音メディアデータの高度アーカイブシステムに関する研究

－芸能音楽の記録と再現のための高度アーカイブシステムの開発－

山下 洋一

Abstract: This paper describes some basic techniques to realize an advanced archive system for audio data. Development of a spoken dialogue prototype system with the Galatea toolkit made it clear that the Galatea toolkit is useful to develop a user friendly interface using an anthropomorphic agent and some improvement is necessary. A new method for recognizing pitch in music sounds is proposed to annotate the music contents. Finally, it is shown that the introduction of prosodic information, such as F0, power, and duration, into the speech summarization based on the extraction of important sentences improves the quality of the speech summary.

1. はじめに

音メディアデータを対象とした高度なアーカイブシステムを実現するには、大きく分けて二つの課題がある。一つは、データへのアノテーションの自動化など、コンテンツ作成に関する支援技術であり、もう一つは、システムを利用する上でのインタフェースの高度化である。本サブプロジェクトでは、音声情報処理技術を用いて、これら両面から音メディアアーカイブシステムの高度化を目指す。本報告では、擬人化エージェントを用いた音声対話システムの構築、楽曲データに対する音高認識、音声データに対する自動要約について述べる。

2. 擬人化エージェントを用いた音声対話システム

機械とのインタフェースを使いやすくするためにさまざまな研究が行われている。人間同士が音声を用いて自然にかつ容易に会話していることから、機械とのインタフェースにおいても音声によるコミュニケーションを可能にしたユーザフレンドリなインタフェースに期待が集まっている。音声認識と音声合成による音声によるコミュニケーションだけでなく、「機械」に対して話しかけるという抵抗感を緩和するために、擬人化エージェントを持った音声対話システムの研究が進められている。このようなインタフェースでは、画面に人物（エージェント）の画像が表示されることにより、ユーザはコミュニケーションをとる具体的な対象を得ることができる。

機械があたかも一個人の人間のように振る舞い、人間の顔や姿を表現し、音声言語で話し聞くような擬人化音声対話エージェントは、今後のヒューマンインタフェース技術において大きな目標の一つである。これまでに、研究開発者が容易に使用・開発参加できる共通の研究プラットフォームを作ることを目的として、擬人化音声対話エージェントツールキット Galatea を開発し、無償で公開してきた。このツールキットは、顔画像が容易に交換可能、多様な韻律での音声合成、対話制御の記述変更が容易、機能モジュール自体を別のモジュールに差し替えることが容易、さらに処理ハードウェアの数への柔軟な対処などの特徴を持っている。

擬人化エージェントツールキットGalatea を利用して音メディアアーカイブシステムのインタフェース部を実現することをめざし、まず、コンテンツがある程度作成されている研究室紹介を具体的なタスクとして対話システムを構築することによって、Galatea の有効性を検証するとともに問題点

を明らかにした。

実装した機能として、

- 研究室ホームページ上の文書の読み上げ
- システム発話中の割り込み
- Webページやコンテンツにあわせて顔画像と声を担当者の顔と声に切り替え

などがあげられる。短期間で様々な機能を持った音声対話システムを構築したことにより、Galateaの有効性を検証できた。本研究では、音声認識における誤認識を避けるために、それぞれの文法ファイルに与える文法数を押さえ、複数の文法ファイルを切り替える手法をとった。そのため認識率は高く誤認識はごく稀であった。一つの文法ファイルにおける語彙数も十分な数と思われる。さらに、今後 Galatea に求められる機能及び課題として、多様な声質による合成音の生成、数字読み上げの指定方法の改善、発話内容が多い場合の発話遅延の解消などを明らかにした。

3. 楽曲データに対する音高認識

楽器によって演奏された楽曲データを保存し、必要に応じて検索し利用するには、物理的な音としての波形データだけでなく、楽譜などその内容を記述したデータが利用できることが望ましい。しかし、楽譜も作成することは、MIDI (Musical Instrument Digital Interface) を利用できない邦楽器などの楽器による演奏や過去の演奏においては、熟練者によるコストのかかる作業が必要となる。

近年、コンピュータの性能向上にともない、コンピュータで音楽データを扱うことが容易になり、採譜を自動化する研究が試みられている。自動採譜システムの実現には、音高認識、調性認識、リズム認識等のいくつかの処理が必要となる。本研究では、西洋音楽を対象としてケプストラムを用いた音高認識手法を提案し、評価を行った。

ケプストラムは、楽曲や音声などの時間波形データをフーリエ変換して得られるパワースペクトルを対数変換してからフーリエ逆変換して得られる特徴量であり、波形に含まれる周期性の検出に用いられる。しかし、ケプストラム上のピークだけに注目した音高認識を行うと、n倍周期成分の出現や検出したいピークへの他のn倍周期成分の重なりという問題があり、認識率は劣化してしまう。そこで、人がある認識したいケプストラムを見て、その他の構成音が既知であるケプストラムと比較することにより、ある程度音名を推測できることに注目し、これをパターン認識の問題と考えて音高認識するという手法を提案した。

認識対象範囲である音高をクラス ω_i とし、そのクラスに対する特徴パラメータを x_i とすると、x_i が生起したときのクラス ω_i の事後確率 $P(\omega_i|x_i)$ はベイズの定理から次式で与えられる。

$$P(\omega_i|x_i)=\frac{P(x_i|\omega_i)P(\omega_i)}{p(x_i)} \quad (1)$$

この確率が大きいときに入力データはクラス ω_i に属す、すなわちこの音高を含んでいると考える. ここで $p(x_i)$、$P(\omega_i)$ が各クラスで同じ値をとると仮定することにより $P(x_i|\omega_i)$ の値で確率の大きさを判断する。

クラスごとの特徴パラメータ x_i は、音高 ω_i を含んだ音のデータのケフレンシ幅(p次元)分を主成分分析しm次元への圧縮を行うことによって得る。なお、対象となる音高をC4～B4の1オクターブとしているため、圧縮後の特徴パラメータもC4～B4に対して7種類生成される。

このようにして得られる x_i を用いて、確率 $P(x|\omega_i)$ は次式の正規分布で表す。

$$p(x_i\mid\omega_i)=\frac{1}{\sqrt{(2\pi)^d\left|\sum_i\right|}}\cdot \exp\left\{-\frac{1}{2}(x_i-m_i)^t\sum_i^{-1}(x_i-m_i)\right\} \quad (2)$$

m_i および $\sum_i$ はそれぞれクラス ω_i における特徴ベクトルの平均ベクトルおよび共分散行列であり、

$$m_i=\frac{1}{n_i}\sum_{x\in X_i}x_i \quad (3)$$

$$\sum_i = \frac{1}{n_i} \sum_{x \in X_i} (x_i - m_i)(x_i - m_i)^t \quad (4)$$

で与えられる。ここで n_i はクラス ω_i のパターン数、X_i はクラス ω_i のパターン集合を表す。入力データ中の和音数が既知の場合には,確率 $P(x_i|\omega_i)$ の大きいクラス ω_i を和音数分上位から選択することで音高認識を行う。

実験に用いたデータはすべて、MIDI音源を用いて16kHzサンプリングで録音したデータである。学習データに使用した音は、表1に示すように、認識対象範囲をC4~B4までとした全音階1オクターブである。

表1: 学習に用いた楽曲データ

音色	単音	和音
Flute	C4～B4の全音階7音	2和音C4G4, D4A4, E4B4の3音
Trumpet	C4～B4の全音階7音	2和音C4G4, D4A4, E4B4の3音
Violin	C4～B4の全音階7音	2和音C4G4, D4A4, E4B4の3音

これらの3音色の単音、和音合わせて計30個の音に対し、それぞれ時刻の違う6フレーム分、計180個のケプストラムを求め、構成音ごとに主成分分析を行って確率密度関数を決定し、音高のモデルを得た。ここで、ケプストラムを求める際のFFTデータ数は512、主成分数は3、ケプストラムの認識対象範囲はB3~C5に対応するケフレンシ37ポイント分を使用した。

表2: 音高認識の結果

音名	認識結果
単音C4～B4	全音認識成功
FluteC4E4G4	C4E4G4
FluteD4F4A4	D4F4A4
FluteE4G4B4	E4G4B4

和音数は既知とし、各音らしさの上位から和音数分認識結果とした結果を表2に示す。すべての評価データにおいて正しく認識されているおり、提案手法の有効性が示された。

4. 音声データの自動要約

情報技術の進歩によって、画像、音声、文字テキスト、さらにそれらを組み合わせたメルチメディア情報が大量に蓄積できるようになっている。情報コンテンツの表現においてマルチメディア化が進むにつれて、音声に代表される言情報を含むコンテンツも増大しており、講演や演説など音声言語が中心的な役割を果たすコンテンツも多い。今後も、文化的／歴史的に意義の大きい講演や演説が多数蓄積されていき、大学での授業や学会講演などの学術的コンテンツのディジタルカーカイブ化も進むものと思われる。蓄積された情報コンテンツのデータベースから欲しいデータを捜し出す時、一般に、蓄積されたデータ量が増えれば増えるほど、欲しいデータを捜し出すことが難しくなり、検索技術やコンテンツへのアノテーションが重要になってくる。

一つ一つのデータがどのような内容なのかを容易に把握できるように、これまでに、文字テキストに対する自動要約の研究が広く行なわれてきている。近年、連続音声認識の性能が向上したことにより、音声データに対する自動要約の研究も始まっている。音メディアで表現されたコンテンツは、文字テキストと比べてスキャニング（拾い読み／拾い聞き）が難しく、講演などの音声に対する要約自動生成に対する期待は大きい。捜し出した音声データが必要なものかどうかを判断するのに、要約された結果を聴いたり読んだりできれば非常に有用である。

音声データの自動要約は、連続音声認識とテキスト要約の単純な組合せによっても実現することが可能である。すなわち、音声を連続音声認識によって文字テキストに変換し、得られた文字テキストに対してテキスト要約を行ない要約結果を得る。しかし、このような処理は音声の持つ言語的な情報のみに注目しており、非言語的な（パラ言語的な）情報を無視されることになる。音声によるコミュニケーションでは、意図、感情、強調、

微妙なニュアンスなどの非言語的情報が韻律情報（声の高さ、声の大きさ、発話速度）によって表現されることがよく知られている。音声の要約でも、音声波形の持つ韻律情報を言語情報と併せて利用することによって、要約の精度を向上させられる可能性がある。そこで本研究では、講演音声を対象として、文境界を既知とした文抽出による音声要約において、言語的な情報に加えて韻律情報を利用することによって要約の精度向上を目指した。

本研究では言語情報と韻律情報を組み合わせて文の重要度を算出する方法を検討する。従って韻律情報だけでなく、文テキストからの言語情報の獲得が必要となる。本研究では公開されているテキスト要約システム Posumを用いた。

文の重要度を決定するために、韻律情報を利用する。時間長、パワー(声の大きさ)、基本周波数(声の高さ)に関して、文中の最小値、 最大値、レンジ、 平均値の4つのパラメータを文ごとに算出する。これに文の長さを加えて、合計13個の韻律パラメータを求め、これから有効なパラメータを選択し、」文の重要度の算出に利用する。文重要度の算出は重回帰モデルによって行った。

講演音声データとしては約10分のＮＨＫ論説番組「あすを読む」の5回分を用いた。講演音声の文単位への分割は、人手で行った。音声認識の性能評価、さらに重要文抽出における認識誤りの影響を分析するために、まず、人手でよる書き起こしテキストを作成した。さらに、言語情報を得るために文ごとに音声認識を行った。音声認識は、CSRC2001年度版のシステムを用いて行った。5つの講演音声データに対する平均単語認識精度は 64.6% であった。

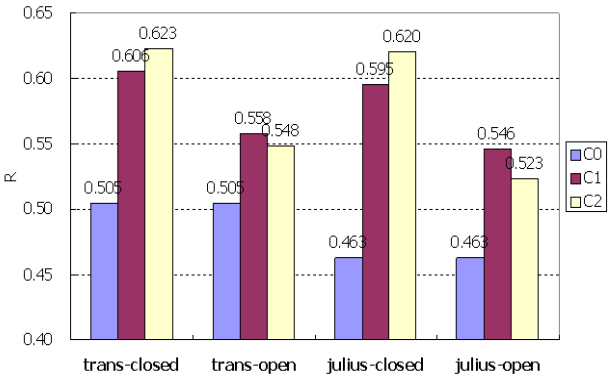

図2　重相関係数による文重要度予測の評価

重要文抽出による要約の精度を、人手による重要文抽出実験で決定した文重要度と自動決定した文重要度の相関によって評価する。評価結果を図2に示す。重相関係数は 0 から 1 の間の値をとり、値が大きいほどモデルによる現象の説明がうまくいっていることを示す。ここで trans- と CSR- はそれぞれ人手による書き起こしテキストと自動音声認識によるテキストから言語情報スコアを算出した場合を示しており、-closed、-open はそれぞれクローズド評価、オープン評価を示している。

図2においてC0は韻律情報を用いず言語情報だけで文重要度を決定した場合、C1とC2は韻律情報を用いた場合で、C1とC2では用いた韻律パラメータの数が異なる。この結果から、C0よりもC1とC2の方が重要関係数が大きくなっており、文重要度の予測がうまくいっていることを示している。さらに、C0、すなわち言語情報だけで重要度を決定する場合には、連続音声認識を用いることによって重相関係数がやや小さくなっている。これは、連続音声認識による認識誤りのために言語情報が劣化したためと考えられる。一方、韻律パラメータを利用したC1とC2では、連続音声認識を用いても重相関係数はそれほど変化しておらず、韻律パラメータを利用することによる効果は連続音声認識を使った場合の方が顕著であることがわかる。

5. まとめ

音メディアデータを対象とした高度なアーカイブシステムを実現するための要素技術に関して、擬人化エージェントを用いた音声対話システムの構築、楽曲データに対する音高認識、音声データに対する自動要約について述べた。今後は、能などの芸能音楽を対象として、アーカイブシステムの開発を進めていく予定である。