2006－MUS－64（12）
2006／2／23

歌唱力評価の聴取者実験と自動評価手法の検討

中野 倫靖[†]　　後藤 真孝[‡]　　平賀 譲[†]

[†] 筑波大学　　[‡] 産業技術総合研究所

[†]{nakano,hiraga}@slis.tsukuba.ac.jp　[‡] m.goto@aist.go.jp

あらまし　本研究では、ポピュラー音楽などの歌曲における歌唱力を、楽譜情報未知で自動的に評価する手法の開発を目指す。また、歌唱力を自動評価するために、聴取者による歌唱力評価が一定であるかどうかを調査する。従来、声楽やオペラ歌唱に関する歌唱力の考察は行われてきたが、それを自動的に評価する手法は研究されておらず、また聴取者の評価が一定であるのかを調査した例はなかった。歌唱力評価システムとしてはカラオケの採点機能が普及しているが、主に評価用の楽譜とのマッチングを行うため、楽譜情報が未知の状況では機能しない。そこで本稿では、歌唱力の評価が複数の聴取者間で一定であることを聴取者実験によって確認し、検討した特徴量を用いて「うまい」「へた」を自動的に識別する。歌唱力評価実験は、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、聴取者実験での結果をもとにラベル付与した 360 フレーズの歌唱音声に対して識別実験を行った結果、82.2%の識別率を得た。

A Preliminary Hearing Experiment towards the Automatic Evaluation of Singing Skills

Tomoyasu Nakano[†]　　Masataka Goto[‡]　　Yuzuru Hiraga[†]

[†]University of Tsukuba

[‡]National Institute of Advanced Industrial Science and Technology (AIST)

Abstract　The aim of this study is to explore a method for automatic evaluation of singing skills, which does not require score information. Our interest is directed towards ordinary, common person's singing, exploring the criteria that human subjects use in judging singing quality, and whether their judgments are stable among each individual. Although previous research on singing evaluation has focused on trained, professional singers (mostly in classic music), there has been no previous attempt that dealt with the problem of automatic evaluation of singing qualities and skills. A singing skills' rating system on Karaoke as an automatic evaluation scheme operates unsuccessfully in our case because it requires score information. In order to achieve our goal, two preliminary experiments, verifying whether the judgments of human subjects are stable, and automatic evaluation of performance by a 2-class classification (*good*/*poor*), were conducted. 22 subjects participated in the experiment. 88.9% of the correlation between the subjects' evaluations were significant at the 5% level. In a classification experiment with 360 singing voices, our method achieved a classification rate of 82.2%.

1　はじめに

本研究では、ポピュラー音楽などの歌曲における歌唱力の特徴を明らかにし、楽譜なしで自動的に評価する手法の実現を目指す。歌唱力を評価するための特徴が明らかになれば、様々なアプリケーションが構築可能となる。例えば、歌唱の何が原因で評価が低下しているかを提示できれば、歌唱指導支援として有用である。また、歌唱力の特徴を検索キーとした新たな音楽情報検索も考えられる。

これまで、歌唱を対象とした研究はあったが、歌唱力を自動的に評価しようとした研究例はなかった。歌唱の特性を明らかにする研究として、基本周波数 (以下、F0 と呼ぶ) の軌跡の特性を明らかにするもの [1,2]、歌唱の F0 制御モデルを提案するもの [3,4]、歌唱音声の特性を総合的に考察したもの [6]、歌唱と朗読音声を自動識別するもの [7] がある。歌唱力評価に関連する研究としては、音響信号からその特徴を見出そうとするものがある [8–13]。しかしこれらは、自動的に評価を行うものではなかった。

また、歌唱力を自動評価するシステムとしてはカラオケの採点機能が普及しており、主に評価用の楽譜情報とのマッチングを行っている。最近では、ビブラート検出 [15]、ポルタメント検出 [15]、声質評価 [14,15] や、メロディーアレンジ検出 [16] などの機能も備えたシステムもある。

楽譜が未知の状況での歌唱力評価が困難なのは、聴取者がどのような特徴をもとに歌唱力を評価しているのかが明らかになっていない点にある。ここで、聴取者によらず評価が一定であるのか、楽譜が未知の状況でもその評価は一定であるのかも分かっていない。そこで本研究では、ポピュラー音楽などの歌曲における歌唱力評価が、聴取者によらずほぼ一定であることを聴取者実験によって明らかにし、楽譜が未知であっても機能する自動評価法を検討する。

以下、**2** で人間による歌唱力評価実験について述べ、**3** で自動評価のための特徴について論じた後、識別実験を行う。最後に、**4** で今後の展開について述べる。

2　人間による歌唱力評価

自動評価手法を考えるために、まずは人間による歌唱力評価が聴取者によらずに一定であるかを調査する。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 (e.g. $n = 5, 7$) の評価値の相関を取る方法が取られていた [17,18]。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が聴取者によって異なる可能性がある。

そこで、それらの問題を解決するために、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、人によって距離尺度が異なっても、その順序関係さえ一定であれば対処できる。

本研究では、二つの順序の類似性を測るために Spearman の順位相関係数 ρ [19] を用いた。要素数が N である順序を$\boldsymbol{r}$=$(1, 2, ..., N)$ のようなベクトルとして表現すると、二つの順位ベクトル$\boldsymbol{a}$と$\boldsymbol{b}$の順位相関係数 ρ は、次式によって定義される。

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

2.2　聴取者実験

聴取者実験用の歌唱音声は、AIST ハミングデータベース (AIST-HDB) [20] と RWC 研究用音楽データベース (RWC-MDB) [21] から抜粋して使用した。今回 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 箇所を抜粋して、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。すなわち、本来は「うまい」歌唱者でも、思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。

表 1: グループ A,B の曲刺激 (40 人による計 80 個)

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

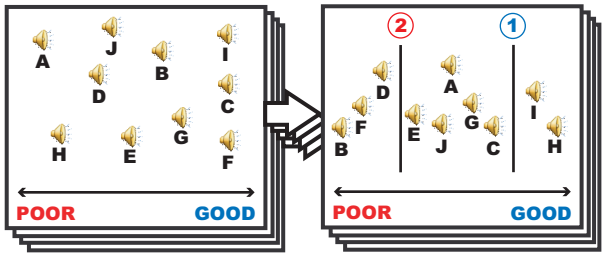

図 1: 実験画面と順序付けの例 (スピーカのマークは、ダブルクリックすると歌声が再生され、ドラッグすると移動する)

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 (20 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 40 個の曲刺激 (10×4) 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 個の曲刺激 (ただし、伴奏のない歌声) を使用した。

呈示する曲刺激は 16kHz, 16bit サンプリングのモノラル音声信号とした。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。

2.2.1　実験手順

図 1 に示すようなインタフェースを用意し、そこに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てもらった (図 1 右: H が最もうまい)。ここで「同一順位がないように並べ替えること」、「横軸には順序とともに歌唱力の差を表現すること」、「縦軸には意味がない」という教示を行った。

全てを並べ替えた後、4 曲それぞれに対し図 1:①②のような 2 本の線を引いてもらった。これは、①より

表 2: 有意に相関があった組の割合

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	38.2% (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)

表 3: うまい/へたのラベル付け結果

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

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り、これをもとにラベルを決定する (図 1 右: うまい=H,I へた=B,F,D)。最後に、歌唱力評価の基準に対して内省をとった。

2.2.2 結果

聴取者間の歌唱力評価が一定であるかを調査する。順位相関係数 ρ について、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 [19]。これらの水準で相関があった組の割合を表 2 に示す。4 曲それぞれに 11 人の評価があるため、それぞれ 55 組 ($= 11 \times 10/2$) の相関を計算した結果である。

また、以下のような基準で、40 人の歌唱者による計 80 個の曲刺激に「うまい」「へた」のラベル付けを行った。

うまい 図 1 右の①より右側に配置した被験者が最も多く、②より左側に配置した被験者が全くいなかった曲刺激。

へた 図 1 右の②より左側に配置した被験者が最も多く、①より右側に配置した被験者が全くいなかった曲刺激。

このような基準でラベル付けした結果を、表 3 に示す。

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 4 に示す。

表 4: 歌唱力評価に関する内省

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
声質	楽しそうな曲は楽しそうに歌って欲しい。 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。
音程	歌いたい音程を歌えているかどうか。 一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。
発音	歌はメッセージを伝えるので、良い発音が必要。
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）。 声が伸びる時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 (3〜5 秒) 分かる。 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。

2.2.3 考察

聴取者による歌唱力の評価実験により、言語と性別、曲の種別一定の条件においては、聴取者間の評価に高い相関があることが示された。すなわち、歌唱力の自動評価という問題設定に対する正当性が得られた。

内省では「声質（声量、声の伸び、声の出し方）」、「音程」、「リズム」、「ビブラート」などを評価の基準としていることが分かった。しかし、どの基準を最も重要視するかは聴取者によってばらつきがあった。また「好み」が評価に影響するというコメントがあった。それにもかかわらず、被験者間の評価には高い相関が見られたことから、歌唱力の評価は個々人の評価基準の違いや好みよりも、歌の「うまさ」が優先されていると考えられる。

今後は、曲や性別、言語が異なる場合についても引き続き調査を行う予定である。

3 歌唱力の自動評価法

聴取者実験により、歌唱力の自動評価を行う正当性と、内省より歌唱力評価の基準が得られた。そこでまずは、歌唱力を自動評価するための着目点について議論する。次いで、それらを良く反映するような特徴量について考察し、最後に、**2.2.2** で得られた結果からラベル付きデータセットを生成して「うまい」「へた」の 2 クラスの識別実験を行い、その結果を考察する。

表 5: 歌唱力評価の着目点

	教則本		学術書
	[22]	[23]	[24]
フォーム (姿勢)	2	2	6
ブレス	2	2	6
発声 (声量拡張、声区)	3,7	2	1,2
声質 (音色)	-	-	4
発音 (滑舌、母音明瞭度)	4	2	5
リズム	5	3	-
音程	6	4	-
テクニック (ビブラートなど)	8	5	3
ジャンルに応じたテクニック	8	5,7	-
声域拡張 (高音発声)	7	7	-
表現力 (曲のイメージ)	8	-	-

3.1 歌唱力評価のための着目点

ポピュラー音楽における歌唱力を自動評価するために、有効な性質は明らかになっていないため、まずは一般的なヴォーカル教則本を参考に、本節で議論する。二つのヴォーカル教則本 [22,23] と、歌唱力の評価という項目が記載された学術書 [24] を取り上げ、歌唱力に直接関係のあると考えられるポイントを抜粋し、それが記載されている章番号を表 5 に示す。教則本における章立ては、最初に記載されているほど重要な基礎である可能性が高く、表 5 から教則本での指導順序は似ていることが分かる。

表 5 及び、**2.2.3** での議論の結果より、以下の 6 種類の評価基準について音響特徴量を **3.2** で議論する。ここで、フォームやブレスは、発声の一部として捉えた。また、リズムも重要なポイントであると考えられるが、現時点では十分な考察が出来ていないため、リズムに関する特徴量の検討は今後の課題とする。

i. 発声（通る声，響く声）
ii. 発音（滑舌・母音明瞭度）
iii. 音程（音高差）
iv. ヴォーカルテクニック
v. 声質・音色
vi. 表現力（フレージング）

3.2 音響特徴量の議論

前節 **3.1** で議論した歌唱力評価基準に対し、それぞれを表す音響特徴量について議論する。本節での議論は全て、16kHz, 16bit サンプリングのモノラル音声信号に対して行う。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。

3.2.1 発声

発声における「響く声」「通る声」「張り・艶のある声」を特徴付ける音響的な性質としては、Singer's Formant が知られている [24–26]。Singer's Formant とは、男性歌手 (オペラ歌手、コンサート歌手) の母音 (有声音) において 2.5kHz から 3kHz の範囲に発生するフォルマントである [25]。物理的 (生理学的) には、喉頭を下げることで咽頭の下部を拡張され、この喉頭の共振周波数が同帯域となることに起因する [27]。Singer's Formant は、第 3〜第 5 フォルマントが互いに近づいてできたもので母音によらず一定である [25]。辰巳らは、その第 3〜第 5 フォルマントが近接しているほど、声に響きがあることを明らかにした [8]。女性歌手についても、中山 他によれば、男性と同様に喉頭を下げて歌唱していると推察されるソプラノやそれほど喉頭を下げない邦楽 (女性) の歌唱において、4kHz 付近に顕著なピークが観察されている [10]。さらに、歌唱以外の声として、男性アナウンサーの声にもこのようなフォルマントの存在が確認されている [28]。

これは必ずしもポピュラー音楽の歌唱に対する知見ではないが、本稿では発声の特徴量として Singer's Formant に関連する特徴量の有効性も検討する。ただし、歌詞の違いによる影響を抑えるために、これを長時間平均スペクトルから算出する。その前処理として、何も行わない場合と歌唱音声信号の 1 次差分によって高域を強調した場合の 2 種類を用いた。

- 長時間平均スペクトル
  - $\boxed{1}$ 最小二乗法に基づく振幅スペクトルの傾斜
  - $\boxed{2}$ 最小二乗法に基づく対数振幅スペクトルの傾斜
  - $\boxed{3}$ 2.5〜5kHz 帯域のパワーの全帯域に対する割合
  - $\boxed{4}$ 2.5〜5kHz 帯域の最大値の周波数 F_g からの 2 次モーメント (全帯域)
  - $\boxed{5}$ 2.5〜5kHz 帯域の最大値の周波数 F_g からの 2 次モーメント (F_g±1.5kHz の帯域)
- 長時間平均スペクトル（高域強調）
  - $\boxed{6}$ 最小二乗法に基づく振幅スペクトルの傾斜
  - $\boxed{7}$ 最小二乗法に基づく対数振幅スペクトルの傾斜
  - $\boxed{8}$ 2.5〜5kHz 帯域のパワーの全帯域に対する割合
  - $\boxed{9}$ 2.5〜5kHz 帯域の最大値の周波数 F_g からの 2 次モーメント (全帯域)
  - $\boxed{10}$ 2.5〜5kHz 帯域の最大値の周波数 F_g からの 2 次モーメント (F_g±1.5kHz の帯域)

$\boxed{4}$、$\boxed{5}$、$\boxed{9}$、$\boxed{10}$は、次の式で与えられる値である。

$$M = \int \left\{ 1 - \frac{S(f)}{S(F_g)} \right\}^2 df \tag{2}$$

ここで $S(f)$ は、周波数 f におけるスペクトル包絡 (ケプストラムから算出) の対数振幅を示す。$\boxed{4}$と$\boxed{9}$では全周波数帯域、$\boxed{5}$と$\boxed{10}$では F_g±1.5kHz の帯域を積分区間としている。

3.2.2 発音（滑舌・母音明瞭度）

歌唱の評価において滑舌の良さや母音明瞭度が挙げられることは多く [24]、**2.2.2** の表 4 でも挙げられていた。発音の明瞭性に関して、桑原 他は、発音が明瞭なアナウンサーは、音韻の変化に伴う第 1・第 2 フォルマント周波数の変動が大きいことを指摘している [28]。

そこで、フォルマント周波数の変動をケプストラムの低次項の分散で表す。ケプストラムを求めるために、歌唱音声信号に対する短時間フーリエ変換 (STFT) を、窓幅 1024 点 (64msec) のハニング窓を 160 点 (10msec) ずつシフトさせて計算した。

- ケプストラム（低次 16 項）

  11 全時刻 (1 フレーズ) における重み付き分散

ここで重みには、時刻 t における F0 の可能性 $P_{F0}(t)$ [29] (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) を用いた。

3.2.3 音程（音高差）

本稿では、楽譜がない状況で音程の良さを評価するために、西洋音楽を前提に、各音が半音（100cent）単位[1] で移動しているかどうかを評価する。すなわち、半音間隔のグリッドを考え、歌唱音声の F0 がどれだけそのグリッド上に存在するかを見る。これを評価するために、コムフィルタの考え方に基づいたフィルタ $p(x;F)$ を用いて、時刻 t において周波数 F がグリッド周波数となる可能性 $P_g(F,t)$ を評価する。ただし、グリッドは 100cent 毎に繰り返すので、$P_g(F,t)$ は $0 < F \leq 100$ についてのみ計算すればよい。仮に音高が 100cent の整数倍で遷移していれば、$P_g(F,t)$ は、それに応じたグリッド周波数 F_g に鋭いピークを 1 つ持つ。しかし、100cent 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。

そこで、次式のように定義する。

$$P_g(F,t) = \int_{-\infty}^{\infty} \int_{t-\tau}^{t} p(x;F) F_{F0}(t) dt dx \tag{3}$$

これは、時刻 t を終端とする窓幅 τ の矩形窓をシフトさせながら算出し、周波数を表す x と F の単位は cent とする。$F_{F0}(t)$ は F0 で、後藤 他の手法 [29] を用いて 10msec 毎に推定した。$p(x;F)$ は最も低いグリッド周波数が F の時に、そこから 100cent 毎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。

$$p(x;F) = \sum_{i=0}^{\infty} \frac{\omega_i}{\sqrt{2\pi}\sigma_i} \exp\left\{ -\frac{(x-F-100i)^2}{2\sigma_i^2} \right\} \tag{4}$$

現在の実装では、全ての i について、重み $\omega_i = 1$、標準偏差 $\sigma_i = 16$cent としている。

$P_g(F,t)$ を求めるために、$F_{F0}(t)$ はカットオフ周波数 5Hz のローパスフィルタによって平滑化[2] を行った後に無音区間を切り詰めたものを利用し、200 点 (2sec) の窓を 5 点 (50msec) ずつシフトさせて算出している。

このようにして計算された $P_g(F,t)$ について、その長時間平均 $g(F)$ を評価する。音程が良い場合には全区間を通して $P_g(F,t)$ が同じ位置に 1 つのピークを持つので、$g(F)$ も同様のピークを持つことが期待できる。

- $P_g(F,t)$ の長時間平均 $g(F)$

  12 $g(F)$ の最大値の周波数 F_g からの 2 次モーメント

  13 $g(F)$ の最大値の周波数 F_g からの傾斜

12 は次のような値である。

$$M = \int_{F_g-50}^{F_g+50} |(F_g - f)\, g(f)|^2 \, df \tag{5}$$

13 は、$0 \leq f \leq 50$ の範囲において、以下の関数を直線近似した傾きとし、それを最小二乗法によって求めた。

$$G(f) = \frac{g(F_g+f) + g(F_g-f)}{2} \tag{6}$$

3.2.4 ヴォーカルテクニック

様々なテクニック [22,23] の中で、特にビブラートは多くの歌曲に共通して重要なテクニックなので、特徴量として求める。再現率よりも精度を重視して検出を行うために、速さ（毎秒に生じる揺らぎの回数）と、振幅 (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 5〜8Hz と 30〜150cent の制限を加える。この値は、ビブラートパラメータに関する声楽とポピュラー音楽（RWC-MDB の歌唱）の調査結果 [26,30] を参考にして決定した。

ビブラート区間は、**3.2.3** で求めた $F_{F0}(t)$ の 1 次差分 $\Delta F_{F0}(t)$ (10msec 毎) に短時間フーリエ変換 (STFT) を行うことで検出する。32 点 (320msec) 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、ビブラートが存在する周波数成分は鋭いピークを持つはずである。そこで、対象とするビブラート周波数の下限を F_L、上限を F_H としたとき、時刻 t に

[1] 本稿では、対数スケールの周波数を cent の単位で表し、Hz で表された周波数 f_{Hz} を、cent で表された周波数 f_{cent} に $1200 \log_2 \frac{f_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}}$ に変換した値を用いる。

[2] FIR フィルタを使用し、ピッチ推定における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、閾値 (300cent) 以上の周波数変化があった箇所のフィルタ係数を 0 として計算した。

おけるピークの鋭さを、

$$S_v(t) = \int_{\mathrm{F_L}}^{\mathrm{F_H}} \left| \frac{\partial X(f,t)}{\partial f} \right| df \quad (7)$$

として定義し、その周波数帯域のパワーを

$$\Psi_v(t) = \int_{\mathrm{F_L}}^{\mathrm{F_H}} X(f,t) df \quad (8)$$

と定義する。そして、時刻 t におけるビブラートらしさ $P_v(t)$ を、

$$P_v(t) = S_v(t)\Psi_v(t) \quad (9)$$

のように定義する。

$P_v(t)$ が大きく、ビブラート振幅が制限内で、さらに、その区間内で $\Delta F_{F0}(t)$ が 5 回以上零交差する区間をビブラートとして判定し、全ビブラートの区間長の総和を特徴量とした。また、ビブラート区間を明示的に検出しなくてもビブラートが存在すれば大きくなるような二つの関数

$$P_{v1} = \max_{0 \le t \le T} (S_v(t)\Psi_v(t)) \quad (10)$$

$$P_{v2} = \frac{1}{T} \int_0^T S_v(t)\Psi_v(t) dt \quad (11)$$

もビブラートが含まれる可能性として利用する。ここで T は、解析したい歌唱音声信号の時間長である。

- ビブラート
  - **14** 推定したビブラートの区間長の総和
  - **15** ビブラートが含まれる可能性 P_{v1}
  - **16** ビブラートが含まれる可能性 P_{v2}

他にも、「声が高い方が評価が高い (**2.2.2**、表 4)」という聴取者実験の内省もあったことから、高いピッチで歌えることもテクニック (スキル) の一つであると考えられる。そこで、F0 の平均値も特徴に追加した。

- F0
  - **17** $F_{F0}(t)$ の時間方向の重み付き平均

ここで、重みは **3.2.2** で用いた $P_{F0}(t)$ とした。

3.2.5 声質（音色）

文献 [14,24] と聴取者実験による内省 (**2.2.2**) から得られた結果から、本稿では「しゃがれ声」と「明るい声/暗い声」の特徴量を設定する。

「しゃがれ声」は、非調波成分の割合が多い声なので [14]、調波成分のパワー (3.2.2 で用いた $P_{F0}(t)$ に相当) が全パワー中に占める割合を評価する。

- 調波・非調波成分比
  - **18** 有音区間中の平均

「明るい/暗い声」に関する特徴として、歌唱音声信号を窓幅 1024 点 (64msec) のハニング窓で STFT した振幅スペクトル $S(f,t)$ から、Spectral Centroid ($C(t)$) と Spectral Rolloff ($R(t)$) を求める [31]。$C(t)$ と $R(t)$ が大きいほど明るい声である。

$$C(t) = \frac{1}{P(t)} \int_0^N (S(f,t) \cdot f)\, df \quad (12)$$

$$R(t) = \underset{r}{\mathrm{argmin}} \left| 0.85P(t) - \int_0^r S(f,t) df \right| \quad (13)$$

$$P(t) = \int_0^N S(f,t) df \quad (14)$$

これらは以下の 2 通りの N で評価する。

- Spectral Centroid
  - **19** 有音区間中の平均（$N = 8$kHz）
  - **20** 有音区間中の平均（$N = 5$kHz）
- Spectral Rolloff
  - **21** 有音区間中の平均（$N = 8$kHz）
  - **22** 有音区間中の平均（$N = 5$kHz）

3.2.6 表現力

本稿では、表現力の一つとして抑揚を考え、パワーの変動と F0 の変動を特徴量とする。ここで、パワーは **3.2.2** で用いた $P_{F0}(t)$ とした。

- 抑揚
  - **23** $P_{F0}(t)$ の有音区間における標準偏差
  - **24** $F_{F0}(t)$ の有音区間における標準偏差

3.3 「うまい」「へた」の 2 クラス識別実験

ここでは、聴取者実験の結果からデータに「うまい」「へた」のラベル付けを行い、提案した特徴量を用いて識別実験を行う。

3.3.1 データセットの生成（ラベル付け）

聴取者実験の結果 (**2.2.2**) から「うまい」歌唱者と「へた」な歌唱者を決定し、彼らが歌った AIST-HDB 中の他のサンプル (曲刺激) についてもラベル付けを行ってデータセットとする。ただし、実際に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 名にのみラベルを付与してデータセットとする。具体的には、平均順位が高く、表 1 右①よりも右側に配置された割合が多いサンプルの歌唱者を「うまい」歌唱者、逆に平均順位が低く、表 1 右②よりも左側に配置された割合が多いサンプルの歌唱者を「へた」な歌唱者とした。そのようにして生成したデータセット 360 サンプルを表 6 に示す。

表 6: 識別実験に用いたデータセット

歌唱者名	ラベル	言語	性別	サンプル数
J054	うまい	日本語	男	50
J002	うまい	日本語	女	50
E017	うまい	英語	男	20
E021	うまい	英語	男	20
E004	うまい	英語	女	20
E008	うまい	英語	女	20
J052	へタ	日本語	男	50
J014	へタ	日本語	女	50
E013	へタ	英語	男	20
E023	へタ	英語	男	20
E001	へタ	英語	女	20
E002	へタ	英語	女	20

3.3.2 実験条件

3.2 で述べた 24 種類の特徴量を全て使用して「うまい」「へた」の識別実験を行う。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 法を用いた。ただし、データセット内には同一曲・同一歌唱者のサンプルが複数含まれているため、評価用を 1/10、訓練用を 9/10 として行う 10-fold Cross-Validation 法と、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・同一歌唱者が含まれないように構成する特殊な Leave-One-Out 法[3] の 2 種類を行った。

識別器には、SVM (Support Vector Machine) と決定木 (C4.5) を用いた。ここで、提案した特徴量が、歌唱者や曲に大きく依存しているような場合、適切な識別が出来ない可能性がある。そこで、全特徴量を使った実験と、歌唱者や曲に依存しないような特徴のみを選別した実験 (SVM による識別のみ) を行った。非依存と考えられる特徴は 12 | 13 | 14 | 15 | 16 である。また、特徴量が男女で異なる振る舞いをする可能性があるため、データセットを男女に分けた場合とまとめた場合についてそれぞれ実験を行う。ただし、男女をまとめた場合は、性別の違いによるものが特に大きいと考えられる 17 (F0 の平均) を除いて実験した。

3.3.3 実験結果

識別結果は、識別率と「うまい」「へた」の精度・再現率によって評価した。

特徴量を全て使った実験結果

10-fold Cross-Validation 法による識別結果を表 7、Leave-One-Out 法による識別結果を表 8 に示す。また、特徴量間の階層関係を考察するために、全データを訓練用として生成した決定木について、根に近い上位のノードに利用されていた特徴量を表 9 に示す。

[3] 実際の Leave-One-Out 法は、訓練用を評価用以外全てで構成するため、ここでは便宜上こう呼ぶことにする。

表 7: 識別率（10-fold Cross-Validation, 全特徴量使用）

		うまい		へタ	
データセット	識別率	精度	再現率	精度	再現率
SVM					
男性のみ	93.3%	94.3%	92.2%	92.4%	94.4%
女性のみ	87.2%	89.4%	84.4%	85.3%	90.0%
男女	85.0%	85.8%	83.9%	84.2%	86.1%
決定木					
男性のみ	90.6%	92.0%	88.9%	89.2%	92.2%
女性のみ	80.6%	80.2%	81.1%	80.9%	80.0%
男女	81.7%	82.4%	80.6%	81.0%	82.8%

表 8: 識別率（Leave-One-Out, 全特徴量使用）

		うまい		へタ	
データセット	識別率	精度	再現率	精度	再現率
SVM					
男性のみ	67.2%	70.1%	60.0%	52.4%	74.4%
女性のみ	41.1%	42.6%	51.1%	63.9%	31.1%
男女	69.4%	69.9%	68.3%	66.8%	70.6%
決定木					
男性のみ	61.7%	67.2%	45.6%	58.8%	77.8%
女性のみ	52.2%	52.3%	51.1%	52.2%	53.3%
男女	71.1%	71.8%	69.4%	70.4%	72.8%

特徴量を選別した実験結果

10-fold Cross-Validation 法による識別結果を表 10、Leave-One-Out 法による識別結果を表 11 に示す。さらに表 11 において、データセットを「男女」とした場合の歌唱者毎の識別率を図 2 に示す。

3.4 考察

識別実験において、特徴量を全て使用した場合、10-fold Cross-Validation 法と Leave-One-Out 法では、明らかに識別率に差があった (表 7, 8)。これは今回提案した特徴量の一部が曲や歌唱者に依存し、異なる曲や歌唱者に有効でなかったためと考えられる。しかし、曲や歌唱者に依存しないと思われる特徴を利用した結果は、10-fold Cross-Validation 法と Leave-One-Out 法において、それほど大きな差はなく (表 10, 11)、Leave-One-Out 法による評価で 82.2%の識別率を得ている。ここで、表 11 の再現率と精度から、誤識別をした割合は「うまい」とラベル付けされた歌唱者が「へタ」にラベル付けされることが多かったことが分かる。これは、利用した AIST-HDB の歌手が思い出しながら歌っていることが原因であると考える。実際に図 2 で誤識別が多かった E008 の歌唱を聴いてみると、思い出しながら歌っていると推察され、特に、音程がうまく取れていなかった。

表 9: 決定木による上位ノードの特徴量

データセット	根	上位の特徴量	
男性のみ	16	17	15
女性のみ	13	14	
男女	14	13	21

表 10: 識別率（SVM, 10-fold Cross-Validation, 特徴量選別）

データセット	識別率	うまい		ヘタ	
		精度	再現率	精度	再現率
男性のみ	85.0%	92.0%	76.7%	80.0%	93.3%
女性のみ	82.2%	89.2%	73.3%	77.4%	91.1%
男女	82.8%	88.8%	75.0%	78.4%	90.6%

表 11: 識別率（SVM, Leave-One-Out, 特徴量選別）

データセット	識別率	うまい		ヘタ	
		精度	再現率	精度	再現率
男性のみ	80.6%	83.1%	76.7%	71.1%	84.4%
女性のみ	67.8%	67.8%	67.8%	67.8%	67.8%
男女	82.2%	89.2%	73.3%	62.3%	91.1%

表 9 の結果を見ると、曲に依存が少ない特徴が上位階層に来ていることが分かる。男女混合のデータセットにおいて 21 が利用されているが、これは、根の 14 で、全 360 サンプルから分岐された残りの 80 サンプル中、4 サンプルを分離させた効果しかなく、有効な特徴量とは言えなかった。

これらより、音程に提案した特徴量（12 13）とビブラートに関する特徴量（14 15 16）は歌唱力評価に有効であり、さらに、曲や歌唱者、性別に依存せずに利用できることが示された。

4 おわりに

本研究では、楽譜情報を用いない歌唱力評価について、言語と曲の種別、歌唱者の性別が一定の条件では、聴取者の評価に高い相関があることを明らかにした。また、言語・性別・曲に依存しない音程とビブラートに関する有効な特徴量を提案した。

今後は、聴取者が歌唱力評価を行う際に「好み」がどれだけ影響するのかを、曲や性別、言語が異なる場合について調査を行い、さらに識別能力を上げるための特徴量を検討していく予定である。

謝辞

本研究に対し有益な議論をして頂いた、亀岡 弘和 氏 (東京大学大学院 情報理工学系研究科) に感謝する。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)、AIST ハミングデータベースを使用した。

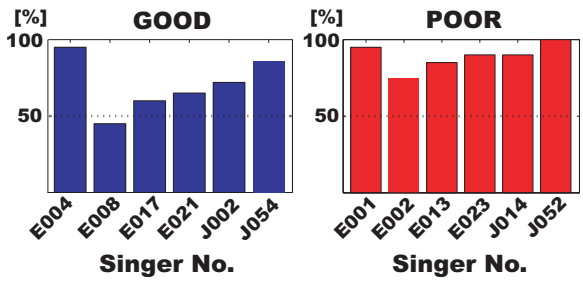

図 2: 表 11 における男女混合データセットの識別率 (歌唱者ごと)

参 考 文 献

[1] 矢田部, 遠藤, 粕谷, 神戸, “歌声の基本周波数の動特性,” 音講論集 3-8-6, pp.383-384, 1998.
[2] 矢永, 河原, “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,” 情処研報 (MUS), Vol.2003 No.082, pp.71-76, 2003.
[3] 柏野, 村瀬, “パート譜を用いたボーカル音分離システム,” 音講論集 2-9-1, pp.625-626, 1998.
[4] 齋藤, 鵜木, 赤木, “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.
[5] 田原, 森勢, 坂野, 入野, 河原, “歌唱音声の音量変化に伴うスペクトル変形の分析について,” 音講論集 3-P-16, pp.271-272, 2005.
[6] 河原, 片寄, “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,” 情処論, Vol. 43, No.2, pp.208-218, 2002.
[7] 大石, 後藤, 伊藤, 武田, 局所的・大局的な特徴を利用した歌声と朗読音声の識別, 情処研報 (MUS), Vol.2005, No.82, pp.1-6, 2005.
[8] 辰巳, 国崎, 樋口, 藤崎, “歌声の響きに関する音響的特徴,” 音講論集 1-2-5, pp.29-30, 1978.
[9] 津田, 森山, 福間, “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] 中山, 小林, “歌の声—音質の魅力と問題点—,” 音響誌, Vol.52, No.5, pp.383-388, 1996.
[11] 池田, “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.
[12] 片岡, 伊東, 池田, 中澤, 米沢, 今関, 橋本, “歌唱支援システム構築のための歌声の分析と評価,” 情処研報 (MUS), Vol.98, No.74, pp.23-30, 1998.
[13] 池田, 伊東, “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.
[14] 株式会社ヤマハ, 安間, 橘, “歌唱音声評価装置、カラオケ採点装置及びこれらのプログラム,” 特開 2005-107088, 2005.
[15] 株式会社ヤマハ, 神谷, 橘, “カラオケ装置,” 特開 2005-107337, 2005.
[16] 株式会社タイトー, 北村, “メロディーアレンジ採点機能を有するカラオケ採点装置,” 特開 2004-102147, 2004.
[17] H. Franco, L. Neumeyer, V. Digalakis and O. Ronen, “Combination of Machine Scores for Automatic Grading of Pronunciation Quality,” *Speech Communication*, Vol.30, pp.121-130, 2000.
[18] 中村, 中川, “日本人の英語発音の評価法,” 信学技報 SP2002-20, pp.51-58, 2002.
[19] M. Kendall and J. D. Gibbons, “Rank Correlation Methods,” *Oxford University Press fifth edition*, 260p., 1990.
[20] 後藤, 西村, AIST ハミングデータベース: 歌声研究用音楽データベース, 情処研報 (MUS), Vol.2005, No.82, pp.7-12, 2005.
[21] 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,” 情処論, Vol.45, No.3, pp.728-738, 2004.
[22] 福島, “これで完璧! ヴォーカルの基礎,” 第 3 版, 123p., 2001.
[23] 高田, “もっとうまく歌える本,” 125p., 2003.
[24] 日本音声言語医学会, 声の検査法: 臨床編, 296p., 1979.
[25] J. Sundberg, “The Science of the Singing Voice,” *Northern Illinois Univ Pr*, 226p., 1987.
[26] D. Deutsch 編（寺西, 大串, 宮崎 監訳）, 音楽の心理学（上）, 334p., 1987.
[27] W. Richards,（石川, 平原 訳）, “ナチュラルコンピュテーション: 聴覚と触覚・力センシング・運動の計算理論,” パーソナルメディア株式会社, 307p., 1994.
[28] 桑原, 大串, “アナウンサー音声の音響的特徴,” 信学論, Vol.J66-A, No.6, pp.545-552, 1983.
[29] 後藤, 伊藤, 速水, “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,” 信学論, Vol. J83-D-II, No. 11, pp.2330-2340, 2000.
[30] 森勢, 平地, 坂野, 入野, 河原, STRAIGHT を用いたビブラート歌唱音声の統計的性質, 音講論集 3-P-15, pp.269-270, 2005.
[31] G.Tzanetakis, G. Essl, and P. Cook, Musical Genre Classification of Audio Signals. *IEEE Trans. on Speech and Audio Processing*. vol.10, no.5, 2002, p.293-302.