歌声を認識するウェアラブルロボットの開発

A Wearable Robot which Recognizes a User's Sung Voice

園田 智也 [1)]　　池長 俊哉 [2)]　　柴田　祐助 [3)]
Tomonari SONODA　Toshiya IKENAGA　Yusuke SHIBATA

1) 2) 3) 早稲田大学理工学部村岡研究室（〒113-6591 東京都新宿区大久保 3-4-1 E-mail: {sonoda, ikenaga, shibata}@muraoka.info.waseda.ac.jp）

ABSTRACT. This paper describes a wearable robot system, recognizes a user's sung melody and tells its music title. Previous systems utilized graphical user interface in recording user's voice and it was difficult to build a system that uses only human voice. We therefore develop a system that a user can talk with a computer and can ask a music title by speaking. The proposed system is designed for not only sound signal processing but also image processing with a video camera or a signal processing of gyro sensor.

１．はじめに

本プロジェクトでは、利用者の肩に乗せて稼動するウェアラブルロボットとして、利用者の歌声を認識するシステムのプロトタイプを開発した。

動物は、視覚・聴覚・嗅覚・触覚・味覚などの五感によって知覚される感覚情報を処理して活動を行っている。我々人間は、さらに、マイクロホンやビデオカメラなどの感覚情報を取得できる装置を伴った装着型計算機（ウェアラブルコンピュータ）によって、本来、知覚できる情報以外の情報の取得や、ある感覚情報に対する概念を想起できないことを回避することが可能となる。例えば、眼前の人物の名前がわからない場合や、現在、聞こえてくる音楽のタイトルを思い出せない場合に、装着型計算機が、利用者の記憶の補助をすることが可能となる。また、このような装着型計算機が常時ネットワークに接続されるユビキタスネットワークのような環境では、利用者が初めて見る光景や音に対しても、計算機が、ネットワーク上の情報源を活用して、利用者に必要な情報を提示することが可能となる。

そこで、本プロジェクトは、このような計算機実現の第一歩として、歌声の認識を行う計算機が、利用者にその曲目を音声で伝えることができるウェアラブルロボットの実現を目的とする。

本システムの最終的な目的は、ロボットが能動的に動き、聴覚信号以外にも、視覚信号など、さまざまな入力インタフェースを有することであり、その設計手法を明らかにすることに重点を置く。このため、本システムには、歌声の認識以外にも、画像認識を利用したジェスチャ解析など、多くの認識技術を利用したアプリケーションを搭載する拡張性を持たせ、広く世の中に普及させることを目的とする。

２．システムの概要

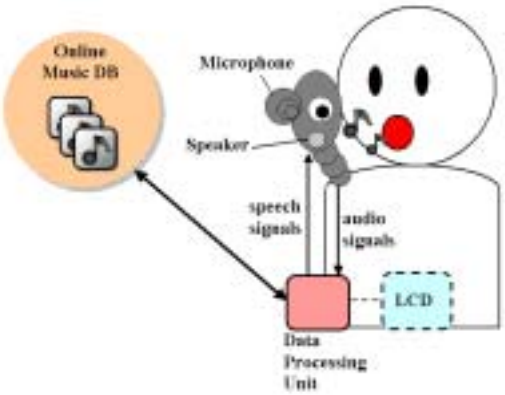

図１：システムの概念図

図２：試作システム

図１、図２に本プロジェクトで開発したシステムの概観

とプロトタイプを示す。利用者の音声は、利用者の肩に設置したロボットの本体に内蔵するマイクによって取得し、装着したデータ処理装置（本実装では Laptop 型 PC）で音声波形解析を行う。波形解析を行った結果は、ネットワーク上のデータベースサーバに送信され、サーバ側では、その曲目を特定し、より似ていると判定された順に曲目を並べた曲リストをデータ処理装置に返送する。最後に、利用者には、計算機の合成音によって、特定された曲目が伝えられる。

本プロジェクトでは、次の（1）～（３）のソフトウェア開発を行った。

(1)ウェアラブルロボットの基盤ソフトウェア
(2) 歌声の旋律認識ソフトウェア
(3) 信号処理ソフトウェア

以下では、以上のソフトウェアの概要を順に述べる。

（1） ウェアラブルロボットの基盤ソフトウェア

本プロジェクトでは、音声・音響信号が処理の対象となるが、上述のように、視覚信号など、様々な入力信号を得られるシステムとなるように、拡張性を持たせた基盤ソフトウェアの設計が必要である。そこで、本プロジェクトでは、この基盤ソフトウェアに対して、

- 状態遷移モデルとして、ロボットの動作を各利用場面に応じて変化させる「動作仕様書モデル」
- 利用者や外界から得られる入力信号を抽象的なイベントデータに変換してから処理する「信号イベントモデル」

の２つを定義・導入することで、将来的に、複数の感覚信号処理プログラムを追加する場合に、基盤ソフトウェアの設計を一切変更することなく、新しい入力インタフェースを構築することができるシステムを実現した。また、このモデルでは、利用者や開発者がプログラムの改変を行うことなく、システムの動作を「動作仕様書」を書き換えるだけでロボットの動作を柔軟に制御することを実現した。本基盤ソフトウェアは、実際には Java で実装されたネットワーク上のサーバモデルを採用している。

（2） 歌声の旋律認識ソフトウェア

上述の基盤ソフトウェア上で動作する信号処理ソフトウェアとして、音声・音響信号を取得し、歌声の旋律を解析する２つの手法を開発した。

①神経のシナプス結合モデルを利用したピッチ抽出の手法（音高の抽出）
②周波数領域のピーク差分を利用した子音検出の手法（音符の長さの抽出）

①は、ウェアラブルロボットで想定される雑音が入る環境下で、人間の音声や音楽音響信号を、優先的に抽出する。この手法では、予め、複数の被験者（本プロジェクトでは３０人）によって入力した音声のみの信号に対して、周波数領域軸上に対応付けた神経素子間のシナプスの結合度を求める。次に、認識対象となる音の解析時に、その結合度を利用することで、雑音環境下の中で特に強く聞こえる周波数信号を抽出する。本モデルでは、**Hopfield** 型神経回路モデルに代表される神経回路モデルにおいて、「同時に発生する信号の出現頻度によって、その信号に対応した神経素子間のシナプス結合度は、強められる」という想定を支持し、ある周波数 A の信号が弱くても、その周波数と結合度の強い周波数が多く存在するときは、その周波数 A も強く聞こえるはずだと仮定する。この仮定に基づいて、結合度の強い周波数を１つのグループとみなして、いわゆる「くし型フィルタ」を構成する。このフィルタを通して、最も強いパワーを持つと判定できたグループの基本周波数を、現在のピッチとする。

図３の(a)～(c) は、複数の被験者の音声を用いて、神経のシナプス結合モデルにおいて、結合度を学習する様子を示したものである。被験者の音声から、周波数領域において、ピッチを求め、そのピッチの５倍音を基準として、ピーク同士の結合度を出現確率によって学習させる。図３中の縦軸の周波数は、あるピッチの５倍音と仮定できた周波数をあらわし、横軸は、そのピッチを検出した場合に強く出現する可能性のある周波数ピークを、確率の高いものをより濃い緑色で表示している。図３(a)では、高周波領域の学習が不完全であるのに対して、(c) では、倍音構造がほとんどの周波数領域に渡って出現している。

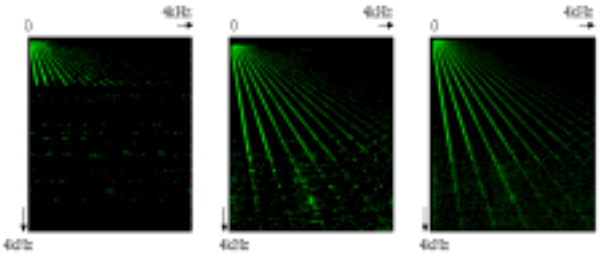

図３：周波数領域に対する神経シナプス結合モデルの学習推移
(a): 10 代 男性１名の 5 回の歌声で学習後
(b):10 代 男女各 3 名で 5 回ずつ計 30 回の歌声で学習後
(c):10,20,30,40,50 代の各世代男女各 3 名(計 30 名)で、5 回ずつ計 150 回の歌声で学習後

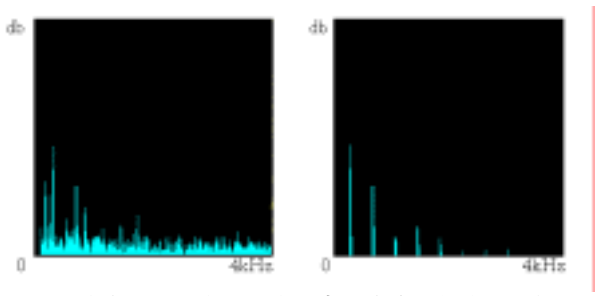

図４：雑音環境の中から利用者の音声を抽出する処理

図４は、神経シナプス結合モデルを利用して、複数のピッチ候補の音が混在する中から、最も強い候補を抽出す

る様子を図示したものである。利用者の音声が周囲の音よりも比較的大きい場合は、上述のシナプス結合モデルを利用して、強い高調波成分を抽出できる。図中では、(a)から(b)を抽出している。利用者の声よりも、周囲のパワーが大きくなる場合に対する手法として、本システムでは、さらに MPEG-AUDIO の圧縮処理でも採用されている聴覚心理モデル（可聴音圧レベル、マスキング効果）を採用している。

②の手法では、歌声に出現する音符の区切りを認識するために、子音検出を行う。歌声の音の高さを形成するのは母音だが、旋律を形成する音符を検出する上で、子音検出は重要な課題である。従来の歌声検索システム [1],[2] では、歌声に現れる子音の検出は、音声波形領域における音量にのみに依存した設計であった。しかし、この手法では「ラララ～」という歌い方に現れる ‘L’の音などを、一般には検出できない。そこで、本ソフトウェアでは、周波数領域における、ピークの差分を手掛かりに母音から子音への変化が発生した時点を特定する。

本手法の予備実験では、子音検出のために、周波数領域のすべてのパワースペクトラム差分を利用したのでは、成功しなかった。これは、人間の神経特性として、強い刺激が与えられた神経をより詳細に分析するために、その周囲の神経の感覚は抑圧される現象と関係していると考えられ、強いピークに対応する神経の周囲の神経が得ているスペクトラムパワーは、実際には知覚されないため（聴覚のマスキング効果）、それらの知覚されないパワーを差分判定に用いることで、誤差が大きくなるためだと考えられる。そこで、本手法では、周波数領域の強いピークのみの位置とパワーの差分のみに着目して、周波数領域上での「乱れ」を検出する。図５では、“ こもった歌い方（はっきりと音符を区切らない歌い方）”をした２０代男性被験者の音符検出を行った結果で、上段は従来の手法、下段は今回の本手法によって得られた結果である。下段の結果は、人間が聞いて音符の区切れと判定する個所と同じ位置で区切れていた。

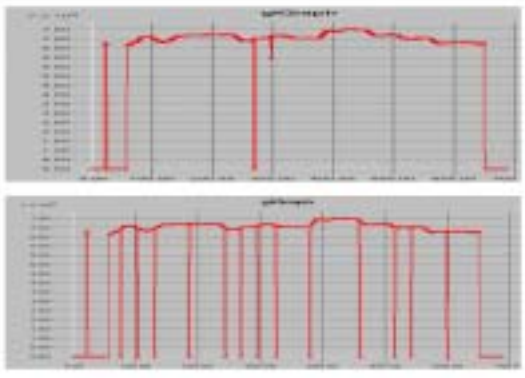

図５:ピーク差分による子音検出処理(上段が従来手法、下段が今回の手法)

（３） 信号処理ソフトウェア

本プロジェクトの最終目的として、歌声認識以外の信号処理にも応用できるシステムを掲げ、基盤ソフトウェアを設計したが､音声以外の信号処理ソフトウェアとして、①動画像入力、②ジャイロセンサー入力、③メール送受信、④音声通信（電話を模した機能）を開発した。
また、既存のソフトウェアとして、音声認識ソフトウェア[3]、音声読み上げソフトウェア（Microsoft Speech API 5.0）、カメラ動作制御（カメラ：キーエンス社 MC-1000）などを、基盤ソフトウェア上で実行できるモジュールを開発した。

３．歌声認識システム全体の構成

以下では、本事業で開発したシステムの全体、および利用者側のシステム構成、曲検索サーバの構成を述べる。

（１） システムのデータ処理

図６に、本事業のシステムにおける、データ処理全体の概観を示す。本システムでは、まず、利用者側が歌って歌声を入力する（システムと会話しながら、曲を尋ねるようなインタフェースに設計した）。このとき、自由なテンポ、調で歌ってよく、曲のどの部分を歌っても良いものとする。録音された音声はデータ処理装置で A/D 変換、音高・音長の抽出処理を行い、ネットワーク上の検索サーバシステムに送信される。検索サーバ側では､SMF（Standard MIDI File）から得られるメロディデータと、利用者に対して音声で読み上げるための曲名・アーティスト名を合わせて、索引データとして格納しており、検索時に利用される。検索の結果、利用者が歌った曲に似ている順に並べられた曲リストが、利用者側のシステムに送信される。そのリストに掲載されている最上位の曲を、本システムは計算機の合成音によって読み上げを行う。

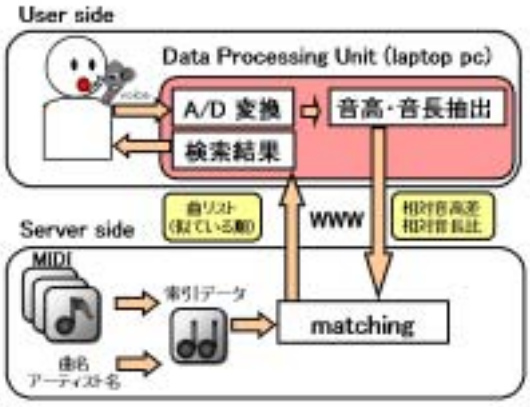

図６：データ処理概観

（２） 利用者側システムの構成

図７に利用者側のシステムの構成を示す。本システムでは、利用者がシステムと会話することを可能とし、歌声を認識するという主目的の他に、将来的に音の信号以外の信号処理を可能とすることを考慮し、全体の設計を行った。具体的には、図中に示す、以下の(a)～(f) の機能に大別できる。

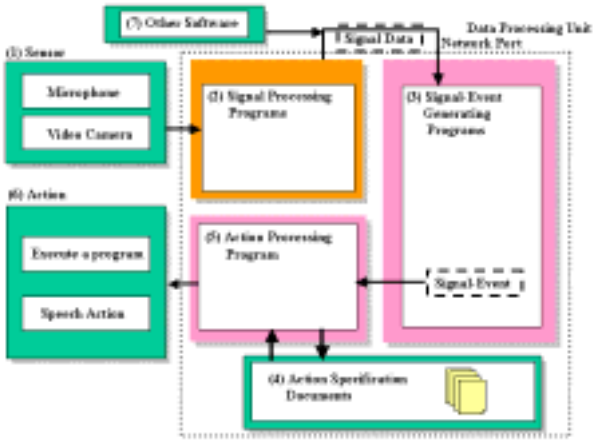

図７：利用者側システムの構成

（a）感覚センサー（Sensor）

本プロジェクトでは、感覚センサーとして、マイクロホンを実装する。また、将来的な実装を見通して、可動式カメラをこの感覚センサーとして設置する。

（b）信号処理(Signal Processing)

マイクロホン、カメラから信号を取得し、ネットワーク経由で「信号イベント」生成プログラムに送信する。ここで、現実装では、送信先は自分自身のデータ処理装置であるが、将来的には、遠隔地の処理装置を利用することも可能となる仕様としている。

（c）信号イベント生成(Signal Event Generating)

本システムでは、取得した音声や他の信号データをそのままプログラムに渡さずに、その信号データをシステムが解析した結果を、「信号イベント」と呼ぶ抽象データで表現し､その信号イベントをシステムの動作に用いる。これは、まったく異なる信号データでも、ある場面では同じ動作をしなければならない場合や、逆に、まったく同じ信号データでも、場面によっては、異なる行動をとらなければならない場合に有効に機能する。

Cf. 人間も、目の前にいる人が、親しい場合と、初対面の場合では、同じ言葉を言われても、受け止め方が異なることがある。

（d）動作仕様書（Action Specification Documents）

本システムでは、ある信号イベントに対して、現在の場面ではどのような行動をすべきかという動作の定義を、簡易なドキュメントファイルによって記述している。このドキュメントファイルを動作仕様書と呼ぶ。動作仕様書は、各利用場面に１つ定義され、次の２種類の記述を有する。

１．現在の利用場面において、各信号イベントが発生した場合の動作記述。（該当する信号イベントに対する記述がない場合は、システムは無視する。）

２．利用場面を変更する動作の記述。（動作場面のハイパーリンク）

（e）動作処理（Action Processing）

動作仕様書に従って、動作決定を行う処理。

（f）動作（Action）

本システムでは、ロボット動作はプログラムの実行と発話処理で行う。また、可動式のカメラを動かす処理を行う。

（７）他のソフトウェア（Other Software）

本システムは、外部の市販ソフトウェアから、上述（３）のプログラムに、ネットワークポート経由で信号データを送り、その信号データを処理することで、任意の信号イベントを発生させることが可能となる。本システムの歌声認識ソフトウェア以外の音声認識、カメラ動作制御プログラムは、この方式で実装される。

４．曲データベースシステムの構成

図８に、本プロジェクトの曲データベースサーバの構成を示す。利用者側のシステムから送信される歌声のデータは、WWW 経由で HTTP サーバの CGI プログラムのサーバを経由して、曲データベースを保有するサーバに渡され、照合処理が行われる。CGI サーバは、負荷分散の役割を果たす。実際には次節で述べるように、図中の２つのサーバは１台の PC 上で稼動させている。

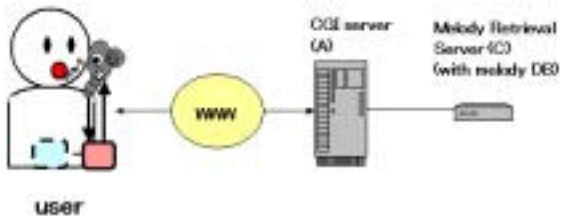

図８：曲データベースサーバの構成

本システムでは、さらに、図９に示すように、並列分散のためのサーバ(B) を導入することで、データベースの規模や利用者のアクセス数に応じたシステムの構築が可能となる設計を採用した。

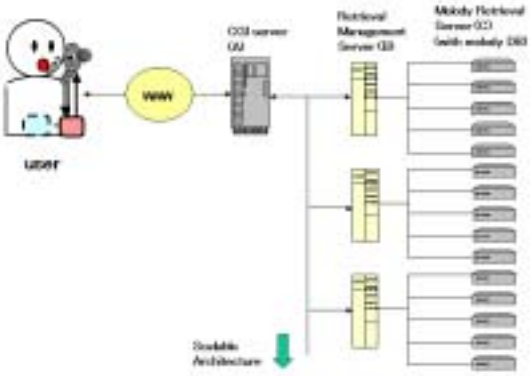

図９：データベースサーバの拡張設計

５．実装と実験結果

利用者側システムと曲データベースサーバ側のシステムの実装は以下のようになっている。

（1） 利用者側システム（ロボット本体）の実装

Laptop PC (CPU:Intel® Pentium® III mobile 866MHz, RAM:392M)
音声入力装置（マイクロホン：MM-SP101VSB）
画像入力装置（ビデオカメラ：Keyence MC-1000）
実行 OS Windows2000 SP2
システム実行環境 Java(j2sdk1.4.0_01)
音声認識ソフトウェア（Julius dictation-kit-v.2.1-sp2）
音声発話エンジン（Microsoft Speech API 5.0）

実装上のパラメータ
歌声認識の音声サンプリングレート（8kHz, 8bit, モノラル）
FFT の窓関数（512 点、ハミング窓、128 点シフト）

（2） 曲データベースサーバ側の実装

データベースサーバ用 PC
（CGI サーバも兼用、CPU: Intel® Pentium® III 866MHz, RAM: 512M）
実装 OS Linux: (kernel version 2.2.18-0vl4.2)
システム実行環境 Java(j2sdk1.4.0_01)

（3） 実験結果

上述したシステムを実装し、認識性能の調査を行った。

本プロジェクトで開発したシステムの認識の正答率を評価するために、20,30 代男女 12 人の 60 回の入力（各被験者につき 5 回入力）によって、日本のポピュラーミュージック 1795 曲を格納しているデータベースに対して、認識実験を行った。このとき、被験者の歌い方は、タタタやラララなど歌詞を伴わない歌い方で、入力時間は 15 秒以内とした（曲データベースの MIDI に適合しやすくするための指示）。実験は、本プロジェクトで開発した子音検出処理の手法を利用したものと、そうでないもの [1][2] を比較した。実験結果を表１に示す。

表１：歌声の認識結果

子音検出処理の有無	正答率（望んだ曲目が出現した頻度）
（1）子音検出処理無し	68.3% (41/60)
（2）子音検出処理有り	80.0 % (48/60)

（4） 考察

本プロジェクトで開発した手法が有効に機能していることが数値的に確認できた。歌詞を伴った歌い方でも、比較的認識できる曲が増えたが、今後は曲データベースの作り方を検討し、より普通に歌って認識できる手法を開発すべきである。

また、本開発の予備実験として、PC 付属のマイクロホンと、会議用のコンデンサーマイクによって、街中の雑踏で録音した曲に対して、旋律の抽出を行う実験を試みたが、録音後のデータを確認したところ、マイクロホンの性能が原因で、環境音や周囲の人の声が大きすぎるか、楽曲の音圧レベルが低かったかのいずれかによって、旋律を確認できるデータは得られなかった。今後の開発では、マイクロホンの性能や指向性に関しては、より研究を行う必要があることが明らかになった。

６．まとめ

本開発では、GUI を用いずに、音声と、歌声から検出されるメロディを手がかりに、数万曲を対象として曲認識を行うことができるシステムを構築した。

また、これまでの歌声検索エンジン[1][2]では、一切、利用者の録音時の音響信号処理に関する検討はまったく行われていなかったが、今回の開発により、聴覚心理モデル、神経のシナプス結合モデルが導入され、より人間らしい認識が可能となった。

本プロジェクトの成果ソフトウェアは､携帯電話上でも、実現可能であることから、現在、携帯電話キャリアと事業化に向けた話を進めている。また、独自に IP 電話に向けた開発を進め、新規に起業する予定である。

参考文献

[1] THE DESIGN METHOD OF A MELODY RETRIEVAL SYSTEM ON PARALLELIZED
COMPUTERS: Tomonari Sonoda, Toshiya Ikenaga, Kana Shimizu and Yoichi Muraoka, WEDEELMUSIC 2002, December, 2002.

[2] WWW 上での歌声による曲検索システム: 園田　智也、後藤　真孝、村岡　洋一, 電子情報通信学会論文誌, D-II VolJ82-D-II No.4, pp.721-731, 1999.4.

[3] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 山本幹雄, 山
田篤, 宇津呂武仁, 鹿野清宏: 日本語ディクテーション基本ソフトウェア（９８年度版）日本音響学会誌, Vol.56, No.4, pp.255-259. 2000.