産業界の技術動向

歌声合成について
―「初音ミク」を支える技術―

ヤマハ株式会社 研究開発センター
音声グループ マネージャ
剣　持　秀　紀

1. はじめに

最近、歌声合成技術が注目を集めています。これは、コンピュータに歌詞と音符を入力するだけで、歌声を合成するという技術です。動画投稿サイト「ニコニコ動画」には、「初音ミク」を筆頭とする歌声合成ソフト Vocaloid を用いてボーカルパートを作成した楽曲の動画が溢れ、市井のクリエータたちが日夜新曲を競って発表しています。人気が出た楽曲は大手レコード会社からリリースされ、ヒットチャートの上位に食い込むようになってきています。実際、2010 年 5 月 19 日に発売された“EXIT TUNES PRESENTS Vocalogenesis feat. 初音ミク”というアルバムは、並み居る人間の歌手を押しのけて、オリコンの週間ランキングで 1 位を獲得しました。合成音声による楽曲がヒットチャートの一画を占めることを 10 年前に誰が予想し得たでしょうか。また、人気楽曲の多くはカラオケとして配信され、カラオケでの人気曲のランキングの上位を占めるようになっています。2010 年の JOYSOUND のカラオケ年間総合ランキングでは、10 位のうち 5 曲が「ボカロ曲」（歌声合成ソフト Vocaloid を用いてボーカルパートを作成したオリジナル曲）となっています。特に若い世代の人々がカラオケボックスでこれらの楽曲を好んで歌っているようです。

図 1　VOCALOID「初音ミク」

本稿では、歌声合成の歴史を振り返りながら、筆者が開発に携わった歌声合成ソフト Vocaloid について紹介し、最後に最近の歌声合成技術をとりまく状況と今後の展望について述べます。

2. 歌声合成の意義

最近では、さまざまな楽器がコンピュータやシンセサイザの上で納得出来る品質で再現できるようになってきています。実際に、商業音楽ではほぼ 100％といって良いほど電子的に作られた音が含まれています。ところが歌声だけは、スタジオに歌手を呼び、それを録音するというスタイルがずっと続けられてきていました。歌声も将来は電子的に合成されたものが「普通に」使用される時代が来ると考え、開発を進めてきました。

歌声はメロディーを奏でるという点が楽器と共通していますが、楽器と違う点があります。それは、歌声は歌詞を持っているということです。歌詞を持っているということは、音色が異なるということであり、楽器に喩えるのであれば、様々な異なる楽器をリアルタイムに切り替えながら演奏していることに等しいことになります。ですから、これまでのシンセサイザと同じような考え方では、「歌うシンセ」

は実現できません。

一方で、歌声には、音声という側面もあります。歌声の音声としての性質とは、発音器官を共有しているという点です。しかし、話し声と歌声とは決定的な違いがあります。歌声は、音程とリズムが楽譜（あるいはそれと同等のもの）に支配されるという点です。したがって、朗読音声と比べると、歌声はリズムや音程が多様性を持ちます。テキスト音声合成（入力した文章をもとに音声を合成する技術）では、ナレーターに長時間朗読音声を読み上げてもらい、入力したテキストで最も似ている部分を取り出して接続する技術（コーパスベース方式）が確立し、自然な音声の読み上げが実現されていますが、歌声の場合は、リズムや音程の多様性と歌詞の組み合わせを考えると、この方式は不可能です。これに加えて、歌声はそれ自体が鑑賞される対象になります。合成という観点からみると、伸ばし音が「美しい」かどうかは重要なポイントになります。合成音の品質はいわゆるハイファイであることが求められます。

3. 歌声合成の歴史

さて、ここで歌声合成の歴史を振り返ってみます。

世界初の合成による歌声は、1962 年にベル研究所の Kelly らによって発表された “Daisy, daisy…” という歌声です [1]。これは、音響管モデル（acoustic tube model）と呼ばれる、滑らかに管の直径が変化するという簡単な形で声道を表現することにより、歌声の生成を物理的にシミュレートしたものです。このときの歌声は文化的にも大きな影響を残し、1968 年に公開された映画「2001 年宇宙の旅」の最後のシーンでコンピュータ HAL9000 が停止する直前に “Daisy, daisy …” と歌う場面にも影響を与えたと言われています。今その合成音を聴くと、1960 年代にこれだけのクオリティで歌声合成が達成されていたことに驚きを禁じ得えません。しかしながら、物理モデルは精密にモデリングしようとすればするほど扱うパラメータの数が膨大になるという欠点もあり、その後も物理モデルによる歌声合成のチャレンジはいくつかあったものの、未だ実用には至っていません。

一方では、歌声も音声の一種であり、音声の分析合成の研究開発の成果も歌声合成に充分に生かされています。ソースフィルタモデル（音声の生成過程をソース（声帯の振動）とフィルタ（声道による調音）に分けて考えるモデル）もその成果の一つです。1980 年に Klatt らが発表した MITalk（のちの DECTalk）は、二次 IIR フィルタ群の並列および直列構成により声道の調音を表現しています [2]。DECTalk はテキスト音声合成を目指して開発されましたが、それを用いた歌声合成もよく知られています。

物理モデルもソースフィルタモデルも音声の生成過程をモデリングしたものとなっていますが、一方では音声の生成過程にとらわれず、発音された音のスペクトルをそのままモデリングする手法も歌声合成に取り入れられています。McAulay らによって発表された正弦波モデリング [3] は、音声信号を短時間 FFT により正弦波の強度、周波数および位相を時間的に変化する関数として表現します。この手法を分析合成方法として使用した歌声合成も提案されています [4]。

研究レベルの歌声合成技術だけではなく、商用のシステムも過去にいくつか市販されています。1997 年にヤマハから発売された PLG-100SG という商品は、MU2000 等の MIDI 音源モジュールに装着して歌声合成機能を提供するものです。FM 音源をベースとした方式により歌声を合成しています。1999 年に KAE Labs（カナダ）より発売された VocalWriter は、Macintosh 用の歌声合成ソフトウェアです [5]。合成方式は不明ですが、音を聞くと、上述の DECTalk の方式に類似した手法であると推測されます。歌声だけでなく、伴奏もこのソフトウェア上で制作することが可能です。2000 年に NTT より発表された正弦波重畳方式により歌声を合成する技術（HORN 法）[6] は、「ワンダーホルン」という名称のソフトウェアとして、NTT アドバンストテクノロジー（株）より発売されています。また、歌声の楽器的な側面に注目したアプローチとしては、2004 年に Virsyn（ドイツ）から発売された CANTOR という

商品が挙げられます。合成方式は、ソースフィルタモデルをベースにしたものと推測されます。音楽制作環境に特化し、各種のソフトウェアプラグイン規格に対応したソフトウェアシンセサイザという形態で発売されています。

しかしながら、過去の歌声合成システムには、いくつかの問題点がありました。それは、(1) 歌詞が聞き取れない、聞き取りにくい、(2) 歌声が自然でなく、ブザー音的に聞こえる、(3) 直感的な入力インタフェースが無く、難しいコマンド列を入力しなければならない、等です。歌声合成技術を楽曲制作に「普通に」使っていただくためには、これらとは逆の条件が必要です。すなわち、(a) 歌詞が聞き取れること（了解性)、(b) 自然な揺らぎや息の音が含まれること（自然性)、(c) 簡単に入力できること（操作性）です。筆者が手がけた歌声合成技術 Vocaloid は、この条件をクリアし、音楽制作の現場で「普通に」使っていただけることを目標として開発を進めてきました。次節では Vocaloid について簡単に紹介します。

4. 歌声合成システム VOCALOID

Vocaloid は、ヤマハが開発し、ライセンシングを行っている歌声合成ソフトウェアです。実際の歌手の歌声から取り出した音声素片を入力された楽譜情報に合うように接続することで合成を行っています [7]。テキスト音声合成で用いられる大規模コーパスベースのものとは異なり、比較的小さな単位の素片（音声の断片）を持ち、それらのピッチ（音程）を変更し、音色を調整して滑らかに接続します。図 2 に全体のブロック図を示します。

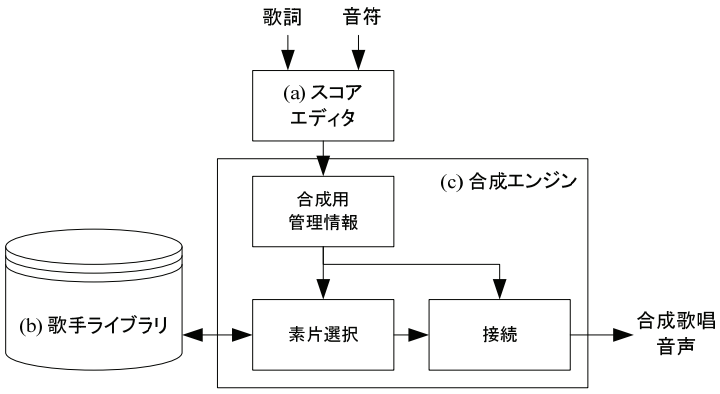

図 2　VOCALOID ブロック図

スコアエディタは、音符と歌詞を分かりやすく入力できるようになっています。図 3 にスコアエディタのスクリーンショットを示します。音符はピアノロール形式（音符の位置を縦に音階、横に時間軸で表したもの）で表現します。歌詞は音符の上に直接入力できるようになっています。

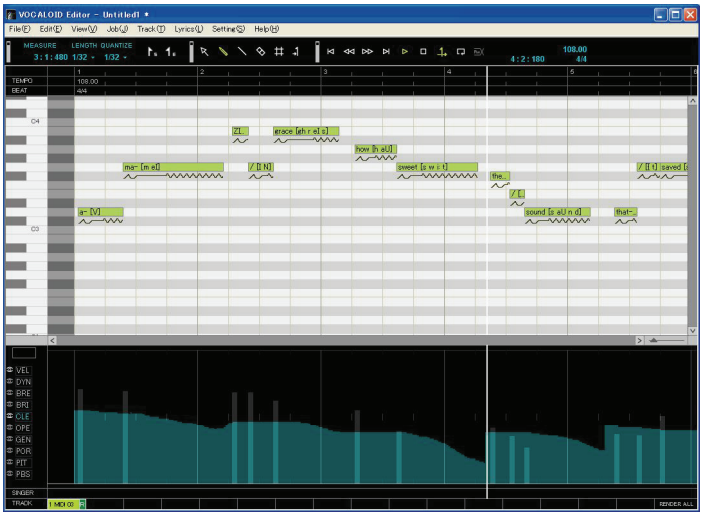

図 3　スコアエディタ

合成エンジンは必要な音声素片を歌声ライブラリから取り出し、連結して合成します。その際の素片の使用タイミングは、C-V（子音―母音）という素片の V（母音）の開始位置と音符開始のタイミングが合うように位置の調整が行われます。このようにしないとリズムやタイミングが明らかに違うものとして感じられるからです。素片の接続には、素片のピッチ（音程）を所望のピッチに変換する必要がありますが、仮に 2 つの素片の接続部分のピッチを合わせたとしても、単純に接続するだけでは二つの素片の音色の差がノイズとな

って現れてしまいます。素片連結時には音色も合わせ込む必要があります。これが合成エンジンの一番のポイントかもしれません。現状では、伸ばし音の区間で隣り合う二音素連鎖のスペクトル包絡を補間することで音色の合わせ込みを行っています。図4にその例を示します。図4では“sing”([sIN])という歌詞の伸ばし音のスペクトル包絡は、伸ばし音に先行するdiphone（音素と音素が変化する部分）すなわち[s-I]の最終フレームと、伸ばし音後のdiphone[I-N]の最初のフレームのスペクトル包絡を時間的に補間することで求められます。スペクトルの微細構造は伸ばし音の音声素片のものを使用します。これにより原理的に連結部分で音色の急激な変化が発生しないようになっています。

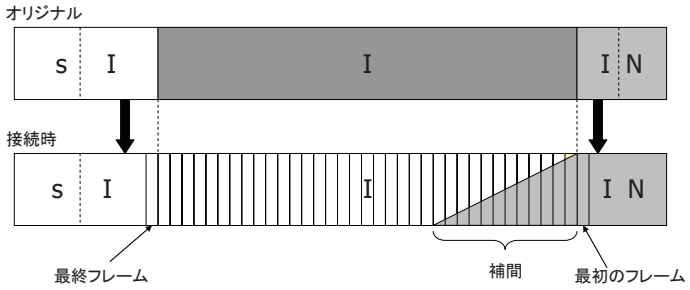

図4　スペクトル包絡の補間

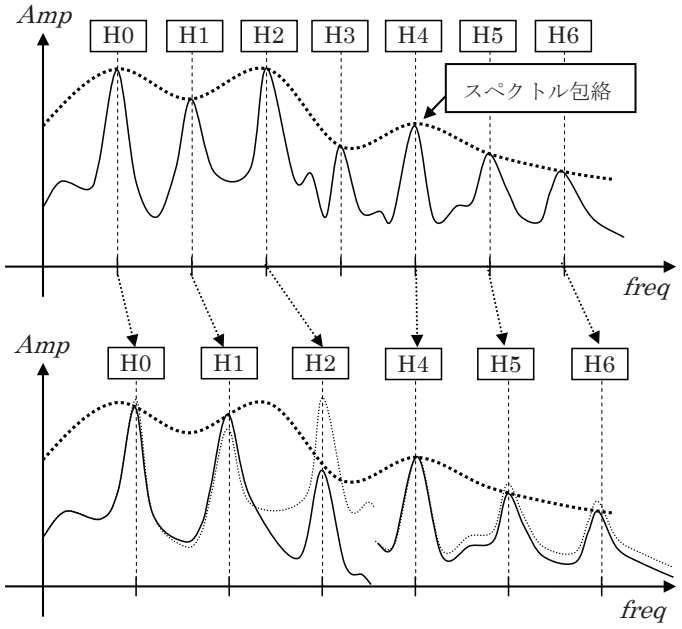

図5　ピッチ変換およびスペクトル包絡の調整

ピッチの変換およびスペクトル包絡の調整は周波数領域で行われます。ピッチ変換は、素片の波形をFFTした後、スペクトルを周波数軸上でスケーリングすることで行われます。（スペクトルの周波数軸上でのスケーリングはピッチを変えることになります。）スペクトルのスケーリングの際、倍音に相当するピーク近傍の微細構造はできるだけ元のものを保つように、非線形なスケーリングが行われます。スペクトル包絡の調整は、倍音に相当するスペクトルのピークの位置が所望のスペクトル包絡に合うように、スペクトルの強度を調整することによって行われます。周波数領域でのピッチ変換と音色の調整の後、IFFTとWindowing & Overlappingを行うことで合成音声が得られます。

歌手ライブラリは、実際の歌手の歌唱データから取り出した音声素片を集めたものです。素片はdiphoneと伸ばし音を使用しています。伸ばし音を素片として持っているところが歌声合成ならではの特徴です。対象となる言語で可能性のある全ての母音、子音の組み合わせと、母音および鼻音の伸ばし音が含まれます。素片用の歌声の録音では、効率的に素片が収集できるように考案された専用の歌詞を歌手に歌ってもらいます。声域によって声質も変化するので、収録は複数のピッチで行います。もちろん、収録するピッチの数が多ければ多いほど合成音のクオリティ向上が期待されますが、歌手への身体的、心理的な負担を考慮して、ある程度のところで妥協が必要となります。歌手のモチベーションを保ち安定した声を出していただくために、さまざまな現実的な工夫（収録する歌声のイメージを伝えるために挿絵を描いて見せる、茶菓によりご機嫌を取る等）も必要になります。収録されたデータは音素セグメンテーションおよび使用する領域のセグメンテーションを自動的に行い、人間の手によるチェックと修正を経て完成となります。

5. 歌声合成の今後

歌声合成は今後、品質がますます向上し、より広い分野で応用されるでしょう。音楽制作の現場では、プロ・アマを問わず、なくてはならないツールとなっていくことでしょう。音楽業界の現場では、すでに、仮歌（作曲家が歌手に歌い方を伝えるために制作する仮の歌）を作る際に、合成音声を使用することも増えていると聞いています。

技術的な面から言えば、特に歌い方や表情付けのモデリングが自然な歌声を作り出す上で重要になってくるでしょう。現状の合成エンジンでも人間の実際の歌声からピッチやダイナミクスを抽出してそのまま再現すると、人間の歌声か合成かわからないぐらいのクオリティの歌声が実現できます。歌声の表現、表情が人間らしい歌声の鍵だと思われます。そうなると、様々な楽曲スタイルに合った歌い方や、特定の歌手の歌い方や癖を自動的に再現できるようなモデルが期待されます。その一方で、クリエイタがさらに細かに自分の思い通りに歌い方や声質をコントロールし、調整できるようなツールや環境も必要になるでしょう。また、現在ではまだ再現が難しいタイプの歌声（例えば vocal fry やいわゆる「ダミ声」等）の合成のために、信号処理の手法の改良も続けられるでしょう。

技術的な発展とともに、単なる楽曲制作のツールとしてだけではなく、さまざまな分野での応用も考えられていくでしょう。Vocaloid の合成エンジンをサーバ上で動作させ、インタネットの接続があれば歌声合成の機能を利用できる NetVocaloid と呼ばれるサービスも実際に運用され、Web 上のプロモーションや携帯電話向けのサービスも行われています。今後は音楽制作にとどまらず、教育分野やエンタテイメント分野への応用も広がっていくことでしょう。

さて、合成された歌声は、これまでの音楽の鑑賞のありかたを変える可能性を秘めています。歌声合成による楽曲の愛好者（特に若い世代）には、クリエイタの思いを直接感じることができるから、ということを楽曲を聴く理由に挙げる場合もあると聞きます。実在の歌手による歌声と異なり、歌手の感情が介在する余地がなく、クリエイタが歌声の表現を自ら作っていくことができるからです。クリエイタにとっては、楽曲を通して自分の感情を（歌手というフィルタを通さずに）直接リスナーに伝えることができるツールを手に入れたと言えるのかもしれません。ただし、もちろんこれは実在の歌手を否定するものではありません。電気楽器、電子楽器、コンピュータ音楽が登場しても生楽器がなくならなかったように、歌声合成は生の歌声と共存しつつ、音楽に新たな可能性と選択肢を提供することになるでしょう。

参考文献

[1] Kelly et al., “Speech Synthesis” , Proceedings of the Fourth International Congress on Acoustics, pp.1-4 (1962).

[2] Klatt, “Software for Cascade/Parallel Formant Synthesizer,” J. of the Acoustical Society of America 67(3) pp.971-995 (1980).

[3] McAulay et al., “Speech Analysis/Synthesis Based on a Sinusoidal Representation” , IEEE Transactions on Acoustics, Speech and Signal Processing 24(4), pp.744-754 (1986)

[4] Macon et al., “A singing voice synthesis system based on sinusoidal modeling,” Proceedings of ICASSP 97, pp. 435-438 (1997)

[5] http://www.kaelabs.com/

[6] http://www.ntt.co.jp/news/news00/0009/000907.html

[7] H. Kenmochi and H. Ohshita, VOCALOID – commercial singing synthesizer based on sample concatenation, Proc. Interspeech, pp. 4009-4010. (2007).