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プロジェクト

能楽を対象としたビデオコンテンツのタグ付け支援

山下洋一 ・ 中川隆広 ・ 岡田一貴
理工学研究科

概要 本報告では、芸能音楽、特に「能」に特化して設計された高度アーカイブシステムについて述べる。本アーカイブシステムは能に対するタグ付けを容易に行う環境を提供する。すでに、PC上で動作するプロトタイプシステムを開発している。また、ユーザフレンドリなインタフェースを実現するために、複数候補の提示に基づく新しい音声インタフェースの手法を提案した。利用者に提示する音声認識結果の数を認識結果に基づいて動的に制御する。評価実験によって提案手法が有効であることを示している。

Development of an advanced archive system for recording and reproducing music of performing arts project

A Support System for Annotating Video Data of "Noh" Performance

Yoichi Yamashita ・ Takahiro Nakagawa ・ Kazuki Okada
Graduate School of Science and Engineering

Abstract: This paper describes an advanced archive system of performing arts, which is specially designed for browsing the contents of "noh". It provides the facility of annotating events in the noh performance. The authors have developed a prototype system on PC and verified that it works well. A novel method of speech interface based on N-best speech recognition is also proposed to build user-friendly interface. The number of candidates of recognition results which are shown to the user is dynamically determined based on the distribution of speech recognition scores of N-best recognition. The evaluation test showed the effectiveness of the proposed method.

1. はじめに

計算機システムの性能向上に伴い、様々なデータがデジタル化され大量に保存されるようになっている。膨大なデジタルデータから欲しいデータを容易に取り出すには、検索に有用なタグ(メタデータ)をデータに付与し、効率よく簡単に閲覧できるシステムが必要となる。本研究テーマでは、能楽を収録したビデオデータを具体的な対象として取り上げ、タグ情報を作成するための支援システムの開発と、ユーザフレンドリなインタフェースを実現するための音声認識について研究を行っている。本報告では、構築中のタグ付け支援システムと複数候補を提示する音声認識インタフェースに関して述べる。

2. タグ付け支援システム

2.1 システムの設計

様々なマルチメディアデータのタグ付けに利用可能な一般的なタグ付けシステムではなく、能のタグ付けに特化させることで効率よくタグ付けの行えるシステムの開発を目指す。そのために、能の特徴を考察し必要な機能を検討する。能では、

(1) 主役(シテ)、相手役(ツレ)などのように、数種類の役がある。
(2) 囃子方(鼓などの奏者)には、笛方、小鼓方、大鼓方、太鼓方という分類がある。

(3) その他に、ワキ方、狂言方といった役がある。
(4) 詞章と呼ばれる台本のようなものがある。
といった特徴が挙げられる。その他に考慮すべき点として、
(1) 台詞ごとに細かくタグを付ける必要がある。
(2) 映像を停止した状態でも細かいタグ付けの作業を可能にする。
などがある。このような能の特徴と、考慮すべき点を踏まえた上で以下のような機能を実現する。
(1) 能には詞章と呼ばれる台本のようなものがあり、それを基に舞台が演じられるので、詞章に対してタグ付けを行う。
(2) 能には主役(シテ)、相手役(ツレ)などのように決まった分類方法と呼び方が十種類あり、タグ付け結果の分類を明確にするためにはタグの名前に統一性を持たせることが必要なので、それぞれをボタンとして配置する。
(3) 実際にタグ付け作業を行う際に、台詞等の発話時間を細かく指定する必要があるので、映像と同期させた音声波形を表示し、映像を停止させた状態で波形の下に付いているスクロールバーを動かすことでポイントを移動させ、タグ付けを行うことを可能にする。

開発言語はBorland C++ Builder 5 を使用し、動画の表示にはMPEG1-LayerⅡ を、波形表示にはwavファイルを、詞章の読み込み、保存にはテキストファイルを使用する。

2.2 システムの機能

2.2.1 詞章の読み込み

図2.1のように行単位で区切られて記述された詞章を入力データとして読み込む。行をクリックすることで行を指定し、その行に対するタグ付けを行う。

2.2.2 役名の選択

各せりふの役名は図2.2に示す役名ボタンを利用して決定する。図2.1のように詞章の行を指定した上で、役名のボタンをクリックすることで図2.3のように詞章におけるせりふの行頭に役名が付加される。図2.3の場合は「シテ」が役名としてタグ付けされている。役名の右にある数字はタグが示している範囲で、左が始点、右が終点の時刻である。また、別の役名のボタンをクリックすることで役名が変更される。

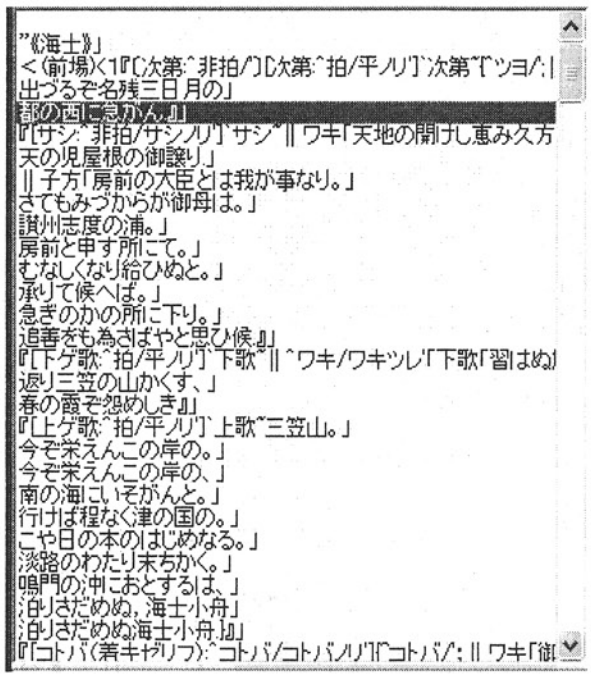

図2.1 詞章の例

図2.2 役名選択ボタン

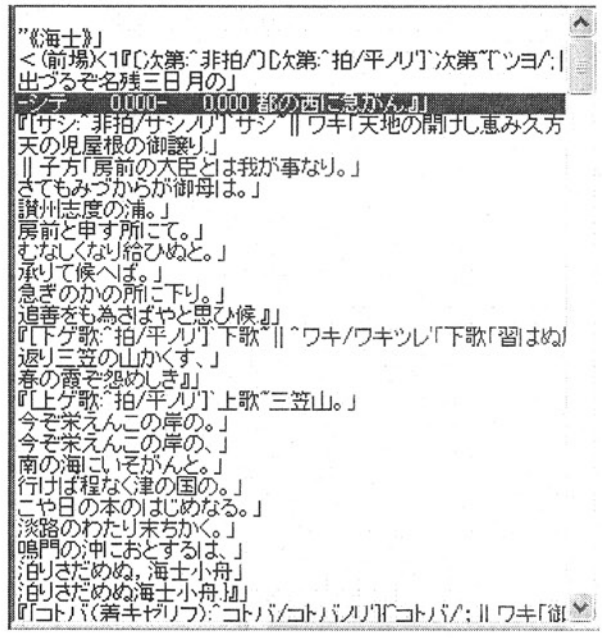

図2.3 役名が付加された詞章の例

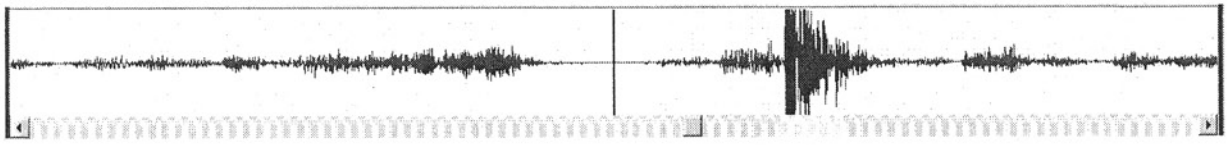

図2.4　音声波形の表示例

2.2.3 音声波形の表示

一つの能の演目は数十分に及ぶため、PCにおけるメモリの制限から、音声波形データをすべて一度にメモリに読み込むことは難しい。できる限りPCの負担を減らすために、現在の表示ポイントの前後数秒間分のデータだけを読み込み表示させる作業を100msごとに行うことで視覚的にも問題のない処理を実現している。

また、波形表示はwavファイルを使用しており映像の表示方法（MPEG1-LayerⅡ）と異なるため、波形と映像のタイミングを同期させる必要がある。そのために、wavファイルのヘッダーファイルからサンプリング周波数を読み込み、wavファイル全体のデータ量を計算することで演目全体の時間長を算出する。映像全体のフレーム数から単位フレーム数でのwavファイルのサンプル数を算出し、ポイントを合わせることで波形と映像の同期を図っている。

実際の波形表示画面は図2.4のようになっており、中央にある黒色の線が現在のポイントを示している。映像を停止させた状態で下部にあるスクロールバーを左右に動かすことで、映像と同期を取りながら表示するデータの時刻が変更できる。

2.3 システムの実装

2.3.1 実装時の問題点と解決方法

当初は波形表示のためにPaint Box関数を使用しており、映像と同期させて波形を表示させる際に高速で一度背景の色で塗りつぶして新たな

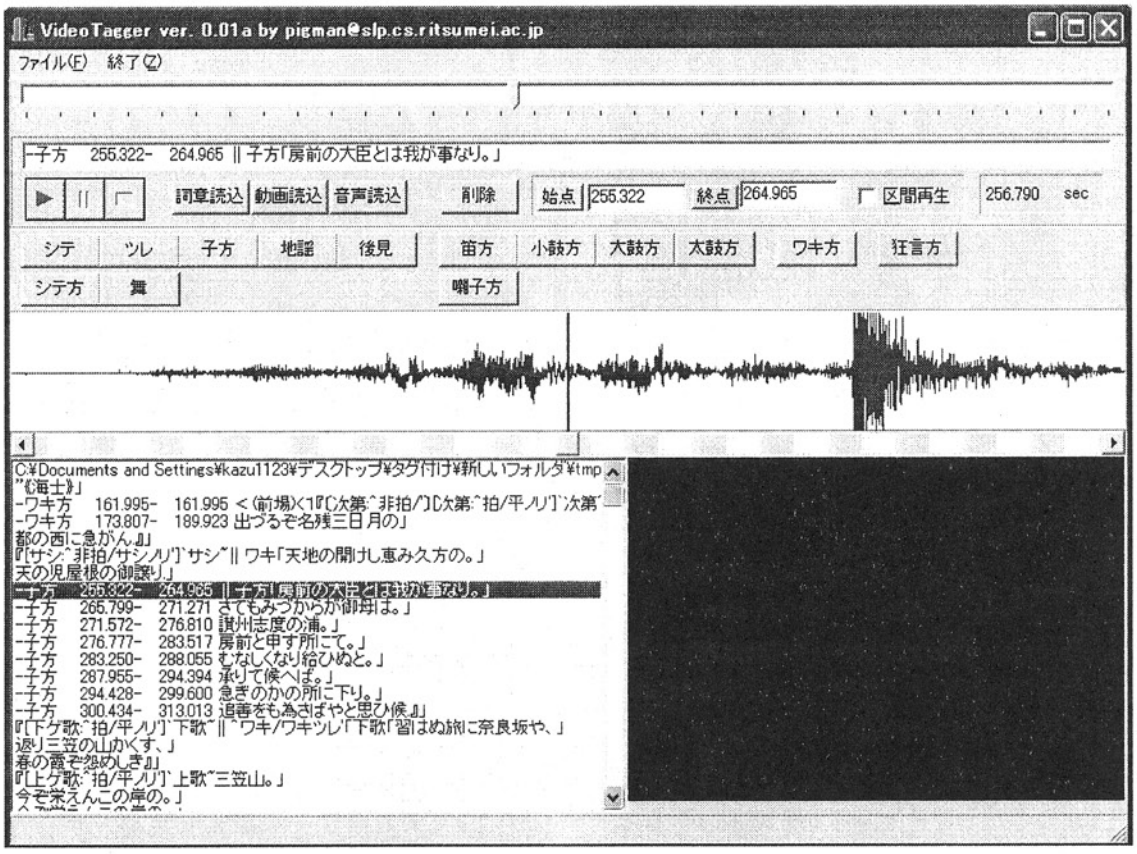

図2.5　タグ付け支援システムの動作画面

波形を書き直すという方法をとっていたため、映像の再生時に波形表示画面が綺麗に表示されないという問題が生じた。また、図形表示関数であるImage関数を利用した実装では、映像再生時は綺麗に表示されるが、映像停止時に波形表示画面下のスクロールバーを動かした時に波形の表示に遅延が生じる問題が生じた。

この問題を解決するために、映像再生時にはImage関数を、停止時にはPaint Box関数を使うことで、映像再生時は波形を綺麗に表示することができ、停止時にもスムーズな作業を行うことができるよう改良された。

2.3.1 システムの動作画面

2.2節で述べた機能を実装したシステムの動作画面を図2.5に示す。動作時の特徴として、

(1) 実際には図2.5の右下の黒い部分に映像が表示され、波形表示部中央の黒い縦線と同期している。
(2) 映像再生時には、波形表示画面下のスクロールバーは固定される。
(3) タグ付け後の詞章をクリックし、再生ボタンを押すことでタグの始点のポイントから再生される。
(4) タグ付け結果が変更された状態で、終了ボタンをクリックすると確認メッセージが表示される。

といった点が挙げられる。

2.4 タグ付けの手順

2.3節で述べたタグ付け支援システムを利用して、タグを付与する作業では、

① 詞章読み込み、動画読み込み、音声読み込みの3つのボタンからそれぞれ詞章データ、映像データ、音声データを読み込む。
② 再生ボタンを押す。
③ 映像と波形を見ながら、詞章に記されている部分を決定する。
④ 停止ボタンを押し、波形のスクロールバーを動かし始点ボタンと終点ボタンを押し、タグの始点と終点を決定する。
⑤ 役名ボタンを押し、役名を決定する。
⑥ ①から⑤を繰り返す。
⑦ 「ファイル」の中にある「名前を付けて保存」を選択し、詞章のテキストファイルを保存する。

の手順で作業を行う。

2.5 タグ付け結果のファイル形式

タグ付け結果のファイルの一部を図2.6に示す。1行目には映像ファイルの場所が示されており、1行空けた後に能の題目が、さらに次の行からタグ付け結果が保存されている。タグ付けされた行はスペースで区切られており左から順に、

役名　タグの始点　タグの終点　詞章データ

となっており、タグの始点、終点は秒単位で表現される。

3. 複数候補を提示する音声認識インタフェースの検討

近年の音声認識技術の向上に伴って、音声認識を用いた対話インタフェースへの期待が高まっている。しかし、現状の音声認識技術では、正しい結果が常に得られるとは限らないことから、複数の認識結果を提示し、利用者に候補の中から正解を選択してもらう N-best 方式に基づいた音声認識インタフェースが検討されている。

N-best の音声認識結果に基づいて複数の正解候補を利用者に提示する場合には、通常、提示する候補数をあらかじめ決めておく。提示する候補数を多くした方が、少ない場合に比べて候補中に正解が含まれている確率が高くなる。しかし、多くの候補の中から正解を探す必要があるため、手間がかかってしまう。提示される候補数が少なく、かつ、候補中に正解が含まれていることが理想的である。このように、候補中で正解が含まれる割合を減少させずに提示する候補数を減らすには、あらかじめ候補数を決めておくのではなく、認識結果に基づいて候補数を動的に決定する必要がある。

そこで、N-best 候補の認識スコアを分析し、認識スコアを利用した候補提示数の決定手法について検討した。

C:¥Documents and Settings¥kazu1123¥デスクトップ¥デモ用¥デモ用.mpg

"《海士》」
-シテ　　11.345-　　14.047 ＜(前場)<1『〔次第:ˆ非拍/'〕[次第:ˆ拍/平ノリ'〕`次第~{ˆツヨ/'; ‖ ˆワキ/ワキツレ'「出づるぞ名残三日月の。」
-シテ　　19.419-　　22.689 出づるぞ名残三日月の」
-シテ　　30.330-　　32.833 都の西に急がん.』」
-ワキ方　　49.283-　　55.556『[サシ:ˆ非拍/サシノリ'〕`サシ~ ‖ ワキ「天地の開けし恵み久方の。」
-ワキ方　　69.469-　　70.537 天の児屋根の御譲り.」
-ワキ方　　80.714-　　89.656 ‖ 子方「房前の大臣とは我が事なり。」
-ワキ方　　116.416-　　152.085 さてもみづからが御母は。」
-ワキ方　　160.994-　　165.465 讃州志度の浦。」
-子方　　174.374-　　192.226 房前と申す所にて。」
-子方　　201.134-　　218.986 むなしくなり給ひぬと。」
-子方　　223.457-　　232.366 承りて候へば。」
-子方　　250.217-　　272.506 急ぎのかの所に下り。」
-子方　　290.357-　　308.175 追善をも為さばやと思ひ候.』」
-ワキ方　　343.877-　　370.637『[下ゲ歌:ˆ拍/平ノリ'〕`下歌~ ‖ ˆワキ/ワキツレ'「下歌「習はぬ旅に奈良坂や、」

(以下略)

図2.6 タグ付け結果の例

3.1 認識スコア分布の分析に基づいた結果の提示

N-best 候補の認識スコアの例を表 3.1 に示す。「この研究室の歴史を知りたい」と発話し、音声認識システムを用いて音声認識を行った結果である。音声認識では、このように文ごとに認識スコアが得られる。広く利用されている統計的手法に基づいた音声認識では、文のスコアは確率に基づいて非常に微小な値として算出されるため、表 3.1 では対数をとって負の数としてスコアを表記している。

研究室のホームページ検索を行う、語彙数の異なる三つのタスクの文章に対して音声認識を行い、認識スコアの分布を分析した結果から、

(1) 第(n+1)位の候補のスコアが、第 n 位の候補のスコアに比べて差が大きい場合には、第(n+1)位以下の候補が正解になることは少ない。
(2) 第 n 位の候補のスコアが、第 1 位の候補のスコアに比べて差が大きい場合には、第 n 位以下の候補が正解になることは少ない。
(3) 第 n 位の候補のスコアが小さい場合は、第 n 位以下の候補が正解になることは少ない。

の三つのヒューリスティックスを得た。

N-best 認識で得られた N 個の認識結果のうち、利用者に提示する結果を上位から順にどの候補までにするかを上記三つのヒューリスティックスに基づいて決定する手法を提案した。

表 3.1: 認識結果の例

n-best	recognition result	recognition score
1	この研究室 の 歴史 が 知りたい	-26.160279
2	この研究室 の 歴史 を 知りたい	-26.161865
3	この研究室 の 研究 を 知りたい	-26.332994
4	この研究室 の 研究 が 知りたい	-26.398474
5	この研究室 の 歴史 が 聞きたい	-26.442401
6	この研究室 の 歴史 を 聞きたい	-26.443183
7	この研究室 の 歴史 が 知りたい です	-26.460549
8	この研究室 の 歴史 を 知りたい です	-26.462135
9	この研究室 の 先生 を 知りたい	-26.471869

The correct answer: この研究室 の 歴史 を 知りたい

3.2 手法の評価

提案手法の有効性を検証するために、連続単語の認識に適用した。タスクは人名検索である。単語辞書に日本人の上位姓(苗字)と人気上位名前(男性)を用いて語彙数を変えて3つのタスクで実験を行った。タスク 4, 5, 6 での単語辞書の語彙数はそれぞれ 710, 840, 1000 単語である。また、タスク 4, 5, 6 のテストセットパープレキシティはそれぞれ 257, 313, 400 であった。

被験者 5 人(男 3 人, 女 2 人)に 20 文ずつ発話してもらい、この合計 100 文をタスク 4, 5, 6 の文法で認識した時の 30-best の認識結果に提案手法を適用した。平均候補提示数と正解提示率をそれぞれ表 3.2(a)と(b)に示す。30-best で得られた認識結果をすべて提示する場合に比べて、提案手法では、正解提示率がやや下がっているものの、提示候補数が劇的に減少していることがわかる。ここで、用いた音声認識システム julian では、認識結果を 30 個作成する 30-best の認識においても、発話文によっては必ずしも 30 個の結果が得られないため、30 個固定の場合でも平均提示候補数が 30 よりもやや小さい値となっている。

さらに、インタフェースにおける提案手法の有効性を検証するため、

method1: 常に1個の候補を提示する。
method2: 常に30個の候補を提示する。
method3: 提案手法に基づいて提示候補数を決定する。

表 3.2: 平均提示候補数と正解提示率

(a) 平均提示候補数

	提示数決定手法	
	30個固定	提案手法
タスク4	27.4	2.2
タスク5	25.9	2.4
タスク6	25.7	2.5

(b) 正解提示率 [%]

	提示数決定手法	
	30個固定	提案手法
タスク4	95	93
タスク5	91	89
タスク6	84	82

の三つの音声認識インタフェースを作成し、実際に被験者に利用してもらい、正解を選択されるまでの所要時間と発話回数を計測した。なお、候補提示数の変化による使いやすさの違いを見るため、自由な発話ではなく、決められた文を発話してもらった。被験者の発話をマイクで入力して音声認識を行い、表示された候補の中から被験者が正しい認識結果をマウスで指示するようにした。もし表示された N-best 候補の中に正しい認識結果が含まれていない場合は何度でも同じ文を発話してもらうこととした。 method1, 2,

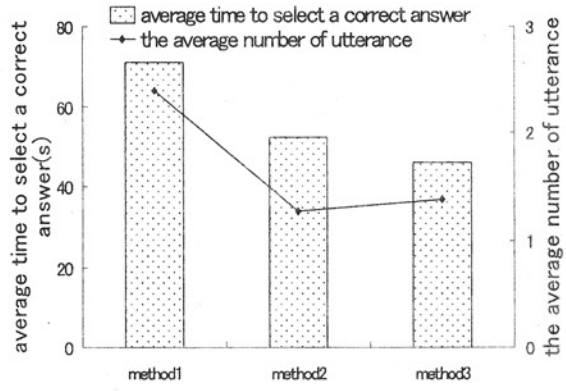

図3.1 正解選択までの平均所要時間と平均発話回数

3 での実際に提示された候補数の平均はそれぞれ、1.0, 28.0, 4.8 であった。結果を図 3.1 に示す。これより、平均発話回数は method2 がもっとも少なかったが，正解選択までの平均所要時間は method3(提案手法)がもっとも短くかかったことがわかる。

4. おわりに

能楽を対象としたタグ付け支援システムのプロトタイプシステムを作成した。今後、タグ付けの作業を行い、音声インタフェースの組み込みも含めたシステムの改良を行っていく。さらに、タグ付けされたデータをもとに、ビデオデータと詞章の対応付けの自動化を検討し、タグ付けの自動化／半自動化を実現することを目指す。

謝辞

本研究を行うにあたり御協力いただいた赤間亮教授、重田みち氏、アートリサーチセンターの皆様に感謝する。