音声認識を用いたドラマのシナリオへの時刻情報付与

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。従来からショットチェンジ検出などの手法により1本のドラマをシーン毎に分割し、各々のシーンに対し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 [4]。最近の大語彙連続音声認識技術の発展により、新たに発話内容の認識が可能となり、発話内容のパターンを対応付けに利用できるようになった。

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。本システムを構成する要素を以下に示す [3]。

(1) 音声トラック分割 音声トラックを発声区間と無発声区間に分離した上で、発声区間の時間がセリフのモーラ数に比例するように分割することにより、近似的にセリフ単位に分割する。

(2) 音声認識 シナリオからセリフを抽出し、形態素解析によって得られたセリフの単語から単語辞書およびn-gram言語モデルを生成する。セリフ単位に分割された音声トラックに対しシナリオから生成された単語辞書とn-gram言語モデルを用いて音声認識をする。認識結果として、音声トラックから認識された単語および音声トラックにおける発話時刻を得る。

(3) DPマッチングによる時刻付与 シナリオのセリフと音声認識によって認識された単語をDPマッチングを用いて対応付け、音声トラックとシナリオのセリフの時間対応付けがとれたシナリオを得る。

(4) フィードバック 時間対応付けの結果を用いて音声トラックの分割点を修正し、分割精度を上げる。

提案するシステムでは、音声認識システムから得た認識単語のうち、正しく認識された単語数が1割程度であっても対応付けができるようにするため、探索する経路が通る可能性のある範囲をあらかじめ絞り込み、認識ミスによる大きな対応付けのずれを抑制している。

以下、各要素システムの動作について説明する。

2 モーラ数を用いた音声トラックの自動分割

テレビドラマにおいては、1人ないしは複数人の役者が1つないしは複数の文から構成されるセリフを連続して発話することが多い。しかし、現在の音声認識システムは1文単位の認識を目標としている。このため、ドラマの音声トラックに対して直接音声認識システムによって発話内容の認識をしようとしても、発話終了を検出した時点で認識を終了してしまい、以降のセリフの開始点を認識できない。音声トラック全体を音声認識システムで処理するためには、音声トラックをセリフ毎に分割した上で音声認識システムへ与える必要がある。

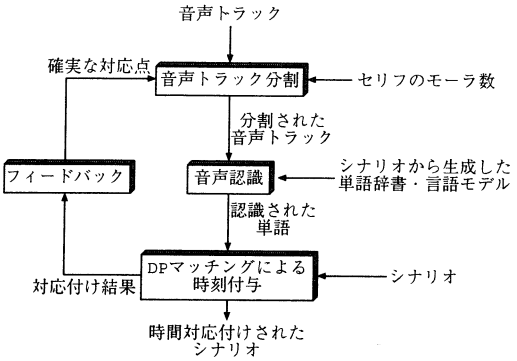

図 1: 対応付けシステムの構造

音声トラックをセリフ毎に分割するためには、各セリフの発話時間を推定しなければならない。本システムでは日本語の発話時間が発話内容のモーラ数に良く比例するという性質に基づき、セリフの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各セリフの発話時間を推定し、音声トラックを分割した。

音声トラックを分割する具体的な手順を図2に示す。まず、音声トラックのパワーを求め、音声トラック全体におけるパワーの最大値よりも閾値以下の場合は発話なし、そうでなければ発話ありとすることにより発話区間を抽出した。抽出した発話区間に対し、発話区間全体の時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発話区間上でのセリフ間の分割点を与えた。セリフのモーラ数は、セリフを形態素解析してセリフの読みを求め、セリフの読みをモーラに変換することにより得た。発話区間上に与えたセリフ間の分割点から元の音声トラックにおける先頭からの時刻を求めることにより、元の音声トラック上でのセリフ間の分割点を得た。

3 音声認識による音声トラックからの単語認識

音声認識においては、発話内容に関する背景知識として、言語モデルと単語辞書を与える必要がある。ドラマの場合、発話内容はシナリオとしてあらかじめ与えられている。このため、汎用の言語モデルおよび単語辞書を用いて音声認識をするよりも、シナリオから発話内容

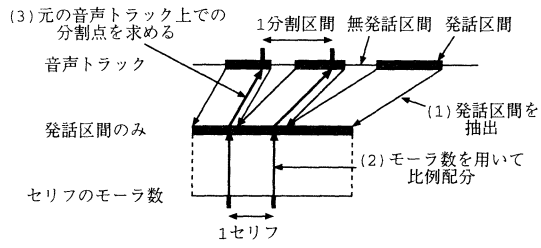

図 2: 音声トラックの分割

を抽出した上で、発話内容に特化した言語モデルおよび単語辞書を生成して音声認識をした方がより高い精度を得られる。

本システムでは、大語彙連続音声認識システムとして、「日本語ディクテーション基本ソフトウェア」の音声認識エンジン JULIUS [2] を用いた。n-gram 言語モデルは、シナリオから抽出した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する。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する。また、音素モデルは「日本語ディクテーション基本ソフトウェア」付属の男女別 3000 状態 16 混合連続分布 HMM を用いた。

4 DP マッチングによるシナリオへの時刻付与

DP マッチングを用いてシナリオと音声トラックを対応付ける際、音声認識システムから得られる単語の性質と日本語の音声学的特徴をスコア関数や探索幅に反映させることにより対応付け精度の向上をはかる。音声認識システムから得られる単語には以下に示すような性質がある。

- **認識された単語のモーラ数と認識の正確さの関係** モーラ数の多い単語は、発話時間が長くなり音声波形から多くの特徴が得られることと、発話された可能性のある単語の候補数が減ることから、正確に認識されやすい。モーラ数の少ない単語には正確に認識された単語もある一方、本来モーラ数の多い単語が発話されたが音声波形からの特徴抽出に失敗し、本来のモーラ数よりも短い単語の連続として認識されたものも含まれる。特に 2 モーラ以下の単語については認識ミスの結果得られた単語であることも多い。
- **発話の有無に対する認識精度の良さ** 発話の有無の認識は発話内容の認識に比べて精度が良い。
- **性別と認識の正確さの関係** 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という性質がある。
- **母音認識の正確さ** 母音と子音を比べると、母音の方が種類が少ないことや、発話中の特徴変化が小さいことから、正確に認識されやすい。

以上の性質に基づき、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発話時間スコアの 2 種を定めた。その上で、経路に対するスコアは、経路上の全単語対に対する単語対スコアと発話時間スコアの和として定義した。また、正しい対応づけの予想経路を考え、予想経路に近づく向きに経路を伸ばす場合は大きな重みを与えた。さらに、探索範囲を予想経路の周辺に限定し、対応付けに大きなずれが生じることを防いだ。話者性別は、男性モデルによる認識と女性モデルによる認識を並行して行なった上で、それぞれ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。以下、スコアの計算法、正しい経路の予想について具体的に説明する。

(1) スコアの計算法

スコアの算出手順を図 3に示す。軸上の目盛は単語であり、丸はセリフの単語と認識結果の単語の対を表す。以下、具体的にスコアの算出手順を説明する。

a. 単語対スコア

ある単語対について、含まれる両単語の発音が同じか似ており、かつ単語のモーラ数が多い場合はその単語対に含まれている両単語は対応していると考えられる。このため、そのような単語対を通るような経路に対して高いスコアを与える必要がある。そこで、対応付けの経路が通過する単語対において、発音の似た単語対に対して高いスコアを与えるための単語対スコアを定めた。その上で、経路が通過する全ての単語対に対する単語対スコアの和を経路のスコアとした。

単語対スコアは、単語対に含まれる両単語の発音をモーラを用いて表した上で、モーラ間の対応の数を求めることにより算出した。モーラ同士が完全に一致している場合だけでなく、モーラ同士の母音だけが一致している場合も、そのモーラ同士は対応しているとした。単語対スコアの定義は、完全一致したモーラ同士の対応に対して 3 点、母音のみ一致したモーラ同士の対応に対して 2 点として、全てのモーラ同士の対応に対して与えた点数の総和とした。これにより、両単語のモーラ数が多く、発音も似ている単語対を通るような経路には高いスコアを与えるようにした。また、全く同じ発音の単語対については単語対スコアを 2 乗し、経路がその単語対を確実に経由するようにした。

単語対スコアの計算例を、図 4の A 点について示す。A 点でのセリフの単語は「ごめんなさい」、認識結果の単語は「ごめん」である。それぞれをモーラで表すと、/go me N na sa i/ および /go me N/ を得る。/go/ の対応と /me/ の対応については、両単語の間でモーラが完全一致しており、それぞれの対応に 3 点ずつ、計 6 点を与える。/N/ は対応しているが、母音を含まないためスコアは与えない。/na/、/sa/、/i/ は対応をもたず、やはりスコアは与えない。以上より、A 点の単語対スコアは 6 点となる。

b. 発話時間スコア

単語対スコアはモーラ数の多い単語が音声認識によって得られていた場合は有効であるが、音声認識にミスがあり、モーラ数の短い単語しか得られなかった場合は単語対スコアは有効に機能しない。このような場合でも対応付けに大きなずれが生じないようにするためには、認識結果から得られた発話時間とセリフから予測さ

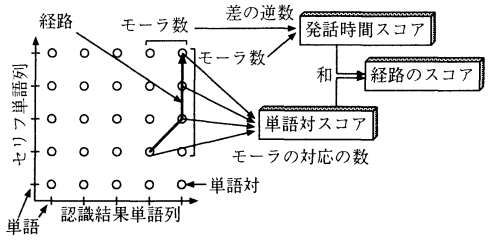

図 3: 経路に与えるスコアの算出手順

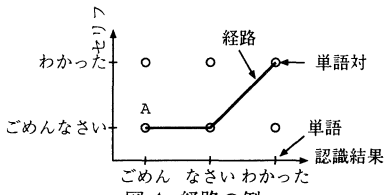

図 4: 経路の例

れる発話時間が等しくなるような経路を選択する必要がある。そこで、セリフと認識結果の各々について、経路上に含まれる全ての単語のモーラ数を持って発話時間を近似した。その上で、セリフと認識結果における全単語のモーラ数が小さい場合に高いスコアを与えるための発話時間スコアを定めた。

発話時間スコアは、セリフと認識結果の各々について、経路に含まれる全ての単語のモーラ数を用い、

$$\frac{\alpha}{|(\text{セリフの全モーラ数}) - (\text{認識結果の全モーラ数})| + 1}$$

と定義した。α は経路のスコアを求める時の重み調整のための係数である。α は経路上に全く同じ発音の単語対が含まれる場合は単語対スコアの方が発話時間スコアより大きくなり、そうでなければ単語対スコアと同程度の値になるように調整する。経路上に含まれるセリフと認識結果の各々の全モーラ数が等しければ、発話時間スコアは α となる。経路上のセリフと認識結果の全モーラ数に大きな差があれば、発話時間スコアは小さくなる。これにより、モーラ数の多い単語が認識結果として得られなかった場合でも、発話時間がほぼ一致するような経路を選択することにより、対応付けに大きなずれが生じないようにした。

発話時間スコアの計算例を、図4の経路について示す。セリフ側の単語のモーラは、表れる順に /go me N na sa i/、/wa ka q ta yo/ となるので、全モーラ数は 11 である。同様に、認識結果側の単語のモーラは /go me N/、/na sa i/、/wa ka q ta yo/ となり、全モーラ数はやはり 11 である。これより発話時間スコアは

$$\frac{\alpha}{|11 - 11| + 1} = \alpha$$

と求められる。

(2) 正しい経路の予想

a. 予想経路と重みへの反映

音声認識では発話の有無は認識しやすいが、それに比べて発声内容を認識するのは困難である。このような場合、音声認識システムから得られる認識結果は助詞などの出現頻度が高く、モーラ数が 1 から 2 程度の短い単語が連続することが多いため、本来の発話内容と比べると得られる単語数が増加する。実験に用いたドラマのシーンでは、セリフの形態素解析の結果得られた単語数が約 250 語であったのに対し、認識システムから得られた単語数は約 400 語に上った。このため、対応付けの経路を図5に示すように認識結果側に傾けながら伸ばす必要がある。

経路が認識結果側に傾くようにするため、図5に示すように正しい対応付けであると予想される経路を定め、それに近付く経路には大きな重みを与えた。予想経路は、各単語列の先頭にある単語の対を始点、各単語列の最後尾にある単語の対を終点とする直線とした。経路の探索時には、経路を伸ばす向きが予想経路に近付く場合はその経路の重みを上げるようにした。これにより、認識結果側に傾くような経路をとりやすいようにした。

b. 探索範囲の限定

認識システムから得られた約 400 語のうち、正しく認識できた単語数は 50 語程度であった。このため、認識結果の単語列には認識ミスによる短い単語が多く含まれ、所々に正しく認識できた単語が現れる。このような単語列を DP マッチングを用いてセリフに対応付けようとすると、認識ミスによる短い単語が連続する区間では経路の探索時に認識結果の単語と似た単語がセリフに含まれていることが多いため、スコア関数が正しく計算されない。その結果、経路が予想経路から大きく外れて対応付けの精度を落す恐れがある。

経路が予想経路から外れないようにするため、図5に示すように対応付けの経路が通る可能性のある範囲をあらかじめ限定した。限定した範囲から外へは経路を伸ばさないようにすることにより、認識ミスによる大きな対応付けのずれが発生することを防いだ。

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に用いた情報はセリフのモーラ数のみであり、分割点の誤差が大きいと音声認識における認識率の低下につながる。これを改善するため、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。

DP マッチングにより得られた対応を全てフィードバックすると、認識ミスにより得られた単語を含む対応もフィードバックされてしまい、精度を改善できない。このため、DP マッチングにより得られた対応のうち確実なものだけを選択し、フィードバックする必要がある。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。よって、音声認識にて 3 モーラ以上の単語

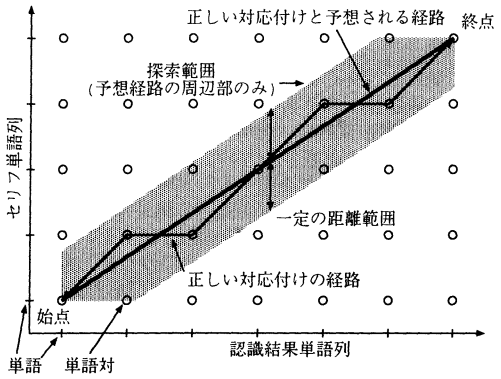

図 5: DP マッチングにおける経路探索範囲

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

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。さらに、3 モーラ程度の単語はほぼ全てのセリフにおいて出現するため、対応付けをとれる可能性が高い。そこで、DP マッチングにより得られたセリフの単語と認識結果の単語の対応のうち、3 モーラ以上で完全一致した単語の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することにより対応付けを固定した。その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。

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%

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、システム概要、セリフのモーラ数を用いた音声トラックの自動分割、セリフから生成した言語モデルおよび単語辞書を用いた音声認識、音声認識および日本語の発話の特徴を利用したスコア関数と経路の絞り込みによる対応付けのずれを防いだ DP マッチングによるシナリオと音声認識結果の対応付け、対応付け結果の自動分割へのフィードバックの手法について述べた。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] 谷村正剛, 中川裕志. 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. インタラクション'99, 1999.

[4] 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.