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

2006－SLP－64 (38)
2006／12／22

講義音声認識における講義スライド情報の利用

山崎 裕紀† 岩野 公司† 篠田 浩一† 古井 貞熙† 横田 治夫†

† 東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail: †yamazaki@ks.cs.titech.ac.jp, ††{iwano,shinoda,furui,yokota}@cs.titech.ac.jp

あらまし 大学などで行なわれる講義に対する音声認識において, 講義中に使用されたスライド資料を用い, 言語モデルを動的に適応する手法を提案する. 認識音声に対応するスライドから抽出した言語情報を適応データとして用いることで適応言語モデルを作成し, 認識に用いる. 大学で開講された講義を対象として認識性能の評価を行ない, 手法の効果を確認した. 講義 1 コース分のスライド全てをグローバルに適応に用いることで, 単語誤り率が 3.1% 削減された. また, キーワード抽出においても性能の改善が見られ, recall にして 21.5% の誤りが削減され, precision にして 13.8% の誤りが削減された. さらに各講義スライドをローカルに適応に用いることで, グローバルな適応のみの結果に対し改善が見られた. 特にキーワード抽出に対して効果が見られ, recall にして 3.1% の誤りが削減され, precision にして 1.4% の誤りが削減された.

キーワード 言語モデル適応, 音声認識, 講義音声

Using presentation slide information for lecture speech recognition

Hiroki YAMAZAKI†, Koji IWANO†, Koichi SHINODA†,
Sadaoki FURUI†, and Haruo YOKOTA†

† Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: †yamazaki@ks.cs.titech.ac.jp, ††{iwano,shinoda,furui,yokota}@cs.titech.ac.jp

Abstract We propose a dynamic language model adaptation method for lecture speech recognition in which the information of text on slides for lectures is used. The speech data corresponding to each slide are recognized with a language model adapted to them by using the slide texts as adaptation data. We evaluated the proposed method by using the speech data of three classroom courses in Japanese, and confirmed its effectiveness. The average speech recognition error was reduced by 3.1% by the global adaptation using all slides used in a cource. The error rates of recall and precision for keywords were also reduced by 21.5% and 13.8% respectively. Furthermore, we achieved the improvement of keyword detection performance by the adaptation using each slide locally. The error rates of recall and precision for keywords were reduced by 3.1% and 1.4% respectively from global adaptation.

Key words Language model adaptation, speech recognition, classroom lecture speech.

1. はじめに

近年の計算機性能と記憶容量の向上および通信回線の広帯域化によって, 大量の音声・映像をディジタルアーカイブとして蓄積・再生することが可能になりつつある. 特に, 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, かつ e-Learning システムへの利用なども期待されている.

これまでに, 講義マルチメディアコンテンツの利用を目的とした多くの研究がなされている [1] [4] [5] [6] [7]. 講義における様々なコンテンツの中で, 講義音声は重要な情報源であり, その書き起こしは講義のインデキシングや検索において有用であると考えられる [3] [4]. 講義音声情

報を効果的に利用するためには, 話し言葉の音声認識技術を用いた精度の高い自動書き起こしが不可欠である.

講義音声の認識は現在盛んに研究されている. Trancoso ら [8] は, ポルトガル語の講義に対し音声認識の検討を行った. 講義の自動書き起こしに関しては多くの研究プロジェクトが存在する. 欧米では CHIL (Computers In the Human Interaction Loop) [9] や, the American iCampus Spoken Lecture Processing project [10] が挙げられる. また, 学会発表を中心とした大規模なデータベースとして, 日本語話し言葉コーパス (CSJ) [11] [12] や, TED コーパス [14] などが構築されたことにより, 話し言葉音声の認識技術が向上してきている. これらのコーパスにより, 学会講演の認識精度は 70~80% に達し, 認識結果を音声要約やインデキシングに用いる研究もなされている [13].

本論文では, 大学などにおける講義の音声認識に焦点を当てる. 講義は講演同様, 一人の話者による発話 (モノローグ) が基本であるが, 発話スタイルは CSJ や TED などにおける学会講演のそれと異なる部分が多いと考えられる. 講義は講演と異なり, 発表練習などの十分な事前準備が常になされているとは限らない. また内容を強調するために同じフレーズが繰り返されるなど, 冗長な部分がより多く見られる. 聴講生からの質問などにより講義が中断されることもしばしばあり, 発話のスタイルは対話音声に近い部分もあると考えられる. このようなスタイルの発声は自発性が高く, 強い調音結合, 非文法的構造, 言い直しや言いよどみなどが多く見られる. これらの理由から, 講義音声の認識は一般に難しく, 学会講演音声の認識よりも性能が低い. さらに, 大学講義を中心とし, 音声認識のためのモデル構築に利用できるよう整備された大規模データベースは現状では存在しない.

講義においては, 聴講者の理解を助けるため, 教科書やスライドをはじめとした様々な補助資料が使われる. これらの資料は講義における重要なキーワードを多く含み, 音声にもそれらのキーワードが多く現れると考えられる. よってこれらの資料は言語モデルの適応において有効であることが期待される. 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. 宮樫ら [15] は発表スライドのテキスト情報を用いた手法を提案している. Cettolo ら [16] は学会の予稿集のテキスト情報を用い, 学会講演音声の認識に関して研究を行った. 何らかの講義資料からテキスト情報を抽出するとき, 講義で用いられる教科書を対象とすると, それが必ずしも電子化されているとは限らず, 利用が難しい. 一方 PowerPoint などを用いて作成された講義スライドからはテキスト情報を簡単に抽出することができる. よってスライド資料を用いた言語モデル適応は講義音声認識において広く適用できると考えられる.

本論文では, スライド情報を用いた動的な言語モデル適応について提案する. 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. 各スライドに依存した言語モデルを構築し, スライドに対応する音声の認識に用いる. 講義の進行に合わせ, 言語モデルを動的に変更することで認識率の向上を図る.

本論文は以下のように構成される. 2 節では研究におけるベースラインシステムの構築について述べる. 3 節では提案する言語モデル適応手法について述べる. 4 節では認識実験を行い, 手法の評価を行う.

2. UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine

UPRISE (Unified Presentation Slide Retrieval by Impression Search Engine) [1] [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. 講義で使用されたテキスト, 音声, 動画やスライドなど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することが可能であり, 講義検索システムとして統合的に利用者に提示することができる. UPRISE は, 利用者が与えたキーワードに対し, 適合する講義のシーンを提示する. UPRISE における音声情報の利用については, 音声情報を検索の重み付けに用いた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強く望まれている.

UPRISE に蓄積された全ての講義コンテンツは講義の進行に対応して格納されている. このコンテンツ間の対応を利用し, 認識音声と強い関連のあるコンテンツを利用することで時間軸に沿ったモデル適応を実現した. 次節では, このようなコンテンツの一つである, 講義スライドを用いた言語モデル適応手法を提案する.

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持つと考えられる. UPRISE は, 各スライドが講義中のどの部分で使用されたかを撮影された映像から自動で検出し, 記録している [17]. これによりスライドが講義中で用いられた時間の情報をモデ

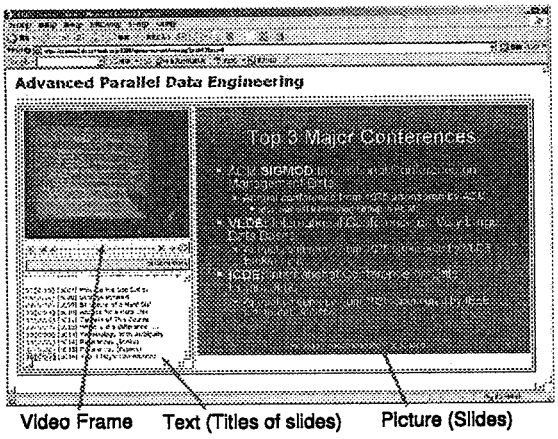

図 1 UPRISE の概観.

ル適応に利用することができる.

本節ではこの時間情報を利用した言語モデルの適応手法を提案する. スライドを用いることで, 講義に特徴的な専門用語などに対し適応することが可能であると考えられる. 言語モデルの適応は, スライドの言語情報から抽出した N グラムの頻度を, 初期モデルの学習データから抽出した N グラムの頻度に重み付けして加えることで行った. 適応言語モデルの語彙は元の学習データの語彙と適応データの語彙を合わせたものとした.

適応手法のアルゴリズムの詳細は以下の通りである. アルゴリズムは 3 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.

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

$F(N_i)$ は初期モデルの学習データにおける N グラム N_i の頻度である. $G(N_i)$ は適応データにおける N グラム N_i の頻度である. w_1 は重み係数であり, 実験により最適化した. 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.

$$F''_j(N_i) = F'(N_i) + w_2 H_j(N_i), \qquad (2)$$

$H_j(N_i)$ は N グラム N_i の j 番目のスライドにおける頻度である. 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. さらにステップ 3 では, j 番目のスライドの直前に使用された k 枚のスライドの情報をローカルに用い, 適応を行なう. この手法は, ある時点での講義の内容が, 講義のより前の部分の内容を含んでいるという予想に基づいている. 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.

$$F'''_j(N_i) = F''_j(N_i) + \sum_{l=1}^{k} w_2 \alpha^l H_{j-l}(N_i), \qquad (3)$$

α^l は j 番目のスライドの重み (w_2) に対する $j-l$ 番目のスライドの重みの比である. ステップ 3 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.

4. 認識実験

4.1 実験条件

講義の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) と, 話者 B による 1 つの講義 (LEC3) から成る. それぞれの講義は 1 コースで 12 回開講され, 一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 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. 実験にはピンマイクで録音された音声を用いた. またこれらの講義において用いられた PowerPoint のスライド資料も, それらが実際の講義の中で使用された時間の情報と合わせ, 収集された. 音声データはその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される. 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途切れた直後に設定した. 評価のためこれらの講義は全て人手で書き起こした. このとき聴講者による質問などの講義の話者以外による発声は除いた. 講義に特徴的な専門用語などを, 講義のキーワードとして, 実験従事者の 1 人が主観で選択した. テストセットとなる講義の詳細を表 1 に示す. また, 各スライドについての詳細を表 2 に示す.

講義音声認識精度を効果的に向上させる最もよい方法は, 書き起こしが整備された大規模な講義データベースを学習データに用いて音響モデルと言語モデルを作成することである. しかしそのような大規模コーパスは現状では存在しない. そこで本研究では, 学会などの講演発表

表 1 テストセット講義の詳細.

	LEC1	LEC2	LEC3
講義名	計算機アーキテクチャ	データベース	計算物理工学
話者	A	A	B
評価対象キーワード数	136	56	53
講義の数	11	10	7
音声長 (時間)	12.8	13.8	9.1
スライドの枚数	265	282	136

表 2 スライド 1 枚あたりの詳細.

	LEC1	LEC2	LEC3
平均単語数	52.1	60.2	52.5
平均キーワード数	9.3	5.5	8.3
平均継続長 (分)	3.1	4.1	5.0

のデータを大量に含む CSJ から構築したモデルをベースラインとした. CSJ の講演データは, 特定のトピックについてのモノローグであるなど, 講義と似た特徴をいくつか持ち, 認識精度に寄与することが期待される.

適応の初期言語モデルとして CSJ の学会講演 967 講演 (3M 形態素) からバイグラムモデルと逆向きトライグラムモデルを構築した. 形態素解析器は ChaSen を用い, 形態素辞書は ipadic を用いた [18]. 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. 言語モデルのバックオフ平滑化には Witten-Bell 法を用いた [19]. この初期言語モデルを, 本論文ではベースライン言語モデルと呼ぶこととする.

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. これらの学習セットは男性話者と女性話者の両方を含む. 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. モデルの構築には HTK [20] を用いた. また音響モデルは MLLR 法 [21] による教師なし適応を行った. 適応データとして, 各回の講義の開始 10 分間の音声を用いた. 回帰クラスの数は 64 とした. MLLR 法による適応により話者および雑音に適応することが期待される. 大語彙連続音声認識デコーダとして Julius [22] を用いた. 音声認識の評価は単語正解精度で行った. また, キーワードに関する認識率の評価を recall と precision で行った. キーワード抽出の性能は, 講義の検索などにおける性能の評価に重要な役割を果たすと考えられる. キーワード抽出の評価は, 認識結果におけるキーワードの開始時間 t_1 と正解開始時間 t_2 を比較して行った. t_1 と t_2 の差が 500 ミリ秒以下であった場合, キーワードが正しく認識されたと判断することとした.

表 3 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 (%).

AM		Word acc.	Recall	Precision	F-measure
Base	LEC1	35.1	30.1	50.5	37.7
	LEC2	33.0	31.2	63.5	41.8
	LEC3	49.4	42.6	75.9	54.5
	Avg.	37.4	33.7	59.7	43.1
Adapted	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	68.0
	Avg.	42.5	42.2	67.3	51.8

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表 3 に示す. 教師なし適応前の音響モデルを使用した場合の単語正解精度の平均値は 37.4% であった. このときキーワード抽出の recall は平均して 33.7%, precision は平均して 59.7% であった. これらの結果は学会講演などの音声認識と比較すると十分な精度とは言えない. 原因としては, 1 章で述べたように, 初期モデルの学習データである CSJ の講演音声とテストセットである講義の発話スタイルの違いが考えられる. MLLR 法を用いた教師なし適応を音響モデルに施したところ, 単語認識精度とキーワード抽出率の両方に改善が見られた. 単語正解精度の平均値は 42.5% であり, 教師なし適応前の音響モデルを用いた結果から 8.1% の誤りが削減された. また recall と precision の誤り率は平均してそれぞれ 12.8%, 18.9% 削減された. 以降の実験において, 音響モデルは教師なし適応を施したものを用いた.

次に先に提案した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 と認識率の関係を図 2 に示した. 結果によると重み係数 w_1 の値は認識率にそれほど影響しない. この結果をもとに w_1 は 20 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表 4 に示した. ステップ 1 の適応により, 単語正解精度は平均して 3.1% の誤りが削減された. またキーワード抽出は recall において平均して 21.5% の誤りが削減され, precision において平均して 13.8% の誤り

表 4 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-measure
LEC1	41.1	50.6	65.9	57.2
LEC2	38.7	49.8	71.6	58.7
LEC3	59.9	65.9	83.0	73.5
Avg.	44.3	54.6	71.8	62.0

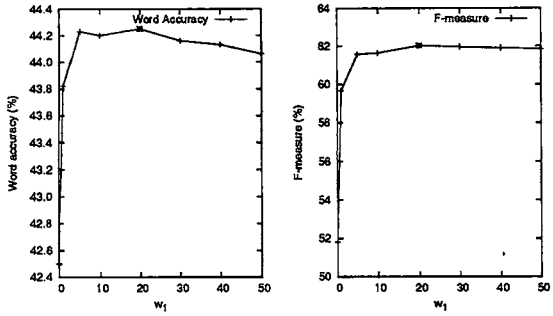

図 2 重み係数 w_1 と認識性能の関係

表 5 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-measure
LEC1	41.1	52.6	66.3	58.6
LEC2	38.7	51.5	72.7	60.3
LEC3	59.4	66.2	83.0	73.6
Avg.	44.2	56.0	72.2	63.1

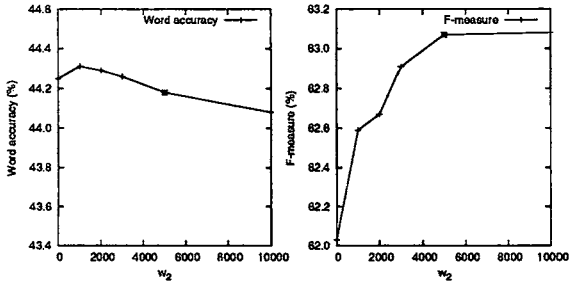

図 3 重み係数 w_2 と認識性能の関係 ($w_1 = 20$).

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 による適応を施した言語モデルについて, 認識による評価を行った. 重み係数 w_2 と認識性能の関係を図 3 に示した. 本実験では重み係数 w_2 の違いは認識性能にそれほど影響がなかった. この結果を受け, 重み係数 w_2 は 5,000 とした. ステップ 2 による適応を施した言語モデルを用いたときの認識結果を表 5 に示した.

表 6 ステップ 1 と ステップ 2 によるキーワード抽出における F 値 (%) (Summary).

LM	Baseline	Step 1	Step 2
LEC1	45.7	57.2	58.6
LEC2	46.7	58.7	60.3
LEC3	68.0	73.5	73.6
Avg.	51.8	62.0	63.1

表 7 ステップ 3 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-measure
LEC1	41.1	52.4	66.4	58.6
LEC2	38.7	51.8	72.7	60.5
LEC3	59.4	66.4	83.0	73.8
Avg.	44.2	56.1	72.2	63.1

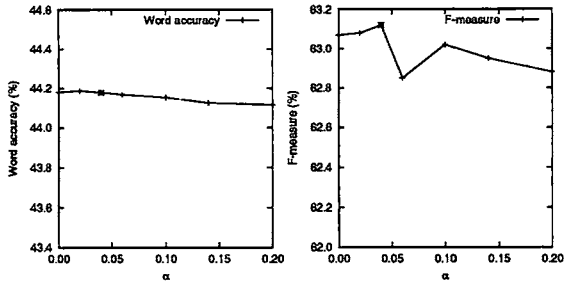

図 4 係数 α と認識性能の関係 ($w_2 = 20$, $w_3 = 5,000$).

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, キーワード抽出の性能に改善が見られた. recall について平均して 3.1% の誤り率が削減され, precision について平均して 1.4% の誤り率が削減された. これらの結果は提案手法がキーワード抽出の性能向上に効果があることを示している.

ステップ 1 とステップ 2 のそれぞれの適応を行ったときのキーワード抽出性能を F 値で表したものを表 6 に示した. 講義ごとに結果を比較したとき, LEC3 におけるステップ 2 の効果が他の講義と比較して小さいという結果であった. この結果の原因に関する調査は今後の課題とする.

続いてステップ 3 による適応を施した言語モデルについて, 認識による評価を行った. 本実験では $k = 2$ とした. 係数 α と認識性能の関係を図 4 に示した. 本実験では係数 α の違いは認識性能にそれほど影響がなかった. この結果を受け, 係数 α は 0.04 とした. ステップ 3 による適応を施した言語モデルを用いたときの認識結果を表 7 に示した. 単語正解精度, キーワード抽出性能共に

ステップ 2 の結果と大きく変わらなかった．この結果から，ある時点での講義音声認識を考えるとき，それ以前の講義内容が含まれるスライドをモデル適応データとして用いたとしても大きな効果は見られないことが分かった．

5. 結論と今後の課題

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．音響モデルの MLLR 法による適応と組み合わせることで，提案手法による音声認識率の改善が得られた．特にキーワード抽出において効果が得られた．

現在は 3 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．本論文ではローカルな適応の際に，発話に対応する 1 枚のスライドだけでなく，さらにその直前のスライドを用いる手法を検討したが，大きな効果は得られなかった．スライド情報をより効果的に利用する枠組みの検討が求められる．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．

文　献

[1] H. Yokota, 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.

[2] H. Yokota, T. Kobayashi, H. Okamoto and W. Nakano, "Unified contents retrieval from an academic repository," Proc. International Symposium on Large-scale Knowledge Resources (LKR2006), Tokyo, Japan, pp. 41-46, 2006.

[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.

[4] A. Fujii, K. Itou and T. Ishikawa, "LODEM: A system for on-demand video lectures," Speech Communication 48, pp. 516-531, 2006.

[5] R. Müller and T. Ottmann, "The Authoring on the Fly system for automated recording and replay of (tele)presentations," Multimedia Systems, vol. 8 no. 3, pp. 158-176, 2000.

[6] Informedia ii digital video library, Carnegie Mellon University The Informedia Project, http://www.informedia.cs.cmu.edu/.

[7] G. D. Abowd, "Classroom 2000: an experiment with the instrumentation of a living educational environment," IBM Systems Journal, vol. 38, no. 4, pp. 508-530, 1999.

[8] I. Trancoso, R. Nunes and L. Neves, "Recognition of classroom lectures in European Portuguese," Proc. INTERSPEECH 2006 - ICSLP, pp. 281-284, 2006.

[9] L. Lamel, G. Adda, E. Bilinski and J. L. Gauvain, "Transcribing Lectures and Seminars," Proc. INTERSPEECH 2005, pp. 1675-1660, 2005.

[10] J. Glass, T. Hazen, I. Hetherington and C. Wang, "Analysis and processing of lecture audio data: Preliminary investigations," Proc. Human Language Technology NAACL, Speech Indexing Workshop, Boston, 2004.

[11] K. Maekawa, H. Koiso, S. Furui and H. Isahara, "Spontaneous speech corpus of Japanese," Proc. LREC2000, Athens, Greece, vol. 2, pp. 947-952, 2000.

[12] The Corpus of Spontaneous Japanese, National Institute for Japanese Language, http://www2.kokken.go.jp/csj/public/.

[13] S. Furui, "Recent progress in corpus-based spontaneous speech recognition," IEICE Transactions on Information and Systems, vol. E88-D, no. 3, pp. 366-375, 2005.

[14] L. Lamel, F. Schiel, A. Fourcin, J. Mariani and H. Tillmann, "The Translanguage English Database TED," Proc. ICSLP, vol. 4, pp. 1795-1798, 2004.

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

[16] M. Cettolo, F. Brugnara and M. Federico, "Advances in the automatic transcription of lectures," Proc. ICASSP 2004, vol. 1, pp. 769-772, 2004.

[17] N. Ozawa, H. Takebe, Y. Katsuyama, S. Naoi and H. Yokota, "Slide Identification for Lecture Movies by matching Characters and Images," Proc. SPIE, vol. 5296-10, Document Recognition and Retrieval XI, pp.74-81, 2004.

[18] ChaSen (ver. 2.2.3) and ipadic (ver. 2.4.4), http://chasen.naist.jp/hiki/ChaSen/.

[19] I. H. Witten and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," IEEE Transactions on Information Theory, vol. 37, no. 4, pp. 1085-1094, 1991.

[20] HMM Tool Kit (HTK) (ver. 3.2), http://htk.eng.cam.ac.uk/.

[21] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol. 9, no. 2, pp. 171-185, 1995.

[22] Julius (ver. 3.5), http://julius.sourceforge.jp/en/julius.html.