Vol.2014-SLP-102 No.3
2014/7/24

講演スライドの文字認識結果を用いた音声認識の改善

童 弋正[1]　秋田 祐哉[1,2]　河原 達也[1,2]

概要：講演の音声認識において言語モデル適応のためのテキストとしてスライドの利用が考えられるが，スライドの電子ファイルを必ず入手できるとは限らない．そこで本研究では，講演映像中のスライドの文字認識結果を利用して言語モデルの適応を行い，音声認識の精度の改善を目指す．文字認識結果には多くの誤りが含まれることから，本研究では形態的・話題的な面からこれらをフィルタリングする手法を提案する．これにより得られたスライド文字認識結果をもとに、関連する新聞記事を用いた適応、またスライドを用いたキャッシュモデルに基づく適応を行う．実際の講演データを用いた評価で、音声認識精度の改善を得ることができた．

1. はじめに

現在多くの大学では、講演や講義の映像をアーカイブして、広く一般に公開するオープンコースウェア（OCW）[1][2] や MOOC[3] の取り組みが進められている．増大する講演・講義映像データを効率的に検索するためには、映像データをテキストに書き起こしてインデクシングすることが有用である．また、書き起こしを用いて映像データに字幕を付与することは、聴覚障害者の視聴支援 [4][5] や、講演・講義の専門的内容に対する理解 [6] にも役立つ．しかし、すべての映像を人手で書き起こして字幕を付与するのはコストが大きく現実的ではない．

これに対して、音声認識を利用して講演・講義を自動的に書き起こすことが考えられる [7][8][9]．しかし、講演・講義では話題により異なる専門用語が多数出現する．これらの専門用語は一般的な大語彙音声認識システムではカバーすることが難しく、認識精度の低下につながる．このため、講演・講義の音声認識では、内容に関連する情報を利用して言語モデルを適応する必要がある．

講演・講義に対する言語モデルの適応では、教科書 [10] や論文・予稿 [11]、Web から収集したテキスト [12]、スライド [13][14][15] などが用いられている．本研究では、このうちスライドに着目する．多くの講演・講義ではスライドを使用しており、講演・講義の発話内容にスライドのテキスト情報が反映していると考えられるからである．スライドを用いた適応手法として、山崎ら [13] はスライドテキスト全体から作った言語モデルと発話時間に対応するスライド一枚から作った言語モデルをベースライン言語モデルと線形補間して適応する手法を提案した．河原ら [14] は確率的潜在意味解析（PLSA）、関連 Web テキストの収集、キャッシュモデルの 3 つの手法による大局的および局所的な適応を提案した．また、Miranda ら [15] は音声認識結果とスライドテキストからラティスを構築して認識精度を改善する手法を提案した．

ただし、これまでの手法はいずれもスライドテキストを入手できることを前提としている．これに対して、たとえば OCW のアーカイブを書き起こす場合のように、講演の映像のみが利用可能で、スライドは必ずしも入手できない場合もある．そこで本研究では、講演の映像に含まれるスライド画像に対して文字認識（OCR）を行い、この結果をもとに言語モデルの適応を行うことを考える．文字認識結果には誤りが含まれるため、誤りの影響を排除し、また断片的な文字認識結果から適応を行うことが求められる．本研究では、このためのフィルタリング手法と適応の枠組みについて検討する．

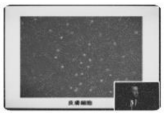
2. スライド文字認識とその有用性

2.1 講演映像のための文字認識システム

本研究では、講演の映像として、スライドのみを映し続けるものではなく、講師やスライドの画面が状況に応じて切り替わるものを想定する．

通常の文字認識の場合、文字が映っている画像から文字部分を切り出した上で、固有の文字パターンとマッチングすることで認識を行う [16]．しかし本研究では、まず映像中でスライドが映っている区間を検出する必要がある．ま

[1] 京都大学　情報学研究科
Graduate School of Informatics, Kyoto University
[2] 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


スライド画像

slide:[スライド番号]
time:[開始時刻]
duration:[経過時間]
num terms:[字数]
text lines:[文字認識内容]

図 1 TalkMiner の出力形式

た、講師の発話に対応するスライドが同定できれば、より細かな適応が可能となるから、各スライドが表示された時刻も得られることが望ましい．

このための文字認識システムとして、富士ゼロックス社の TalkMiner[*1] を用いる．このシステムでは、まず一定時間変化のない映像のセグメントをスライドとして検出し、映像内からスライド画像を抽出する．次に抽出されたスライド画像について文字認識処理を行い、テキストを抽出する．その後、テキストとスライド画像を自動で紐付けてインデックス化し、データベースに保存する．認識結果の出力形式は**図 1** の通りである．

出力には、スライドテキストに加えてスライドの開始時刻と経過時間が含まれ、スライドと音声区間の時間的対応をとることができる．なお、TalkMiner ではシステムのパラメータを利用者が調整することができない．また、出力結果には信頼度は付与されていない．

2.2 文字認識結果の有用性の検証

講演映像から文字認識を行う際には、図や写真を文字と間違えて認識する、映像がはっきりしていないため認識に失敗するなどのために、認識精度が低下する．このような文字認識結果の有用性について、実際の講演映像をもとに検証を行った．本研究では、京都大学 OCW で配信されている iPS 細胞研究所のシンポジウム（2010 年）の 3 つの講演データ（**表 1**）を対象として、TalkMiner により映像の中のスライド文字情報を抽出した．3 講演のスライドの文字認識精度は正解率（recall）が 67.14%であるものの、湧出し誤り（false alarm）が 42.99%である（**表 2**）．また、これらの講演に対して 5.2 節で述べるベースラインシステムで音声認識を行った．この精度は 79.32%である．

音声認識の言語モデルに含まれていない未知語や、音声認識を誤った部分がスライドの中に含まれている場合、スライドテキストを用いて音声認識を改善できる可能性がある．そこで、この検証では、未知語のうち OCR テキストでカバーできる割合と、音声認識誤りとなったキーワードのカバレージを調べた．

講演に出現する未知語の割合と、このうちスライドでカバーできる未知語の割合を**表 3** に示す．ここでは人手による講演の書き起こしを用いて集計を行った．言語モデルは「日本語話し言葉コーパス」（CSJ）から学習したため、「iPS」などの専門用語はカバーされずに未知語となった．3 講演の平均の未知語率は 3.6%である．このうち 48%の未知語が正しいスライドテキストでカバーできる．文字認識精度が低いスライド OCR テキストでも、38%の未知語をカバーできる．これより、スライド OCR テキストを適応に用いることで未知語率を削減することができる．

一方、キーワードは講演特有の単語として定義した．ここでは、キーワードらしさの指標に、式 (1)〜(3) で定められる tf-idf スコアを用いる．

$$tf(t,d) = \frac{n_{t,d}}{\sum_{s\in d} n_{s,d}} \quad (1)$$

$$idf(t) = \log \frac{N}{df(t)} + 1 \quad (2)$$

$$tfidf(t,d) = tf(t,d) \cdot idf(t) \quad (3)$$

$n_{t,d}$ はある単語 t の文書 d 内での出現回数で、$\sum_{s\in d} n_{s,d}$ は文書 d 内のすべての単語の出現回数の和である．N は全文書数で、$df(t)$ はある単語 t が出現する文書の数である．

idf の計算に用いる文書集合には、毎日新聞 2011 年分の 81768 記事を用いた．講演ごとに書き起こしテキストの各単語の tf-idf スコアを計算して、助詞・言い淀み・頻度が大きい名詞（「もの」、「こと」など）を除去した上で、スコアが大きな順に全単語数の 10 分の 1 まで残してキーワードとした．これらのキーワードのカバレージは**表 4** の通りで

表 1 利用する講演データの仕様

講演時間の合計	78 分 50 秒
書き起こしの総単語数	17046
キーワード数	1720
音声認識精度	79.32%

表 2 講演スライドの文字認識結果

置換誤り（substitution error）	28.73%
削除誤り（deletion error）	4.13%
正解率（recall）	67.14%
湧出し誤り（false alarm）	42.99%

表 3 未知語のカバレージ（3 つの講演の合計）

未知語率	3.57%
上記のうちスライド OCR テキストに含まれる割合	38.16%
(cf.) 上記のうち正しいスライドテキストに含まれる割合	47.86%

表 4 キーワードのカバレージ（3 つの講演の合計）

音声認識誤りとなったキーワードの割合	39.30%
上記のうちスライド OCR テキストに含まれる割合	88.60%
(cf.) 上記のうち正しいスライドテキストに含まれる割合	93.34%

[*1] http://www.fujixerox.co.jp/company/technical/main_technology/capturing/talkminer.html

ある．キーワードの 39%が誤認識されたが、これに対して正しいスライドテキストとスライド OCR テキストのカバレージはそれぞれ 93%と 89%であり、高い割合となった．文字認識結果を用いて適応することでキーワードの音声認識精度の改善も期待できるといえる．

3. 文字認識結果と関連新聞記事による言語モデル適応

本節では、文字認識結果を用いて言語モデルを適応する手法を提案する．本手法は文字認識誤りのフィルタリングとその結果を用いた適応の 2 つに分けることができる．以下ではこれらについて詳しく述べる．

3.1 文字認識結果のフィルタリング

文字認識結果には誤認識が多数含まれているため、これらをフィルタリングする必要がある．本研究では 2 段階のフィルタリングを行う．まず、図や写真を誤って文字として認識した場合には、記号や一般には使われていない文字が多く出力されるため、形態素解析器 Kytea[*2] により検出、除去を行う．次に、新聞記事データベースを用いて、関連する新聞記事に出現する単語のみを残すようフィルタリングする．

Kytea によるフィルタリングでは、文字認識結果を形態素解析して、認識結果を単語単位まで分割する．この際、記号には「NA」というラベルが付与される．また、一般的でない文字の列は未知語として扱われ、「UNK」というラベルが付与される．「NA」および「UNK」ラベルが付与された単語を除去する．

新聞記事によるフィルタリングでは、毎日新聞（2011 年）の 81768 記事の中で、スライドテキストとの類似度が大きな新聞記事を利用する．類似度はベクトル空間モデルに基づき計算する．2.2 節と同様に、これらの記事で文字認識結果の各単語の tf-idf を計算して、以下の 3 つの条件を満足する単語をスライドキーワードとする．

（1）tf-idf の値が最大の tf-idf の 1/10 より大きい．

（2）$df <$（1/10× 新聞記事数）．$df >$（1/10× 新聞記事数）になると単語の特殊性があまり見られないため．

（3）$df > 1$.

（4）単文字のひらがな、カタカナ、アルファベットではない．

スライドテキストに対して、キーワードの tf-idf の値を要素とするキーワードベクトル K_S を作成する．新聞記事 $n_1, n_2, \cdots, n_{81768}$ に対しても、キーワードの tf-idf を計算してキーワードベクトル $K_{n_1}, K_{n_2}, \cdots, K_{n_{81768}}$ を作成する．最後に式 (4) でベクトル間のコサイン距離に基づく類似度を記事ごとに計算し、類似度の上位 5000 記事を選択して関連記事とする．

$$similarity = \frac{K_S \cdot K_n}{|K_S||K_n|} \tag{4}$$

これにより抽出された新聞記事は講演との関連度が大きい記事であり、これらの記事の中に出現しない単語はスライドテキストや講演で出現する可能性が低いと想定する．これに基づいて、スライドの OCR 結果のうち、関連新聞記事に出現した単語のみを残す．

[*2] http://www.phontron.com/kytea/index-ja.html

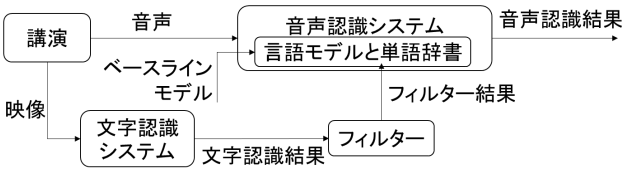


図 2 文字認識結果による適応

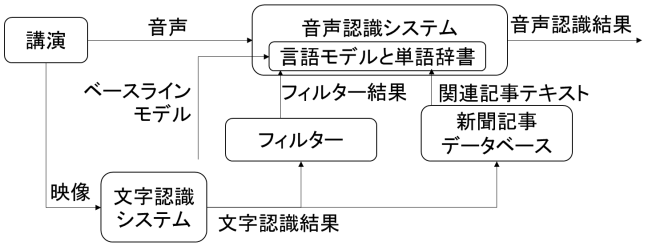


図 3 文字認識結果と新聞記事による適応

3.2 言語モデルの適応

フィルタリングしたスライド OCR テキストを図 **2** のようにベースライン言語モデルと線形補間 [13][17] して適応を行う．線形補間の重みは、後述する開発セットにおいてパープレキシティが最小になるように設定する．

フィルタリングした後のスライドテキストの量は小さくなるため、スライドテキストとの類似度が大きな新聞記事を適応に用いる手法も検討する．適応に使う新聞記事は 3.1 節のフィルタリングにおける類似度の上位 200 記事である．選択した新聞記事は図 **3** に示すように、スライド OCR テキストとともにベースライン言語モデルと線形補間する．

4. キャッシュモデルによる適応

ある発話の近傍で表示されているスライドテキストの中に出現する単語は、その発話で出現する可能性があると考えられる．このような単語に高いスコアを与えるために、キャッシュモデル [18] を導入する．

本来キャッシュモデルは、直前に発話された単語をキャッシュして、これに含まれる単語が再び使用される可能性があるとして言語スコアを大きくするものである．本研究では、単語の発話履歴をスライドテキストの単語に置き換えて [14]、発話に対応するスライドとその前後のスライドの単語をキャッシュに入れる．なお、キャッシュモデルは内

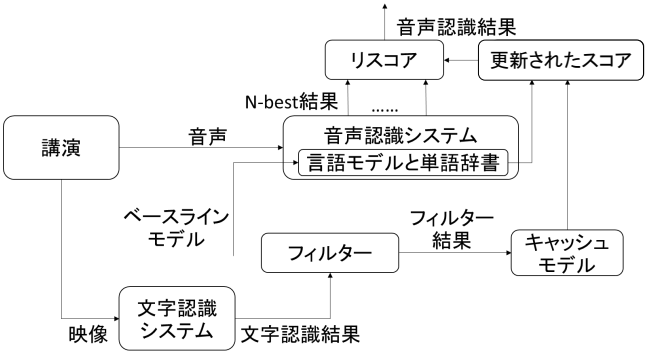


図 4　キャッシュモデルによるリスコアリング

容語に限定することとし、フィルタリングした後の OCR テキストから以下の単語をさらに除去する．

（1）助詞、助動詞などの機能語

（2）頻度が大きい名詞（「こと」、「もの」など）

（3）単文字のひらがな、カタカナ、アルファベット

本研究では、文献 [14] と異なり、認識尤度に対するリワードの形でキャッシュモデルを適用する．まず、キャッシュに基づく各単語のリワードスコアを式 (5) で定める．

$$C(w|p) = \sum_{i=p-1}^{p+1} \sum_{w_s \in S_i} \delta(w, w_s) \qquad (5)$$

$C(w|p)$ は現在の発話に対応するスライド p およびその前後のスライドにおける単語 w のリワードスコアである．$\delta(w, w_s)$ はクロネッカーのデルタ関数であり、式 (6) で定義される．

$$\delta(w, w_s) = \begin{cases} 1, & w = w_s \\ 0, & w \neq w_s \end{cases} \qquad (6)$$

このリワードスコアを用いて、式 (7) に示すように認識結果の尤度を再計算する．

$$score = score_{AM} + \alpha \cdot score_{LM} + \beta \cdot n + \gamma \sum_{i=1}^{n} C(w_i|p) \qquad (7)$$

$score$ は認識結果の尤度で、$score_{AM}$ と $score_{LM}$ は音響尤度と言語尤度である． n は認識結果の単語数である．α, β, γ はそれぞれ言語モデルの重み、単語挿入ペナルティ、リワードスコアの重みである．

本研究では、キャッシュモデルを用いた適応を N-best 仮説のリスコアリングにより適用する．すなわち、**図 4** のように、ベースラインモデルによる N 個の仮説のスコアを式 (7) で再計算して尤度を更新する．そして、更新された尤度をもとにスコアが最も高い候補を出力する．

5. 評価実験

提案する文字認識結果のフィルタリング手法と言語モデル適応手法について評価を行った．評価セットは、2.2 節でも用いた京都大学 OCW の 2010 年の 3 講演（表 1）である．

表 5　フィルタリング前後の単語の再現率と適合率

	再現率	適合率	F 値
フィルタリング前	62.59%	33.94%	44.01
フィルタリング後	60.72%	47.19%	53.11

表 6　フィルタリング前後の未知語とキーワードの再現率

	未知語	キーワード
フィルタリング前	79.73%	94.93%
フィルタリング後	74.57%	93.98%

5.1 フィルタリングの有効性

3 講演のスライドの文字認識結果の、フィルタリング前後におけるすべての単語の再現率と適合率を求めた（**表 5**）．また、未知語とキーワードについて、フィルタリング前後の再現率を計算した（**表 6**）．

表 5 から、フィルタリングにより再現率は 1.9 ポイント低下したものの、適合率が 13.3 ポイント改善した．未知語とキーワードに限った場合（表 6）でも、再現率の低下はそれぞれ 5.2 ポイント、1.0 ポイントである．これらより、提案するフィルタリング手法が正しい文字認識結果を損なうことなく誤りを除去できているといえる．

5.2 文字認識結果と新聞記事を用いた適応

フィルタリング後のテキストを図 2 に示したようにベースライン言語モデルと線形混合する．また、図 3 のように関連新聞記事もあわせて用いた適応を行う．参考のために、文字認識結果の代わりに正しいスライドテキストを用いて適応した場合の評価も行う．

音声認識は Julius 4.1.5 デコーダを用いる．ベースライン言語モデルは「日本語話し言葉コーパス」（CSJ）のすべての学会講演と模擬講演を用いて学習したモデルである．学習データの総単語数は 7.66M 単語で、モデルの語彙サイズは 37K の単語 3-gram モデルとなっている．なお、モデルの線形補間に際しては、開発セットとして別の年（2009 年）の iPS 細胞研究所のシンポジウムの 3 講演を使用し、これらの書き起こしにおけるパープレキシティから重みを定めた．

適応なし、スライドテキストだけで適応した言語モデル、スライドテキストと関連新聞記事で適応した言語モデルについて、講演全単語とキーワードの音声認識精度を**表 7** に示す．正しいスライドテキストを用いて適応した場合もあわせて示す．ここで用いるキーワードは 2.2 節で定めたものである．

表 7 より、スライドテキストだけで適応した言語モデルと、関連新聞記事も用いて適応した言語モデルのいずれによっても認識精度が改善された．関連新聞記事も用いて適応した場合、スライドテキストだけで適応した場合より

表 7　フィルタリング前後の未知語とキーワードの再現率

	適応なし（ベースライン）	スライドテキストのみによる適応		スライドテキストと関連新聞記事による適応	
全単語（文字単位の精度）	79.32%	スライド OCR テキスト	83.75% (+4.43%)	スライド OCR テキスト	85.07% (+5.75%)
		(cf.) 正しいスライドテキスト	84.77% (+5.45%)	(cf.) 正しいスライドテキスト	85.56% (+6.24%)
キーワード（単語単位の精度）	60.16%	スライド OCR テキスト	86.88% (+26.72%)	スライド OCR テキスト	90.22% (+30.06%)
		(cf.) 正しいスライドテキスト	90.26% (+30.10%)	(cf.) 正しいスライドテキスト	91.03% (+30.87%)

も、認識精度の改善が大きくなっている．特にキーワードにおいて効果が大きく、認識精度は 30 ポイント改善されて 90%に達した．スライドテキストを用いた適応により、「iPS」のような未知語や専門用語が言語モデルでカバーされて認識できるようになったことが大きな要因である．

適応の効果が言語モデルと単語辞書のいずれにみられたのかを確かめるため、どちらか一方のみの適応も試みた．文字認識結果のスライドテキストで言語モデルのみを適応し、単語辞書は新たな単語を登録しない場合、認識精度は 82.60%（+3.28 ポイント）であった．一方、単語辞書への登録のみ行い、言語モデルは適応しない場合、精度は 80.27%（+0.95 ポイント）であった．両方とも適応した場合（表 7、83.75%）の改善はそれぞれの改善の和とおおむね一致しており、双方が加算的に機能しているといえる．

文字認識結果で適応した場合では、全単語の認識精度の改善幅は正しいテキストを用いた場合の改善幅の 81.3%〜92.1%である．文字認識結果のフィルタリング後の F 値が 53.1 であることを考えると、適応が十分に機能しているといえる．キーワードの認識精度でも同様の傾向であった．

5.3　キャッシュモデルによる適応

ここでは、適応なし、スライドテキストのみによる適応、スライドテキストと新聞記事で適応した言語モデルのそれぞれにキャッシュモデルを導入して評価した．音声認識システムに最大 100 個の文仮説を出力させ、これらをリスコアする．比較のため、正しいスライドテキストを用いたキャッシュモデルによる適応も行う．

図 5 と**図 6** はリスコアによる全単語およびキーワード認識精度である．リワードスコアの重みは 0 から 1 まで 0.1 ごとに変化させ、認識精度を計算した．

全単語の音声認識精度（図 5）については、適応なしの言語モデルでは、正しいスライドテキスト・OCR テキストのいずれの場合も、キャッシュモデルにより認識精度がそれぞれ最大 0.78 ポイント、0.51 ポイント改善された．一方、スライドテキストだけで適応した言語モデルでは、正しいスライドテキスト・OCR テキストの場合の改善がそれぞれ最大 0.15 ポイント、0.10 ポイントであった．スライドテキストと新聞記事で適応した言語モデルでも同様に最大 0.09 ポイント、0.09 ポイントとなった．

図 6 のキーワード認識精度でも同様の傾向が見られた．スライドテキストのみ、またはスライドテキストと新聞記事で適応した言語モデルでの改善幅（OCR キャッシュの場合はそれぞれ 0.25 ポイント、0.50 ポイント）は、適応なしの言語モデルでの改善幅（OCR キャッシュの場合は 3.80 ポイント）より小さい．スライドテキストや関連新聞記事で適応した言語モデルでは、スライドの中のキーワードなどの重要単語についてすでに強化されているので、ほとんど改善が得られなかった．

6.　まとめ

本研究では、スライドの文字認識結果をもとに言語モデルを適応する手法について検討した．まず、精度の低い講演スライドの文字認識結果でも、多くの未知語とキーワードをカバーできることを示した．そして、形態素解析器と新聞記事データベースを利用して、文字認識誤りをフィルタリングする手法を提案した．フィルタリングしたスライドテキストおよび類似度の大きな新聞記事を用いた言語モデル適応により、音声認識精度の改善を得ることができた．

参考文献

[1] 土佐尚子，美濃導彦：京都大学 OCW：オープンエデュケーションがもたらすもの，工学・工業教育研究講演会講演論文集，pp. 204–205 (2010).
[2] 竹村治雄：大阪大学 OCW の現状と課題，工学教育研究講演会講演論文集，pp. 720–721 (2012).
[3] Daradoumis, T., Bassi, R., Xhafa, F. and Caballe, S.: A Review on Massive E-Learning (MOOC) Design, Delivery and Assessment, *in Proc. 3PGCIC* (2013).
[4] 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，日本音響学会春季研究発表会講演論文集，2-4-5 (2014).
[5] Cerva, P., Silovsky, J., Zdansky, J., Nouza, J. and Malek, J.: Real-Time Lecture Transcription using ASR for Czech Hearing Impaired or Deaf Students, *in Proc. Interspeech* (2012).
[6] Ferdiansyah, V.，Nakagawa, S.：Effect of Captioning Lecture Videos For Learning in Foreign Language, 情報処理学会研究報告，2013-SLP-97-13 (2013).
[7] Togashi, S. and Nakagawa, S.: A Browsing System for

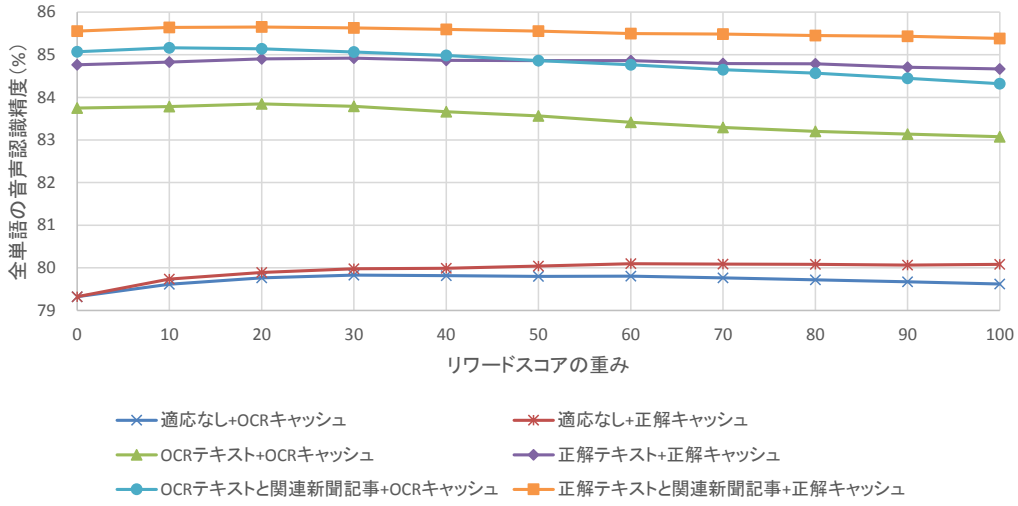


図 5 キャッシュモデルで適応した場合の全単語の音声認識精度（文字単位）

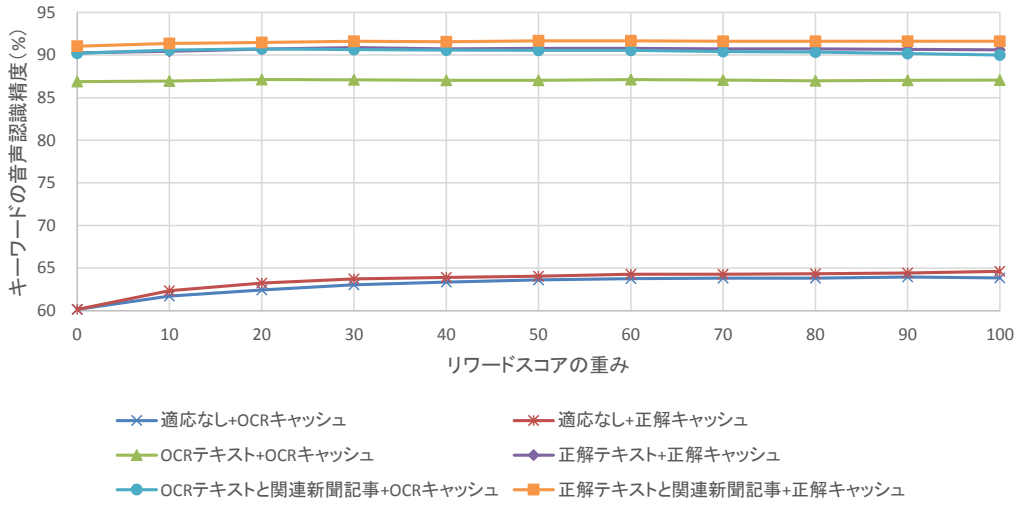


図 6 キャッシュモデルで適応した場合のキーワードの音声認識精度（単語単位）

Classroom Lecture Speech, *in Proc. Interspeech* (2008).
[8] Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. and Duerstock, B.: Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom, *IEEE Trans. Learning Technologies*, Vol. 6, No. 4, pp. 299–311 (2013).
[9] Glass, J., Hazen, T. J., Cyphers, S., Malioutov, I., Huynh, D. and Barzilay, R.: Recent Progress in the MIT Spoken Lecture Processing Project, *in Proc. Interspeech* (2007).
[10] Park, A., Hazen, T. and Glass, J.: Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling, *in Proc. ICASSP* (2005).
[11] 渡邉真人，秋田祐哉，河原達也：予稿の話し言葉変換に基づく言語モデルによる講演音声認識，情報処理学会研究報告，2011-SLP-89-1 (2011).
[12] Masumura, R., Hahm, S. and Ito, A.: Training a Language Model Using Webdata for Large Vocabulary Japanese Spontaneous Speech Recognition, *in Proc. Interspeech* (2011).
[13] 山崎裕紀，岩野公司，篠田浩一，古井貞熙，横田治夫：講義音声認識における講義スライド情報の利用，情報処理学会研究報告，2006-SLP-64-39 (2006).
[14] 河原達也，根本雄介，勝丸徳浩，秋田祐哉：スライド情報を用いた言語モデル適応による講義音声認識，情報処理学会論文誌，Vol. 50, No. 2, pp. 469–476 (2009).
[15] Miranda, J., Neto, J. and Black, A.: Improving ASR by Integrating Lecture Audio and Slides, *in Proc. ICASSP* (2013).
[16] 黄瀬浩一，大町真一郎，内田誠一，岩村雅一：カメラを用いた文字認識・文書画像解析の現状と課題，電子情報通信学会技術研究報告, PRMU2004-246 (2005).
[17] Martinez-Villaronga, A., del Agua, M., Andres-Ferrer, J. and Juan, A.: Language Model Adaptation for Video Lectures Transcription, *in Proc. ICASSP* (2013).
[18] Kuhn, R. and De Mori, R.: A Cache-based Natural Language Model for Speech Recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 12, No. 6, pp. 570–583 (1990).