Vol.2010-MUS-86 No.5
2010/7/28

実時間歌唱力補正に基づく新たなカラオケエンタテインメントの創出

森勢将雅[†1]　中野皓太[†1]　西浦敬信[†1]

我々は，実時間で歌唱力を補正することにより「誰が」「どんな曲でも」「簡単に」歌える新たなエンタテインメントについて研究を進めている．これまで使われてきたカラオケは，歌唱力の高い使用者は楽しめるが，歌唱力の低い使用者が楽しめるものでは無い．本稿では，事前に収録されたプロ歌手の歌声から歌唱力に相当するデータを抽出し，使用者の歌唱にリアルタイムで転写することにより，実時間歌唱力補正を実現するシステムについて述べる．本システムにより，歌唱力の低い使用者は歌唱力補正機能により楽しむことができ，歌唱力の高い使用者にとっても，他者に自らの歌唱力を提供するという新たな楽しみを与える利点がある．ここでは，実時間歌唱力補正を実現するための基盤技術，および計算速度を指標とした客観評価により有効性について論じる．

A new Karaoke entertainment based on the real-time singing style correction

Masanori Morise ,[†1] Kota Nakano [†1]
and Takanobu Nishiura[†1]

A new Karaoke entertainment to happily sing any music is proposed based on the real-time singing style correction. Although the conventional Karaoke application has entertained the skilled user, the unskilled user cannot enjoy it. The proposed system can solve the problem by the real-time singing style correction based on a professional singer's singing. In this paper, the method to extract the parameter about singing style is proposed to correct the singing style of the user. The entertainment for both the skilled users and unskilled users is also discussed.

1. はじめに

カラオケは，日本発祥の娯楽・文化であり，現在では世界各国で親しまれている主要なエンタテインメントの 1 つといえる．初期のカラオケは，カラオケ音源に併せて使用者が歌う機能のみであったが，よりエンタテインメント性を高めるため，使用者の歌唱力を評価する採点機能や，使用者が歌える楽曲の幅を増やすことを狙った楽曲のキーを変更する機能，入力された歌唱の音色を変える機能など，様々な工夫がなされてきた．さらに近年では，アマチュアクリエイタが作曲した楽曲を，投票によりカラオケに登録できるシステムの実現，自分の写真をカラオケの画像に用いる機能，プロ歌手と仮想的にデュエットさせるバーチャルデュエット機能など，より多くの使用者を満足させるサービスが提案されている．その一方で，歌唱力の低い使用者を満足させるために必要な歌唱力の制御に関する機能は存在しないのが現状である．

人間の歌声を対象とした変換技術では，従来 AutoTune や Melodyne 等のソフトウェアが用いられてきた．特に近年では，話し声を歌声に変換する SingBySpeaking[1)] や，2 名の歌声から，その中間的な歌声を合成できる歌唱モーフィング[2),3)] 等が提案されている．これらの研究は，主に「人間らしさを損なわない自然な歌声」を合成することを目的とし，SingBySpeaking では，歌声に存在する固有の成分を話し声に付与することで，どのような話し声からも自然な歌唱が合成することが可能とした．音声モーフィングでは，歌唱から「声質」と「歌い回し」を取り出し，他者のものと混合して再合成することで，声質を変えることなく歌い回しだけを他者のものと入れ替えることを可能にしている．しかしながら，人間の音声を対象としたこれらの歌唱変換技術は，実時間処理を行うことが出来ないため，カラオケにおける歌唱力の実時間補正は不可能であった．

本稿では，事前にテンプレートとして保存されたプロ歌手の歌唱力 (以下では，歌唱力テンプレートとする) を，リアルタイムで使用者の歌唱に転写することによる歌唱力の実時間補正技術について述べる．本技術を用いることで,「誰が」「どんな曲でも」「簡単に」歌えるようになるため，歌唱力の低い使用者を含め，従来のカラオケより多くの使用者を満足できる新たなエンタテインメント創出に繋がると考えられる．

以下，2 章では，本稿で提案するカラオケシステムのコンセプトと技術的な課題について

†1 立命館大学
Ritsumeikan University

Vol.2010-MUS-86 No.5
2010/7/28

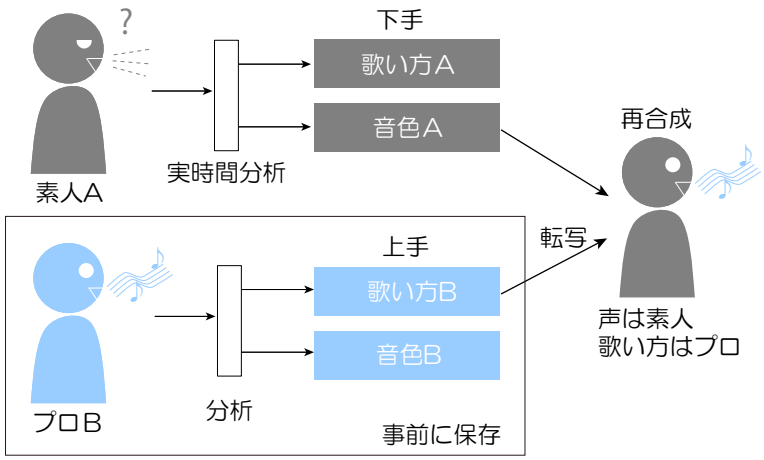

図 1 提案するカラオケシステムのアウトライン
Fig. 1 Outline of the proposed Karaoke system

述べる．3 章では，実時間歌唱力補正を実現する高品質歌唱合成法について説明し，本システムを実現するための問題点について述べる．4 章では，3 章で示した問題点を解決するための方法について述べる．5 章では，4 章で提案した方法が本稿で示した問題を解決しているか否か客観評価を行い，結果に基づいてカラオケシステム実現に対する考察を行う．最後に 6 章で，本稿のまとめを述べる．

2. 実時間歌唱力補正により実現されるカラオケエンタテインメントの概要

ここでは，提案する実時間歌唱力補正により創出されるエンタテインメント性の向上について述べる．提案技術は，図 1 のように，歌唱力テンプレートを事前に保持し，使用者の歌唱にリアルタイムで歌唱力テンプレートを転写することにより歌唱力の補正を行う．本稿では，歌唱力テンプレートを用いた新たなカラオケシステムのコンセプトを提案し，実時間歌唱力補正を実現するための課題について述べる．

2.1 歌唱力補正によるエンタテインメント性の向上

近年のカラオケは，プロ歌手とのバーチャルデュエットや，歌詞の表示される画面に使用者が好きな画像を設定できるなど多くの工夫が実用化されている一方，使用者の歌唱力そのものに対する工夫は無く，歌唱力の低い使用者にとって必ずしも楽しめるものではないという問題が改善されていない．提案するカラオケシステムは，誰でも簡単に歌えるため，歌唱力が低く人前で歌うことが出来ない使用者を楽しませることが可能になる．

使用者は，歌唱力テンプレートによる歌唱力補正を行うために，歌唱力テンプレートと同一のタイミングで歌うことが要求される．提案するカラオケシステムでは，このタイミングを併せる作業を使用者の責任とする．このタイミングを併せる作業をゲーム的要素と考え，視覚的にタイミングを指示する工夫を行うことで，エンタテインメント性を高めることが可能と考えられる．海外における家庭用カラオケゲームである SingStar においても，発話タイミングによりスコアを加算するシステムを導入することで，ゲームとしての楽しさを高めている[⋆1]．

また，ある楽曲を対象に用意された歌唱力テンプレートは，その楽曲に対し固有ではなく，同一の歌手が同じ楽曲を表情を変えて歌った場合，それぞれが異なる歌唱力テンプレートとして使用可能である．すなわち，使用者は，同一楽曲に対しても，様々な歌唱力テンプレートを用いることで，表情を変えて歌うことが可能となる．

2.2 歌唱力テンプレートを用いた新たな楽しみの創出

歌唱力テンプレートは，必ずプロ歌手の歌唱から作られるわけではなく，使用者の歌唱に基づいて歌唱力テンプレートを作ることも可能である．そのため，歌唱力に自信のある使用者は，自らの歌唱の歌唱力テンプレートを他者に配布することが可能となる．

歌唱力の低い使用者は，様々な歌唱力テンプレートを用いた実時間歌唱力テンプレートを用いて楽しく歌うことができる．さらに，歌唱力の高い使用者は，自らの歌唱力を歌唱力テンプレートとして配布し，多くの使用者に使ってもらうという楽しみが生まれる．本システムの実現により，歌唱力の低い使用者と高い使用者の両方にとって楽しめる，新たなエンタテインメントが実現されるだろう．

2.3 プロトタイプの実装と課題

これまでの検討[4] では，実時間処理を行うため，高精度な分析法ではなく簡易的な分析合成法を用いて実時間歌唱力補正が可能であることを示した[⋆2]．しかし，歌唱力に相当する物理パラメタの推定結果に大きな誤差が生じた場合，合成される歌唱の品質が大きく低下する問題があった．より高品質な実時間歌唱力補正を行うためには，各推定法の推定精度を，

⋆1 SingStar は，2004 年に海外で発売されたプレイステーション 2 用のゲームである．歌唱力補正機能は有さないが，家庭用ゲームとしてシリーズ累計で 1200 万本という驚異的なヒットを生み出した．数人単位のチームで順番に歌いスコアを競う「パス・ザ・マイク」等，ゲーム性を追及している．
⋆2 実時間歌唱力補正のデモ：http://www.youtube.com/watch?v=GtzeDAJQ-oU

実時間処理が可能な範囲で向上させる必要がある．本稿では，計算コストを削減するための分析法について述べ，実時間分析合成が可能であることを，音声分析に必要な時間を指標として行われた客観評価により明らかにする．

3. 高品質な歌唱合成を実現する基礎技術

SingBySpeaking や歌唱モーフィングによる歌唱変換は，高品質な音声合成が可能な方式である STRAIGHT[5),6)] や TANDEM-STRAIGHT[7)] を用いて行われる．本稿では，両方を区別せず示す場合は STRAIGHT とし，区別する場合，Legacy-STRAIGHT，TANDEM-STRAIGHT と表記する．STRAIGHT は，1939 年に提案された Vocoder[8)] と同一の機構を有する．具体的には，図 2 のように，音声から基本周波数，スペクトル包絡，非周期性指標と呼ばれる 3 つの物理パラメタを推定する分析法と，3 つの物理パラメタから音声波形を再合成する合成法から構成される．SingBySpeaking では，基本周波数に関する歌唱固有の特徴として，主に音高遷移とビブラートを付与する特徴がある．変換対象となる楽曲の譜面を既知とし，入力歌唱の対応付けを行うことで時間的な伸縮と歌唱の特徴付与が行われる．歌唱モーフィングは，基本周波数と時間的な発声のタイミングを「歌い回し」，スペクトル包絡と非周期性指標を「声質」と定義し，2 名の歌い回しと声質をモーフィングすることで歌い回しを直接制御できる．本稿における歌唱力テンプレートは，歌唱モーフィングにおける「歌い回し」と同一である．しかしながら，これらの方法は，特に音声の時間的な構造を対応付ける前処理が存在することから，カラオケにおける実時間歌唱力補正を行うことができない．

そのため，提案するカラオケシステムでは，入力音声のタイミングを歌唱力テンプレートと同一であると仮定し，歌唱力テンプレートの時間情報に併せて入力歌唱の基本周波数を補正する．使用者の歌唱の時間的タイミングが歌唱力テンプレートと一致する場合，対応付けを行う必要がない．よって，実時間歌唱力補正を実現するためには，STRAIGHT の計算コストに関する問題を解決すれば良い．

4. STRAIGHT の計算コスト削減

STRAIGHT による 3 つの分析法により，音声から基本周波数，スペクトル包絡，非周期性指標が推定される．表 1 は，各分析に必要な計算時間を分析時間比で示している．現在の STRAIGHT では，1 秒の音声を分析合成するために，その 5 倍以上の時間を必要としていることから，これを，1 秒未満にすることが，実時間歌唱力合成を行うための要求事項

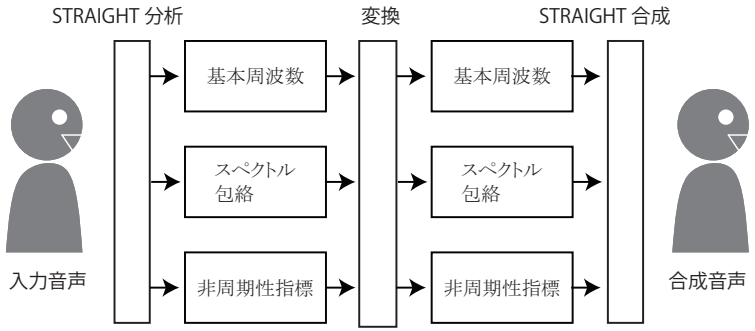

図 2 STRAIGHT(TANDEM-STRAIGHT) の処理の流れ
Fig. 2 Overview of STRAIGHT (TANDEM-STRAIGHT works the same way).

表 1 TANDEM-STRAIGHT を構成する各分析法の計算時間比
Table 1 Elapsed time ratio for each method of TANDEM-STRAIGHT

分析法	分析時間比
基本周波数	21.7 %
スペクトル包絡	5.5 %
非周期性指標	72.8 %
合計	100 %

となる．また，表 1 より，実時間処理の実現には，非周期性指標の分析コストを削減することが重要であるといえる．

4.1 DIO: 高 SNR の音声を対象とした高速な基本周波数推定法

基本周波数は，有声音に含まれる声帯振動の間隔を示すパラメタであり，この間隔を正確に推定することが基本周波数推定の目標といえる．人間の音声における基本周波数は時間とともに変化することため，周期性を仮定できる短い区間の波形を切り出して，時間軸における相関や，周波数軸におけるスペクトルから抽出された特徴量に基づいて推定する方法が一般的である[9)]．

従来法では，信号の時間波形における相関や[10)] や，ケプストラムを用いた調波構造[11),12)] を特徴量としてきた．特に，相関に基づく特徴量を用いて推定を行う YIN[13)] や，スペクトルの調波構造に基づく特徴量を用いて推定を行う SWIPE′[14)] は，1 %以下のエラー率を達成できることが各文献により示されている．

DIO (Distributed Inline-filter Operation) は，音声の SNR が十分に高いという制約を

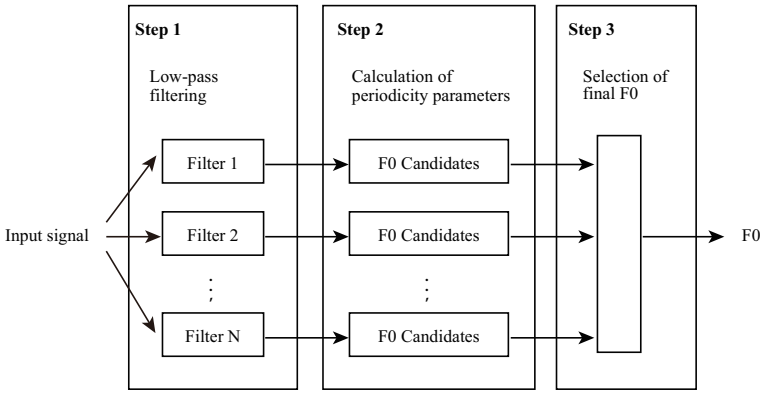

図 3 DIO のアウトライン
Fig. 3 Outline of DIO

満たす音声のみを対象に，極めて高速に F0 を推定できる方法として提案された[15]．本手法は，音声の SNR が十分に高い場合，基本周波数に起因する成分が，スペクトルに存在する最も低いピークを抽出すれば良いことに着目した方法であり，図 3 に示されるように，3 つのステップにより推定を行う．Step 1 では，基本波に相当するピークのみを抽出するため，様々なカットオフ周波数を持つローパスフィルタにより波形を処理する．Step 2 では，各信号から,「基本波らしさ」に相当するパラメタを取り出す．Step 3 では，計算された基本波らしさから，最終的な基本周波数の選定を行う．基本波らしさや方法の方法の詳細については，文献[15] を参照されたい．

4.2 STAR: ピッチ同期分析と対数パワースペクトル平滑化を組み合わせたスペクトル包絡推定法

音声のスペクトル包絡推定は，短時間フーリエ変換[16] を用いたパワースペクトル推定を基盤として発展してきた．有声音のスペクトルには，目的とするスペクトル包絡の情報だけではなく，基本周波数に起因するスペクトルの変動が含まれる．窓関数による波形切り出しが必要な短時間フーリエ変換ベースの方法を用いた場合，本来変化しないはずのスペクトル包絡が，窓関数により切り出す時刻に応じて変動するという問題が生じる．この影響される成分を取り除く方法として，ケプストラム[11),17] を用いた方法や，最尤スペクトル推定法[18]，線形予測[19] に基づく方法が提案されてきた．STRAIGHT では，基本周波数に起因するスペクトルの変動が,「分析時刻に依存した時間方向の変動成分」と,「パワースペクトル上に含まれる周波数方向の変動成分」であることに着目し，まず分析時刻に依存して変動する成分を除去し，その後パワースペクトルに含まれる変動成分を除去するという 2 段階の処理により，分析時刻に依存しないスペクトル包絡推定を行っていた．

STAR (Synchronous Technique and Adroit Restoration)[20] は，STRAIGHT を超える推定精度を達成しつつ，STRAIGHT よりも計算コストを削減できる方法として提案されたスペクトル包絡推定法である．STRAIGHT では，時間方向の変動成分，周波数方向の変動成分を独立して取り除く 2 段階処理で構成されるが，STAR では，時間-周波数方向の変動成分を 1 段階の処理で除去できる特徴を有する．

4.2.1 Step 1. ピッチ同期窓によるパワースペクトルの推定

Vocoder の考え方では，基本周期 T_0 の有声音 $y(t)$ は，音源となる声帯振動を模した周期 T_0 のパルス列 $x(t)$ と，調音フィルタのインパルス応答 $p(t)$ の畳み込みで表現される．音声のスペクトル包絡推定は，以下の式に示される有声音の波形 $y(t)$ から，調音フィルタのパワースペクトルを推定する問題と等しい．ここでは，議論を簡単にするため，$p(t)$ と基本周期 T_0 は，時間に対して変化しないものとする．

$$y(t) = p(t) * x(t), \tag{1}$$

$$x(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_0), \tag{2}$$

t は，時間を表現し，$*$ は，畳み込みを表現する記号とする．このとき，$y(t)$ のパワースペクトル $Y(\omega)$ は，以下となる．

$$\begin{aligned} Y(\omega) &= P(\omega)X(\omega) \\ &= P(\omega) \sum_{n=-\infty}^{\infty} \delta(t - n\omega_0), \end{aligned} \tag{3}$$

$$\omega_0 = \frac{2\pi}{T_0} = 2\pi f_0, \tag{4}$$

ω は角周波数を表し，$\delta(t)$ は $t = 0$ の場合のみ値を有するデルタ関数を示す．式 (3) より，有声音のスペクトルは，基本周波数 f_0 の整数倍の成分にしかパワーを持たないことが分かる．ディジタル信号処理では，波形を窓関数により切り出すため，周波数軸上では窓関数のスペクトルが畳み込まれることとなる．このとき，分析時刻に依存して各調波の位相が変化するため[7]，窓関数により切り出されたスペクトルも分析時刻に依存して変化する．

ピッチ同期分析では，基本周期の整数倍 (2 倍以上) のハニング窓を用いることで，基本

周波数 f_0 の整数倍のパワーに対し分析時刻に依存した成分を取り除くことができる[21]．音声のスペクトル包絡は短時間で変化するため，STRAIGHT においても，短い時間窓を用いて切り出しを行っていた．Legacy-STRAIGHT では，相補的時間窓と呼ばれる 2 つの窓関数のパワースペクトルを適切な重みで加算することで時間方向の変動成分を除去し，TANDEM-STRAIGHT では，ピッチ同期した窓関数を用い異なる時間に設定された 2 つの窓関数のパワースペクトルを平均することで，この時間方向の変動成分を除去していた．時間方向の変動成分を除去したパワースペクトルを周波数方向に平滑化することで，分析時刻に依存せず高い精度のスペクトル包絡を推定している．STAR では，パワースペクトルを計算する段階では，時間方向の変動成分，周波数方向に存在する変動成分の両方を除去せず，次節で示す対数パワースペクトルのピッチ同期平滑化により両方の変動成分を同時に取り除く．

4.2.2 Step 2. 対数パワースペクトルのピッチ同期平滑化

ケプストラムでは，リフタリング処理によりケフレンシー軸の低域成分のみを残すことでパワースペクトルの平滑化を行う．しかしながら，パルス列を畳み込むことによる影響のほとんどがケフレンシー軸における T_0 の整数倍であることから，リフタリング処理は，必要な情報まで欠落させるといえる．

STAR による平滑化後のパワースペクトル $|S_s(\omega)|^2$ は，以下の式により得られる．この式は，周波数軸において基本周波数の幅を持つ矩形窓を対数パワースペクトルに畳み込むことに相当する．

$$|S_s(\omega)|^2 = \exp\left(\frac{2}{\omega_0}\int_{\omega-\omega_0/2}^{\omega+\omega_0/2} \log\left(|Y(\omega)|\right) d\omega\right) \tag{5}$$

$Y(\omega)$ における負の周波数のスペクトル値は，折り返しにより得られる値をそのまま用いることとする．この平滑化処理により，時間・周波数方向の変動が同時に取り除かれる．

時間軸上の畳み込みは，周波数軸上では積となり，対数パワースペクトルは，以下の式に示される通り加算となる．

$$Y(\omega) = P(\omega)X(\omega) \tag{6}$$

$$\log\left(Y(\omega)\right) = \log\left(P(\omega)\right) + \log\left(X(\omega)\right) \tag{7}$$

式 (1) の説明で述べた通り，$\log\left(P(\omega)\right)$ は，分析時刻に依存せず一定の値である．すなわち，時間・周波数の変動は，$\log\left(X(\omega)\right)$ に含まれることが分かる．

周波数方向の変動成分は，ケプストラム法でも用いられるように，ケフレンシー軸上の高次に存在するピークとして現れる．一方，$\log\left(X(\omega)\right)$ は，窓関数の影響を除き，ほとんどの成分は T_0 の整数倍に生ずることとなる[⋆1]．式 (6) による平滑化は，ケフレンシー軸上の T_0 の自然数倍に 0 を持つフィルタをパワースペクトルに畳み込むことと等価である．したがって，ピッチ同期平滑化により，0 次ケプストラム以外の変動成分は取り除かれる．0 次ケプストラムは，概ね信号のパワーに一致することから，STAR で用いられる基本周期の 3 倍の窓長を持つハニング窓により切り出された波形から求められた対数パワースペクトルを平滑化することにより，時間，周波数方向に存在する変動を一度に除去できる．詳細については，文献[20] を参照されたい．

4.3 非周期性指標推定

STRAIGHT における非周期性指標とは，有声音中に含まれる周期信号と非周期信号とのエネルギー比率のスペクトルと定義されている．すなわち，非周期性成分のエネルギーが大きくなるよう非周期性指標を操作することで，擦れ声を合成することが可能となる．非周期性指標は，STRAIGHT において最も計算時間を必要とする一方で，品質に与える影響は，基本周波数・スペクトル包絡に比べると小さい．また，一般的に雑音のパワースペクトルを推定するためには，Welch 法[22] 等を用い，短時間スペクトルを平均する方法が用いられる．しかしながら，音声は時間と共に特徴が変化する信号であることから，短時間の波形からスペクトルを推定することが要求されるため，雑音のパワースペクトルを正確に推定することは不可能であり，何らかの近似を用いる必要がある．

TANDEM-STRAIGHT では，この問題に対処するため周波数軸を任意の個数の帯域に分割し，各帯域の非周期性成分を計算している．得られた各帯域の非周期性指標を線形補間することで，各離散周波数の非周期性指標を求める．本稿では，声帯振動のスペクトルが低域強調フィルタに近いことから，高域ほど雑音に脆弱であるという特性に着目し，非周期の成分が，周波数に依存して増加する (すなわち高い周波数ほど非周期的である) という仮定をする．非周期性指標は，0 から 1 の間を取るため，本稿では，シグモイド関数を基準とした以下の式により非周期性指標を与える方法を提案する．

$$A_p(\omega) = 1 - b - u(1 - s(\omega)) \tag{8}$$

$$s(\omega) = \frac{1}{1 + \exp(-\omega_0(\omega - \omega_c))}, \tag{9}$$

⋆1 実際には，窓関数により切り出される影響が含まれるため，T_0 の整数倍だけではなく，周辺のケフレンシー軸上に広がりを持って分布することとなる．

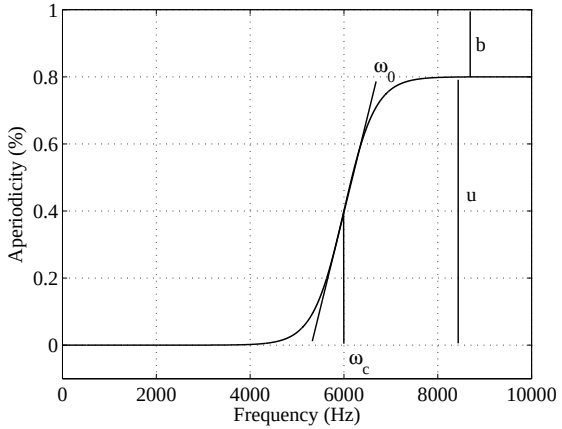

図 4 シグモイド関数に基づく非周期性指標の設計
Fig. 4 Design of the aperiodicity based on a sigmoid function

各パラメタと設計される非周期性指標との関係を図 4 に示す．ω_c は，非周期性指標が 0.5 となる周波数，ω_0 は，傾斜の急峻さを決定するパラメタ，u は，非周期性指標の最小値と最大値とのダイナミックレンジ，b は，最大となる非周期性指標の値 ($b < 1.0$) を表す．我々は，予備的検討により ω_c が基本周波数と概ね比例関係にあること，10 kHz 以上の成分は，概ね非周期性成分であることを確認した．本稿では，予備的検討による値により各パラメタの設定を行っているが，各パラメタの最適化が必要であり，今後検討を行う予定である．

4.4 無声音の分析合成

STRAIGHT では，基本周期を持たない無声音は，基本周期を一定値としてスペクトル包絡を計算していた．合成は，スペクトル包絡より得られるインパルス応答と，ホワイトノイズを畳み込むことで行われていた．無声音のスペクトル包絡を求めることで，音色制御が可能になるという利点がある一方，時間的にエネルギーが分散するホワイトノイズを音源とすることから，特に破裂音に関する品質低下が問題となる．

カラオケのような歌唱を分析合成の対象とした場合，無声音の音色変換は必ずしも必要ではないため，無声音の分析・合成処理は不要である可能性がある．提案するカラオケシステムでは，はじめに音声が有声音か無声音か判定を行い，無声音と判定された場合，入力波形をそのまま出力波形とすることにより，無声音の処理に係る計算コストを削減することとした．この方法は，音色制御が不可能になるというデメリットがある一方，無声音の品質低下が完全に 0 になること，計算コストが大幅に削減できるメリットがある．

5. 実験と考察

ここでは，提案法が必要とする計算時間に関する評価と，合成された音声の品質に関して行われた予備的検討の結果について示す．提案システムは，Matlab，および C 言語にて実装されており，C 言語版については，Web を通じて配布を行っている[⋆1]．計算速度の比較として，Matlab 版の TANDEM-STRAIGHT と提案法との計算時間比を求めた．分析には，男女数名分の音声試料を用い，1 音声について 10 回分析を行い，計測時間を求めた．得られた計算時間の比により比較評価を行っている．なお，10 回分析の計算時間を 10 回計測し，分散が 1 % 以下であることを確認している．

音声分析に必要な絶対時間については，C 言語版を用いて評価した．C 言語版の高速フーリエ変換には FFTW[⋆2] を利用して実装した．音声試料については，約 36 秒の歌唱を入力とし，計測に係った時間をそのまま評価値とした．なお，分析時刻の計測には，Core 2 Duo P9600 2.66 GHz の CPU，および 4.00 GB のメモリを登載した，OS が Windows Vista 64 ビットのノート PC を用いた．

評価結果を表 2 に示す．分析速度比は，値が大きいほど提案法のほうが高速であることを示す．DIO に基づく基本周波数分析は，TANDEM-STRAIGHT と比較すると 100 倍以上高速であることが分かる．また，36 秒もの長時間の音声を 500 msec 未満で推定できることからも，実時間処理が十分に可能であるといえる．スペクトル包絡推定に関しては，TANDEM-STRAIGHT が 2 つの窓関数のパワースペクトルを統合する処理，および，周波数平滑化後の後処理により，高速フーリエ変換を 4 回行う必要があるが，STAR は，高速フーリエ変換 2 回でほぼ同等の結果を計算することが出来る．非周期性指標に関しては，現在の実装が簡易的であるため，36 秒の音声を対象とした場合でも 1 msec 以下の分析時間となり，正確な計算速度比を求めることが出来なかった．しかしながら，基本周波数推定よりも圧倒的に高速に推定可能であることから，本稿で目標とした計算コスト削減は十分に行われているといえるだろう．

⋆1 提案システムのソースコードを配布している．現在は，C 言語版のプロトタイプのみが配布されている．
http://www.aspl.is.ritsumei.ac.jp/morise/world/
⋆2 FFTW: FFT を実装しているフリーのライブラリを配布している．http://www.fftw.org/

表 2 分析時間比と約 36 秒の音声分析に必要となった計算時間
Table 2 Elapsed time ratios and the elapsed times for a singing

分析法	TANDEM-STRAIGHT との分析速度比	36 秒の音声資料の分析に必要な時間
F0	103.1	409 msec
スペクトル包絡	1.7	12300 msec
非周期性指標	計測不能	1 msec 以下

5.1 計算コスト削減に関する考察

提案法により，音声分析に関する計算コストの問題は解決されたといえる．非周期性指標推定に関しては今後各係数の最適化が必須であるが，基本周波数に関する計算コストを大幅に削減できたことからも，実時間処理を達成することは十分可能だろう．

DIO は，十分に高い SNR の音声を対象とした評価では，TANDEM-STRAIGHT よりも高い精度で推定が可能である．しかしながら，通常の室内で録音された歌唱の分析では，特に高域の基本周波数において大きな推定誤差が生じることが指摘されている．これは，通常室内における暗騒音は低域が支配的であること，DIO は，基本波のみを取り出すため，基本波のエネルギーが暗騒音よりも十分に大きいことを条件としていることが原因であろう．そのため，低域通過フィルタにより波形を処理する前に，低域の雑音を抑圧することが必要であるといえる．具体的には，低域通過フィルタではなく，帯域通過フィルタを併用することにより，より雑音に頑健な推定が可能になると考えられる．

STAR は，スペクトル推定精度に関して，TANDEM-STRAIGHT よりも高い精度であることが示されている．TANDEM-STRAIGHT は，2 つの窓関数で計算されたパワースペクトルを平均する必要がある一方，STAR は，1 つの窓関数で計算されたパワースペクトルに基づいてスペクトル包絡を推定できる．すなわち，分析に対する時間分解能は，STAR のほうが高い．

5.2 品質に関する考察

提案法により合成された音声に関する非公式な品質評価を，様々な音声サンプルを対象として数名の被験者を対象として行った．ここでは，スペクトル包絡推定に関する評価として，基本周波数と非周期性指標については，TANDEM-STRAIGHT により推定された値を用いた．その結果，男性声の再合成音の品質は，TANDEM-STRAIGHT と STAR とでほぼ変わらないという結果が得られた一方，女性声の再合成音に関しては，TANDEM-STRAIGHT のほうが STAR よりも高いことが確認された．

この原因は，TANDEM-STRAIGHT では行われていた周波数平滑化後の後処理を STAR では行っていないことが原因であると考えられる．スペクトル距離に基づく客観評価では，TANDEM-STRAIGHT よりも STAR が優れているが，この評価に用いるスペクトル包絡の真値が単一極から構成される単純なものであったため，後処理による影響が生じなかった可能性がある．また，スペクトル距離による客観評価結果は，主観評価と相関はあるものの，完全な対応関係に無いことが原因といえるだろう．すなわち，TANDEM-STRAIGHT により行われる対数パワースペクトル平滑化後の後処理は，スペクトル距離の評価に関する貢献は少ないが，主観的な品質には大きな影響を与えていた可能性が示唆される．そのため，STAR により得られたスペクトル包絡に品質向上のための後処理を加えることを今後検討する予定である．

また，C 言語版を用いた合成音声は，一部の音声では TANDEM-STRAIGHT と等価である一方，特に女性声に関して高域に雑音が混入することが確認されている．これは，簡易的に計算された非周期性指標が不適切であることが原因である．品質を向上させるため，非周期性指標のパラメタ推定を高精度化することは今後の重要な課題である．

5.3 カラオケシステムへの適用

実装されたプロトタイプを用いて使用者に実時間歌唱力補正を体験させる評価を行った．システムの遅延は 40 msec 以下であり，テンポの速い曲ではやや遅延が感じられるものの，提案システムに適したビブラートなどを用いたスローテンポの曲に関しては，遅延は許容範囲であるという評価であった．歌唱力転写に関しては，使用者がタイミングよく歌唱された場合は，特にビブラートやこぶしなどまで表現できると好評であった．一方，タイミングを外した場合や，ブレスが大きく基本周波数推定が正しく行われなかった場合，大きく品質が損なわれることが指摘された．

これらの問題に対しては，SingStar で実装されている発生タイミングを視覚的に提示する機能の実現，DIO の推定精度を改善するための後処理により対応する予定である．また，カラオケのように背景に楽曲が存在する場合，非周期性指標の推定ミスによる知覚的影響は大きく緩和された．さらに，一般的なカラオケのように，合成歌唱にエコーを付与して出力することで，品質の劣化を覆い隠せる可能性がある．今後は,「カラオケシステム」として品質を高めるための検討を行う予定である．

6. おわりに

本稿では，カラオケにおけるエンタテインメント性を向上させる工夫として，歌唱力テン

プレートを用いた実時間歌唱力補正を提案した．実時間歌唱力補正を実現するためには，既存の音声分析変換合成システムの計算コストを削減する必要があることを示し，基本周波数分析法 DIO，スペクトル包絡推定法 STAR について述べ，実時間分析が可能であることを示した．

今後は，システムの本実装を行うと共に，基本周波数分析の精度を向上させる処理を追加する予定である．また，非周期性指標推定に必要となる推定パラメタの最適化を行う．さらに，実時間歌唱力補正に必須となる歌唱のタイミングを同期させるための工夫や，歌唱力テンプレートの自動生成について検討してゆきたい．

謝辞

本研究の一部は，IPA 未踏ユースプロジェクト，日本学術振興会 科学研究費補助金，科学技術振興機構 戦略的創造研究事業のデジタルメディア領域 CrestMuse プロジェクトの支援を受けて行われた．

参 考 文 献

1) 齋藤毅，後藤真孝，鵜木祐史，赤木正人，“SingBySpeaking：歌声知覚に重要な音響特徴を制御して話声を歌声に変換するシステム，” 音楽情報科学研究会，vol.2008, no.12, pp.25-32, 2008.
2) H. Kawahara, H. Matsui, “Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation,” Proc. ICASSP 2003, vol.I, pp.256-259, 2003.
3) 河原英紀，生駒太一，森勢将雅，高橋徹，豊田健一，片寄晴弘，“モーフィングに基づく歌唱デザインインタフェースの提案と初期的検討，” 情報処理学会論文誌，vol.48, no.12, pp.3637-3648, 2007.
4) 中野皓太，森勢将雅，西浦敬信，“音声合成を目的とした励起信号抽出に関する初期的検討，” 日本音響学会 2009 年秋季研究発表会, pp.415-416, 2009.
5) H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, “Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneus-frequency-based F0 extraction,” Speech Communication, vol.27, pp.187-207, 1999.
6) 河原英紀，“Vocoder のもう一つの可能性を探る–音声分析変換合成システム STRAIGHT の背景と展開–，” 日本音響学会誌，vol.63, no.8, pp.442-449, 2007.
7) H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, H. Banno, “A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation,” Proc. ICASSP 2008, pp.3933-3936, Las Vegas, 2008.
8) H. Dudley, “Remaking speech,” J. Acoust. Soc. Am., vol.11, no.2, pp.169-177, 1939.
9) W. Hess, “Pitch determination of speech signals,” Springer-Verlag, Berlin, 1983.
10) M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, H.J. Manley, “Average magnitude difference function pitch extractor,” IEEE Transactions on acoustic, speech, and signal processing, vol.ASSP-22, no.5, 1974.
11) A.M. Noll, “Short-time spectrum and “cepstrum” techniques for vocal pitch detection,” J. Acoust. Soc. Am., vol.36, no.2, pp.269-302, 1964.
12) A.M. Noll, “Cepstrum pitch determination,” J. Acoust. Soc. Am., vol.41, no.2, pp.293-309, 1967.
13) A. Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” J. Acoust. Soc. Am., vol.111, no.4, pp.1917-1930, 2002.
14) A. Camacho and J. G. Harris, “A sawtooth waveform inspired pitch estimator for speech and music,” J. Acoust. Soc. Am., vol.124, no.3, pp.1638-1652, 2008.
15) 森勢将雅，河原英紀，西浦敬信，“基本波検出に基づく高 SNR の音声を対象とした高速な F0 推定法，” 電子情報通信学会 論文誌 D，vol.J93-D, no.2, pp.109-117, Feb. 2010.
16) Rabiner, L. R., and Shafer, R. W., “Digital Processing of Speech Signals,” Prentice-Hall, NJ, 1978.
17) A.V. Oppenheim, “Speech analysis-synthesis system based on homomorphic filtering,” J. Acoust. Soc. Am., vol.45, no.2, pp.458-465, 1969.
18) 板倉文忠，斉藤収三，“統計的手法による音声スペクトル密度とホルマント周波数の推定，” 電子情報通信学会論文誌，vol.53-A, no.1, pp.35-42, 1970.
19) B.S. Atal, S.L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” J. Acoust. Soc. Am., vol.50, no.2B, pp.637-655, 1971.
20) 森勢将雅，中野皓太，西浦敬信, “歌唱合成システムの実現を目的とした高品質音声分析合成法の提案，” 信学技報, vol.110, no.71, pp.89-94, June 10-11, 2010.
21) M.V. Mathews, Joan E. Miller, and E.E. David, “Pitch synchronous analysis of voiced sounds,” J. Acoust. Soc. Am., vol.33, no.2, pp.179-185, 1961.
22) P.D. Welch, “The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, ” IEEE Trans. Audio Electroacoust. vol.15, no.2, pp.70-73, 1967.