ドラマにおけるシナリオのセリフと音声トラックの同期システム

谷村 正剛　中川 裕志
横浜国立大学 工学部

1　はじめに

テレビドラマにおいて、視聴者の興味に沿ったシーンを手早く検索したいという要求が強まっている。中でも役者のセリフや演技、シーンの雰囲気など意味内容に基づいた検索システムの需要が期待されている。ドラマのシーン検索ではシナリオを検索し、それに対応する動画像および音声を利用者に提示するというシステム構成が考えられている [1] が、シナリオのセリフなどにはそれらが対応する動画像や音声トラックにおける先頭からの時間が記述されていない。従来から画像特徴や音量パターンを用いてシナリオに対応する音声トラックにおける時刻を求める試みがなされていた [3] が、最近の音声認識技術の発展により、新たに発話内容のパターンを利用することが可能となった。我々は、音声トラックの発話内容を認識することによりシナリオに対応する音声トラック上の時刻を自動付与するシステムを提案する。本システムの構造を図 1に示す。以下、各要素システムの動作について説明する。

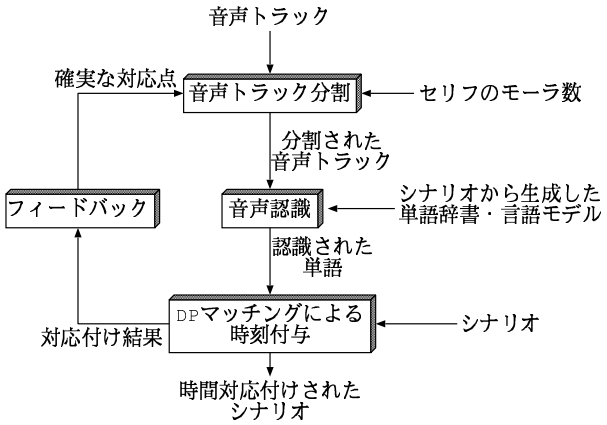

図 1: 対応付けシステムの構造

2　モーラ数を用いた音声トラックの自動分割

現在の音声認識システムは文単位の認識を目標としているため、音声トラックは音声認識にかける前にセリフ毎に分割する必要がある。本システムでは各セリフの発話時間によく比例するモーラ数に応じて音声トラックの発話時間を比例配分することにより音声トラックを分割した (図 2)。音声トラックから発話区間を抽出し、発話区間を各セリフのモーラ数に応じて比例配分することにより分割点を求め、発話時間の抽出により落された無発話区間を復元することにより元の音声トラック上での分割時刻を算出した。セリフの読みは、セリフの形態素解析により求めた。発話区間の検出は、音声トラックのパワーを求め、パワー最大値よりも閾値以下に下回った場合は発話なし、そうでなければ発話ありとした。

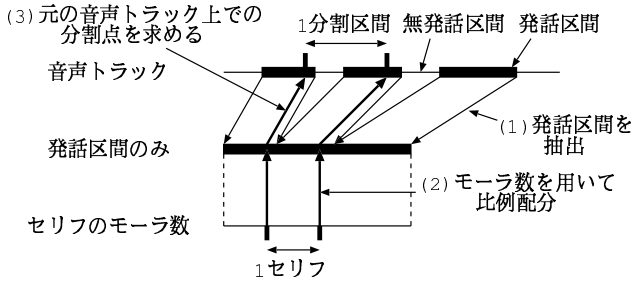

図 2: セリフのモーラ数を用いた音声トラックの自動分割

3　音声認識による音声トラックからの単語認識

本システムにて用いた大語彙連続音声認識システム JULIUS [2] では、認識のために単語 n-gram 言語モデルおよび単語辞書が必要となる。本システムでは汎用言語モデルおよび単語辞書を用いた場合よりも精度を向上させるため、シナリオから発話内容を抽出することにより言語モデルおよび単語辞書を生成した。言語モデルはシナリオのセリフを形態素解析した上で単語 n-gram 出現確率を計算することにより生成した。単語辞書は形態素解析の結果得られた単語の読みを音素列に変換することにより生成した。また、音素モデルは男女別 3000 状態 16 混合連続分布 HMM を用いた。話者性別は音声認識時には判別せず、男声モデルを用いた認識と女声モデルを用いた認識をそれぞれ行い、後述する DP マッチングを用いた対応づけにおいて話者性別を判定した。

4　DP マッチングによるシナリオへの時刻付与

音声認識の性質として、発話時間が長い単語は認識しやすいこと、母音は子音に比べて認識しやすいこと、発話の有無は発話内容に比べて認識しやすいこと、話者性別と音素モデルの性別が異なると認識率が著しく低下することがある。これらの性質に基づき、本システムでは単語の長さおよび母音を中心にした単語対の一致度を評価する単語対スコアと、発話時間の一致度を評価する経路スコアの 2 種を求めた。その上で両者の重みつき線形和をもって DP マッチングのスコア値とし、時刻付与および話者性別の判定を行なった (図 3)。単語対スコアは、単語対を構成する音声認識結果およびシナリオの各単語に対し、単語を構成するモーラの系列に対して DP マッチングをとり、得られた解経路の持つスコアを持って単語対スコアとした。経路スコアは、経路上に現れる音声認識結果およびシナリオの時間一致度を、モーラ数の一致度をもって近似することにより求めた。話者性別は、各経路のスコアを求める際に男声モデルと女声モデルの認識結果のうちスコアが高いものを選択することにより判別した。

Synchronization System of Script Lines and Sound Track in Drama

Seigo Tanimura and Hiroshi Nakagawa
Faculty of Engineering, Yokohama National University
79-5 Tokiwadai, Hodogaya-ku, Yokohama-shi, Kanagawa 240-8501, Japan
Tel: 045(339)4137 Fax: 045(338)1157
E-mail: {tanimura,nakagawa}@naklab.dnj.ynu.ac.jp

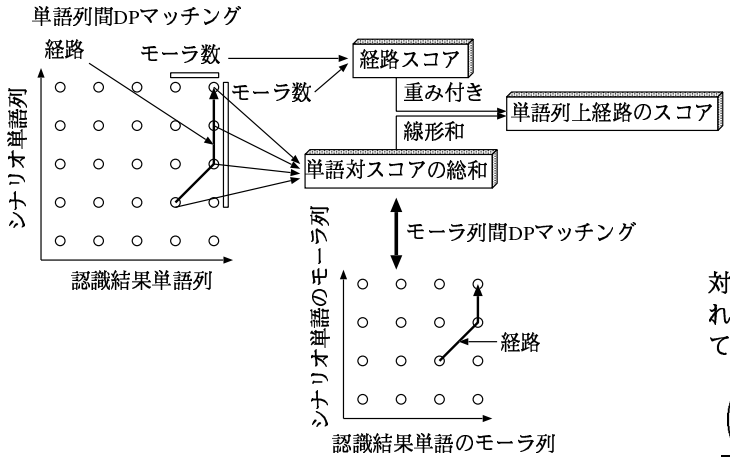

図 3: DP マッチングにおけるスコア算出手順

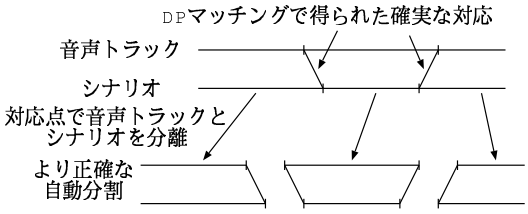

図 4: 対応付け結果を用いた音声トラック分割へのフィードバック

5 音声トラック自動分割へのフィードバック

音声トラックの自動分割精度を改善するため、DP マッチングによって付与された時刻を音声トラックの自動分割にフィードバックさせた (図 4)。一般的に、3 モーラ以上の単語については、考えられるモーラの組合せに対して実際に意味を持つ単語の数が急に少なくなるため、音声認識にて 3 モーラ以上の単語が認識された場合、その単語は確実に認識されたと考えられる。本システムでは、3 モーラ以上の完全一致単語を含む対応は確実なものであるとし、それらの対応点にて音声トラックとシナリオを分離した上で音声トラックを再分割することにより、より正確な分割点を求めた。

6 ドラマのシーンを用いた時刻付与性能の評価

本システムの性能を評価するため、ドラマのシーンを用いて評価実験を行なった。シーンの特徴を表 1、付与時刻の正解率を表 2 に示す。表 2 における「一致」は、時刻付与によってセリフに対して短時間でも正しい音声トラックが

表 1: シーンの特徴

セリフ数	23
音声トラック時間	120 秒
場所	家の居間
無発話時間	短
発話者性別	女性 / 男性
BGM	なし

表 2: 付与時刻の正解率)

フィードバック		なし	1 回
正解率	一致	74%	61%
	±1 ずれ	77%	82%
	±2 ずれ	83%	91%

対応付けられていれば正解とした場合の値である。「±1 ずれ」および「±2 ずれ」は、セリフが ±1 ないしは ±2 ずれていても正解とした以外は「一致」と同じである。正解率は

$$\frac{\left(\begin{array}{c}\text{音声トラックのセリフと正しく対応づけられた}\\ \text{シナリオのセリフ数}\end{array}\right)}{(\text{シーンに含まれる全セリフ数})}$$

と定めた。

フィードバックなしの場合の正解率は「一致」で 74% となり、モーラ数を用いた音声トラックの自動分割が有効であることを示した。フィードバックをかけた後の正解率は「±1 ずれ」および「±2 ずれ」にてそれぞれ 5% および 8% 増加し、フィードバックによって音声トラック自動分割の精度が改善されていることを示した。

7 まとめ

音声認識を用い、ドラマのシナリオに対応する音声トラックの時刻情報を付与するシステムを提案し、ドラマのシーンを用いた実験によってセリフ単位での時刻付与における本システムの有用性を評価した。

今後の課題としては、音声トラックの自動分割における発話時間の推定精度の改善や、DP マッチングへのフィードバックによる時刻付与精度の改善が考えられる。

謝辞

この研究は文部省科学研究費補助金 (創成的基礎研究: 課題番号 09NP1401) の援助を受けている。また、東京大学の坂内正夫教授には大変有益な御助言を頂いた。

参考文献

[1] 三浦健仁, 中川裕志. シナリオを用いたドラマのシーン検索システム. 情報学シンポジウム, 1999.

[2] 伊藤克亘, 河原達也, 武田一哉, 鹿野清宏. 日本語ディクテーション基本ソフトウェア. 人工知能学会全国大会 (第 12 回) 論文集, pp. 449–452, 1998.

[3] 柳沼良知, 和泉直樹, 坂内正夫. 同期されたシナリオ文書を用いた映像編集方式の一提案. 信学論 (D-II), Vol. J79-D-II, No. 4, Apr 1996.