音 楽 情 報 科 学 22－1
（1997. 10. 18）

音声認識とピッチ検出を併用した歌声の自動伴奏

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

歌い手の意志を反映させるため、意識的に揺らした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、歌声のテンポ抽出を行うことが必須となる。そこでケプストラム法から母音とピッチを検出することで、歌唱位置推定する試みを行っている。製作したシステムでは、音声入力にMacintosh内蔵のSoundInputDeviceを用い、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、DSP、ローパスフィルタ等の特別なハードウェアが不要となっている。ここではシステムの概要と幾つかの評価実験について報告する。

Automated Accompaniment using Sound Recognition and Pitch Extraction

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. In order to realize the system, it analyze the singing position in detail by two keys, the vowels of lyric and pitch of singing. The singing voice is obtained by "SoundInputDevice" in Macintosh, and the singing pitch is detected by the harmonic structure in real time. Therefore, the system consist of only software not using special hardware such as DSP and LowPassFilter.

1．はじめに

現在商用化されているカラオケシステムは選択された曲に対し、規定の時間、規定の速さで人の意志に係わらず演奏をする。もし、これらのカラオケの演奏に人の意志などが反映されれば、聴き手、歌い手ともに心地良い演奏を楽しめるであろう。そこで我々は、人が歌う際に意識的にテンポ揺らして歌いたい、ここのフレーズはゆっくり歌いたいなどの要望に答えるべく、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を製作している。それまでにも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[1][2][3]が、メロディが歌唱などのいわゆるアコースティックサウンドの場合[4][5][6]、一般に歌唱位置やテンポを細かく抽出するのは難しい。そこで本システムにおいてはテンポ抽出の手法としてケプストラム法を利用し歌い手の発声する母音とピッチを実時間で獲得することで楽譜情報から歌唱位置、テンポを割り出すことにした[7]。その際、ピッチ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。又、歌声の入力をMacintosh内蔵の“Sound-Input-Device”で行うことで、ソフトウェアだけでの構成が可能になった。本稿の前半ではシステムの概要を中心に説明する。後半では本システムの実験について述べ、母音とピッチを併用した方法が有効であること等を確認する。

2．システム概要

システムの概要を図1に示す。まず、Macintoshに内蔵されている“Sound-Input-Device”で歌声をAD変換し、それに対しケプストラム法を用いリアルタイムで母音認識とピッチ検出の両方を行う。得られた母音、ピッチの情報から歌唱の位置を判定し、歌い手のテンポを割り出す。そのテンポを用いてMIDI音源により伴奏を出力することで人のテンポに合わせた伴奏が可能になる。

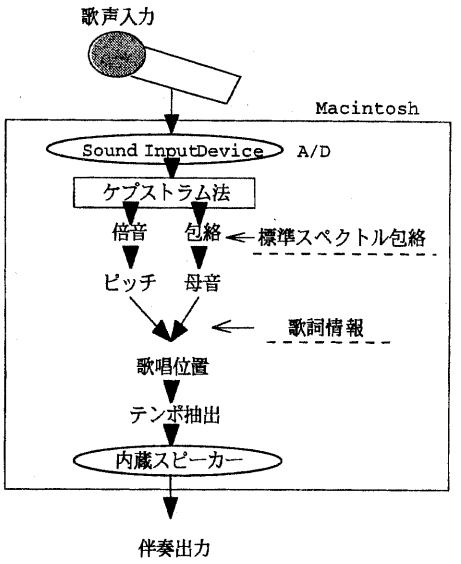

図1　システム概要

3．テンポ解析

3-1　母音認識

歌唱位置判定のために母音認識を行う。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。そのため特定話者認識を採用することにした。

まず、自動伴奏を行う前に歌い手に5種類の母音（a,i,u,e,o）を発声してもらい、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として保持しておく。そして実際の自動伴奏中にフレーム長約23.2ms（サンプリング周波数11.025kHz 量子化ビット数16bit、サンプル数256点）で低ケフレンシー部をFFTすることにより随時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する。マッチングには式1を用いた。この式は分散の大きい周波数成分ほどマッチングの荷重を軽減するよう配慮されている。その結果、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。

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

i: 母音 (i=1/a,2/i,3/u,4/e,5/o)
j: j番目の周波数 ($f_j = j\,\Delta f$)
k_i: iの母音との類似値
α_{ij}: jの母音の f_j におけるスペクトル包絡の平均
y_j: 随時得られる f_j におけるスペクトル包絡
σ_{ij}^2: jの母音の f_j におけるスペクトル包絡の分散
N: サンプリング数 (256)

3-2 ピッチ検出

母音認識と同時にピッチ検出も行う。ピッチの検出としては様々な方法が考えられるが、ここでは高ケフレンシー部をフーリエ変換して得られる倍音構造を利用する。但し、半音を区別するのに必要な周波数分解能を考慮すると窓長が長くなるため実時間処理には適さない (C(48)とC#(49)の半音差を区別するにはフレーム長約128ms が必要)。そこで整数倍の倍音のピークの位置を用いピッチ検出行えば、周波数分解能をある程度高くし、窓長を短くしても差し支えない。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。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、パワーが小さいフレーム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。本システムではスペクトルの周波数刻みは約43.1Hz (フレーム長約23.2ms) とし、n=12 とした。したがって、C(48) からF(65) まで半音の区別ができる。

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

f_1 : 基本周波数

f_k : 第k倍音目の周波数

3-3 歌唱位置判定

実時間で得られる母音、ピッチと楽譜情報をマッチングさせ、歌っている箇所を判別する。楽譜情報としては歌詞 (母音)、音程、音符の長さを入力してある。具体的には抽出された母音、ピッチのいづれかが楽譜情報と一致した場合その音符が歌われたと判断する。但し、下記の条件などにより、歌唱位置の誤判定を防ぐことにした。

(1) 一つの言葉を歌った直後から200 ms以内に次の言葉が歌われることはない
(2) 急激なテンポ変化はない
(3) ピッチ変化が半音で、母音も変わる箇所についてはなるべく母音のみで歌唱位置を判定する

3-4 伴奏出力

歌唱位置から歌い手のテンポを計算する。実際には1拍の長さをms単位で計算する。しかし、歌われたテンポで伴奏を流してしまっては、抽出しているテンポが1つ前の音符間のテンポであるため、歌い手の変化に遅れてしまい歌い手と伴奏が一致しない。そこで、本システムでは次の音符で歌と伴奏のタイミングが一致するように伴奏のテンポを随時変化させるようにした。

4. 評価実験

4-1 母音認識率とピッチ認識率

システムで扱われる母音認識とピッチ検出の精度について実験を行った。まず母音認識については、自動伴奏の場合と同様に予め作成しておいた標準スペクトル包絡パターンとのマッチングを行う。自由な音程で指定した母音を約2秒間発声してもらった (被験者10名)。表1に結果を示す。平均して90%程度の認識率であるが、特に標準パターン作成の際発声した音程と大きく異なる音

程の場合に認識率が悪くなることが判った。

表1　母音の認識率

A	I	U	E	O
90%	89%	87%	84%	93%

ピッチ検出においてシステムの対象となるのは“歌い手がある特定の音程を出しているつもり”の状態に対してであり、その音程が実際よりも低いなど、正しいとは限らない。人の声と楽器音を比較するために、入力をキーボードの音（正確な音程）とした場合のピッチの認識率と、入力を人の声（ある特定の音程を出してもらうように依頼した場合）とした場合のピッチの認識率の2パターンについて実験を行った。被験者1人に対し、ある音程を約2秒間発声してもらった。音階を48から65の間（MIDIノートナンバーで表示）で行った結果が図2である。

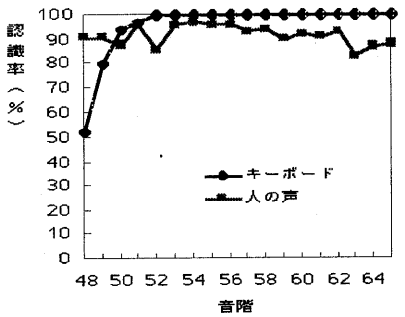

図2　ピッチの認識率

ピッチの高い部分では、正確な音程で発声することは難しいためにピッチが変動している可能性がある。

4-2　ピッチの外れた歌声に対する自動伴奏

実際の歌唱において楽譜どおり正しい音程で歌うことは難しい。しかし、本システムは母音認識とピッチ検出の2つを利用しているため、たとえ音程が楽譜情報と合致しない時があったとしても理論上、母音からメロディを認識することは可能である。そこで、キーをある程度外して歌ったもの（1名の男性）を録音しておき、それをシステムに聴かせることで自動伴奏を行った。但し、入力する歌のテンポは一定になるようにした。

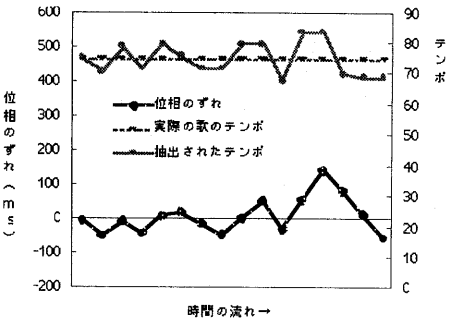

図3　ピッチを外した歌に対するテンポ抽出と位相のずれ

位相の大幅なずれの原因としては、譜面上同じ母音の続く箇所において母音認識のみでは細かい歌唱位置判定が出来なくなること、又、母音が約25msに1回認識するという時間はテンポの抽出にとってやや長いと言えること、などが挙げられる。

4-3　正しいピッチで歌われた場合の自動伴奏

次の3つ場合について実験を行った。

システム(1)　母音認識のみで歌唱位置推定を行う自動伴奏システム
システム(2)　ピッチ検出のみで歌唱位置推定を行う自動伴奏システム
システム(3)　前述の母音認識とピッチ検出を併用し歌唱位置推定を行う自動伴奏システム

正しいピッチで歌ったものを録音し、システム(1)(2)(3)に対して自動伴奏を行いテンポ追従の比較をする。図4に実際の歌のテンポと各システムにおいて抽出されたテンポを示す。図5では歌声と伴

奏とのずれを示した。図４において一見、テンポ抽出には差がないように見えるが、図５のようにシステム(1)(2)とシステム(3)を比較すると、（３）の位相のずれが他に較べ小さい。これは母音単独もしくはピッチ単独でのテンポ抽出はフレーズによっては同じ音、同じ母音で発声されるために歌唱位置を推定が出来ないためと推測される。

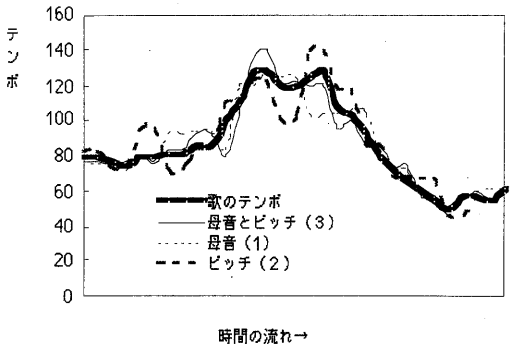

図４　歌声のテンポ抽出

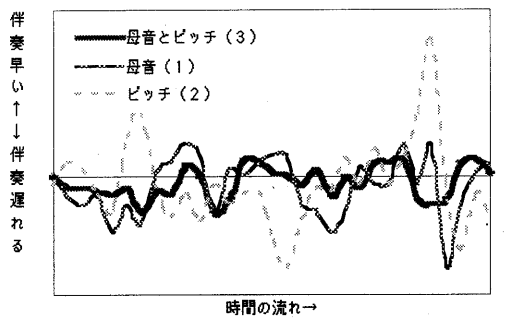

図５　歌声と伴奏の位相差

5. おわりに

自動伴奏の歌い手のテンポ抽出に対し、母音認識とピッチ検出を併用することにより細かい歌唱位置を推定でき、テンポに適応した伴奏を実現した。特にある程度音程がずれていても主に音声認識（母音認識）を用いることによりテンポ追従が可能になった。しかしテンポの追従が機械的になってしまうため、テンポ追従のアルゴリズム等の改善を行う必要がある[8]。又、本研究室において提案されたリアルタイムで声質を保存したままピッチを変化させることが可能なアルゴリズム[9]を導入することにより、音程を外した場合に実時間で歌声を正しい音程に変換させる機能を付加させることなども合わせて検討していきたい。

参考文献

[1] Dannenberg,RB. An On-Line Algorithm for Real-Time Accompaniment , Proc.ofICMC, pp.193-pp.248 (1984)

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

[3] 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)

[4] Vercoe,B. The Synthetic Performerin the Context of Live Performance , Proc.ofICMC, pp.199-200(1984)

[5] 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)

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

[7] 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)

[8] Horiuchi,Y. and Tanaka,H. A Computer Accompaniment System With Independence , Proc.of ICMC, pp418-pp.420(1993)

[9] 笹平宜誠、橋本周司、“声質を保存した歌声のピッチ補償”情報処理学会第54回全国大会講演論文集(2)、pp.281-pp.282(1997)