Vol.2010-MUS-84 No.14
2010/2/16

歌うロボット
ーVOCALOID とサイバネティックヒューマン HRP-4C のコラボレーションー

橘　誠[†]　中岡慎一郎[††]　剣持秀紀[†]

本稿では，2009 年 10 月に開催された CEATEC JAPAN 2009 にて，ヤマハブースに出展した歌声合成システム VOCALOID と産業技術総合研究所が開発したサイバネティックヒューマン HRP-4C“未夢”のコラボレーションによる歌うロボットの演出について，その概要とデモシステムの技術を紹介する．

A Singing Robot:
Collaboration between VOCALOID and Cybernetic Human HRP-4C

Makoto Tachibana[†], Shin'ichiro Nakaoka[††]
and Hideki Kenmochi[†]

This paper describes the technique applied to a robot to perform singing-voice as exhibited at CEATEC JAPAN 2009. To achieve a realistically robot-singing performance, facial motions such as lip-sync and facial gestures are required. This work is a result of a collaboration between the technology “VOCALOID” (developed by YAMAHA) and the cybernetic Human HRP-4C named “Miim” (developed by AIST). We report the technical overview of the system developed for the mentioned exhibition.

1. はじめに

2009 年 10 月 6 日～10 日に千葉･幕張メッセで開催されたアジア最大級の最先端 IT･エレクトロニクス総合展である CEATEC（Combined Exhibition of Advanced Technologies）JAPAN 2009 において，我々は経済産業省主催「ライフコンテンツフロンティア」内のヤマハブースにおいて歌声合成システム VOCALOID とサイバネティックヒューマン HRP-4C のコラボレーションによる歌うロボットの展示をおこなった．

「VOCALOID」はヤマハが開発した歌詞とメロディーを入力することで歌声音声を合成する技術，およびそれを応用したアプリケーションソフトウェアの総称である[1]．実際の人の歌声から収録したデータより作成された「歌手ライブラリ」を用いて「合成エンジン」にて合成を行うことで，リアルな歌声を合成することが可能であり，これまでに，クリプトン･フューチャー・メディア（株）[2]の「初音ミク」に代表されるキャラクターボーカルシリーズ，（株）インターネット[3]の「がくっぽいど」に代表されるアーティストボーカルシリーズなどが発売されており，動画投稿サイトなどネット上を中心にVOCALOID使用した楽曲が数多く発表され，世界中で幅広く利用されている．

また，「サイバネティックヒューマンHRP-4C」は，産業総合技術研究所（産総研）が開発した日本人青年女性の平均的体型を参考にして設計した身長 158cm，体重 43kg の人間に非常に近い外観を有する二足歩行ヒューマノイドロボットである[4][5]．HRP-4Cの大きな特徴として，人間に近い全身のプロポーションと，二足歩行を含む全身動作，およびリアルな顔の表情変化を組み合わせることが可能であることが挙げられる．これにより，従来のヒューマノイドロボットでは実現できなかったエンタテイメント分野への応用が期待されており，これまでにも「第８回東京発日本ファッション・ウィーク」のファッションショーのひとつにて司会を務めるといった活躍が注目されてきた[6]．

歌を歌うことのできるロボットとしては，NECのPaPeRo[7]，（株）ビジネスデザイン研究所のifbot[8]，SONYのQRIO[9]などが挙げられるが，任意の歌声を合成可能なVOCALOIDとリアルな表情が表現できるHRP-4Cのコラボレーションにより，歌うサイバネティックヒューマンを実現することを目的とした初の試みとなった．

リアルに歌うことを実現するためには．顔の動作，表情，リップシンクなどの動きも必要となる．そこで，本稿では CEATEC JAPAN 2009 での VOCALOID と HRP-4C のコラボレーションを実現したシステムについての技術的な詳細を紹介する．

† ヤマハ株式会社　サウンドテクノロジー開発センター
Center for Advanced Sound Technologies, Yamaha Corp.
†† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST),

Vol.2010-MUS-84 No.14
2010/2/16


声：CV-4C β
©Crypton Future Media Inc.


初音ミク
©Crypton Future Media Inc.

Megpoid
© INTERNET Co. Ltd.

図 1 演出に使用した衣装と歌声 DB

2. デモンストレーション

2.1 デモの概要

CEATEC開催期間の 10 月 6 日～10 日の間，ライフコンテンツフロンティア内のヤマハブースでは１日に４回，“**PLAY iT**”と題して「時を超える」，「空間を超える」，「演奏の枠を超える」，「リアリティーを超える」，「ソウゾウを超える」技術を紹介する 30 分程度のデモンストレーションを行なった．その中の「ソウゾウを超える」として行なわれたのが，このVOCALOIDとロボットのコラボレーションデモである．デモでは，10 分弱の間に 2～4 曲の歌を披露した．また，ステージによって通常のHRP-4Cのスタイルだけでなく，キャラクターに合わせた衣装を着け，声も変えて登場するという演出を行なった．通常のHRP-4C「未夢(ミーム)」の声はクリプトン･フューチャー・メディア（株）が提供した「CV-4C(β)」が担当した．またクリプトン･フューチャー・メディア（株）の「初音ミク」および（株）インターネットの「Megpoid」の歌声ライブラリを使用した場合には，それぞれ「初音ミク」，「GUMI」の衣装を着けて登場した．図 1にそれぞれの衣装と歌声DBを示す．

2.2 デモの流れ

デモの具体的な流れは以下のとおりである．

(1) **曲のリクエストと，リクエストをされた方の名前を受け取る**
(2) **リクエスト者へのお礼と曲タイトルを紹介するしゃべりを披露する**
(3) **自動演奏ピアノ「Disklavier™（ディスクラビア）E3」[10]の伴奏に合わせた歌を披露する**

まず，(1)では，ブースに来ていただいたお客様からのリクエストを受け付ける．曲候補はブース内のメインディスプレイに表示し，その中からリクエスト曲を受け付けた．図 2 にリクエスト時にメインディスプレイに表示したデモ曲一覧を示す．リクエストの候補曲は延べ 12 曲あるが，各キャラクターに向いている曲をそれぞれ 5～6 曲選んだ組合せを用意し，ステージ毎に登場するキャラクターに合わせてパターンを変えて使用した．リクエストの受け付けには，セカイカメラ[a]のエアシャウト機能[b]を利用した．また，電波状況などによって，セカイカメラから送信されたエアシャウトを受け取れなかった場合には，ブース内のお客様から直接リクエスト曲を受け付けた．受け取った情報は，VOCALOIDシステムの担当者が，リクエスト者の名前のアクセント記号付き「かな」とリクエスト曲のID番号をiPodTouch端末上のWebフォームから入力し，VOCALOIDシステムに渡す．ここで，アクセント付き「かな」とは，名前のアクセント核（高さアクセントが高から低へ変化する箇所）の位置を「’」で指定した「かな」文字で，「佐藤」は「さ’とう」，「高橋」は「たか’はし」といった入力になる．

未夢（パターン 1）

未夢（パターン２）

Megpoid

初音ミク

図 2 使用したデモ曲

(2)では，リクエスト曲とリクエスト者の名前に応じたしゃべりを披露する．例えば，リクエスト者が「佐藤」さんで，リクエスト曲が「ワールドイズマイン」の場合には，

「佐藤さん，リクエストありがとうー．ワールドイズマインを歌います」
「佐藤さんからリクエストいただきました．ありがとうございます．
それでは歌いまーす，ワールドイズマイン」

a) 頓智・（トンチドット）が開発した iPhone 上で動作する拡張現実テクノロジーで iPhone 上のデジタルカメラ上に実際の景色と「エアタグ」というテキストや画像などを重ねて表示する．
b) エアタグを周囲のセカイカメラユーザーに向かって飛ばす機能．

といったしゃべりを行なう．中には，
「佐藤はん，リクエストおおきにですー」
といったバリエーションもあり，これらはランダムに選択される．
そして，(3)では，自動演奏ピアノ「ディスクラビア」の伴奏に合わせて，歌を披露する．歌が終わったら(1)に戻り，次のリクエストを待つ．

3. VOCALOID合成システム

図 3 に今回のデモ用に新たに開発した VOCALOID 合成システムのデータ処理の流れを示す．このシステムでは，VOCALOID の合成エンジンだけでなく，同期演奏に必要な MIDI の送信機能, HRP-4C の動きを制御させるための姿勢制御データの送信機能などを備えており，それぞれに異なったプロトコルでの通信・制御を行なうことが可能である．また，歌声やしゃべりなどは，あらかじめ作成された WAV ファイルではなく，合成シーケンスファイルあるいはその場で生成されたシーケンスから VOCALOID 合成エンジンを用いてその場で合成している．

システムで処理するデータは，曲 ID などからテンプレート処理される部分と，アクセント付き「かな」文字に従って動的に生成される部分の二つに分けられる．テンプレート処理部では，曲 ID に従って，必要な合成シーケンスファイル，コーラスパートなどの WAV ファイルとディスクラビアの伴奏用の SMF，HRP-4C の動作を指示する姿勢制御ファイルを呼び出し，タイミングを合わせてそれぞれを送信する．VOCALOID の合成シーケンスファイルは，歌だけでなく，しゃべりの一部にも使用している．しゃべりの合成シーケンスファイルは，実際の発話を録音した音声データから合成に必要な韻律情報（声の高さと各音素の継続時間）を求め，それに基づいて作成した．姿勢制御ファイルは，HRP-4C の「キーポーズシーケンス」機能を利用し，HRP-4C のコントローラ上にあらかじめキーポーズを定義しておき，どの時刻にどの姿勢になるかを，時刻とキーポーズから成る CSV ファイルを用いて指定している．本システムと HRP-4C のコントローラは LAN で繋がっており, コントローラへのリモートコールによってまず姿勢制御データを 1 曲分全て送信した後，動作を開始するコマンドを他のデータとタイミングを合わせて送信することによって同期を行なっている．また, ディスクラビアへは MIDI インタフェースを介して直接繋がっており, 逐次 MIDI 情報を送信し制御している．

動的生成部では，アクセント付き「かな」入力から，まず VOCALOID の合成エンジンで音声を生成するために必要な韻律情報を生成し，得られた各音素の継続時間情報に基づきリップシンクのタイミングを生成する．そしてタイミングを合わせて VOCALOID 合成エンジンの駆動と HRP-4C へのリップシンク情報の送信を行なう．なお，韻律情報によって，テンプレート情処理部の情報も更新される．例えば，「○○さん，リクエストありがとー」という台詞の場合，テンプレート処理される「リクエストありがとー」のタイミングは，リクエスト者の名前（「佐藤さん」，「高橋さん」）の長さによって異なってくるため，それを考慮した処理を行なっている．

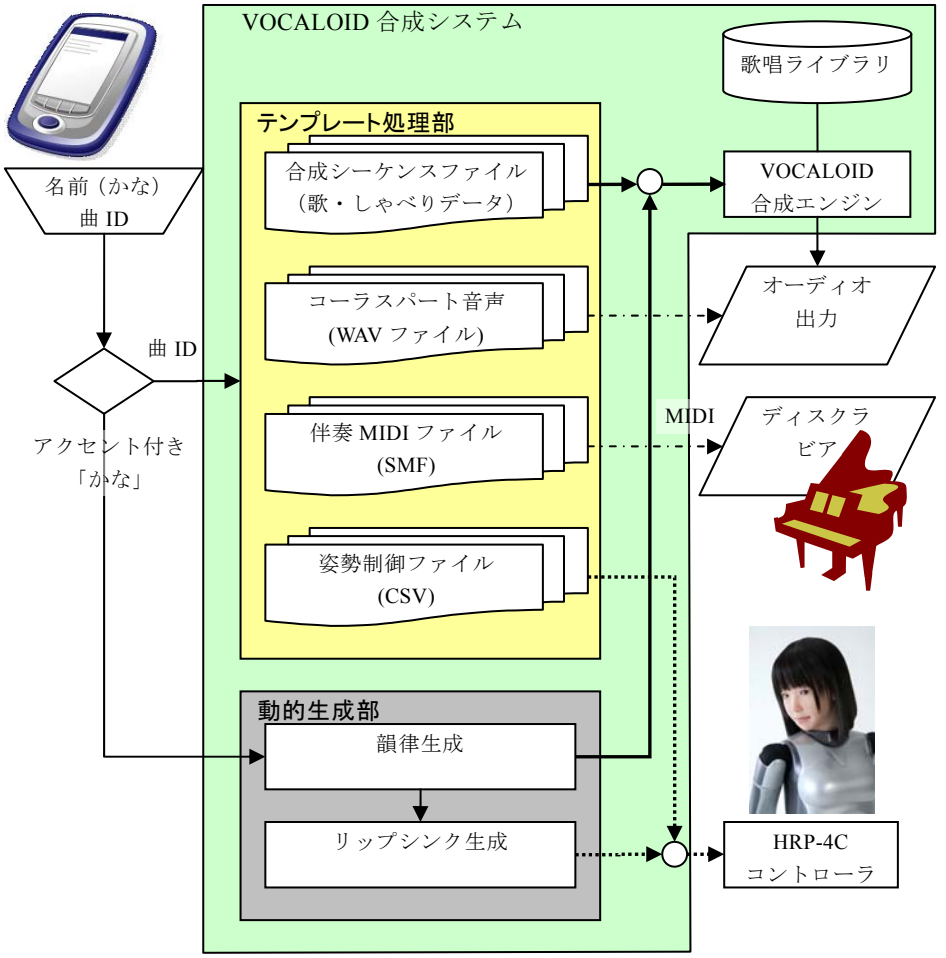


図 3 VOCALOID 合成システム

Vol.2010-MUS-84 No.14
2010/2/16

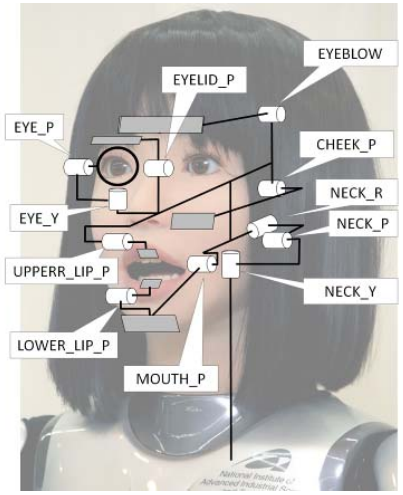


図 4 HRP-4C の頭部機構

4. 表現豊かに歌うための動作生成

人間と同じように，身振り・手振りを交えて歌い上げるロボットが理想的ではあるが，今回のデモでは，耐久時間や安全性を考慮し，首から上の動作を歌と合わせて作成した．

4.1 HRP-4Cの頭部機構

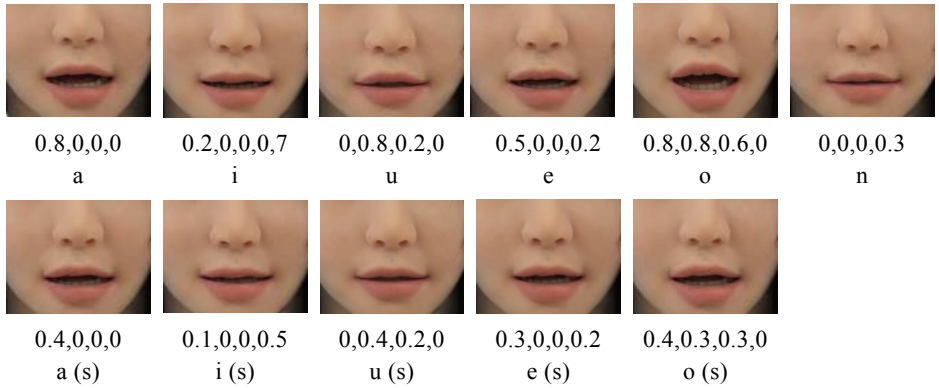
HRP-4C の頭部機構を図 4 に示す[6]. HRP-4Cは首の動作に 3 自由度と顔の各部位の動作に 8 自由度の関節を備えており，これらの関節を用いて頭部全体の姿勢や顔の表情を変化させることが可能である．顔の各部位の動作としては，眉の上下，瞼の開閉，頬の上下，上唇，下唇，顎の開閉にそれぞれ 1 軸が与えられており，眼球についてはその方向を変える 2 軸が与えられている．ただし，眉，瞼，眼球，頬といった左右のペアをもつ部位は，それぞれのペアが同一の関節によって駆動される設計となっており，ウィンクのように左右で独立した動作を行うことはできない．

4.2 キーポーズによる制御

HRP-4C では「キーポーズシーケンス」を扱うインタフェースを備えており，あらかじめいくつかの顔の状態をキーポーズとして作成しておき，それらを時間軸上に並べることで一連の動作を制御することが可能である．最終的な関節角軌道はキーポーズ間を補間し，HRP-4C コントローラが自動的に生成する．これは「キーフレームアニメーション」と基本的には同様の方式である. さらにキーポーズ列の階層化と共有化，および部分的な顔部位に対するキーポーズの定義とそれらの重ね合わせを導入することで，同じ表情が何回も現れたり部分的に反復的な動作を行ったりすることが多い顔の動作に対して，効率的に制御可能となっている．今回のデモにおいては，キーポーズとそれに対応する各機構の位置情報は全てのデモ中の動作において共有して使用している．また，リップシンク，目蓋，首の 3 つの階層分けて，それぞれの制御部位を定義し，それらを重ね合わせることで頭部全体の動きを作り出している．

4.3 リップシンク

唇の動き（リップシンク）を制御するために，5 母音の形状と口を閉じた状態の 6 つのキーポーズを作成した．また，5 母音に対しては，口の開きを小さくした形状もそれぞれ用意し，延べ 11 種類の形状を使用した．形状は口の開き方（MOUTH_P），上



数字はそれぞれ MOUTH_P, UPPER_LIP_P, LOWER_LIP_P, CHEEK_P の最大移動量に対する割合

図 5 作成した口の形状

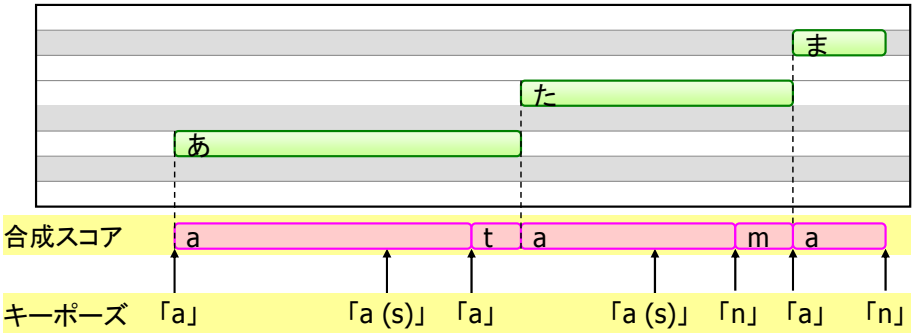


図 6 リップシンクのキーポーズ生成例

下唇の突出し具合（UPPER_LIP_P, LOWER_LIP_P），頬の位置（CHEEK_P）を使用して変化させている．図 5に作成した口の形状を示す．「(s)」で示したものは，それぞれの母音の口の開き方を少し小さくして登録したものである．図中の数字は，MOUTH_P, UPPER_LIP_P, LOWER_LIP_P, CHEEK_Pの最大移動量に対する各キーポーズでの変化の割合となっている．なお各関節の稼動範囲角はMOUTH_P: 0.0 ～ 10.0 UPPERLIP_P: -25.0 ～ 0.0，LOWERLIP_P: 0.0 ～ 25.0，CHEEK_P: -3.3 ～ 0.0（単位はdegree）となっている．これらを用いてリップシンクを生成するため，VOCALOIDの合成シーケンスファイルから各音素の発音されるタイミングを抽出し，唇形状のキーポーズシーケンスの作成を行なった．VOCALOIDの合成エンジンでは，音節の母音の開始部分が音符開始のタイミングに合うように調整した「合成スコア」を持ってお

り[1]，この情報を利用して音符を構成する音節の最初の音素の開始時刻に，その音節の母音の形状を出力することとした．なお，「ま，ば，ぱ」などの子音（唇音）を発音する際には，「n」のラベルを使用して，子音の区間で口を閉じ，母音の区間になってから開くようにした．さらに，歌声で音を伸ばしたときの表現を自然に歌っているように見せるため，伸ばし音の後に同じ母音の音節が続くときには，少し口の開きの小さいものを次の母音の手前に挿入し，口を少し小さくしてからまた大きく開く動作を表現した．図 6 に歌のスコアとリップシンクのキーポーズ生成の関係例を示す．

4.4 目蓋動作の生成

目蓋の動作として最初に挙げられるのが，瞬きである．HRP-4C のコントローラには瞬き生成機能が実装されており，瞬きの速度や最大時間間隔などをパラメータで指定すれば，あとはロボット内部の制御プログラム側で最大時間間隔を越えないランダムな時間間隔で自動的に瞬きが生成される．今回のデモにおいても瞬きの生成はランダムに行なっている．

さらに，歌を歌っている表情をより自然に表現するため，目を閉じる動作を姿勢制御ファイル中で指定した．目を閉じる動作は，歌の音符と対応付け，その音で目を閉じている確率（P(eyelid = close)）が，音符長と，母音の種類に依存して決定されると仮定した．具体的には，音符長を BPM（Beats Per Minute）で正規化した長さ T が設定した閾値（threshold）を超える場合に，その音符で目を閉じる確率が

$$P(\text{eyelid} = \text{close}) = \begin{cases} 1.0 & (\text{vowel\_type} = \text{"i"}, T \geq \text{threshold}) \\ 0 & \text{else} \end{cases}$$

で与えられるとした．また threshold の値を 1～1.5 倍の間でランダムに変化させることで，似たようなフレーズが続いた場合などに，常に同じ箇所で目を閉じてしまうことが起こらないようにした． 実際の人間の歌う表情に近づけるには，音高や前後のフレーズの流れなども考慮した複雑なモデルを作成する必要があると考えられるが，今回の非常にシンプルな制御でも，ある程度自然な動作を生成することができた．

なお，歌い終わった後にお辞儀をする動作に合わせて目を閉じる，曲に合わせて前奏や後奏の間は目を閉じるといった動作命令も手動で作成し，付与している．

4.5 首の動作の生成

歌に合わせた首の動作は自然で表情豊かに歌い上げるためには欠かせない動作である．デモでは，延べ 12 曲（未使用曲を含めると 17 曲）の首の振り付けを用意した．これらの首の動作を短期間で効率的に作成するために，簡単な首の状態遷移モデルを仮定し，首の動作を半自動で作成した．

首のキーポーズとして，首の 3 自由度を考慮して，上下 3 段階，左右 3 段階，回転 3 段階を組合せた 27 種類を作成し，このキーポーズを繋ぎ合わせることで，首の動きを生成した．動作を生成する手法として，文献 [11][12]などのように，音響特徴から

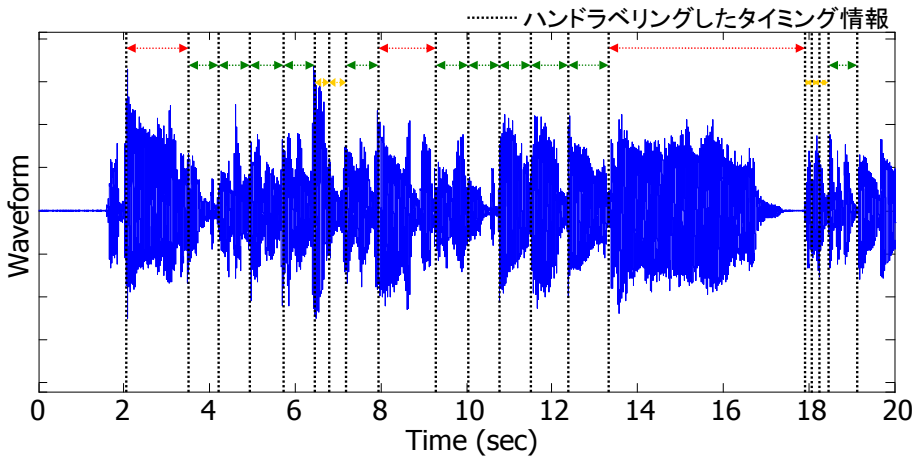


図 7 ハンドラベリングしたタイミング情報と首動作の関係

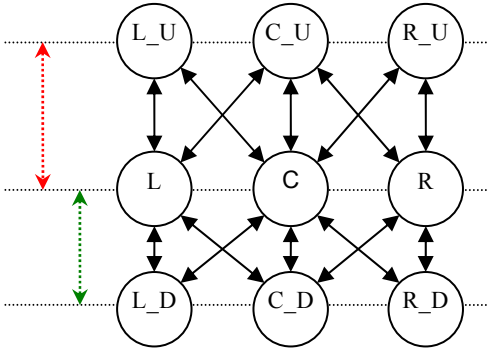


図 8 上下左右方向の首動作の状態遷移

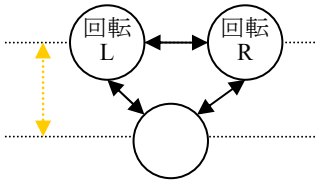


図 9 回転動作の状態遷移

自動的に作成する手法も提案されているが，今回は，ある程度，意図的に動きを制御することによって，曲中の「キメ」などのタイミングで，適切な動作が行なえることを目的とし，手動で指定したタイミング情報から，首の動きを半自動生成する方法を用いた．具体的には，動作を行なうポイントを，曲を聴きながら手動で指定し，そのポイント間の長さを 3 種類に分類し，それぞれで異なる動きを生成するようにした．図 7 に楽曲へのハンドラベリングと動作の関係を示す．まず，最も基本的な動作として，曲の拍に合わせて首を上下に振る動作を，ポイント間の長さが曲の 1 拍とほぼ同じくらいの長さの場合（図中の◀……▶）に行なうこととした．左右の動きは，ランダム

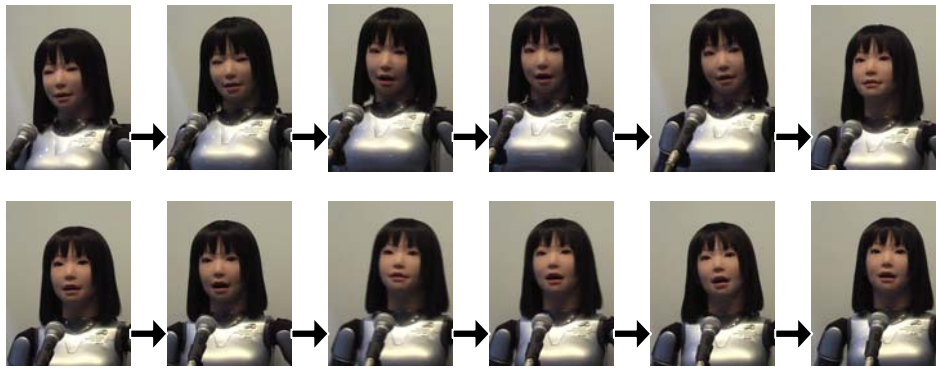


図 10 首動作の生成例

に変化させたが，急激な変化を避けるために，左右に振れた場合には一度中央を経てから反対の方向に向くように状態遷移に制約を設けた．

また，比較的長い間隔があいているポイント間（◀┈┈▶）では，首がゆっくりと上を見上げてから水平位置に戻る動きを行なうこととした．見上げる動作は曲のテンポに応じて変化させた．図 8 に上下左右方向の状態遷移図を示す．図中の◀┈▶，◀┈▶は，それぞれ，手動で付与したタイミング情報間の時間に対応しており，状態中の L，C，R は左，中央，右，U，D は上，下にそれぞれ対応している．各ノードから複数のノードに遷移可能な場合には，ランダムにどちらかの遷移が選ばれるようにした．例えば，中央「C」からスタートした場合には，◀┈▶の時間内に「C」→「C_D」→「C」といった状態遷移を行い，首を上下させる動作が生成される．

さらに，曲中の「キメ」などのタイミングでは，手動で付与するポイント情報において，短い間隔でポイントを打つことで，その間に首を左右に回転（傾ける）状態遷移を行なう（図 9）．これは，首が上下左右方向のどの状態にあっても，その位置を基準に遷移する．例えば首が「右上」の状態にあった場合には，「R_U」の状態から「右上（R_U）位置で右か左へ傾く（回転R／回転L）」へ動く動作が◀┈▶の時間内に行なわれる．なお，次のポイント間の長さが上下左右方向への状態遷移が可能な長さであった場合には，傾きの情報はリセットされる． 図 10に実際の首動作の生成例を示す．

5. おわりに

本システムは CEATEC 期間，大きなトラブルもなく動作し，多くの来場者の注目を集めた．また，ヤマハブース全体では，テレビ放映 23 番組，オンラインニュース記事 80 本以上（海外メディア 20 本以上を含む）に加えて，ネット上には関連動画が 100 本以上投稿され，延べ 50 万回以上再生されたほか，3,500 を超える BLOG で紹介されるなど大きな反響があった．

今後の課題としては，首から上だけではなく，体全体を使った動作表現が挙げられる．また，今回の動作生成は非常に単純なロジックを用いて行なっており，実際の人間の歌う姿などを反映させて，より共感できる人間らしい動作をさせる必要があると考えられる．さらに，VOCALOID の合成システムに関しても，よりリアルに表現豊かな声が再現できるようにすることで，PC 上のソフトウウェアだけでなく様々な分野への応用・発展を目指していきたい．

謝辞 本発表および紹介した展示に関して，（独）産業技術総合研究所知能システム研究部門ヒューマノイド研究グループをはじめ，ご協力いただいた関係者の方々並びにご参加いただいた皆様方に感謝の意を表します．

参考文献

1) 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, Vol. 2008-MUS-74-9, No. 12, pp.51-58, 2008.
2) http://www.crypton.co.jp/
3) http://www.ssw.co.jp/
4) 梶田秀司, 金子健二, 金広文男, 原田研介, 森澤光晴, 中岡慎一郎, 三浦郁奈子, 藤原清司, Neo Ee Sian, 原功, 横井一仁, 比留川博久: サイバネティックヒューマン HRP-4C の開発- プロジェクト概要-, 第 27 回日本ロボット学会学術講演会予稿集, 2009.
5) 金子健二, 金広文男, 森澤光晴, 三浦郁奈子, 中岡慎一郎, 梶田 秀司: サイバネティックヒューマン HRP-4C の開発 -システム設計-, 第 27 回日本ロボット学会学術講演会予稿集, 2009.
6) 中岡慎一郎, 金広文男, 三浦郁奈子, 森澤光晴, 藤原清司, 金子健二, 梶田秀司, 比留川博久: サイバネティックヒューマン HRP-4C の開発 - 顔動作作成システム-, 第２７回日本ロボット学会学術講演会予稿集, 2009.
7) 大中慎一, 安藤友人, 岩沢透:人とのインタラクション機能を持つパーソナルロボット PaPeRo の紹介, 情報処理学会研究報告, Vol. 2001-SLP-37-7, No. 68, pp.37-42, 2001.
8) http://www.business-design.co.jp/product/ifbot/
9) T. Sawada, T. Takagi, M. Fujita: Behavior selection and motion modulation in emotionally grounded architecture for QRIO SDR-4XII, Proc. IROS 2004, Vol. 3, pp.2514-2519, 2004
10) http://www.yamaha.co.jp/product/piano-keyboard/disklavier/e3/
11) 室伏空, 中野倫靖, 後藤真孝, 森島繁生: ダンス動画コンテンツを再利用して音楽に合わせた動画を自動生成するシステム, 情報処理学会研究報告, Vol. 2009-MUS-81, No. 21, pp.1-7, 2009.
12) 白鳥貴亮, 中澤篤志, 池内克史: 音楽特徴を考慮した舞踊動作の自動生成, 電子情報通信学会論文誌D, Vol. J90-D, No. 8, pp. 2242-2252, 2007