能楽ビデオデータに対するタグ付けの自動化

岡田 一貴[1] 高橋清人[2] 山下 洋一[3] 重田 みち[4] 赤間 亮[5]
[1]立命館大学大学院 理工学研究科
[2]立命館大学 理工学部
[3]立命館大学 情報理工学部
[4]立命館大学 21 世紀 COE 推進機構
[5]立命館大学 先端総合学術研究科 / アートリサーチセンター

マルチメディアデータのアーカイブでは，単にデータを蓄積するだけでなく，適切なタグ情報（メタデータ）を付与することによって，検索や閲覧での利便性を高めることが重要である。本報告では，能楽のビデオコンテンツを対象としたタグ情報の自動生成について述べる。能ビデオの音声データをもとに演者の発話区間を自動検出し，あらかじめ作成されている台詞との対応付けを自動化することによって，能楽の台本である詞章に時間情報を付与する。発話区間の自動検出では音声の帯域分割とクラスタリングに基づいた手法を，台詞との対応付けでは発声速度で正規化された発声時間長と台詞長に基づいた手法を提案する。

A Study of Automatic Annotation for Noh Performance Video

Kazuki Okada[1] Kiyohito Takahashi[2] Yoichi Yamashita[3] Michi Shigeta[4] Ryo Akama[5]

Graduate School of Science and Engineering, Ritsumeikan University[1]
Science and Engineering, Ritsumeikan University[2]
Information Science and Engineering, Ritsumeikan University[3]
Research Center of 21th COE, Ritsumeikan University[4]
Graduate School of Core Ethics and Frontier Sciences, Ritsumeikan University[5]

Annotating with useful tag information (meta-data) is very important to facilitate accessing and browsing multimedia data as well as collecting data. This paper describes a method of automatic generation of tag information annotating "Noh" performance video. It provides "shisho", which is a script for Noh performance, with time information using two-stage processing. In the first stage, a clustering technique automatically extracts spoken documents from the audio track data using feature parameters of the filter bank. The second stage aligns word sequences in the script with the extracted spoken segments based on matching the word length in the script with the length of spoken segments that is normalized by the speaking rate.

1.はじめに

近年、遺跡の劣化、書物など文化財の老朽が懸念される中で、デジタルアーカイブが注目されている。デジタルアーカイブとは、散在している有形・無形文化財、膨大な遺跡、自然環境などをデジタル映像や文書として記録・保管したものであり、様々な文化財・遺跡・書物などの永久的保存に役立つ手段として注目されてきている[1]。デジタルアーカイブ化をするにあたり、データの管理や容易な閲覧・研究のために、データに対してタグ（メタデータ）を付けることが好ましい。しかしながら、膨大なデータに対してすべて手作業でタグやメタデータを付けることは非常に困難であり、自動化が求められている。

様々なデータを対象にしたデジタルアーカイブの研究が進められる中、21 世紀 COE プロジェクトのひとつのプログラムである「京都アート・エンターテイメント創成研究」が生まれた。これは、京都を核とする日本の有形・無形文化・芸術を、最先端情報技術を応用して、コンテンツをデジタルアーカイブ

化するプロジェクトである[2]。本報告では、古典芸能である能に焦点を当て、能楽ビデオデータに対するタグ付けの自動化について述べる。

2. 能楽ビデオデータに対するタグ付け

2.1能楽

能楽とは、能舞台といわれる特別な舞台上で、地謡や囃子の演奏に合わせ、演者が舞い謡うという様式を取る[3][4]。能楽において舞台を演出する構成員は、舞い、又は謡いを担当する役柄として、シテ、ツレ、ワキ方、子方、地謡、後見といわれる演者と、楽器演奏を担当する役柄として囃し方と言われる奏者となっている。囃し方においては、笛方、小鼓方、大鼓方、太鼓方と各楽器毎に役柄が設定されている。

能楽には、舞台演出を記した台本として「詞章」があり、演者の台詞、囃し方の演奏する目安となる記号が記されている。図1に演目「海士」の詞章の一部を示す。

2.2 詞章をもとにしたタグ付け

タグ付けとは、対象となるファイル、またはファイルの一部に属性情報を付与することである。例えば、ファイルが動画である場合、一部の場面に対し、どのような特徴を持つか、その特徴を持つ場面が、いつ始まり、いつ終わるのかという時間情報などの属性情報を付与することが考えられる。

能楽のビデオデータでは、図 1 に示す未編集の詞章に対し、図 2 のように役柄情報、時間情報を付与することにより、必要な場面の検索や閲覧したい場面の再生に非常に有用な情報を与える。図 2 では、詞章の本文に対し、役柄の名称、開始フレーム位置と終了フレーム位置が付与されている。

……
讃州志度の浦。」
房前と申す所にて。」
むなしくなり給ひぬと。」
承りて候へば。」
急ぎのかの所に下り。」」
……

図 1：詞章の例

……
-シテ 100-1300 讃州志度の浦。」
-シテ 1320-2333 房前と申す所にて。」
-ワキ 2590-3455 むなしくなり給ひぬと。」
-笛方 3599-3700 承りて候へば。」
-ツレ 3800-3950 急ぎのかの所に下り。」
……

図 2：タグ付与された詞章の例

3.タグ付け支援システム

能楽データを対象として、タグ付け処理を支援し、タグ付けデータをもとにコンテンツを閲覧するGUI環境をこれまでに開発してきている[5]。開発言語は C++ 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している。

システムの動作画面及び各部位名称を図 3 に示すと共に以下に詳細を示す。

(1)時間表示用トラックバー

ビデオファイルの時間情報の全体を示す。

ドラッグすることで、再生位置を変更することも可能となっている。主に再生位置の決定に使われる。

(2)詞章表示部

詞章一覧から選択した詞章が表示される。

リストから別表示することで、タグ付け処理の際の見易さを考慮して配置されている。

(3)ツールボタン郡

各ファイル読み込みの為のボタン、役柄と時間情報を設定する為のボタン、ビデオファイル操作の為のマルチメディアボタンが配置されている。

(4)音声波形表示部

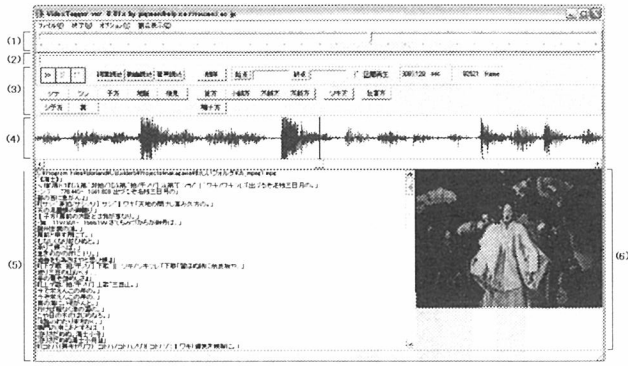
読み込まれた音声ファイルを視覚化し、音声波形として表示する。

(5)詞章一覧表示部

読み込みボタンを使用し、読み込まれた詞章のテキストファイルがリスト表示される。タグ付けを行う際に、このリスト郡からタグ付け対象候補をクリックすることで選択する。

(6)ビデオデータ表示部

読み込まれたビデオファイルが表示される。ビデオファイルの表示サイズに合わせて、自動で表示部の大きさも変わる



(1)時間表示用トラックバー
(2)詞章表示部
(3)ツールボタン郡
(4)音声波形表示部
(5)詞章一覧表示部
(6)ビデオデータ表示部

図3.タグ付け支援環境動作画面と各名称

4.タグ付けの自動化

3章で示したタグ付けシステムを用いてタグ付け作業を行うことはできるが、能のすべての演目に対して手動でタグ付けを行うことはタグ付け作業者に大きな負担を強いる作業となる。そこで、タグ付け作業の自動化をすることを目指す。本システムにビデオファイル、詞章ファイルを読み込むと同時に、詞章に時間情報を自動的に付与することを目指している。

タグ付けの自動化の手順は図4のようになっており、(1)発話区間の自動抽出と(2)詞章と発話区間のマッチングの2段階の処理で構成される。以下にそれぞれの処理を説明する。

(1)発話区間の自動抽出

詞章に対して時間情報を自動的に生成するためには、まず詞章に対応する発話部分の抽出を行うことが必要がある。本報告では発話している区間を「発話区間」、それ以外の区間を「非発話区間」と呼ぶ。タグ付け自動化では、まず能楽の音データを自動的に発話区間と非発話区間に分類する。

(2)詞章と発話区間のマッチング

詞章に時間情報を付与するには、(1)で抽出された発話区間に対して、詞章中の各台詞と結びつける必要がある。ここでは、詞章と発話区間のマッチングを行い、詞章に対応した発話区間の時間情報を付与する。

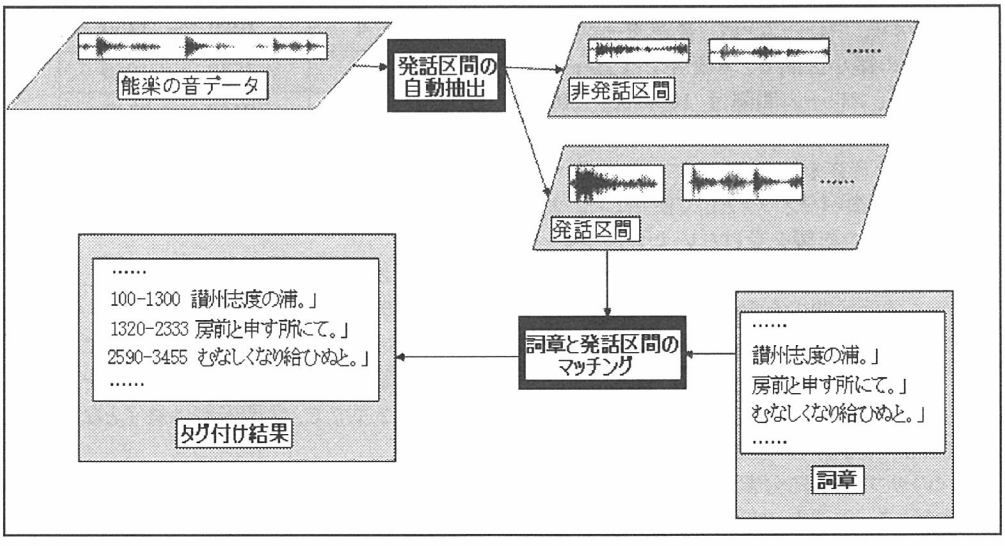


図4:タグ付けの流れ

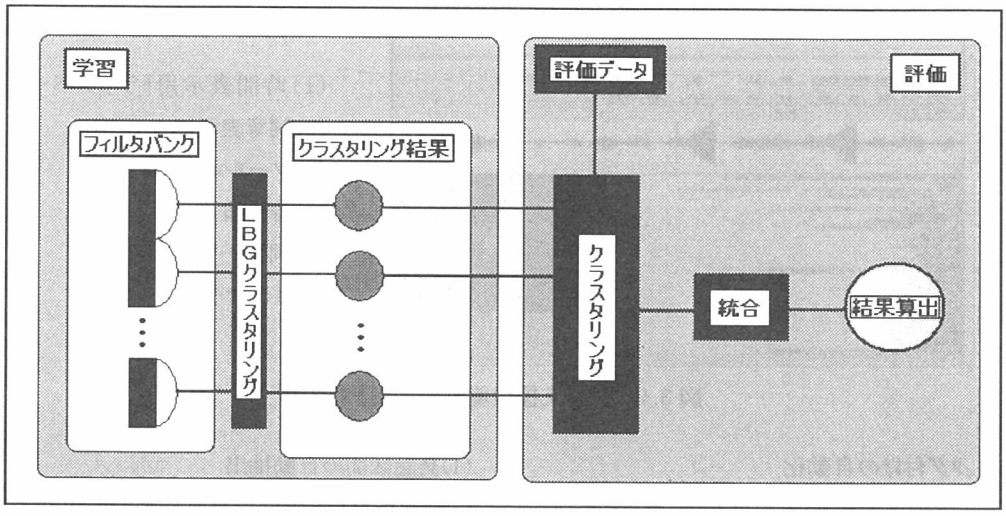


図5:発話区間の自動抽出の流れ

5.発話区間と非発話区間の自動分類

4章で示した発話区間の自動抽出を行うにあたり、発話区間と非発話区間はそれぞれの類似した性質を持つことが推測できるため、音データ全体をクラスタリングによって、発話区間と非発話区間に自動分類することを試みる。

5.1 評価実験

5.1.1 手法

フレーム単位で分割された音データをクラスタリングする。フレームとは、音声データを分析するための微小区間で、本研究では20[msec]とした。また、フレーム間隔は 10[msec](100[フレーム/sec])とした。また、発話と同時に演奏される笛や鼓の音声を抑えるために、フィルタバンクの出力を特徴パラメータとして使用した。また、音声の大小の影響を受けないように、フレーム毎の値の平均が0になるように正規化する。

図5に発話区間の自動抽出の流れを示し、以下に手順を示す。

＜手順 1＞

100Hz～5kHz の周波数区間をメルスケールで24のサブ帯域に分割し、メルフィルタバンクを構成する。周波数の低い方から3個ずつのサブ帯域のフィルタ出力を順にまとめて24個のフィルタ出力を表1に示す8つの帯域に分割する。各帯域ごとにクラスタリングを行い8種類のクラスタリング結果を得る。クラスタリング手法として、LBG アルゴリズム[6]を用いている。

表1:分割した帯域の周波数

帯域	周波数(Hz)
1	100 ～ 250
2	330 ～ 530
3	640 ～ 880
4	1020 ～ 1350
5	1530 ～ 1940
6	2180 ～ 2720
7	3020 ～ 3710
8	4100 ～ 5000

＜手順 2＞

学習データのラベル情報をもとに、それぞれのクラスタにおいて、そのクラスタ中に含まれる発話区間の割合(クラスタ発話確率)を求める。

手順 2 までで、学習過程は終了となる。

＜手順 3＞

i 番目の帯域のクラスタリング結果に対し、評価データをフレーム毎にもっとも類似するクラスタを決定し、そのクラスタにおけるクラスタ発話

確率をS_iとする。式(1)、(2)でそのフレームのフレーム発話確率S、フレーム非発話確率NSをそれぞれ求め、SとNSを比較し、Sが大きければ発話フレーム、そうでなければ非発話フレームとする。

$$S = \prod_{i \in C} S_i \qquad (1)$$

$$NS = \prod_{i \in C} (1 - S_i) \qquad (2)$$

ここでCはフレーム発話確率を求めるのに用いる帯域の集合である。

5.1.2 使用データ

評価実験では、演目「海士」の収録データを用いた。学習データとして、開始から500[sec]までの500[sec]の区間(50000フレーム)を用いた。評価データとして500[sec]から公演終わりまで(5224[sec])の4724[sec]の区間(472400[フレーム])を使用した。

5.1.3 実験結果

クラスタ数を2〜1024まで変化させたときに、正しく抽出された発話区間の割合を精度と検出率で求め、F値を計算する。

$$\text{精度} = \frac{\text{正しく抽出された発話フレーム数}}{\text{発話フレームとして抽出したフレーム数}} \times 100 \qquad (3)$$

$$\text{検出率} = \frac{\text{正しく抽出された発話　フレーム数}}{\text{発話フレームの総数}} \times 100 \qquad (4)$$

$$\text{F値} = \frac{2 \times \text{精度} \times \text{検出率}}{\text{精度} + \text{検出率}} \times 100 \qquad (5)$$

図6に実験結果を示す。フレーム発話確率の算出に用いる帯域の集合Cとして以下の2種類を試みた。

$C1$：帯域1〜8

$C2$：帯域2, 3, 4, 6

ここでの帯域の番号は表1に示した各帯域を表す。$C1$では全帯域を使用する。$C2$では、発話の特徴が現れやすく、笛・鼓の音が抑制されやすい4つの帯域を使用する。また、比較のために、帯域の分割を行わず、24個のフィルタバンク出力を一度に使用した場合についても発話区間の抽出精度を調べた。これを便宜上$C0$と表記する。

すべてを一度に使用した結果$C0$に比べ、帯域に分割した$C1$の結果のほうが全てのクラスタにおいて精度が上回った。さらに、発話の特徴が現れやすい帯域を使用した$C2$では、クラスタ数8においてF値が$C1$に比べ4.6向上した。

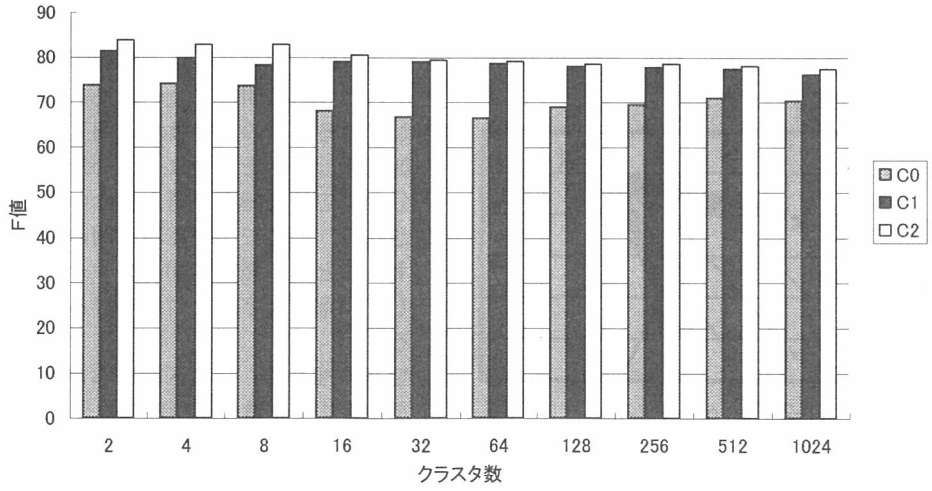


図6：発話区間の抽出結果

6.詞章と発話区間の対応付け

能楽ビデオデータに対するタグ付けを自動化するにあたり、発話区間と非発話区間を自動分類した後、詞章と発話区間とのマッチングを行う必要がある。能楽では発声方法が通常と大きく異なり、また発話と楽器演奏が同時に行われる場面が多数存在するため、音声認識による発話内容の自動決定は非常に困難である。そこで、発話区間のフレーム長と詞章の文字数から算出したフレーム長を利用して、長さ情報のみを利用した対応付けを行う。長さの対応付けにはDPマッチングを用いる。

6.1DPマッチングの導入方法

2つの入力系列を
発話区間のフレーム長:

$$a(I) = (a_1, a_2, a_3 \cdots a_i \cdots a_I) \tag{3}$$

台詞のフレーム長:

$$b(J) = (b_1, b_2, b_3 \cdots b_j \cdots b_J) \tag{4}$$

とする。音データ中の一文字あたりのフレーム数Nは、

$$N = \frac{\text{音データ中の総発話フレーム数}}{\text{詞章の全文字数}} \tag{5}$$

で表されるのでj番目の台詞のフレーム長は、

$$b_j = j\text{番目の台詞の文字数} \quad \times N \tag{6}$$

で表すことができる。また、DP パスは以下の図7のように設定する。

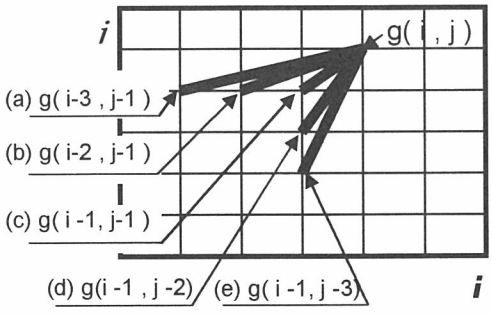


図7：DPパスの設定

DPマッチングにおける部分最適距離 $g(i,j)$は

$$g(i,j) = \min \begin{pmatrix} g(i-3,j-1) & + & d_1(i,j) \\ g(i-2,j-1) & + & d_2(i,j) \\ g(i-1,j-1) & + & d_3(i,j) \\ g(i-1,j-2) & + & d_4(i,j) \\ g(i-1,j-3) & + & d_5(i,j) \end{pmatrix} \tag{7}$$

となり、図7において、
(a)のパスを通る場合：

$$d_1(i,j) = \left|(a_i + a_{i-1} + a_{i-2}) - b_j\right| \tag{8}$$

(b)のパスを通る場合：

$$d_2(i,j) = \left|(a_i + a_{i-1}) - b_j\right| \tag{9}$$

(c)のパスを通る場合：

$$d_3(i,j) = \left|a_i - b_j\right| \tag{10}$$

(d)のパスを通る場合：

$$d_4(i,j) = \left|a_i - (b_j + b_{j-1})\right| \tag{11}$$

(e)のパスを通る場合：

$$d_5(i,j) = \left|a_i - (b_j + b_{j-1} + b_{j-2})\right| \tag{12}$$

である。

6.2 話速の推定に基づいた発話区間長の補正

能楽では、各台詞の話速が均等でないことから、発話区間のフレーム長と実際に発声している台詞の文字数が比例しないため、発話区間の長さをそのまま用いたのでは、十分な対応付け精度が得られないことが予想される。そこで、各発話区間の話速を推定することで、発話区間のフレーム長を補正して詞章の文字数と対応づけることを考える。

話速の推定は、発話区間におけるスペクトルの平均変化量に基づいて行う。i 番目の発話区間における k 番目のフレームのスペクトルを $\mathbf{S}_i$ としたとき、i 番目の発話区間のスペクトル平均変化量 V_i を

$$V_i = \frac{1}{N_i - d} \sum_{k=1}^{N_i - d} \left| \mathbf{s}_i(k) - \mathbf{s}_i(k+d) \right| \tag{13}$$

として求める。ここで、N_i は i 番目の発話区間のフレーム数を表す。スペクトル差の平均値である V_i の値が小さいなら話速は速く、逆に V_i の値が小さいなら話速は遅いと考えられる。そこで、この V_i を各発話区間の時間長に乗じることにより発話区間長を補正する。今回の実験ではフレーム長を 20msec、フレームシフトを 10msec、d を 10 とした。

6.3 対応付け実験

6.3.1 使用データ

演目「海士」の初めから 33 番目までの詞章と、初めから 30 番目までの発話区間を使用した。1番目の詞章と発話区間が、33番目の詞章と30番目の発話区間がそれぞれ対応している。なお、発話区間の抽出は手作業で行った結果を用いる。

また、1フレーム 20msec とし、非発話区間が 100msec 以上続く区間を発話区間の境界として使用している。

6.3.2 実験結果

話速推定に基づいた発話区間長の補正を行わない場合と補正を行った場合の実験結果を図8に示す。発話区間長の補正をしない場合、33個の詞章の内完全に発話区間と対応付けした台詞は8個に止まったが、正解の対応付けとは最大で2発話区間だけ離れており全体的に正解との差は狭いと考えられる。話速の推定に基づいて発話区間長の補正をした場合、33個の詞章の内、発話区間と完全に対応付けした台詞は9個になった。対応付け全体を見てみると、正解の対応付けとの差は発話区間長を補正しない場合に比べて、良くなっている部分と悪くなっている部分が見られる。

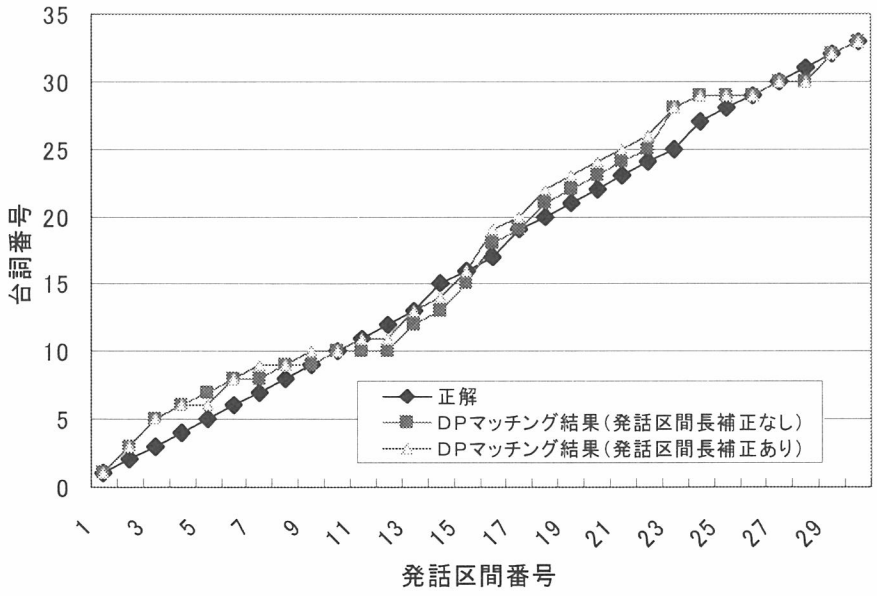


図8: 発話区間と詞章の対応付け結果

6.3.3 スペクトル差の平均と話速の相関

DP マッチングの対応付けで発話区間と正しく対応付けされた台詞が 9 個に止まった理由として、話速の推定において実際の話速とスペクトル差の平均値との相関がとれていないためだと考えられる。話速との相関を調べた結果を図 9 に示す。スペクトル差の平均値と話速との相関がとれているなら、発話区間を示すグラフ上の点が傾きが正の直線上に集まるはずであるが、広い範囲に分布しており、相関が不十分であると思われる。

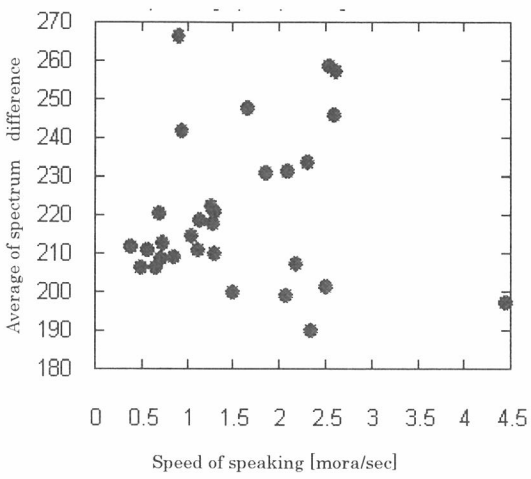


図 9 ： 発話区間のスペクトル差の平均値と実際の話速との相関関係

7. まとめ

能楽データを対象としたタグ付け作業の自動化を目指して、詞章に対する時間情報の自動決定について検討した。

まず、能の公演の音データをもとに、クラスタリングによって人の声の含まれている区間とそれ以外の区間との自動分類を試みた。さらに、抽出された発話区間と詞章とを対応付けるために、話速により修正された発話区間長と台詞の長さをDPマッチングにより対応付ける手法を試みた。今後は、発話区間抽出に用いる周波数帯域の検討、発話区間における話速の推定制度の改善などが挙げられる。

謝辞

本研究は，21世紀COEプログラム「京都アート・エンタテインメント創成研究」の支援を受けて行われた。

参考文献

[1] 関西デジタルアーカイブ
http://www.kiis.or.jp/kansaida/index.html

[2] 21 世紀 COE プロジェクト「京都アート・エンターテイメント創成研究」
http://www.ritsumei.ac.jp/acd/re/k-rsc/krc/21st_coe/index_coe.html

[3] 大阪能楽会館
http://www.pp.iij4u.or.jp/~rohnishi/

[4] 社団法人 能楽協会
http://www.nohgaku.or.jp

[5] 中川隆弘 他、“能楽ビデオデータに対するタグ付け支援環境の開発”, 情報処理学会シンポジウムシリーズ 人文科学とコンピュータシンポジウム論文集, Vol.2005, No.21, pp.143-149 (2005.12).

[6] Y.Linde, A.Buzo, and R.M.Gray, “An algorithm for vector quntizer design,” IEEE Trans.Commun.,vol.COM-28,no.1,pp.84-95,Jan.1980.