類似音素行列を用いた音声認識結果とキャプション文字列との自動対応付けに関する検討

高橋 伸弥† 森元 逞† 西本 由之††
†福岡大学工学部 ††(株) ジェイ・フィット

1 はじめに

近年、聴覚障害者のための字幕付き番組放送の普及が重要課題とされ、官民を挙げた字幕放送の拡充が急ピッチで進んでいる。また，Web 上のストリーム映像に対しブラウザ上で付加情報を簡単に表示させることが出来るような仕組みも規格化されており (図 1)、今後はテレビ放送だけでなく、Web 上の映像番組にも字幕サービスが普及していくものと予想される.

図 1: 字幕対応付けの例

映像情報に付随した原稿もしくは書き起こし文が利用できるケースも少なくないことから，映像情報と予め用意された字幕 (キャプション) 文字列とを自動的かつ高精度に対応付けする方法として、音声認識結果の単語と字幕文字列中の単語を対応付ける方法 [1] や音声単位，文単位で対応づけする方法 [2] などが提案されている．しかし，字幕文字列は不要な語が取り除かれて簡潔な表現になっているため，冗長語や言い淀みなどを多く含む自然な発話に対しては対応付けが簡単でない場合も多い．また背景雑音を含んだり不明瞭な発音の発話であったりする場合には，発話区間の切り出しが難しいという問題や誤認識による影響が大きいという問題がある.

そこで本研究では，誤認識しやすい音素の組合せを予め求めておき，これを用いて音声認識結果の音素時系列と字幕文字列の発音時系列との対応づけを行う方法を提案する.

2 自動字幕対応付けアルゴリズム

2.1 基本的な考え方

基本的な考え方は，正解となる字幕文字列のローマ字読みと音声認識結果の音素系列との対応付けを DP マッチングを用いて行い，音声波形の時間軸上のどの点と字幕文字列の開始点とが対応するかを求めるというものである．具体的には，(1) 映像から音声波形を抽出する，(2) 抽出した音声波形を適当な長さに分割する，(3) 分割されたそれぞれの音声波形を音声認識する，(4) 音声認識結果の音素と対応する時刻とを取得する，(5) 全ての認識結果の音素系列を連結する，(6) 字幕文字列をローマ字読みに変換する，(7)DP マッチングを用いて字幕音素時系列と認識結果音素時系列間で対応付けを行う，という手順で処理を行う.

ここで，上記の処理 (2) における音声波形の分割は，純粋に音声認識プログラムの動作環境上の制約であり，発話区間を抽出しているわけではないことを断っておく[1]．また処理 (3) では，ユニグラムのみを用いて単語音声認識を行った[2].

2.2 DP マッチング

字幕音素時系列を $\mathbf{A} = \{a_i | (i = 0, \ldots, N)\}$，認識結果音素時系列を $\mathbf{B} = \{b_j (j = 0, \ldots, M)\}$ とする．このとき，$\mathbf{A}$ $\mathbf{B}$ 間の DP 距離は，以下の漸化式を計算することで求めることが出来る.

$$D(i,j) = \min \begin{cases} D(i,j-1) + p_{\text{ins}}, \\ D(i-1,j-1) + d(i,j), \\ D(i-1,j) + p_{\text{del}}. \end{cases}$$

ここで $d(i,j)$ は，音素 a_i と b_j 間の距離であり，正解が a_i であったとき認識結果が b_j である混同確率 $p(b_j|a_i)$ を用いて，$d(i,j) = 1 - p(b_j|a_i)$ で求めるものとする．混同確率の計算は，Julius 付属の日本語音素 HMM で定義された音素にショートポーズを追加した全 41 種の間で行った．この混同確率を行列形式で表現したものを，本稿では類似音素行列と呼ぶ．類似音素行列の学習用音声データとしては，国立国語研究所の話し言葉コーパスに収録されている講演音声データ計７４分を用いた.

Automatic Closed Caption Alignment Using Confusion Matrix

† Shin-ya TAKAHASHI (takahasi@tl.fukuoka-u.ac.jp)
† Tsuyoshi MORIMOTO (morimoto@tl.fukuoka-u.ac.jp)
Department of Electronics Engineering and Computer Science, Fukuoka University (†)
†† Yoshiyuki NISHIMOTO (nishimoto@jfit.co.jp)
J-FIT Co. LTD.(††)

1) 予め発話区間を切り出しておき，発話ごとに対応付けを行う手法もあるが，背景雑音等の影響で発話区間切り出しの精度が著しく低くなってしまうケースも多いため，本手法では発話区間切り出しを行わずに対応付けを行っている.

2) バイグラム／トライグラムを用いた場合，言語モデルの重みにより実際に発話された単語とかけ離れた単語が認識結果となる場合があるためである.

表 1: 実験に使用した映像

データ	放送日	放映時間	字幕文章数	字幕単語数	音声単語数	音声認識単語正解率
A	2007/6/8*	8 分 19 秒	49	1284	1429	26.7%
B	2007/6/16**	3 分	19	489	562	32.2%
C	2007/7/21**	3 分	22	536	585	21.0%
D	2007/8/3*	8 分 12 秒	49	1175	1281	28.0%

* コミュ! ふくおか　　** ギモン解決! ふくおか Q

表 2: 書き起こし文と字幕文章の例

書き起こし文	なるべく、えっと、デパートとか、あの、販売店の紙袋とかね、ビニール袋も
字幕文章	デパートとか販売店の紙袋やビニール袋は

また，挿入および削除の場合のペナルティ $p_{\mathrm{ins}}, p_{\mathrm{del}}$ は，母音の場合と子音の場合とで確率が異なることが予想されることから，それぞれ $p_{\mathrm{ins}}^{(\mathrm{v})}, p_{\mathrm{ins}}^{(\mathrm{c})}, p_{\mathrm{del}}^{(\mathrm{v})}, p_{\mathrm{del}}^{(\mathrm{c})}$ として与える．

3 実験

3.1 実験条件

音声認識エンジンとして Julius Ver.3.4 を用いた．汎用言語モデル及び音響モデルには，Julius ディクテーションキット Ver. 3.0 付属の ウェブテキストから学習した 6 万語のトライグラム言語モデルと性別非依存 PTM モデルを使用した．

実験には，福岡市公式ウェブサイト[3]に公開されている市公報テレビ番組を使用した (詳細は表 1)．ここで，字幕単語数とは字幕文章に含まれる単語延べ数であり，音声単語数とはもともとの音声に含まれている単語の延べ数である．また表には，音声認識時の単語正解率も併せて示している．この単語正解率は，書き起こしテキストを正解文として求めたものである．また入力音声は雑音や非音声部分 (音楽など) の除去等を一切行わず，トライグラム言語モデルを用いて認識を行った．

表を見ても分かる通り，音声認識の精度が非常に低くなっている．これは，対象とする音声が，雑音や重畳音などを多く含むだけでなく，話し言葉特有の言い淀みや言い直し，冗長語を多く含んでいるためである．表 2 に，音声の書き起こし文とそれに対応する字幕文章の例を示す．

3.2 対応付け実験結果

比較のために、類似音素行列を使用した場合と使用しない場合の実験を行った．類似音素行列を使用しない実験では，音素間の局所距離 $d(i,j)$ を音素が一致すれば 0，一致しなければ 1 とした．字幕文章の開始時刻と人手で与えた正解の開始時刻との平均誤差を表 3 に示す．表には，挿入及び削除のペナルティを固定した場合 ($p_{\mathrm{ins}}^{(\mathrm{v})} = p_{\mathrm{ins}}^{(\mathrm{c})} = p_{\mathrm{del}}^{(\mathrm{v})} = p_{\mathrm{del}}^{(\mathrm{c})} = 1.0$) と変化させた場合の結果も併せて示している．ここで各ペナルティの値は，0.25 から 1.0 まで 0.25 刻みで変化させた場合に，4 つのデータに対する誤差の平均が最良の結果となったものを用いた．

表 3: 実験結果

データ	類似音素行列無し		類似音素行列有り	
	ペナルティ		ペナルティ	
	固定	変更 *	固定	変更 **
A	0.556	0.502	0.493	0.475
B	0.305	0.318	0.273	0.259
C	0.838	0.422	0.827	0.532
D	1.035	1.000	1.399	1.400
平均	0.684	0.561	0.748	0.546

* $p_{\mathrm{ins}}^{(\mathrm{v})} = p_{\mathrm{ins}}^{(\mathrm{c})} = 0.75, p_{\mathrm{del}}^{(\mathrm{v})} = 1.0, p_{\mathrm{del}}^{(\mathrm{c})} = 0.5$

** $p_{\mathrm{ins}}^{(\mathrm{v})} = p_{\mathrm{ins}}^{(\mathrm{c})} = p_{\mathrm{del}}^{(\mathrm{v})} = 0.75, p_{\mathrm{del}}^{(\mathrm{c})} = 0.5$

表から分かるようにデータ A,B,C に対しては，類似音素行列を用いた手法が比較的良い結果となっている．データ D に関しては，認識音素数と字幕音素数との比が 1 割以上他よりも大きく，字幕と対応がつかない不要な音素が多く含まれていたことが理由で，類似音素行列の効果が十分得られなかったものと思われる．

4 おわりに

類似音素行列を用いて音声認識結果音素時系列と字幕文字列発音時系列との対応づけを行う方法を提案した．評価実験により，字幕が音声内容を十分に反映しているケースにおいては提案手法が有効であることを示した．本稿では，字幕を表示するタイミングのみを評価対象とし，字幕表示を消すタイミングについては評価していない．字幕を表示すべきでない区間を検出するには，必要な音声と不要な音とを切り分ける必要があり，音素認識結果だけを用いたアプローチでは実現が難しいと思われる．解決法としては，後処理として字幕文字列の長さ (モーラ数) などで消すタイミングを計算する方法などが考えられる．

謝辞

映像データを提供下さった福岡市広報課に感謝します．本研究の一部は，文部科学省科学研究費補助金 (若手研究 (B)，課題番号 19700184) による．

参考文献

[1] C-W Huang *et. al*, "Automatic Closed Caption Alignment Based on Speech Recognition Transcripts," Technical Report, Columbia ADVENT,2003, 12.

[2] 西沢，杉山，"音声特徴と言語情報を用いた音声とテキストの自動対応付け，" 音学講論，1-P-3，pp.167-168 (2004)

[3] 福岡市広報テレビ番組: http://www.pr-fukuoka-city.stream.jfit.co.jp/index.php