テレビドラマのシナリオと音声トラックの時間対応付け

谷村 正剛　　中川 裕志
東京大学 情報基盤センター

ビデオコンテンツにおいては、コンテンツに付随するテキストを利用することにより、検索などがより容易になると考えられる。しかし、テレビドラマのシナリオ等のように、テキスト中のある部分がビデオのどの区間に対応するのかについては記述されていないことがあり、そのままでは検索に用いることは難しい。我々はこの問題を解決するため、音声認識を利用してシナリオのセリフとビデオの音声トラックを自動的に対応付けるシステムを提案する。

Alignment of a Script with a Sound Track in a TV Drama

Seigo Tanimura　Hiroshi Nakagawa
Information Technology Center, The University of Tokyo
{tanimura,nakagawa}@r.dl.itc.u-tokyo.ac.jp

Text accompanied with a video content allows us to retrieve desired video contents with ease. One example is retrieval of a TV drama. However, a script of a drama generally has no annotation of times and durations in a video for sentences of the speech lines and directions, making it difficult to use a script in retrieval. To solve this problem, we propose a system to align the sound track with the speech lines of the script in a TV drama.

1　はじめに

テレビドラマのようにテキストの付随したビデオコンテンツを検索する場合、テキストすなわちドラマのシナリオを対象とした検索を行なうことにより、ビデオの動画像や音声を直接検索の対象とするよりも使いやすくなると考えられる。しかし、テレビドラマのシナリオには各々のセリフないしはト書きが対応するドラマのビデオ中での時刻が記述されていない。シナリオのセリフやト書きに対応したシーンを再生するためには、シナリオのセリフを音声トラックおよび動画像と時間的に対応付ける必要がある。これを解決するために、動画像のショットチェンジ、音声トラックの音量、シナリオのセリフに含まれる文字数などを用いたシナリオのセリフと音声トラックおよび動画像の時間的な対応付けが提案されている [1] が、セリフ単位での対応付け精度は7割程度に留まっている。

対応付けの精度を改善する方法として、ドラマの音声トラックに対する音声認識により、セリフの発話時刻をより正確に得るという方法が考えられる。しかし、単純に音声トラックに対して音声認識を実行することはできない。現在の音声認識は1文程度を認識単位としており、1本のドラマについて音声トラック全体を1度に認識することは困難である。このため、音声認識の前に音声トラックをセリフ単位にまで分割する必要がある。

音声トラック全体を1度にセリフ毎に分割することも考えられるが、1本のドラマには多くのセリフが含まれているため、1度にセリフ毎に分割すると精度を落す恐れがある。これを防ぐには、セリフよりも時間的に長い論理シーンを単位としてまず大まかな対応付けをとり、その上で論理シーンをセリフ毎にさらに細かく分割するという2段階の分割方法をとることができる。そのためには、ドラマ全体から論理シーンおよびセリフを抽出する必要がある。

セリフの抽出は、日本語の発話時間がモーラ数に比例するという性質 [2] を用いることにより容易に実現できる。また、論理シーンは動画像へのショットチェンジ検出により得られた物理ショットが複数連続したものである。物理ショットは、市販のツール [6] を用いてドラマのビデオから簡単に抽出することができる。しかし、ドラマでは1本のシーンから大量のショットチェンジが得られることが多い。このため1度の対応付けだけで論理シーンの境界となるショットチェンジを正確に抽出することは困難である。したがって、対応付け結果のフィードバックにより論理シーン境界とショットチェンジの対応を修正することが不可欠となる。

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および動画像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。

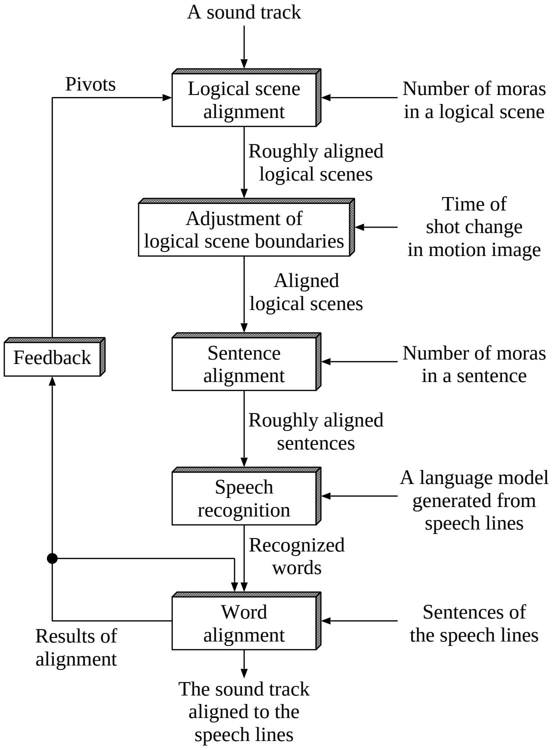

図 1: 対応付けシステムの構造

1. 論理シーン対応付け

   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、シーン全体を近似的に論理シーン単位に分割する。更に、動画像からショットチェンジを検出 [6] した上で、各論理シーンの境界を最も近い動画像上のショットチェンジ時刻へ補正する。

2. セリフ対応付け

   論理シーン対応付けとほぼ同様の方法により、論理シーンを近似的にセリフ単位に分割する。

3. 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。

4. 単語対応付け

   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。

5. フィードバック

   論理シーン分割および対応付けの精度を上げるため、時間対応付けの結果得られた確実な対応点を用いて論理シーン分割の誤りを修正する。修正は時間対応付けおよび論理シーン対応付けに反映する。

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

音声認識により得られた単語とセリフの単語の対応付けでは、数分程度を超える長いシーンを直接対応付けようとすると精度を保ったまま計算量を減らすことが難しくなる。効率良く対応付けをとるため、最初にシーン全体について論理シーンを単位とした対応付けをとり、それにより得られた各論理シーンに対してさらに細かい対応付けをとる。論理シーンは動画像上の複数の連続した物理ショットから構成される。したがって、各々の論理シーンの境界に対応している物理ショット境界を求めることにより、動画像から物理ショットのみを用いて論理シーンを抽出することが考えられる。しかしドラマに含まれる物理ショット数は論理シーンの数倍程度と多いため、前述の方法で論理シーンの対応付けをとるのは容易ではない。

また、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。

以上の問題に対し、我々はセリフの発話時間に基づいて音声トラックとシナリオを対応付け、その上で論理シーンまたはセリフ単位にシーンを分割する方法をとった。セリフの発話時間に基づいてシーンを論理シーンまたはセリフ毎に分割するためには、各論理シーンまたは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 [2] に基づき、論理シーンまたはセリフのモーラ数に応じて音声トラッ

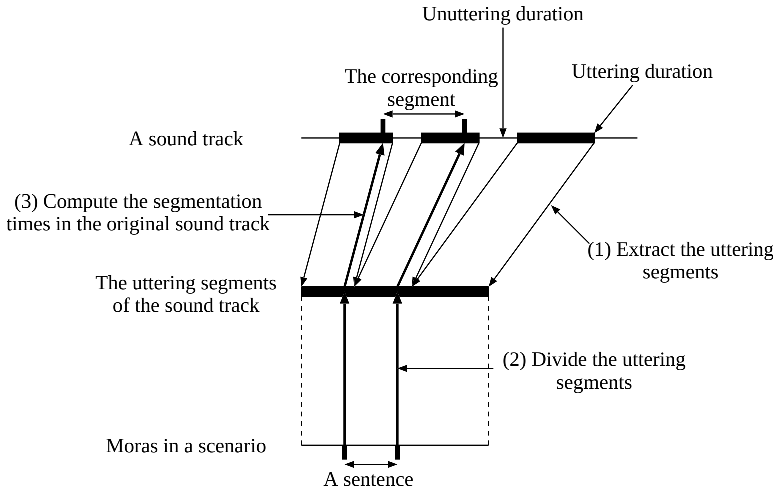

図 2: 音声トラックの分割

クの発話時間を比例配分することにより各論理シーンまたは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。図2および以下の説明はセリフ対応付けについてである。論理シーンについてはセリフの場合と同じ処理を実行した後、更に後処理をする。

セリフ単位の対応付けは以下のようにして行なった。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。

論理シーン単位の対応付けでは、セリフ単位の対応付けと同じ処理を論理シーンに対して行なった後、論理シーンの境界を最も近い動画像上のショットチェンジ時刻に合わせる補正をした。ショットチェンジ時刻の検出には市販の動画像編集ツール [6] を用いた。

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」[3] の音声認識エンジン JULIUS [4] 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 [5] 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。

音声認識において、誤認識した単語が認識結果に含まれていると単語単位での対応付けにおいて精度が悪化するのみならず、フィードバックにて用いる確実な単語対応にも誤認識した単語を含むものが現れてフィードバックの効果が得られなくなることもある。誤認識した単語と正しく認識した単語の違いとして、正しい単語ではその前後に現れる単語が N-gram 言語モデル $(N \geq 2)$ にも現れていることが多いのに対し、誤認識した単語ではその前後にランダムな単語を持つことが多い。そこで、言語モデルの重みを

JULIUS のデフォルト値の 2 倍とすることにより誤認識を防いた。

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単語対応付け

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。

初期条件

$$g(a_1, b_1) = 0 \quad (4)$$

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

漸化式

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

最終解

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

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

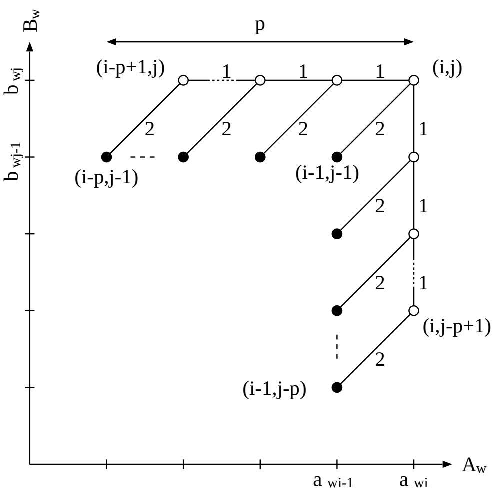

図 3: 一般的な DP マッチングにおける傾斜制限

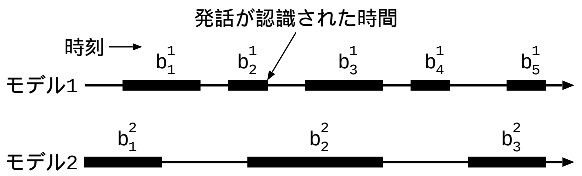

図 4: 2 種類の音響モデルを用いた認識結果の例

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

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

と書き換えればよい。

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。

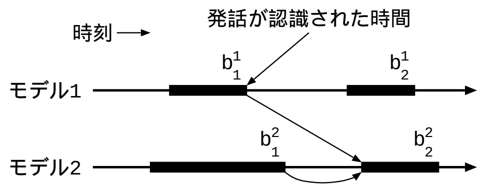

図 5: 複数のモデルにおける直前の単語の選択法

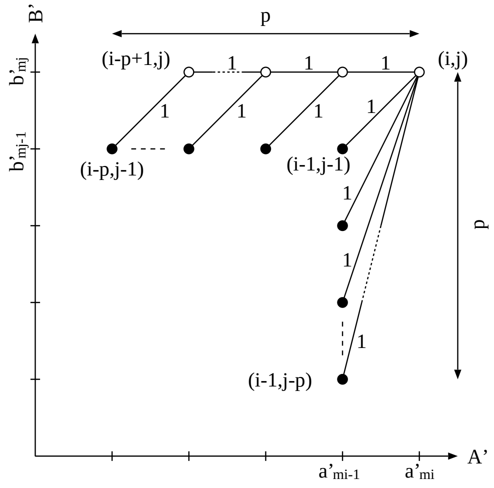

図 6: 時間正規化 DP マッチングにおける傾斜制限

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 $g\left(a_i, b_j\right)$ 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 $g\left(a_i, b_j\right)$ 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。

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

$$A' = \left\{a'_1, a'_2, \ldots, a'_{i'}, \ldots, a'_{I'}\right\} \tag{9}$$

$$B' = \left\{b'_1, b'_2, \ldots, b'_{j'}, \ldots, b'_{J'}\right\} \tag{10}$$

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 $s'\left(a'_{i'}, b'_{j'}\right)$ を

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

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、$g\left(a'_{I'}, b'_{J'}\right)$ が得られるまでの $g\left(a_{i'}, b_{j'}\right)$ に対する $s\left(a'_{i'}, b'_{j'}\right)$ 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 $g\left(a'_{I'}, b'_{J'}\right)$ 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 $g\left(a'_{I'}, b'_{J'}\right)$ 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 [2]。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、$g\left(a'_{i'}, b'_{j'}\right)$ に対する $s\left(a'_{i'}, b'_{j'}\right)$ の加算回数は J' 回となり、$g\left(a'_{I'}, b'_{J'}\right)$ 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

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

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 $g'\left(a'_{I'}, b'_{J'}\right)$ を直接用い、

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

とする。

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。また、論理シーンの対応付けに誤差があると単語対応付けの開始点および終了点が正しく対応付けられず、確実な単語対応を得ることができない。そこで、単語対応付けによって得られた対応点をフィードバックし、論理シーンの対応付けを修正することが考えられる。

単語対応付け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、単語対応付け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、単語対応付け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た。以下、そのような対応をピボットと呼ぶことにする。

論理シーン対応付けの修正は、以下の2通りの方法で行なった。

1. 単語対応付けの改善

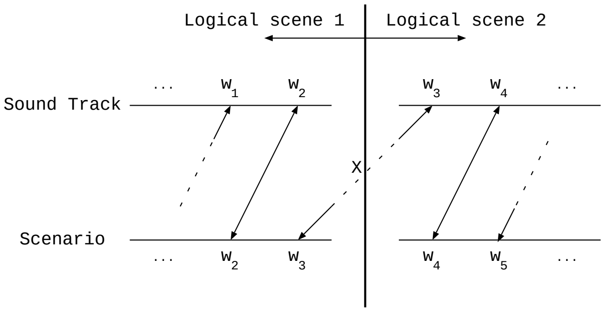

図 7: 論理シーン開始点 / 終了点のずれによる影響

単語対応付けにおいて、対応付ける論理シーンの開始点や終了点が正しく対応付けられていないと、図7に示すように論理シーン境界付近で単語の対応がとれなくなり、ピボットを得ることができなくなる。そこで、図8に示すように、ピボッ

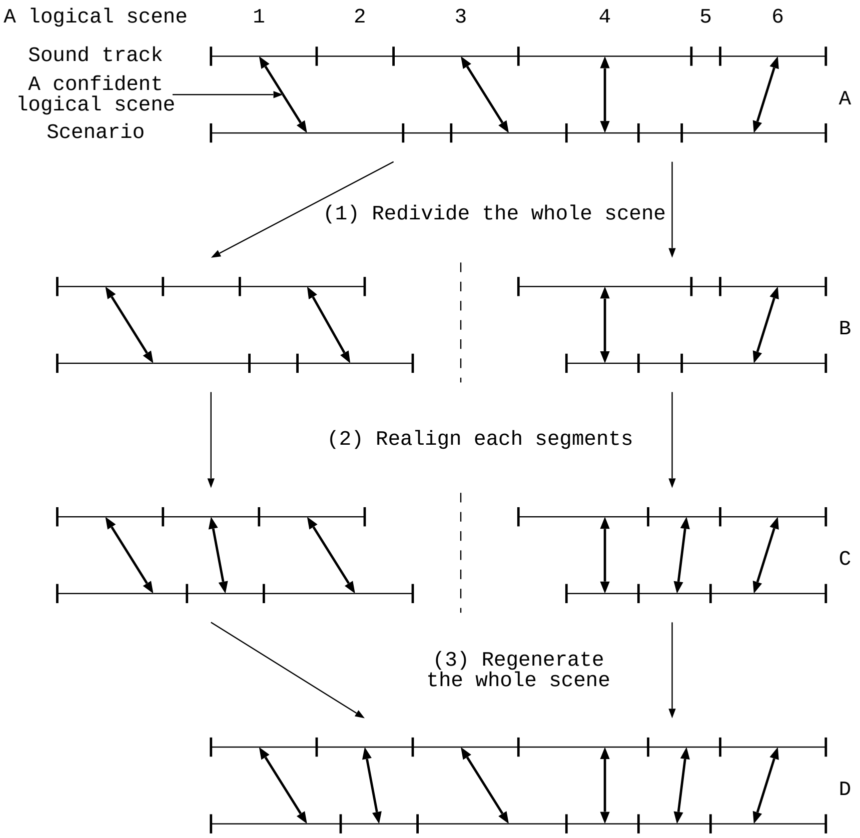

図 8: 単語対応付けへのフィードバック

トが1つ以上得られた論理シーンは正しく対応付けがとれているものとした上で、そのような論理シーンを開始点および終了点として単語対応付けをやり直した。

2. 論理シーン対応付けの改善

単語対応付けにて得られたピボットを固定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。

表 1: 対応付け実験に用いたシーン

論理シーン数	24
セリフ数	91
音声トラック時間 [分: 秒]	14:44

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。表1に実験に用いたドラマシーン、図9および10に各セリフの言い出しと言い終りの時刻のずれ、表2に5回のサイクルにより2秒以内に対応付けがとれたセリフの比率を示す。

表 2: 2秒以内に対応付けがとれたセリフの比率

言い出し	36.3%
言い終り	15.4%

フィードバックを繰り返すことにより時間のずれがゼロに近付き、数回のサイクルによりシナリオ全体の内7割程度のセリフは数秒以内のずれで対応付けをとることができた。

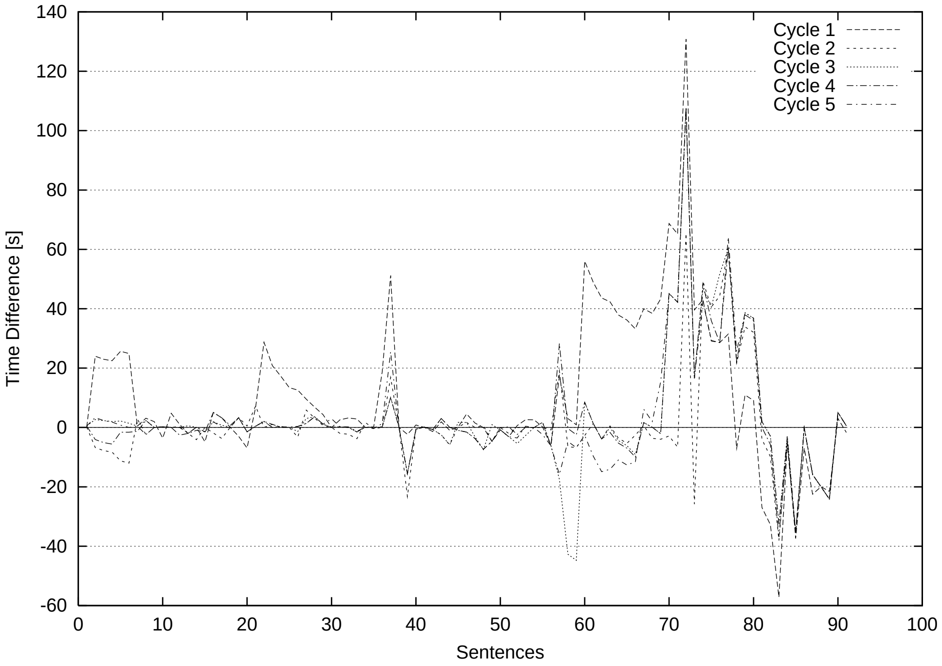

図 9: 各セリフの言い出しのずれ時間

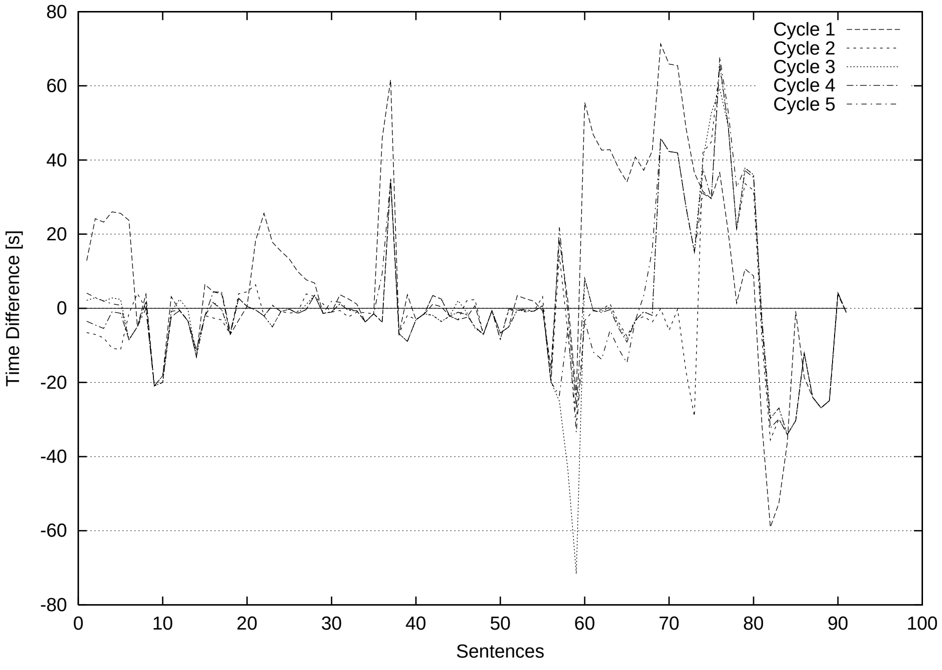

図 10: 各セリフの言い終りのずれ時間

7 まとめ

音声認識および動画像のショットチェンジを用い、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本システムの有用性を確認した。今後の課題としては、ニュースと新聞などドラマ以外の音

声とテキストの対応付けへの応用や、パラレルコーパスなどへのアラインメント手法の応用などが考えられる。

参考文献

[1] 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

[2] 田窪 行則, 前川 喜久雄, 窪薗 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

[3] 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

[4] 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

[5] 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

[6] 日立製作所: Mediachef/かっと for Windows 95, 1997