情報処理学会論文誌 Vol. 50 No. 2 469–476 (Feb. 2009)

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

河原達也[†1] 根本雄介[†1]
勝丸徳浩[†1] 秋田祐哉[†1]

大学などの講義で使用されるスライドの情報を用いて，言語モデルを動的に適応することにより，音声認識の高精度化を実現する方法を提案する．まず，当該講義のスライド全体のテキストを用いて，PLSA（Probabilistic Latent Semantic Analysis）により N-gram モデルの話題への適応を行う．次に，発話に対応する個々のスライドの情報を用いて，キャッシュモデルによりスライドに現れる単語の確率を強化し，認識結果のリスコアリングを行う．京都大学で行われた技術講習会と正規の講義を対象とした音声認識において評価を行った結果，PLSA による大域的な適応とキャッシュモデルによる局所的な適応を組み合わせることにより，認識精度の有意な改善が得られた．特に，キーワードの検出で大きな改善が得られ，大学の講義でも 80%に近い精度（F 値）を実現した．

Automatic Lecture Transcription by Exploiting Slide Information for Language Model Adaptation

Tatsuya Kawahara,[†1] Yusuke Nemoto,[†1]
Norihiro Katsumaru[†1] and Yuya Akita[†1]

We investigate several language model adaptation methods which exploit presentation slide information for automatic lecture transcription. First, N-gram probabilities are re-scaled with lecture-dependent unigram probabilities estimated by PLSA (Probabilistic Latent Semantic Analysis) using all slides of the lecture. Then, N-best hypotheses of the initial speech recognition results are re-scored using word probabilities enhanced with a cache model using the slide corresponding to each utterance. Experimental evaluations on real lectures show that the proposed method with the combination of the global and local slide information achieves a significant improvement of recognition accuracy, especially in the detection rate of content keywords.

1. はじめに

近年，講義や講演などの音声・映像をディジタルアーカイブとして蓄積し，ネットワークを通じて配信する取り組みが進められている．アーカイブには検索のためのインデックスやユニバーサルアクセスのための字幕を付与することが望ましいが，手間のかかる作業であり，半自動化できることが望ましい．このため，音声認識技術の利用が検討されている[1)–4)]．

音声認識技術は，聴覚障害のある学生のためのノートテイク（＝リアルタイムのメモ・字幕付与）支援においても有用と考えられる．現在，多くの大学において講義のノートテイクがボランティアの学生によって提供されているが，これらのノートテイカは速記者のような能力があるわけではなく，講師の発話のすべてを書き取ったり，タイプしたりすることはできない[⋆1]．また大学の講義は専門性が高いので，当該講義とノートテイカの専門分野が一致していないと，用語を正確に聞き取ることも困難になる．

このような講義を対象とした自動音声認識において十分な精度を確保するためには，音響モデル・言語モデルを話し言葉や当該ドメインに対応させる必要がある．本研究では言語モデルに焦点を置くが，話し言葉のスタイルとともに，当該分野・話題の専門用語や固有の表現をモデル化する必要がある．話題とスタイルが適合したコーパスを用いて統計的言語モデルを学習することが理想的であるが，実際にはこのようなテキストは容易に得られないので，類似のコーパスを用いてモデルを構築することが一般的である．ただし，このようなモデルは多数の話者や話題を包含するので，特定の講義の話題に対する予測能力は低下する．個別の講義に言語モデルを適応するために，講義で使用される教科書や講義の一部書き起こしを利用して，言語モデルを補間し適応する手法も提案されているが[6),7)]，このようなテキストが電子的に得られることは限定的である．

これに対して，近年多くの講義でプレゼンテーションスライド（Microsoft PowerPoint など）が使用されつつあるので，スライドの情報を利用して言語モデルの適応を行うことを考える．我々が外国語の講演を聴講する際に，スライドがあると聞き取りが容易になるのと同様に，スライドのテキストから音声認識の言語モデル適応に有用な情報が得られるものと期待される．まず，当該講義（1 回分）のスライドファイル全体から，その講義で使用される専門用語や話題に関する情報が得られる．さらに，講義途中で提示されている個々のスラ

†1 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
⋆1 一般に書き取れるのはせいぜい 2 割程度といわれる[5)]．

© 2009 Information Processing Society of Japan

イドから，その時点で発話されている内容語を予測することができる．本稿では，前者を大域的情報，後者を局所的情報とよぶ．

ただし，教科書や書き起こしのようなテキストと異なり，講義スライドはキーワードを主とする断片的な記述が中心で，テキストサイズも小さいことから，適応ということを考慮しても，信頼に足る N-gram モデルを推定することは困難である．山崎ら[8] は，スライドテキストから学習した N-gram モデルにより言語モデルの補間を行う手法を試みているが，1 コース（講義約 10 回分）全体のスライドのテキストを用いており，必ずしも現実的な前提でない．逆に，富樫ら[9] も試みているように，当該講義（1 回分）のスライドのみで N-gram モデルを直接適応しても，ほとんど効果が得られない．したがって，スライドに含まれる話題やキーワードといった情報に着目した頑健な適応の枠組みが必要である．

そこで本稿では，講義スライドから得られる大域的情報および局所的情報を効果的に活用することで，言語モデルの適応を行う手法を提案する．スライドのような少量のデータから効果的な適応を実現するために，内容語の集合から話題をモデル化する PLSA（Probabilistic Latent Semantic Analysis）[10] と，キーワードの出現をモデル化するキャッシュモデル[11] の枠組みを導入する．具体的には，講義スライド全体の情報を用いて PLSA による N-gram モデルの大域的な適応を行い，さらに講義スライドと講義音声の時系列の対応に基づいてキャッシュモデルを適用し，局所的な適応を行う．

そのうえで，実際に大学で行われた講義において，これらの手法の評価を行う．特に，スライドから得られる情報を用いた適応の効果があるか，内容の把握において重要となるキーワードの認識において有効であるか，などの点に着目して評価する．

2. PLSA に基づく言語モデルの大域的適応

PLSA [10] は，単語の出現頻度を用いて文書集合中の文書と単語を特徴づける枠組みであり，文書 d，単語 w に対して式 (1) で定式化される．

$$P(w|d) = \sum_{j=1}^{N} P(w|t_j)P(t_j|d) \tag{1}$$

あらかじめ大規模な文書コーパスを用いて，文書の特徴（たとえば話題）を表す N 次元部分空間 $\{t_j\}$ に関する確率 $P(w|t_j)$，$P(t_j|d)$ を EM アルゴリズムにより推定する．そして，この部分空間に特定の文書 d を射影することにより，これに依存した単語 w の生起確率 $P(w|d)$ が求められる．単語頻度に基づく射影であるため，短いフレーズを中心に記述されている講義スライドでも有効であると期待される．また，潜在意味空間を介して，文書 d に出現していない単語でも，同じ話題 t_j に属していれば，その確率が更新されるため，出現単語が限定される講義スライドにおいて頑健な適応ができるものと期待される．

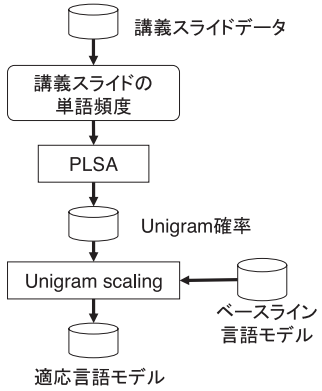

図 1　PLSA に基づく言語モデル適応
Fig. 1　Language model adaptation based on PLSA.

本研究では，適応用の文書として講義スライドを使用する．適応対象となる講義において使用されたスライドファイル全体から抽出したテキスト[⋆1]を S_{all} とし，S_{all} を PLSA による部分空間へ射影することで，講義内容に依存した単語確率 $P(w|S_{all})$ を求める．

ただし，この推定式を 3-gram 確率に直接適用するには膨大なパラメータと計算量が必要となり，現実的でない．そこで，PLSA では unigram 確率のみを推定し，これに基づいて，ベースライン言語モデルの 3-gram 確率に対して式 (2) によるリスケーリング[12)–14)] を行う．

$$P(w_i|w_{i-2}, w_{i-1}; S_{all}) \propto \frac{P(w_i|S_{all})}{P(w_i)} P(w_i|w_{i-2}, w_{i-1}) \tag{2}$$

上記のスライド情報を用いた言語モデル適応の流れを**図 1** に示す．なお，スライドの話題に対してのみ適応を行うため，内容語と考えられる名詞（サ変動詞の語幹含む），具体的には，接頭，接尾，非自立，数，代名詞を除く名詞に限定して上記を適用し，その他の機能語や汎用的な単語に対してはベースライン言語モデルによる確率をそのまま用いた．数（数

⋆1 Microsoft PowerPoint の「アウトライン」を抽出．図・表・式のテキストは抽出されない．タイトルなどの区別はしていない．

© 2009 Information Processing Society of Japan

詞）については，種類が多く，話題との関連づけが難しいうえに，スライドでは箇条書き表現で用いられる場合が多いので除外した．

3. 関連 Web テキスト収集による言語モデルの適応

上記の手法に加えて，当該講義に関連する Web テキストを収集して言語モデルを補完する方法[15),16)]も検討した．Web テキストの収集による言語モデル適応の概要を**図 2** に示す．

まず，講義で使用される各スライドの話題に合致した文書を，学習に十分な数だけ収集する．そのために，各スライドからこれを特徴づけるキーワードを抽出し，検索クエリを生成する．具体的には，講義で使用された各スライドに含まれる名詞から $tf{\cdot}idf$ 値の上位 3 単語を選択し，これらの AND 検索を実行する．これを各スライドに対して繰り返す．このとき，各クエリによる収集ページ数の上限を 500 とした．

収集された Web テキストは，話題に関しては対象講義におおむね適合するものの，話し言葉や文章のスタイルになっていないものを多数含むので，言語モデルの学習に適した文を選択する必要がある．選択に先立って，タグや記号，1 文が一定の長さ（=10 文字）より短いものや，一定の長さ（=7 文字）以上のアルファベット系列が含まれる文を除去する．この前処理を経たテキスト中の各文に対して，ベースライン言語モデルによりパープレキシティを計算し，この値が小さい文を選択する[15)]．

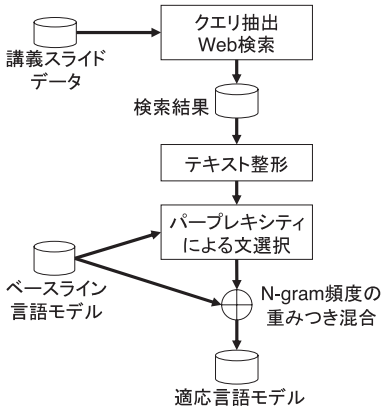

図 2　Web テキスト収集による言語モデル適応
Fig. 2　Language model adaptation using Web text retrieval.

このように収集・選択された Web テキストの N-gram 頻度を，ベースライン言語モデルの N-gram 頻度に重み付きで混合することで適応を行う．

4. キャッシュモデルに基づく局所的適応

キャッシュモデル[11)]では，単語 w_i の直前の単語履歴をキャッシュ$H = \{w_{i-|H|}, \cdots, w_{i-1}\}$として記憶し，これに含まれる単語が再び使用される確率が高いと予測する．このキャッシュに基づく単語 w_i の出現確率 $P_c(w_i|H)$ は，式 (3) により与えられる．ただし，$|H|$ は単語履歴 H の長さ，δ はクロネッカーのデルタである．

$$P_c(w_i|H) = \frac{1}{|H|} \sum_{w_h \in H} \delta(w_i, w_h) \tag{3}$$

これにより，繰り返し発話されるキーワードなどが認識されやすくなる効果が期待できる．

本研究では，キャッシュモデルの枠組みに基づいて，講義スライド中の単語の出現情報を用いることで単語の生起を予測する．単語 w_i が含まれる発話に対応するスライドに含まれる単語のリスト S をキャッシュに置き換えることにより，式 (3) は式 (4) のようになる．ただし，$|S|$ はスライド S に含まれる総単語数である．

$$P_s(w_i|S) = \frac{1}{|S|} \sum_{w_s \in S} \delta(w_i, w_s) \tag{4}$$

これにより，スライドに出現するキーワードが認識されやすくなる効果が期待できる．

さらに，発話の履歴 H と対応するスライド S の両方を考慮した場合のキャッシュモデルも構成できる．

$$P_{cs}(w_i|H,S) = \frac{1}{|H|+|S|} \left\{ \sum_{w_h \in H} \delta(w_i, w_h) + \sum_{w_s \in S} \delta(w_i, w_s) \right\} \tag{5}$$

このキャッシュモデルによる確率 $P_c(w_i|H)$，$P_s(w_i|H)$，$P_{cs}(w_i|H,S)$（のいずれか）を，3-gram 言語モデルによる確率 $P(w_i|w_{i-2}, w_{i-1})$ と重み付き線形補間することで適応を行う．

キャッシュモデルに基づく言語モデル適応の流れを**図 3** に示す．ベースライン言語モデルを用いた音声認識により得られた N-best 仮説中の各単語に対して，式 (3)〜(5) による推定を行い，ベースライン言語モデルとの線形補間を行う．得られた適応確率を用いて，N-best 仮説の言語モデル尤度を更新し，リスコアリングを行う．ここで，ベースライン言語モデルの代わりに，PLSA による適応を行った言語モデルを用いることで，講義で使われるスライドファイル全体の大域的情報と個々の発話に対応するスライドの局所的情報を組み合わせた

© 2009 Information Processing Society of Japan

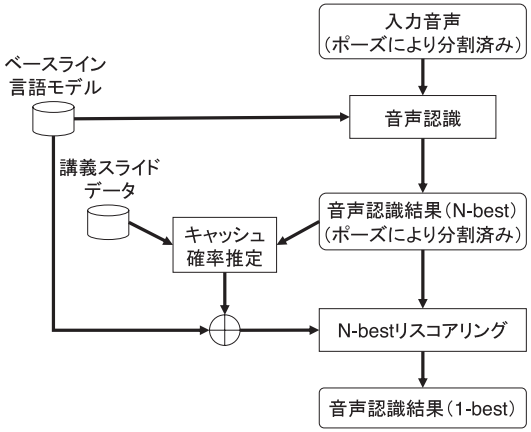

図 3　キャッシュモデルに基づく言語モデル適応
Fig. 3　Language model adaptation based on cache model.

適応が実現される．

5. 評 価 実 験

5.1 実験データと条件

技術講習会と大学講義の 2 種類の講義音声を対象に評価実験を行った．講習会の音声は，2004 年と 2005 年に京都大学学術情報メディアセンターで行われた音声認識・音声対話技術講習会[⋆1]における 12 回の講義の分である．大学講義の音声は，京都大学の学部および大学院向けに行われた 3 科目（画像処理論，パターン認識，パターン認識特論）の各 1 回分である．各講義の時間はおおむね 90 分である．講習会・大学講義あわせて講師の重複は 1 名のみである．音声は，ワイヤレスのピンマイク（SONY WRT-824）を通して，自動アーカイブシステム[17] で収録された．

講習会の各講義音声に含まれる単語数はおおむね 15〜20 K であり，テストセット 12 回分の合計で 215 K 単語である．大学講義の方はやや発話速度が遅く，各講義おおむね 15 K 単語であり，テストセット 3 回分の合計で 47 K 単語である．また，これらにおいて使用された講義スライドとその時間情報が利用可能である．各講義で使用されたスライドの数は，講習会では大半が 40 以上（最大 97）であったのに対して，大学講義では 20〜40 程度であった．スライド 1 枚あたりの単語数は平均約 50 であった．

音声認識には，Julius 3.5.2 デコーダを用いた．音声はあらかじめ無音区間により発話ごとに区切っている．使用した音響モデルは，『日本語話し言葉コーパス』（CSJ）に含まれる 257 時間の学会講演から話者適応学習（SAT: Speaker Adaptive Training）した 3,000 状態・64 混合の状態共有 triphone HMM に対して，教師なしで MLLR（Maximum Likelihood Linear Regression）話者適応を行ったものである．ベースライン言語モデルは，CSJ の学会・模擬 2,720 講演（7 M 単語）から学習した語彙サイズ 50 K の単語 3-gram モデルである．

ベースライン言語モデルによるテストセットパープレキシティは講習会が 115.3，大学講義が 152.8 であった．講習会の方が大学講義に比べてパープレキシティが小さいのは，講習会における発話スタイルや話題が大学講義に比べて，CSJ の学会講演に近いためと考えられる．具体的には，学内の学生に話しかける大学講義に比べて，社会人や他大学の学生を対象とした講習会の方が，学会講演のように丁寧でかつ緊張感のある発話スタイルになっている．話題についても，日本音響学会の研究発表会など，音声認識・音声対話に関する講演が CSJ に多く含まれている．

また，ベースライン言語モデルによる講習会音声の平均単語認識精度は 71.60%，大学講義音声の平均単語認識精度は 58.61%であった．講義ごとの単語認識精度は，講習会では 65〜80%に分布していたが，大学講義では平均値から 2%以内であった．スライド中に現れた未登録語を単語辞書に追加したところ，講習会において 0.23%，大学講義において 0.19%の単語認識精度の改善（絶対値）が得られた．以後の評価実験においては，このスライドから未登録語を追加した単語辞書を使用する．

5.2 大域的適応の評価

まず，PLSA による大域的適応の実装・評価を行った．PLSA の部分空間はベースライン言語モデルの学習に用いたものと同一の CSJ 学会講演により構築した．部分空間の次元数は，予備実験によって 100 と定めた．

表 1 に，講習会と大学講義それぞれに対するテストセットパープレキシティの平均値を示す．スライドファイルに出現したすべての単語を用いて PLSA を行った場合は，パープレキシティが 20%以上も増大した．内容語に絞って適応を行うことにより，講習会に対してはベースライン言語モデルからの改善が得られた．これにより，スライドのような断片的なテキストから適応を行う際には，内容語に絞ることが不可欠であることが確認された．

⋆1 http://seminar.media.kyoto-u.ac.jp/archive/2004_onsei.php,
http://seminar.media.kyoto-u.ac.jp/archive/2005_onsei.php

© 2009 Information Processing Society of Japan

表 1　大域的適応によるテストセットパープレキシティ
Table 1　Test-set perplexity after global adaptation of language model.

	講習会	大学講義
ベースライン	115.3	152.8
PLSA（スライド；全単語）	139.0	237.0
PLSA（スライド；内容語のみ）	110.2	165.5
PLSA（認識結果）	101.6	133.3
PLSA（正解書き起こし；参考）	97.7	126.4
Web text（20 M）	105.6	124.3
Web text（50 M）	108.6	122.2

比較対象として，ベースライン言語モデルによる音声認識結果を用いた PLSA 適応[13] も行った．表 1 には，正解の書き起こしを用いた場合の結果も示している．この場合は，内容語だけでなく付属語も含めてすべての単語で適応を行うことが効果的であり[13]，全体としてパープレキシティのより大きな削減が得られている．しかしながらこの手法は，講義全体を 2 回認識（デコーディング）する必要があるため，リアルタイムの字幕付与・ノートテイクには利用できない．

次に，関連 Web テキスト収集による適応を行った．ここでは，特定領域研究「情報爆発 IT 基盤」において開発されている検索エンジン Tsubaki [⋆1] を使用した．Web 検索のクエリ生成のためのキーワード選定に使用する tf·idf 値の計算の際には，tf として各スライドにおける単語頻度を，idf として前記の CSJ の 1 講演を 1 文書と見なした文書頻度に基づく値を使用した．収集したテキストから，ベースライン言語モデルによるパープレキシティを用いて文を選択するが[15]，その際の閾値を変化させて，20 M 単語および 50 M 単語のテキストが選択されるようにした．収集したテキストによる N-gram 頻度をベースライン言語モデルの N-gram 頻度と混合する際の重みは，予備実験により 0.1 と定めた．

テストセットパープレキシティによる評価を表 1 に示す．講習会においては収集テキストサイズが 20 M 単語のとき 8.5%，大学講義においては 50 M 単語のとき 19%，パープレキシティが削減された．大学講義において効果が大きいのは，CSJ の学会講演と講習会の内容にかなりの重なりがあるのに対して，大学講義の話題は CSJ で十分にカバーされていないためと考えられる．

次に，音声認識実験による評価を行った．各手法による単語認識精度（Word Accuracy）による評価を**表 2** にまとめる．

表 2　各手法による単語認識精度
Table 2　Word recognition accuracy after language model adaptation.

手法	講習会 acc.(%)	大学講義 acc.(%)	認識条件
ベースライン	71.60	58.61	○
未登録語追加	71.83	58.80	○
PLSA（スライド）	72.40	59.41	○
PLSA（認識結果）	72.83	60.37	×
テキスト混合（認識結果）	71.72	59.09	×
テキスト混合（Web，20 M）	72.37	60.50	○
テキスト混合（Web，50 M）	72.45	60.91	○
キャッシュ（**スライド**）	72.44	60.11	△
キャッシュ（認識結果）	72.24	59.63	△
スライド+キャッシュ	72.66	60.30	△
PLSA+スライド	72.98	60.68	△
PLSA+キャッシュ	72.80	60.42	△
PLSA+スライド+キャッシュ	73.11	60.97	△

（○：認識 1 回，×：認識 2 回，△：リスコアリング）

スライドからの PLSA 適応では，講習会・大学講義の双方において単語認識精度が 0.80%改善された．これは，有意水準 1%で統計的に有意な改善である．音声認識結果を用いた PLSA 適応による改善は講習会において 1.23%，大学講義において 1.76%であった．音声認識結果にはスライドにはない情報も含まれることが多く，これが精度に影響したと考えられる．しかし前述のとおり，この手法はリアルタイム認識には適用できない．

なお比較のため，音声認識結果の単語列から推定した N-gram 頻度をベースライン言語モデルのものに混合して音声認識を行ったところ（表 2 のテキスト混合（認識結果）），単語認識精度の改善は講習会において 0.12%，大学講義において 0.48%であり，上記の手法に及ばなかった．

Web テキスト収集による適応では，収集テキストサイズが 50 M 単語のとき，講習会において 0.85%，大学講義において 2.30%，単語認識精度が向上した．講習会と大学講義の間で効果に大きな差（1.5%）が見られたのは，テストセットパープレキシティによる評価と同様である．

5.3　局所的適応の評価

キャッシュモデルに基づく言語モデル適応に必要なスライドと発話の対応関係は，講義収

⋆1 http://tsubaki.ixnlp.nii.ac.jp/se/index.cgi

© 2009 Information Processing Society of Japan

録時に記録されたスライドの切替え時間情報に基づいて与えた[⋆1]．キャッシュの長さ $|H|$ と線形補間の重みは，講習会音声に対するクロスバリデーションにより $|H| = 60$，重み 0.1 と決定し，大学講義における実験でも同一の値を使用した．単語認識精度を表 2 に示している．

スライド（すなわち P_s）のみ使用して適応を行ったところ，講習会において 0.84%，大学講義において 1.50%の単語認識精度の改善が得られた．これはスライドを利用しない通常のキャッシュ（すなわち P_c）の使用による改善（講習会で 0.64%，大学講義で 1.02%）を上回っている．さらに，通常のキャッシュをスライドと併用することで，講習会，大学講義でそれぞれ 1.06%，1.69%の改善となった．4 章で述べたように，スライドをキャッシュとすることにより発話に対応するスライド中の単語が，通常のキャッシュモデルにより直前に発話された単語が，それぞれ認識されやすくなる効果が確認された．

5.4　大域的適応と局所的適応の統合

スライド全体を使用して PLSA による言語モデル適応を行ったもの（表 2 の 3 行目）に対して，キャッシュモデルによる 3 種類のリスコアリングを適用した結果を表 2 の下段に示す．ここでは，ノートテイク支援にも応用できるように高速な適応を指向して，テキスト収集に時間を要する Web テキストによる適応は行わなかった．

PLSA による言語モデル適応と，発話に対応するスライドとキャッシュを併用したリスコアリングを組み合わせることで，講習会，大学講義のそれぞれにおいて 1.51%，2.36%の単語認識精度の改善が得られた．PLSA による適応とキャッシュモデルによる効果がほぼ加算的に現れており，講義全体の大域的な情報と発話周辺の局所的な情報の組合せが有効であることが示された．

各手法および組合せについて，講義ごとにその効果のばらつきは見られるが，効果の得られないものはなかった．ただし，効果の程度とスライド数などとの明確な関連は見いだせなかった．

5.5　キーワード認識精度による評価

次に，講義内容を把握するうえで重要であり，インデックスとしても有用であると考えられるキーワードの認識精度を調べた．ここでは，スライド中に出現する名詞のうち，接頭，接尾，非自立，数，代名詞を除いたものをキーワードとした．キーワードの数は各講義で 1〜3 K，テストセット（15 講義）全体で 30 K 単語である．その検出精度を F 値で**表 3** に示す．

表 3　各手法によるキーワード検出精度
Table 3　Keyword detection accuracy after language model adaptation.

手法	講習会 F 値 (%)	大学講義 F 値 (%)
ベースライン	85.00	70.78
未登録語追加	85.28	70.87
PLSA（スライド）	86.64	74.78
PLSA（認識結果）	86.75	75.03
テキスト混合（認識結果）	85.67	72.05
テキスト混合（Web，20 M）	85.78	75.09
テキスト混合（Web，50 M）	86.01	75.92
キャッシュ（**スライド**）	87.36	75.93
キャッシュ（認識結果）	86.40	73.07
スライド+キャッシュ	87.59	75.96
PLSA+スライド	88.18	78.37
PLSA+キャッシュ	87.39	76.76
PLSA+スライド+キャッシュ	88.24	78.60

（◯：認識 1 回，×：認識 2 回，△：リスコアリング）

手法間の比較については，単語認識精度とおおむね同様の傾向が見られるが，ベースラインと比べた場合の改善幅は全般に大きく，提案手法をすべて統合した場合（表 3 の最下段），講習会において 3.24%，大学講義において 7.82%の F 値の改善が得られた．なお，講習会では，再現率 84.12%，適合率 92.78%，大学講義では，再現率 71.71%，適合率 86.96%となった．適合率はベースラインでも約 9 割であったのに対して，再現率で大きな改善が得られている．これらの改善幅が前節で述べた単語認識精度の改善幅より著しく大きいことから，提案する適応手法が，スライドに出現するようなキーワードの認識において特に有効であるといえる．講習会に比べて，大学講義において改善幅が大きいのは，講習会の話題が CSJ によるベースライン言語モデルでかなりカバーされていたためと考えられる．

6. おわりに

本稿では，講義で使用されるスライドの情報を活用して，PLSA とキャッシュモデルの枠組みに基づいて，音声認識用の言語モデルの適応を行う手法を提案した．京都大学で行われた技術講習会と正規の講義の音声を対象に評価を行った結果，PLSA による大局的な適応とキャッシュモデルによる局所的な適応を組み合わせることにより，単語認識精度の有意な改善が得られ，特にキーワードの検出に限定すると，大学の講義で約 8 ポイントも向上した．

⋆1 発話がスライドの切替えに重なる場合は，重なり時間が長い方のスライドに対応づけた．

© 2009 Information Processing Society of Japan

今後の研究課題として 2 つあげられる．第 1 に，大学の講義に対する音声認識精度の改善である．今回の評価でも，講習会と比較して大学講義の方が全般に低い認識精度となった．大学講義の方が，発話スタイルがあまりフォーマルでなく，話題も CSJ でカバーされていないためと考えられる．ただし，同じ講師が何回も講義を行うので，書き起こしはコスト的に現実的でないとしても，事前に適応データを収集できる可能性がある．また，発話スタイル/話者と話題については，ある程度独立にモデル化することも考えられる[13]．

第 2 の課題として，音声認識結果から字幕を生成・編集する方法について検討し，ノートテイク支援を行うプロトタイプシステムを開発することである．NHK では，音声認識に基づいて放送用字幕を作成・編集するシステムを構築・運用しているが[18]，編集作業には相当の訓練を要している．講義の場合は放送番組ほど完璧な字幕を要求されないが，学生ボランティアでも容易に操作できるようなインタフェースを模索する必要がある．

謝辞　本研究の一部は，総務省 SCOPE「音声認識技術を用いた会議録及び字幕の作成支援システム」の支援により行われた．本研究で使用した講義データの収録にご協力頂いた京都大学美濃研究室の皆様，西口敏司先生に感謝します．

参 考 文 献

1) 岡本拓明，仲野　亘，小林隆志，直井　聡，横田治夫，岩野公司，古井貞熙：音声情報を統合したプレゼンテーションコンテンツ検索，信学論，Vol.J90-D, No.2, pp.209–222 (2007).
2) 北出　祐，河原達也：講義の自動アーカイブ化のためのスライドと発話の対応付け，情報処理学会研究報告，2005-SLP-55-11 (2005).
3) 富樫慎吾，山口　優，北岡教英，中川聖一：講義音声の認識・要約・インデックス化の検討，情報処理学会研究報告，2006-SLP-62-11 (2006).
4) Glass, J., Hazen, T.J., Hetherington, L. and Wang, C.: Analysis and Processing of Lecture Audio Data: Preliminary Investigations, *Proc. HLT-NAACL* (2004).
5) 吉川あゆみ，太田晴康，広田典子，白澤麻弓：大学ノートテイク入門，人間社 (2001).
6) Park, A., Hazen, T. and Glass, J.: Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling, *Proc. ICASSP* (2005).
7) Trancoso, I., Nunes, R., Neves, L., Viana, C., Moniz, H., Caseiro, D. and Mata, A.I.: Recognition of Classroom Lectures in European Portuguese, *Proc. Interspeech* (2006).
8) 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-64-38 (2006).
9) 富樫慎吾，北岡教英，中川聖一：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，日本音響学会春季講演論文集，1-P-24 (2006).
10) Hoffman, T.: Probabilistic Latent Semantic Indexing, *Proc. SIG-IR* (1999).
11) Kuhn, R. and De Mori, R.: A Cache-based Natural Language Model for Speech Recognition, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.12, No.6, pp.570–583 (1990).
12) Gildea, D. and Hoffman, T.: Topic-based Language Models using EM, *Proc. Eurospeech* (2003).
13) Akita, Y. and Kawahara, T.: Language model adaptation based on PLSA of topics and speakers for automatic transcription of panel discussions, *IEICE Trans.*, Vol.E88-D, No.3, pp.439–445 (2005).
14) 栗山直人，鈴木基之，伊藤彰則，牧野正三：情報量基準で語彙分割した PLSA 言語モデルによる話題・文型適応，情報処理学会研究報告，2006-SLP-64-41 (2006).
15) 翠　輝久，河原達也：ドメインとスタイルを考慮した Web テキストの選択による音声対話システム用言語モデルの構築，電子情報通信学会論文誌，Vol.J90-D, No.11, pp.3024–3032 (2007).
16) Suzuki, M., Kajiura, Y., Ito, A. and Makino, S.: Unsupervised Language Model Adaptation Based on Automatic Text Collection from WWW, *Proc. Interspeech* (2006).
17) 西口敏司，亀田能成，角所　考，美濃導彦：大学における実運用のための講義自動アーカイブシステムの開発，信学論，Vol.J88-DII, No.3, pp.530–540 (2005).
18) 安藤彰男：ニュース音声自動字幕化システム，情報処理学会研究報告，SLP-34-28 (2000).

(平成 20 年 6 月 3 日受付)
(平成 20 年 11 月 5 日採録)

© 2009 Information Processing Society of Japan

河原　達也（正会員）

1987 年京都大学工学部情報工学科卒業．1989 年同大学院修士課程修了．1990 年同博士後期課程退学．同年京都大学工学部助手．1995 年同助教授．1998 年同大学情報学研究科助教授．2003 年同大学学術情報メディアセンター教授．現在に至る．この間，1995 年から 1996 年まで米国ベル研究所客員研究員．1998 年から ATR 客員研究員．1999 年から 2004 年まで国立国語研究所非常勤研究員．2001 年から 2005 年まで科学技術振興事業団さきがけ研究 21 研究者．音声言語処理，特に音声認識および対話システムに関する研究に従事．京都大学博士（工学）．1997 年度日本音響学会粟屋潔学術奨励賞受賞．2000 年度情報処理学会坂井記念特別賞受賞．情報処理学会連続音声認識コンソーシアム代表，IEEE SPS Speech TC 委員，IEEE ASRU 2007 General Chair，言語処理学会理事，を歴任．情報処理学会音声言語情報処理研究会主査．日本音響学会，人工知能学会各評議員．電子情報通信学会，言語処理学会，IEEE 各会員．

根本　雄介

2005 年京都大学工学部情報学科卒業．2007 年同大学院情報学研究科修士課程修了．現在，トヨタ自動車株式会社に勤務．在学中，音声言語処理の研究に従事．

勝丸　徳浩

2007 年京都大学工学部情報学科卒業．現在，同大学院情報学研究科修士課程在学中．音声言語処理の研究に従事．

秋田　祐哉（正会員）

2000 年京都大学工学部情報学科卒業．2002 年同大学院情報学研究科修士課程修了．2005 年同博士後期課程修了．京都大学博士（情報学）．2005 年より京都大学学術情報メディアセンター助手（現，助教）．音声言語処理の研究に従事．2007 年日本音響学会粟屋潔学術奨励賞受賞．電子情報通信学会，日本音響学会，IEEE 各会員．

© 2009 Information Processing Society of Japan