授業音声字幕化のための学習データ分類に基づく話者依存音響モデル学習

穂坂　圭一 †1　伊藤　信義 †2
西崎　博光 †3　関口　芳廣 †3

本研究では，授業音声の自動字幕化のための音声認識率改善方法として音響モデルの構築方法について検討を行った．字幕化を目的とした場合，大量の教員音声から特定話者モデルを学習するべきである．しかし，現状ではコストの問題から，それが難しい．そこで２つの方法を提案する．１つは，既存の講演音声データベースを用いて，データベース内の講演をクラスタリングし，クラスタ毎の音響モデルを学習する方法，もう１つは，教員音声と似た講演をデータベースから選別し，それを用いて学習を行う方法である．これらの手法により，類似話者モデルを学習し，教員音声の認識率改善を図る．提案手法を大学授業，小学校授業音声に適応したところ，有効性を確認できた．

Speaker Dependent Acoustic Modeling Using Training Data Classification for Subtitling Lecture Speech

Keiichi Hosaka,†1 Nobuyoshi Itoh,†2
Nishizaki Hiromitsu†3 and Sekiguchi Yoshihiro †3

This paper describes acoustic modeling to improve speech recognition performance of lecture speech for subtitling lecture speech. Specific speaker model is very useful for improving recognition performance, but it is very difficult because it takes high-cost of collecting so much lecture speech. We propose two techniques of specific speaker acoustic modeling. One is to train the model from a speaker group made by clustering a lot of speakers. The other is to select the speeches from the database, and train the model from them. In the speech recognition experiment on university and elementary school lecture speech, we showed that the acoustic modeling technique we proposed improved the recognition performance.

1. はじめに

近年，聴覚障害のハンデを持った子どもの数が年々増加している．学校の授業で教員の音声を聞くことができなければ，授業内容を把握するのが困難である．そのため，教員は教材などを工夫する必要がある．そこで，小学校・中学校・高校・大学の授業において，聴覚障害児・生徒・学生のための授業内容保証のために，教員音声を字幕化するプロジェクトが盛んに行われている[1)2)3)]．

授業音声の字幕のためには，教員音声を即座にタイプライターに入力して字幕にする方法や，音声認識を用いて認識した結果を人間が訂正し，字幕として出力する方法がある．

授業音声の字幕を実現するためには，このように多くの人の力が必要となる．現在は聴覚障害者が増加しており，すべての障害者が支援体制の整った学校へ進学できるとは限らないため，一般学校へ入ることを余儀なくされた．従って一般学校に，字幕システムが導入されることが望ましいが，復唱者や字幕訂正者の人手不足の問題もあり，多くの学校で授業音声の字幕化を実現することはほぼ不可能である．このような背景から，（ほぼ）全自動で教員音声を字幕化するシステムの構築が望まれている．

教員音声を全自動で字幕化するためには，現状では音声認識技術が必要不可欠である．しかし，授業音声は話し言葉であり，かつ学会講演よりもより砕けた話し方である[4)]．そのため，授業音声を高い音声認識精度で認識することが難しく，専用の音声認識誤り訂正者がいなければ字幕化が難しい．日本放送協会（NHK）の放送音声の字幕化でも，音声認識率95%以上が得られていたとしても訂正者が必要である[5)] ⋆1．

近年，日本語話し言葉コーパスや日本語講義音声コンテンツコーパス（CJLC）[4)] のように大量の話し言葉音声を収録した音声コーパスが公開されている．多くの研究機関が，これ

†1 山梨大学大学院医学工学総合教育部コンピュータ・メディア工学専攻
Dept. of Computer Science and Media Engineering, Educational Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi
†2 山梨大学工学部コンピュータ・メディア工学科
Dept. of Computer Science and Media Engineering, Faculty of Engineering, University of Yamanashi
†3 山梨大学大学院医学工学総合研究部
Dept. of Research Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi
⋆1 ただし，授業音声の場合は，音声認識誤りが多少含まれていても問題はないと思われる．

らのデータを利用することで，授業音声の音声認識率を改善するための手法を提案している[6)7)]．

本稿でも，授業音声の字幕化に向けて，音声認識を改善する手法を提案する．授業音声認識を改善する方法としては，先程挙げた文献で提案されている手法や，音響モデルの話者適応化，話者依存モデル化など様々の方法が考えられる．

我々も，言語モデル，音響モデルに関してのアプローチを試しており[8)]，本研究では特に音響モデリングについて紹介する．

授業を音声認識する際には，その授業者の十分な音声データがあれば特定話者音響モデルを利用すると最も認識率が良くなるはずである．認識対象教員の十分な量の音声を予め録音し（そのときの録音環境は実際に字幕システムを使う環境と同じにしておくことも重要），その書き起こしを行うことによって，音響モデルの学習データが用意できる．しかし，大量のデータの書き起こしには膨大な時間が必要であり，現場の教師がそれを行うことは得策ではない．

そこで，本研究では特定話者音響モデルを学習する際に問題となるデータ収集の問題を解決する方法を提案する．提案手法では，既存の音声コーパスを話者クラスタリングによって分類し，認識対象話者に適したモデルを使用する方法（これを“ 話者クラスタリング法 ”と記す）と，認識対象話者に似た音声を既存の音声コーパスから選別しそこから学習したモデルを使用する方法（これを“ 類似話者選別法 ”と記す）を用いる．これら 2 つの手法を用い，特定話者音響モデルの学習に関する問題の解決を図る．

伊藤ら[9)] は，少量の認識対象音声と，話者選択用の特定話者モデルを用いることで音声コーパスから類似話者選別を行っている．そして，学習話者数を自動的に変更することで，講演音声認識に対して高い認識性能を持つモデルを構築している．これに対し，本研究では，認識対象音声の書き起こしが事前に用意できる場合などを考慮した類似話者選別や，話者クラスタリング法によるモデル構築との比較を行っている．

本研究では，提案手法を大学の授業音声と小学校の授業音声に適応した．実験の結果，提案手法，特に類似話者選別法に基づく音響モデルが，大学・小学校授業音声の認識率改善に有効であることが分かった．しかし，小学校授業音声の認識に関しては，改善したとしても字幕化に耐えうる精度には程遠く，課題が残ることも明らかとなった．

本稿の構成は次の通りである．まず 2 節では授業音声認識における特定話者モデルの有効性について説明する．3 節と 4 節では，提案手法である話者クラスタリング法と類似話者選別法について説明し，大学授業音声に対し提案手法の有効性の確認を行う．5 節で小学校授業音声認識実験について述べ，最後に 6 節でまとめを述べる．

表 1　特定話者モデルの学習に使用した音声
Table 1　Training speech data for specific speaker modeling.

モデル	学習音声	学習音声の総時間数
不特定話者モデル	CSJ に収録されている講演音声 2517 個 （男性講演 1513 個，女性講演 1004 個）	約 560 時間
特定話者モデル	教員 A 授業音声 3 授業分	約 4 時間
	教員 B 授業音声 3 授業分	約 1.5 時間
	教員 C 授業音声 3 授業分	約 4.5 時間

2. 特定話者モデルの有効性の確認

2.1 特定話者モデル

本稿でいう特定話者モデルとは，音声認識対象となる教員本人の音声から学習した音響モデルである．認識対象教員の音声を用いて学習を行うので，この教員に対しては不特定話者モデルを用いるよりも高い認識率を得ることができる．

まず，認識対象教員本人の音声とその書き起こしから特定話者モデルを作成した．そして，不特定話者モデルと特定話者モデルを使った場合の認識率を比較し，特定話者モデルを使用した方が，認識率が高くなることを実験的に確認する．

2.2 実 験 条 件

不特定話者モデルは，日本語話し言葉コーパス[10)](以下“ CSJ ”と略す) に収録されている 2517 講演から学習を行った．特定話者モデルは，2005 年度から 2008 年度に山梨大学工学部コンピュータ・メディア工学科で開講された，数学，物理，コンピュータサイエンス等に関する授業のうち，教員 3 名の授業（各教員につき 3 授業分の音声を使用）を用いて学習を行った．

音声認識に用いる言語モデルは，CSJ の全講演音声の書き起こしから学習した，語彙数約 42,000 語のトライグラムモデルを使用し，連続単語音声認識を行う．

評価に用いる音声は，特定話者のモデル学習に使用した教員 3 名の授業音声である．これは，特定話者モデル学習の際に使用した音声とは，異なる授業音声を用いる．

各音響モデルの学習に使用した音声の詳細を表 1 に示す．また，モデルの学習条件を表 2 に示す．

2.3 実験結果と考察

不特定話者モデルと特定話者モデルの認識率を表 3 に示す．これは 3 名の授業の平均単

表 2　音響モデルの学習条件
Table 2　Training condition for acoustic modeling.

サンプリング周波数	16[kHz]
量子化ビット数	16[bit]
分析窓長	25[ms]
窓間隔	10[ms]
特徴パラメータ	$MFCC(12) + \Delta MFCC(12) + \Delta\Delta MFCC(12)$ $+\Delta Pow(1) + \Delta\Delta Pow(1)$(計 38 次元)
音響モデルの単位	syllable
HMM の状態数	母音 3 状態，子音 5 状態の Left-to-Rignt HMM
HMM の混合数	64 混合

表 3　特定話者モデルの認識率 [%]
Table 3　Recognition rates using specific speaker model.

教員		不特定話者モデル	特定話者モデル
教員 A	正解率	44.01	53.70
	正解精度	36.55	45.20
教員 B	正解率	48.75	51.01
	正解精度	40.64	37.34
教員 C	正解率	52.50	67.12
	正解精度	46.27	61.08
平均	正解率	48.42	57.28
	正解精度	41.15	47.87

語認識率である．

表 3 から，不特定話者モデルよりも特定話者モデルの平均認識率の方が正解率で約 9%，正解精度で約 6%高いことが分かる．従って，特定話者モデルを使用することは有用である．しかし，特定話者モデルは，3 コマ分の授業音声より学習されているが，1 コマあたりの授業音声が短い教員 B では，それでも十分な学習データが得られなかったためか，不特定話者モデルよりも正解精度が悪くなっていた．従って，ある程度の十分な学習データを確保しておく必要がある．

2.4　特定話者モデルの問題点

実験結果より，特定話者モデルを使用することで認識率が高くなることが実験的に示された．従って，教員の音声認識を行う場合は，その教員の音声から学習した特定話者モデルを用いた方が認識率は高くなる．

しかし，特定話者モデルを作成するには，問題点がある．それは，前述したように，事前に認識対象教員の十分な長さの音声とその書き起こしが必要となってくる点である．

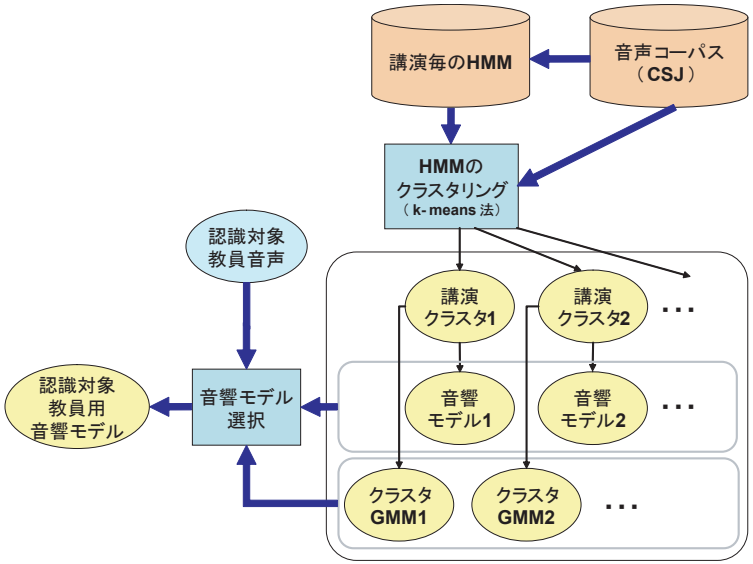

図 1　話者クラスタリング法の流れ
Fig. 1　Block diagram of speaker clustering.

学習に十分な認識対象教員のデータを集めることは，非常に困難である．しかし，近年では，CSJ や CJLC のように多数の話者の講演・講義音声を収録した音声データベースが利用可能となった．認識対象教員のデータを大量に集めることは難しいが，このような音声コーパスから認識対象教員に類似した音声を集めることは可能である．

そこで本研究では，話者分類や類似話者選別を行い，認識対象教員に適した音響モデルを構築する手法について検討を行った．

3.　話者クラスタリング法

本節では，話者クラスタリングによって認識対象教員に適したモデルを構築する手法（話者クラスタリング法）を説明する．

3.1　話者クラスタリングによるモデル学習

話者クラスタリング法の手順を図 1 に示す．

まず，CSJ の講演の話者クラスタリングを行う．あらかじめ，講演ごとに HMM を学習

表 4　話者クラスタリング法による音響モデルの認識率 [%]
Table 4　Recognition rates using acoustic model by speaker classification method.

モデル	正解率	正解精度
不特定話者モデル	44.83	37.72
クラスタリング	47.78	38.29

しておく．HMM の学習条件は，音響モデルの単位と混合数を除いて表 2 と同じである．単位は monophone，混合数は 1 混合である．HMM の分布間距離（バタチャリヤ距離[11)]）に基づく k-means クラスタリングを行うことで，講演のクラスタリングを行う．そして，各クラスタごとに音響モデルを作成する．また，各クラスタごとの GMM も作成しておく．作成した複数の音響モデルの中から，認識に用いるモデルを選び，音声認識を行う．音声認識に使用する音響モデルの選択は，作成した GMM と認識対象音声の音響尤度を計算して行う．

3.2　話者クラスタリング法により構築したモデルの評価実験

3.1 節の手順に基づき，実際にモデルを構築し，授業音声を用いて評価実験を行った．そして，不特定話者モデルと比較を行い，本手法が有効であるか確認を行った．

3.2.1　実 験 条 件

音声認識に用いる音響モデルは，CSJ の 2517 講演から学習した不特定話者モデルと，話者クラスタリング法により作成した音響モデルである．クラスタリング法では，CSJ の講演を 29 個のクラスタに分類し（各クラスタの講演数が 100 程度になるようにした），それぞれのクラスタに分けられた講演音声から 29 種類の音響モデルを学習した．また，音響モデルの学習と同時に，GMM の学習も行った．

モデルの学習条件は，2.2 節の表 2 と同じである．評価に用いる音声は，山梨大学コンピュータ・メディア工学科の教員 7 名の授業音声である．各教員音声に最も近いクラスタの音響モデルを，教員音声と GMM の音響尤度により選択し，認識を行った．

3.2.2　実験結果と考察

7 個の授業音声の平均認識率を表 4 に示す．これは 7 名の授業の平均単語認識率となっている．

表 4 より，不特定話者モデルと比べて，話者クラスタリング法により学習したモデルを使用した方が，平均正解率で約 3%，平均正解精度で約 0.5%改善していた．従って，本手法が有効であることが確認できた．

しかし，大きな改善は見られなかった．これは，CSJ 講演のクラスタリングを行うことによって，似た講演同士のクラスタを作成することができたが，クラスタに含まれる講演音声すべてが認識対象教員に近い音声であるとは限らないためである．

そこで次節では，認識対象教員の音声データを用意できることを前提とした音響モデリング法を提案する．これは，教員音声と似た講演音声を直接選択する方法であり，話者クラスタリング法よりもより認識対象教員の音声特徴を反映する音響モデルを学習できるはずである．

4.　類似話者選別法

本節では，認識対象教員に類似した音声を選別することで話者依存モデルを構築する手法（類似話者選別法）を説明する．

前述した通り，教員の特定話者モデル構築のために，授業音声を大量に用意し，その書き起こしを用意することはコストの面で非常に困難である．しかし，数コマ分の授業音声ならば，録音し，書き起こすことができるのではないかと考えられる．

類似話者選別法では，まず，認識対象教員の少量の音声データとその書き起こしから，教員の HMM を学習する．これを用いて，CSJ から教員に類似した講演音声を選別する．選別した講演音声から音響モデルを学習することで，類似話者モデルを学習する．このとき，類似話者選別用 HMM を学習する際には，正解の書き起こし（ラベル）を用いて学習する場合（これを“ 教師あり話者選別 ”と記す）と，音声認識により自動生成したラベルを用いて学習する（これを“ 教師なし話者選別 ”と記す），の 2 通りが考えられる．本研究では，教師なし・教師ありの両方の場合の実験を行った．

4.1　教師なし話者選別

教師なし話者選別に基づく類似話者選別法の手順を図 2 に示す．

まず，不特定話者モデルを用いて，認識対象教員 A の音声を認識し，認識結果を得る．この認識結果から音節ラベルを作成し，教員 A の類似話者選別用 HMM を学習する．一方で，CSJ の各講演毎の HMM を学習しておく．教員 A の類似話者を選別するため，教員 A 選別用 HMM と各講演毎の HMM とのバタチャリア距離を計算し，距離が短い順に 100 講演の講演音声を選別する．選別された講演音声は，教員 A の音声と似ているはずである．この教員 A に似た講演音声を用いて，教員 A 音声の認識用 HMM を学習する．

従来手法では，類似話者選別用のモデルとして GMM を用い，音響尤度により類似話者選別を行っている[12)]．しかし，GMM は音声全体の特徴をモデル化したものなので，音韻毎の特徴をモデル化する HMM の方が，より精度の高い選別が可能となる．

話者選別でも，直接，教員 A の音声と各講演毎の HMM との音響尤度を計算する方法も

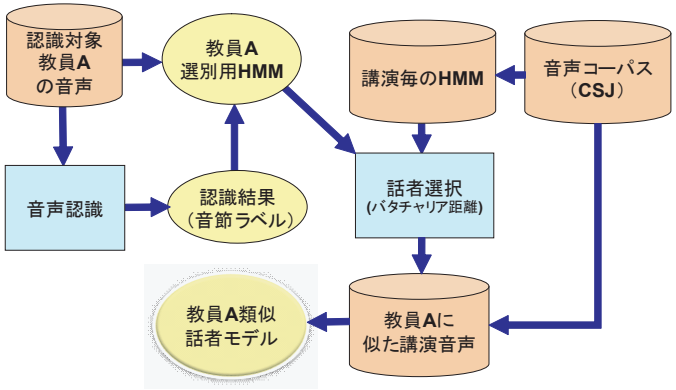

図 2　教師なし話者選別に基づく類似話者選別法の流れ
Fig. 2　Block diagram of similar speaker selection with unsupervised modeling.

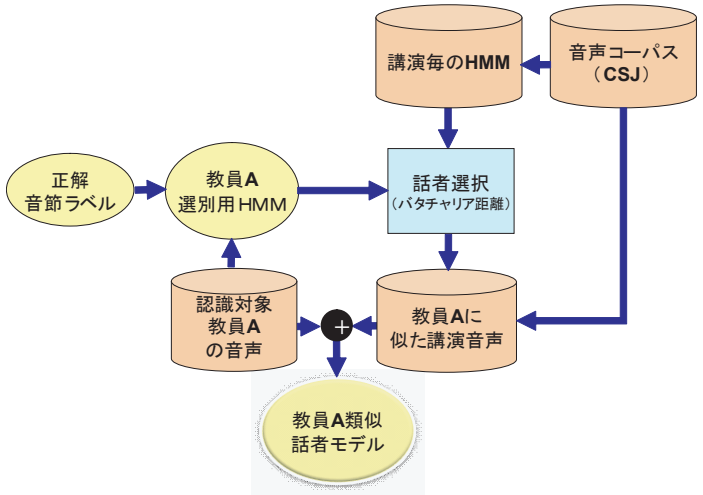

図 3　教師あり話者選別に基づく類似話者選別法の流れ
Fig. 3　Block diagram of similar speaker selection with supervised modeling.

考えられるが，HMM を用い音響尤度を計算すると多大な時間が掛かってしまう．そこで本研究では，計算量が比較的少ない HMM 間のバタチャリヤ距離を用いて類似話者選別を行うことにした．

4.2　教師あり話者選別

教員音声とその正解書き起こしがある場合の教師あり話者選別に基づく類似話者選別法の手順を図 3 に示す．

まず，認識対象教員を教員 A とすると，教員 A の音声とその正確な音声ラベルから選別用 HMM を学習する．その後は，教師なし選別と同様の処理により，教員 A の類似話者モデルを作成する．ただし，教員 A の音声と正確な書き起こしがあるならば，それを認識用音響モデル学習に利用できる．そこで，選別用の教員 A の音声を選別された教員 A に似た講演音声グループに加え，音響モデルを学習する（加えない場合も実験した）．

4.3　類似話者選別法により作成したモデルの評価実験

4.1 節，4.2 節の手順に基づき，認識対象の教員に対する類似話者モデルを学習した．この節では，本手法を大学授業音声を用いて音声認識実験により評価する．

4.3.1　実 験 条 件

音声認識に用いる音響モデルは，次の 5 種類である．

- 不特定話者モデル，
- 教師なし話者選別（GMM と HMM）を用いて学習した類似話者モデル（2 種類），
- 教師あり話者選別による類似話者モデル（学習の際に教員音声を追加した場合と追加しなかった場合の 2 種類）．

作成した音響モデルの詳細を表 5 に示す．モデルの学習条件は，2.2 節の表 2 と同じである．

評価に用いる音声は，山梨大学工学部コンピュータ・メディア工学科の教員 7 名の授業音声である．もちろん，各教員において，類似話者モデルを学習する際に用いる授業音声と異なる音声を評価対象としている．

4.3.2　実 験 結 果

7 個の授業音声の平均認識率を表 6 に示す．これは 7 名の授業の平均単語認識率である．

表 6 より，教師なし話者選別と教師あり話者選別のどちらにおいても，不特定話者モデルの認識率を大きく改善している．教師なし話者選別では，類似話者を選択するモデルとして GMM と HMM を用いたが，HMM を用いた方が 2%ほど正解率・正解精度が良い．このことからも，HMM を使って類似話者を選択した方が，認識対象教員により近い音声データを選別できていることが言える．

表 5　類似話者選別法により作成した音響モデル
Table 5　Acoustic models trained by the similar speaker selection method.

モデル	学習音声
不特定話者	CSJ の学会・模擬講演 2517 個
教師なし話者選別（GMM）	認識対象音声から作成した GMM によって選別した講演音声 100 個
教師なし話者選別（HMM）	教師なし話者 HMM によって選択された講演音声 100 個
教師あり話者選別（HMM）	教師あり話者 HMM によって選択された講演音声 100 個
教員音声利用（HMM）	教師あり話者選別法による学習音声 ＋ 認識対象教員の授業音声 1 コマ分

表 6　類似話者選別法により作成した音響モデルの平均認識率 [%]
Table 6　Recognition rates using acoustic models by the similar speaker selection.

モデル	正解率	正解精度
不特定話者モデル	44.83	37.72
教師なし話者選別（GMM）	49.31	40.77
教師なし話者選別（HMM）	51.55	42.50
教師あり話者選別（HMM）	52.48	43.41
教員音声利用（HMM）	55.39	47.74

表 7　特定話者モデルと話者依存モデルの平均認識率 [%]
Table 7　Recognition rates using specific speaker model and similar speaker models.

モデル	正解率	正解精度
特定話者モデル	57.28	47.87
教員音声利用（HMM）	59.32	51.67

教師なし／ありの比較では，正解率・正解精度ともに 1%程度の差しかない．よって話者選別の際には，教師なしモデルでも十分に類似話者を選択できることが分かった．

さらに，教師あり話者選別で選ばれた学習データに教員音声を加えることで，さらに大きく認識率が向上した．最終的には，不特定話者モデルと比べて正解率と正解精度ともに約 10%改善している．

4.4　特定話者モデルとの比較

認識対象教員の音声から学習した特定話者モデルと，教師あり類似話者選別法（＋教員音声も HMM 学習データに加える）による類似話者モデルを用いたときの教員 3 名の授業音声の平均認識率を表 7 に示す．この音声は，2 節の授業と同じものである．

表 7 より，提案手法である類似話者選別法の認識率が，特定話者モデルよりも高くなった．これにより，事前に十分な特定話者音声データを準備できない場合に，本提案手法が有効であることが分かる．

4.5　考　　察

事前に，教員音声とその正確な書き起こしを持っている／持っていないどちらの場合においても，提案手法である類似話者選別法 (HMM) に基づいて学習した音響モデルの方が，認識率が高くなった．

表 4 の話者クラスタリング法のモデルの認識率と比べても，本手法を用いることでより高い認識率の改善ができている．

また，3 授業分の音声から学習した特定話者モデルよりも，1 授業分の音声から類似話者選別を行い作成したモデルの方が認識率が高くなった．従って，提案手法を用いることで，音響モデル作成のコスト (特定話者の学習データ収集，その書き起こしの準備等の労力) をあまりかけずとも，高い精度での音声認識を行うことが可能である．

大学授業音声の実験結果より，話者クラスタリング法を使用するよりも，類似話者選別法を使用する方が認識率改善に有効であることが分かった．そこで，次節の小学校授業音声に対しての評価は，類似話者選別法に基づいてモデルを構築し評価実験を行う．

5.　小学校授業音声に対する提案手法の評価

これまでに，話者クラスタリング法および類似話者選別法に基づいて学習した音響モデルの評価を，大学の授業音声で行った．その結果，大学の授業音声に対しては，事前に認識対象教員の 1 コマ分の授業音声が用意さえできれば，大きく音声認識率を改善できることが分かった．

1 節で述べたように，小学校においても授業音声の字幕化の要望が強い．これは，聴覚障害を背負った児童の増加，アスペルガー症候群に代表される高機能広汎性発達障害児の増加という背景によるものである[⋆1]．

本節では，小学校授業向け字幕提示システム開発に向け，小学校音声認識に対しても提案する音響モデル学習手法が有効であるかどうかを確かめる．

5.1　小学校授業音声

岐阜県内の 2 つの公立小学校 (1 つは特別支援学校) に協力して頂き，小学校の授業音声を収録した．授業音声の録音は，リニア PCM で録音できる IC レコーダーを用い，ピンマイクで収録した．授業内容は，国語，算数，ホームルームなど様々な科目を含んでいる．

⋆1 発達障害児の多くは，記憶力が乏しく，教員音声を理解できない．そのため，授業中に何を指示されているのか理解できず，落ち着きがなくなってしまうことから，字幕提示が発達障害児の支援に効果があると考えられている．

表 8　小学校授業音声認識に使用する音響モデル
Table 8　Acoustic models used to recognize lecture speech of elementary school class.

モデル	学習音声もしくは適応に用いる音声
不特定話者	CSJ の学会・模擬講演 2517 個
教師なし話者選別（HMM）	教師なし話者 HMM によって選択された講演音声 20〜40 個
教師あり話者選別（HMM）	教師あり話者 HMM によって選択された講演音声 20〜40 個
教員音声利用（HMM）	教師あり話者選別法による学習音声 ＋ 認識対象教員の授業音声 1 コマ分
適応化	教員音声利用（HMM）による音響モデルを教員音声を用いて教師あり MLLR 適応

収録した小学校授業音声と大学授業音声との大きな違いは，次の通りである．

- 発話速度が遅い
- 話し方が丁寧
- 問い掛け，命令調の言い方が多い
- フロアの雑音が多い

5 名の小学校教員（男性 1 名，女性 4 名）の授業音声を評価に用いる．各教員音声において，類似講演を探すために利用する音声と評価に使う音声を違うものにするため，2 授業ずつ用意した．ただし，1 授業しか録音されていない教員音声の場合は，1 つの授業を前半・後半に 2 分割し，片方を学習用に，片方を評価用とした．

5.2 実験条件

音声認識実験では，下記の 5 種類の音響モデルを用いた．

- CSJ の 2517 講演から学習した不特定話者モデル，
- 教師なし話者選別法 (HMM) により学習した類似話者モデル，
- 教師あり話者選別法 (HMM) により学習した類似話者モデル．類似話者選別用の教員音声を認識用 HMM 学習に用いた場合と用いない場合の 2 種類，
- 教師あり話者選別法＋教員音声によるモデルに対し，MLLR による教師あり話者適応を行ったモデル．

表 8 に使用した音響モデルについてまとめる．音響モデルの学習条件は，混合数を除いて表 2 と同じである．混合数は 32 混合となっている．

今回，小学校授業音声を認識するための適切な言語モデルを用意できなかったため，3 節，4 節の認識実験とは異なり，連続音節認識による評価となっている．評価は，単語ではなく音節認識率で行う．

5.3 実験結果

実験結果を表 9 に示す．これは 5 名の教員の平均音節認識率となっている．

表 9　小学校授業音声の平均認識率 [%]
Table 9　Recognition rates of elementary school lecture speech.

モデル	音節正解率	音節正解精度
不特定話者	27.61	-9.063
教師なり話者選別（HMM）	41.47	-0.080
教師あり話者選別（HMM）	41.50	1.476
教員音声利用（HMM）	45.20	8.335
適応化	49.32	10.86

まず，CSJ による不特定話者モデルにおいては，正解率で 28%，正解精度が-9%となり，非常に悪い結果となった．また，挿入誤りが非常に多い．

不特定話者に対し，教師あり・教師なし話者選別手法で学習した音響モデルでは，約 14%の正解率の改善が得られ，大学授業と同様に，提案手法の有効性を示せた．表 9 からも分かるように，初期認識結果が非常に悪い教師なし話者選別を行っても，教師あり話者選別の場合と差がほとんどない．

認識対象教員の音声を学習データに追加すると，さらに認識率が改善した．また，MLLR による適応化を行うことで，最終的には，不特定話者モデルに比べて約 22%（27.6%→ 49.3%）正解率が向上した．

5.4 考　　察

提案手法（と話者適応）を用いることで認識率を大きく改善することができた．小学校授業音声においても提案手法が有効であると言える．また，教師なしと教師ありの話者選別に基づく音響モデルを使った認識率は，ほとんど差がなく同程度であった．この結果から，事前に認識対象教員の授業が 1 コマ分あれば，その書き起こしが無くても，認識率を改善できることが分かった．ただ，授業音声の書き起こしを人手で準備さえすれば[⋆1]，教師あり学習によってさらに高い音声認識率を得ることができる．

提案手法の有効性は確認できた．しかし，小学校授業音声を認識させたい場合，不特定話者モデルでは正解率が 28%，正解精度が-9%であり，CSJ の講演音声を使った音響モデル学習では不十分であることが明らかとなった．前述したとおり，小学校音声と CSJ の講演音声との特徴がかなり異なる．CSJ の講演音声と小学校授業との録音環境が大きく違ったり，話し方が異なったり，ノイズが多いことが認識率の大幅な低下を招いている．特に，小学校授業音声の発話速度が CSJ の講演音声よりも遅めであることから，挿入誤りが多く

⋆1 指定したテキストを教員に読み上げてもらうことでも対処できるかもしれない．

なっている．これは，全てのモデルにおいて言えることである．

6. ま と め

本稿では，授業音声字幕化のための，音響モデルの構築方法について検討を行った．本研究では，話者クラスタリングや類似話者選別によって，認識対象教員音声認識に適した音響モデルを構築する方法を提案した．

大学授業音声や小学校授業音声を用いて提案手法が有効であるか評価を行った．提案手法に基づいてモデルを構築することで，不特定話者に比べて認識率が改善できた．特に，教師あり話者選別による類似話者音響モデルを用いることで，音声認識率を大きく向上させることに成功し，提案手法の有効性を示すことができた．

しかしながら，小学校授業音声認識に関しては，絶対的な音声認識率が低い等の問題点が多く残っている．今後は，話者ごとに必要な音響モデルの学習数などの詳しい調査を行い，さらに音響モデルの改善を行うこと，言語モデルの話題適応化を行うなどをして，小学校授業音声認識率のさらなる改善を目指していく予定である．

参 考 文 献

1) 京都大学，総務省，全日本難聴者・中途失聴者団体連合会近畿ブロック，速記科学研究会，速記懇談会, “『聴覚障害者のための字幕付与技術』シンポジウム 2009 ”, http://www.ar.media.kyoto-u.ac.jp/jimaku/jimaku09.html，2009.11
2) 若月大輔，加藤伸子，河野純大，塩野目剛亮，村上裕史，皆川洋喜，西岡知之，三好茂樹，黒木速人，内藤一郎, “ 遠隔地リアルタイム字幕支援のためのキーワード入力ソフトウェアの研究開発 ”，筑波大学テクノレポート Vol.16 2009.3
3) 中野聡子，金澤貴之，牧原功，黒木速人，上田一貴，井野秀一，伊福部達, “ 音声認識技術を利用した字幕呈示システムの活用に関する研究―聴覚障害者のニーズに即した呈示方法― ”，メディア教育研究，Vol.5，No.2，pp.63-72, 2008
4) 土屋雅稔，小暮悟，西崎博光，太田健吾，山本一公，中川聖一, “ 日本語講義音声コンテンツコーパスの作成と分析 ”，情報処理学会論文誌，Vol.50，No.2，pp.448-459，2009
5) 今井亨，小林彰夫，佐藤庄衛, “ 放送用リアルタイム字幕制作のための音声認識技術の改善 ”，第 2 回音声ドキュメント処理ワークショップ，pp.89-94，豊橋技術科学大学メディア科学リサーチセンター，2008
6) A. Park， T.J. Hazen， and J.R. Glass, “ Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling ”，Proc. of the ICASSP2005，Vol.I，pp.I-497-500，2005
7) 根本雄介，秋田祐哉，河原達也, “ 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 ”，第 1 回音声ドキュメント処理ワークショップ， pp.89-94， 豊橋技術科学大学メディア科学リサーチセンター，2007
8) 藤原裕幸，西崎博光，関口芳廣, “ 話題依存言語モデル構築のための LSA と単語発音情報を用いた語彙推定 ”，第 8 回情報科学技術フォーラム（FIT）講演論文集第 2 文冊，pp.35-42，2009
9) 伊藤貴，奥山洋平，加藤正治，小坂哲夫，好田正紀, “ 話者クラス音響モデルを用いた講演音声認識の性能向上 ”，信学技報，SP2009-42，pp.7-12，2009.7
10) K.Maekawa, “ Corpus of Spontaneous Japanese: Its design and evaluation ”，Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)，pp.7-12，2003
11) 中川聖一, “ パターン情報処理 ”，丸善株式会社，1999
12) 芹澤伸一，馬場朗，松浪加奈子，米良祐一郎，山田実一，李晃伸，鹿野清宏, “ 十分統計量と話者距離を用いた音韻モデルの教師なし学習法 ”，電子情報通信学会論文誌，Vol.J85-D-Ⅱ，No.3，pp.382-389，2002