高を中心とした変動の幅）☆に，それぞれ制限を加える．速さと変調幅は，**図 5** に示すパラメータ（R_n [sec]，E_n [cent]）を抽出し，

☆ extent は振幅と訳されることがあるが，amplitude との混同を避けるために，本論文では変調幅とした．

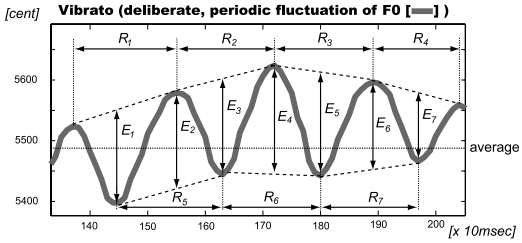

図 5 ビブラートパラメータの抽出
Fig. 5 Extraction parameters for vibrato detection.

$$\frac{1}{\text{rate}} = \frac{1}{N} \cdot \sum_{n=1}^{N} R_n \tag{8}$$

$$\text{extent} = \frac{1}{2N} \cdot \sum_{n=1}^{N} E_n \tag{9}$$

として求め，それぞれの制限範囲を 5〜8 Hz と 30〜150 cent とした．これらの値の範囲は，ビブラートパラメータに関する声楽での調査結果[22] と，ポピュラー音楽での調査結果[23] を参考にして決定した．

ビブラート区間は，3.1.1 項で求めた $F_0(t)$ の 1 次差分 $\Delta F_0(t)$（10 msec ごと）に短時間フーリエ変換（STFT）を行うことで検出する．32 点（320 msec）のハニング窓を用いた STFT で得られる振幅スペクトルを $X(f,t)$ とすると，ビブラートの速さに対応する周波数成分が他の周波数成分よりも支配的で，かつ鋭いピークとなるはずである．そこで，速さの下限を $\mathrm{F_L}$，上限を $\mathrm{F_H}$ としたとき，時刻 t におけるビブラート速さの周波数帯域のパワー $\Psi_v(t)$ とピークの鋭さ $S_v(t)$ を

$$\Psi_v(t) = \int_{\mathrm{F_L}}^{\mathrm{F_H}} \hat{X}(f,t) df \tag{10}$$

$$S_v(t) = \int_{\mathrm{F_L}}^{\mathrm{F_H}} \left| \frac{\partial \hat{X}(f,t)}{\partial f} \right| df \tag{11}$$

として定義する．ここで，$\hat{X}(f,t)$ は，次式に示すように，各時刻 t ごとに全周波数帯域のパワーで正規化したものである．

$$\hat{X}(f,t) = \frac{X(f,t)}{\int X(f,t) df} \tag{12}$$

これらを用いて，時刻 t におけるビブラートらしさ $P_v(t)$ を，

$$P_v(t) = S_v(t) \Psi_v(t) \tag{13}$$

のように定義する．

$P_v(t)$ が大きく，速さと変調幅が制限内で，さらに，$F_0(t)$ がその平均音高と 5 回以上交差する区間をビブラートとして判定した．ビブラートの検出例を**図 6** に

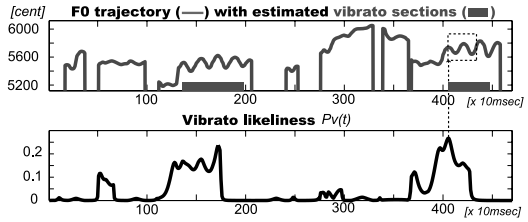

図 6 ビブラートの検出例
Fig. 6 An example of vibrato detection.

示す．特徴量には，検出された全ビブラートの区間長の総和（3），$P_v(t)$ の最大値（4）と平均値（5）を用いた．これらは，ビブラートが存在する歌唱において大きな値をとると考えられる．

3.2 識別実験

聴取実験より，聴取者間の歌唱力評価には高い相関があることが明らかになったが，その評価は聴取者間で完全には一致しなかった（表 3）．これは，高い（低い）歌唱力を持つ歌唱音声が逆の評価をされることはほとんどないが，同程度の歌唱力を持つ歌唱音声はその評価が細かく変動することを意味する．したがって，明らかに異なる歌唱力の差を識別できる特徴量が歌唱力評価には有効だといえる．このような識別手法の必要性は，2.7 節において内省報告の結果からも考察した．そこで本節では，提案した特徴量（1–5）の有効性を，「うまい」「へた」の 2 クラス識別実験によって評価する．

3.2.1 データセットの生成（ラベル付け）

識別実験のデータセットとして，2.4 節の表 1 のサンプルの中から「うまい」「へた」にラベル付けされたサンプルだけを抜き出すと，表 4 に示すように，各メロディに対して 5，6 サンプルしか使用できず，サンプル数が計 43 サンプルと少なくなってしまう．

そこで，AIST ハミングデータベース（AIST-HDB）[18] では，同一歌唱者が複数のメロディを歌っている特長を生かして，サンプル数を増やす．すなわち，聴取実験で「うまい（へた）」とラベル付けされた歌唱音声を歌った歌唱者をうまい（へたな）歌唱者として，彼らが歌った AIST-HDB 中の他のサンプルについても同様のラベル付けを行う．ただし，すべてのデータが聴取者によって評価されていないため，このようにして付けたラベルが正しいとは限らない．そこで，そのようなラベル誤りをできるだけ減らすために，評価の高かった（低かった）歌唱者 1，2 人にのみラベルを付与してデータセットとする．そのようにして生成した 12 人の歌唱者による歌唱全 600 サンプルを**表 6** に示す．これらは，RWC 研究用音楽データベース[19] のポピュラー音楽データベース（RWC-MDB-P-2001）と音楽ジャンルデータベース（RWC-MDB-G-2001）から 50 曲 100 カ所（日本語曲 25，英語曲 25）を抜粋して，その曲を初めて聴く被験者が 5 回聴いた後に，思い出しながら歌う音声を収録したものである．

表 6 識別実験に用いたデータセット
Table 6 Dataset for classification experiment.

歌唱者名	クラス	言語	性別	サンプル数
E004	うまい	英語	女	50
E008	うまい	英語	女	50
E017	うまい	英語	男	50
E021	うまい	英語	男	50
J002	うまい	日本語	女	50
J054	うまい	日本語	男	50
E001	へた	英語	女	50
E002	へた	英語	女	50
E013	へた	英語	男	50
E023	へた	英語	男	50
J014	へた	日本語	女	50
J052	へた	日本語	男	50

データセットの歌唱音声を実際に聴取したところ，「うまい」にラベル付けされた歌唱のいくつかは，思い出し歌唱であることが原因でピッチやリズムがやや不安定なものもあるが，「うまい」と感じられるものであった．逆に，「へた」にラベル付けされた歌唱は，へたといえる歌唱がほとんどであった．

3.2.2 実験条件

識別実験には，データセットのサンプルすべてを訓練用とする実験（Closed 実験）と，データセットを評価用と訓練用に分割し，評価用を順に変えながら識別率を評価する Cross-Validation 法（Open 実験）の 2 種類を行った．Open 実験では，評価用を 1 サンプルずつ変えながら，訓練用は残りの全サンプルから評価用と同一曲ないし同一歌唱者が含まれないように構成した．

これらの実験では，提案した特徴量が歌唱音声の「うまい」「へた」を，原曲のメロディや歌唱者に依存せずに識別できることを明らかにする．Closed 実験で得られた識別率は，提案した特徴量での識別性能の上限を表すので，これが低ければそもそも提案した特徴量は「うまい」「へた」を分離できていないことになる．そのうえで，Open 実験で得られた識別率が Closed 実験と同程度に高ければ，提案した特徴量により原曲のメロディや歌唱者に依存せずに歌唱力を評価できるといえる．

識別器には，ソフトマージン法による線形 SVM（Support Vector Machine）[24] を用い，LIBSVM [25] によって実装した．

3.2.3 実験結果

特徴量の有効性は，識別率（C）と「うまい」「へた」の精度（P_i）・再現率（R_i）によって評価した．ここで $i = \{good, poor\}$ はクラスを表す変数であり，class_{good} がうまい歌唱サンプルが属するクラス，class_{poor} がへたな歌唱サンプルの属するクラスである．P_i，R_i，C は，それぞれ以下のように定義する．

$$P_i = \frac{\mathrm{class}_i \text{ へ正しく識別したサンプル数}}{\mathrm{class}_i \text{ として識別したサンプル数}} \times 100 \quad (14)$$

$$R_i = \frac{\mathrm{class}_i \text{ へ正しく識別したサンプル数}}{\mathrm{class}_i \text{ の総サンプル数}} \times 100 \quad (15)$$

$$C = \frac{\text{正しく識別したサンプル数}}{\text{総サンプル数}} \times 100 = \frac{R_{good} + R_{poor}}{2} \quad (16)$$

データセットを「男性のみ（300）」「女性のみ（300）」「全サンプル（600）」と変えて実験を行った結果を，**表 7**（Closed 実験）と**表 8**（Open 実験）にそれぞれ示す．さらに，データセットを「全サンプル」とした，Open 実験における歌唱者ごとの再現率を**図 7** に示す．

3.2.4 考察

データセットを「全サンプル」とした識別実験では，Closed 実験で 86.2%（表 7），Open 実験では 83.5%（表 8）の識別率が得られた．どちらの実験でも高い識別性能が得られたことから，提案した特徴量は歌唱力の「うまい」「へた」を識別するうえで有効であり，原曲のメロディや歌唱者に依存せずに歌唱力の評価ができることが確認された．

表 7 Closed 実験の結果
Table 7 Results of the Closed experiment.

Dataset	C	P_{good}	R_{good}	P_{poor}	R_{poor}
男性のみ	91.3%	98.4%	84.0%	73.3%	98.7%
女性のみ	82.7%	90.2%	73.3%	61.8%	92.0%
全サンプル	86.2%	90.3%	81.0%	73.4%	91.3%

表 8 Open 実験の結果
Table 8 Results of the Open experiment.

Dataset	C	P_{good}	R_{good}	P_{poor}	R_{poor}
男性のみ	87.7%	93.8%	80.7%	70.8%	94.7%
女性のみ	71.7%	74.8%	65.3%	58.0%	78.0%
全サンプル	83.5%	87.6%	78.0%	70.3%	89.0%

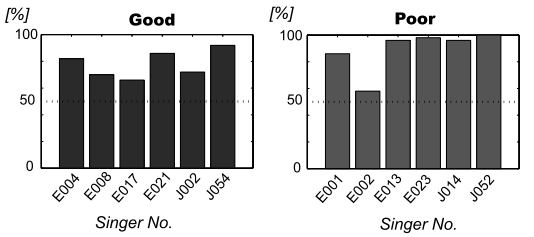

図 7 データセットを「全サンプル」とした場合の歌唱者ごと再現率（Open 実験）
Fig. 7 Recall rates (by singer) of “all samples” dataset in the Open experiment.

精度（P_i）・再現率（R_i）の関係において，全条件で P_{good} と R_{poor} が高い値となり，$P_{good} > P_{poor}$，$R_{good} < R_{poor}$ であった．これは，「うまい」歌唱を高い精度で識別できるが，本来「うまい」歌唱を「へた」として誤識別する場合も多いことを意味する．すなわち，提案した特徴量は「うまい」の判定に有効であるが，それのみでは歌唱力評価に必要な他の要因（たとえば，発声やリズムなど）を反映できない場合があるために，このような結果となったと考えられる．

誤識別の原因としては，3.2.1 項で述べたように歌唱音声の収録条件が思い出し歌唱であるために，「うまい」歌唱者でもピッチが不安定となる場合や，ビブラートがなかったり範囲の制限を満たさずに検出されない場合が多かった．「へた」な歌唱者が音を伸ばした際に意図せずに声が揺れて，ビブラートと検出されて誤識別することもあった．

データセットを「女性のみ」とした場合だけ，他の場合に比べて性能が大きく低下していたので，ここではさらに，その原因を詳しく考察する．6 人の女性歌唱者中，2 つの実験での識別率において，「うまい」歌唱者 E008 の性能の低下（68% → 54%）と，「へた」な歌唱者 E002 の性能の低下（84% → 56%）が大きかった．E008 は，「うまい」歌唱でビブラートは多く検出されたものの，F_0 が適切に半音間隔で遷移していないことが多く，E002 は，「へた」な歌唱でピッチは不安定だったものの，声が震えてビブラートが検出されることが多かった（きれいな揺れではないが，ビブラートとして聴こえるサンプルが多かった）．そのため，この 2 人の歌ったサンプルの多くが，特徴空間上で重なった分布となってしまったことが，Open 実験での性能低下の原因と考えられる．

データセットを「全サンプル」としたときに識別率の低下が抑えられたのは，サンプル数が増えたことで安定した識別面が学習できたためだと考えられ，今後このような性能の低下を回避するためには，訓練用サンプル数を増やすか，相対音高やビブラート以外の特徴量を用いる必要がある．

4. おわりに

本研究では，曲の種別（メロディ）・歌唱者の性別・

歌詞の言語が一定の条件では，聴取者の評価に高い相関があることを明らかにし，人間が着目する特徴について考察した．また，楽譜情報を用いずに抽出できる，曲や歌唱者に依存しない相対音高とビブラートに関する有効な特徴量を提案した．

なお，このほかにも，本研究の一環として長時間平均スペクトルの傾斜や，Singer's Formant の周波数帯域のパワー，低次元ケプストラムの変動，調波成分が含まれる割合，Spectral Centroid や Spectral Rolloff の平均，F_0 やパワーの変動などを特徴量として試みてきたが，いずれも本論文での識別率を上回っていない．

今後は，メロディ・性別・言語の組合せが異なる場合について聴取者の歌唱力評価を調査し，また，曲や歌唱者に依存しない発声やリズムに関する音響特徴量を検討していく予定である．

謝辞 本研究の聴取実験に関してご助言をしていただいた山本那美氏，自動評価手法に関して有益な議論をしていただいた亀岡弘和氏，聴取実験に参加していただいた被験者の方々に感謝いたします．本研究では，RWC 研究用音楽データベース（ポピュラー音楽 RWC-MDB-P-2001，音楽ジャンル RWC-MDB-G-2001），AIST ハミングデータベースを使用しました．

参 考 文 献

1) Sundberg, J.: *The Science of the Singing Voice*, p.226, Northern Illinois University Press (1987).
2) 中山一郎，小林範子：歌の声―音質の魅力と問題点，日本音響学会誌，Vol.52, No.5, pp.383–388 (1996).
3) 矢田部学，遠藤康男，粕谷英樹，神戸孝夫：歌声の基本周波数の動特性，日本音響学会平成 10 年度秋季講演論文集 3-8-6, pp.383–384 (1998).
4) 矢永龍一郎，河原英紀：会話音声と歌声音声の基本周波数制御の動特性について，情報処理学会研究報告音楽情報科学（SIGMUS），Vol.2003, No.082, pp.71–76 (2003).
5) 西内美登里，大串健吾：専門家と非専門家の歌声の評価，日本音響学会聴覚研究会資料，H-90-1, pp.1–6 (1990).
6) 辻 直也，赤木正人：歌声らしさの要因とそれに関連する音響特徴量の検討，日本音響学会聴覚研究会資料 H-2004-8, Vol.34, No.1, pp.41–46 (2004).
7) 齋藤 毅，鵜木祐史，赤木正人：歌声の F0 動的変動成分の抽出と F0 制御モデル，日本音響学会聴覚研究会資料，Vol.31, No.10, pp.683–690 (2001).
8) 大石康智，後藤真孝，伊藤克亘，武田一哉：局所的・大局的な特徴を利用した歌声と朗読音声の識別，情報処理学会研究報告音楽情報科学（SIGMUS），Vol.2005, No.82, pp.1–6 (2005).
9) 津田弘樹，森山 峻，福間 彰：3D 解析による歌声の評価に関する研究，電子情報通信学会情報・システムソサイエティ大会 D-458, p.461 (1996).
10) 池田 操：音響分析による歌曲「赤とんぼ」の歌唱評価，上越教育大学研究紀要，Vol.17, No.1, pp.395–407 (1997).
11) 片岡靖景，伊東一典，池田 操，中澤達夫，米沢義道，今関義弘，橋本昌己：歌唱支援システム構築のための歌声の分析と評価，情報処理学会研究報告音楽情報科学（SIGMUS），Vol.98, No.74, pp.23–30 (1998).
12) 池田 操，伊東一典：音楽科学生と一般学生の歌声の音響分析と評価―シンガーズ・フォルマントを指標として，上越教育大学研究紀要，Vol.19, No.2, pp.493–509 (2000).
13) 柏野邦夫，村瀬 洋：パート譜を用いたボーカル音分離システム，日本音響学会平成 10 年度春季講演論文集 2-9-1, pp.625–626 (1998).
14) 河原英紀，片寄晴弘：高品質音声分析変換合成システム STRAIGHT を用いたスキャット生成研究の提案，情報処理学会論文誌，Vol.43, No.2, pp.208–218 (2002).
15) Franco, H., Neumeyer, L., Digalakis, V. and Ronen, O.: Combination of Machine Scores for Automatic Grading of Pronunciation Quality, *Speech Communication*, Vol.30, pp.121–130 (2000).
16) 中村直生，中川聖一：日本人の英語発音の評価法，電子情報通信学会研究報告 SP2002-20, pp.51–58 (2002).
17) Kendall, M. and Gibbons, J.D.: *Rank Correlation Methods*, 5th edition, p.260, Oxford University Press (1990).
18) 後藤真孝，西村拓一：AIST ハミングデータベース：歌声研究用音楽データベース，情報処理学会研究報告音楽情報科学（SIGMUS），Vol.2005, No.82, pp.7–12 (2005).
19) 後藤真孝，橋口博樹，西村拓一，岡 隆一：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，情報処理学会論文誌，Vol.45, No.3, pp.728–738 (2004).
20) 株式会社ヤマハ，神谷伸悟，橘 聡：カラオケ装置，特開 2005-107337 (2005).
21) 後藤真孝，伊藤克亘，速水 悟：自然発話中の有声休止箇所のリアルタイム検出システム，電子情報通信学会論文誌 D-II, Vol.J83-D-II, No.11, pp.2330–2340 (2000).
22) Seashore, C.E.: A Musical Ornament, the Vibrato, *Psychology of Music*, pp.33–52, McGraw-Hill (1938).
23) 森勢将雅，平地由美，坂野秀樹，入野俊夫，河原英紀：STRAIGHT を用いたビブラート歌唱音声

の統計的性質，日本音響学会 2005 年春季講演論文集 3-P-15, pp.269–270 (2005).
24) Vapnik, V.N.: *Statistical Learning Theory*, p.736, John Wiley & Sons (1998).
25) Chang, C.-C. and Lin, C.-J.: LIBSVM: A Library for Support Vector Machines (2001). Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm

(平成 18 年 5 月 9 日受付)
(平成 18 年 10 月 3 日採録)

中野 倫靖（学生会員）

2003 年図書館情報大学卒業．2005 年筑波大学大学院図書館情報メディア研究科博士前期課程修了．現在，同大学院図書館情報メディア研究科博士後期課程．日本音響学会，日本音楽知覚認知学会各会員．

後藤 真孝（正会員）

1993 年早稲田大学理工学部電子通信学科卒業．1998 年同大学大学院理工学研究科博士後期課程修了．同年電子技術総合研究所（2001 年に独立行政法人産業技術総合研究所に改組）に入所し，現在に至る．2000 年から 2003 年まで科学技術振興事業団さきがけ研究 21「情報と知」領域研究員，2005 年から筑波大学大学院システム情報工学研究科助教授（連携大学院）を兼任．博士（工学）．音楽情報処理，音声言語情報処理等に興味を持つ．1997 年情報処理学会山下記念研究賞（音楽情報科学研究会），2000 年 WISS2000 論文賞・発表賞，2001 年日本音響学会粟屋潔学術奨励賞・ポスター賞，2002 年情報処理学会山下記念研究賞（音声言語情報処理研究会），2002 年日本音楽知覚認知学会研究選奨，2003 年インタラクション 2003 ベストペーパー賞，2005 年情報処理学会論文賞等 18 件受賞．電子情報通信学会，日本音響学会，日本音楽知覚認知学会各会員．

平賀 譲（正会員）

1983 年東京大学大学院理学系研究科（博士課程）中退，同年図書館情報大学助手．現在，筑波大学大学院図書館情報メディア研究科教授．日本認知科学会，日本音楽知覚認知学会，ACM 等各会員．