論文 / 著書情報
Article / Book Information

論題(和文)	スライド資料を用いた講義音声認識のための言語モデル適応
Title(English)	
著者(和文)	山崎 裕紀, 岩野 公司, 篠田 浩一, 古井貞熙, 横田 治夫
Authors(English)	Koji Iwano, Koichi Shinoda, SADAOKI FURUI, Haruo Yokota
出典(和文)	日本音響学会2007年春季講演論文集, Vol. , 3-9-8, pp. 79-80
Citation(English)	, Vol. , 3-9-8, pp. 79-80
発行日 / Pub. date	2007, 3

スライド資料を用いた講義音声認識のための言語モデル適応*

◎山崎 裕紀, 岩野 公司, 篠田 浩一, 古井 貞熙, 横田 治夫 (東工大)

1 はじめに

大学などで行われる講義の音声・映像は, 学術的に有用な知識資源であり, e-Learning システムへの応用も期待されている. 講義における音声情報を効果的に利用するために, 話し言葉の音声認識のより一層の性能向上が求められている. 本論文では, 講義で用いられたスライドの情報を用いた動的な言語モデル適応手法を提案する. スライドなどの補助資料を用いたモデル適応の手法はこれまでにいくつか提案されている (e.g. [1]). 本手法では各スライドが実際の講義中に使用された時間の情報を利用することで, 各スライドに依存した言語モデルを構築する.

2 プレゼンテーション蓄積検索システム UPRISE

UPRISE (Unified Presentation Slide Retrieval by Impression Search Engine) [2] は聴講者の自主学習を支援することを目的とした講義のプレゼンテーションシステムである (Fig. 1). 講義で使用されたテキスト, 音声, 動画, スライドをはじめとした多種多様なマルチメディアコンテンツを格納し, 効果的な講義検索を可能とする. 検索の重み付けに音声情報を用いる手法が提案されているが [3], その適用に際し, 音声認識精度の一層の向上が望まれている.

3 動的言語モデル適応

講義中に用いられる各スライドは, それが使用されたときの発話の内容と深い関連をもつと考えられる. UPRISE では, 各スライドが講義中で使用された時刻の情報が記録されている. この情報をモデル適応に利用することができる.

言語モデルの適応は N グラムの頻度の重み付け加算によって行う. 適応手法のアルゴリズムは以下の 2 つのステップに分かれる. ステップ 1 として, 1 コース分の講義で使用されたスライドの全てを適応データとした, グローバルな適応を行なう. 各 N グラム N_i の頻度 $F(N_i)$ は, 適応により次式のように更新される.

$$F'(N_i) = F(N_i) + w_1 G(N_i), \quad (1)$$

ここで $F(N_i)$ は初期モデルの学習データにおける N グラム N_i の頻度. $G(N_i)$ は適応データにおける N グラム N_i の頻度, w_1 は重み係数である. 続いてステップ 2 では, ステップ 1 により適応された言語モデルを, さらに, 個別のスライドを用いてローカルに適応する. 適応モデルにおける各 N グラム N_i の頻度 $F''_j(N_i)$ は以下のように求められる.

$$F''_j(N_i) = F'(N_i) + w_2 H_j(N_i). \quad (2)$$

ここで $H_j(N_i)$ は N グラム N_i の j 番目のスライドにおける頻度, w_2 は重み係数である. ステップ 2 において, j 番目のスライドに対応する適応モデルは, 各 N グラムについて頻度 $F''_j(N_i)$ を用いることで構築される.

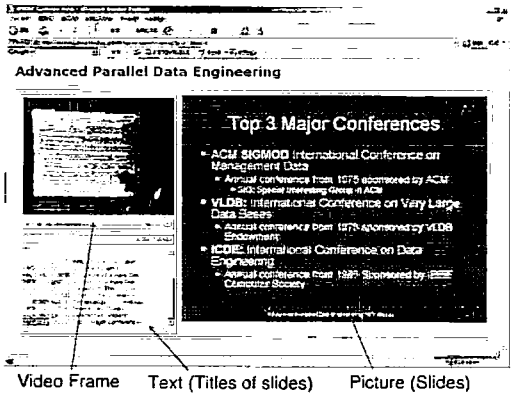

Fig. 1 UPRISE の画面イメージ.

4 認識実験

4.1 実験条件

講義データベースとして. 東京工業大学において日本語で行われた講義の音声と映像を収集した. それらは, 話者 A による 2 つの講義 (LEC1, LEC2), 話者 B による 1 つの講義 (LEC3), 話者 C による 1 つの講義 (LEC4) から成る. それぞれの講義は計 12 回開講され, 1 回ごとの講義の長さは 80 分前後である. ただし, 各回の講義のうち, 録音に問題があった回は実験から除外した. LEC1 及び LEC4 では 1 回分, LEC2 では 2 回分, LEC3 では 5 回分が上記の理由から除外された. 音声データはスライド時間情報を用いて区切った. これにより各発話の区切りが, 対応するスライドの区切りと同期する. 発話が 2 つのスライドにまたがって行われた場合は, 音声の区切りはその発話が終了した直後に設定した. また, 講義のキーワードとして, 講義に特徴的な専門用語を, 実験従事者の 1 人が主観で選択した. Table 1 に講義に関する統計を示す.

音響モデルは CSJ の学会講演 953 講演と模擬講演 1,543 講演から構築した. 音響特徴量として 12 次元の MFCC とその ΔMFCC 12 次元, および Δ パワーの計 25 次元を用いた. 各発話ごとに CMS 処理を行った. モデルとして left-to-right 型の 3 状態トライフォン HMM を用い, 3000 状態, 16 混合の状態共有モデルとした. また各回の講義の開始 10 分間の音声を用い. MLLR 法による教師なし適応を行った. 言語モデル適応の初期モデル (ベースライン言語モデル) は日本語話し言葉コーパス (CSJ) の学会講演 967

*Language model adaptation using presentation slides for lecture speech recognition. by Hiroki Yamazaki, Koji Iwano, Koichi Shinoda, Sadaoki Furui. and Haruo Yokota (Tokyo Institute of Technology)

Table 1　講義に関する統計量.

	LEC1	LEC2	LEC3	LEC4
講義時間 (分)	766.4	830.7	544.7	951.8
スライド使用時間 (分)	754.0	803.0	352.4	948.4
1 スライドの単語数 (平均)	52.1	60.2	52.5	13.0
1 スライドのキーワード数 (平均)	9.3	5.5	8.3	3.4

Table 2　ベースライン言語モデルによる音声認識とキーワード抽出の結果 (%).

	Word acc.	Recall	Precision	F 値
LEC1	39.2	37.2	59.1	45.7
LEC2	37.3	36.1	66.2	46.7
LEC3	57.7	56.4	82.4	67.0
LEC4	60.1	49.6	71.9	58.7
Avg.	47.4	45.0	69.1	54.5

講演 (3M 形態素) を用いて構築した. 認識語彙は学習データにおいて出現頻度の高い順から 25,000 語彙を選択した.

大語彙連続音声認識デコーダとして Julius を用いた. 音声認識の評価は単語正解精度で行った. また, キーワード抽出の評価を recall と precision を用いて行った. キーワード抽出の性能は, 講義の検索などにおける性能評価に役立つ.

4.2　実験結果

ベースライン言語モデルを用いた認識結果を Table 2 に示す. 単語正解精度は 47.4% であった. このときキーワード抽出の recall は 45.0%, precision は 69.1% であった. 次に講義スライドを用いた言語モデル適応手法の効果について検証を行った. 重み係数 w_1, w_2 の値はそれぞれ 20, 9000 とした. ステップ 1 で作成した言語モデルを用いたときの結果を Table 3 に示す. ステップ 1 の適応により, 単語正解精度においては平均して 3.0% の誤りが削減された. またキーワード抽出では recall において 19.0%, precision において 12.0% の誤りが削減された. これらの結果はステップ 1 によるグローバルな適応が音声認識とキーワード抽出の両方に効果があったことを示している. 続いてステップ 2 の評価結果を Table 4 に示す. 単語正解精度はステップ 1 の結果と大きく変わらなかったが, キーワード抽出に改善が見られ, F 値にして 2.4% の誤り率が削減された. recall に関しては 2.7% の誤り率が削減されたが, precision に関しては 1.1% の誤り率削減に留まった.

ステップ 2 の効果を講義ごとに結果を比較したとき, LEC1, LEC2 は F 値にして 3.8% の誤りが削減されているのに対し, LEC3, LEC4 においては 1.2% の削減に留まっている. Table 1 によると LEC3 は講義中にスライドが用いられている時間帯の割合が他の講義と比較し少ない. 本実験ではスライドが用いられていない部分の音声認識ではステップ 2 においてもステップ 1 と同じモデルを用いたため, 効果が薄れていると考えられる. また LEC4 は スライド 1 枚あたりの平均単語数及び平均キーワード数が少ない. そのため, 局所的な適応において効果が得にくいと考えられる.

Table 3　ステップ 1 による適応後の音声認識とキーワード抽出の結果 (%).

	Word acc.	Recall	Precision	F 値
LEC1	41.1	50.6	65.9	57.2
LEC2	38.7	49.8	71.6	58.7
LEC3	59.8	65.8	83.1	73.4
LEC4	61.4	57.1	74.4	64.6
Avg.	49.0	55.5	72.8	63.0

Table 4　ステップ 2 による適応後の音声認識とキーワード抽出の結果 (%).

	Word acc.	Recall	Precision	F 値
LEC1	41.1	52.5	66.3	58.6
LEC2	38.6	51.9	73.1	60.7
LEC3	59.4	66.3	83.3	73.8
LEC4	61.2	57.7	74.3	65.0
Avg.	48.8	56.7	73.1	63.9

5　おわりに

講義音声認識精度向上のためのモデル適応手法として, 発話に対応したスライドの言語情報を用いて作成した言語モデルを動的に用いる手法を提案した. 実験により効果が確認され, 特にキーワード抽出において効果が得られた.

現在は 4 つの講義によって実験を行っているが, 講義数が少なく, 十分信頼できる評価結果が得られていない. 今後より多くの講義データを収集し, 検討を行なう必要がある. また, 大学講義だけでなく, テレビ講義など他のコンテンツにも実験の対象を拡大していくことも今後の課題である.

謝辞　本研究は 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」による支援を受けた.

参考文献

[1] 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 2006 年春季講演論文集, 1-P-24, pp. 191-192, 2006.

[2] H. Yokota, et al., T. Kobayashi, T. Muraki and S. Naoi, "UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine," IEICE Transactions on Information and Systems, vol. E87-D, no. 2, pp. 307-406, 2004.

[3] 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, "プレゼンテーション蓄積検索システムにおける講義・講演音声情報を利用した適合度の改善," 第 17 回電子情報通信学会データ工学ワークショップ (DEWS2006) 論文集, 6c-o1, 2006.