音楽情報処理最前線！

未来の音楽の楽しみ方、作り方はどう変わるのか？
コンピュータは音楽を理解できるようになるのか？
コンピュータを使って音楽を研究する「音楽情報処理」
という研究分野が、世界的に注目を集めています。
本連載では、そうした最先端の研究事例を紹介していきます。

あなたの話声を歌声に変換する歌声合成システムSingBySpeaking

言葉を話すのと同じくらい簡単に色々な歌声を出してみたい。
でも、歌を歌うのは上手ではないし、ボイストレーニングを受けるのも大変だし…。
今回はそういった問題を解決する技術、「SingBySpeaking」を紹介します。

1. 言葉を話すだけで歌声を合成したい

「歌を歌うのは大好きだけど、自分の思い通りに歌うのは難しい」こんな悩みを持つ人は少なくないだろう。「言葉を話すように簡単に歌声を出してみたい！」と思った人はいないだろうか？　歌声（うたごえ）も、話声（はなしごえ）も、人間の声に他ならないのだから、このような希望を抱いてもおかしくはない。

もしも、歌詞を朗読するだけで自動で歌声に変換してくれるような技術があれば、上記の希望を少しでも満たしてくれるのでは…。そこで登場するのが、今回紹介する歌声合成システム「SingBySpeaking」である。

2. 話すだけで自分の声で歌ってくれる「SingBySpeaking」

歌声は、話声と同じ人間の声ではあるものの、実はその音響的な構造はかなり異なっていて、いろいろな「歌声特有の特徴」が存在する。「SingBySpeaking」は、歌詞を朗読した音声にこの「歌声特有の特徴」を付与することで、歌声へ変換することができる。

図1のように歌詞の朗読音声と、それに対応したメロディの譜面（MIDIデータ）を入力し、

①話声の高さF0（エフゼロ）（注1）
②時間長
③音色

の3つの情報を、それぞれに対応する3つの変換処理部で変換し、歌声を合成する。この変換処理が歌声特有の特徴を付与することに相当する。具体的には、まず本連載の2008年9月号でも紹介された音声分析変換合成法「STRAIGHT」を用いて、歌詞の朗読音声を声の高さに対応する「F0情報」と、音色に対応する「スペクトル情報」に分解。次に、F0の変化情報を譜面のMIDIデータで変換。話声の時間長や、音色を歌声のものへ変換。最後に、変換した各情報を使って「STRAIGHT」によって再合成することで、歌声を合成する。

以下では、各変換処理部の概要を紹介する。

① 声の高さを変換する（正確すぎる変化は人間らしくならない！）

歌声は、その高さの変化（F0変化）が楽曲のメロディに対応している点で話声とは大きく異なる。しかも、実際の歌声のF0変化は、MIDIデータのような階段状の矩形的な変化をするのではなく、図2に示すように、以下の4種の変動が含まれた複雑で滑らかな変化をしている。

○**オーバーシュート**：滑らかな音高変化、及び変化直後に目標音高を越える瞬時的な変動
○**ヴィブラート**：同一音高区間での準周期的な変動
○**プレパレーション**：音高変化直前に、変化とは逆方向に振れる瞬時的な変動
○**微細変動**：発声区間全体で観測される不規則で細かい変動

図1　SingBySpeakingの処理の流れ
自分が発話した歌詞の朗読音声と、メロディのMIDIデータから歌声を合成する。声の高さ（F0）、時間長、音色に対し歌声特有の特徴を付与する処理を段階的に施すことで、歌詞の朗読音声が歌声へと変換される。各変換処理部では、図2以降で示す歌声特有の音響特徴が付与される。歌詞の朗読音声の分析と、歌声を合成する処理において「STRAIGHT」を使用している。

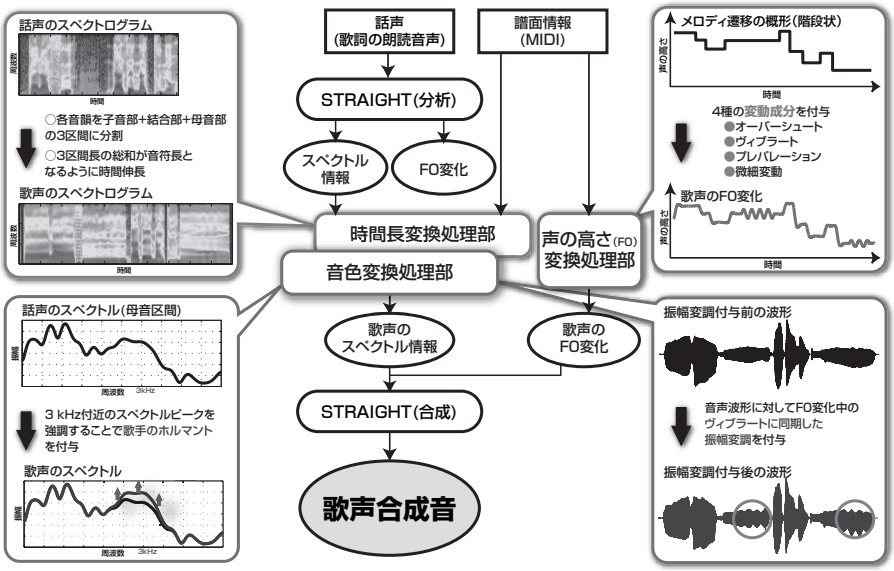
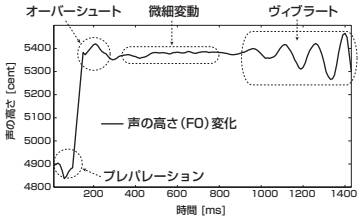


図2
声の高さ（F0）変化中の4種の変動成分の一例



ヴィブラートはプロの歌唱者に多い特徴だが、それ以外は歌唱技術に関係なく歌声共通に観測される。「SingBySpeaking」では、MIDIデータによるメロディ遷移の概形（階段状の変化）に対して、4種のF0の変動を付与することで歌声のF0変化を生成している。

齋藤 毅
（さいとう たけし）

2006年北陸先端科学技術大学院大学博士後期課程修了。博士（情報科学）。ATR認知情報科学研究所を経て、現在、産業技術総合研究所特別研究員。歌声知覚及び歌声合成、音声中の個人性知覚に関する研究に従事。「人間が如何にして歌声を発声し、聴いているのか？」という問題に対して、情報科学、心理学、生理学の多角的立場から取り組んでいる。

後藤 真孝
（ごとう まさたか）

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主任研究員。計算機によって実世界の音楽・音声コンテンツを自在に扱える技術の確立を目指し、音楽・音声の音響信号の自動理解と、それに基づくユーザインタフェースの研究を中心に、様々な研究課題に取り組んでいる。

「音楽情報科学研究会」へ参加してみませんか？

情報処理学会 音楽情報科学研究会（SIGMUS）は、コンピュータと音楽とが関わり合うあらゆる場面を活動対象とする学際的研究会で、年5回の研究発表会を開催しています。研究会に会員登録すると、参加できなかった研究発表会の論文集の郵送、過去の全研究発表会の論文のダウンロードなどの特典があります。研究会の登録方法や研究発表会の開催に関する最新情報などは http://www.ipsj.or.jp/sigmus/ をご覧ください。

読者の中には「ヴィブラートは歌声を表情豊かにしてくれるが、それ以外の変動は却って邪魔になるのでは？」と考える人がいるかもしれない。確かにヴィブラートは重要だが、それだけでは不十分である。実際、その他の3種の変動は、歌唱技術や歌唱スタイルに関係なく生み出される「歌声共通の成分」であることが、さまざまな研究でわかってきた。つまり、これらの変動の存在こそが、自然で「人間らしい歌声」としての条件と言え、歌声合成においても重要な役割を担っているのだ。

「SingBySpeaking」では、この4種の変動に着目し、MIDIデータとして入力されるメロディ遷移の概形に対して各変動を付与することで、歌声のF0変化を自動生成する。

② 声の時間長を変換する（子音・結合部・母音に分割！）

歌詞を朗読した場合と比べ、歌声中の個々の音韻長は大きく異なることになる。日本語の歌声の場合、一つの音韻（ひらがな1文字で表記できる音声）に一つの音符が割り当てられることが多い。つまり、歌声中の個々の音韻の長さ（音韻長）は、楽曲のテンポが与えられると、自動的に決定される。

そこで「SingBySpeaking」では、図3に示すように、音韻一つを子音部＋結合部＋母音部の3区間に分割し、個々の区間を個別に時間伸長する。具体的には、子音部はごく僅かに伸長し、結合部は40msに固定して、母音部は音符長からこれら2区間を引いた長さとなるように伸長している。これは面倒な処理なので、一見すると音韻全体を一様に伸長すれば良いのでは？と思うかもしれない。しかし、子音部や結合部をむやみに伸長すると人間の声として不自然になってしまう。つまり、声の自然性を保持した状態で、歌声の長さを自在に変換するためにこのような処理が必要になると言える。

③ 声の音色を変換する（歌声には話声にはない「響き」がある！）

F0と同様に、スペクトルにも歌声特有の特徴が観測され「SingBySpeaking」では2種類の特徴を話声に付与している。

一つ目の特徴は、図4に示した「歌手のホルマント（特定の周波数帯域にエネルギーが集中している部分）」（注2）である。これは、スペクトルの2.4～3.2 kHz付近に存在する顕著なピーク成分で、主に男性オペラ歌唱で（一部の邦楽歌唱においても）観測される。うまい歌唱にはこの特徴があることが多く、「歌声特有の響き」を与えることも明らかとなっている。ここでは、話声中の母音区間におけるスペクトルに対して、3 kHz付近のピークを強調することで、歌手のホルマントを付与している。これにより、自分の声の音色を、より歌声らしく変化させることができる。

二つ目の特徴は、音声振幅（声の大きさ）及びホルマントの、ヴィブラートに同期した振幅変調である。図5を見ると、F0変化におけるヴィブラートに同期して、音声波形全体とホルマントが変動していることが分かる。ここでは、F0変化を生成する際に付与したヴィブラートに同期させて声の大きさの変化に振幅変調を付与することで、この特徴を表現している。これにより、単にヴィブラートを付与した時より、はるかに表情豊かな歌声に変換される。

注1：音響信号処理分野では、声の高さを基本周波数と呼び、更にこれをF0と表記することが多い。
注2：singer's formant や singing formant と呼ばれることも多い。

3.「SingBySpeaking」がもたらすもの

「SingBySpeaking」は、話声を驚くほど自然で人間らしい歌声に変換してくれ、「言葉を話すように簡単に歌声を出してみたい」という希望を叶えてくれる。しかし、この研究がもたらすものはこれだけでは無い。歌声特有のさまざまな音響特徴を段階的に付与していきながら歌声を合成して聴くことで、自分の話声が歌声に徐々に変化していくのを、直感的に理解できる。つまりは、歌声を合成し、更には歌声の仕組みを理解できる歌声合成システムともいえるのだ。「SingBySpeaking」で合成した音声のサンプルは、**http://staff.aist.go.jp/saitou-t/singbyspeaking.html** 上で楽しむことができるので、是非とも聴いて確認してみて頂きたい。

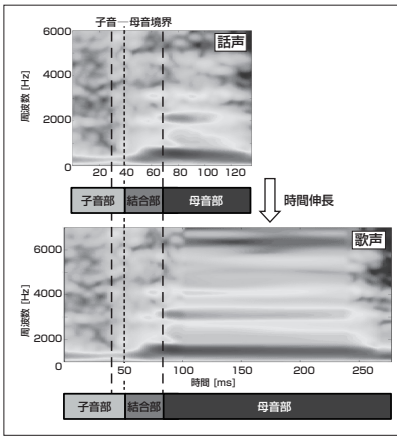


図3　時間長の変換処理
歌詞の朗読音声中の一つの音韻のサウンドスペクトログラムを示しており、子音部、結合部、母音部の各区間を独立に時間伸長できる。結合部の長さは40msで固定しており、母音部を最も長く伸長している。その際、音声波形を直接伸長するのではなく、スペクトルの時間系列を伸長していることに注意してもらいたい。

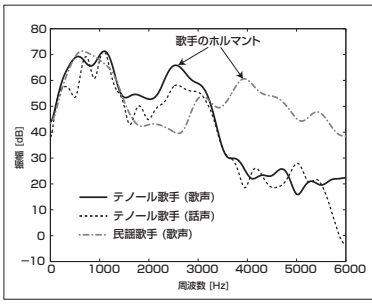


図4　「歌手のホルマント」の一例
テノール歌手の歌声と話声を比較すると、2.5kHz付近において歌声の方が大きなピークを持っていることが分かる。また、民謡歌唱のおいても、4kHz付近に同様にピークがある。いずれも「歌手のホルマント」であり、歌声の音色に大きな変化をもたらす。

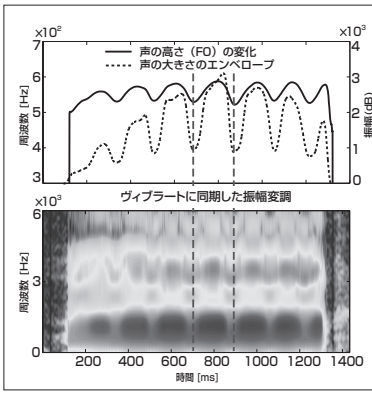


図5　ヴィブラートに同期した声の大きさ及びホルマントの振幅変調成分
プロのソプラノ歌手が発声した母音／a／の声の大きさの変化と高さ（F0）変化（上図）、及びサウンドスペクトログラム（下図）。F0変化中のヴィブラートに同期して、声の大きさとホルマント（サウンドスペクトログラム中の色が濃い帯状の箇所）が振幅方向に変調されていることが分かる。