2008－SLP－72 (5)
2008／7／18

大学講義のノートテイク支援のための音声認識用言語モデルの適応

勝丸 徳浩　秋田 祐哉　森 信介　河原 達也

京都大学 情報学研究科
〒606-8501 京都市左京区吉田本町

我々は，大学の講義におけるノートテイク支援を目標とした音声認識システムの研究開発を行っている．本研究では，専門性の高い講義に対して，言語モデルを効率的に適応する方法を検討する．大学の講義では，同一の講師が同一科目を一定期間担当することが通例であるので，以前の講義のデータを活用することを考える．ノートテイカが音声認識結果を評価・選別する応用場面を想定して，認識結果を教師ありで言語モデル適応に用いる方法と，音声認識結果の信頼度に基づいて教師なしで適応する方法を検討する．さらに，講義スライドを用いて，PLSA や Web テキスト収集に基づいて適応する手法との統合も行い，効果を確認した．

Language Model Adaptation for Automatic Speech Recognition to Support Note-Taking of Classroom Lectures

Norihiro KATSUMARU　Yuya AKITA　Shinsuke MORI　Tatsuya KAWAHARA

School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract We are developing an automatic speech recognition (ASR) system to assist note-taking in the classroom. In this work, we focus on an efficient method to adapt the language model (LM) for ASR to university lectures, in which a number of technical terms are used. We assume that one lecturer teaches a specific course subject through a certain period (a semester), and exploit the data of the lectures previously given by the same lecturer. Specifically, we propose an LM adaptation scheme supervised by the note-takers, who verify the ASR results and filter the well-recognized hypotheses. We also investigate an unsupervised adaptation method based on the confidence score of ASR. The methods are combined with other LM adaptation methods based on PLSA and Web text collection using the lecture slides.

1 はじめに

近年，情報保障が重要視されている．聴覚に障害があっても，健常者と同じ情報に接するために，講義内容を文字化することは1つの情報保障である．現在，聴覚障害の学生に対する講義の補助として多くの大学で，ノートテイクが行われている．ノートテイクは，学生のボランティアなどにより講義内容を書き起こすものである．通常，話す速さは書く速さよりも速いため，2人以上で交代で行うものの，講義の全発話を書き起こすこは非常に困難である．

特に，大学の講義は専門性が高く，専門分野が同一の人でないと聴き取り自体が困難である．しかし，専門分野が同一の人を継続的に確保するのは困難である．そこで，音声認識技術を活用することでノートテイクを全自動化，もしくは，半自動化することができれば有益である．

自動音声認識を実現するためには，認識対象の音響的・言語的特徴を適切にモデル化する必要がある．現在の音声認識のモデルは，統計的モデルであり，高い認識精度を確保するには十分な量のデータからモデルを学習する必要がある．しかし，大学講義の発話に関してそのような大規模なコーパスは存在しない．また，専門用語が多数出現し，音声認識システムで未登録となる問題が起こりうる．したがって，言語モデルや単語辞書を適応する必要がある [2][3]．そこで我々は，講義で使われるスライドを用いて適応する方法を提案している [1]．ただし，用いる講義スライドは，単体で言語モデルが構築できるほどテキストの分量が十分あるわけではないので，PLSA による適応や Web テキスト収集による適応といった手法を提案した．

また，大学の講義では，通常1人の講師が続けて同じ科目の授業を行う．そのため，同一講師が以前行った一連の講義の内容は，認識対象の講義と親和性が高いと期待される．ただし，人手による書き起こしはコスト的に困難であるため，認識結果を基に適応学習用のデータを作成することを考える．本研究では具体的に，音声認識の信頼度を用いる手法 (教師なし)，及びノートテイカによる正解判定による手法 (教師あり) を提案する．

以下，第2章では，ノートテイクにおける音声認識技術の利用について概観する．第3章では，言語モデル適応の概要と個々の手法の詳細について述べる．第4章では，評価実験について述べる．第5章では，全体を総括する．

2 ノートテイクの現状と音声認識の利用

2.1 ノートテイクの現状

現在，ノートテイクは人手で行われている．通常は，支援を受ける人を左右から挟むように2名のノートテイカが座り，数分毎に交代しながら講師の発話を書き起こしていく．発話速度の方が，書き起こす速度よりも速いので，ノートテイカは要約筆記を行う．ペンによる手書きが依然主流であるが，最近は，IPtalk[9] と呼ばれるソフトウェアを用いたパソコン入力を用いる場合が増えてきている．ただし，この場合も手動による入力であることに変わりはない．

2.2 音声認識を利用したノートテイク

音声認識システム単独でノートテイカの代わりがつとまる認識精度を確保するのは容易ではない．そこで，音声認識技術はあくまで入力の補助手段ととらえて，ノートテイカの負担・人数を減らすことを目標とする．

提案するシステムの図1に示す．

音声認識システムを利用する場合，ノートテイカの役割は大きく2点ある．まず，音声認識システムによる認識文の成否を判定し，成功した場合はそのままユーザに提示することである．次に，認識が失敗した場合に，できる範囲で訂正を行うことである．

本研究では，ノートテイカによる認識結果の選別，及び修正の結果を利用し，音声認識性能の改善を図ることを考える．

3 言語モデル適応

3.1 処理の概要

最近の講義では，電子化されたスライドを用いる場合が増えている．そこで，講義スライド

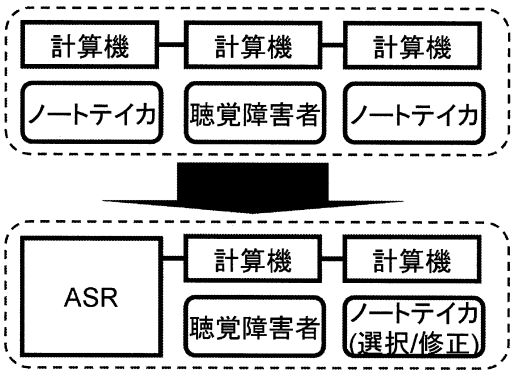

図 1: ASR を利用したノートテイクシステム

を用いて，講義ごとの言語モデル適応を行う．通常，スライドに記述されるテキストの分量は言語モデルの構築に十分ではない．そこで，少数のテキストから言語モデル適応を行う手法として，PLSA による適応と，Web テキスト収集による適応を行う [1].

さらに，同じ講師の同一科目の以前の講義の書き起こしから言語モデルを構築し，ベースライン言語モデルと混合することも考える．その際に前記の通り，ノートテイカの作業の情報と連携する．

3.2 話題語の定義と再現率・適合率

ノートテイカの支援をする音声認識システムは，講義を理解する上で重要な語を正しく認識することに重点を置くべきである．そこで，本研究では，講義資料をもとにその講義において重要な語 (話題語) を定め，話題語の再現率・適合率も評価尺度とする．

話題語は，講義資料中に含まれる名詞 (接頭，接尾，非自立，数，代名詞を除く) として定義する．話題語の適合率は，音声認識システムが認識した話題語のうち，正しく認識できたものの割合である．話題語の再現率は，書き起こし中に存在する話題語のうち，音声認識システムが正しく認識した話題語の割合である．

3.3 PLSA による適応

PLSA[4] は文書に依存した単語の生起確率を用いて，文書集合中の文書を特徴づける枠組みである．単語頻度に基づく部分空間への射影を行うため，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．

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全スライドから摘出したテキストを S_{all} とし，これを PLSA による部分空間へ射影することで，講義スライドの内容に依存した単語確率 $P(w|S_{all})$ を求める．求めた $P(w|S_{all})$ をもとに，ユニグラムスケーリングを行うことでトライグラムモデルの適応を行う (式 1).

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

3.4 Web テキストを用いた言語モデルの適応

言語モデルの学習テキストを補完するために Web を利用する．講義で使用されたスライドから話題語を選別し，tf•idf 値の上位 3 単語で AND 検索を行う．tf として，各スライドにおける単語頻度を，idf として，CSJ の学会・模擬講演の 1 講演を 1 文書とみなした文書頻度に基づく値を使用する．収集ページの上限を 1 クエリあたり 500 件とした．収集された Web テキストには，言語モデルの構築に適さないものも含まれる．そこで，ベースライン言語モデルによるパープレキシティの小さいものから 50MB 分のテキストを選択して言語モデルを構築する．これとベースライン言語モデルを重み付きで混合することで，適応言語モデルを構築する [1].

3.5 以前の講義の書き起こし・認識結果を利用する適応

同一講師による以前の講義は，専門分野・話題と話者性の両面で認識対象の講義に最もマッ

チしていると考えられる．認識対象と親和性の高い以前のテキストを利用する枠組みとして [6] がある．しかしながら，講義音声を忠実に書き起こすのはコストが高い．そこで，音声認識結果の利用を検討する．ただし，音声認識結果は誤りが多いので，信頼できる部分を選別することが必要となる．教師なしで信頼できる部分を選別する手法と，ノートテイカーを教師として信頼できる部分を選別する手法を検討する．

教師なしの場合，信頼できる部分の選別は，認識結果の内容語の音声認識結果の信頼度をもとに行う [7]．具体的には，信頼度が閾値以上の内容語の割合が多い発話単位から言語モデルを構築し，ベースライン言語モデルと混合する．閾値は，学習用の講義において正解・不正解別に内容語の信頼度の分布をとり，正解の分布と不正解の分布の谷にあたる値とする．

ノートテイカを教師とする場合，信頼できる部分の選別は，発話単位の内容語の正解率をもとに行う．正解率の高い発話単位のみから言語モデルを構築し，ベースライン言語モデルと混合する．

内容語は，名詞，動詞，形容詞とした．ここでは，ポーズで区切られたの発話単位を一文とする．

4 評価実験

4.1 実験条件

本学で 2007 年度後期に開講された地球工学科 3 回生向けの,「資源工学のための材料学-線形破壊力学-」の中から 3 回連続する講義に対して実験を行った．収録は，ピンマイクを講師の襟元に着用してもらった．非圧縮，48kHz でサンプリングしたものを 16kHz にダウンサンプリングし認識を行った．

3 回目の講義をテストセットとし，1 回目，2 回目の講義をモデル適応に用いた．

音声認識には，Julius 3.5.3 を用いた．音声はあらかじめ発話ごとに区切られている．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 (SAT) 学習した 3,000 状態，64 混合の状態共有トライフォン HMM である．前 2 回の講義で教師あり MLLR 話者適応を行った．ベースライン言語モデルは，CSJ の学会・模擬講演 2720 講演から学習した語彙サイズ 50K のトライグラムモデルである．言語モデルの混合には [5] の手法を用いた．Web テキスト収集による適応を行う際，検索エンジンは goo[8] を用いた．

4.2 PLSA 及び Web テキスト収集による適応の評価

PLSA による適応と Web テキスト収集による適応の組合せ方法は，ベースラインに PLSA による適応を行った後，Web テキストから構築した言語モデルを混合することで行った．

単語認識精度及び話題語の再現率・適合率をそれぞれ表 1，表 2 に示す．

表 1: PLSA，Web 適応における単語認識精度

	単語認識精度
ベースライン (Base)	56.27
PLSA	57.21
Web	57.70
PLSA+Web	58.11

表 2: PLSA，Web 適応における話題語の検出精度

	再現率	適合率	F 値
ベースライン (Base)	47.13	95.23	63.05
PLSA	54.56	94.36	69.14
Web	58.72	97.57	73.31
PLSA+Web	62.43	96.02	75.66

PLSA による適応により，認識精度が 0.94%，話題語の検出精度の F 値が 6.09%それぞれ改善した．Web テキスト収集による適応により，認識精度が 1.43%，話題語の検出精度の F 値が 10.26%それぞれ改善した．PLSA 適応と Web テキスト収集を組み合わせた場合，認識精度が 1.84%，話題語の検出精度の F 値が 12.61%改善した．

PLSA による適応と Web テキスト収集による適応のいずれも認識精度，話題語の検出精度の F 値ともに改善しており適応の効果が見られる．また両者を組み合せた場合の効果も確認できた．この結果は以前に我々が行った実験 [1] とも整合性がとれている．

4.3 以前の講義の書き起こし・認識結果を利用した適応の評価

音声認識の信頼度による文選択の効果を調べるために，各発話単位ごとに，認識結果に含まれる内容語の信頼度が閾値を超えるものの割合が多いものを，40%から90%まで10%おきに選択し混合した．閾値は，前2回の講義のうち，一方を学習用として閾値を決めた．これを互いに行うことで2回分の講義を用いた．信頼度の閾値は両方とも0.3となった．

人手(ノートテイカ)による文選択の効果を調べるために，認識結果に含まれる内容語の正解率の高いものを，40%から90%まで10%おきに選択した．

結果を表3，表4に示す．

人手書き起こし(Trans)を用いることにより，ベースライン(Base)から単語認識精度が7.64%，話題語の検出精度のF値が20.06%改善した．これは，同一講師の以前の講義の書き起こしデータが言語モデルの適応に大きな効果を持つことを示している．一方，音声認識結果をそのまま用いた場合(Recog)はベースライン(Base)より，単語認識精度が2.30%，話題語の検出精度がF値4.91%改善したが，人手書き起こし(Trans)とは大きな差がある．これは，音声認識結果では，十分な効果が得られず，信頼できる部分を選別する必要性を示唆している．

認識の信頼度による選択(Cmscore40)と人手による選択(Oracle50)を導入することにより，話題語の検出精度のF値において，より大きな改善が得られ，一定の効果が確かめられた．しかし，人手による選択(Oracle50)の方が，認識の信頼度による選択(Cmscore40)と比較して改善幅がかなり大きいことから、人手による文選択の効果が大きいといえる．人手による文選択を50%行うときの内容語の正解率は50%であった．

4.4 各手法の統合

PLSAによる適応，Webテキスト収集による適応，以前の講義の認識結果を利用する適応を統合した．具体的には，ベースライン言語モデルにPLSAによる適応を行い，Webテキストから構築した言語モデル，以前の講義の認識結果から構築した言語モデルを順次混合することで適応を行った．結果を表5，表6に示す．

いずれの場合もPLSA及びWebテキスト収集と組み合せることで，話題語の検出精度のF値が改善した．よって，PLSA，Webテキスト収集による適応と，以前の講義の書き起こし・認識結果を利用する適応を組み合せる効果が示された．しかし，話題語の検出精度のF値において，人手による選択(Oracle50)と人手による選択(Oracle50)+PLSA+Webでは0.22%の改善にとど

表 3: 以前の講義データを用いた場合の単語認識精度

	単語認識精度
ベースライン (Base)	56.27
認識結果 (Recog)	58.57
人手書き起こし (Trans)	63.91
信頼度による選択 (Cmscore)40%	57.70
信頼度による選択 (Cmscore)50%	57.80
信頼度による選択 (Cmscore)60&	57.69
信頼度による選択 (Cmscore)70%	57.74
信頼度による選択 (Cmscore)80%	57.66
信頼度による選択 (Cmscore)90%	57.73
人手による選択 (Oracle)40%	59.26
人手による選択 (Oracle)50%	59.21
人手による選択 (Oracle)60%	58.63
人手による選択 (Oracle)70%	58.42
人手による選択 (Oracle)80%	58.28
人手による選択 (Oracle)90%	57.69

表 4: 以前の講義データを用いた場合の話題語の検出精度

	再現率	適合率	F 値
ベースライン (Base)	47.13	95.23	63.05
認識結果 (Recog)	53.04	94.57	67.96
人手書き起こし (Trans)	73.69	95.30	83.11
信頼度による選択 (Cmscore)40%	53.21	97.73	68.90
信頼度による選択 (Cmscore)50%	52.31	97.89	68.18
信頼度による選択 (Cmscore)60%	51.07	97.22	66.96
信頼度による選択 (Cmscore)70%	51.74	97.87	67.70
信頼度による選択 (Cmscore)80%	50.84	97.62	66.86
信頼度による選択 (Cmscore)90%	51.07	97.63	67.06
人手による選択 (Oracle)40%	57.14	96.21	70.70
人手による選択 (Oracle)50%	57.59	97.52	72.42
人手による選択 (Oracle)60%	55.46	97.62	70.73
人手による選択 (Oracle)70%	54.11	97.57	69.61
人手による選択 (Oracle)80%	52.98	97.52	68.67
人手による選択 (Oracle)90%	52.53	97.49	68.27

表 5: 各手法の統合による単語認識精度

	単語認識精度
Base	56.27
Recog	58.57
Trans	63.91
Cmscore50	57.80
Oracle50	58.20
Base+PLSA+Web	58.11
Recog+PLSA+Web	57.98
Trans+PLSA+Web	63.52
Cmscore50+PLSA+Web	57.46
Oracle50+PLSA+Web	58.45

表 6: 各手法の統合による話題語の検出精度

	再現率	適合率	F 値
Base	47.13	95.23	63.05
Recog	53.04	94.57	67.96
Trans	73.69	95.30	83.11
Cmscore50	52.30	97.89	68.18
Oracle50	58.27	97.92	73.06
Base+PLSA+Web	62.43	96.02	75.66
Recog+PLSA+Web	58.49	96.65	72.88
Trans+PLSA+Web	75.70	93.60	83.71
Cmscore50+PLSA+Web	57.03	94.41	71.11
Oracle50+PLSA+Web	59.84	94.49	73.28

まっており，元の認識結果の質が向上するにつれて PLSA，Web テキスト収集による適応の効果は小さくなっている．

最後に，認識精度の低かった文を人手で書き起こす場合を想定して，人手による選択 (Oracle50) の言語モデル構築に使われなかった文を書き起こし，それを含めて言語モデルを構築した．その結果，単語認識精度において 63.04%，話題語の検出精度において 80.81%となり，人手書き起こしの場合 (Trans) にほぼ近づいた．これは，効率的に言語モデルを改善する方法といえる．

このことから，書き起こしが全くない，もしくは認識結果が著しく悪い場合は，PLSA 及び Web テキスト収集による適応を適用し，書き起こしが増えてから，これを選別して利用する枠組みが考えられる．

5 まとめ

本稿では，大学講義における情報保障の一環として，音声認識技術をノートテイク支援に用いることを目的とし，言語モデル適応の手法を検討した．

今後は，本手法による言語モデル適応を組み込んだノートテイク支援システムを開発し，評価していきたいと考えている．

参考文献

[1] Tatsuya Kawahara,Yusuke Nemoto and Yuya Akita: “Automatic Lecture Transcription by Exploiting Presentation Slide Information for Language Model Adaptation”,*Proc. ICASSP*,pp4929-4932(2008).

[2] A.Park, T.Hazen and J.Glass: “Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Laguage Modeling”, *Proc. ICASSP*, Vol.1,pp497-500(2005).

[3] I.Trancoso, R.Nunes, L.Neves, C.Viana, H.Moniz, D.Caseiro and A.I.Mata. Recognition of Classroom Lectures in European Portuguese. *Proc. Interspeech*, 2006.

[4] T.Hoffman: “Probabilistic Latent Semantic Indexing”,*Proc. SIG-IR*, (1999).

[5] 長友健太郎，西村竜一，小松久美子，黒田由香，李晃伸，猿渡洋，鹿野清宏:“相補的バックオフを用いた言語モデル融合ツールの構築”，情報処理学会論文誌 Vol. 43. 7, pp2098-2107(2002).

[6] Marcello Federico, Nicola Bertoldi: “Broadcast news LM adaptation over time”, *Computer Speech and Language* 18. pp417-435(2004).

[7] 渡邉友裕，西崎博光，宇津呂武仁，中川聖一: “講演音声における認識結果の高信頼度部分の抽出とそれを用いた教師なし話者適応”，情報処理学会研究報告 2003-SLP-49-2(2003).

[8] http://www.goo.ne.jp/

[9] http://iptalk.hp.infoseek.co.jp/