特別研究報告書

複数文の連結を考慮した
会話音声とテキストの対応付け

指導教員　　椋木 雅之 准教授

京都大学工学部情報学科

清水 渚佐

平成24年2月2日

複数文の連結を考慮した会話音声とテキストの対応付け

清水 渚佐

内容梗概

近年，学校教育において，外国語によるコミュニケーション能力の育成が重視されている．より実践に近い語学力を身につけるために，授業に映像を取り入れることが考えられる．そこで，本研究では，教師が授業で取り扱っている単語や熟語でテキストを検索し，その文に対応する映像部分を再生することを想定する．このような機能を実現するためには，各文が映像のどの部分に対応しているかを正確に知る必要がある．

ここで，映像は音声と動画から構成されており，音声とテキストの対応付けが得られれば，自動的に映像とテキストの対応付けが求まる．そこで，本研究では，音声とテキストをテキスト中の文単位で対応付けることを目的とする．語学学習番組は，日常会話を取り扱ったものが多く，授業では実際に日常会話で使われる表現を学ぶ必要があるため，教材映像として適している．そのため，本研究では，語学学習番組の会話シーンを対象とする．

従来研究として，映像・音声・テキストから共通するパターンを 0.5 秒ごとに抽出して DP マッチングを用いてドラマ映像とテキストを対応付ける手法が提案されている．また，字幕から抽出したテキスト情報に対応する音素・音節単位の音声モデルと音声をマッチングすることでニュース放送の音声と字幕を対応付ける手法が提案されている．これらの研究では，音素単位など，非常に細かい粒度での対応付けを求めることができる．しかし，これらの研究では，1 文の中に短いポーズが含まれている場合や，言い淀みなどのテキストに書き起こされない発話がある場合には，音声とテキストの不一致が生じる．また，前者の研究では，文と文の間には非発話区間があることを想定しているが，実際の日常会話では，文と文が連続して発話される場合がある．これらの問題により，音声とテキストの正しい対応付け結果が得られない可能性がある．

本研究で想定している文単位での再生という利用法を考えると，文単位での対応付けが取れていれば十分である．そこで，本研究では，音声とテキストの文単位での対応付けを考える代わりに，文中のポーズや言い淀み，複数文の連続発話といった従来研究における問題点に対処する．

複数の文が連続して発話されたために，文と文の間に対応すべき非発話区間

が存在しない場合，1 つの発話区間に対して複数の文が対応付けられる必要がある．逆に，1 文を発話する間に非発話区間が存在する場合，1 文に対して複数の発話区間が対応付けられる必要がある．よって，本研究では，文及び発話区間の隣り合うもの同士のあらゆる連結パターンの対応付けを行うことにより，これらの問題点に対処する．さらに，文と発話区間の特徴から適合度を計算し，適合度が高い組み合わせから対応付けを決定していくことにより，対応すべき文がテキストに存在しない発話が，他の文と間違って対応付けられることを防ぐ．適合度の計算には，テキスト及び音声から抽出した発話継続長とキーワードの特徴を用いる．過度に文同士が連結して対応付けられないように，連結した文の数のペナルティと，文の発話継続長及び文中のキーワードの類似度の重み付き和を適合度とし，対応付けを行う．最後に，複数文を連結して発話区間に対応付けた部分に対して，各文の推定発話継続長の比により発話区間を分割することで，発話区間における各文の区切り位置を推定する．

本手法の有効性を調べるために，語学学習番組の会話シーンに対して，抽出された文と発話区間の集合をそのまま対応付ける場合と，あらゆる連結パターンを生成して対応付ける場合とで，音声とテキストの文単位の対応付け精度の比較を行った．各文の文頭・文末が対応付けられた音声の時刻と手動で与えた文頭・文末の正解時刻のずれが 0.5 秒及び 1 秒以内のものは対応付けに成功したものとし，文単位の対応付け精度を求めた．15〜28 文を含む 24〜71 秒の会話シーン 12 サンプルについて評価を行った結果，文及び発話区間を単独に対応付ける場合に比べ，連結パターンを対応付けることにより，0.5 秒以内のずれを許容する場合は 38.5%から 47.3%に 8.8 ポイント，1 秒以内のずれを許容する場合は 48.3%から 59.6%に 11.3 ポイント精度が向上した．

今後の課題として，文と発話区間の順序関係は変わらないため，文と発話区間の全体における位置を適合度に反映させることで，対応付けが大きくずれる問題を回避することができると考えられる．また，今回は英会話を扱う 1 番組のみで検証を行ったが，様々な語学学習番組の会話シーンに対して本手法を検証する必要がある．

Matching Conversational Speech and Text Considering Connection of Sentences

Nagisa SHIMIZU

Abstract

In school education, it is emphasized to train the ability to communicate in foreign languages. In our application, a teacher searches a part of educational language video that addresses syntax or a word that the teacher wants to teach, and show the students the video as a practical example. In order to achieve this application, we propose a method to find a part of a video clip corresponding to each sentence in the given text book.

In this paper, we aim to find audio segments corresponding to each sentence in the text. Because an audio data is always synchronized with an image sequence in a given video clip, to find the audio segment means to find the video segment. The target video clips in our method are learning language TV programs about everyday conversational talk.

A previous research proposed a method for matching a drama movie with its scenario text every 0.5 seconds using DP matching based on the features extracted from the image sequence, the audio data and the text data. Another previous research proposed a method for matching captions with speeches of a TV news program by applying the syllable/phoneme HMM constructed through the captions. In these researches, they achieved the matching in fine grained unit, such as phonetic units.

However, in these researches, the matching tends to fail when a speaker makes a pause while speaking one sentence, or when there are speeches not transcribed in the text, such as a falter. Additionally, in the former previous method, it is assumed that there are always pauses between adjacent sentences. However, it is common that more than two sentences are spoken continuously in conversational talk.

In order to achieve our application, it is enough to find the matching between the speeches and the text data in sentence unit. Therefore, we aim to get the matching between the text and the audio in sentence unit and we cope with the problems in previous methods.

In this paper, we match any patterns generated by concatenating more than two neighboring sentences or speech segments. If the speaker speaks more than two sentences continuously without a pause, one speech segment needs to be matched to multiple sentences. Conversely, if the speaker makes a pause during speaking one sentence, one sentence needs to be matched to multiple speech segments. Therefore we need to match multiple sentences and multiple speech segments. Additionally, we decide the matching between the speech and sentence pair in the order of high fitness, so that non-transcribed speeches are not incorrectly matched with sentences. The fitness is weighted sum of similarity of speech duration and keywords and the penalty of number of concatenated sentences. Finally, even when multiple sentences are matched to speech segments, the start and the end times are found for each sentence by dividing the speech segments according to the ratio of the estimated speech duration of each sentence. In order to investigate the concatenation effects on the accuracy of the matching, we compared the accuracy when sentences and speech segments are isolated with the one when these concatenated patterns are generated. We regard that the matching is correct when the difference of the start and the end time between correct and estimated speech segments is within 0.5 or 1 seconds.

In this paper, we applied our method to 12 conversational video clips. Each video clip is from 24 to 71 seconds long and contains from 15 to 28 sentences. Compared with the method of matching single sentence and speech segment, our method improves the total accuracy from 38.5% to 47.3% by 8.8 points in case the acceptable range of error is 0.5 seconds, and from 48.3% to 59.6% by 11.3 points in case that is 1 second.

As a future work, because order of sentences and speech segments is not change, the position of them in the video clip should be used for calculating fitness. Moreover, in this paper, we apply our method to only one learning language TV program in English. It is necessary to evaluate our method with others.

複数文の連結を考慮した会話音声とテキストの対応付け

目次

第1章　緒論

近年，学校教育において，外国語によるコミュニケーション能力の育成が重視されている．より実践に近い語学力を身につけるためには，テキストのみで勉強するよりも映像を使った方が理解が促進されると考えられる．このとき，単語の難易度や取り扱う構文など，学習したい内容に適した映像を用意する必要がある．また，生徒の興味を引くような映像を使うほうが効果的である．黒田ら [1] は，語学学習番組の構造化を行い，その結果から教材映像を作成した．語学学習番組は，日常会話を取り扱ったものが多く，正しい文法や発音を含むため，魅力的な教材となり得る．

そこで，本研究では，語学学習番組を授業用教材として次のように利用することを想定する．教師は授業中，構文や単語を教えながら，その時々で，それらを含む映像を学生に見せて会話における使い方を紹介する．このとき，映像全体ではなく，単語や熟語でテキストを検索し，映像中のその文に対応する部分を再生することで，授業の要点をより的確に指し示すことができる．このような機能を実現するためには，会話シーン全体のような大まかな単位ではなく，各文が映像のどの部分に対応するかを正確に知る必要がある．

語学学習番組のように，外国語による日常会話を取り扱った映像コンテンツは古くから作られているため，大量に存在する．その中には，台詞と映像との同期が取られているものもあるが，特に古い映像コンテンツには，台詞と映像との同期は全く取られておらず，単にその映像中の会話を書き起こしたテキストが添付されているものも多い．また，今後新しく作成される映像コンテンツすべてが，台詞と映像の同期が取られているとは限らない．

ここで，映像は音声と動画から構成されており，音声と動画は同期が取られていることから，音声とテキストの対応付けが得られれば，自動的に映像とテキストの対応付けが求まる．そこで，本研究では，映像ではなく音声とテキストをテキスト中の文単位で対応付けることを目標とする．テキストとは，音声中の台詞を書き起こした文の列のことを指す．

語学学習番組では，1 分程度の長さの映像に 20 文程度の会話のやりとりが録画された会話シーンが取り扱われており，その中で発話された台詞を書き起こしたテキストが付与されている．また，映像と台詞との同期は取られていない．本研究では，このような語学学習番組の会話シーンを対象とする．

テレビ番組におけるテキストと映像の対応付けに関する従来研究には，次のようなものがある．柳沼ら [2] は，映像，音声，シナリオ文書に共通するパターンを 0.5 秒ごとに抽出し，DP マッチングを用いて対応付けることにより，ドラマ映像を構造化・データベース化する手法を提案した．シナリオ文書とはシーンごとの場所，人物名，それぞれの人物の台詞が書かれているテキストのことを指す．小林ら [3] は，字幕から音素・音節単位 HMM の連結を構成し，入力音声とマッチングすることで音声と字幕の同期を取り，字幕付きニュース放送から語学学習教材を半自動的に作成するシステムを提案した．これらの研究は，0.5 秒ごとや音節・音素単位といったように非常に細かい粒度での対応付けを求めることができる．しかしながら，テキストでは一文になっている部分の音声が，実際には文中に無音区間を含む場合や，言い淀みや間投詞など，テキストに表記されない音声があると，音声とテキストの不一致が生じ，正しい対応付け結果が得られないという問題点がある．

一方，上記で述べたような本研究の利用法を考えると，音節や音素ほど細かい単位で対応付けられている必要はなく，文単位での対応付けが取れていれば十分である．そこで，本研究では，従来より大きい単位である 1 文単位で音声とテキストの対応付けを行うことを目的とする代わりに，次のような従来研究における問題に対処する．

1 つ目の問題は，音声において，文と文の間に無音区間が存在するとは限らず，複数の文が連続して発話されることである．2 つ目の問題は，1 文の中にも無音区間が存在することである．3 つ目の問題は，言い淀みなど，テキストに書き起こされていない発話を行った発話区間が存在することである．そのため，音声における切れ目とテキストにおける切れ目が必ずしも一致しない．

以上の 3 つの問題に対し，本研究では，複数文の連結を考慮した多対多の対応付け問題として音声とテキストの対応付けを行う手法を提案する．隣り合う文同士が連結する可能性を考慮して対応付けを行うことにより，連続して発話されるために文と文の間に無音区間がない問題に対処する．さらに，1 文に対して複数の発話区間が対応付けられることを考慮することにより，1 文の中に無音区間が存在する場合にも対処する．また，音声の発話区間とテキストの文の適合度が高い組み合わせから対応付けを決定していくことにより，テキストに書き起こされない発話区間に対処する．最後に，得られたテキストと発話区間との多対多の対応付け結果から，一文単位での対応付けを推定し，音声とテ

キストの文単位での対応付けを求める．

以下，2 章では音声とテキストの対応付け問題と従来研究について述べる．3 章では，複数文の連結を考慮した対応付け手法について述べる．4 章では，提案手法の検証実験について述べる．5 章では，本研究のまとめと今後の課題について述べる．

第 2 章　1 文再生のための映像インデキシング

2.1　音声とテキストの対応付け問題の定義

本研究で対象とする音声とテキストの対応付け問題について述べる．

本研究では，語学学習番組の外国語による会話シーンを対象とする．

一般的に語学学習番組の会話シーンには，字幕やクローズドキャプションなどが付与されていることが多い．字幕やクローズドキャプションは，映像とある程度の同期が取れているため，これらを利用すれば比較的容易に映像とテキストを対応付けられると考えられる．しかし，背景ノイズによる影響や，字幕の表示方法が映像によって異なるため，字幕からテキストを抽出することは難しい [4]．また，クローズドキャプションが付けられた映像は近年増加してきているが，大量にある過去の映像にも対応するためには，クローズドキャプションが存在しない場合も利用できる手法は有用である．

一般的に，語学学習番組には番組に対応したテキスト教材があり，このテキスト教材には映像中の台詞が書かれている．テキスト教材では，どの文がいつ話されたかという明示的な時間情報が付与されていないのが一般的であるため，各文が映像や音声のどの部分に対応するかはわからない．授業では，教師が必要な単語や文を検索し、対応する部分の映像や音声が再生できることが求められるが，この機能を実現するためには，文と映像や音声が対応付けられていなければならない．映像や音声とテキストの対応を知るためには，映像や音声とテキストからそれぞれ特徴を抽出し，対応付けを求める必要がある．

以下に，本研究で対象とする語学学習番組の映像について述べる．図 1 は，語学学習番組の映像とテキストの例である．この映像は，自然な会話に近い条件で撮影されたものであり，言い淀みや発話中の短いポーズが含まれている．話者の声以外に含まれている音は，車の音や周囲の人の声といった環境音であり，BGM は含まれていない．図 2 は，言い淀みを含む部分の音声波形の例である．

また，図 3 は，文中に短いポーズを含む部分の音声波形の例である．さらに，話者の声が重複している部分や，発話が連続して行われるような部分も存在する．


```
Excuse me?
Where is the nearest subway station?
It's the 7th St/Metro center Station.
It's the 7th St/Metro center Station.
How can I get there from here?
Is it far?
No.
It's within walking distance.
Go straight.
Take a right on your second block.
It's on the right.
You'll see a Metro sign above you.
That sounds easy.
Thanks a lot.
You can't miss it.
```

図 1: 語学学習番組の映像とテキストの例

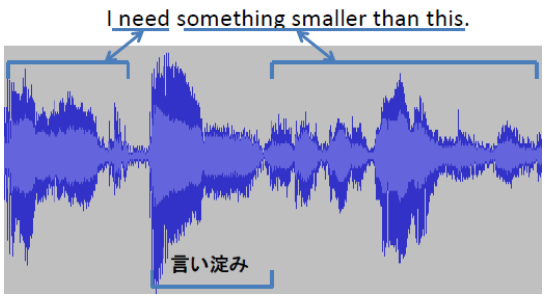

図 2: 言い淀みを含む文の音声波形

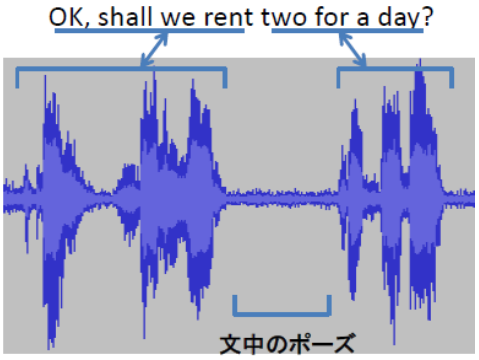

図 3: 短いポーズを含む文の音声波形

2.2 従来研究の問題点

テレビ番組映像を有効に再利用するために，映像や音声とテキストの対応付けに関して，従来から様々な研究が行われている．

文献 [2] では，ドラマ映像を対象として，次のような手順により映像とシナリオ文書の対応付けを行う手法を提案している．複数のメディアから参照できる

0.5 秒ごとの特徴パターンを，映像・音声とシナリオ文書からそれぞれ抽出し，各特徴パターンに対して DP マッチングを用いて対応付けを行い，それらの結果の重み付け和をメディア間の対応付け結果とする．各メディアから抽出する特徴パターンは，台詞パターン，女性の存在パターン，場面の変わり目パターンの 3 つを用いている．台詞パターンは，音声からは振幅レベルの大小パターン，シナリオ文書からは文字数を利用して推定した時間を用いている．女性の存在パターンは，音声からはピッチを抽出し，シナリオ文書からは各台詞の話者の名前を抽出することで求めている．場面の変わり目パターンは，映像からはカット検出の結果，シナリオ文書からはシーンの変わり目を用いている．しかし，この研究では，文と文の間には均等に無音区間が存在すると仮定しているため，複数の文が連続して発話される場合や，1 つの文の中に無音区間が存在する場合には，ずれが大きくなる可能性がある．

文献 [3] は，字幕付きテレビニュース放送を対象として，音声と字幕の同期を行う手法を提案している．字幕は，文字放送を通して得られたもの，もしくは，映像からの書き写しや Web から取得したものを利用している．字幕のかな文字列や音素記号列から，それに対応する音節・音素単位 HMM の連結を構成し，入力音声とのマッチングを行う．無音区間はパワーの閾値処理によって除かれている．

この 2 つの研究は，言い淀みや間投詞，環境音といったテキストに記述されていない発話や音が存在する場合，音声とテキストの照合結果の尤度が低くなり，正しい対応付けが得られないという問題点がある．

2.3 複数文の連結を考慮した対応付けによる精度向上

本研究では，文単位でその対応する映像部分を再生することを想定しているため，従来研究のような細かい単位ではなく，文単位の対応付けがわかっていれば十分である．そこで，本研究では，テキストにおける文単位の対応付けを目的とし，以下の問題点に対処する．

まず，文と文の間に無音区間が存在するとは限らないという問題がある．文献 [2] では，文と文の間には無音区間が存在すると仮定し，文字数による台詞パターンの推定を行っている．しかし，今回対象とする映像には，文と文の間に無音区間が存在しない場合があり，間違って対応付けられる可能性がある．また，文と文の間に無音区間が存在しないと，テキストにおける文と文の切れ目

に対応する部分が，音声においては切れ目となっていないため，各文に対応する音声区間を求めるのは難しい．このような場合，連続して発話される文は連結して音声区間と対応付けることにより，各文ごとに独立して対応付けるよりも，正しい対応付けが求まると考えられる．次に，1 文の中に無音区間が存在する場合があるという問題がある．先ほどの問題点とは逆に，テキストにおける 1 文に対応する部分の途中に，音声の切れ目が存在しているため，文と発話区間は一対一の対応にはならない．このような場合，テキストの 1 文に対して，音声における複数の発話区間が対応付けられる一対多の対応づけを行う必要がある．本研究では，文及び発話区間の連結を考慮して対応付けを行うことにより，これらの問題点に対処する．

また，テキストに書き起こされていない発話区間が存在するという問題点がある．発話区間検出では，言い淀みや間投詞のような，テキストに書き起こされていない発話や，話者の発話ではないノイズが発話区間として検出される場合があるため，音声にはテキストの文と対応していない発話区間が存在する．加えて，文と文の間の無音区間には偏りがあり，文の文字数から推定した時間が必ずしも音声における時間と一致するとは限らない．そのため，従来研究 [2] のように，全体における文の位置を考慮して対応付けを求めると失敗することがある．そこで，文と発話区間のすべての組み合わせの中で適合度が最も高い対応付けから順に確定していくことによって，全体における位置によらずに対応付けを決めることを考える．ここで問題となるのは，全体における位置を考慮しないで対応付けることにより，最初の方の文と最後の方の発話区間が対応付けられるといったように，位置を考慮して対応付けを行う場合には起こらなかった問題が生じる可能性があることである．しかし，会話の話題は時間と共に変化し，会話の最初と最後では台詞に含まれる単語が異なると考えられるため，台詞中の単語を適合度の計算に用いることでこの問題に対処する．

複数文の連結を考慮してテキストの文と音声の発話区間の対応付けを求めるため，この対応付け結果から，さらに 1 文単位での対応付けを求める必要がある．複数文を連結して対応する発話区間が求められた部分は，各文の文字数によって推定された発話継続長を利用して対応付けられた発話区間を分割して各文に対応付けることにより，1 文に対する対応付けを求める．文字数から推定される発話継続長には誤差があるが，本研究で想定するような選択した文単位の再生を行うアプリケーションの場合，少なくとも選択した文は対応付けられた

発話区間に欠損なく含まれていることが求められる．このような要求を満たすために，文頭は推定時刻より前に，文末は推定時刻より後ろにずらすことを許せば，推定された文に対応する発話区間に欠損のないようにすることができる．

さらに，ノイズの影響のため，無音区間を発話区間として誤検出する場合や，発話区間を無音区間として誤棄却する場合がある．対応する発話区間が見つからなかった文については，発話区間が誤棄却されたと考えて，その前後の文に対応する発話区間にはさまれた無音区間に対応付ける．このようにすることにより，発話区間検出が上手くいかない音声データに対しても，前後の文との対応付けを利用して，文に対応する発話区間を発見することが可能になる．

第3章　会話音声とテキストの1文単位での対応付け

3.1　処理の流れ

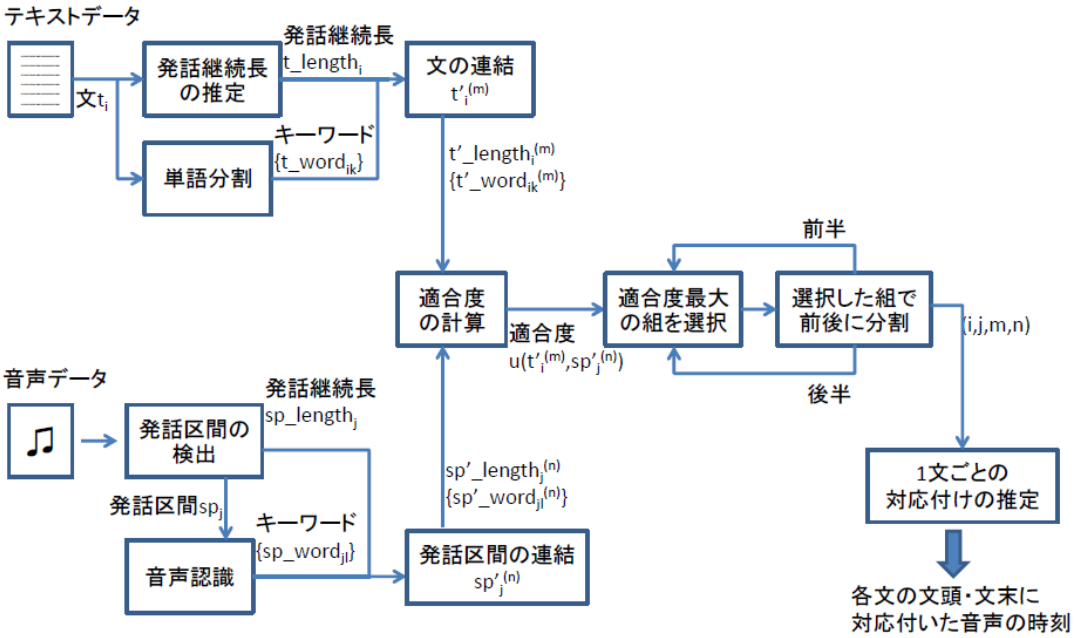

図 4: 提案手法の処理の流れ

提案手法の全体の処理の流れは以下のようになる (図 4).

1. テキストデータからの特徴抽出

   テキストデータでは複数文が一連になっているため，これを文に区切り，文集合 $T = \{t_i\}\ (i = 1, \ldots, I)$ を求める．ただし，I は文の数とする．文 t_i の

文字数から発話継続長 t_length_i を推定する．また，t_i に含まれる単語を，キーワード $\{t\_word_{ik}\}(k = 1, \ldots, K_i)$ とする．ただし，K_i は t_i に含まれる単語数とする．t_length_i と $\{t\_word_{ik}\}$ を，文 t_i を特徴づける特徴量とする．手順 1 の処理は，3.2 節で説明する．

2. 音声データからの特徴抽出
   音声データに対して発話区間検出を行い，発話区間集合 $SP = \{sp_j\}$ $(j = 1, \ldots, J)$ を求める．ただし、J は検出された発話区間の数とする．発話区間 sp_j の音声認識結果 $Rec(sp_j)$ に含まれる単語を，キーワード $\{sp\_word_{jl}\}(l = 1, \ldots, L_j)$ とする．ただし，L_j は $Rec(sp_j)$ に含まれる単語数とする．sp_length_j と $\{sp\_word_{jl}\}$ を，発話区間 sp_j を特徴づける特徴量とする．手順 2 の処理は，3.3 節で説明する．
3. 文および発話区間の連結
   隣り合う文及び発話区間同士を連結する場合のあらゆる連結パターンを生成する．文と発話区間が 1 対 1 に対応付けられる部分で連結が行われると，対応付け精度が低下する可能性がある．1 つの発話区間に対応する文の数や 1 文に対応する発話区間の数は限られており，不必要な連結を防ぐために，文及び発話区間の連結する最大数は制御すべきであると考えられる．そこで，連結する文の最大数 t_max と連結する発話区間の最大数 sp_max を設定し，連結パターンの生成を行う．
   まず，文 t_i について，隣り合う文同士を連結する場合のあらゆる連結パターンを生成する．$m + 1$ 個の隣り合う文 t_i〜t_{i+m} を連結したものを $t_i'^{(m)}(m = 0, \ldots, t\_max-1, i = 1, \ldots, I-m)$ とする．隣り合う文同士のあらゆる連結パターンの文集合を $T' = \{t_i'^{(m)} | (m = 0, \ldots, t\_max - 1, i = 1, \ldots, I - m)\}$ とする．同様にして，発話区間 sp_j について、隣り合う発話区間同士を連結する場合のあらゆる連結パターンを生成する．$n+1$ 個の隣り合う文 sp_j〜sp_{j+n} を連結したものを $sp_j'^{(n)}(n = 0, \ldots, sp\_max - 1, j = 1, \ldots, J - n)$ とする．隣り合う発話区間同士のあらゆる連結パターンの発話区間集合を $SP' = \{sp_j'^{(n)} | (n = 0, \ldots, sp\_max - 1, j = 1, \ldots, J - n)\}$ とする．手順 3 の処理は，3.4 節で説明する．
4. あらゆる連結パターンの文集合 T' とあらゆる連結パターンの発話区間集合 SP' の対応付け

手順 1 と 2 で抽出された特徴を用いて，$t_i'^{(m)}$ と $sp_j'^{(n)}$ の全組み合わせに対して適合度 $u(t_i'^{(m)}, sp_j'^{(n)})$ を計算する．そして，適合度が最も高い対応付けの組み合わせを確定する．次に，すでに確定された対応付けの前後の発話区間・文において，それぞれ適合度が最も高い対応付けの組み合わせを求め，それらの対応付けを確定する．以上の処理を，対応付けられる発話区間または文がなくなるまで繰り返す．手順 4 の処理は，3.5 節で説明する．

5. 連結された文の区切り位置の推定

手順 4 において複数文が連結された文が音声区間と対応付けられたものについて，その各文に対応する発話区間を求める．また，これまでの処理で発話区間が対応付けられなかった文について，対応する音声区間を推定する．手順 5 の処理は，3.6 節で説明する．

3.2 テキストデータからの特徴抽出

3.2.1 文の文字数による発話継続長の推定

文の文字数からその文の発話継続長を推定するために，発話継続長と文字数の相関性を調べた．NHK の語学学習番組「3ヶ月トピック英会話」の 2010 年 6 月 30 日と 7 月 7 日の放送回に含まれる英語の会話音声を対象とした．これらの長さは 24.50 秒と 63.35 秒であり，文の数は 15 文と 28 文であった．1 文字単位での音声の発話継続長を直接測定するのは困難であるため，文 t_i の文字数 t_num_i とその文に対応する音声の発話継続長 $true\_length_i$ を手動で抽出し，文ごとに 1 文字あたりの平均発話継続長を求め，それを用いて映像における全文の 1 文字あたりの平均発話継続長 $letter\_mean$ とその標準偏差 $letter\_stddev$ を以下の式により計算した．ただし，I は各映像に含まれる文の数とする．

$$letter\_mean = \frac{1}{I}\sum_{i=1}^{I}\frac{true\_length_i}{t\_num_i} \tag{1}$$

$$letter\_stddev = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}\left(\frac{true\_length_i}{t\_num_i} - letter\_mean\right)^2} \tag{2}$$

表 1 に結果を示す．

全データの 1 文字当たりに必要な時間は 0.052 秒であり，有効数字を考慮して，1 文字当たりに必要な時間は $letter\_time = 0.05$ とした．標準偏差は平均 0.027 秒であり，文によって発話の速さにずれがあることがわかる．このため，

表 1: 1 文字にかかる時間の平均と標準偏差（秒）

映像 ID	平均	標準偏差
6/30(1)	0.049	0.018
7/7(1)	0.054	0.031
全データ	0.052	0.027

文字数から推定された発話継続長には多少の誤差が生じる．

各文の発話継続長 t_length_j は，文 t_j の文字数 t_num_j と 1 文字当たりに必要な時間 $letter\_time$ を用いて，式 (3) により求める．

$$t\_length_j = letter\_time \times t\_num_j \tag{3}$$

テキストデータをピリオドで区切ることにより文 t_j を抽出する．文 t_j の文字数 t_num_j は，t_j の文字数を数えることにより求めた．ただし，スペースやカンマは文字数として含み，ピリオドは文字数として含まない．

3.2.2 単語分割によるキーワード抽出

文 t_j は英文であるため，単語と単語の間はスペースやカンマで区切られている．そこで，スペースやカンマで文 t_j を分割し，キーワード $\{t\_word_{jl}\}$ を得る．

3.3 音声データからの特徴抽出

3.3.1 発話区間検出

音声データにおいて，発話区間の話者の声の大きさは，非発話区間のノイズによる音の大きさに比べて大きいことが予想される．よって各時刻における音圧レベルを算出し，それが閾値以上の時は発話区間であると考えられる．ただし，摩擦子音 (主に/s/) はパワーが小さいため，音圧レベルだけを見ると，非発話区間に分類されてしまうことがある．ここで，摩擦子音は高周波数成分を多く含むため，音声信号の波形が 0 と交差する頻度は大きくなるという特徴がある．そこで，音声信号の振幅レベルとゼロ交差数により一定幅の時間フレームごとに音声/非音声を判定し，その結果を元に発話区間の検出を行う [5][6].

まず，一定時間幅の時間フレーム $f(f = 1, \ldots, F)$ ごとに，振幅レベルとゼロ交差数を求める．ただし，F は音声データの全フレーム数である．また，時間方向の順序を表すパラメータを t，時間フレーム内のサンプル数を N，音声信号を $x_f(t)(t = 1, \ldots, N)$ とする．

振幅レベルは，フレーム内の音声信号の二乗和の対数であり，式 (4) で表される．

$$lp_f = log\sum_{t=1}^{N} x_f(t)^2 \tag{4}$$

また，ゼロ交差数は，フレーム内の音声信号 $x_f(t)$ が 0 と交差する回数であり，式 (5) で表される．

$$Z_f = \sum_{t=1}^{N-1} (neg(x_f(t)x_f(t+1))) \tag{5}$$

ただし，

$$neg(x) = \begin{cases} 0\ (x \geq 0) \\ 1\ (x < 0) \end{cases} \tag{6}$$

である．

本研究では，振幅レベルとゼロ交差数を音声データ全体におけるそれぞれの最大値で割ることにより，0 ∼ 1 に正規化する．

振幅レベルの閾値 $Th_{lp}(0 \leq Th_{lp} \leq 1)$ とゼロ交差数の閾値 $Th_Z(0 \leq Th_Z \leq 1)$ を設定し，各フレームにおける振幅レベル lp_f とゼロ交差数 Z_f が式 (7) の条件を満たすならば，そのフレームを音声フレームと判定し，そうでなければ非音声フレームと判定する．

$$lp_f \geq Th_{lp} \text{ または } Z_f \geq Th_Z \tag{7}$$

この音声/非音声が判定されたフレームに対して，発話区間検出を行う．非発話区間が短すぎる場合は，検出誤りとして棄却する．また，意味のある発話はある程度の長さがあると仮定して，検出された発話区間が短すぎる場合にはノイズとして棄却する．

具体的には，以下のように発話区間を検出する (図 5)．まず，連続する非音声フレーム数 nsp_count が非発話区間の最小フレーム数 Th_{nsp} 以下であれば，前後の発話区間は連続しているとみなす．次に，連続する音声フレーム数 sp_count が，発話区間の最小フレーム数 Th_{sp} 以下であれば，その音声フレーム群は非発話区間として棄却する．

このようにして抽出された発話区間 sp_j に対して，時間フレーム幅 sp_width と sp_j のフレーム数 sp_frame_j を用いて，式 (8) により発話継続長 sp_length_j を計算する．

$$sp\_length_j = sp\_frame_j \times sp\_width \tag{8}$$

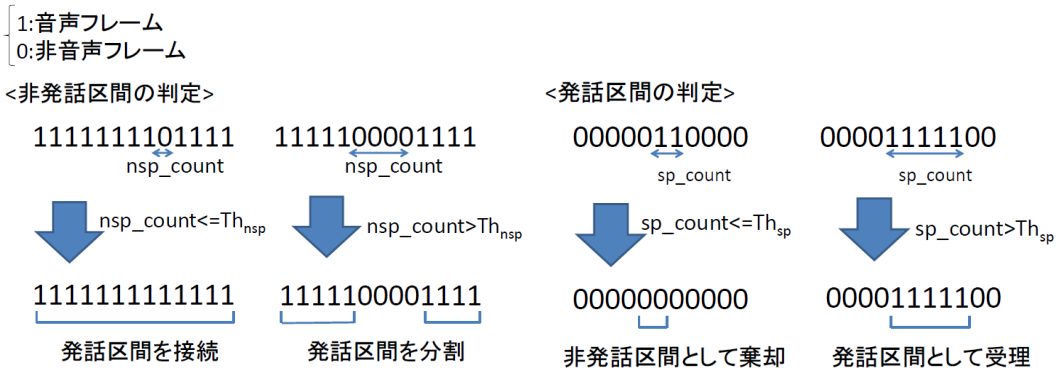

図 5: 発話区間検出

3.3.2 音声認識によるキーワード抽出

次に，各発話区間 sp_j に対して，その発話区間における発話内容をキーワードとして抽出するため，sp_j の音声認識を行う．本研究では，Google の音声認識エンジン [7] を利用した．発話区間 sp_j を手作業で音声認識エンジンに入力し，認識結果 $Rec(sp_j)$ をテキストとして得た．認識結果 $Rec(sp_j)$ は，単語ごとにスペースで区切って出力される．よって，$Rec(sp_j)$ をスペースにより単語単位に区切ったものを，発話区間 sp_j のキーワード $\{sp\_word_{jl}\}$ とする．音声認識に失敗して認識結果が得られなかった発話区間は，1 つもキーワードを持たないものとした．

3.4 複数文および複数発話区間の連結

3.2 節，3.3 節では，1 文 t_i や単一の発話区間 sp_j から特徴抽出を行う処理について述べた．3.4 節では，複数の文や発話区間を連結したパターンを生成する処理について説明する．

連結する文の最大数を t_max とし，隣り合う文 t_i から $t_{i+m}(m = 0, \ldots, t\_max-1)$ を連結したものを $t'^{(m)}_i$ とする (図 6)．$t'^{(m)}_i$ の発話継続長 $t'\_length_i{}^{(m)}$ は，連結した各文の発話継続長の和とし，式 (9) により求めた．また，$t'^{(m)}_i$ のキーワード $\{t'\_word_{ik'}{}^{(m)}\}$ は，連結した各文のキーワードの和集合とし，式 (10) により求めた．

$$t'\_length_i{}^{(m)} = t\_length_i + \cdots + t\_length_{i+m} \tag{9}$$

$$\{t'\_word_{ik'}{}^{(m)}\} = \bigcup_{\substack{p=i,\ldots,i+m \\ k=1,\ldots,K_p}} t\_word_{pk} \tag{10}$$

また，同様にして，連結する発話区間の最大数を sp_max とし，隣り合う発話区間 sp_i から $sp_{i+n}(n = 0, \ldots, sp\_max - 1)$ を連結したものを ${sp'_j}^{(n)}$ とする (図 7)．${sp'^{(n)}}_j$ の発話継続長 ${sp'_j}^{(n)}$ は，その発話区間の開始から終了までとし，式 (11) により求めた．また，${sp'_j}^{(n)}$ のキーワード $\{{sp'\_word_{jl'}}^{(n)}\}$ は，連結した各発話区間のキーワードの和集合とし，式 (12) により求めた．ただし，発話区間 sp_j の開始時刻を sp_start_j，終了時刻を sp_end_j とする．

$$ {sp'\_length_j}^{(n)} = sp\_end_{j+n} - sp\_start_j \tag{11} $$

$$ \{{sp'\_word_{jl'}}^{(n)}\} = \bigcup_{\substack{q=j,\ldots,j+n \\ l=1,\ldots,L_q}} sp\_word_{ql} \tag{12} $$

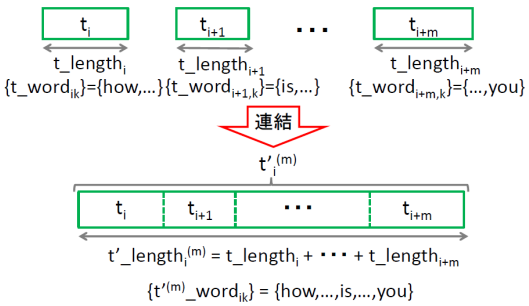

図 6: 文の連結

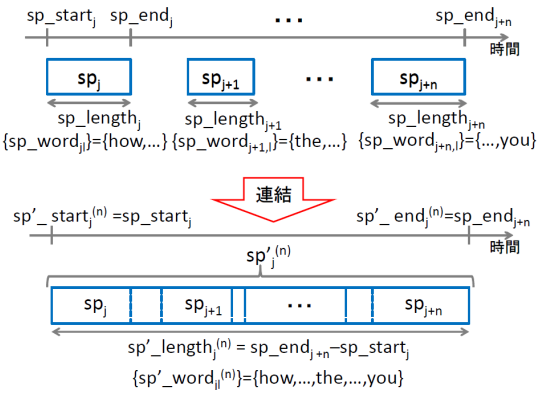

図 7: 発話区間の連結

3.5 音声データとテキストデータの適合度の計算

3.4 節で求めた特徴を用いて，文 ${t'_i}^{(m)}$ と発話区間 ${sp'_j}^{(n)}$ の適合度 $u({t'_i}^{(m)}, {sp'_j}^{(n)})$ を計算する．

まず，音声とテキストの発話継続長の差が小さいほど，適合度は高くなる．発話継続長の差による適合度 $u_{length}({t'_i}^{(m)}, {sp'_j}^{(n)})$ は，式 (13) で定義する．

$$ u_{length}({t'_i}^{(m)}, {sp'_j}^{(n)}) = -|{t'\_length_i}^{(m)} - {sp'\_length_j}^{(n)}| \tag{13} $$

また，音声とテキストのキーワードの一致数が多いほど，適合度は高くなる．キーワードの一致数による適合度 $u_{word}({t'_i}^{(m)}, {sp'_j}^{(n)})$ は，式 (14) で定義する．

$$ u_{word}({t'_i}^{(m)}, {sp'_j}^{(n)}) = \sum_{k'} \sum_{l'} match({t'\_word_{ik'}}^{(m)}, {sp'\_word_{jl'}}^{(n)}) \tag{14} $$

ただし，

$$match(t\_word_{ik}, sp\_word_{jl}) = \begin{cases} 1\ (t\_word_{ik} = sp\_word_{jl}) \\ 0\ (t\_word_{ik} \neq sp\_word_{jl}) \end{cases} \tag{15}$$

文や発話区間を連結することにより，その文や発話区間に含まれるキーワードが増えるため，連結するほどキーワードによる適合度が高くなりやすい．そのため，連結せずに正しい対応付けが取れていた部分に対して，不必要な連結が起こり，間違った対応付けが行われてしまう可能性がある．また，連結する文や発話区間に含まれるキーワードが多いほど，連結による影響が大きいと考えられる．そこで，$t_i'^{(m)}$ の連結数 m とキーワード数 $K_i'^{(m)} = \sum_{a=i}^{i+m} K_a$($K_i$ は文 i のキーワード数)、$sp_j'^{(n)}$ の連結数 n とキーワード数 $L_j'^{(n)} = \sum_{b=j}^{j+n} L_b$($L_j$ は発話区間 j のキーワード数) を用いて，連結によるペナルティ $P(t_i'^{(m)}, sp_j'^{(n)})$ を，式 (16) で定義する．

$$P(t_i'^{(m)}, sp_j'^{(n)}) = m \times K_i'^{(m)} + n \times L_j'^{(n)} \tag{16}$$

$t_i'^{(m)}$ と $sp_j'^{(n)}$ の適合度を $u(t_i'^{(m)}, sp_j'^{(n)})$ とする．発話継続長の差による適合度 $u_{length}(t_i'^{(m)}, sp_j'^{(n)})$ に対する重みを w_{length}，キーワードの一致数による適合度 $u_{word}(t_i'^{(m)}, sp_j'^{(n)})$ に対する重みを w_{word}，連結によるペナルティ $P(t_i'^{(m)}, sp_j'^{(n)})$ に対する重みを $w_{connect}$ とすると，適合度 $u(t_i'^{(m)}, sp_j'^{(n)})$ を式 (17) で定義する．

$$\begin{aligned} u(t_i'^{(m)}, sp_j'^{(n)}) \quad = \quad & w_{length} \times u_{length}(t_i'^{(m)}, sp_j'^{(n)}) + w_{word} \times u_{word}(t_i'^{(m)}, sp_j'^{(n)}) \\ & - w_{connect} \times P(t_i'^{(m)}, sp_j'^{(n)}) \end{aligned} \tag{17}$$

連結を含む文 T' と発話区間集合 SP' の対応付けは，以下のようなアルゴリズムで行われる（図 8）．

1. $t_i'^{(m)}$ と $sp_j'^{(n)}$ のすべての組み合わせについて適合度 $u(t_i'^{(m)}, sp_j'^{(n)})$ を計算する．
2. $u(t_i'^{(m)}, sp_j'^{(m)})$ が最大となる組み合わせ $(i_{max}, j_{max}, m_{max}, n_{max})$ を選択する．
3. 選択された組み合わせより前の部分と後ろの部分に分割する．
4. 前後に対してそれぞれ上記の処理を繰り返す．処理の範囲内に発話区間または文のどちらかがなければ，その範囲の処理は終了する．

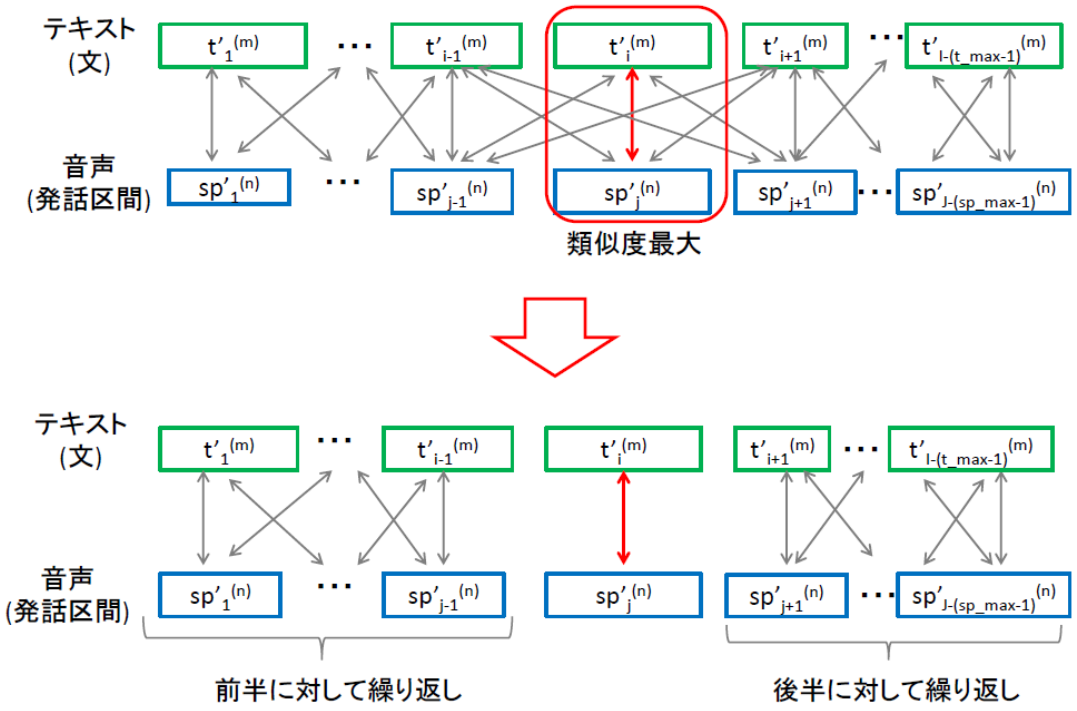

図 8: 音声の発話区間とテキストの文の対応付け

上記の手順 1 において，適合度が同じ値である $t_i'^{(m)}$ と $sp_j'^{(n)}$ の組み合わせが 2 つ以上存在した場合，文及び発話区間の順序は変わらないため，正しい対応付けの組み合わせの文と発話区間は，全体に対して同じような相対位置位置にあると考えられる．そのため，適合度が同じ値の組み合わせは，文 t_i の全文における相対位置（i/I）と発話区間 sp_j の全発話区間における相対位置（j/J）の距離が最も小さいものを選択する．

3.6 音声データにおける文の区切り位置の推定

前節の処理では，複数文が連結されたパターンに対して発話区間が対応付いている場合がある．そこで，それを構成する複数文のうち，各文の文頭と文末が，その複数文に対応付けられた発話区間においてどの時刻に対応しているか求めることにより，音声とテキストの文単位の対応付けを行う．

以下，$t_i'^{(m)}$ と $sp_j'^{(n)}$ が対応付けられた場合を考える．文 t_i の開始時刻及び終了時刻を t_start_i と t_end_i，発話区間 sp_j の開始時刻及び終了時刻を sp_start_j と sp_end_j とする．

$n = 0$ のとき，1 文に対する発話区間が求められているため，$t_i'^{(0)} = t_i$ の文頭

に $sp'_j{}^{(n)}$ の最初の発話区間 sp_j の開始時刻を対応付け，t_i の文末に $sp'_j{}^{(n)}$ の最後の発話区間 sp_{j+n} の終了時刻を対応付ける (式 (18)，式 (19))．

$$t\_start_i = sp\_start_j \tag{18}$$

$$t\_end_i = sp\_end_{j+n} \tag{19}$$

$n \neq 0$ のとき，複数文の連結に対する発話区間が求められているため，さらに 1 文単位の対応付けを推定する必要がある．まず，$t'_i{}^{(m)}$ の最初の文 t_i の文頭に $sp'_j{}^{(n)}$ の最初の発話区間 sp_j の開始時刻を対応付け，$t'_i{}^{(m)}$ の最後の文 t_{i+m} の文末に $sp'_{j(n)}$ の最後の発話区間 sp_{j+n} の終了時刻を対応付ける（式 (20)，式 (21)）．

$$t\_start_i = sp\_start_j \tag{20}$$

$$t\_end_{i+m} = sp\_end_{j+n} \tag{21}$$

さらに，$t_i{}^{(m)}$ と $sp_j{}^{(n)}$ の発話継続長の比と各文の発話継続長の推定値 t_length_i を用いて，文の連結部分に対応する音声データの時刻を推定する．発話継続長の比 $length\_rate\left(t'_i{}^{(m)}, sp'_j{}^{(n)}\right)$ は，式 (22) により計算される．

$$length\_rate\left(t'_i{}^{(m)}, sp'_j{}^{(n)}\right) = \frac{sp'\_length_j{}^{(m)}}{t'\_length_i{}^{(n)}} \tag{22}$$

$0 \leq c < n$ に対して以下の計算 (式 (23)，式 (24)) を行い，文の連結部分に対応する音声データの時刻を推定する．

$$t\_end_{i+c} = t\_start_{i+c} + t\_length_{i+c} \times length\_rate\left(t'_i{}^{(m)}, sp'_j{}^{(n)}\right) \tag{23}$$

$$t\_start_{i+c+1} = t\_end_{i+c} \tag{24}$$

また，発話区間が対応付けられなかった文は，発話区間検出に失敗したものとみなし，前後の文に対応付けられた発話区間にはさまれた区間を対応付ける．

第 4 章　実験及び考察

4.1　実験環境

本研究の提案手法を用いた音声とテキストの対応付け実験について述べる．3 章の処理により，各文の文頭及び文末の与えられた映像クリップ中での時刻が出力されるため，この時刻と手動で抽出した正解時刻のずれを評価した．本研

究で想定するような授業での文単位の再生を考えると，ある程度の誤差を許容し，文頭は許容誤差の分だけ前へ，文末は許容誤差の分だけ後ろにずらして対応付けることにより，少なくとも該当する文は損失なく含むようにすることができるため，多少のずれであれば実用可能であると考えられる．そこで，0.5 秒及び 1 秒以内のずれを許容した場合の対応付け精度を文単位の対応付け精度として評価する．テキストデータの文の数を $text\_num$ とすると，文頭及び文末の数は $text\_num \times 2$ で表される．さらに，対応付け結果の時刻と正解時刻のずれが許容範囲内にある文頭及び文末の数を $correct\_match$ とすると，対応付け精度 $accuracy$ は式 (25) で求められる．

$$accuracy = \frac{correct\_match}{text\_num \times 2} \tag{25}$$

表 2: 実験に用いた映像の長さと文の数

映像 ID	映像の長さ (秒)	文の数
6/30(1)	24.5	15
6/30(2)	30.9	17
6/30(3)	37.9	15
6/30(4)	37.7	15
7/7(1)	63.3	28
7/7(2)	32.4	18
7/7(3)	54.3	21
7/7(4)	71.3	26
7/14(1)	48.6	23
7/14(2)	53.3	23
7/14(3)	35.1	16
7/14(4)	51.7	23

実験には，NHK の語学学習番組「3ヶ月トピック英会話」の英語による会話シーンを用いた．1 回の放送につきこのような会話シーンが 4 本含まれており，2010 年 6 月 30 日 ~2010 年 7 月 14 日に放送された 3 回分，計 12 本の映像クリップを対象に，実験を行った．この各会話シーンの長さと会話文の数を表 2 に示す．

発話区間検出における閾値を定めるために予備実験を行い，振幅レベルの閾値 $Th_{lp} = 0.5$，ゼロ交差数の閾値 $Th_Z = 0.5$ とし，時間フレーム f の幅は 0.1

秒とした．発話区間検出，音声認識，文字数による推定発話継続長の結果を表3，表4，表5にそれぞれ示す．

表3: 発話区間検出の精度 (%)

映像ID	適合率	再現率	F値
6/30(1)	75.3	89.2	81.7
6/30(2)	92.9	83.6	88.0
6/30(3)	94.1	90.7	92.4
6/30(4)	46.2	100.0	63.2
7/7(1)	71.5	89.0	79.3
7/7(2)	92.4	93.2	92.8
7/7(3)	66.0	87.0	75.1
7/7(4)	81.4	96.9	88.5
7/14(1)	89.3	86.7	87.9
7/14(2)	87.4	85.8	86.6
7/14(3)	90.0	95.6	92.7
7/14/(4)	79.5	96.1	87.0
平均	80.5	91.1	84.6

表4: 音声認識の結果 (%)

映像ID	適合率	再現率	F値
6/30(1)	29.4	7.7	12.2
6/30(2)	35.3	14.1	20.2
6/30(3)	37.9	10.5	16.4
6/30(4)	50.0	4.1	7.6
7/7(1)	46.1	29.7	36.1
7/7(2)	62.3	40.9	49.4
7/7(3)	54.5	28.2	37.2
7/7(4)	54.6	37.9	44.7
7/14(1)	17.6	6.6	9.6
7/14(2)	35.2	17.2	23.1
7/14(3)	39.2	29.5	33.7
7/14(4)	41.1	27.0	32.6
平均	41.9	21.9	26.9

発話区間検出では，複数文が1つの発話区間として検出された部分や，1文が複数の発話区間として検出された部分がある．6/30(4)のデータは，高周波数成分のノイズが全体に渡って含まれているため，他のデータに比べて，発話区間検出の精度が悪かった．音声認識の精度は低く，ある1文に対する認識率のみ高いといったように，偏りが見られた．文字数による発話継続長の推定結果は，ずれの大きさが平均0.29秒であり，6文字ほどのずれであった．また，標準偏差は0.36秒であり，文ごとに発話の速さが異なることがわかる．

4.2 1文単位の対応付けの精度評価

4.2.1 適合度の重みの検証

発話継続長の差による適合度に対する重みw_{length}，キーワードの一致数による適合度に対する重みw_{word}，連結によるペナルティに対する重み$w_{connect}$の3つの重みを定めるために，重みを変化させたときの対応付け精度を調べた．

表 5: 文字数による発話継続長の推定結果のずれ (秒)

映像 ID	平均	標準偏差	最大値
6/30(1)	0.20	0.19	0.60
6/30(2)	0.30	0.97	4.10
6/30(3)	0.23	0.22	0.80
6/30(4)	0.20	0.26	1.00
7/7(1)	0.29	0.25	0.85
7/7(2)	0.29	0.28	1.00
7/7(3)	0.21	0.17	0.60
7/7(4)	0.38	0.40	1.40
7/14(1)	0.38	0.48	2.45
7/14(2)	0.36	0.49	2.05
7/14(3)	0.18	0.16	0.60
7/14(4)	0.48	0.50	2.40
全体	0.29	0.36	1.49

まず，w_{length} と w_{word} の割合を定めるために，$w_{length} = 1$ とし，w_{word} を変化させたときの対応付け精度を調べた．ここでは文及び発話区間の連結を考慮せず，連結する文及び発話区間の最大数 t_{max} と sp_{max} は 1 に固定した．表 6 に結果を示す.

表 6: 重み w_{word} を変化させたときの対応付け精度 (%)

w_word	0.5 秒以内	1 秒以内
0.1	27.7	34.1
0.5	38.1	47.3
1	38.1	47.7
1.5	37.9	47.5
2	37.9	47.5

w_{word} を増やしていくと，$w_{word} = 1$ が最も対応付け精度が良く，1 を境に対応付け精度が低下する傾向に変わることがわかる．よって，$w_{length} = 1$、$w_{word} = 1$ とする.

次に，$w_{connect}$ を定めるために，他のパラメータの値を固定し，$w_{connect}$ のみ

を変化させたときの対応付け精度を調べた．ここでは，文及び発話区間の連結を考慮する最大数は，$t_{max} = 5$、$sp_{max} = 5$ とした．結果を表 7 に示す．

表 7: 重み $w_{connect}$ を変化させたときの対応付け精度 (%)

w_connect	0.5 秒以内	1 秒以内
0.03	34.3	46.2
0.04	36.3	46.7
0.05	43.3	55.2
0.06	43.9	53.7
0.07	40.9	50.0

$w_{connect}$ を増やしていくと，$w_{connect} = 0.05$ のとき 1 秒以内のずれを許容した場合の対応付け精度が最も高くなっている．よって，$w_{connect} = 0.05$ とする．

4.2.2 連結を考慮する長さに対する検証

文の連結を考慮する最大数 t_{max} と発話区間の連結を考慮する sp_{max} を定めるために，t_{max} 及び sp_{max} を変化させたときの対応付け精度を調べた．

表 8: 文の連結の最大数 t_{max} を変化させたときの対応付け精度 (%)

t_{max}	0.5 秒以内	1 秒以内
1	38.1	47.7
2	40.1	49.6
3	42.6	53.9
4	45.9	56.8
5	45.5	57.4
6	44.0	56.2
7	43.8	56.2

表 9: 発話区間の連結の最大数 sp_{max} を変化させたときの対応付け精度 (%)

sp_{max}	0.5 秒以内	1 秒以内
1	41.1	51.4
2	42.1	52.6
3	43.5	54.6
4	43.8	55.7
5	43.8	55.7
6	43.8	55.7
7	43.8	55.7

まず，sp_{max} を固定し，t_{max} を変化させたときの対応付け精度を調べた．検出された発話区間の最大数は 25 なので，発話区間をすべて連結する場合まで考慮して $sp_{max} = 25$ とした．結果を表 8 に示す．

同様にして，t_{max} を固定し，sp_{max} を変化させたときの対応付け精度を調べ

た．文の最大数は 28 なので，文をすべて連結する場合まで考慮して $t_{max} = 28$ とした．結果を表 9 に示す．

t_{max} の値を 1 から順に増やしていくと，$t_{max} = 5$ のとき対応付け精度が最大となり，その後は徐々に低下する．また，sp_{max} の値を 1 から順に増やしていくと，$sp_{max} = 4$ のとき最大となり，その後は変化しない．よって，$t_{max} = 5$, $sp_{max} = 4$ と定める．

4.2.3 提案手法の精度評価

4.2.1 節，4.2.2 節の結果により定めたパラメータの値を用いて，提案手法による音声とテキストの対応付け実験を行った．連結を考慮することによる対応付け精度への影響を調べるために，連結を考慮しない場合 ($t_{max} = 1$，$sp_{max} = 1$) の結果と，連結を考慮した場合 ($t_{max} = 5$，$sp_{max} = 4$) の結果を比較する．結果を表 10 に示す．

表 10: 提案手法の対応付け精度 (%)

映像 ID	0.5 秒以内		1 秒以内	
	連結なし	連結あり	連結なし	連結あり
6/30(1)	23.3	26.7	30.0	36.7
6/30(2)	32.4	38.2	47.1	47.1
6/30(3)	20.0	20.0	30.0	30.0
6/30(4)	3.3	3.3	6.7	6.7
7/7(1)	60.7	69.6	69.6	82.1
7/7(2)	75.0	69.4	86.1	86.1
7/7(3)	57.1	52.4	61.9	54.7
7/7(4)	13.5	36.5	26.9	51.9
7/14(1)	10.9	39.1	13.0	65.2
7/14(2)	54.3	60.9	71.7	69.6
7/14(3)	78.1	84.4	87.5	100.0
7/14(4)	28.3	45.7	41.3	58.7
全体	38.5	47.3	48.3	59.6

今回の実験では，12 本の映像クリップのうち 7 本に対して，連結を考慮することによって対応付け精度が向上した．また，全体としての精度も 0.5 秒以内のずれを許容した場合は 38.5%から 47.3%に 8.8 ポイント向上，1 秒以内のずれ

を許容した場合は 48.3%から 59.3%に 11.3 ポイント向上した．

実際に，複数文が 1 つの発話区間として検出された発話区間の数を，表 11 に示す．

表 11: 複数文が連結して検出された発話区間の数

映像 ID	1 文単独	複数文を含む発話区間の数						複数の発話区間を含む文の数
		2 文	3 文	4 文	5 文	6 文	7 文	
6/30(1)	4	0	3	0	0	0	0	2
6/30(2)	5	4	1	0	0	0	0	1
6/30(3)	6	1	0	0	1	0	0	2
6/30(4)	0	0	1	0	1	0	1	0
7/7(1)	17	1	1	1	0	0	0	1
7/7(2)	6	4	1	0	0	0	0	1
7/7(3)	11	1	0	0	0	1	0	0
7/7(4)	11	0	1	1	0	1	0	2
7/14(1)	11	1	1	0	0	0	0	4
7/14(2)	9	1	0	1	0	0	0	4
7/14(3)	12	1	0	0	0	0	0	0
7/14/(4)	12	6	0	0	0	0	0	0

各文ごとの対応付け結果を調べると，複数文が 1 つの発話区間として検出された部分や，1 文が複数の発話区間に分割されて対応付けられた部分で対応付けの改善が見られた．また，連結部分の対応付けが改善されることで，その前後の部分も正しい対応付けとなる場合があった．

例として，7/14(1) のデータの対応付け結果の前半部分を図 9 に示す．図中の矢印は対応付け結果を表し，番号は対応付けが決定した順番を表す．連結を考慮しなかった場合，t_7 が sp_{14} に間違って対応付けられたため，t_8 は本来対応付けられるべき sp_5 に対応付けられることはなく，sp_{15} 以降の間違った範囲で対応付けが行われた．同様にして，t_1 が sp_5 に間違って対応付けられたため，t_2〜t_6 は間違った対応付けが行われた．この結果に対し，連結を考慮した場合，t_7，t_8 が sp_4，sp_5 に正しく対応付けられており，t_8 以降は正しい範囲で対応付けが行われた．また，t_3 が先に対応付けられたことにより，t_1 が sp_5 に対応付けられることなく，正しい対応付けが求められた．

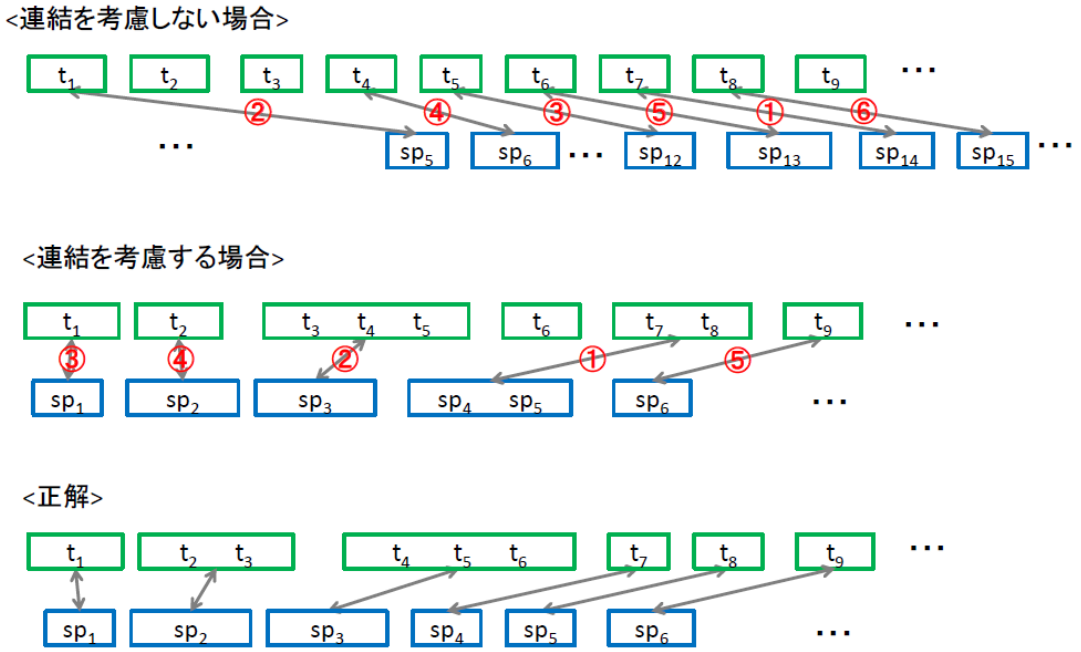

図 9: 7/14(1) の対応付け結果

一方，連結を考慮した場合の方が精度が低下したデータが 12 本のうち 2 本あった．これは，”I”や”this”といったようなキーワードが異なる複数の文に含まれており，その文を連結することにより，正しい対応付けを行った場合よりも適合度が高くなり，間違った対応付けが行われたことが原因であると考えられる．

連結を考慮しない場合も考慮する場合も 6/30(4) のデータの対応付け精度がかなり低いのは，発話区間検出の精度が低いことが原因である．6/30(4) のデータは 15 文からなっているが，それらの文が 3 つの発話区間として検出されていた．このため，1 つの発話区間はたくさんの文を含み，音声認識の精度も低く，正しく対応付けられなかった．

また，表 4 と表 10 より，音声認識の精度が比較的高い 7/7(1)∼(4) のデータは対応付け精度が高く，音声認識の精度が低い 6/30(1)∼(4) のデータは対応付け精度が低くなっており，対応付け結果が音声認識の精度に依存していることがわかる．例えば，6/30(1) のデータは，後半の文に対応する音声区間において，正しく音声認識されたキーワードが 1 つもないため，発話継続長が同じような発話区間が間違って対応付けられている．また，7/14(1) のデータでは，音声認識で誤認識されたキーワードが，その発話区間に対応する文とは異なる文に含

まれており，間違った対応付けの適応度が高くなった．

さらに，正しい対応付けが行われた場合でも，検出された発話区間の先頭または末尾が欠損していたために，正解時刻とのずれが生じている文頭・文末があった．

提案手法では，音声認識の認識結果が得られなかった場合，発話継続長のみを用いて対応付けるため，大きくずれた対応付けが行われていた．ある対応付けが大きくずれてしまうと，それ以降の対応付けが間違った範囲で行われ，精度が低下する．このような大きくずれた対応付けを防ぐためには，テキスト全体における文の位置と音声全体における発話区間の位置の差を考慮することが必要であると考えられる．そこで，文及び発話区間の位置を考慮した適合度を追加し，実験を行った．これらの詳細は，付録 A.1 に示した．

第 5 章　結論

本研究では，複数文の連結を考慮した会話音声とテキストの文単位の対応付け手法を提案した．テキストの文及び音声の発話区間の連結を考慮することにより，文と文の間に無音区間が存在しない場合や，1 文の中に無音区間が存在する場合に対処した．また，文と発話区間に対して，それぞれ発話継続長とキーワードの抽出を行い，適合度が最大となる対応付けから決定していくことにより，言い淀みなどテキストには書き起こされない発話が存在するという問題に対処した．

今後の課題としては，まず，発話区間検出の精度向上が求められる．発話区間検出の精度が低いと音声認識の精度も低く，正しい対応付けを求めることができない．次に，本手法では，すべての単語をキーワードとして同等に適合度に反映させたが，複数の文に含まれる単語がある場合，間違った対応付けの適合度も高くなる．また，完全に一致したキーワードの数のみを用いてキーワードによる適合度の計算を行ったが，発音が似ているために音声認識で間違われやすい単語や，単語の一部が一致した場合についても適合度に反映させる必要があると考えられる．さらに，DP マッチングを用いて文の全体における位置を考慮した場合の対応付け手法と本手法の比較をすることも重要である．

今回は，1 番組に対して本手法の検証を行ったが，他の番組に対しても本手法をの検証を行う必要がある．また，本手法によってテキストと音声の対応が

取られた映像を用いてシステム構築を行い，実際の授業で試用してもらい，評価を行う必要がある．

謝辞

本研究を行うにあたり，多くのご教示と熱心なご指導を賜りました美濃導彦教授，椋木雅之准教授に深く感謝いたします．また，本報告書の作成において多くの助言をいただきました社会情報学専攻の山肩洋子准教授に深く感謝いたします．最後になりましたが，日頃より様々な面でご協力いただきました美濃研究室の皆様に深く感謝いたします．

参考文献

[1] 黒田智也, 椋木雅之, 浅田尚紀: 語学学習番組の映像構造化に基づく教材映像提示インタフェースの作成, 電子情報通信学会技術研究報告, Vol. 109, No. 149, pp. 13–18 (2009).

[2] 柳沼良知, 坂内正夫: DP マッチングを用いたドラマ映像・音声・シナリオ文書の対応付け手法の一提案, 電子情報通信学会論文誌 D, Vol. J79-D-2, No. 5, pp. 747–755 (1996).

[3] 小林聡, 田中敬志, 森一将, 中川聖一: 字幕付きテレビニュース放送を素材とした語学学習教材作成システム, 人工知能学会論文誌, Vol. 17, No. 4, pp. 500–509 (2002).

[4] 柳沼良知, 鈴木一史, 清水康敬: 教育用映像コンテンツのデータベース化のためのテロップ認識手法の検討, 電子情報通信学会技術研究報告, Vol. 105, No. 374, pp. 13–18 (2005).

[5] 木田祐介, 河原達也: 複数特徴の重み付き統合による雑音に頑健な発話区間検出, 電子情報通信学会論文誌 D, Vol. J89-D, No. 8, pp. 1820–1828 (2006).

[6] 久富慎二, 松本哲也, 竹内義則, 工藤博章, 大西昇: 事前学習を用いないオンラインでの話者識別, 電子情報通信学会技術研究報告, Vol. 107, No. 551, pp. 145–150 (2008).

[7] Wichary, M. and the Google Chrome team: HTML5 Presentation. http://slides.html5rocks.com/.

付録

A.1　文及び発話区間の位置関係を考慮した対応付け

A.1.1　位置関係を考慮した適合度の計算

本研究の実験結果より，対応付け精度に改善の余地が残されているデータに対して，さらに詳しく考察を行った．例として，6/30(1) のデータの対応付け結果の 1 部を図 A.1 に示す．

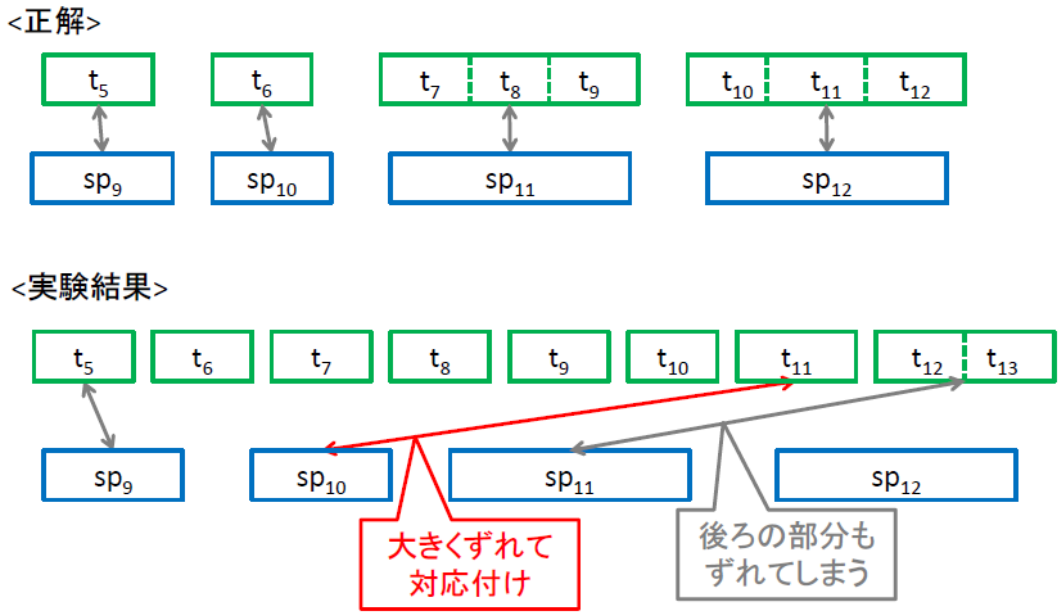

図 A.1: 大きくずれた対応付け結果

これらのデータでは，ある部分が大きくずれて対応付けられたため，それ以降も間違った範囲で対応付けが行われ，対応付け精度が低下している．この問題は，文や発話区間の位置を考慮していないことが原因であると考えられる．そこで，文字数から推定した文の開始時刻 t_start_i と発話区間の開始時刻 sp_start_j の差を適合度に反映させる．

t_start_i は，文字数による推定発話継続長を用いて，以下のようにして計算する．文 t_{x_1}〜t_{x_2} と発話区間 sp_{y_1}〜sp_{y_2} の対応付けを求める場合について説明する．

まず，t_{x_1}〜t_{x_2} の推定発話継続長の総和 $t\_total\_time_{x_1x_2}$ を求める．

$$t\_total\_time_{x_1x_2} = \sum_{i=x_1}^{x_2} t\_length_i \tag{A.1}$$

次に，対応付けを考える音声区間の長さ $sp\_total\_time_{y_1y_2}$ を求める．

$$sp\_total\_time_{y_1y_2} = sp\_start_{y_2+1} - sp\_end_{y_1-1} \tag{A.2}$$

ただし，sp_{y_1} が最初の発話区間のとき $(y_1 = 1)$，$sp\_end_{y_1-1} = 0$ とする．また，sp_{y_2} が最後の発話区間のとき $(y_2 = J)$，$sp\_start_{y_2+1} = total\_time$ とする．ここで，sp_end_j は sp_j の終了時刻，$total\_time$ は映像の長さ，J は検出された発話区間数とする．

文と文の間の時間 $pause$ は，式 (A.3) により求め，均等に割り当てる．

$$pause = \frac{sp\_total\_time_{y_1y_2} - t\_total\_time_{x_1x_2}}{x_2 - x_1 + 1} \tag{A.3}$$

各文の開始時刻 $t\_start_i (i = x_1, \ldots, x_2)$ は，式 (A.4) により求める．

$$t\_start_i = \begin{cases} sp\_end_{y_1-1} + pause \ (i = x_1) \\ t\_start_{i-1} + t\_length_{i-1} + pause \ (i = x_1 + 1, \ldots, x_2) \end{cases} \tag{A.4}$$

文 $t_i'^{(m)}$ と発話区間 $sp_j'^{(n)}$ の開始時刻の差による適合度 $u_{position}(t_i'^{(m)}, sp_j'^{(n)})$ を式 (A.5) で定義する．

$$u_{position}(t_i'^{(m)}, sp_j'^{(n)}) = -|t\_start_i - sp\_start_j| \tag{A.5}$$

$u_{position}(t_i'^{(m)}, sp_j'^{(n)})$ は，1 つ対応付けを決定する度に計算し直す．

開始時刻の差を反映させた新しい適合度 $u_{new}(t_i'^{(m)}, sp_j'^{(n)})$ は，式 (A.6) で定義する．

$$u_{new}(t_i'^{(m)}, sp_j'^{(n)}) = u(t_i'^{(m)}, sp_j'^{(n)}) + weight_{position} \times u_{position}(t_i'^{(m)}, sp_j'^{(n)}) \tag{A.6}$$

ただし，$weight_{position}$ は，開始時刻の差による適合度に対する重みとする．

A.1.2 実験結果

文及び発話区間の位置関係を考慮した新しい適合度を用いた場合，検出された発話区間と各文を手動で対応付けた場合の対応付け精度を，表 A.1 に示す．手動で対応付けた場合の精度は，提案手法の上限値となる．

位置関係を反映させた適合度を用いることにより，12 本中 8 本のデータで対応付け精度が改善した．全体としての精度も，0.5 秒以内のずれを許容する場合は 47.3%から 52.7%へ 5.4 ポイント，1 秒以内のずれを許容する場合は 59.6%か

ら 69.2%へ 9.6 ポイント改善した．特に大幅な改善が見られたデータは，大きくずれて対応付けが行われたため，それ以降の対応付けも間違っていたものであった．

逆に，12 本中 2 本のデータでは，対応付け精度が低下した．これは，文字数から推定した各文の開始時刻の誤差が大きく，位置を考慮しない場合に正しく対応付けられていたものが，間違って対応付けられたためであると考えられる．

また，正しい音声認識結果が得られなかった場合，発話継続長と開始時刻による適合度を用いて対応付けを行う．このとき，文字数から推定された文の発話継続長と開始時刻にずれが生じ，間違った対応付けが起こる場合がある．そのため，上限精度に達しなかったと考えられる．

表 A.1: 位置を考慮した適合度による対応付け精度 (%)

映像 ID	0.5 秒以内				1 秒以内			
	連結なし	連結あり			連結なし	連結あり		
	位置なし		位置あり	手動	位置なし		位置あり	手動
6/30(1)	23.3	26.7	43.3	83.3	30.0	36.7	50.0	90.0
6/30(2)	32.4	38.2	52.9	97.1	47.1	47.1	67.6	100.0
6/30(3)	20.0	20.0	56.7	90.0	30.0	30.0	76.7	93.3
6/30(4)	3.3	3.3	13.3	36.7	6.7	6.7	13.3	43.3
7/7(1)	60.7	69.6	78.6	76.8	69.6	82.1	87.5	83.9
7/7(2)	75.0	69.4	77.8	91.7	86.1	86.1	86.1	97.2
7/7(3)	57.1	52.4	35.7	73.8	61.9	54.7	40.5	92.9
7/7(4)	13.5	36.5	28.8	80.8	26.9	51.9	50.0	96.2
7/14(1)	10.9	39.1	41.3	78.3	13.0	65.2	73.9	87.0
7/14(2)	54.3	60.9	58.7	76.1	71.7	69.6	91.3	97.8
7/14(3)	78.1	84.4	84.4	93.8	87.5	100.0	100.0	100.0
7/14(4)	28.3	45.7	56.5	84.8	41.3	58.7	78.3	97.8
全体	38.5	47.3	52.7	80.2	48.3	59.6	69.2	90.6