2 歌声合成の過去・現在・未来
「使える」歌声合成のためには

剣持秀紀（ヤマハ（株））

歌声合成に対する関心の高まり

近年，歌声合成に対する関心が高まっている．「ニコニコ動画」をはじめとする動画サイトでは，主にアマチュアのクリエータが，歌声合成ソフトウェアを用いて制作したオリジナル楽曲を文字通り日夜投稿している．アマチュアのクリエータがオリジナル楽曲を発表する道具として，クリプトン・フューチャー・メディア（株）の「初音ミク」を筆頭とする歌声合成ソフトウェア VOCALOID は必要不可欠なツールとなっているといっても過言ではない．また，主に若い世代を中心として，歌声合成ソフトウェアを用いて作成された楽曲を好んで聴く層も出現しており，音楽業界にとっても歌声合成技術は重要な存在となっている．本稿では，歌声合成技術の歴史を振り返りながら，筆者が開発にかかわった歌声合成技術 VOCALOID について紹介し，そして歌声合成技術の今後について，主に技術的な側面から論じる．

過去～歌声合成の歴史

■ 歌声合成の研究開発

世界で初めてコンピュータによって歌声を合成した例は，1962 年にベル研究所の Kelly らによって発表された “Daisy，daisy，…” という歌声であると言われている[1]．これは音響管モデル（Acoustic Tube Model）と呼ばれるものであり，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して歌声の生成を物理的にシミュレートしたものである．この合成音は今聞いても，1960 年代にこれだけのクオリティで歌声の合成が実現されていたことに驚きを禁じ得ない．こ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」でも最後のシーンで HAL9000 が停止する直前に ”Daisy，daisy，…” と歌う場面に影響を与えたと言われている．

それ以来，産業での応用はあまり行われないまま，いわば「細く長く」歌声合成に関する研究はさまざまな研究機関や企業によって行われてきた．実際に商用のシステムとして発売されたものもある．

しかしながら，「飛び道具」として商業音楽に使われる場合はあったにせよ，実際の音楽制作のシーンで，歌声合成技術が広く使われることはなかった．このことは，他の楽器の場合に，商業音楽でコンピュータを利用した演奏（いわゆる「打ち込み」）が広く行われるようになっていることと対照的であった．

■ 歌声合成技術の難しさ

歌声の合成が通常の楽器音と異なっている点は，歌声には歌詞があるという点である．つまり歌声には音声としての性質が伴う．歌詞があるということは，音符ごとに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，さまざま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相当することになる．通常音色が極端に変化しない楽器からのアプローチだけでは，歌声はうまく再現できない．また，歌詞があることはユーザインタフェースにも特別な考慮が必要である．歌詞をどのように入力するかということは，歌声合

成技術を考える上で避けては通れない．

また一方で，歌声には，音声という性質だけでなく楽音という性質もあるのも事実である．すなわち，韻律は楽譜（あるいはそれに相当するもの）によって支配され，また音自体の「美しさ」が審美の対象となる．過去の歌声合成技術では，話し声の合成の延長として考案されたものもあり，1990 年代に国内で市販されたテキスト音声合成ソフトウェアでは歌声機能が付いていたものも多い．しかし合成音自体，とりわけ伸ばし音の美しさが歌声では重要であることから，実際の音楽制作シーンで利用されることはほとんどなく，「おまけ機能」にとどまっていた．

歌声が音声としての性質と楽音としての性質を両方備えることが，歌声合成を難しくしている理由だと考えられる．

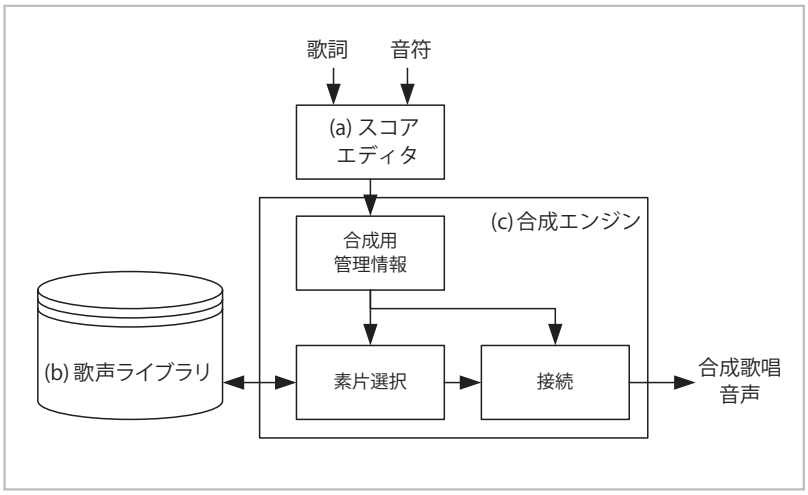

図 -1　VOCALOID の構成

現在～ VOCALOID 歌声合成システム

筆者らは，以上を踏まえ，実際に音楽制作シーンで利用されるために，歌声合成技術に要求される条件として，以下の 3 つを念頭に置き VOCALOID 歌声合成システムの開発を行ってきた[2]．

(1) **了解性**：歌詞が聞き取れること
(2) **自然性**：できるだけ人間の歌唱が持つ特性が再現されること．ブザー音的にならないこと．
(3) **操作性**：音符と歌詞を効率的に入力できること．伴奏を含めて楽曲制作を行いやすいこと．

もちろん，現在においてもそのすべてが達成されているとは言いがたい面もあるが，今後もこれらは重要な条件であると考えている．

VOCALOID 歌声合成システムの構成を図 -1 に示す．図 -1 に示されるように，(a) ユーザが歌詞や音符を入力するスコアエディタ（VOCALOID Editor），(b) 実際の歌声の録音をもとに取り出した音声素片の集まりである歌声ライブラリ，(c) 歌声ライブラリ中の音声素片を連結・変換する合成エンジンから構成される．

以下それぞれの要素について説明する．

■ (a) スコアエディタ（VOCALOID Editor）

スコアエディタの外観を図 -2 に示す．複数のトラックを効率的に管理するためのトラックエディタがあり，トラック上の歌声合成のパートをエディットするためにはミュージカルパートエディタを使用する．ミュージカルパートエディタは，歌声に必要な基本要素，すなわち歌詞と音符（音高，タイミング，長さ）を効率的かつ直感的に入力できるように，ピアノロール（横軸が時間，縦軸が音高の 2 次元平面上で，発音中の部分に着色する表示方法）を基本にしたユーザインタフェースとなっている．歌詞に関しては，日本語の場合は平仮名または片仮名，英語・スペイン語の場合は単語をそのまま入力し，内部で自動的に音声記号（発音を表記した記号）に変換される．韓国語の場合はハングル，中国語の場合はピンインを入力する．ユーザが音声記号をエディットして変更することも可能である．歌声に必要な歌い出しや音符間の歌い回し，伸ばし音中のビブラ

ートなども簡単に調整できるようになっている．その他合成音の声質も時変的にコントロールできる機能も備えている．

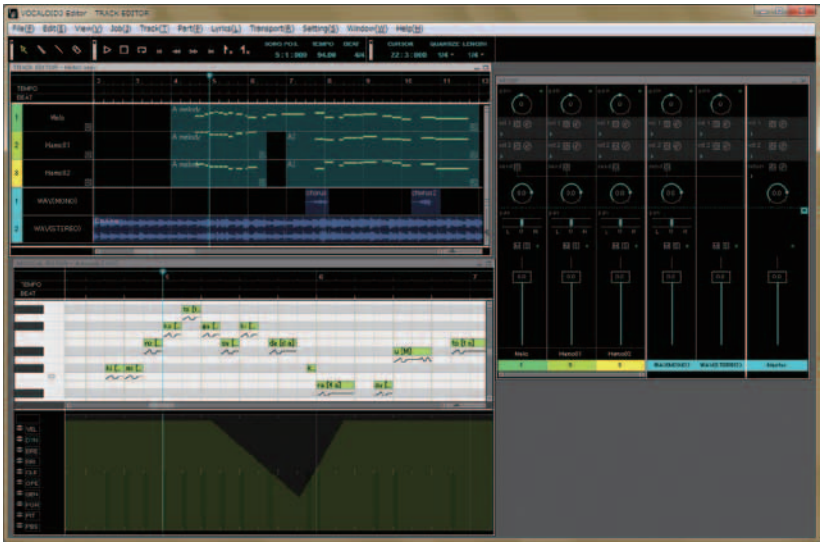

図 -2　スコアエディタ（VOCALOID Editor）

■（b）歌声ライブラリ

歌声ライブラリは，ある音素から別の音素の変化部分と母音の伸ばし音が音声素片として含まれる．これに加えて，2011 年発売の VOCALOID3 からは，3 音素の連鎖も含めることが可能となった．これにより，たとえば母音間に挟まれた場合に変化しやすい子音（たとえばハ行の子音など）がより自然に再現できるようになっている．

歌手の歌声の音色は音域によって異なるため，同じ音素の組合せの素片であっても複数の異なるピッチについてそれぞれ素片を持つ．

歌声ライブラリ制作にあたっては，実際の歌手に特別な歌詞を歌ってもらい，その録音から必要な部分を切り出して登録する．ライブラリ制作はある程度自動化されているが，最終的には人間の耳と目による調整が必要である．

■（c）合成エンジン

合成音声は歌声ライブラリから必要な素片を選択し，接続する．たとえば「あさ」（音声記号で [asa]）という歌詞を合成するためには，#-a，a，a-s，s-a，a，a-#（# は無音）という素片を接続する．もちろん単純に素片を接続しただけでは歌にならない．ユーザが指定した音符の音高（ピッチ），音符のタイミングや長さに合うように素片を加工して接続する必要がある．

タイミングに関しては，歌声特有の制御が必要になる．音符を構成する音節の，開始部分（最初の子音の開始部分）を音符の開始時刻に合わせたのでは，人間の耳には遅れて聞こえる場合が多い．特に子音が無声摩擦音等の継続時間が長い（数 10 ミリ秒以上の）場合にはその遅れが顕著である．人間は無意識に音節内の母音の位置でタイミング合わせを行っているからである．そのため素片の母音部分の開始位置が音符の開始時刻に合うように，素片の使用タイミングを早める調整を行っている．

ピッチに関しては，ユーザの指定した音符や表情に合うように内部的にピッチ曲線が描かれる．このピッチ曲線で指定されるピッチに合うように素片のピッチ変換を行う．ピッチ変換は波形を高速フーリエ変換（FFT）し，周波数軸上でスケーリングすることで行う．

ピッチ変換を行っただけでは，素片間に音色の違いがあるため，そのまま接続すると音色が突然変化するためにノイズとなる．そこで伸ばし音の間で音色の補間を行うことで音色の突然の変化を避けるようにしている．「あさ」の最初の「あ」という音節の場合で言えば，#-a の最後の [a] の音色のスペクトル包絡と a-s の最初の [a] の音色のスペクトル包絡を補間することで a の伸ばし音区間のスペクトル包絡を作る．#-a および a-s の区間では素片が持つスペクトル包絡をそのまま用いる．このようにして求めたスペクトル包絡に沿うように，ピーク近傍の強度の調整を行い，最終的に逆 FFT を行い，時間領域の波形が得られる．

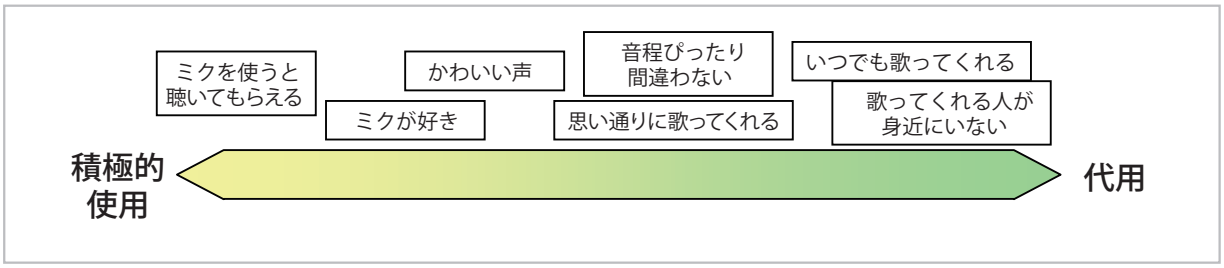

図 -3　歌声合成を使う理由

歌声合成技術の未来

さてここで，なぜ人々が歌声合成を利用するかについて考えてみたい．

このことは，メトロノームと電子メトロノームとの関係で考えると分かりやすい． 電子メトロノームが最初に発売された当時は，「視認性が悪い」「聞こえにくい」などの悪評もあったが，それらは技術の進歩により克服され，今では持ち運びやすく，正確であるというメリットにより，また三連符や変拍子などの機械式メトロノームでは実現不可能な機能も実現され，楽器の練習ではごく普通に使われるようになってきている．これと同様に歌声合成も，人間の歌声の単なる代用（**図 -3 の右側**）ではなく，歌声合成の積極的な利用（**図 -3 の左側**），すなわち歌声合成でなければできないことも目指していかなければならないと考えられる．最近の，いわゆる VOCALOID 楽曲を聴くと，いままでのポップスではあり得なかったような独創的で新鮮な表現の歌詞が現れる楽曲も多い．普通であれば恥ずかしくて歌えないような歌詞の曲であっても，歌声合成によっていったん歌として存在してしまうと，それは新しい表現として受け入れられるようになっていく．このような新しい表現の手段として用いられるということも重要な「積極的な利用」であろう．

このように歌声合成技術の重要性は今後もますます高まっていくであろう．ここでは今後の歌声合成技術について，合成音のバリエーション拡大，利用場面の拡大，ユーザ層拡大 の 3 つの視点から考えてみたい．

■ 合成音のバリエーション拡大

今後も合成音そのものの品質をさらに向上し，実際の人間の声にさらに近づけていく必要があるだろう．合成音には合成音なりの良さがあり，現状での合成音を好む人々がいるのも事実ではあるが，人間の声に近づけていくことは，合成技術全般（CG なども含む）の持つ宿命とも言えよう．

現状の VOCALOID では，いわゆる「ダミ声」などのピッチがきれいに抽出できない声の再現ができない．歌声をより人間に近づけるためには，この種の音声が再現できるようになることも重要である．また，たとえば中野・後藤による VocaListener[3] が示すように，実際の人間の歌声が持つピッチやダイナミクスなどの韻律を抽出し，合成音で再現すると，人間の声と区別できないほどの合成音が得られることから，歌声の表情（すなわち韻律）をいかに自然に作り出せるかということも今後の課題の 1 つであるといえよう．現状では，選択する歌声ライブラリにかかわらず，同一のピッチモデル，ダイナミクスモデルにより韻律が決められるが，たとえば統計的な手法により特定の歌手の歌いまわしを再現するような手法[4] も今後重要となっていくことであろう．

合成音のバリエーションといえば，利用できる言語を増やしていくことも重要である．2012 年 1 月現在，VOCALOID 合成エンジンは日本語，英語，スペイン語，韓国語に対応しているが（中国語は対応中），今後も対応言語を増やしていきたい．

さらに，VOCALOID のユーザの中には，歌声合成エンジンを利用して無理やり話し声を合成し，そ

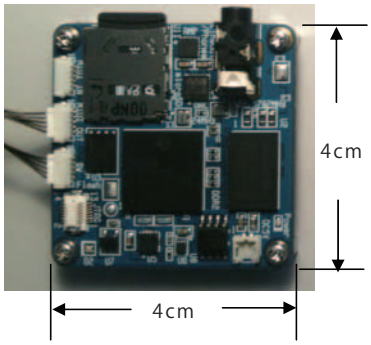

図 -4
VOCALOID-board

れを用いたコンテンツを動画サイトに投稿する人々も生まれてきている．これは既存のテキスト音声合成システムでは不可能な，感情をこめた話し声を手作業で作り出そうという試みである．実際我々の日常会話について内省してみると，歌声に近い韻律を持つものも少なくない（たとえば「行ってきまーす」等）．このような朗読音声ではない話し声を合成できる技術もコンテンツ制作に必要となってくると考えられる．

■ 利用場面の拡大

現状で歌声合成技術が使用されるのは，いわゆる「打ち込み」による音楽コンテンツ制作に限られている．今後は，ライブやコンサートでの使用も求められるようになっていくだろう．そのためには，リアルタイムに，いわば楽器を演奏するように，歌声合成エンジンをコントロールできるようなユーザインタフェース，中でも特に歌詞を入力するインタフェースが必要となってくると思われる．

パソコン用のアプリケーションソフトウェアだとどうしても利用場面が限られてくることから，さまざまなハードウェア環境への移植も必要となってくるであろう．筆者らは VOCALOID 合成エンジンを DSP（Digital Signal Processor）上に移植し，4cm × 4cm の小型の専用ボード（図 -4）上で動作させている（VOCALOID-board）[5]．また，iOS 等への VOCALOID 合成エンジンの移植と商品開発も行っている．

■ ユーザ層の拡大

VOCALOID を含む歌声合成ソフトウェアのユーザは，動画コンテンツ向けに楽曲制作をする人々が中心である．しかしながら，音楽を供給する側のツールとしてだけ用いることは，歌声合成技術の可能性を半減させてしまっているともいえる．たとえば（著作権者の許諾を得た上で），替え歌を楽しんだり，メロディを変えて楽しむなどのカジュアルな使い方も歌声合成技術の応用の 1 つであろう．そのためには，「楽しむ」ために適したインタフェースも必要となる．いずれにしても，生の歌声ではできない新しい楽しみ方が歌声合成技術によって実現されていくことになるであろう．

参考文献
1）Lochbaum, K. : Speech Synthesis, Proc. of the Fourth International Congress on Acoustics, pp.1-4（1962）.
2）剣持秀紀，大下隼人：歌声合成システム VOCALOID －現状と課題，情報処理学会研究報告，2008-MUS-74-9, 12, pp.51-58（2008）.
3）中野倫靖，後藤真孝 : VocaListener：ユーザ歌唱の音高および音量を真似る歌声合成システム，情報処理学会論文誌，Vol.52, No.12, pp.3853-3867（2011）.
4）Saino, Tachibana and Kenmochi : A Singing Style Modeling System for Singing Voice Synthesizers, Proc. of INTERSPEECH-2010, pp.2894-2897（2010）.
5）剣持秀紀，吉岡靖雄 : 歌声合成技術 VOCALOID とその組み込み機器への応用可能性，人工知能学会研究会資料，SIG-Challenge-B002-1（2010）.

（2012 年 1 月 21 日受付）

剣持秀紀 ▌kenmochi@beat.yamaha.co.jp
1993 年京都大学大学院工学研究科電気工学第二専攻修士課程修了，同年ヤマハ（株）入社．入社以来音声信号処理の研究開発に従事．