情報処理学会論文誌 Vol.57 No.5 1375–1388 (May 2016)

歌声の印象評価尺度の構築に基づく多様な印象の自動推定手法

金礪 愛[1,a)] 中野 倫靖[2,b)] 後藤 真孝[2,c)] 菊池 英明[1,d)]

受付日 2015年7月31日, 採録日 2016年2月8日

概要：本論文では，ポピュラー音楽におけるアマチュア女性歌唱者の歌声を対象として，音響信号から歌声の印象を自動推定する手法を提案する．従来，歌声を音響信号から自動的に評価する際，特定の印象（歌唱力や熱唱度など）を対象とした研究は多く行われてきた．しかし，歌声の印象評価尺度を一から構築した研究はなく，歌声が与えうる多様な印象についての包括的な調査は行われていない．本論文では，主観評価実験や因子分析を経たうえで，歌声の多様な印象を適切に評価可能な評価尺度を構築した．そして，47 語の印象評価語と歌声の印象評価に関わる 3 因子（迫力性，丁寧さ，明るさ）の強度を音響特徴量から推定するため，印象得点と音響特徴量を用いた重回帰分析を行う．60 個の歌声データを用い，各印象を推定する重回帰モデルを構築したところ，3 因子のモデルの決定係数について 0.958（迫力性），0.551（丁寧さ），0.643（明るさ）という結果を得た．また本手法によって，60 歌唱それぞれにおける，50（= 47 + 3）種の印象得点の実測値と推定値の重相関係数を求めた結果，それらの平均は 0.720 であった．

キーワード：歌声情報処理，歌声印象推定，印象評価尺度，因子分析，重回帰分析

An Automatic Estimation Method of Various Impressions Based on Scale Construction for Singing Impressions

Ai Kanato[1,a)] Tomoyasu Nakano[2,b)] Masataka Goto[2,c)] Hideaki Kikuchi[1,d)]

Received: July 31, 2015, Accepted: February 8, 2016

Abstract: This paper presents a method for estimating the impression of a singing voice using acoustic features in popular Japanese songs sung by amateur female singers. Many previous researches on automatic singing voice evaluation using acoustic features dealt with specific impressions such as “singing skill” and “singing enthusiasm”. However, none of them constructed impression scales in a bottom-up manner, and comprehensively investigated various impressions of singing voices. An impression scale, which can be used to evaluate various singing impressions properly, was consequently constructed via factor analysis using the results of a subjective evaluation. Multiple regression analysis using acoustic features and impression scores was conducted for estimating the impression score of 47 impression words and 3 factors (“powerful”, “cautious”, “cheerful”) that were extracted by factor analysis. Using the multiple regression model to estimate the impression score for 60 recordings, the coefficients of determination for the 3 factors were found to be 0.958 (powerful), 0.551 (cautious), and 0.643 (cheerful). Using our method, the average of the multiple correlation coefficients was calculated as 0.720 for the observed values for 50 (= 47 + 3) impression scores and the estimated values for each recording.

Keywords: singing information processing, singing impression estimation, impression scale, factor analysis, multiple regression analysis

[1] 早稲田大学大学院人間科学研究科
Graduate School of Human Science, Waseda University, Tokorozawa, Saitama 359–1192, Japan
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305–8568, Japan

[a)] kanato.w@gmail.com
[b)] t.nakano@aist.go.jp
[c)] m.goto@aist.go.jp
[d)] kikuchi@waseda.jp

© 2016 Information Processing Society of Japan

1. はじめに

本研究では，ポピュラー音楽における歌声を対象として，人がその歌声を聴いた際に感じる多様な印象を，音響信号から自動推定することを目的とする．印象を歌声の音響特徴量から推定できれば，実際に歌声を聴くことなく歌声に対する印象の自動付与が可能となり，歌声の印象に基づいた楽曲検索の実現につながる．また，自身の歌声の印象を自動推定することにより，歌声の表現力向上のための客観的な指標として利用できる．さらに，歌声に対する主観評価と音響的な特徴との関連性を明らかにすることで，人間の歌声認知のメカニズム解明にも貢献できると考えられる．

歌声が聞き手に与える印象は，「歌唱者の外面的な特性（性別，年齢，体格など）」「歌唱者の内面的な特性（性格など）」「歌唱者の感情（嬉しそうな，一生懸命ななど）」「歌声そのものを形容する評価語（明るい，透き通ったなど）」「歌唱力（うまいなど）」などの多様な要因に基づいており，それぞれに対応する評価語が存在すると考えられる．以降，本論文では歌声の印象を形容する語を「印象評価語」と呼ぶ．

従来，歌声の印象の自動推定においては，特定の印象の強度を推定する研究が多く行われている．たとえば，Nakano らは，歌唱された楽曲の楽譜情報を用いずに歌声の歌唱力を自動推定する手法を提案している [1]．また，Tsai らはオリジナル楽曲と歌唱者の歌声の類似性に基づいた歌唱力評価を行っている [2]．歌唱力以外の自動評価では，Daido らが歌声の熱唱度の自動推定手法を提案している [3]．歌声ではなく，話声においては，感情の自動推定に関する研究が多く行われている．たとえば，Luengo らの調査では，感情の自動推定には韻律情報よりもスペクトル包絡が重要であることが明らかにされている [4]．また，Vlasenko らはスペクトルのフォルマントに関する特徴が感情推定に有効であると示している [5]．そして，Scherer は話声に対する感情推定手法が歌声にも適用可能だと述べている [6]．

一方，自動推定を目的としていない，歌声の印象と音響特徴量の関係性を考察する研究も行われている．たとえば，Kotlyar らは歌声における感情表現について，11 人のプロの歌唱者が歌唱した歌声を用い，音響特徴量との関係を調査している [7]．このように，従来の研究は「感情」「歌唱力」といった特定の印象を対象としており，包括的な歌声の印象について調査したり，歌声の印象評価尺度を一から構築したりした研究はなかった．歌声の印象を包括的にとらえられる尺度が存在していなかったことにより，特定の印象のみを扱う傾向があったと考えられる．

そこで，本論文ではまず歌声の印象を適切に評価可能な多様な印象評価語の収集を行った．それらの語を用いて主観評価実験を行い，その結果を因子分析することで，印象評価尺度を構築した（2 章）．そして，歌声の印象を自動推定するため，歌声の音響特徴量に基づき，多様な印象評価語と印象評価に関わる因子の強度を自動推定する重回帰モデルの構築を行った（3 章）．**図 1** は，本研究の印象推定手法の概略を示している．

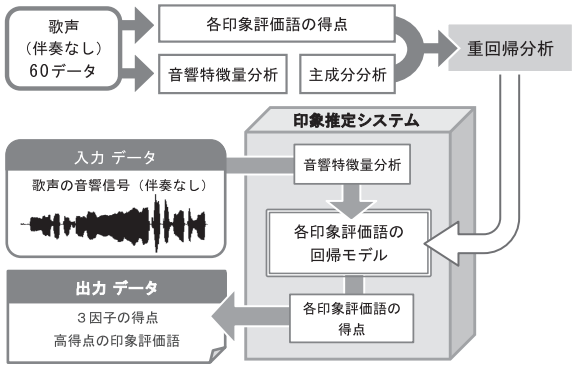


図 1 システムの概要
Fig. 1 Overview of estimation system.

1.1 本研究が対象とする歌声

本論文では，「歌声の印象」を「歌声を聴いた際にその歌声に対して生じる主観的な感覚」と定義する．歌声の印象は，歌唱する楽曲の歌詞，メロディ，テンポなど様々な要素に影響されるが，本論文では声質や抑揚の表現など，発声表現の違いから生じる歌声の印象を対象とする．

聴取者が，ある歌声からその印象を認知する際には，歌声そのものの情報のほかに，楽曲の属性（邦楽曲，合唱曲など）や歌唱者の属性（「女性の歌声である」「プロの歌手の歌声である」など）による印象の影響を受ける可能性がある．また，評価者の属性によっても，認知される印象は大きく異なると考えられる．

本論文では，一般人が聴いたり，一般人が歌ったりすることを考慮し，各要素の属性について，以下のように定めた．まず，楽曲は，日本人の多くが聞き慣れているであろう「日本のポピュラー音楽」を想定する．次に，歌唱者は，多くの一般歌唱者に適用できるよう，プロではなくアマチュア歌唱者を対象とする．より詳細な印象の差異をとらえるため，性別を限定し，「アマチュア女性歌唱者」のみを対象とした．今後は男性歌唱者を対象とし，検討する必要がある．最後に，専門知識を持たない一般人が認知する印象の調査を目的とし，歌声の評価者は「音楽的な専門知識を持たない一般人」とした．なお，本論文で述べる実験の被験者（歌唱者，評価者）は，すべて大学生である．

これらの条件により，本論文で提案する印象推定手法では，日本のポピュラー音楽におけるアマチュア女性歌唱者の歌声に対して，音楽的な専門知識を持たない一般人が認知する印象を推定することが可能である．

© 2016 Information Processing Society of Japan

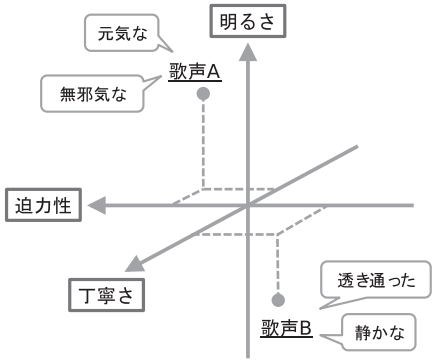


図 2 印象空間の例
Fig. 2 An example of singing impression space.

1.2 印象推定手法

本研究では，ある歌声の印象を推定する際に，(1) 印象空間内における位置を推定する，(2) 印象を示す印象評価語を選定する，という 2 種類の推定を行う．まず，推定対象の歌声から音響特徴量を算出する．次に，因子分析によって得られた歌声の 3 因子（迫力性，丁寧さ，明るさ），および印象評価語 44 語それぞれに対応する重回帰モデルを用い，音響特徴量を独立変数とし，従属変数として得点を算出する．44 語の印象評価語に対応する重回帰モデルでは，得点が大きくなるほどその印象評価語の強度が高いことを示している．そして，3 因子の得点を用いて印象空間内における位置を示し，44 語のうち得点の高かった印象評価語を出力することで，印象の推定を行う．

図 2 は 3 因子によって表現される，3 次元の歌声の印象空間を表している．歌声 A の 3 因子の得点が，迫力性：0.5，丁寧さ：−0.5，明るさ：0.8 だとすると，これらの得点を用いることで，印象空間内での位置が特定される．その結果，複数の歌声の印象の違いを，3 軸の因子得点をもとに把握することが可能となる．ただし，3 因子の得点を用いるだけでは歌声の印象を直感的に理解することが難しい．そのため，得点の高かった印象評価語を出力することで，歌声の印象を直感的に理解しやすく明示する．

2. 印象評価尺度の構築

歌声の印象評価に関わる因子，また，それらの印象を表現する言葉を明らかにするため，主観評価実験と因子分析により歌声の印象評価尺度を構築した．まず，多様な印象評価語を収集し，歌声を用いない主観評価実験を行ったうえで仮尺度（2.1 節）を構築する．その後，実際の歌声を用い，歌声の印象評価尺度に用いる語の選定（2.2 節）を行い，尺度を構築した．

2.1 仮尺度の構築

仮尺度構築の目的は，歌声の多様な印象を適切に形容できる語を選定することである．そのために，多様な印象評価語を収集（2.1.1 項）し，了解性調査（2.1.2 項）および同義性調査（2.1.3 項）を行った．以上の実験では歌声を用いないため，本論文ではこれを仮尺度と名付けた．

2.1.1 印象評価語の収集

この調査では，歌声の印象を形容しうる多様な評価語を収集するため，「A. 学術的に重要な語」「B. 専門的に使用される語」「C. 日常的に使用される語」という 3 つの観点に基づき，4 種類の調査対象（以降，ソースと呼ぶ）から，「歌声を評価していると考えられる語」を抽出した．

A. 学術的に重要な語 音楽を対象とした先行研究で扱われている語は有用だと考え，文献 [8], [9] で用いられていた評価語を抽出した．

B. 専門的に使用される語 音楽や歌声の評価を専門的に行う際には，多用な評価語が使用されると考えられる．そこで，音楽情報サイト[*1]における，専門家による CD レビューから「歌声を評価していると考えられる語」を手作業で抽出した．2010 年 6 月から 2012 年 6 月までの 350 枚分の CD レビューを対象としている．

C. 日常的に使用される語 音楽的知識を持たない一般人が歌声を評価する際に用いる評価語は，多くの評価者の間で判断基準が似通っている重要な語であると考えられる．そのため，SNS サービスと動画共有サイトから評価語の抽出を行った．

まず，SNS サービスとして Twitter[*2] を対象とし，「な歌」「い歌」という文字列を含む投稿を自動収集し，該当する評価語を手作業で抽出した．収集は 2012 年 8 月 1 日から 8 月 28 日まで，4 週間継続して行った．ツイートを取得する時間帯を日によってずらすことで様々な評価語が集まるよう工夫した．平日は 300 ツイート，ツイート数の増える休日は 500 ツイート，4 週間で計 10,000（$= (300 \times 5 + 500 \times 2) \times 4$）ツイートを調査した．

動画共有サイトとしてはニコニコ動画[*3]を対象とし，「歌ってみた」というタグがつけられている動画へのコメントを収集した．まず，「歌ってみた」というタグがつけられている動画から「再生が多い順」に 35 名の歌唱者を選出した（2012 年 8 月 1 日時点）．そして，各歌唱者による初投稿動画と最新の投稿動画から，それぞれ最新 500 コメントを収集し，その中から，「歌声を評価していると考えられる語」を手作業で抽出した．ここで，それぞれの歌唱者の動画を 2 種類用いたのは，人気のある歌唱者の場合，最新の動画であるほど，評価語に偏りがみられる可能性があるためである．

調査は 2012 年 8 月 1 日から 8 月 7 日の期間に行った．

収集した語の数を**表 1** に示す．収集の結果より，のべ 11,905 語の中からソースごとに重複を除外した 898 語を

[*1] ロッキング・オン：http://ro69.jp/
[*2] https://twitter.com/
[*3] http://www.nicovideo.jp/

© 2016 Information Processing Society of Japan

表 1 収集した語の数
Table 1 Number of collected words.

収集元	のべ数	異なり数
A. 先行研究 [8], [9]	180	162
B. CD レビュー	699	372
C. SNS サービス	10,000	294
C. 動画共有サイト	1,026	232
合計	11,905	898

選定した．そして，本尺度に不適切と考えられる語を除外し，590 語を了解性調査の対象とした．ここで，対象外とした 308 語は，聴取者に生じる反応を表したもの（癒される，泣きそうになるなど），固有名詞が含まれるもの（ジャイアンみたいななど），形容詞で代替できない動詞，印象ではなく歌唱技術を直接評価しているもの（高音がでている，ビブラートがきれいなど）である．

2.1.2 了解性調査

次に，選定された 590 語を用い，評価者 20 名に「評価語があてはまる歌声を想像できる」「評価語があてはまる歌声を想像できない」の二択で回答を求めるアンケート調査を行った．「想像できる」と答えた人数の割合を了解性とし，了解性が 0.85 以上であり，評価語収集の際に 2 種類以上のソースで観察された 64 語を次の同義性調査の対象とした．了解性が低かったために除外した語としては，「湿った」「ヌケの良い」「のんきな」があった．

2.1.3 同義性調査

同義性調査は 2 段階に分かれている．まず，評価者 10 名に対し，選定された 64 語に対応する縦横 64 マスの表を用い「ある評価語と似たような歌声を表す評価語」を選択するよう求めた．全 2,016（$= 64 \times (64-1)/2$）通りの評価語対のうち，3 割以上の評価者が「似ている」と回答した 562 対について，さらに詳細な調査を行う．

562 対の評価語を用い，評価者 10 名に「2 つの評価語が表す歌声はどの程度似ている」か歌声を用いずに想像させ，1 から 7 の 7 段階評価で回答を求めた．各評価者の回答を平均した値を同義性とし，同義性が 5 以上であった二対の語は，互いに同義性が高いと判断した．

同義性が高い語を統合した結果，44 語の評価語を仮尺度として選定した．その際，3 語以上にわたって同義性が高かった場合は，それらの評価語の多様性を保つよう配慮し，語の選択を行った．たとえば，「澄んだ」「透き通った」「クリアな」という 3 語は互いに同義性が高かったため，このグループからは了解性が最も高い「透き通った」という語を選択した．

2.1.4 結果

構築した仮尺度を**表 2** に示す．この仮尺度に含まれている評価語は「多様な歌声の印象を表す」「歌声の印象を想像しやすい」「同義性が極端に高い語を含まない」という条件を満たしているといえる．

表 2 印象推定に用いた歌声の印象評価語（47 語）
Table 2 The word set used for impression estimation (47 words).

仮尺度（44 語）		
甘い	心のこもった	（ドスが効いている）
安定している	こもっている	（伸びやかな）
勢いがある	（爽やかな）	激しい
（一生懸命な）	静かな	ハスキーな
色気のある	声量のある	鼻にかけたような
美しい	シャープな	響きのある
嬉しそうな	少女のような	（不安定な）
落ちつきのある	少年のような	ぶりっこみたいな
かっこいい	女性的な	（震えている）
悲しい	芯のある	真っすぐな
軽やかな	透き通った	無邪気な
可愛い	繊細な	優しい
聴きやすい	男性的な	（陽気な）
気持ち良さそうな	（中性的な）	弱い
元気な	特徴的な	

※括弧内は，因子分析（2 章）の際に除いたが，推定モデルの構築（3 章）で用いた評価語を示す

歌声評価に重要であると考えられる語（3 語）		
好きな	うまい	曲に合ってる

また，この 44 語とは別に，評価語収集の際に頻出していた表現のうち，歌唱技術に対する評価である「うまい」，個々人の好みを表す「好きな」，背景音楽との適合度合いを示す「曲に合ってる」の 3 語を，以降の分析では追加して用いる．これらの語は尺度構築においては対象外としたが，歌声の評価において重要な語であると考えられるため，因子分析には用いずに，印象推定モデルの構築（3 章）において分析対象とする．

2.2 歌声の印象評価尺度に用いる語の選定

構築された仮尺度をもとに，歌声の印象評価に適した評価語の選定を行う．まず，同一の楽曲を歌唱した歌声を収録（2.2.1 項）し，仮尺度を用いた印象評定実験（2.2.2 項）を行う．その結果を用いて因子分析（2.2.4 項）を行い，評価尺度として適切な評価語を選定する．さらに，歌声の印象空間についての考察も行う．

2.2.1 歌声収録

尺度構築に向けた印象評定実験では，以下の条件を満たす歌声を用いた．各条件を設定した理由を以下に示す．

歌詞，メロディ，テンポ，キー，が統一されていること
声質や抑揚の表現など，発声表現による印象の違いをとらえることを目的としているため．

評価者にとって未知のメロディ・歌詞であること 既存曲では，評価者の経験や知識が印象評価に影響を与える可能性があるため．

© 2016 Information Processing Society of Japan

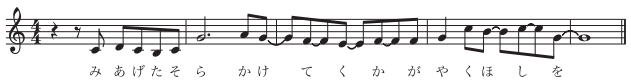


図 3 実験に用いたオリジナルメロディ
Fig. 3 An original song used at the experiments.

認知できる印象が多様であること 歌声刺激の印象の多様性が低いと，部分的な印象空間しか表現できなくなってしまう可能性があるため．

研究用の代表的な歌声コーパスとして，RWC 研究用音楽データベース [10] や，AIST ハミングデータベース [11] などがあげられるが，上記条件に該当するデータは存在しないため，印象評定実験の歌声刺激は新規に収録を行った．

9 秒程度のオリジナルメロディおよび歌詞（**図 3**）を作成し，21 名のアマチュア女性歌唱者の歌声を収録した．その際，多様な印象を与える歌声の収録を目指し，「一番うまく聴こえるように」「自分が歌いやすいように」「できるだけ平らに」「表現豊かに」「なるべく地声で」「なるべく裏声で」「誰かの歌い方を真似して（明確に歌い方を変えて）」という 7 種類の歌唱条件での歌唱を求めた．

収録は防音室で行い，ヘッドフォンで伴奏を聞きながら歌唱を行ってもらった．計 147（$= 21 \times 7$）歌唱を収録したが，後の印象評定実験での評価者の負担を抑えるため，同一歌唱者の中で聴取印象に大きな差が見られないデータを除外した．最終的に，60 データを印象評定実験の刺激として選定した．選定された 60 データは，21 名の歌唱者全員の歌声を 2〜5 データずつ含んでいる．

2.2.2 印象評定実験

選定した 60 データの歌声を対象とし，44 語の仮尺度および歌声評価に重要な 3 語による印象評定実験を行った．歌声を評価者に呈示する際，収録の際に用いた伴奏音は除外している．また，歌声の聴取回数に制限は設けなかった．評価者は 20 代で正常な聴力を有する一般大学生 19 名（男性 9 名，女性 10 名）である．ウェブ上のアンケートページを用い，各評価語がどの程度あてはまるか，7 段階での評価を求めた．

2.2.3 因子分析に用いる語の選定

印象評定の結果を用い，各評価語における「評価者間の相関」および「評価語間の相関」を求める．まず，歌声データ 60 種類における，各評価者同士，全 171（$= 19 \times (19-1)/2$）通りの相関の値を求め，平均値を算出した．この値が低い評価語は，評価者間の評価傾向が似ていないため，尺度の評価語としては不適切であるといえる．加えて，各評価語どうし，全 946（$= 44 \times (44-1)/2$）通りの相関の値を求める．この値が大きい評価語どうしは似たような評価傾向にあるため，評価尺度の語数を削減するために統合を行った．

「評価者間の相関」が 0.2 以下の評価語のうち，「評価語間の相関」が 0.75 以上であり，他の語で代替できると考えられる 7 語を除外した．また，「ドスが効いている」という語は，評価者によって多くの歌声に高い得点をつける評価者と，多くの歌声に低い得点をつける評価者に分かれていたため，除外した．評価者により評価傾向が大きく異なる語は尺度に不適切であると考えられるためである．

除外した評価語を，表 2 において括弧を付与して示す．

2.2.4 因子分析

印象評定実験の結果を評価者ごとに標準化し，歌声データごとに各語の平均値を算出した．36（$= 44-7-1$）語の印象評価得点を用い，因子分析を行う．因子数はスクリー基準に基づいて決定し，分析には最尤法，プロマックス回転を用いた．その結果，因子負荷量がどの因子においても 0.35 以下である評価語，また，独自性の値が極端に高い評価語を，尺度に不適切と見なし除外した．

さらに，各因子の内的一貫性の高さの指標となる Cronbach の α 係数 [12] を求め，すべての因子において $\alpha > 0.85$ となるまで，因子分析と評価語の除外を繰り返した．

2.2.5 結果

最終的に 12 語が尺度として適切であると判断された（**表 3**）．また，抽出された 3 因子に対し，各因子の因子負荷量が高い評価語を参考に，それぞれ「迫力性」「丁寧さ」「明るさ」と命名した．これらの因子の因子間相関の値を**表 4** に示す．この結果から，3 因子はある程度独立して，歌声の印象評価に寄与しているといえる．

表 3 完成した尺度の評価語と因子負荷量
Table 3 Impression words with the factor loadings in the impression scale.

	第 1 因子（迫力性）	第 2 因子（丁寧さ）	第 3 因子（明るさ）
勢いがある	**0.932**	0.044	0.024
声量のある	**0.917**	0.188	−0.192
弱い	**−0.898**	0.023	−0.008
静かな	**−0.752**	0.466	−0.166
聴きやすい	0.146	**1.001**	0.271
透き通った	−0.127	**0.886**	0.236
落ちつきのある	−0.286	**0.775**	−0.232
響きのある	0.387	**0.756**	−0.161
嬉しそうな	0.246	0.092	**0.923**
軽やかな	−0.037	0.358	**0.854**
可愛い	−0.286	0.145	**0.830**
無邪気な	−0.085	−0.359	**0.777**
寄与率	0.292	0.292	0.262
信頼性係数 α	0.926	0.893	0.877

表 4 3 因子の因子間相関
Table 4 Correlation between 3 factors.

	第 1 因子（迫力性）	第 2 因子（丁寧さ）	第 3 因子（明るさ）
第 1 因子（迫力性）	1.000		
第 2 因子（丁寧さ）	0.189	1.000	
第 3 因子（明るさ）	0.229	−0.132	1.000

© 2016 Information Processing Society of Japan

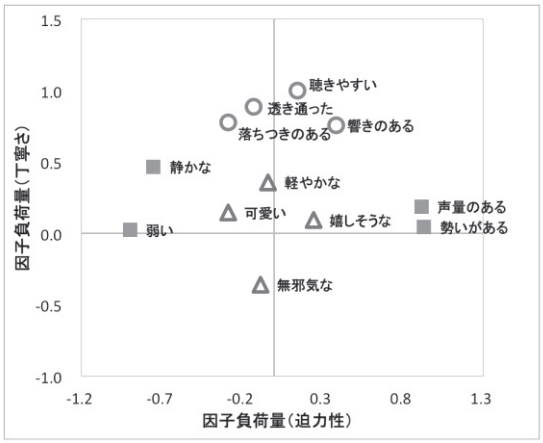


図 4 12 語の第 1 因子（迫力性），第 2 因子（丁寧さ）の因子負荷量
Fig. 4 Factor loadings of 12 words in “Powerful” and “Cautious”.

各因子における α 係数の値はすべて $\alpha > 0.85$ であり，各因子の内的一貫性が保たれているといえる．また，評価者の男女別に因子分析を行ったところ，男女ともに評価語の構成が似通った因子が得られたため，評価者の性別によらず，この尺度は有効であると考えられる．

図 4 では，迫力性と丁寧さの次元における 12 評価語の因子負荷量を示している．このように，12 語の評価語は 3 因子と対応しているため，各因子に関わる 4 語の得点を合計することで，3 因子の得点を算出することが可能である．

2.2.6 印象評価尺度および仮尺度の使用方法

印象評価尺度を構築したことにより，44 語で構成された仮尺度，および最終的に 12 語で構成された評価尺度が得られた．以下に，これら 2 つの尺度の特徴について述べる．

まず，12 語の評価尺度の特徴は，「歌声の印象を適切に評価できること」である．ここでいう評価とは，「聴取者が，ある歌声の印象を印象空間内のある位置に一意に定める」ことを指している．あくまでも，印象をある一点に定めることを目的としているため，聴取者ごとの主観的な評価の違いが少ない 12 語が選ばれている．

次に，44 語の仮尺度の特徴は，「歌声の印象を多様な語を用いて表現できること」である．この仮尺度は，幅広く収集された 898 語において，「歌声の印象として想像できる」という条件を満たした語で構成されている．また，極端に似通った語を除外しているため，44 語という限られた語数の中で，様々な歌声の印象を表現することが可能だといえる．ただし，この 44 語には聴取者ごとに認識が異なる語も含まれている．つまり，印象をある一点に定めるという本来の目的とは合致しないことからも，本論文ではこれを仮尺度と呼んでいる．

以上の特徴より，本研究では印象を推定する際，「印象空間内における位置把握を目的とした，3 因子の得点の提示」および，「印象の直感的な理解を目的とした，44 語における得点上位 5 語の提示」を行うこととする．後者で得点上位 5 語を提示するのは，複数の語を提示することにより，聴取者ごとの認識の違いを補うためである．

3. 印象推定モデルの構築

歌声の印象を音響特徴量から自動推定するためのモデル構築を行う．印象と音響特徴量を対応づけるための手法には様々な種類があるが，本論文では歌声の印象を線形モデルで表現することを試みる．そこで，印象の強度を連続的な値で推定可能である重回帰モデルの構築を行った．

重回帰モデルでは，説明変数どうしの相関によって引き起こされる多重共線性について考慮しなければならない．そこで，本論文では重回帰モデルの説明変数どうしの独立性を保つため，次の手順で重回帰モデルを構築した．まず，歌声から音響特徴量を抽出（3.1 節）し，それらの特徴量を用いて主成分分析（3.2 節）を行う．その合成得点（59 次元）を説明変数として扱い，2 章で求めた印象得点を推定するモデルを，各印象評価語ごとに構築する（3.3 節）．

3.1 音響特徴量の抽出

印象評定実験で用いた歌声データ 60 歌唱から，音響特徴量の抽出を行う．多様な楽曲に適用することを想定し，調査対象とする音響特徴量は，楽譜情報や歌詞の情報を用いずに抽出できるものとした．

分析に用いた歌声データは 44.1 kHz，16 bit サンプリングのモノラル信号である．まず，STRAIGHT [13] を用いて 1 ms ごとに F_0（基本周波数），スペクトル包絡，非周期性指標を推定する．分析フレームは 1 ms ごととし，それらを用いて計 221 種類の音響特徴量の抽出を行った．この節では，抽出した各特徴量の詳細について述べる．

本論文で抽出した音響特徴量は，抽出方法により次の 3 種に大別できる．なお，本論文では，1 歌唱ごとに，その有声区間における平均値，標準偏差，中央値，四分偏差を求め，これを統計特徴量と呼ぶ．

静的特徴量 1 フレームごとに抽出した特徴量を用い，統計特徴量を抽出（**図 5**）．

動的特徴量 複数のフレームにおける変動量を求め，統計特徴量を抽出（3 もしくは 4 種類のフレーム数を対象として，それぞれで変動量を計算）．

F_0 特徴量 ビブラートなど，F_0 に関わる特徴量を抽出．

抽出した特徴量については，**表 5** にまとめて示した．

本論文では，動的特徴量などの算出において回帰係数を用いるが，すべて以下の式に基づく．ここで y は分析対象とする特徴ベクトルであり，$2K+1$ はベクトルの長さを表している．たとえば，y にはスペクトル包絡や F_0 軌跡などが相当する．

© 2016 Information Processing Society of Japan

表 5 抽出した音響特徴量一覧
Table 5 Calculated acoustic features.

静的特徴量における統計特徴量			
対象とするスペクトル包絡		S_{lin}	S_{log}
スペクトル重心		○	○
スペクトル傾斜	0–22.05 kHz	○	○
	0–3 kHz	○	○
	0–6 kHz	○	○
	0–9 kHz	○	○
倍音構成	H1/H2	○	○
	奇数・偶数倍音の比	○	○
歌唱フォルマントらしさ		○	○
スペクトルフラックス		○	○
フォルマント	F_1	○	○
	F_2	○	
非周期性指標の総和		○	
非周期性指標の傾斜	0–22.05 kHz	○	
	0–3 kHz	○	
	0–6 kHz	○	
	0–9 kHz	○	

動的変動量における統計特徴量					
フレーム幅 K [ms]		10	25	50	100
フォルマント	F_1	○	○	○	
	F_2	○	○	○	
スペクトル	0–3 kHz	○	○	○	
	0–22.05 kHz	○	○	○	
F_0	$\Delta f0(t)$	○	○	○	○
	$\Delta\Delta f0(t)$	○	○	○	○
パワー		○	○	○	○

F_0 に関する特徴量
相対音高のピークの鋭さ，ピークの傾斜
フレーズ全体における cent の傾き（1 ms，1,000 ms）
フレーズ全体における cent の標準偏差（1 ms，1,000 ms）
ビブラートの速さに該当するパワーの最大値，平均，標準偏差
ビブラートらしさの最大値，平均，標準偏差
ビブラートと認定された区間における，上記の特徴量
ビブラートの速さ，深さの最大値，平均，標準偏差
有声区間中のビブラートと認定された区間の割合
F_0 の安定度（K=10, 25, 50, 100）

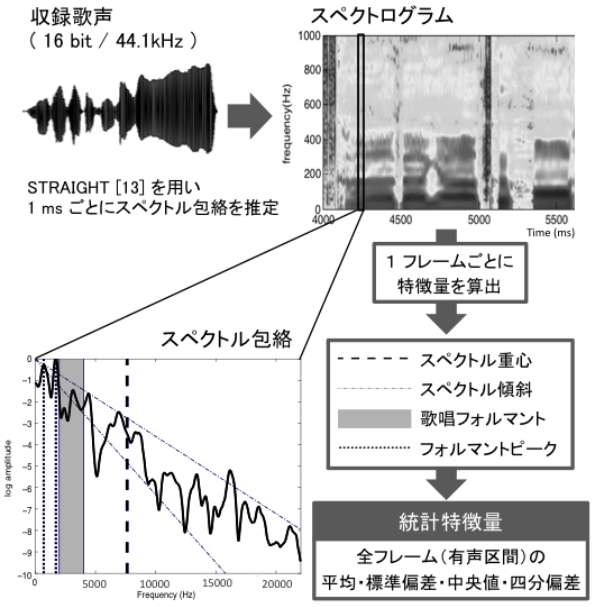


図 5 静的特徴量抽出の例
Fig. 5 Example of static features extraction.

$$R(y) = \frac{\sum_{k=-K}^{K} k \cdot y_k}{\sum_{k=-K}^{K} k^2} \tag{1}$$

3.1.1 スペクトル包絡に関する音響特徴量

スペクトル包絡は，歌声の声質を特徴づける重要な特徴量の 1 つであり，先行研究においても様々な検討がなされている（文献 [14] など）．本調査では，各時刻 t におけるスペクトル包絡 $S_{\mathrm{lin}}(f,t)$ および対数スペクトル包絡 $S_{\mathrm{log}}(f,t) = \log|S(f,t)|$ における以下の特徴量の抽出を行う．ここで，f は周波数ビンの番号を示している．

スペクトル重心 スペクトル重心は，*Timbral Texture Feature* として知られている [15]．スペクトル包絡 $S_{\mathrm{lin}}(f,t)$，対数スペクトル包絡 $S_{\mathrm{log}}(f,t)$ から，各時刻におけるスペクトル包絡の重心 $S_c(t)$ を，以下の式を用いて求め，統計特徴量を算出する．B は，周波数ビンの数を示している．

$$S_c(t) = \frac{\sum_{f=1}^{B}(f \cdot S_{\mathrm{lin|log}}(f,t))}{\sum_{f=1}^{B}(S_{\mathrm{lin|log}}(f,t))} \tag{2}$$

スペクトルフラックス スペクトルフラックスも *Timbral Texture Feature* として知られており，局所的なスペクトル変化の指標とされている [15]．時刻 t のフレームにより標準化されたスペクトル包絡 $S_{\mathrm{lin}}(f,t-1)$，対数スペクトル包絡 $S_{\mathrm{log}}(f,t-1)$ を用い，以下の式によりスペクトルフラックス $S_f(f,t)$ を求め，統計特徴量を算出する．

$$S_f(t) = \sum_{f=1}^{B}(S_{\mathrm{lin|log}}(f,t) - S_{\mathrm{lin|log}}(f,t-1))^2 \tag{3}$$

スペクトル傾斜 式 (1) を用いてスペクトル包絡 $S_{\mathrm{lin}}(f,t)$，対数スペクトル包絡 $S_{\mathrm{log}}(f,t)$ から，時刻ごとの傾きを求める．4 種類の帯域（0–3 kHz，0–6 kHz，0–9 kHz，0–22.05 kHz）におけるスペクトル傾斜を求め，統計特徴量を算出する．

Singer's Formant 歌声らしさや声の響きを評価する特徴量として Singer's Formant（歌唱フォルマント）が

© 2016 Information Processing Society of Japan

知られている [14], [16], [17]．本論文では，スペクトル包絡，対数スペクトル包絡の 2–4 kHz の帯域におけるパワーの全帯域に対する割合を歌唱フォルマントらしさの特徴量として求め，統計特徴量を抽出する．

スペクトルの倍音構造 基本波の強さ（F_0 に該当する周波数におけるパワー）は気息性の指標として知られているため，統計特徴量を算出する．また，倍音のパワー比は，歌声の声区の判別に有効であると報告されている [18], [19]．

本論文では，基本波のパワー H1 と第 2 倍音に該当するパワー H2 の比（H1/H2），および奇数倍音と偶数倍音に該当するパワーの総和の比を，スペクトル包絡から求め，統計特徴量を抽出する．

3.1.2 音韻性の知覚に関する音響特徴量

スペクトル包絡にはフォルマントに関する情報も含まれており，音韻の知覚や歌声の印象にも影響を及ぼすと考えられるため，関係する特徴量を抽出する．

フォルマントに関わる特徴量 フォルマントに関係する特徴量として，スペクトル包絡のピーク周波数を求める．まず，各時刻 t のスペクトル包絡のケプストラムの低次成分に対して逆フーリエ変換を行い，文献 [20] を参考に，フォルマント周波数である可能性が高いと考えられる帯域（$F_1 < 900\text{Hz}$，$900\text{Hz} < F_2 < 3{,}300\text{Hz}$）に制限したうえでピークの検出を行い，第 1 ピーク $F_1(t)$，第 2 ピーク $F_2(t)$ を求めた．$F_1(t)$，$F_2(t)$ の値を用い，統計特徴量を抽出する．

3.1.3 非周期性成分

STRAIGHT [13] では，スペクトル包絡の全体のエネルギーに対する非周期成分の割合を，0 から 1.0 の値で求めることができる．値が 1 に近づくほど，非周期成分の割合が多いことを示しており，歌声に含まれている非周期成分の大きさを評価することができる．

非周期性成分 スペクトル包絡全帯域における非周期性成分の値の総和を求め，統計特徴量を抽出する．

非周期性成分の傾斜 非周期性成分を式 (1) の $y(k)$ に代入し傾きを求める．4 種類の帯域における傾きを用い，統計特徴量を抽出する．

3.1.4 動的な特徴量

3.1.3 項までで扱った特徴量は，歌声の「声質」に関係する静的な特徴量である．歌声の印象の評価には，スペクトル包絡やフォルマントに関わる特徴量の動的な変動も関与していると考えられるため，以下の特徴量の算出を行う．それぞれ，分析フレーム幅を 1 フレームずつシフトさせながら回帰係数を求めるが，ある時刻の前後 K フレーム内に無声区間が含まれていた場合，その時刻は分析対象外とする．

パワーの動的変動量 以下の式により，各時刻におけるパワー $P(t)$ を求め，式 (1) を用い，回帰係数を求める．4 種類のフレーム幅（K=10, 25, 50, 100）を用い，有声区間の統計特徴量を抽出する．

$$P(t) = \sum_{f=1}^{B} S_{\text{lin}}(f,t) \tag{4}$$

スペクトル包絡の形状の動的変動量 スペクトル包絡 S_{lin} および対数スペクトル包絡 $S_{\log}$ の各周波数ビンにおける回帰係数 $\Delta S_{\text{lin}}(f,t)$ および $\Delta S_{\log}(f,t)$ を式 (1) を用いて求め，時刻 t における全周波数ビンの回帰係数の絶対値の総和を算出する．4 種類のフレーム幅（K=10, 25, 50, 100）を用い，有声区間の統計特徴量を抽出する．

フォルマントに関わる動的特徴量 $F_1(t)$ および $F_2(t)$ を用い，式 (1) により回帰係数を求める．3 種類のフレーム幅（K=10, 25, 50）における，統計特徴量を抽出する．

3.1.5 F_0 に関する特徴量

本論文で扱う周波数は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（$= 440 \times 2^{\frac{3}{12}-1} = 261.62\ldots\text{Hz}$）の cent 値を 4,800 cent とすると，周波数 f_{Hz} の音の cent 値 f_{cent} は

$$f_{\text{cent}} = 1200 \log_2\left(\frac{f_{\text{Hz}}}{f_c}\right) + 4800 \tag{5}$$

で表される．今後，本論文では基本周波数を $F_0(t)$ で表す．ここで，t は時間軸を示している．

相対音高 本論文では，楽譜情報を用いない特徴量を扱うため，歌声の相対音高に関する 2 種類の特徴量 [21] を算出する．この特徴量は，音高が半音（100 cent）単位で遷移しているかどうかを評価する指標である．具体的には，文献 [21] における相対音高の正確さ（$g(F)$）のピークの鋭さ，およびピークの傾斜を直線近似した傾き [21] を特徴量として扱う．また，半音ごとの遷移を評価するための異なる指標として，式 (6) を用いて $c(t)$ を求める．$c(t)$ から 50 ms ごとに平均を算出して $c(\bar{t})$ とする（平均算出のための分析フレームは 1,000 ms とした）．$c(t)$ および $c(\bar{t})$ を用い，有声区間の標準偏差を求めた．

$$c(t) = \text{mod}(f_{\text{cent}}, 100) \tag{6}$$

加えて，$c(t)$ および $c(\bar{t})$ を平均値が 0 になるよう標準化し，式 (1) に代入することで，歌声の有声区間における傾斜を求めた．時間経過による $c(t)$ のずれを評価する指標として用いる．

ビブラート ビブラートは歌唱力の評価に影響する重要な特徴量である [22]．そのため，文献 [22] と同様に時刻 t におけるビブラートの速さ（5–8 kHz）に相当する周波数帯域のパワー $\Psi_v(t)$ とビブラートらしさ $P_v(t)$ を求める．ビブラートの深さが 30–150 cent であり，分析区間（320 ms）の平均音高と 5 回以上交差する区間

© 2016 Information Processing Society of Japan

をビブラートであると定め，その区間における $\Psi_v(t)$ および $P_v(t)$ の最大値，平均値，標準偏差を算出する．また，有声区間においてビブラートであると判断された区間の割合，ビブラートの速さ（毎秒に生じる揺らぎの回数），深さ（平均音高からの音高の変動幅）も特徴量として扱う．本論文では，$F_0(t)$ から次式のようにビブラートを含む変動成分を抽出して $f_d(t)$ とした後，上記特徴量を抽出する．

$$f_d(t) = F_0(t) - f_l(t) \tag{7}$$

ここで，$f_l(t)$ は，$F_0(t)$ にカットオフ周波数 5 Hz のローパスフィルタをかけて変動を除去したものである．

F_0 の動的特徴量 歌声の $F_0(t)$ における重要な要素として，プレパレーションやオーバーシュート [23] など，異なる音高へ遷移する際の動的特徴がある．本論文では，式 (1) の $y(k)$ に $F_0(t)$ を代入して回帰係数 $\Delta F_0(t)$ を求め，F_0 の動的特徴量として扱う．4 種類のフレーム幅（K=10, 25, 50, 100）を用い，有声区間の統計特徴量を算出する．また，求めた $\Delta F_0(t)$ を式 (1) の $y(k)$ に代入して同様に $\Delta\Delta F_0(t)$ も求め，有声区間の統計特徴量を算出する．

F_0 の安定度 $\Delta F_0(t)$ において，有声区間中で変動がきわめて小さい部分（$|\Delta F_0(t)| < 0.0005$）の割合を求め，どの程度 $F_0(t)$ がぶれずに歌えているかを評価する．4 種類のフレーム幅（K=10, 25, 50, 100）を用いた．

3.2 音響特徴量の主成分分析

算出した 221 種類の音響特徴量を用い，主成分分析を行う．主成分分析により得られる合成得点を重回帰分析の説明変数として用いることにより，多重共線性などの問題を回避することができると考えられるためである．

音響特徴量を特徴量ごとに標準化し，主成分分析を行った結果，第 20 主成分までで累積寄与率が 90%に達した．主成分分析では，分析に用いたサンプル数（歌唱データ 60 歌唱）より一次元少ない数の主成分を得ることができるため，重回帰分析では，全 59 主成分を説明変数として用いることとする．

3.3 重回帰分析および交差検定

歌声の音響特徴量から印象得点を自動推定するため，重回帰分析を用いて重回帰モデルを構築する．用いた印象得点は，表 2 に示した 44 語，「好きな」「うまい」「曲に合ってる」という 3 語，および印象評価尺度をそれぞれの因子に対応づけて得点化した「迫力性」「丁寧さ」「明るさ」の 3 因子の得点，計 50（$= 44 + 3 + 3$）種類である．44 語の印象得点は，「得点上位 5 語を提示して印象の直感的な理解を助ける」ために，3 因子の印象得点は「印象空間内における位置を提示する」ために用いる．

計 50 種類の印象得点を標準化した値を目的変数，音響特徴量より合成した計 59 主成分の得点を標準化した値を説明変数とし，重回帰分析を行う．標準化された主成分得点を説明変数として用いたのは，各モデルの説明変数ごとの回帰係数を偏回帰係数として得るためである．偏回帰係数を得ることにより，各説明変数がどの程度印象推定に寄与しているかを表す指標として用いることが可能となる．説明変数の数が多いため，ステップワイズ変数選択法を用い，各印象得点ごとに計 50（$= 44 + 3 + 3$）種類のモデルを構築した．

モデルの評価は，自由度調整済み決定係数 $\hat{R}^2$，Leave-one-out 交差検定（LOO）による重相関係数 R_{LOO} で行う．さらに，特定の歌唱者を除いたデータでの交差検定を Leave-one-singer-out 交差検定（LOSO）と呼び，その重相関係数 R_{LOSO} も分析する．$\hat{R}^2$，R_{LOO}，R_{LOSO} の値が 1 に近いほど，モデルの推定精度が高いことを意味する．分析に用いた歌声データ数は 60 であり，以降 N で表す．

自由度調整済み決定係数 $\hat{R}^2$ 重回帰モデルでは説明変数が増えるほどモデルの説明力が高まるため，説明変数の数の多さを考慮した自由度調整済み決定係数 $\hat{R}^2$ を式 (8) により求める．ここで，m_n は印象評定実験による実測値，e_n はモデルによる推定値，$\bar{m}$ は実測値の平均値，N は歌声データ数（$= 60$），P はモデルに含まれる説明変数の数を表す．

$$\hat{R}^2 = 1 - \frac{\sum_{n=1}^{N}(m_n - e_n)^2/(N-P-1)}{\sum_{n=1}^{N}(m_n - \bar{m})^2/(N-1)} \tag{8}$$

重相関係数 R_{LOO} Leave-one-out（LOO）交差検定では，特定の歌声データを除外し，残りのデータを用いて重回帰モデルを作成する．その際，全データを用いて構築されたモデルで，印象推定に有効だと判断された特徴量を説明変数として用いる．そして，作成した重回帰モデルから．除外した歌声データの印象を推定することで，実測値と推定値の比較を行う．この分析を 60 データの歌声すべてに対して行い，全 60 データの歌声における印象得点の実測値 m_n（$n = 1, 2, \ldots, N$）と推定値 e_n（$n = 1, 2, \ldots, N$）におけるピアソンの積率相関係数（以降，相関係数と呼ぶ）を求める．得られた相関係数を二乗し，重相関係数 R_{LOO} を求めた．

重相関係数 R_{LOSO} Leave-one-singer-out（LOSO）交差検定では，同一歌唱者による歌声データの影響を排除するため，特定の歌唱者の歌声データを除き，LOO と同様の手順で重相関係数 R_{LOSO} を求めた．

3.4 結果と考察

重回帰分析の結果を示し，考察を行う．また，実際の歌声に対する印象推定例を示す．

© 2016 Information Processing Society of Japan

表 6 各印象推定モデルにおける自由度調整済み決定係数および重相関係数
Table 6 $\hat{R}^2$ and R of each constructed impression estimation model.

印象評価尺度の 12 語

	$\hat{R}^2$	R	
評価語		LOO	LOSO
勢いがある	0.840	0.791	0.811
声量のある	0.865	0.820	0.818
弱い	0.929	0.875	0.869
静かな	0.887	0.838	0.824
聴きやすい	0.800	0.703	0.691
透き通った	0.723	0.640	0.598
落ちつきのある	0.688	0.597	0.582
響きのある	0.698	0.600	0.612
嬉しそうな	0.534	0.454	0.419
軽やかな	0.518	0.449	0.447
可愛い	0.728	0.628	0.617
無邪気な	0.855	0.790	0.798
平均	0.755	0.682	0.674

歌声の印象評価における 3 因子

	$\hat{R}^2$	R	
因子		LOO	LOSO
迫力性	0.958	0.923	0.931
丁寧さ	0.551	0.471	0.462
明るさ	0.643	0.574	0.593
平均	0.717	0.656	0.662

歌声評価に重要であると考えられる語

	$\hat{R}^2$	R	
評価語		LOO	LOSO
好きな	0.555	0.454	0.387
うまい	0.386	0.302	0.217
曲に合ってる	0.238	0.176	0.131

$\hat{R}^2$ の値が大きかった 10 語

	$\hat{R}^2$	R	
評価語		LOO	LOSO
繊細な	0.938	0.900	0.896
弱い	0.929	0.875	0.869
激しい	0.925	0.887	0.872
気持ち良さそうな	0.888	0.809	0.812
静かな	0.887	0.838	0.824
声量のある	0.865	0.820	0.818
鼻にかけたような	0.862	0.766	0.786
無邪気な	0.855	0.790	0.798
優しい	0.851	0.791	0.800
勢いがある	0.840	0.791	0.811
参考：44 語の平均	0.685	0.607	0.595

表 7 各印象推定モデルにおける第 1 主成分から第 8 主成分の偏回帰係数
Table 7 Partial regression coefficient of typical principal component in estimation model.

印象推定モデル	各主成分に対応する偏回帰係数							
	1	2	3	4	5	6	7	8
迫力性	−1.15	−0.35	0.95				0.39	−0.45
丁寧さ				−0.65		−0.51		0.80
明るさ	0.69			0.67				0.64
勢いがある	−0.29		0.32					−0.10
声量のある	−0.33		0.24				0.12	
弱い	0.28		−0.18				−0.15	0.12
静かな	0.25	0.13	−0.21	−0.20				0.14
聴きやすい				−0.12				0.26
透き通った	0.13	0.16		−0.15				0.28
落ちつきのある			−0.19	−0.19		−0.14		0.16
響きのある	−0.15	0.15		−0.19	−0.12	−0.15		
嬉しそうな	0.10		0.10	0.14				0.10
軽やかな	0.15		0.11					0.14
可愛い	0.31			0.16				0.28
無邪気な	0.13			0.26				0.12

※空白箇所は，その主成分がモデルに採用されなかったことを示している

表 8 3.4.3 項で考察を行った各主成分の特徴と根拠となる特徴量
Table 8 Explanation of principle components considered at 3.4.3.

1	基本周波数の遷移の多様さ（$\Delta f0$ の標準偏差）
2	スペクトル傾斜（0–22.05 kHz における中央値） パワーの変動の小ささ（ΔP の平均*）
3	スペクトル包絡の変動の多さ（$\Delta S_{\log}$（全帯域）） 基本周波数の変動の多さ（$\Delta\Delta f0$ の中央値）
4	口唇の開口度合い・声道長の短さ（F_1 の平均） 立ち上がりの速さ（ΔS_{lin} の標準偏差）（0–3 kHz）
5	声質の多様性（スペクトル重心の標準偏差） ビブラートの少なさ（ビブラートの割合*）
6	対数スペクトル傾斜（0–6，0–9 kHz における中央値） 口唇の開口のメリハリのなさ（F_1 の四分偏差*）
7	調音運動の変動の多さ（ΔF_2 の四分偏差，中央値） 対数スペクトル傾斜（0–22.05 kHz における標準偏差）
8	ビブラートらしさ（最大値，平均値） 調音運動のメリハリ（ΔF_2 の標準偏差）

3.4.1 重回帰分析および交差検定の結果

重回帰分析および交差検定の結果を**表 6** に示す．各モデルは，すべて $p < .001$ で有意であった．印象評価尺度においては，「迫力性因子」や迫力性に関わる「勢いがある」「声量のある」「弱い」「静かな」といった語，および「聴きやすい」「無邪気な」という評価語では決定係数が $\hat{R}^2$ が 0.8 を超えており，特徴量からの印象推定精度が高いといえる．特に，「迫力性因子」に関しては R_{LOO} と R_{LOSO} の結果においても 0.9 を上回っており，モデル学習に用いていない歌唱者の歌声でも十分に印象推定が可能といえる．そのほかには「透き通った」「可愛い」といった評価語の推定精度が比較的高く，決定係数 $\hat{R}^2$ が 0.7 以上であった．44 語の評価語全体においては，$\hat{R}^2$ が 0.8 以上の語が 14 語，$\hat{R}^2$ が 0.7 以上の語が 25 語であった．

また，各重回帰モデルにおいて，どの主成分得点が大きく寄与していたかを，**表 7** に示す．それぞれ，各モデルの偏回帰係数が大きかった上位 5 主成分得点のうち，第 1 主成分から第 8 主成分に含まれている偏回帰係数を示している．加えて，各主成分得点がどのような発声方法に起因していると考えられるか，**表 8** にそれぞれの特徴を示した．

3.4.2 印象推定モデルについての考察

推定モデルにおける R_{LOSO} の値が R_{LOO} の値よりも小さい評価語では，歌声の印象が歌唱者に依存していると考えられる．たとえば，印象評価尺度における「透き通っ

© 2016 Information Processing Society of Japan

た」「嬉しそうな」といった評価語がこれに該当する．「うまい」「好きな」という評価語においても，R_{LOO} と比較して R_{LOSO} の値が小さい．歌唱技術の差はそれぞれの歌唱者に依存すると考えられるため，この推定結果は妥当といえる．また，「好きな」という評価語に関しても同様に，評価者の歌声の好みが歌唱者に依存していると考えられる．

3.4.3 主成分得点ごとの考察

表 7 では，印象評価尺度 12 語の推定モデルに採用された主成分の偏回帰係数を示している．数値が記載されていない主成分は，その印象の推定モデルでは用いられていないことを表しており，第 5，6，7 主成分では，寄与していた推定モデルが 12 語中 3 語以下であった．そこで，上位 8 主成分のうち，12 語中 4 語以上の印象推定に大きく寄与していると考えられる第 1，2，3，4，8 主成分についての考察を以下に述べる．それぞれの主成分において負荷量の高かった音響特徴量上位 10 種類（**表 9**）を取り上げ，考察を行う．表中，および本文中の*は，負荷量の値が負であったことを示す．

第 1 主成分　「少女のような」「可愛い」「伸びやかな*」「声量のある*」「男性的な*」「中性的な*」「優しい」「ドスが効いている*」「繊細な」「少年のような*」「軽やかな」「弱い」「芯のある*」「静かな」「迫力性因子*」「明るさ因子」の各モデル内で，偏回帰係数が最も高い説明変数となっている．負荷量が高い音響特徴量が F_0 の動的特徴量で占められており，特に $\Delta f_0(t)$ の標準偏差が重要であることから，F_0 が遷移する速度の多様さを反映していると考えられる．

第 2 主成分　「美しい」「心のこもった」「透き通った」「繊細な」「陽気な*」の各モデル内で，偏回帰係数が大きい説明変数である．この主成分では，パワーの動的変動量の小ささおよびスペクトル傾斜が大きく影響しており，ある種の声質表現を行おうとした結果，それにともないパワーの変動も小さくなっていると考えられる．主に合唱経験のある歌唱者の歌において得点が高くなっていたため，合唱における発声練習により習得可能な声質が関係していると考えられる．

第 3 主成分　「かっこいい」「激しい」「元気な」「勢いがある」「落ちつきのある*」「一生懸命な」「気持ち良さそうな」「シャープな」の各モデル内で，偏回帰係数が最も大きい説明変数である．スペクトル包絡の動的変動に大きく関与しており，また，F_0 の安定度が負の負荷量になっている点が特徴である．F_0 の動的特徴量も関与しているが，値の分散ではなく中央値が関係しているため，第 1 主成分とは異なり，1 歌唱中における変動の多さを反映させていると考えらえる．

第 4 主成分　「ぶりっこみたいな」「嬉しそうな」「響きのある*」「色気のある*」「悲しい*」「無邪気な」の各モデル内で，偏回帰係数が最も大きい説明変数である．第

表 9　各主成分において負荷量が高かった音響特徴量
Table 9　Acoustic features with high loading in each principal components.

第 1 主成分	F_0 の動的変動量 (K=10) の標準偏差
	ΔF_0 の動的変動量 (K=100) の中央値
	F_0 の動的変動量 (K=25) の標準偏差
	ΔF_0 の動的変動量 (K=10) の平均
	F_0 の動的変動量 (K=10) の平均
	F_0 の動的変動量 (K=25) の平均
	ΔF_0 の動的変動量 (K=25) の平均
	ΔF_0 の動的変動量 (K=10) の標準偏差
	F_0 の動的変動量 (K=50) の平均
	ΔF_0 の動的変動量 (K=50) の平均
第 2 主成分	スペクトル傾斜 (0–22.05 kHz) の中央値
	パワーの動的変動量 (K=50) の平均 *
	スペクトル傾斜 (0–22.05 kHz) の平均
	パワーの動的変動量 (K=25) の平均 *
	パワーの動的変動量 (K=50) の標準偏差 *
	スペクトル包絡全体の動的変動量 (K=50) の四分偏差 *
	スペクトル包絡全体の動的変動量 (K=50) の平均 *
	パワーの動的変動量 (K=100) の平均 *
	パワーの動的変動量 (K=25) の標準偏差 *
	スペクトル傾斜 (0–9 kHz) の平均
第 3 主成分	対数スペクトル包絡全体の動的変動量 (K=25) の中央値
	ΔF_0 の動的変動量 (K=10) の中央値
	F_0 の動的変動量 (K=10) の中央値
	対数スペクトル包絡全体の動的変動量 (K=50) の中央値
	F_0 の動的変動量 (K=25) の中央値
	F_0 の安定度合い (K=10) *
	F_0 の安定度合い (K=50) *
	スペクトル傾斜 (0–3 kHz) の平均
	F_0 の動的変動量 (K=10) の四分偏差
	スペクトル包絡 (0–3 kHz) の動的変動量 (K=25) の中央値
第 4 主成分	歌唱フォルマントの平均
	歌唱フォルマントらしさの中央値
	歌唱フォルマントらしさの標準偏差
	F_1 の平均
	歌唱フォルマントらしさの四分偏差
	スペクトル包絡 (0–3 kHz) の動的変動量 (K=50) の標準偏差
	スペクトル包絡 (0–3 kHz) の動的変動量 (K=25) の標準偏差
	F_1 の動的変動量 (K=50) の標準偏差
	スペクトル傾斜 (0–3 kHz) の標準偏差 *
	倍音構造 (H1/H2) の平均 *
第 8 主成分	対数スペクトル重心の四分偏差
	F_2 の動的変動量 (K=25) の標準偏差
	f_{cent} の標準偏差 (1000 ms 区間)*
	F_2 の平均
	F_2 の中央値
	F_2 の動的変動量 (K=50) の標準偏差
	対数スペクトルの重心の標準偏差
	ビブラートらしさの最大値
	F_2 の動的変動量 (K=10) の標準偏差
	ビブラートらしさの平均

© 2016 Information Processing Society of Japan

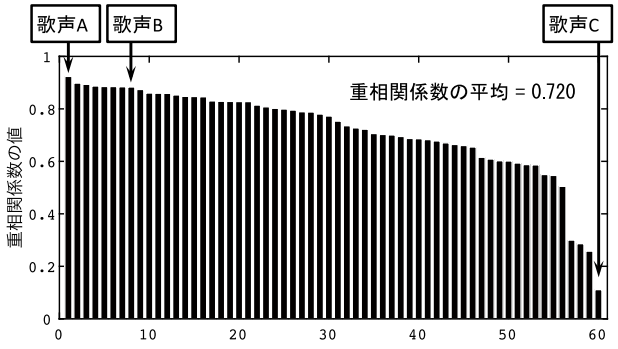


図 6 60 個の歌声データそれぞれにおける，50 種の推定値の重相関係数 $R_s^{I=50}$．横軸は重相関係数の値が高い順に歌声データを並べた番号である

Fig. 6 Distribution of $R_s^{I=50}$ with 50 scores for each recordings.

1 フォルマントと歌唱フォルマントが大きく関与している．歌唱フォルマントの有無は，一般的には「歌声らしさ」や「響き」と関連付けられており [24]，本分析での結果とは一致していない．本分析では 2–4 kHz の帯域のパワーの強さを歌唱フォルマントの指標としたが，F_1，つまり第 1 フォルマントの上昇により口唇の開口度が大きくなるのにともない，該当帯域のパワーも強調されたのではないかと考えられる．

第 8 主成分 「甘い」「女性的な」「聴きやすい」「透き通った」「丁寧さ因子」の各モデル内で，偏回帰係数が最も大きかった説明変数である．先行研究において，ビブラートは歌唱力評価に関係することが明らかにされており [1]，「うまい」という評価語と相関の高い「聴きやすい」が含まれていることなどから，先行研究と一致する結果が得られたといえる．また，F_2 に該当する第 2 フォルマントの変動は口内の舌の位置に影響されることが知られており，F_2 の変動の分散が大きいということは，調音運動にメリハリがある，と解釈できる．その結果，丁寧さに関わる評価語に関与していたのではないかと考えられる．

3.4.4 印象の自動推定例

歌声データごとの推定精度の指標として，重相関係数 R_s^I を求める．本論文で扱った 50 (= 44+3+3) 種類の印象得点を対象とし，印象評定実験による実測値 m_i $(i = 1, 2, \ldots, I)$ とモデルによる印象得点の推定値 e_i $(i = 1, 2, \ldots, I)$ の相関係数を求め，二乗することにより重相関係数 R_s^I を求める．ここで，I は対象とした印象得点の数を表す $(I = 50)$．60 個の各歌声データにおける重相関係数 $(R_s^{I=50})$ の分布を**図 6** に示す．この値が 1 に近いほど，推定値と実測値との誤差が少ないといえる．

ここで，全 60 歌唱における平均は 0.720 であり，高い精度で印象の自動推定ができているといえる．また，「うまい」「好きな」「曲に合ってる」という 3 語と 3 因子の得点を除いた重相関係数 $(R_s^{I=44})$ においては，全 60 歌唱の平均が 0.772 であった．つまり，44 語の印象評価語に限定した印象推定では，より高い精度で歌声の印象を自動推定できているといえる．

表 10 印象の自動推定例

Table 10 Examples of impression estimation.

	実測値		推定値	
歌声 A	美しい	1.181	美しい	1.294
	女性的な	1.103	伸びやかな	1.132
	響きのある	0.906	透き通った	1.097
	伸びやかな	0.893	落ちつきのある	0.978
	優しい	0.860	女性的な	0.862
歌声 B	かっこいい	1.805	声量のある	1.356
	芯のある	1.379	伸びやかな	1.098
	声量のある	1.310	かっこいい	1.095
	勢いがある	1.253	勢いがある	1.048
	安定している	1.069	芯のある	0.858
歌声 C	女性的な	1.228	女性的な	1.191
	ぶりっこみたいな	1.070	伸びやかな	0.733
	少女のような	0.966	真っすぐな	0.708
	特徴的な	0.778	気持ち良さそうな	0.677
	甘い	0.687	無邪気な	0.557

実際の推定例として，重相関係数が最も高かった歌声（歌声 A），歌声 A とは異なる印象であり 8 番目に重相関係数が高い歌声（歌声 B），最も低かった歌声（歌声 C）について，印象の自動推定結果を**表 10** に示す．ここでは，印象評価語 44 語における，印象得点上位 5 語を記載している．歌声 A では，上位 5 語のうち 3 語が重複しており，類似した印象を自動推定できているといえる．また，歌声 B についても，歌声 A とは異なる印象の傾向であるが，上位 5 語中 4 語が重複しており，印象の傾向によらず自動推定が行えていると考えられる．

一方，$R_s^{I=50}$ が最も小さかった歌声 C では，最も印象得点が高かった語は重複しているものの，ほか 4 語は大きく異なっている．図 6 において，特に相関の小さい（$R_s^{I=50} = 0.4$ 以下の）歌声データが 4 つあるが，それらは歌声 C と同様に「ぶりっこのような」「特徴的な」という印象の実測値が高い歌声であった．この 2 つの評価語の $\hat{R}^2$，R_{LOO}，R_{LOSO} の値は全 44 語中最も低く，自動推定が難しい語であるということが分かる．これらの歌声を聞いてみると，無理やり声を作っているような，歌声としての不自然さが感じられる歌声であった．このような，ある種の不自然さが生じることにより，主観評価が適切に行われず，モデルによる推定精度との誤差が大きくなったのだと推測される．

3.4.5 異なる歌唱者に対する印象推定精度

本論文では，日本のポピュラー音楽におけるアマチュア女性歌唱者の歌声に対して，音楽的な専門知識を持たない一般人が認知する印象を推定可能なモデルの作成を行った．様々な楽曲に対応できるよう，歌声の特徴量分析では

© 2016 Information Processing Society of Japan

表 11 異なる歌唱者に対する印象推定精度の評価に関する詳細
Table 11 Details of evaluation for impression estimation model in different singers.

楽曲に関する情報				
歌声	時間長	最低音	最高音	音域
D	7.4	59	66	7
E	12.4	62	66	4
F	9.6	58	73	15
G	8.7	60	75	15
H	10.9	56	68	12
I	12.4	59	71	12

最低音，最高音は MIDI ノートナンバーに換算した値を示している

印象得点に関する情報				
歌声	重相関係数 $R_S^{I=50}$	評価者間相関	レンジ	分散
D	0.256	0.114	2.20	0.47
E	0.251	0.455	3.50	0.97
F	0.643	0.602	3.94	1.19
G	0.640	0.474	4.21	1.03
H	0.802	0.558	3.92	1.18
I	0.593	0.462	3.16	0.97

重相関係数は 50 種の得点，それ以外は 44 種の得点を用いている

楽曲に依存しない特徴を扱っている．また，様々な歌唱者に広く対応できるよう，多様な印象が認知される歌声をモデル作成に用いた．しかし，印象推定精度の評価では「モデル作成に用いた歌声と同一の歌唱者」「同一の楽曲」における推定精度が評価対象であった．

そこで，前節（3.3 節）で作成したモデルを用い，異なる歌唱者に対する印象推定精度を評価する実験を行った．「モデル作成に用いた歌声とは異なる女性歌唱者」6 名に「歌唱者が歌い慣れている既存曲」を歌唱してもらい，10 名の一般大学生による主観印象評価の結果と，提案手法での印象推定結果を比較した．

表 11 は各歌声の楽曲情報および印象得点に関する情報を示している．$R_S^{I=50}$ は，2 章における印象評定実験（2.2.2 項）と同様，評価者ごとに 47 語の印象得点を標準化し，10 名分の平均点と本提案手法による推定得点の重相関係数を表している．全 6 データにおける $R_S^{I=50}$ の平均は 0.531 であり，同一楽曲，同一歌唱者の歌声を用いた場合よりも印象推定精度（重相関係数の値）が下がっていた．こうした異なる歌唱者が他の楽曲を歌った場合にも高い精度で推定するためには，今後の研究で，本提案手法を改良していく必要がある．

4. おわりに

本論文では，歌声の多様な印象を自動推定することを目的とし，まず，歌声の印象を適切に評価可能な評価尺度を構築した．その結果，歌声の印象評価に関わる因子として「迫力性」「丁寧さ」「明るさ」という 3 因子，およびこれらの因子の得点を算出するために必要な 12 語の評価語を得た．次に，歌声の音響特徴量と印象を対応づけるため，重回帰分析を行った．その結果，3 因子のモデルの決定係数 $\hat{R}^2$ について 0.958（迫力性），0.551（丁寧さ），0.643（明るさ）という結果を得た．また，60 歌唱それぞれにおける，50 種の印象得点の実測値と推定値の重相関係数 R_s を求めたところ，平均で 0.720 という値が得られたため，提案手法により歌声の多様な印象を高い精度で自動推定できていると考えられる．ただし，「ぶりっこみたいな」「特徴的な」という印象に関わる歌声では極端に R_s の値が小さくなってしまうため，これらの印象にどう対応するか，検討しなければならない．

本論文では，アマチュア女性歌唱者のみを対象としていたため，今後は男性歌唱者，あるいはプロ歌唱者を対象とした調査を行っていく．本研究の手順にならい，歌声収録（2.2.1 項）以降の調査をそれぞれの歌声に対応させることで，適切な印象推定モデルを構築できると考えられる．歌唱者の性別ごとにモデルを作成すると，歌声を入力する際に性別が限定されてしまうため，男女の歌唱音声が同一印象空間内にある場合の印象推定モデルについても検討が必要である．今後は音域やテンポなどが異なる多様な楽曲を用い，より頑健な印象推定モデルの構築を目指していきたい．

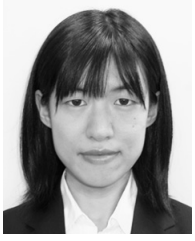
謝辞 本論文の一部は，科学技術振興機構 OngaCREST プロジェクトによる支援を受けました．

参考文献

[1] Nakano, T., Goto, M. and Hiraga, Y.: An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features, *Proc. INTERSPEECH 2006*, pp.1706–1709 (2006).
[2] Tsai, W.-H. and Lee, H.-C.: Automatic Evaluation of Karaoke Singing Based on Pitch, Volume, and Rhythm Features, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.20, No.4, pp.1233–1243 (2012).
[3] Daido, R., Ito, M., Makino, S. and Ito, A.: Automatic Evaluation of Singing Enthusiasm for Karaoke, *COMPUTER SPEECH AND LANGUAGE*, Vol.28, No.2, SI, pp.501–517 (2014).
[4] Luengo, I., Navas, E. and Hernaez, I.: Feature Analysis and Evaluation for Automatic Emotion Identification in Speech, *IEEE Trans. Multimedia*, Vol.12, No.6, pp.490–501 (2010).
[5] Vlasenko, B., Prylipko, D., Philippou-Hbner, D. and Wendemuth, A.: Vowels Formants Analysis Allows Straightforward Detection of High Arousal Acted and Spontaneous Emotions., *Proc. INTERSPEECH 2011*, pp.1577–1580 (2011).
[6] Scherer, K.R.: Expression of Emotion in Voice and Music, *Journal of Voice*, Vol.9, No.3, pp.235–248 (1995).
[7] Kotlyar, G.M. and Morozov, V.P.: Acoustical Correlates of the Emotional Content of Vocalized Speech, *Sov. Phys. Acoust.*, Vol.22, No.3, pp.208–211 (1976).
[8] 谷口高士：音楽作品の感情価測定尺度の作成および多面的感情状態尺度との関連の検討，心理学研究，Vol.65, No.6,

© 2016 Information Processing Society of Japan

pp.463–470 (1995).
[9] 平江 遼，西 隆司：感性に基づくクラシック音楽の分類，日本音響学会誌，Vol.64, No.10, pp.607–615 (2008).
[10] 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).
[11] 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，Vol.2005, No.82, pp.7–12 (2005).
[12] Cortina, J.M.: What is Coefficient Alpha? An Examination of Theory and Applications, *Journal of Applied Psychology*, Vol.78, No.1, pp.98–104 (1993).
[13] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: Restructuring Speech Representations Using a Pitch-adaptive Timefrequency Smoothing and an Instantaneous-frequency-based F0 Extraction: Possible Role of a Repetitive Structure in Sounds, *Speech Communication*, Vol.27, No.3-4, pp.187–207 (1999).
[14] Sundberg, J.: *The Science of the Singing Voice*, Northern Illinois University Press (1987).
[15] Tzanetakis, G. and Cook, P.: Musical Genre Classification of Audio Signals, *IEEE Trans. Speech and Audio Processing*, Vol.10, No.5, pp.293–302 (2002).
[16] 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).
[17] エリクソンドナ，齊藤 毅，細川久美子，岸本宏子，羽石英里：女声の「歌唱フォルマント」の音響学的研究：その 1，昭和音楽大学研究紀要，Vol.29, pp.13–26 (2010).
[18] 平山健太郎，伊藤克亘：ポピュラー歌唱における高音域の声区と発声状態の判別手法，情報処理学会研究報告音声言語情報処理，Vol.2012-SLP-90, No.16, pp.1–6 (2012).
[19] 小島 俊，齋藤 毅，中野倫靖，後藤真孝，三好正人：歌声における裏声と地声を識別するための音響特徴量の検討，電子情報通信学会技術研究報告：信学技報，Vol.112, No.266, pp.67–72 (2012).
[20] 田窪行則，前川喜久雄，窪薗晴夫，本多清志，白井克彦，中川聖一：岩波講座言語の科学 2，岩波書店 (1998).
[21] 中野倫靖，後藤真孝，平賀 譲：楽譜情報を用いない歌唱力自動評価手法，情報処理学会論文誌，Vol.48, No.1, pp.227–236 (2007).
[22] Nakano, T., Goto, M. and Hiraga, Y.: Subjective Evaluation of Common Singing Skills Using the Rank Ordering Method, *Proc. ICMPC 2006*, pp.1507–1512 (2006).
[23] Saitou, T., Unoki, M. and Akagi, M.: Development of an F0 Control Model Based on F0 Dynamic Characteristics for Singing-voice Synthesis, *Speech Communication*, Vol.46, No.34, pp.405–417 (2005).
[24] 齋藤 毅，辻 直也，鵜木祐史，赤木正人：歌声らしさの知覚モデルに基づいた歌声特有の音響特徴量の分析，日本音響学会誌，Vol.64, No.5, pp.267–277 (2008).



金礪 愛 （学生会員）

2012 年早稲田大学人間科学部人間情報科学科卒業．2014 年早稲田大学大学院人間科学研究科修士課程修了．修士（人間科学）．現在，同大学院人間科学研究科博士後期課程．

中野 倫靖 （正会員）

2008 年筑波大学大学院図書館情報メディア研究科博士後期課程修了．博士（情報学）．現在，産業技術総合研究所主任研究員．日本音響学会会員．2009 年情報処理学会山下記念研究賞（音楽情報科学研究会），2013 年 Sound and Music Computing Conference（SMC2013）The Best Paper Award 等各受賞．

後藤 真孝 （正会員）

1998 年早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．現在，産業技術総合研究所情報技術研究部門首席研究員兼メディアインタラクション研究グループ長．IPA 未踏 IT 人材発掘・育成事業プロジェクトマネージャー，情報処理学会理事等を兼任．日本学士院学術奨励賞，日本学術振興会賞，ドコモ・モバイル・サイエンス賞基礎科学部門優秀賞，科学技術分野の文部科学大臣表彰若手科学者賞，情報処理学会長尾真記念特別賞，星雲賞等，42 件受賞．

菊池 英明 （正会員）

1991 年早稲田大学理工学部電気工学科卒業，1993 年早稲田大学大学院修士課程修了．同年（株）日立製作所中央研究所入社．早稲田大学理工総研助手，国立国語研究所非常勤研究員，早稲田大学人間科学部非常勤講師・専任講師・准教授を経て，2012 年より早稲田大学人間科学学術院教授．博士（情報科学）．音声言語，音声対話，ヒューマン・エージェント・インタラクションの研究に従事．人工知能学会，日本音響学会，ヒューマンインタフェース学会，電子情報通信学会等各会員．

© 2016 Information Processing Society of Japan