K-027

語学学習番組を教材利用するための会話音声とテキストの対応付け
Aligning Conversational Speeches to Textbook Contents for Using Skits in Language Study TV Programs

清水 渚佐† Nagisa Shimizu　山肩 洋子† Yoko Yamakata　椋木 雅之‡ Masayuki Mukunoki　美濃 導彦‡ Michihiko Minoh

1. はじめに

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．コミュニケーションの基本は会話である．会話表現を効果的に学ぶために，語学の授業に会話映像を取り入れることが考えられる．このとき，授業で取り扱っている単語や熟語を含む文に対応する部分の会話映像を選択して再生することで，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，各文が映像のどの部分に対応しているかを知る必要がある．

映像は音声と動画が時間的に同期したものなので，文と映像の対応付けでは，映像中の音声と会話音声を書き起こした文の列（以降テキストと呼ぶ）の対応付けが得られれば，自動的に映像と文の対応付けも求まる．また，会話表現を学ぶには，単語のような細かな単位ではなく，文単位での学習が適している．そこで，本研究では，音声とテキストを文単位で対応付けることを目的とする．本研究で対象とする映像は，日常会話を取り扱った語学学習番組の会話シーンとし，会話シーンのテキストは与えられているとする．

2. 1 文再生のための映像インデキシング

2.1 従来研究の問題点

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[1]は，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を提案した．小林ら[2]は，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を提案した．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，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，文と文の間にはポーズ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．

2.2 複数文の連結を考慮した対応付け

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．

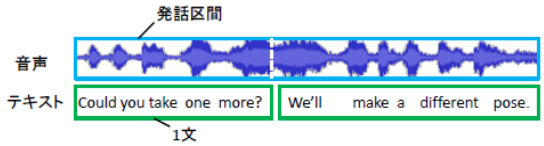

図1: 複数文の連続発話

図 2: 文中のポーズ

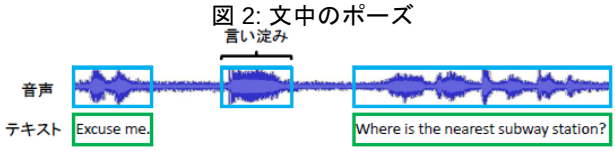

図 3: テキストに書き起こされない発話

複数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ポーズが存在しない場合，発話区間検出によって検出された 1 区間(発話区間)に対して複数の文を対応付ける必要がある(図 1)．逆に，1 文を発話する間にポーズが存在する場合，1 文に対して複数の発話区間を対応付ける必要がある(図 2)．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を生成し，それらの間で対応付けを行うことにより，連続発話や文中のポーズの問題に対処する．

さらに，言い淀み等の対応すべき文がテキストに存在しない発話に対しては，文と発話区間の適合度を計算し，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，他の文と間違って対応付けられることを防ぐ(図 3).

3. 会話音声とテキストの 1 文単位の対応付け

3.1 テキスト及び音声からの特徴抽出

テキストと音声を文および発話区間に分割し，それぞれから発話継続長とキーワードを抽出して特徴として用いる．以下，分割と特徴抽出の方法について説明する．

テキストは，ピリオドで区切ることで文に分割する．文の発話継続長は，1 文字当たりに必要な時間と文の文字数を乗算することによって推定する．文のキーワードは，その文に含まれる単語の集合とする．

音声は，対数パワーとゼロ交差数の閾値処理によって，発話区間に分割する．発話区間の発話継続長は，その発話区間の開始時刻と終了時刻の差とする．発話区間のキーワードは，Google の音声認識エンジン[3]を使ってその発話区間を音声認識した結果得られる単語の集合とする．

† 京都大学大学院情報学研究科　Graduate School of Informatics, Kyoto University
‡ 京都大学学術情報メディアセンター　Academic Center for Computing and Media Studies, Kyoto University

3.2 複数文及び複数発話区間の連結

文及び発話区間の連結を考慮して対応付けを行うために，隣り合う複数の文及び発話区間を連結したパターンを生成する．文の連結パターンの発話継続長は連結された文の発話継続長の総和とし，キーワードは連結された文のキーワードの和集合とする．発話区間の連結パターンの発話継続長は連結された発話区間の開始から終了までとし，キーワードは連結された各発話区間のキーワードの和集合とする．

3.3 適合度によるパターンの対応付け

適合度は，文及び発話区間の発話継続長や開始位置の差，キーワードの一致度，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に応じて課したペナルティの重み付き和とする．

連結パターンの対応付けは，以下の手順で行う．まず，文と発話区間の連結パターンのすべての組み合わせについて適合度を計算する．その中で，適合度が最大となる組み合わせを対応付ける．対応付けが決定した文及び発話区間を含む連結パターンは取り除く．同時に，対応付けが決定した文および発話区間より前の部分と後ろの部分にテキスト及び発話区間を分割する．分割した各部分に対してこれらの処理を再帰的に繰り返す．処理の範囲内に文または発話区間の連結パターンがなくなれば，その範囲の処理を終了する．

3.4 音声データにおける文の区切り位置の推定

3.3 節の処理では，対応付けられた連結パターンには複数の文が含まれている場合がある．そこで，連結パターンに含まれる各文が対応付けられた発話区間のどの部分に対応するかを推定する必要がある．

まず，連結パターン全体の発話継続長に対する各文の発話継続長の比を用いて対応付けられた発話区間を分割する．分割された発話区間に，最初の文から順番に文を割り当て，各文に対応する音声データの時刻を推定する．

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．

4. 実験

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，連結パターンを生成して対応付ける場合の，音声とテキストの文単位の対応付け精度を比較した．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許容範囲内のものは対応付けに成功したとみなし，対応付け精度を求めた．ずれの許容範囲は 0.5 秒と 1 秒の 2 種類を設定して評価した．

実験には，NHK の語学学習番組「3 カ月トピック英会話」の英語による会話シーンを用いた．1 回の放送につき会話シーンが 4 本含まれている．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．会話シーンは，15～28 文を含む 24～71 秒の映像であった．結果を表 1 に示す．

実験の結果，連結パターンを考慮することにより，文及び発話区間を単独に対応付ける場合に比べて平均精度が向上した．

表 1: 対応付け結果(%)

映像	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)	43.3	56.7	50.0	66.7
6/30(2)	44.1	73.5	64.7	88.2
6/30(3)	36.7	63.3	56.7	83.3
6/30(4)	6.7	6.7	10.0	13.3
7/7(1)	60.7	66.1	69.6	82.1
7/7(2)	75.0	83.3	80.6	91.7
7/7(3)	47.6	52.4	52.4	54.8
7/7(4)	44.2	55.8	53.8	78.8
7/14(1)	39.1	60.9	65.2	87.6
7/14(2)	56.5	67.4	76.1	89.1
7/14(3)	78.1	78.1	87.5	96.9
7/14(4)	63.0	47.8	80.4	52.2
平均	49.6	59.3	62.3	73.7

精度が大きく向上したデータでは，複数文を 1 つの発話区間として検出している部分があり，連結を考慮することによりこの部分の対応付けが正しく求められた．また，連結部分が正しく対応付けられたため，その前後の間違っていた対応付けも改善された．

連結パターンを考慮することにより平均精度は向上したが，実用レベルとしてはまだ不十分である．対応付けに失敗した原因の一つは，発話区間から得られた発話継続長と，文の文字数から計算した発話継続長の推定値の差が大きかったことである．この差が大きい場合，連結パターンの対応付けや，連結パターンから音声を 1 文単位に分割する処理で失敗することがあった．

5. おわりに

文及び発話区間の連結パターン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，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，言い淀みなどの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，全体精度を向上するため，文の発話継続長の推定を工夫する必要がある．また，提案手法による対応付けを用いてシステム化を行い，実際の授業での有効性を評価する必要がある．

謝辞

本研究は科研費 22500919 および 23300311 の助成を受けて実施した．

参考文献

[1]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位置提案”, 電子情報通信学会論文誌 D, Vol.J79-D-2, No.5 (1996).

[2]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”, 人口知能学会論文誌, Vol.17, No.4 (2002).

[3]Marcin Wichary and the Google Chrome team, “HTML5 Presentation”, < http:// slides.html5rocks.com/>.