5P-06

講演音声認識結果の誤り箇所の復唱入力を用いたノートテイクシステム

大田 健翔 † 秋田 祐哉 † 河原 達也 †

† 京都大学大学院情報学研究科

1. はじめに

講演などの場における聴覚障がい者への情報保障として，ノートテイクやパソコン要約筆記などが行われている．これらは講師の発話をリアルタイムに書き起こし，視覚情報として聴覚障がい者に提示するもので，主にノートに手書きする方法とパソコンにキー入力する方法がある．

しかし，これらの方法ではノートテイカーに訓練が必要となったり，書き起こせる量に限度があるという問題がある．その上，講師の発話内容が専門的になるほど，テイカーにも専門的な知識が必要となる．

これらの問題を解決するために，我々は音声認識を用いたノートテイクシステムの研究を進めている [1][2]．これは講師の発話を音声認識して自動的に書き起こし，字幕として提示するものである．しかし，講師の発話は音声認識を意識していない自然なものであるために，認識精度が低下する．したがって，認識誤りの修正作業がかなり必要となり，リアルタイム性を損なうこととなっている．

本研究では，速やかな修正のために，復唱（リスピーク）を用いた認識誤り修正手法を提案する．音声認識誤りの部分だけ復唱し，自動的に上書きすることで，作業者に負担の小さいシステムを目指す．

2. 関連研究

自動認識の難しい音声に対して字幕を作成する手法として，**復唱 (リスピーク) 方式**がある．これは発話者の音声を直接認識するのではなく，作業者が明瞭に復唱することで，元の話者の発話内容を正しく認識させる手法である．これにより発話者の話し方の違いを吸収できると共に，キーボードよりも早く入力できる．NHK では幅広いジャンルの番組で使用している [3]．しかし，音声を聞きながら，長時間明瞭に話す訓練が必要となり，NHK でもプロのアナウンサーが行っている．

一方，熟練者でなくとも可能な修正手法として**複数仮説の統合**が提案されている [4]．これは，音声認識結果に加えて，不特定多数の視聴者により投稿された発話の書き起こしの断片を統合して字幕を作成する方法である．ここでの入力はキーボードで行われる．視聴者は自分のペースで投稿すればいいので，作業に特別な技術は要しない．しかし，統合してひとつの完全な字幕を得るには多くの作業者が必要であり，ノートテイクの現場では実現が難しい．

3. 提案システム

本研究では，復唱方式と仮説統合を組み合わせた手法を提案する．具体的には，講師の音声認識結果をベースの字幕として用い，認識が誤っている部分のみ復唱して，その認識結果と統合することにより，より正確な字幕を出力する．作業者は講師の音声認識誤りのみを復唱すればいいので少人数で済み，熟練者でなくとも運用が可能であると考えられる．

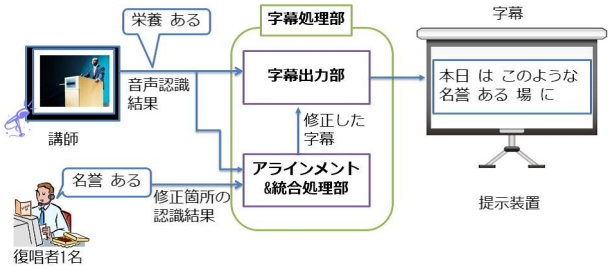

図 1: 提案システムの全体図

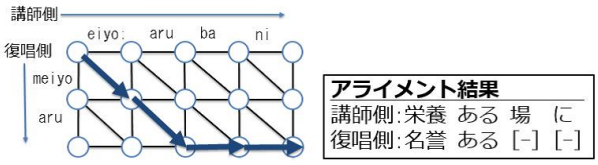

図 2: アライメントの例 ('[-]' はギャップを表す)

3.1 システムにおける処理の流れ

システムの全体像を図 1 に示す．まず，講師のマイクから音声を入力して認識する．その結果を字幕処理部内の字幕出力部とアライメント&統合処理部に送る．講師の音声は認識結果とともに復唱者側でモニターされ，音声認識誤りが発生した場合，復唱者はその部分を復唱する．復唱の音声認識結果は講師の音声認識結果履歴とアライメントされる．これにより該当部分が復唱結果で修正され，字幕出力部に送られ，字幕へ反映される．復唱者の人数は 1 名である．

3.2 仮説のアライメント処理

復唱者と講師の音声認識結果は単語 (形態素) 単位で動的計画法によりアライメントされる．例を図 2 に示す．

動的計画法における各パスのコストは，講師側と復唱側それぞれの単語における音素列の編集距離である．すなわち，二つの単語間の音素列がマッチしているほどコストは小さくなり対応しやすくなる．また，対応する単語がない場合はギャップと呼ばれる空文字に対応させる．ギャップ対応時のコストは定数値のペナルティである．すなわち，ある単語と単語が対応するかどうかは，編集距離がギャップのコストより小さくなるかによる．

3.3 統合処理

アライメントが取られた単語列の対を次のヒューリスティックの規則で統合することにより誤りを修正する．

1. 原則として，復唱側の単語を用いる．ギャップの場合は講師側の単語を用いる (図 3①)

A note-taking system using respeaking errors in automatic speech recognition of lectures. *Kensho OTA, Yuya AKITA, and Tatsuya KAWAHARA (Graduate School of Informatics, Kyoto University.)*

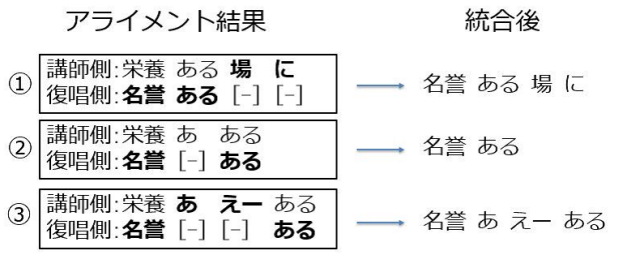

図 3: 統合の例．太字の単語が字幕に採用される．

2. ただし，復唱側の，単語と単語に挟まれた 1 つのギャップは無視する (図 3②)
3. 復唱側でギャップが 2 つ以上続く場合は，ギャップに対応する講師側の単語を用いる (図 3③)

4. 予備実験

シミュレーションによる予備的な実験を行い，復唱による修正の効果を検証した．実験方法は次の通りである．

1. あらかじめ収録した講演に対して音声認識を行い，単語単位で時間同期した字幕を作成する．
2. 字幕付き講演音声を視聴しながら誤りの部分を復唱する．この時に復唱認識結果のログを取っておく．
3. 復唱の音声認識結果で字幕を修正する．
4. 修正前と修正後における字幕の文字正解率，正解精度を測る．

使用した講演音声は京都大学 iPS 細胞研究所 (CiRA) の 2011 年シンポジウムにおけるもので．講演の長さは 44 分 9 秒であった．

講演の音声認識には，我々が開発・公開しているオンライン字幕作成システム [5] を用いた．このシステムでは音声認識エンジン Julius[‡] を用いて字幕を作成しており，音響モデルは日本語話し言葉コーパス (CSJ) の講演音声 257 時間で学習した DNN-HMM，言語モデルには CSJ 学会・模擬講演書き起こし (合計 7.7M 単語) と CiRA の web サイトから収集したテキスト (53K 単語) から学習した単語 3-gram モデルを用いた．

復唱側でも音声認識に Julius を用いた．音響モデルは，復唱が読み上げ調であることを考慮して，日本音響学会新聞記事読み上げ音声コーパス (ASJ-JNAS)86 時間で学習した GMM-HMM，言語モデルは講演音声用の言語モデルにさらに家庭医学書「メルクマニュアル」(1.4M 単語) を加えて学習したものを使用した．

アライメントは本実験では手作業で行った．復唱認識結果の末尾 3 文字について，講演音声認識結果とのマッチングを行い，マッチした箇所から前へさかのぼり，妥当であると考えられる部分まで復唱認識結果で置換した．

5. 結果と考察

今回の講演音声における，講師の発話数は 954 となった．ここでの発話とは，Julius の音声区間検出による自動切り出しが行われた単位である．講師音声の認識率（ベースライン）は，文字正解率で 85.88%，文字正解精度で 83.30%であった．誤り箇所の総数は 1117 で，その誤り文字総数は 2427 であった．

[‡]http://julius.sourceforge.jp/

復唱の発話数は 379 で，合計時間は 14 分 30 秒であった．復唱自体の認識率は文字正解率で 89.21%，文字正解精度で 85.12%であった．NHK の復唱者で単語正解精度は 84.4%と報告されている [3] ことを考えると妥当な精度といえる．

アライメントは精度（Precision），再現率（Recall）で評価した．これらは次式で定義される．

$$Precision = \frac{\text{誤りを修正できた箇所の数}}{\text{アライメントされた箇所の数}}$$

$$Recall = \frac{\text{誤りを修正できた箇所の数}}{\text{講師音声認識結果の誤り箇所の数}}$$

アライメントにより修正できた箇所は 279 であった．ただし，ここでは 1 文字以上修正できたものをカウントしている．逆に，アライメントにより誤りが発生した箇所は 10 であった．したがって，アライメントの Precision は 0.965 で，Recall は 0.250 であり，Recall が小さいことがわかる．すなわち、復唱者側の見過ごしが改善のネックなっているといえる．これに関しては，字幕提示を 1 発話分のみとしていたため，どう修正すべきか迷っている間に次の発話へ切り替わってしまい，修正できなかったものと考えている．

アライメントにより，修正後の認識率は文字正解率で 89.57%で 3.69 ポイントの改善，文字正解精度で 87.02%で 3.72 ポイントの改善となった．また，アライメントできた 238 発話に限定すると，講師側の誤り文字総数は 798 個であったのに対し，復唱で修正できた文字総数は 642 個であり，復唱でかえって生じた誤り文字の数は 73 個であった．これにより，復唱文字修正率は 73.7%となった．

6. まとめと今後の課題

本稿では，講師の音声認識の誤り部分を復唱入力で修正することで，非熟練者でも運用できるノートテイクシステムを提案した．

上記のさらなる改善のためには Recall の改善が求められる．今後は，ユーザビリティを考慮したインタフェースの開発や，修正すべき箇所とそうでない箇所を自動で判別する機能を追加していくことを考えている．

参考文献

[1] 勝丸徳浩, 河原達也, 秋田祐哉, 森信介, 山田篤. 講義音声認識に基づくノートテイクシステム. 電子情報通信学会技術研究報告,WIT-109-260, pp. 25–30, 2009.

[2] 桑原暢弘, 秋田祐哉, 河原達也. 音声認識結果の有用性の自動判定に基づく講義のリアルタイム字幕付与システム. 日本音響学会研究発表会講演論文集, 2-4-5 春季, pp. 57–60, 2014.

[3] 本間真一. 生字幕制作のための音声認識. NHK 技研 R&D, No. 122, pp. 25–31, 2010.

[4] 浮田俊輔, 緒方淳, 後藤真孝, 小林哲則. ライブストリーミングのための協調的音声書き起こしシステム. 情報処理学会研究報告, 2011-SLP-85, 2011.

[5] 秋田祐哉, 河原達也. 音声認識を用いたオンライン自動字幕作成・編集システム. 日本音響学会研究発表会講演論文集, 2-8-4 秋季, pp. 65–66, 2013.