知能と複雑系 118－4
（1999．10．15）

ドラマのビデオ音声トラックとシナリオのセリフの時刻同期法

谷村 正剛
横浜国立大学 工学部 電子情報工学科

中川 裕志
東京大学 情報基盤センター 図書館電子化研究部門

我々は、音声トラックへの音声認識によりドラマのシナリオのセリフに対して音声トラックを自動的に同期させるシステムを提案する。本システムは、セリフのモーラ数を用いた音声トラックの自動分割、シナリオを背景知識として用いた音声認識、日本語の音声学的特徴および音声認識の性質に基づいた音声認識結果とセリフの対応付け、対応付け結果を用いた自動分割へのフィードバックの4要素から構成される。

Time Synchronization Method between a Sound Track and Speech Lines in a Drama

Seigo Tanimura
Division of Electrical and Computer Engineering,
Faculty of Engineering,
Yokohama National University
tanimura@naklab.dnj.ynu.ac.jp

Hiroshi Nakagawa
Digital Library Division,
Information Technology Center,
The University of Tokyo
nakagawa@r.dl.itc.u-tokyo.ac.jp

We propose a system to synchronize a sound track to a script in a TV drama by speech recognition to the sound track automatically. Our system consists of four elements: automatic segmentation of a sound track by the number of moras in a script, speech recognition with the script as the background knowledge, alignment of the words recognized in a sound track to the speech lines based on the phonetic properties and the characteristics of speech recognition, and feedback of the alignment result to the segmentation of the sound track.

1 はじめに

我々は日常、テレビや新聞などさまざまなメディアからの情報を、テキスト、静止画像、動画像、音声など複数のモダリティを通して受けとっている。そのような情報の理解支援をする上では、各モダリティごとに意味内容を解析するだけでなく、あるモダリティにおいて得られた意味内容を他のモダリティにおいても利用できるようにすることが必要となる。それを実現するためには、異なるモダリティにて表現された情報の間の対応付けをとる手法が不可欠である。テレビドラマにおいてはテキストとしてシナリオがあり、これを音声トラックと時間的に対応付けることにより、シナリオを解析して得た意味情報を音声トラックや動画像の解析でも利用することにより、より詳細な意味解析が可能になると考えられる。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 [2] が、最近の音声認識技術の発達により、発話内容を用いてさらに精度の良い対応付けをとることが可能になってきた。

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とそれに対応する音声トラックを時間的に対応付け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。

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 マッチングによる対応付け

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、音声トラックに対する音声認識において複数の音響モデルを用いることにより認識率を改善している。また、複数のモデルを用いた音声認識では認識に用いたモデル数だけの認識結果が得られるが、セリフと認識結果を対応付ける DP マッチングにおいて、セリフの各単語に最もよく類似したモデルを局所的に選択することにより、複数の認識結果を有効に利用している。また、音声認識により得られる単語数はセリフから得られる単語数よりも多くなることが多いため、音声認識により得られた単語の一部を除去することにより、DP マッチングの対象となる単語の数が傾斜制限から外れて対応付けに影響を与えることを防止している。さらに、DP マッチングにおける計算量を削減するため、正しい対応付けの予想経路を考え、予想経路の周辺のみで計算を実行する。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。

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 [3] 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。

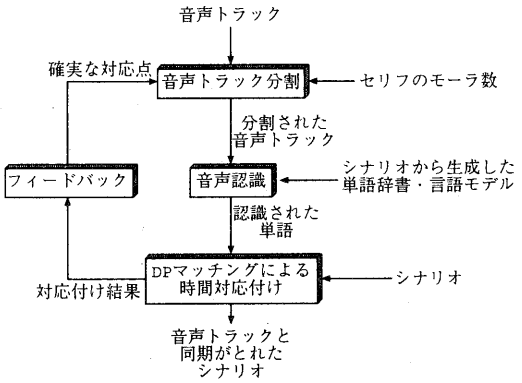

図 1: 対応付けシステムの構造

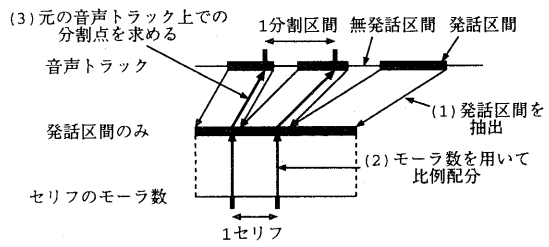

図 2: 音声トラックの分割

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」[4]の音声認識エンジン JULIUS [5] 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、音響モデルは「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 [6] を用いた。音響モデルは男性モデルと女性モデルの 2 種類が用意されていたが、音声トラックの音質がよくないため音声トラックに対する話者性別の判定が困難であった。そこで、本システムでは男性モデルおよび女性モデルを用いた音声認識を並行して行い、後述する DP マッチングを用いた対応付けにおいてそれぞれにおいて得られた認識結果を統合してセリフに対応付けた。

4 DP マッチングを用いたシナリオのセリフと音声認識結果の対応付け

ドラマの音声トラックには BGM や物音などの雑音が含まれることが多いため、音声認識から得られた認識結果の認識率は 3 割程度しか得られない。このため、認識結果を直接音声トラックの発話内容として用いることは難しく、より正確な発話内容の表現となっているセリフとの対応付けをとることが必要となる。また、各役者の音声特徴には個人差があるため、音声認識において単一の音響モデルを用いるよりも、性別や年齢、話速、発声様式などごとに複数のモデルを利用した方がより高い対応付けの精度を得ることができる。その場合、モデルごとに複数の音声認識結果が得られるため、認識結果をセリフに対応付ける際、セリフから得られた単語に最も類似した認識結果を選択するような工夫が必要である。

また、パターン単語列に対する入力単語列の DP マッチングでは、パターン単語列 A に含まれる単語 $a_i \in A$ に対する入力単語列 B の単語 $b_j \in B$ の類似度 $s(a_i, b_j)$ を求める必要がある。このとき、音声認識により得られた単語の性質として、モーラ数の多い単語が認識された場合は、モーラ数の少ない単語に比べて認識ミスである可能性が少ないという傾向があるため、マッチングにおいてこの性質を利用することによりより正確な対応付けがとれると考えられる。

以上の要求を満たすため、本システムではまずパターン単語 a_i のモーラ列 A'_i に対する入力単語 b_j のモーラ列 B'_j の時間正規化 DP マッチングをとることにより単語間の類似度 $s(a_i, b_j)$ を求め、その上で単語列間のマッチングをとった。その際、男女別に得られた認識結果を有効に利用するため、一般的に用いられている 1 パターン系列 A に対する 1 入力系列 B の DP マッチングを拡張し、1 パターン系列 A に対する 2 入力系列 $B^m (m = 1, 2)$ の DP マッチングを行った。パターン系列はセリフから得られた単語列とし、入力系列は音声認識結果から得られた単語列としている。

以下、一般的な DP マッチングについて説明した上で、単語列間のマッチングの方法および単語間の類似度 $s(a_i, b_j)$ の求め方について説明する。また、単語列間のマッチングにおける工夫について述べる。

4.1 一般的な DP マッチング

まず、拡張された DP マッチングとの比較のため、単一の認識結果をセリフに対応付けるための DP マッチングについて説明する。

対応付ける単語を、

$$A = \{a_1, a_2, \ldots, a_i, \ldots, a_I\} \quad (1)$$

$$B = \{b_1, b_2, \ldots, b_j, \ldots, b_J\} \quad (2)$$

とする。ここに

A セリフから得られた単語列

a_i セリフから得られた単語

B 音声認識から得られた単語列

b_j 音声認識から得られた単語

である。B を A に対応付ける問題は、単語 a_i と b_j の類似度を $s(a_i, b_j)$ としたとき、対応付け系列の類似度

$$S(a_I, b_J) = \frac{\sum_{k=1}^{K} w(k)\, s(a_{i_k}, b_{j_k})}{\sum_{k=1}^{K} w(k)} \quad (3)$$

を最大化するような a_{i_k} と b_{j_k} を求める問題と考えられる。ここに、$w(k)$ は $s(c_k)$ の重みである。局所的な類似度の平均がどのような対応付け系列についても等しくなるように、認識結果の単語を 1 語挿入または削除する場合の重み $w(k)$ を 1、それ以外の場合を 2 とすると、この問題は以下のような初期条件および漸化式を用いて解くことができる。

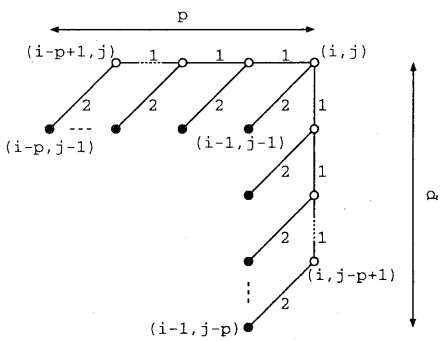

図 3: 一般的な DP マッチングにおける傾斜制限

初期条件

$$g(a_1, b_1) = 0 \tag{4}$$

$$g(a_i, b_1) = g(a_1, b_j) = -\infty \quad (i > 1, j > 1) \tag{5}$$

漸化式

$$g(a_i, b_j) = \max \begin{cases} g(a_{i-1}, b_j) + s(a_i, b_j) \\ g(a_{i-1}, b_{j-1}) + 2s(a_i, b_j) \\ g(a_i, b_{j-1}) + s(a_i, b_j) \end{cases} \tag{6}$$

最終解

$$S(a_I, b_J) = \frac{g(a_I, b_J)}{I + J} \tag{7}$$

以上の式を用いて $S(a_I, b_J)$ を求めることにより、(a_1, b_1) から (a_I, b_J) に至る最適経路とその類似度を求めることができる。

また、極端に連続した伸縮を防ぐ方法として、傾斜制限を用いることができる。p 回以上の連続した挿入および削除を禁止するためには、図 3に示すような傾斜制限を用いればよい。線分上の値は、その線分に沿った経路に対する重みである。(i, j) で示される点への最適な経路は、黒丸で示される点から至る各経路のうち、類似度が最大であるものを選択することにより求められる。具体的には、漸化式 (6) を

$$g(a_i, b_j) = \max_{q=1,2,\ldots,p-1} \begin{cases} g(a_{i-q-1}, b_{j-1}) + 2s(a_{i-q}, b_j) \\ \quad + \sum_{r=1}^{q} s(a_{i-q+r}, b_j) \\ g(a_{i-1}, b_{j-1}) + 2s(a_i, b_j) \\ g(a_{i-1}, b_{j-q-1}) + 2s(a_i, b_{j-q}) \\ \quad + \sum_{r=1}^{q} s(a_i, b_{j-q+r}) \end{cases} \tag{8}$$

と書き換えればよい。

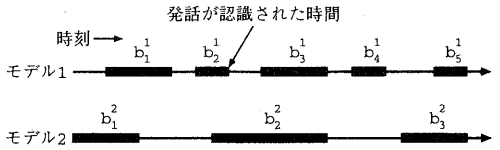

図 4: 2 種類の音響モデルを用いた認識結果の例

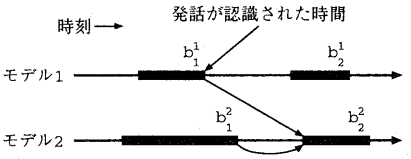

図 5: 複数のモデルにおける直前の単語の選択法

4.2 複数の音響モデルを用いた認識結果を対象とする DP マッチング

本システムでは男女別の音響モデルを用いて音声認識をするため、認識結果としては男性モデルの場合と女性モデルの場合の 2 種類が得られる。しかし、4.1節にて説明した一般的な DP マッチングでは入力単語列を 1 本のみとしている。このため、複数の入力単語列を取り扱えるように DP マッチングのアルゴリズムを拡張する必要がある。

図 4は、音声認識において 2 種類の音響モデルを用いて得られた認識結果の単語を時間軸上にて表現した例である。この例において認識結果をセリフに対応付けることを考えると、例えば b_1^1 と b_1^2 のように同時に複数のモデルで発話が認識されることがあるため、どのモデルにて認識された単語がセリフの単語と最も良く類似しているのか求めなければならない。

以上の問題に対し、本システムでは、b_j に対する直前の単語 b_{j-1} をモデルごとに求めるようにした。具体的には、各モデルごとに b_j の発話開始時刻の直前に発話終了を持つような単語を 1 語ずつ選択し、それらを b_{j-1} として漸化式 (8) を計算し、最も高い類似度 $g(a_i, b_j)$ を与えるようなモデルを選択する。図 5は、モデルが 2 種類の場合にて b_2^2 の直前に発話された単語を選択する例である。b_1^1 および b_1^2 は b_2^2 発話開始の直前に発話が終了しているため、b_1^1 と b_1^2 をそれぞれ式 (8) の b_{j-1} に代入し、類似度 $g(a_i, b_j)$ を最も高くする方を選択する。また、b_2^1 は発話終了が b_2^2 の発話開始よりも遅いため、b_2^2 の直前の単語とはしない。以上により、DP マッチングを用いて最適なモデルを動的に選択する。

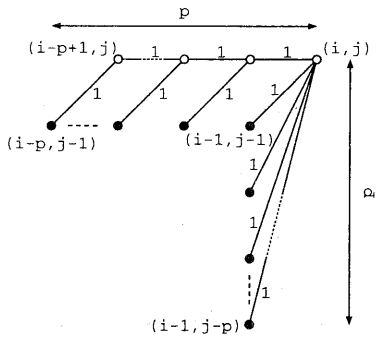

図 6: 時間正規化 DP マッチングにおける傾斜制限

4.3 単語間類似度計算のための時間正規化 DP マッチング

単語間類似度は、セリフから得られた単語のモーラ列に対する認識結果から得られた単語のモーラ列の DP マッチングを行い、その結果得たモーラ間の対応付け系列の類似度を求めることにより計算する。それには、式 (1) および式 (2) をそれぞれ

$$A' = \{a'_1, a'_2, \ldots, a'_{i'}, \ldots, a'_{I'}\} \quad (9)$$

$$B' = \{b'_1, b'_2, \ldots, b'_{j'}, \ldots, b'_{J'}\} \quad (10)$$

のように置き換え、式 (9) をセリフ、式 (10) を認識結果から得られた単語のモーラ列とする。また、モーラ間の類似度 $s'(a'_{i'}, b'_{j'})$ を

$$s'(a'_{i'}, b'_{j'}) = \begin{cases} 3 & (a'_{i'} = b'_{j'}) \\ 2 & (a'_{i'}\text{の母音と}b'_{j'}\text{の母音のみが等しい}) \\ 0 & (\text{それ以外}) \end{cases} \quad (11)$$

と定め、母音のみが一致した場合でも小さい類似度は与えるようにする。以上のようにした上で DP マッチングを用いて B' を A' に対応付ければよい。

ところで、式 (6) や式 (8) に示したような漸化式を用いた DP マッチングでは、$g(a'_{I'}, b'_{J'})$ が得られるまでの $g(a_{i'}, b_{j'})$ に対する $s(a'_{i'}, b'_{j'})$ の加算回数は $I' + J'$ 回である。このため、式 (6) や式 (8) を用いた DP マッチングでは $g(a'_{I'}, b'_{J'})$ がセリフから得られた単語のモーラ数に依存する。このため、セリフから得られた単語が変わると $g(a'_{I'}, b'_{J'})$ も変化してしまい、類似度の比較が困難になる。

単語間類似度がセリフから得られた単語に依存しないようにする方法として、時間正規化 DP マッチングがある [3]。これは挿入に対する重みを 0 とした DP マッチングである。これにより、$g(a'_{i'}, b'_{j'})$ に対する $s(a'_{i'}, b'_{j'})$ の加算回数は J' 回となり、$g(a'_{I'}, b'_{J'})$ がセリフのモーラ数に依存しなくなる。時間正規化 DP マッチングでは図 6 に示すような傾斜制限を用いるのが普通である。この時の漸化式は、式 (8) を

$$g'(a'_{i'}, b'_{j'}) = \max_{q=1,2,\ldots,p-1} \begin{cases} g'(a'_{i'-q-1}, b'_{j'-1}) + s'(a'_{i'-q}, b'_{j'}) \\ g'(a'_{i'-1}, b'_{j'-1}) + s'(a'_{i'}, b'_{j'}) \\ g'(a'_{i'-1}, b'_{j'-q-1}) + s'(a'_{i'}, b'_{j'-q}) \\ \quad + \sum_{r=1}^{q} s'(a'_{i'}, b'_{j'-q+r}) \end{cases} \quad (12)$$

に変更すればよい。

また、音声認識により得られた単語の性質として、モーラ数の多い単語が認識された場合は、モーラ数の少ない単語に比べて認識ミスである可能性が少ないということがある。このため、認識結果の単語のモーラ数が多い場合は、単語間類似度も大きくなることが望ましい。したがって、単語間類似度に対しては式 (7) の右辺分母にあるような加算回数 J' を用いた正規化はせず、式 (12) から求めた $g'(a'_{I'}, b'_{J'})$ を直接用い、

$$s(a_i, b_j) = g'(a'_{I'}, b'_{J'}) \quad (13)$$

とする。

4.4 単語数の調整

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。例えば、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。

セリフから得られた単語の数と音声認識により得られた単語の数の間に大きな差があると、極端な伸縮が多くなり、傾斜制限があると最適な対応付けが難しくなる。これを防ぐため、音声認識により得られた単語の一部を除外し、音声認識により得られた単語の数がセリフから得られた単語の数にほぼ等しくなるように調整した。

単語数の調整のために除外する単語は、無発話時間に相当する単語とした。具体的には、句読点を除外対象とした。音声認識では発話中に現れる無発話時間を単語の一種として取り扱い、句点や読点として認識する。しかし、発話時に生じる無発話時間は常にセリフに含まれる句読点の通りに現れるとは限らない。具体的な例としては、セリフに句読点がないにもかかわらず、発話時には息つぎなどにより無発話時間が入ることがある。そのような無発話時間は、音声認識により句読点として得られる。逆に、セリフに句読点が含

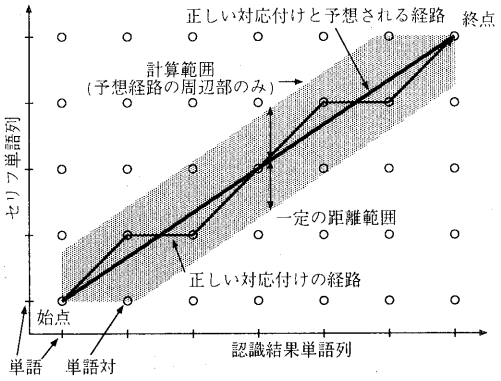

図 7: DP マッチングにおける計算範囲

まれていても連続発話されることもある。この場合音声認識からは句読点を得ることはできない。このように、音声認識により得られた句読点はセリフと対応しないことがあるため、対応付けにおいて利用することは難しい。そこで、音声認識により得られた単語から無発話時間に相当する単語を除外した。これにより、音声認識により得られた単語の数をセリフから得られた単語の数とほぼ同程度になるようにした。

4.5 予想経路を用いた計算量の削減

一般的な DP マッチングでは、考えられるパターン系列の要素 a_i と入力系列の要素 b_{jm}^{m} の全ての組合せに対して、式 (6) などに示されるような漸化式を計算する必要がある。組合せの数はパターン系列の長さ I と入力系列の長さ J の積のオーダであるため、対応付けの対象とするシーンが長くなると急に計算量が増えてしまう。これに対処するため、対応付けの予想経路を考えた上で、その周辺のみで漸化式の計算を実行することにより、計算量を削減した。

予想経路の例を図 7に示す。セリフの単語列と認識結果の単語列の間の正しい対応付け経路は、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し、両者の間をほぼ一定の傾きを保ちながら結ぶと考えることができる。そこで、始点と終点を結ぶ直線を対応付けの予想経路とし、予想経路から一定の単語数の距離以下の範囲内でのみ漸化式の計算を行なった。なお、音声認識結果が男女別に得られるため、予想経路も男女別に求めた。

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対応付け性能の評価

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、図 7に示した予想経路の正解率を表 2、対応付け結果の正解率を表 3 に示す。表 2および表 3 における「一致」は、対応付け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正解率である。「±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。

シーン 1 およびシーン 2 の両者において、予想経路はセリフ 2-3 個程度の範囲内に収まっていた。シーン 1 では、フィードバックなしの場合にすべてのセリフについて前後 1 セリフ以内の対応付けをとることができた。また、フィードバックにより全てのセリフについて対応する音声トラックを対応付けることができた。さらに、各セリフの言い出しおよび言い終りにつ

表 1: シーンの特徴

	シーン 1	シーン 2
セリフ数	18	37
音声トラック時間	135 秒	248 秒
場所	家の玄関	職場、屋外など
無発話時間	短	中
発話者性別	男性	主に男性
BGM	なし	一部あり

表 2: 予想経路の正解率

	シーン 1		シーン 2	
	女性	男性	女性	男性
一致	61.1%	61.1%	32.8%	56.8%
±1 ずれ	83.3%	83.3%	64.9%	94.6%
±2 ずれ	100%	94.4%	86.5%	100%

いても平均して 1 単語程度にずれを抑えることができた。シーン 2 では、フィードバックにより対応付けの正解率を 10 から 15% 程度改善することができた。

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全ての単語を対応付けられるように対応付け経路の向きを工夫した 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、複数の音響モデルを用いた音声認識により得られた認識結果に対する DP マッチングの方法のより詳細な検討がある。今回はモデルを男女別の 2 種類としたが、複数の発話方法や複数の話者など、より多くのモデルを用意して音声認識をすることにより認識精度を改善できると考えられる。3 種類以上のモデルから得られた認識結果のマッチングをとる方法については、現在検討中である。この他、音声トラックの自動分割における発話時間の検出法や分割点の算出法を改善することがある。特に、ドラマにおいては BGM がよく使われるため、BGM の影響を受けにくい発話時間の検出法を検討する必要がある [1]。また、音声トラックの分割点を無発話時間中にとることにより、音声認識の精度を改善することも考えられる [7]。さらに、対応付け結果のフィードバックによる音声トラック自動分割の修正の改善や、より多くのシーンを用いた実験なども考えられる。

表 3: 対応付けの正解率

	シーン 1		シーン 2	
フィードバック	なし	1 回	なし	1 回
一致	94.4%	100%	75.0%	86.5%
±1 ずれ	100%	100%	84.6%	97.3%
±2 ずれ	100%	100%	86.5%	100%

参考文献

[1] Kenichi Minami, Akihito Akutsu, Hiroshi Hamada, Yoshinobu Tomonaga: Video Handling with Music and Speech Detection, IEEE Multimedia, Vol. 5, No. 3, 1998

[2] 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), J79-D-II, No. 4, 1996

[3] 田窪 行則, 前川 喜久雄, 窪園 晴夫, 本多 清志, 白井 克彦, 中川 聖一: 岩波講座 言語の科学 2 音声, 岩波書店, 1998

[4] 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

[5] 李 晃伸, 河原 達也, 堂下 修司: 単語トレリスインデックスを用いた大語彙連続音声認識エンジン JULIUS, 電子情報通信学会技術研究報告, SP98-3, 1998

[6] 武田 一哉, 峰松 信明, 伊藤 彰則, 伊藤 克亘, 宇津呂 武仁, 河原 達也, 小林 哲則, 清水 徹, 田本 真詞, 荒井 和博, 山本 幹雄, 竹沢 寿幸 , 松岡 達雄, 鹿野 清宏: 大語彙日本語連続音声認識研究基盤の整備 – 汎用音素モデルの作成 –, 情報処理学会研究報告, 97-SLP-18, 1997

[7] Katsumi Tadamura, Eihachiro Nakamae: Synchronizing Computer Graphics Animation and Audio, IEEE Multimedia, Vol. 5, No. 4, 1998.