～2013 年「音の匠」記念講演より～

歌声合成技術 VOCALOID™と新しい音楽

ヤマハ株式会社 事業開発部 yamaha+推進室 VOCALOID プロジェクト

剣持 秀紀

1. はじめに

ここ数年、歌声合成技術に注目が集まっています。ニコニコ動画や YouTube などの動画サイトには、合成された歌声による楽曲が数多く投稿されており、若い人を中心にそのような楽曲を楽しむ人々が増えています。ここでは、この音楽の新しい動きを支えている歌声合成技術 VOCALOID とその仕組みについて、歌声合成技術の歴史にも触れながら説明します。

2. 自己紹介～私の原点～

私は、昔から機械いじりや電気工作、あるいはコンピュータをいじることが好きで、中学校や高校の頃には「ラジオの製作」「I/O」「マイコン BASIC Magazine」などをよく読んでいました。中学の頃に PC-8001 というパソコンが発売されましたが、友達のお父さんがこれを持っていたので、友達の家に遊びに行ってはこれを触っていました。PC-8001 は Z80 互換の CPU を使っていたと思いますが、今でも Z80 の機械語で覚えている命令もあります。また、その後 PC-8001mkIISR というパソコンを家で買いましたが、そのパソコンには FM 音源が搭載されており、PLAY 文で MML を入力し簡単な音楽を演奏することが可能でした。このあたりが私の原点の一つです。

もう一つの原点は音楽です。小さいころピアノを習っていましたが、高校の部活でヴァイオリンを始め、大学でもオーケストラで弾いていました。とはいえ、私が入った大学のオーケストラはプロの音楽家になる人を何人も輩出した「名門」で、ついていくために必死に練習していました。しかし、その頃に培った音楽に対する知識や考え方は、間接的ではありますが今に生かされているのではないかと最近になって思えるようになってきました。

一方、大学院での修士論文のタイトルは「あけぼの衛星で観測された赤道域 ELF 波動の伝搬特性に関する研究」です。人工衛星で観測された自然発生の電波を信号処理で解析して到来方向を推測するという研究です。音ではなく電波の研究をしていたわけですが、その当時使っていた横軸が時間、縦軸が周波数という図は、音声の

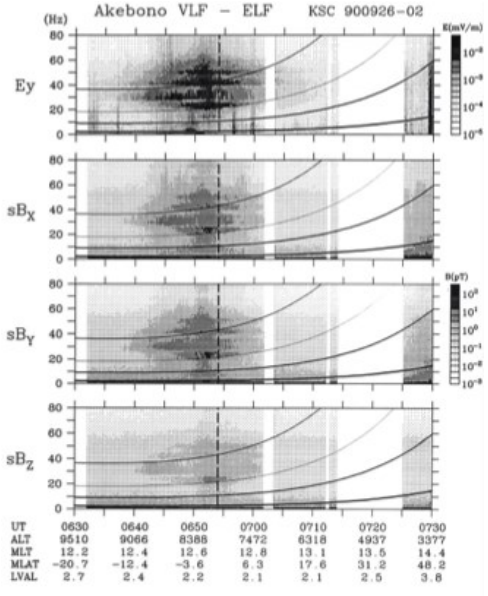

図 1 修士論文より

世界ではよく使われる図なので、こちらも間接的には今の仕事に役立っているといえると思います。このあたりが私の第 3 の原点と言えると思います。

そして、1993 年にヤマハに入社しました。入社当時はアクティブ・ノイズ・コントロールという、騒音に対して逆の波を出して軽減するという技術の研究開発をしていました。3 年後の 1996 年、音声合成や音声認識の技術を開発しているベルギーの L&H という会社とヤマハとの合弁会社に出向し、そこでいわば「音声屋」としての素養を身につけ、1999 年に復職し、それ以来 VOCALOID を含む、歌声や音声に関する技術開発に従事しています。仕事以外では、ヴァイオリンをアマチュアオーケストラや弦楽四重奏で演奏したり、アナログレコードを鑑賞するのが趣味です。カートリッジは DL-103、ターンテーブルはヤマハの GT-2000、そしてアンプは真空管のアンプ(C.E.C. Tube53)を使っています。仕事ではデジタルの世界ですが、趣味の世界では完全なアナログ人間です。

前置きはこの程度にして本題に入りたいと思います。

3. 歌声合成システム VOCALOID

VOCALOID とは、ヤマハが開発した歌声合成技術およびその応用ソフトウェアを表します。歌詞と音符を入力するだけで高品質な歌声を合成することができます。システムの構成は、図 2 のようになります。

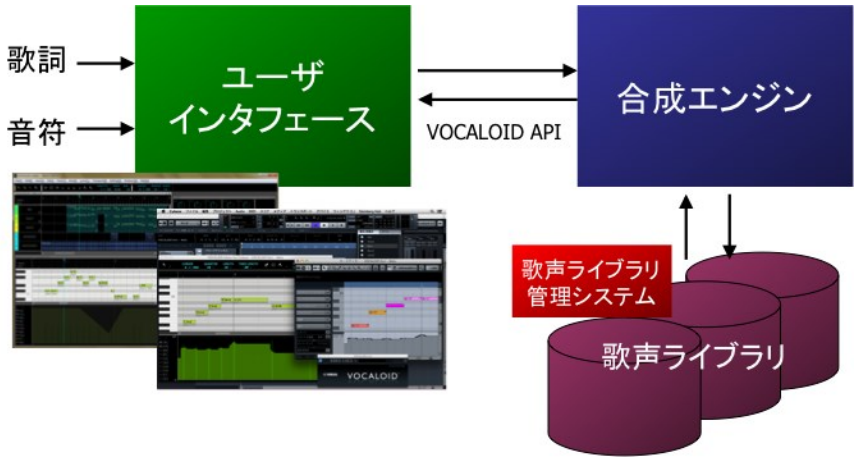

図 2　VOCALOID 構成図

歌詞と音符を入力するので、何らかのユーザインタフェースが必要になります。このために、歌詞と音符を関連づけられる形で入れられるようなインタフェースを持つ専用のエディタ(VOCALOID Editor)が用いられます。また、歌詞と音符を入力する機能を Cubase という DAW(Digital Audio Workstation)上に組み込み、伴奏で用いられる他の楽器と同等に歌声を扱うことができるようにした VOCALOID Editor for Cubase という製品も販売されています。このユーザインタフェースから「合成エンジン」に、音符と歌詞の情報を送り、合成エンジンは歌声を合成してその結果をユーザインタフェース側に返し、ユーザは歌声を聴くことができます。

合成エンジンでは、何もないところから歌声を作れれば良いのですが、まだそこまでは進んでいませんので、実際の人間の歌声から歌声の断片を集めたもの（歌声ライブラリと読んでいます）

から、必要な断片を取ってきて、それを加工してつなげることで歌声を作り出しています。

歌声ライブラリを作る技術や権利はヤマハからパートナー企業にライセンスしており、各社から様々な歌声ライブラリ製品が出ています。有名な「初音ミク」は、クリプトン・フューチャー・メディアさんが歌声ライブラリを開発し、製品化したものです。

ところで、この VOCALOID という名前ですが、Vocal に “-oid”という接尾辞をつけたものです。“-oid”というのは「～のような」という意味を作る接尾辞ですから、VOCALOID とは「Vocal のような」という意味になります。この名前は、「いつかは人間の声と区別がつかないくらい品質を高めたい」という願いと、「人間の声と同じではないことによる新しい表現を追求」という 2 つの意味が込められる良い名前だと思います。ヤマハが発表する前には世の中に存在していなかった名前なので、この名前はもちろんヤマハの登録商標になっています。

開発は 2000 年にスタートして、2003 年に新技術に関するプレス発表を行い、2004 年に最初のバージョンがリリースされました。2007 年には VOCALOID2 にバージョンアップし、これを用いたクリプトン・フューチャー・メディアさんの「初音ミク」が大ヒットし、これを用いた楽曲がニコニコ動画などの動画サイトに数多く投稿されるようになりました。2011 年には更にバージョンアップして VOCALOID3 となり、それを使った歌声ライブラリも数多く発売されています。数え方にもよりますが、最初のバージョンから VOCALOID3 まで合わせると、50 種類以上のものが発売されており、言語も日本語だけでなく、英語、中国語、韓国語、スペイン語に対応しています。

これらの歌声ライブラリとソフトウェアを用いて、多くの皆さんがオリジナル曲を作り、競い合うようにニコニコ動画などに投稿しています。そして、そのような楽曲を多くの若い人々が好んで聴いています。人気曲は大手レコード会社から CD として発売され、オリコン 1 位になったものも複数あります。カラオケでのランキングでも VOCALOID を使って作られた楽曲が上位に来ることもあります。アスキー総合研究所が 2012 年に行った調査によると、女子中学生・高校生の 54%はボーカロイドの曲が好きという結果が出ています。このように若い人々を中心に大きな音楽ムーブメントとなっています。

4. 歌声合成技術開発の背景

VOCALOID の開発を始めたのは 2000 年です。当時は音楽をシーケンサで「打ち込み」で行うのは当たり前になっていました。また音源は外部のハードウェア音源を用いるのが主流でしたが、一方で、コンピュータの中で演算により音源を実現する「ソフト音源」も発売され始めた頃です。いろいろな楽器が電子的に再現できるようになってきた中で、「歌声」だけはそういう世界とは無縁でした。歌声も「打ち込み」で制作できるようになれば、いろいろな可能性が広がると考え、開発を始めました。

もちろん、VOCALOID 以前にも歌声を合成する研究はいろいろなところで行われておりました。世界で初めてのコンピュータによる歌声は、1960 年代のベル研の Kelly らによる研究の成果です。“Daisy Bell”という歌を歌わせたものですが、今聴いても 1960 年代にこれだけの歌声を合成できていたことは驚きです。（インターネットで “daisy bell computer” などのキーワードで検索すると見つかります。）この歌声は、文化的にもさまざまな影響を与えました。スタンリー・キューブ

リック監督の映画「2001 年宇宙の旅」の最後の方で、人工知能 HAL 9000 がシャットダウンしていくところで「昔こんな歌を歌った」ということで、この歌を歌う場面があります。

その後も色々な研究機関で歌声を合成しようという試みが行われてきました。また、コンシューマ向け商品として発売されたものもあります。

歌声合成では、歌声の 2 つの性質（つまり、音声としての性質と楽器としての性質）の両方を考慮しなければなりません。音声としての性質としてまず考えられるのは、他の楽器に比べて圧倒的に音色のバリエーションが広いという点です。音韻による音色の違いは、めまぐるしく楽器が変わっていくことに相当するかもしれません。その他にも個人性による音色の違いもあります。また、発音機構を話し声と共用していることから、ピッチが急には変えられない（常にポルタメントがかかる）ということも音声としての特徴として挙げられます。一方、楽器としての性質とは韻律（音の高さの変化とタイミング）が、楽譜あるいはそれに相当するものによって支配されるということです。またビブラートなどの表現も楽器としての性質です。いずれにせよ、この音声としての性質と楽器としての性質の両方を考慮しなければならない点が歌声合成の難しい点です。

先人の業績に敬意を払いつつ、「音楽制作の現場で使っていただくこと」を目標に、新たに 2000 年から開発を始めたのが VOCALOID です。

5. VOCALOID の仕組み

VOCALOID は実際の歌手の歌声から取り出された声の「断片」（音声素片と呼びます）をつなぎ合わせることで歌声を合成しています。そして、その音声素片を集めたものを「歌声ライブラリ」と呼んでいます。歌声ライブラリに含まれる音声素片は、ある音素から次の音素への移り変わりの部分と、母音の伸ばし音です。例えば、「あさーー」という歌詞の歌（「あ」は短く「さ」は長い）を合成するためには、#-a, a-s, s-a, a（伸ばし音）、a# （#は無音を示す）という音声素片が必要となります。これをつなぎ合わせることで歌声を作り出します。

しかし、単に音声素片をペタペタとはりつけただけでは歌声になりません。音声素片の音の高さが、楽譜から要求される音の高さとは異なることと、音の高さを合わせたとしても素片と素片の間の音色の微妙な音色の違いがノイズとなって聞こえるからです。

VOCALOID では、以下のような方法で音色を調整して、素片と素片の境界での音色が急激に変化しないようにしています。

(a) 時間があまりない場合には、音色を合わせていく（音色をクロスフェード）。

(b) 時間が十分にある伸ばし音については、直前の音色（「あさーー」の場合だと s-a の最後の音色）を引き伸ばし、最後のところで次の音色（a-#の最初の音色）に徐々に変化させる。

この音色の調整の様子を図 3 に示します。

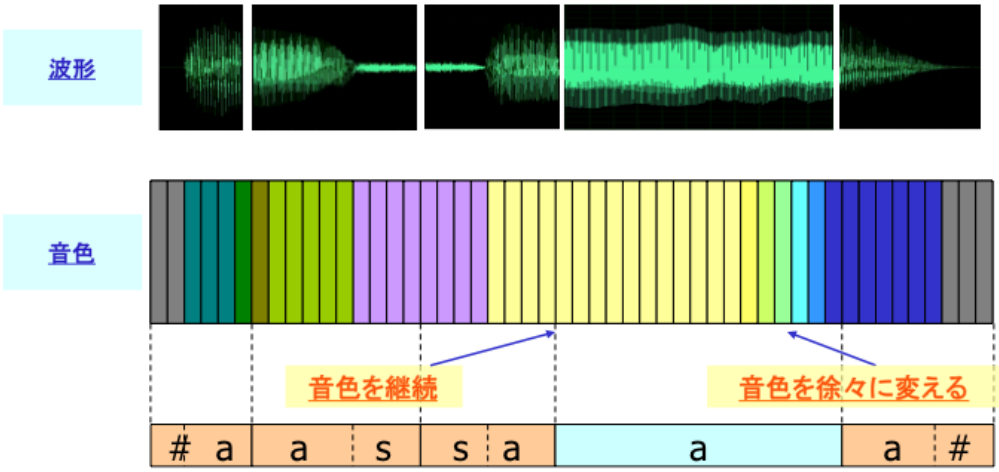

図 3　音色の調整

音の高さの調整と音色の調整は、周波数領域での信号処理によって行っています。波形をいったん FFT（高速フーリエ変換）した後、周波数軸上でスケーリングすることで音の高さを変えることができます。また、各倍音のレベルを上げ下げすることで音色を調整することができます。このようにして滑らかに音声素片をつないで歌声を作り出します。

また歌声では、タイミングも重要です。簡単のため「さ」という歌詞が４分音符で連続する場合を考えます。このときに、「さ」の発音開始を４分音符の頭のタイミングで行うと、どうしても遅れて聞こえてしまいます。これは人間が歌うときに、音節の中の母音の位置でタイミングを合わせているからです。つまり、合成する場合には、音節の母音の位置を音符のタイミングに合うように音声素片の位置を調整する必要があります。この様子を図 4 に示します。

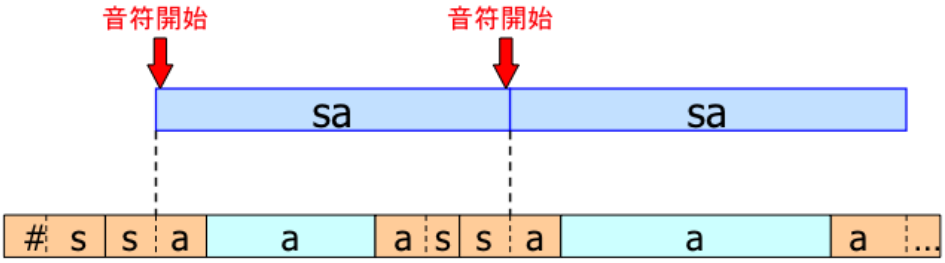

図 4 素片使用のタイミングの調整

実際の実装ではここが一番苦労した点です。ユーザインタフェース側から音符開始の指示を受けた合成エンジンは、その音符開始よりも前に発音を開始しなければならないという、因果律に反することを行わなければならないからです。現実的な解決法として、音符開始の指示を前もって送るということをしています。例えば「今から 500 ミリ秒後に『さ』という音節をもつ音符を発音しなさい」というという指示を合成エンジン側に送ると、合成エンジンはタイミングを合わせて、500 ミリ秒後には「さ」の中の「a」の発音が始まるように調整します。

さてここで、歌声ライブラリについても簡単に説明したいと思います。歌声ライブラリを作るには、歌手や声優の声を録音し、その中から音声素片を取り出すことが必要です。しかし、どんな歌詞が来ても合成できるようにするためには、対象となる言語で可能性がある全ての音素の組み合わせを効率よく収録する必要があり、そのために特別な歌詞を考案しています。その歌を収

録した後、データ処理の作業になりますが、波形を見ながら必要な音声素片をひたすら切りだしていく作業になります。ここの部分は合成品質のクオリティを左右する大事な作業ですので、おろそかにはできません。

6. 歌声合成と新しい音楽

最後に、なぜ若い人を中心に VOCALOID を使った楽曲が人気になっているのかを考えてみたいと思います。ここからは私の主観的な分析になりますが、私は、生身の人間の歌手が歌っていないことそのものがポイントだと考えています。音符と歌詞を入力するという作業は、オフラインの演奏行為そのものです。出来上がった歌声の「演奏者」は音符と歌詞を入力した人になります。しかし、キャラクタが与えられた場合は、そのキャラクタに「歌ってもらっている」ような錯覚があるのも事実です。その錯覚を受け入れることで、新しい世界観が広がり、新しい表現が生まれ、それを若い人の心をとらえたのではないでしょうか。VOCALOID で作られた楽曲を聴くと、特に歌詞に私ははっとします。今までの商業音楽にはない、商業音楽ではありえないような、粗削りではあるけれども斬新な歌詞をもつ楽曲が多いのです。ネットの発達により、アマチュアでも自分の楽曲を世界中に届けることができるようになりました。作った人が直接聴く人に届ける「産地直送」のようなところが魅力の一つなのかもしれません。

歴史を振り返ると音楽が変化するタイミングには社会の変化がありました。例えばモーツァルトの最後の交響曲である交響曲 41 番ハ長調 K.551 が作曲されたのは 1788 年、一方ロマン派の入り口であるベートーヴェンの交響曲第 3 番変ホ長調「英雄」作品 55 はそれからわずか 16 年後の 1804 年です。その間にはフランス革命という社会の変化がありました。音楽はこのように短期間に変化します。音楽が変化するもう一つの要素として、楽器の変化があります。ベートーヴェンはピアノソナタ 32 曲を作曲する中で、徐々に音域を広げてきました。これは何も初期に出し惜しみをしていたわけではなく、当時のフォルテピアノの音域がだんだんと拡大してきたことに対応していると言われています。このように楽器の変化は音楽そのものも変化させます。歌声合成という新しい道具（＝instrument すなわち楽器）も（ベートーヴェンのピアノソナタほどの大きな話ではないかもしれませんが）音楽を変化させていると言えると思います。

ネットの発達という社会の変化と、歌声合成という新しい楽器により、新しい音楽がこれからも生み出されていくことを願っています。

筆者プロフィール

剣持 秀紀（けんもち ひでき）
1967 年：静岡県生まれ
1993 年：京都大学大学院工学研究科修士課程修了、同年ヤマハ（株）入社
1996 年：ヤマハとベルギー企業との合弁会社 L&H ジャパン（株）に出向。
1999 年：ヤマハ（株）に復職
2000 年：VOCALOID 開発を開始。以降、VOCALOID を含む歌声、音声信号処理に関する研究開発を行う。