音 楽 情 報 科 学 26－3
（1998．8．7）

母音認識とピッチ検出を用いた歌声のテンポ抽出　第3報

東　英司　　橋本 周司
早稲田大学理工学部
E-mail : {azuma , shuji}@shalab.phys.waseda.ac.jp

人が意識的あるいは無意識的に揺らして歌った歌のテンポに合わせて演奏する自動伴奏システムの実現が本研究の目的である。そのために、ケプストラム法から母音とピッチを検出することで、歌唱位置及び歌声のテンポ抽出を行う。また音声入力をMacintosh内蔵のサウンドデバイスで行い、ピッチ検出に倍音構造からの推定法などを使用することで、伴奏システムにはマイクの他に特別なハードウェアを必要としない。現在MAXでの構成にすることで、より正確かつ自然な追従演奏を目指している。ここではその概要を報告する。

Tempo Extraction of Human Singing Using Vowel Recognition and Pitch Detection　III

Eiji Azuma　Shuji Hashimoto
School of Science and Engineering, Waseda University

An overview and experimental results of an automated accompaniment system for singing are described. The purpose of this system is to produce an adaptive accompaniment in real time to follow the human singing in an arbitrary tempo. The system analyzes the singing position in detail by two keys; the vowels of lyric and the pitch of singing. The singing voice is obtained by the sound-device in Macintosh, and the singing pitch is detected from the harmonic structure in real time. Therefore, the system is realized by using software but not any special hardware such as DSP and LowPassFilter.

1．はじめに

我々は言語を媒介としてコミュニケーションを交わす。しかし互いに扱う言語が異なる場合、その情報伝達は極めて難しくなる。この場合、ジェスチャーなど試行錯誤で意志伝達をするであろう。ところで、音楽は世界共通の言語と言われている。確かに、伝わりにくい歌詞のニュアンスや雰囲気を、リズム、音量、テンポ、メロディなどに載せることで、多くの人々の心に訴えかけることができると言える。最近はカラオケの出現により、歌を「聴く」という受け身の姿勢ではなく、歌を「歌う」「聴かせる」という能動的姿勢に移り変わってきた。カラオケのように伴奏に合わせて歌うのではなく、自由かつ表現豊かな歌唱というものが機械による伴奏においてもで可能になるとすれば、聴き手、歌い手共により心地よい演奏が楽しめるであろう。そのような背景から我々は人の歌のテンポに合わせて伴奏を出力する自動伴奏システムについて研究してきた[1][2][3][4]。つまり規定テンポという制約条件を取り除くことで「ここのフレーズはゆっくりと歌いたい」などの要求に答えることができると考えたのである。今までに様々な自動伴奏システムの研究について多くの報告がなされている[5][6][7] が、メロディ入力が歌声そのもの、つまりアコースティックサウンドである場合[8][9][10][11] 、一般に歌唱位置の検出は容易ではない。過去の研究においては歌声の音程をトラッキングに用いた自動伴奏システムが数多くあるが、歌い手が音程を外して歌ってしまうケースが多々あるために、この問題を解決する何らかの手法が必要となる。そこで、我々は歌声の音程情報の他に、歌声の歌詞情報（母音）を扱うことで、歌声のトラッキングの精度を上げることに成功している。音程と歌詞の検出については主にケプストラム法を利用している。特に音程情報の検出には、倍音構造を用いた比較的小さな窓長でも正確に検出できる方法を用いている。現在このシステムをMAX上での構成に移植し、より自然な伴奏の実現を図っている。ここではその経過と特にテンポ追従実験について報告をする。

2．システム概要

システムの概要を図1に示す。まず、歌声をAD変換し、それに対しケプストラム法を用いてリアルタイムで母音認識とピッチ検出の両方を行う。これにより、随時得られる母音とピッチのデータから歌唱の位置、テンポを割り出し、歌唱に合わせて伴奏をMIDIシグナルで同期出力させる。

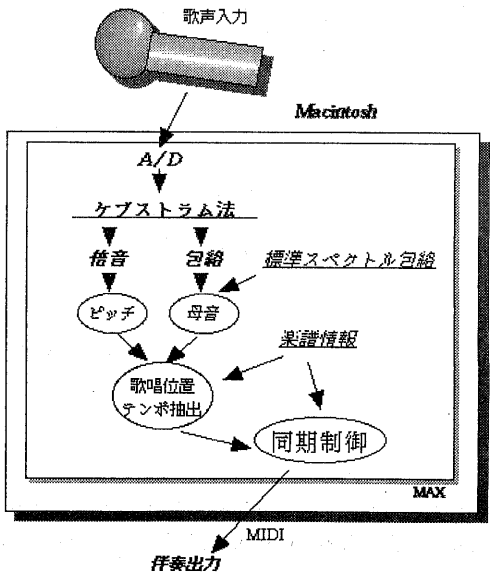

図1　システム構成

3．解析部

3-1　母音認識

歌唱位置判定のために母音認識を行う。実用性という観点から見れば不特定話者認識が理想と言えるが、本システムにおいて望まれる母音認識

は実時間での高い認識率を必要とする。そのため現段階では特定話者認識を採用している。

オフライン処理

自動伴奏を行う前に歌い手に5種類の母音（あ～お）をそれぞれ数秒間ずつ発声してもらい、あらかじめ各母音のスペクトル包絡（100フレーム分）の平均と分散を計算し、これを標準スペクトル包絡パターン（平均包絡パターン、分散包絡パターン）として保持しておく。音声分析のフレーム長は23.2msとする。スペクトル包絡の検出には、パワースペクトルの対数をとって逆フーリエ変換したケプストラムの低ケフレンシ部分をフーリエ変換する一般的な手法を用いている。

オンライン処理

オフライン処理と同様に歌声の低ケフレンシ部を随時フーリエ変換することにより得られるスペクトル包絡と、5つの母音の標準スペクトル包絡パターンとのマッチングを行う。

$$k_i = \sum_{j=0}^{N-1} \left[-\frac{\log(2\pi\sigma_{ij}^2)}{2} - \frac{(y_j - \alpha_{ij})^2}{2\sigma_{ij}^2} \right] \cdots (式1)$$

k_i ：母音 i との類似値
α_{ij} ：母音 i、周波数 $f_j(=j\Delta f)$ におけるスペクトル包絡平均
σ_{ij}^2 ：母音 i、周波数 f_j におけるスペクトル包絡分散
y_j ：随時得られる周波数 f_j におけるスペクトル包絡
N ：1フレーム中のサンプル数

i ：母音（i = 0,1,2,3,4）
j ：（0,1,……,N-1）

分散の小さい周波数帯は音響的特徴を良く表すので、マッチングの荷重を大きくするべきである。そのため、本システムではマッチングに式1を用いた。標準スペクトルとの類似度が最も高く（k_i が最大）、閾値を超えたものを歌われた母音と判定する。

3-2　ピッチ検出

母音認識と同時にピッチの検出を行う。ピッチ検出としては様々な方法が考えられるが、ここではケプストラムの高ケフレンシ部をフーリエ変換して得られるスペクトルを利用する。このスペクトルは対数パワースペクトルの微細構造部分に該当するため、包絡情報が欠如した倍音構造を持つことになる。これをピッチ検出に用いるわけであるが、必要最低限な半音の区別をするには、長いフレーム長が必要となり、実時間でのピッチ検出が不可能になってしまう。そこで第1ピークのみの推定ではなく整数倍の倍音のピークの位置を用いれば、フレーム長を短くしても半音の区別が可能となる。つまり第n倍音の周波数をnで割った周波数が基本周波数と推定できるということである。しかし精度向上のため特定の倍音のみで推定せず、すべての倍音に対し、重みをつけて基本周波数の推定を行うことにした(式2)。尚、パワーが小さい音声区間については無音もしくは子音の部分と判断し、母音認識、ピッチ検出の対象としない。n=12、表1のサンプリング条件(23.2ms)の場合、式3の条件からF(87.3Hz)からA(880Hz)までの基本周波数に対し、半音区別が可能となる。

$$f_1 = \sum_{k=1}^{n} f_k \Big/ \sum_{k=1}^{n} k \quad \cdots (式2)$$

$$2\Delta f < f_1 < \frac{F}{2n} \cdots (式3)$$

f_1 ：基本周波数
f_k ：第k倍音目の周波数
F ：サンプリング周波数

サンプリング周波数	22.05 (kHz)
周波数分解能	43.1 (Hz)
フーリエ窓長	23.2 (ms)
サンプル	512 (point)

表1　サンプリング条件

3-3　歌唱位置判定

上で述べた手法から、23.2ms 周期で次々と得られる母音とピッチの情報を用いて、楽譜情報とのマッチングを行うことで、歌っている箇所を判別していく。楽譜情報には図3のように歌詞（母音）、音程（ピッチ）、音符の長さ、歌唱推定ポイントが含まれている。歌唱推定ポイントとは歌っている場所を監視するポイントのことで、歌声と伴奏の同期をとる部分に相当する。また、歌唱推定ポイントは1拍もしくは2拍おきに指定している。実際の伴奏において、その歌唱推定ポイント付近での歌い手の歌唱情報と楽譜情報とが一致した場合、その音符が歌われたと判断する。音程を外す頻度が多い場合には母音とピッチのいづれかが一致した場合にその箇所が「歌われた」と判定する。

	み	や	こ	の	せ	い	ほ	く	わ	せ	だ
母音	2	1	5	5	4	2	5	3	1	4	1
ピッチ	55	60	60	60	60	60	59	55	52	53	55
音長	24	24	24	24	24	24	24	24	18	6	48
歌唱推定ポイント	1	1	1	1	1	1	1	1	1	0	1

図3　楽譜情報

3-4　伴奏出力

歌唱推定ポイントにおいて歌声と伴奏を同期させなければならない。このため、Max の seq オブジェクトを利用した。seq オブジェクトには四分音符の 24 分の1だけ伴奏をすすめる機能があるため、これにより拍の頭などで伴奏と歌との同期を取って行くことになる。例えば歌唱のテンポが伴奏のテンポよりも遅い場合、２４分の２３拍目でいわゆる「待ち」の状態に入り、「歌われた」瞬間に伴奏も次の演奏へ展開していく（図4-1）。すなわちこの時完全に歌と伴奏は同期した状態になる。また、ある程度歌のテンポが遅くなった際には次の拍で同期できるよう、伴奏のテンポを更新する（テンポを遅くする）。逆に歌唱が伴奏よりも早い場合は、無理矢理同期させようとすると伴奏を途中で切る必要があるため音楽的に不自然になってしまう。したがって、その次の歌唱推定ポイントでの同期がスムーズに取れるように処理を行っている（図4-2）。そのため「歌われた」瞬間は完全には同期しない状態になるが、伴奏のテンポを早くすることにより、このずれを徐々に回復していく。

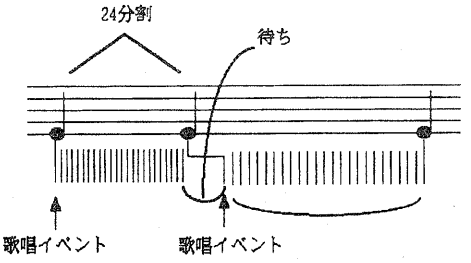

図4-1　歌唱のテンポが伴奏より遅い場合

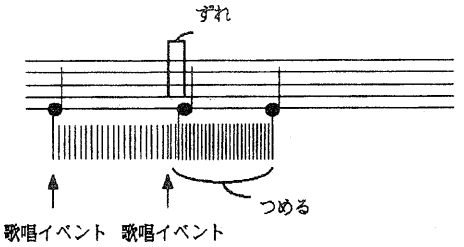

図4-2　歌唱のテンポが伴奏より早い場合

4. 実験

伴奏と歌唱タイミングの同期についての実験を行った。実験では歌唱推定位置を１拍毎に設け、歌唱タイミングをマウスのクリックで与えて自動伴奏を行った。実験で使ったMAXプログラムを図５に示す。

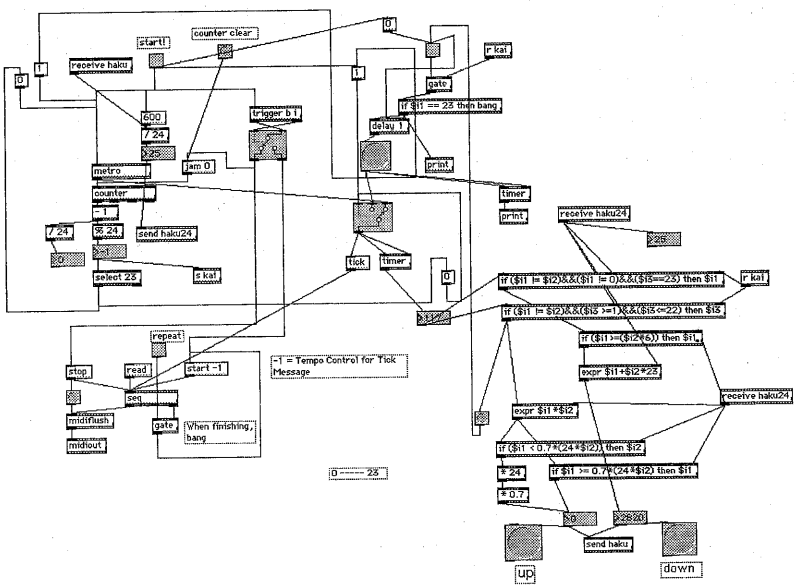

図５　Max における同期プログラム

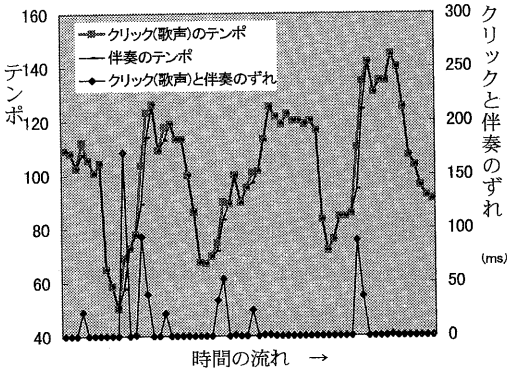

図６　テンポ変化に対する伴奏の位相ずれ

実験結果は図６である。図から歌のテンポが遅くなる場合にはアルゴリズム通り、ずれが生じていない。またそれほど急激なテンポアップさえしなければ、位相差は 100ms 以内におさまっており、更に次のポイントでは、ずれが減少する方向に向かっている。

5. おわりに

本論文において人の歌声に合わせた自動伴奏システムについて述べた。今回は特にMAX上での構成に移行しオブジェクトを用いることで、自然かつ滑らかに追従することができた。マウスクリックでの実験からわかるように精度のよい追従が行えた。

一方、問題点として「待ち」が挙げられる。極端にためを作る場合はともかく、中途半端にテンポを遅くした場合、最後の２４分の２３拍目から１拍目が間延びしてしまい、音楽的に良い印象を与えない。この点については今後の課題である。しかし、位相差が拡大しないという利点がある。また現在、母音認識とピッチ検出を用いた全体システムの公開を検討している。

参考文献

[1] 井上、橋本、大照、“適応型歌声自動伴奏システム”、情報処理学会論文誌、vol.37 pp.31-pp.38 (1996)

[2] 東、尾上、橋本、“母音認識とピッチ検出による歌声のテンポ抽出”情報処理学会第 54 回全国大会講演論文集(2)、pp.283-pp.284 (1997)

[3] 東、橋本、“音声認識とピッチ検出を併用した歌声の自動伴奏”情報処理学会 音楽情報科学 97-MUS-22 pp.1-pp.5(1997)

[4] 東、橋本、“母音認識とピッチ検出による歌声のテンポ抽出２”情報処理学会第 56 回全国大会講演論文集(2)、pp.52-pp.53(1998)

[5] Dannenberg, RB. "An On-Line Algorithm for Real-Time Accompaniment", Proc.of ICMC, pp.193-pp.248 (1984)

[6] Dannenberg, RB. and Mont-Reynaud,B. "Following an Improvisation in Real-Time", Proc. of ICMC, pp.241-pp.248 (1987)

[7] 直井、大照、橋本、“実時間拍検出機能を用いた自動伴奏システム”、日本音響学会講演論文集、pp.465-pp.466 (March,1989)

[8] Vercoe, B. "The Synthetic Performer in the Context of Live Performance", Proc.of ICMC, pp.199-200 (1984)

[9] Katayose, H., Kanamori, T., Kame, K., Nagashima, Y., Sato, K., Inokuhci, S. and Simura,S. "Virtual Performer ", Proc.of ICMC, pp138-pp.145 (1993)

[10] Horiuchi, Y. and Tanaka, H. "A Computer Accompaniment System With Independence", Proc.of ICMC, pp.418-pp.420 (1993)

[11] Grubb,L. and Dannenberg,R. "A Stochastic Method of Tracking a Vocal Performer", Proc.of ICMC, pp.301-pp.308 (1997)