# 講義音声認識のためのスライド情報を用いた言語モデル適応

根本 雄介† 秋田 祐哉† 河原 達也†

† 京都大学 情報学研究科 知能情報学専攻

e-mail: nemoto@ar.media.kyoto-u.ac.jp

## 1 はじめに

近年，講演や講義などの音声・動画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字幕を付与することが考えられるが，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，音声認識の活用が検討されている．

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は得られないため，類似のコーパスなどを用いている．ただし，このようなモデルは多数の話題を含むので，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 [1]．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 [2]．ただし，教科書や書き起こしのようなテキストとは異なり，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，PLSA（Probabilistic Latent Semantic Analysis）[3] による N-gram 確率の推定および Web テキストの収集による補間 [4] を検討する．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 [5] を導入する．本研究では，講義スライド全体の情報を用いて，PLSA と Web テキストの収集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．

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

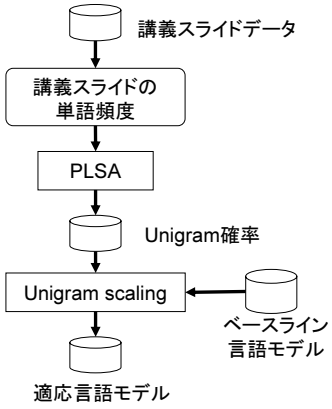
PLSA では，大規模な文書コーパスを用いて，文書の特徴（例えば話題）を表す $N$ 次元部分空間 $\{t_j\}$ を EM アルゴリズムによりあらかじめ構築する．この部分空間に文書 $d$ を射影することにより，文書に依存した単語 $w$ の生起確率 $P(w|d)$ を得ることができる．単語頻度に基づく射影であるため，短いフレーズを中心に記述されている講義スライドでも有効であると期待される．また，PLSA では式 (1) に基づく推定により文書中に存在しない単語確率の推定も行われるため，出現単語が限定される講義スライドにおいても有効と期待される．

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．

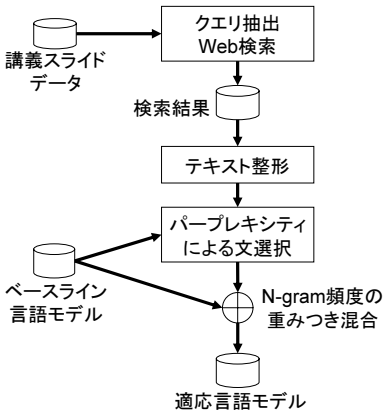
こうして $P(w|S_{all})$ が推定されるが，3-gram 確率に対する適用は莫大な計算量が必要となり，現実的でない．そこでベースラインの 3-gram 確率に対して式 (2) による unigram スケーリング [6] を行い，3-gram 言語モデルの適応を行う．上記のスライド情報を用いた言語モデル適応の流れを図 1(a) に示す．

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

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．




(a) PLSA に基づく言語モデル適応




(b) Web テキストを用いた言語モデル適応

図 1: 講義スライド 全体を利用した言語モデル適応

### 2.2 Web テキストを使用した言語モデルの適応

Web テキストの収集による言語モデル適応の概要を図 1(b) に示す．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 件とする．

収集された Web テキストには，言語モデルの学習に適さないスタイルの文が含まれるので，適応に使用する文を選択する．選択に先立って，タグや記号，1 文が一定の長さより短いものやアルファベットの数が一定の比率を超える文を除去する．整形された Web 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，この値が閾値より小さい文を選択する．

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学習に用いたテキストの N-gram の頻度に重みつきで混合することで適応言語モデルを構築する．

### 2.3 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，$\delta$ 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手法を提案する．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言語モデルの適応が実現される．

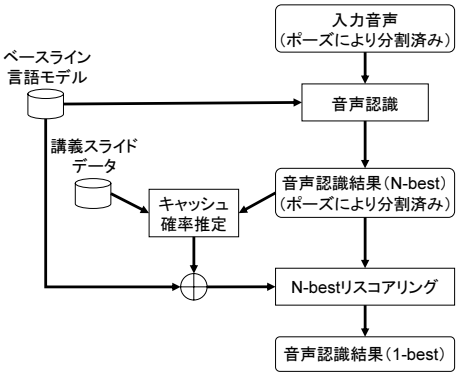



図 2: キャッシュモデルに基づく言語モデル適応

# 3 評価実験

## 3.1 実験条件

本研究では，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, 2005 年に京都大学学術情報メディアセンターにて行われた音声認識・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講義 3 回分（画像処理論，パターン認識特論，パターン認識）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・大学講義ともに講師の重複は 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．

また，Web 検索のクエリ生成時に使用する $tf{\cdot}idf$ 値の計算には，$tf$ として各スライドにおける単語頻度を，$idf$ として上記の 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．

## 3.2 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と人手による書き起こしを用いた適応も行った．

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．

なお，比較のため認識結果の N-gram 頻度を言語モデル学習テキストに混合して作成した言語モデルにより音声認識を行ったが，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，上記の手法には及ばなかった．

### 3.2.2 Web テキスト収集

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 を使用した．収集したテキストから文を選択する際の閾値を変化させて実験を行なった．なお，収集したテキストによる N-gram 頻度の混合重みは予備実験により 0.1 と定め，重みつき頻度の端数は切り上げた．

音声認識を行ったところ，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，スライドを用いて PLSA による適応を行った場合を上回る結果となった．講習会と大学講義の間で適応の結果に 1.5%の差が生じたが，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

表 1: 各適応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 | × |
| PLSA(書き起こし) | 73.22 | 61.15 | - |
| テキスト混合 (認識結果) | 71.72 | 59.09 | × |
| 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**(認識結果) | 72.24 | 59.63 | △ |
| キャッシュ(**スライド**) | 72.44 | 60.11 | △ |
| スライド+キャッシュ | 72.66 | 60.30 | △ |
| PLSA+キャッシュ(認識結果) | 72.80 | 60.42 | △ |
| PLSA+キャッシュ(スライド) | 72.98 | 60.68 | △ |
| 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）

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 各手法の統合

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 による言語モデルの適応と，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．

講義ごとの実験結果を図 3，4 に示す．講義によって言語モデルの適応に伴う認識精度の向上にばらつきが見られた．多くの講義で PLSA とキャッシュを併用した場合に加算的な効果が得られたが，講義 K, Y, Z では PLSA による適応の効果が見られなかった．これはスライドの中で数式が占める割合が大きかったり，関連の薄い例文を使用した説明が頻出したためと考えられる．この影響で，講義 K, Y においては PLSA とキャッシュを併用した場合にキャッシュのみ使用した場合を下回る結果となった．また，講義 F, I, K のように，Web 検索により不要なテキストも収集されてしまい，認識精度が改善されない場合も存在した．

## 4 おわりに

本研究では，講義で使用されたスライド情報をもとに PLSA，Web テキスト収集，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

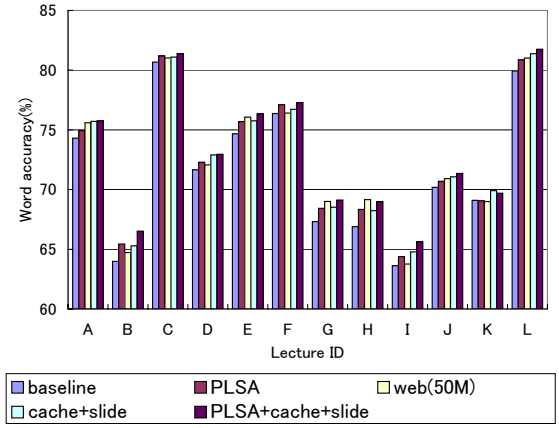



図 3: 各講義の単語認識精度（講習会）

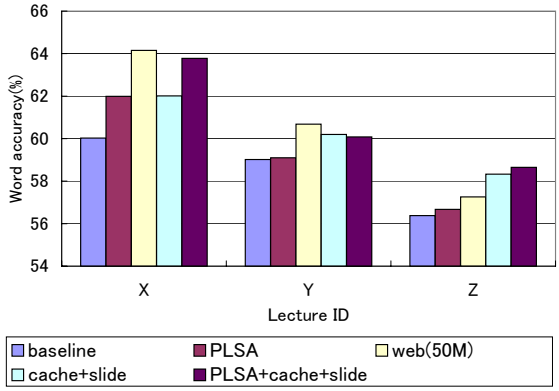



図 4: 各講義の単語認識精度（大学講義）

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と京都大学で行われた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，講習会において 1.5%，大学講義において 2.4%単語認識精度が改善した．

## 参考文献


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

[2] 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.

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

[4] 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.

[5] 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

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