社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－179 (16)
2007－SLP－ 66 (16)
2007／5／25

スライド情報を用いた言語モデル適応による講義の音声認識と字幕付与

根本 雄介　河原 達也　秋田 祐哉
京都大学 情報学研究科 知能情報学専攻
〒606-8501 京都市左京区吉田本町

あらまし　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述べる。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 モデルのスケーリングを行う。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見られ、大学の講義でも 80%に近い精度を実現した。これに基づいて、講義に字幕を付与する試みを行った。

Automatic Lecture Transcription by Exploiting Slide Information for Language Model Adaptation

Yusuke Nemoto　Yuya Akita　Tatsuya Kawahara
School of Informatics, Kyoto University, Kyoto 606-8501, Japan

Abstract　We investigate several language model adaptation methods which exploits presentation slide information for automatic lecture transcription. First, N-gram probabilities are re-scaled with lecture-dependent unigram probabilities estimated by PLSA 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 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][2][3][4]。

高精度な音声認識を実現するためには、認識対象音声の言語的特徴をよく反映した言語モデルが必要となる。講義音声の場合、話題に依存した専門用語を多数含み、話し言葉のスタイルをもつという特徴がある。したがって、これに適合した大量のテキスト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。しかし、実際にはこのようなテキストは容易に得られないため、類似のコーパスを用いてモデルを構成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題を包含するので、特定の講義の話題に関する予測能力は低下する。

この問題に対して、講義で使用された教科書や講義の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が提案されている [5][6]。このようなテキストが得られることは限定的であるため、講義で使用されるスライドを利用して言語モデルの補間を行う手法も提案されている [7]。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、単純な補間に基づく手法では効果は限られている。

そこで本稿では、講義スライドを効果的に利用し、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その効果的な適応を実現するために、PLSA (Probabilistic Latent Semantic Analysis)[8] の枠組みによる N-gram 確率の推定 [9][10] を行う。さらに、スライドの記述に沿って講義の内容が移り変わっていくため、キャッシュモデル [11] も導入する。本研究では、講義スライド全体の情報を用いて PLSA による言語モデルの適応を行い、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。

2 スライド情報を利用した言語モデルの適応

2.1 PLSA に基づく言語モデルの適応

PLSA は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。

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

PLSA では、大規模な文書コーパスを用いて、文書の特徴 (例えば話題) を表す N 次元部分空間 $\{t_j\}$ を EM アルゴリズムによりあらかじめ構築する。この部分空間に文書 d を射影することにより、文書に依存した単語 w の生起確率 $P(w|d)$ を求める。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、PLSA では潜在意味空間を介して文書 d に存在しない単語の確率も推定されるため、出現単語が限定される講義スライドにおいても有効と期待される。

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抽出したテキストを 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。

こうして $P(w|S_{all})$ が推定されるが、3-gram 確率に対する適用は膨大な計算量が必要となり、現実的でない。そこでベースラインの 3-gram 確率に対して式 (2) による unigram スケーリング [12][13][14] を行い、3-gram 言語モデルの適応を行う。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。

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

なお、スライドの話題に対してのみ適応を行うため、話題語と考えられる名詞、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、$P(w|S_{all})$ を推定し、その他の汎用語や機能語に対しては $P(w|S_{all})$ としてベースライン言語モデルによる確率を用いた。

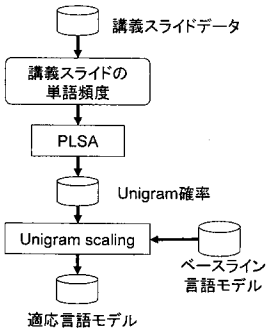

図 1: PLSA に基づく言語モデル適応

2.2 キャッシュモデルに基づく言語モデルの適応

キャッシュモデルでは、単語 w_nの直前の単語履歴をキャッシュ$H = \{w_{n-|H|}, \cdots, w_{n-1}\}$ 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_nの出現確率 $P_c(w_n|H)$ 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。

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

本研究では、キャッシュモデルの枠組みに基づき、講義スライド中の単語の出現頻度情報を用いることで単語の生起確率を推定する。単語 w_nが含まれる発話に対応するスライド S における単語の出現頻度を考慮した単語 w_nの生起確率を式 (4) により定義する。ただし、$|S|$ はスライド S に含まれる総単語数とする。

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

さらに、単語 w_nの発話中の単語履歴 H と、対応するスライド S を結合した単語のセットを $H \cup S$ とするとき、式 (5) により単語履歴 H とスライド S のもとでの単語 w_nの生起確率 $P_{cs}(w_n|H,S)$ を定義する。

$$P_{cs}(w_n|H,S) = \frac{1}{|H|+|S|} \sum_{w_x \in H \cup S} \delta(w_n, w_x) \qquad (5)$$

単語 w_nの生起確率 $P_c(w_n|H)$、$P_s(w_n|H)$、$P_{cs}(w_n|H,S)$ のいずれか一つを N-gram 言語モデルによる

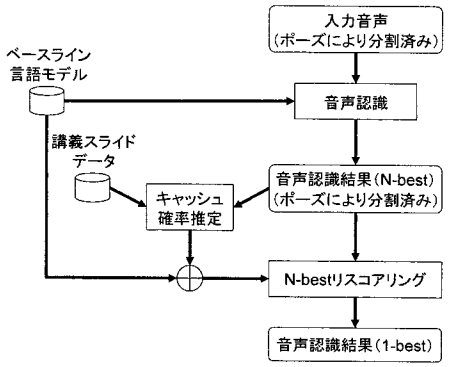

図 2: キャッシュモデルに基づく言語モデル適応

確率と線形補間することで、スライドや単語履歴、もしくはその両方を用いて適応した単語 w_nの生起確率が得られる。

キャッシュモデルによる適応の流れを図 2 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、スライド情報を用いて式 (4)(5) による生起確率の推定を行い、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 による適応後の言語モデルを用いることで、講義スライド全体の情報と発話に対応するスライドの局所的情報を組み合せた適応が実現される。

講義音声認識のためのキャッシュモデルによる適応は富樫ら [15] によっても検討されているが、本研究でいう大域的な適応を行っており、効果が限定されている。本研究では、大域的な適応にはより頑健な適応手法である PLSA を用い、キャッシュモデルは局所的な適応に用いるものとする。

3 評価実験

3.1 実験データと条件

本研究では、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における 12 回分の講義音声である。大学講義の音声は、京都大学の大学院および学部向けに行われた 3 科目 (画像処理論、パターン認識特論、パターン認識) の

各1回分である。これらはすべて異なる科目の講義であり、時間は全て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。また、これらにおいては、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。

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SAT学習した3000状態・64混合の状態共有triphone HMMに対して、教師なしMLLR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・模擬講演2720講演(単語数7M)から学習した語彙サイズ50Kの3-gramモデルである。

3.2 音声認識実験結果

表1にベースライン言語モデルおよび各手法により適応を行った言語モデルにおける単語認識精度(word accuracy)を示す。ベースライン言語モデルによる講習会音声の単語認識精度は71.60%、大学講義音声の単語認識精度は58.61%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%の単語正解精度の改善(絶対値)が得られた。以後、適応言語モデルによる音声認識を行う際には、スライドから未登録語を追加した単語辞書を使用する。

3.2.1 PLSA

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学会講演により学習した。部分空間の次元数は予備実験においてテストセットパープレキシティの値が最適化された100とした。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適応[13]も行った。

スライドによる適応では、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。音声認識結果を用いた場合の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、スライドによる適応では音声認識を1回だけ行えばよいので、認識結果を使用する場合に比べて高速な処理が可能である。

3.2.2 関連Webテキストの収集による言語モデル適応

提案手法に加えて、当該講義に関連するWebテキストを収集して、言語モデルを構築する方法も評価した。まず、講義で使用されたスライドからこれを特徴づける語句を選択し、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、検索クエリとする。生成されたクエリごとにAND検索を行い、Webテキストを収集する。このとき、収集ページ数の上限を1クエリあたり500件とした。

Webテキストの収集には、特定領域研究「情報爆発」において開発が進められている検索エンジンTsubaki[1]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[16]が、その際の閾値を変化させて実験を行なった。収集したテキストによるN-gram頻度とベースライン言語モデルのN-gram頻度を重み付き混合する際の混合重みは予備実験により0.1と定めた。

音声認識の結果、収集テキストサイズが50M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%)が見られたが、CSJにより講習会の内容のかなりの部分がカバーされているのに対して、大学講義の内容がカバーされていないことがこの要因として考えられる。

3.2.3 キャッシュモデル

キャッシュモデルに基づく言語モデルの適応に必要なスライドと発話の対応関係は、講義収録時に記録されたスライドの切り替え時間情報に基づいて与えた。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。

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%単語認識精度が向上した。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)の使用による講習会で0.64%、大学講義で1.02%の向上を上回った。さらに、キャッシュをスライドと併用する

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

表 1: 各手法による単語認識精度

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 (Web, 20M)	72.37	60.50	○
テキスト混合 (Web, 50M)	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)

表 2: 各手法によるキーワード検出精度

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 (Web, 20M)	85.78	75.09
テキスト混合 (Web, 50M)	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)

ことで講習会、大学講義でそれぞれ 1.06%、1.69%単語認識精度が改善された。発話に対応するスライド中の単語や直前に出現した単語といった局所的情報による適応の効果が確認された。

3.2.4 PLSA とキャッシュモデルの統合

スライド全体を使用して PLSA による言語モデルの適応を行った結果 (表 1 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 1 の下段に示す。スライド全体を用いた 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行うことで、講習会、大学講義のそれぞれにおいて 1.51%、2.36%単語認識精度が改善した。PLSA による適応と、スライドやキャッシュによる適応の効果がほぼ加算的に現れており、講義全体の情報と発話周辺の局所的な情報の組み合わせによる適応が有効であることが示された。

3.3 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上で重要であり、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。その検出精度を F 値で表 2 に示す。

ベースラインと比べて、講習会において 3.24%、大学講義において 7.82%の改善が得られた。これは前節の単語認識精度の改善幅より著しく大きく、提案する適応手法が、話題語の認識において特に有効であることが示された。大学の講義の方が改善幅が大きいのは、技術講習会の話題が 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。

なお、すべてを統合した場合 (表 2 の最下段) に、講習会では、再現率 84.12%、適合率 92.78%、大学の講義では、再現率 71.71%、適合率 86.96% となっており、特に再現率で大きな改善が得られている。

4 字幕の付与

大学の講義についても、近年ユニバーサルアクセスの観点から、聴覚障害のある学生のために字幕に近いメモをリアルタイムでとったり (=ノートテイク)、講義のアーカイブに字幕を付与する必要性が高まっている。これらは、放送コンテンツほど完璧な書き起こしでなくても許容されるので、音声認識技術の利用の可能性があると考えられる。

そこで、音声認識結果に対して、文節単位の抽出と節境界・文境界の検出 [17][18] を行い、フィラーを除去した上で、字幕の作成を行った。

字幕は、1 行 16 文字程度・最大 5 行程度を想定し、文境界で画面をクリアし、文境界以外の節境界 (弱境界) では改行するが、おさまる範囲で同じ画面に表示し、それ以外の改行は文節境界で行うようにした。

その例を図 3 に示す。今後、評価を行っていきたいと考えている。

音声 認識 を 実現 する 為 に は
色んな 技術 が 入っ て い ます

で 研究 分野 各 音 親 も 非常 に
幅広く て

言語 音 を 使う と いう こと で
音声 学 や 音韻 学
それから
言語 学 が 含ま れ て い ます し

工学 的 に は 分野 的 に は 電気 電子
なり ます けども

デジタル 信号 処理 で A D 変換 と か です ね
デジタル 信号 処理 そういう もの を
フーリエ 変換 それで 変換 と
それ で

で 更に その パターン を 扱う と いう
で パターン 認識 で

(空行箇所で画面をクリア)

図 3: 音声認識結果から自動生成された字幕の例

5 おわりに

講義で使用されたスライド情報を用いて、PLSA とキャッシュモデルに基づいて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ところ、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、認識精度の有意な改善が得られ、特にキーワードに限定すると、大学の講義で 8 ポイントも向上した。

今後、京都大学学術情報メディアセンターで構築されている大規模な講義アーカイブに適用して、認識精度の改善を図るとともに、字幕の付与についても一層の検討と評価を進めていく予定である。

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. 講義音声認識における収録装置とケプストラム正規化法の検討. 情報処理学会研究報告, 2006-SLP-64-38, 2006.

[4] J.Glass, T.J.Hazen, L.Hetherington and C.Wang. Analysis and Processing of Lecture Audio Data: Preliminary Investigations. In Proc. HLT-NAACL, 2004.

[5] A.Park, T.Hazen and J.Glass. Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling. In Proc. ICASSP, 2005.

[6] 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.

[7] 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.

[8] T.Hoffman. Probabilistic Latent Semantic Indexing. In Proc. SIG-IR, 1999.

[9] T.Misu and T.Kawahara. A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts. In Proc. Interspeech, 2006.

[10] M.Suzuki, Y.Kajiura, A.Ito and S.Makino. Unsupervised Language Model Adaptation Based on Automatic Text Collection from WWW. In Proc. Interspeech, 2006.

[11] R.Kuhn and R.De Mori. A Cache-based Natural Language Model for Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(6), pp. 570-583, 1990.

[12] D.Gildea and T.Hoffman. Topic-based Language Models using EM. In Proc. Eurospeech, 2003.

[13] Y.Akita and T.Kawahara. 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] 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.

[16] 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対話システム用言語モデルの構築. 電子情報通信学会技術研究報告, SP2006-126, NLC2006-70 (SLP-64-42), 2006.

[17] Y.Akita, M.Saikou, H.Nanjo, and T.Kawahara. Sentence boundary detection of spontaneous Japanese using statistical language model and support vector machines. In *Proc. Interspeech*, pp. 1033–1036, 2006.

[18] 西光雅弘, 河原達也, 高梨克也. 隣接文節間の係り受け情報に着目した話し言葉のチャンキングの評価. 情報処理学会研究報告, SLP-61-4, 2006.