Vol.2015-NL-221 No.2
Vol.2015-SLP-106 No.2
2015/5/25

初等教育における授業音声の収集と音声認識の基礎的検討

南條 浩輝[1,a)]　西崎 博光[2,b)]

概要：初等教育における授業音声の音声認識の研究を行う．これまでの授業の音声認識は主に大学などの高等教育における学習支援を対象として行われており，初等中等教育を対象としたものはほとんど行われていなかった．初等教育（小学校）での授業の音声認識とそれを用いた学習・教育支援の研究は社会的意義が大きく重要である．小学校授業での教師発話は児童向けの発話であり，大人向けの発話と音響的にも言語的にも特徴が異なる．さらに，児童向け発話やテキストのデータが十分に整備されていない．このため，音声認識の音響モデルと言語モデルの学習が難しい問題があった．この問題に対し，本研究ではまず小学校授業の収録を行い，次に，そのデータを用いた音声認識の検討を行った．具体的には 44 件の授業音声の収録を行い，これらを用いて音響モデルと言語モデルを学習して，音声認識システムを構築した．小学校授業音声は大人向け発話やテキストだけでのモデル化は難しいこと，および実際の授業データを用いて学習することの有効性を確認できた．

1. はじめに

会議や学術講演，大学講義などの音声を保存した上で，字幕や話者情報などのメタデータを付与して保存・活用しようとする音声ドキュメント処理の研究が盛んに行われている [1][2][3][4]．大学などの高等教育現場においては，このような音声ドキュメント処理技術について多くの研究が試みられている [5][6][7][8]．一方，初等中等教育においては，ICT（Informationand Communication Technology）の活用による学校教育の情報化が推進されている [9] ものの，音声ドキュメント処理に基づく教材作成などの試みはほとんど行われていない．このような背景に基づき，本研究では，初等中等教育における授業音声，特に低年齢向けの授業音声（小学校授業）を対象とし，音声ドキュメント処理の重要な基礎技術である音声認識について研究を行う．

我々は，初等中等教育における授業（以下，子供向け授業）の音声認識について研究をこれまでにも行っている．教師発話の音響的特徴のモデル化（音響モデル）の研究 [8] および，言語的特徴のモデル化（言語モデル）の研究 [10][11][12][13][14] などを，2009 年度に収録した 14 件の初等中等教育（特別支援学校含む）における授業データを用いて行ってきた．ただし，データ数が少ないことと，収録時の問題（音声が振り切れている/小さすぎる，残響音が大きいなど）があることが問題であった．

これに対し，2013 年度に新たに 44 件の小学校授業の収録を行った．本稿ではまずこのデータについて述べる．次にこのデータを用いた音声認識のための音響・言語モデルの学習とそれを用いた音声認識の結果について述べる．

2. 初等中等教育における授業音声の特徴とモデル化の難しさ

初等中等教育においては，授業は児童・生徒（以下，子供）が理解できるように進められる．子供に馴染みがない語（難しい語）は基本的に使用されず，呼びかけや確認が多い傾向がある．

子供向け授業音声は話し言葉ではあるものの，大人向けの話し言葉（大学講義，学術講演，会議）とは言語的にも音響的にも大きく異なっており，大人向けテキストから学習した言語モデルおよび大人向けの発話で学習された音響モデルを子供向け授業の音声認識に用いるのは難しい [8][14]．実際に，呼びかけ調の話し言葉のモデル化の問題や発声方法に起因する音響的な問題，具体的には発話速度，発話速度やパワーの変動，強調発声（hyper articulate）の問題が大きそうなことを確認している [14]．

また，いずれの研究も，子供向けの言語表現や子供向けの発話（授業データ）が整備されていない問題（利用できる授業データが少ない問題）が大きかった．これに対し，本研究では新たに授業データを収集し，それを用いることで，子供向け授業をどの程度モデル化できるかについて検討を行う．

[1] 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku University
[2] 山梨大学
University of Yamanashi
[a)] nanjo@rins.ryukoku.ac.jp
[b)] hnishi@yamanashi.ac.jp

表 1 収録した授業音声（44 授業，14 名，26.3 時間）の内訳

	男性	女性
低学年	0 名，0 授業，0 時間	5 名，16 授業，10.4 時間
中学年	2 名，6 授業，2.5 時間	2 名，7 授業， 5.0 時間
高学年	3 名，9 授業，4.3 時間	2 名，6 授業， 4.1 時間
合計	5 名，15 授業，6.8 時間	9 名，29 授業，19.5 時間

表 2 収録した授業内容の内訳

	算数	国語	社会	理科	生活	道徳	総合（英語）	図工
低学年	6	6	0	0	2	2	0	0
中学年	3	3	2	3	0	1	1	0
高学年	5	3	3	3	0	0	0	1
合計	14	12	5	6	2	3	1	1

3. 初等教育授業データの収集

3.1 授業データの内訳

2013 年度に山梨県内の小学校の協力を得て，小学校における授業音声の収集を行った．音声収録はピンマイク（ソニー ECM-CS10）を用いて 48kHz，16 ビット量子化，ステレオで行ったのちに 16kHz, 16bit モノラル音声にダウンサンプリングして保存した．

1 名あたり 2 件から 4 件の授業音声を収録し，男性と女性教員あわせて 14 名による 44 の授業（26.3 時間）を収録した．小学校低学年，中学年，高学年の授業音声を偏らないように収録した．収録した授業音声の内訳を表 1 を示す．男性教員による低学年向け授業は収録されていないが，それ以外は複数名による複数授業のデータが収録されている．男性教員授業は，5 名，15 授業分（6.8 時間），女性教員授業は，9 名，29 授業分（19.5 時間）であり，女性教員の授業が多くなっている．

次に，授業内容の内訳を表 2 に示す．算数と国語の授業は各学年 3〜6 件の合計 10 件以上，理科と社会（生活を含む）は各学年 2〜3 件収録されている．その他，道徳，総合（英語），図工の授業も 1〜3 件収録されている．

3.2 授業音声データの分析

まず，44 件（26.3 時間）の授業データを無音区間に基づいて区切った．このように無音で区切られた音声区間を本稿では発話とよぶ．この各発話に対して，書き起こしテキストを付与した．さらに，日本語話し言葉コーパス（CSJ: Corpus of Spontaneous Japanese）[15] の書き起こし基準 [16] を参考にタグをつけた．本稿では，これを「授業コーパス 2013」と記述する．収録授業データには，子供発話や課題待ち時間などで教師が発話していない時間が相当数含まれる．本コーパスの中でも実際に教師発話が含まれる部分は 12.8 時間分であった．本稿では，この 12.8 時間分の発話音声を用いて種々の音声認識の実験を行った結果について述べる．

本節では，まず各発話にどのようなタグ（ここでは発話イベントとよぶ）が付与されたかを調べた．具体的には，低学年，中学年，高学年ごとに，各発話イベントを含む発話数と全発話に対する割合を調査した．結果を表 3 に示す．

母音の引き延ばしは，通常よりも長く発声される長音，例えば「でーーーーーはー」の「でー」などにつけられている．母音の引き延ばしは，平均的には全発話の 20%〜30%に含まれることがわかった．授業ごとにみると，0%〜10%に含まれている授業が 3 件，10%〜20%に含まれている授業が 8 件，20%〜30%に含まれている授業が 12 件，30%〜40%に含まれている授業が 11 件，40%〜50%に含まれている授業が 8 件，それ以上含まれている授業が 2 件（百人一首の授業では 90%），となっており，10%〜50%程度の発話で引き延ばし発声が行われていることがわかった．

固有名詞については 6%〜10%程度の発話に含まれており，ほとんどが児童の名前（呼びかけ）である．授業ごとのばらつきも大きく，数%程度のものもあれば 20%を超えるものもある．

フィラー（えー，んーと，などの有声休止）については全体の 5%から 10%程度，言い直しについては全体の 3%から 6%程度であった．

不明瞭な発話については低学年で多いことがわかった．この原因としては，授業スタイルが録音環境とマッチしていない可能性や，そもそも低学年の子供向け発話は特殊であるために人間にとっても聞きにくかったという可能性が考えられる．

その他，ピンマイクで録音しているものの，小学校授業では教師と児童のやりとりが多く，教師発話中に児童音声が背景音として含まれていることが多いことがわかった．

これらの特徴，特に子供向け授業に起因する特徴については，今後，詳細な分析および大人向け授業との比較などを通じ，明らかにしていきたい．

4. 小学校授業音声の認識のための音響モデル

乳幼児に向かって大人が話す場合に，特殊な発話（infant/child-directed-speech（IDS/CDS）と呼ばれる発話 [17]）を行う場合があることが知られている．典型的には，高めの声，抑揚が大きいなどの特徴を含む発話であ

表 3 授業コーパス 2013 における発話イベントを含む発話数とその割合

	全発話	母音引き伸ばし	フィラー	固有名詞	言い直し	不明瞭
低学年	6649	2229 (33.52%)	329 (4.95%)	529 (7.96%)	172 (2.59%)	761 (11.45%)
中学年	3368	651 (19.33%)	389 (11.55%)	356 (10.57%)	200 (5.94%)	205 (6.09%)
高学年	4984	1498 (30.06%)	549 (11.02%)	286 (5.74%)	174 (3.49%)	167 (3.35%)

る．子供向けの教師発話は乳幼児向けの発話ではないものの，大人向けの発話と大きく特徴が異なっており，IDS/CDSと同じような特徴を含んでいる可能性がある．どのような特徴が一致する，または異なるといった分析は今後行っていきたい．

このように特殊な発声が行なわれる子供向け授業の音声は，大人向けの発話から学習した音響モデルでは適切に認識できない可能性が高く，子供向け授業音声のための音響モデルの検討が必要である．本節では，大人向けの発話から学習した音響モデルと子供向けの発話から学習した音響モデルを用いて，子供向け授業音声のための音響モデルの検討を行う．

4.1 日本語話し言葉コーパス（CSJ）からの音響モデル学習

音声認識においては，認識対象音声と類似した音声を学習データとして音響モデルを学習することが有効である．授業音声は話し言葉であるため，話し言葉でモデル学習を行うことが重要である．話し言葉の音声を大量に集めたコーパスとしてCSJがある．これは大人向けに話された講演音声などからなるコーパスである．CSJと小学校授業音声は話し言葉という観点では一致しているものの，発話の対象が大人/子供と異なる．このため，小学校の授業音声認識の音響モデルのためにはCSJ全体ではなく，CSJの中でも比較的話し方が対象となる教師音声に近い音声のみで学習することが重要といえる．実際に我々は，CSJ全体を用いるよりも一部の講演を選択して音響モデルを学習する有効性を確認している [8]．

このような背景に基づき，本研究では，認識対象の教師音声ごとに，教師音声と音響的特徴の類似度が高いCSJの講演音声を選択し，そこから音響モデルを学習する [8] ことを検討する．具体的には，CSJの各講演と認識対象授業音声それぞれに対しモノフォン音響モデルを作成し，授業音声の母音モデルとバタチャリア距離の近い母音モデルをもつCSJの講演音声を選択して学習する．

バタチャリア距離が小さい順にCSJ講演を60講演集め，GMM-HMM音響モデルを学習した．各時刻（フレーム）ごとに39次元（MFCC(12)+power(1)+ΔMFCC(12)+Δpower(1)+ΔΔMFCC(12)+ΔΔpower(1)）の特徴量を取り出し，音節ごとに3状態left-to-rightのGMM-HMM（各状態32混合）を学習した．音節数は134（無音モデルも含む），フレーム幅は25ミリ秒，フレームシフトは10ミリ秒とした．

学習データ量については，CSJ（60講演）は少なめに見積もって10時間（10分 × 60講演），平均的には13時間程度（2702講演を600時間として単純に60講演分の時間）と考えることができ，授業コーパス2013（12時間程度）と同程度と考えられる．

表 4 音響モデル評価用テストセット: 2009年度収録授業

ID	科目	教師性別	発話数
AF1	国語	女性	202発話
AF2	社会	女性	135発話
AM1	社会	男性	338発話
AM2	SHR	男性	46発話

4.2 小学校授業からの音響モデル学習

認識対象と同じ発話スタイルである授業コーパス2013の教師発話から音響モデルを学習する．ここでは2種類のモデルを学習する．

4.2.1 GMM-HMM音響モデル

フレームごとに39次元の特徴量を取り出し，音節ごとに3状態left-to-rightのGMM-HMM（各状態32混合）を学習した．音節数は134（無音モデルも含む），フレーム幅は25ミリ秒，フレームシフトは10ミリ秒とした．

4.2.2 DNN-HMM音響モデル

DNNの入力は，各フレームで，対象フレームと前後5フレームを合わせた11フレーム分の429次元の音響特徴量（39次元 × 11）とした．隠れ層は8層とした．隠れ層の次元数は512，出力層の次元数は402（= GMM-HMMの状態数 = 134 × 3）とした．

4.3 連続音節認識による音響モデルの比較

連続音節認識を行って，学習した3種類の音響モデルの比較を行う．具体的には，学習した各種音響モデルと音節3-gram言語モデルおよびWFST版SPOJUSデコーダ [18] を用いて音声認識システムを構成し，連続音節認識を行うことで比較を行う．

本研究では，認識対象の子供向け授業音声として2009年度に収録した4件の授業を用いた．テストセットの詳細を表4に示す．

4.3.1 子供向け発話で学習した音響モデルの評価

大人向け発話（CSJの一部）で学習したGMM-HMM音響モデルと子供向け発話（授業コーパス2013）で学習したGMM-HMM音響モデルを用いて音声認識（連続音節認識）を行った．結果（音節正解精度）を表5に示す．

CSJ中の類似60講演から学習したGMM-HMM音響モデルでの音節正解精度は平均5.6%（最大でも20%）と低い．これに対し，授業コーパス2013で学習したGMM-HMM音響モデルを用いた場合は，CSJから学習した音響モデルを用いる場合よりも高かった．CSJ（大人向け発話）から学習した音響モデルよりも，同程度の学習データ量の小学校授業データ（子供向け発話）から学習した音響モデルを用いたほうが，高い音声認識精度を得られることが確認できた．なお我々は，子供向け授業の音声認識において，CSJ全体から学習した音響モデルよりも，CSJの一部から学習した音響モデルを用いる有効性を確認している [8]．

これらのことは，小学校授業音声認識のための音響モデ

表 5 音響モデルの評価（連続音節認識時の音節正解精度 (%SylAcc.)）

AM 学習データ		CSJ（60 講演）	授業コーパス 2013	
AM タイプ		GMM-HMM	GMM-HMM	DNN-HMM
テストセット	AF1	-24.2	8.8	22.4
	AF2	13.3	30.0	45.0
	AM1	20.1	28.2	41.8
	AM2	13.2	21.3	34.2
	平均	5.6	22.1	35.9

ル学習には，大人向けの発話をたくさん集めても効果はなく，子供向け発話を集めることが必要であることを示しており，小学校授業音声が大人向け発話と異なる特徴を持つことを示している．

4.3.2 GMM-HMM 音響モデルと DNN-HMM 音響モデルの比較

次に，音響モデルのモデルタイプの比較を行った．具体的には子供向け発話（授業コーパス 2013）で学習した GMM-HMM 音響モデルと DNN-HMM 音響モデルの比較を行った．結果は表 5 に示されている．DNN-HMM を用いた場合は，GMM-HMM よりも高い音声認識精度が得られた．実際の子供向け授業音声を集めて DNN-HMM 音響モデルを学習するのが効果的であることがわかった．

4.3.3 話者クローズド DNN-HMM 音響モデルを授業音声の認識

次に，授業コーパス 2013 の授業音声の認識について述べる．授業コーパス 2013 中の 4 件の授業を選び音声認識のテストセットとした．テストセットを表 6 に示す．

これらの 4 つの授業を除いた 40 授業で DNN-HMM 音響モデルを学習して，連続音節認識を行った．4 名の話者の別の授業音声は学習データに含まれているため，音響モデルは話者クローズドなものとなっている．実際に各教師は何度も授業を行うため，このような音響モデル学習は授業音声認識システム構築において十分に実現可能といえる．

結果を表 7 に示す．認識率は平均 42%（各授業 54%，58%，52%，32%）であり，2009 年度授業データの音声認識率と比べて高い．当該話者の音声を含めて音響モデルを学習することで，高精度な音響モデルが実現できることがわかる．

ただし，まだ認識率は十分でない．この主な原因として，子供向け発話の特徴のモデル化がまだ不十分であること，および，実際の授業における教師発話の収録においてはクリーンな収録環境を得ることができず，教師発話に背景音として児童音声や残響音が混入することを避けられないこと，が挙げられる．実際に，テストセット M1 の授業では，教師と児童が同時に英語を話す，児童同士が会話をするなどが行われており，教師音声への児童音声の混入が大きい問題が顕著である．

今後は，子供向け授業発話の特徴の分析とそのモデル化を進めるとともに，教師発話と背景音の分離の技術（教師発話の明瞭化）を進めていく予定である．

表 6 授業コーパス 2013 テストセット

ID	対象	科目	教師性別	発話時間	発話数
F1	低学年	国語	女性	28.6 分	4897
F2	低学年	算数	女性	17.2 分	2619
F3	低学年	国語	女性	17.4 分	2612
M1	中学年	総合（英語）	男性	9.2 分	1008

表 7 DNN-HMM 音響モデルを用いた授業コーパス 2013 テストセットの連続音節認識結果（音節正解率（%SylCorr.）と正解精度 (%SylAcc.)）

ID	Corr.	Acc.
F1	63.7	54.2
F2	67.9	58.2
F3	72.1	52.0
M1	39.4	32.2
平均	60.8	49.2

5. 授業音声の認識のための言語モデル

次に授業音声のための言語モデルについて述べる．子供向けの教師発話は大人向けの発話とは言語的特徴も大きく異なる．例えば，呼びかけ表現が多く使われる，難しい語はあまり使用されない，子供向けの表現が用いられるなどの特徴があり，これらの特徴をモデル化する必要がある．さらに，小学校授業は話し言葉であるため，言語モデルにおいて話し言葉表現もモデル化する必要がある．

このような背景に基づき，本研究では，CSJ，子供向け WEB サイトのテキストおよび授業データの書き起こしを用いた言語モデルの学習を検討する．

5.1 CSJ と子供向け WEB サイトを用いた言語モデル

小学校授業における教師発話は子供向け表現を含む話し言葉である．したがってこの両者を同時にモデル化する必要がある．

我々はこれまでに子供向け表現の学習テキストコーパスとして，「子供向け WEB サイトコーパス」を収集している（表 8）[10][11]．しかし，このコーパス中のテキストは，子ども向けに書かれてあるものの，文体は基本的に書き言葉である．このため，小学校授業での話し言葉的な表現は十分にモデル化できない．話し言葉表現の学習テキストコーパスとしては，CSJ がある．ただし，これは子供向け表現を含んでいない．

我々は，子供向け授業の音声認識のための言語モデル学習にこの両コーパスを用いる有効性を確認している [11]．まずこの CSJ と子供向け WEB サイトを用いた言語モデルについて述べる．

子供向け WEB サイトコーパスと CSJ では，同じ表現であってもそれぞれ，ひらがなと漢字が使われるなど表記が異なる．この両者をうまく融合させるには，表記の統一が重要である．「かな」を漢字に変換するよりも漢字を「か

表 8 子供向け WEB サイトコーパス

収集元	NHK 週刊こどもニュース	Yahoo!きっず ニュース
件数	475 件	2507 件
テキストサイズ	427k 単語	733k 単語
収集期間	2006/1〜2010/12	2010/9〜2011/6

な」に変換するほうが誤りが少ないと考える．また，音声認識結果を将来的に児童が利用する場合は，未学習の漢字が含まれないほうが望ましい．このような背景から，本研究では全ての漢字を「かな」に変換して「かな」表記の語彙を作成し，言語モデルを学習する．

本研究では，形態素（厳密には異なるが，以後，本稿では単語という）を単位とする単語 3-gram 言語モデルを学習する．単語とそのよみは，Chasen-2.4.4+Unidic-1.3.12 を用いて決定する．その際，地名および人名は 1 つのクラスとしてモデル化する．

CSJ と子供向け WEB サイトコーパスの併用においては，テキストベースでコーパスを混合するとサイズの大きいコーパス（CSJ）での単語 N-gram の出現カウントの影響が大きいため，それぞれのコーパスで言語モデルを学習し, それらを確率ベースで補間する [19]（CSJ:WEB=8:2）．

このようにして，言語モデルエントリ数 19871 の単語 3-gram 言語モデル（CSJ+WEB_LM）を学習し，この言語モデルを用いてテストセット（表 6）のパープレキシティ（PP），補正パープレキシティ（APP），未知語の数（#OOVs）を調べた．結果を表 9 にまとめる．テストセット F1，F2，F3 に対しては未知語率（#OOVs/#word）は 2%未満であり，これまでの別の小学校授業（2009 年収録授業）を対象とした実験 [12] でのカバー率とほぼ同等である．APP の値もほぼ同等である．テストセット M1 は未知語率が 25%と高く，APP も高い．これは総合（英語）の授業であることに起因する．

5.2 2013 授業音声コーパスを用いた言語モデル

次に実際の授業音声の書き起こしテキストを使った言語モデルの学習を検討する．具体的には 2013 授業音声コーパスのうち，テストセットの 4 つの授業を除いた 40 授業の書き起こしから言語モデルを学習し，その評価を行う．40 授業の書き起こしテキストサイズは 115,539 単語（無音モデル除く．含む場合は 150,169）であり，データ量は非常に少ないものの，実際の授業における言語表現は多く含まれる．また本実験では，話者クローズドのモデルにもなっている．実際に各教師は何度も授業を行うため，このような話者クローズドの言語モデル学習は授業音声認識システム構築において十分に実現可能といえる．

この 40 授業の書き起こしテキストから，言語モデルエントリ数 5221 の単語 3-gram 言語モデル（School_LM）を学習し，この言語モデルを用いてテストセット（表 6）のパープレキシティ，補正パープレキシティ，未知語の数を調べた．

表 9 CSJ と WEB 言語モデル（CSJ+WEB_LM）によるテストセットパープレキシティと未知語数

	#word	#OOVs（種類）	PP	APP
F1	4897	96 (48)	622.7	658.3
F2	2619	31 (25)	721.8	740.1
F3	2612	42 (33)	282.4	296.9
M1	1008	249 (62)	616.7	1015.3

表 10 2013 授業音声言語モデル（School_LM）によるテストセットパープレキシティと未知語数

	#word	#OOVs（種類）	PP	APP
F1	4897	92 (58)	157.2	170.7
F2	2619	58 (27)	140.6	153.5
F3	2612	54 (45)	96.1	106.1
M1	1008	262 (64)	218.4	574.4

表 11 混合言語モデル（CSJ+WEB+School_LM）によるテストセットパープレキシティと未知語数

	#word	#OOVs（種類）	PP	APP
F1	4897	34 (21)	201.1	207.6
F2	2619	8 (8)	191.1	192.8
F3	2612	21 (18)	121.7	126.0
M1	1008	227 (48)	359.6	841.5

表 10 に結果を示す．未知語率は CSJ+WEB_LM とほぼ同等であることがわかる．PP および APP は低いが，これは言語モデルのエントリサイズ，および学習データが小さいことに起因するものである．CSJ+WEB_LM との補正パープレキシティの比較から，音声認識にとってどちらが適しているかを判断するのは難しいと考える．

5.3 CSJ+WEB+2013 授業音声コーパス混合言語モデル

次に，CSJ と子供向け WEB サイトコーパス，授業書き起こしすべてを用いて言語モデルを学習することを考える．

ここでも，それぞれのコーパスで言語モデルを学習し, それらを確率ベースで補間する．具体的には，CSJ_LM と子供向け WEB サイト LM を 8:2 で混合した言語モデル（CSJ+WEB_LM）と授業書き起こしから学習した言語モデル（School_LM）を 5:5 で混合する．

こうして，言語モデルエントリ数 20825 の単語 3-gram 言語モデル（CSJ+WEB+School_LM）を学習した．この言語モデルを用いてテストセット（表 6）のパープレキシティ，補正パープレキシティ，未知語の数を調べた．

表 11 に結果を示す．未知語率が大きく改善し，F1, F2, F3 では未知語が半分以下となった．M1 に対しても未知語率が減少していることがわかる．CSJ+WEB と授業音声書き起こしに含まれる単語が異なり，うまく語彙をカバーしていることがわかる．APP も CSJ+WEB_LM に比べて低くなっており，極端に語彙サイズが小さい School_LM の値と近いこともわかる．ただし M1 については，総合（英

表 12 各言語モデルによるテストセットの n-gram カバレージ

	LM	3-gram	2-gram	1-gram
F1	CSJ+WEB	38.6%	39.9%	21.5%
	School	37.2%	35.5%	27.4%
	CSJ+WEB+School	48.3%	35.7%	16.0%
F2	CSJ+WEB	48.2%	35.9%	15.9%
	School	50.3%	31.6%	18.1%
	CSJ+WEB+School	61.1%	29.2%	9.7%
F3	CSJ+WEB	55.9%	31.1%	13.1%
	School	50.5%	28.9%	20.6%
	CSJ+WEB+School	65.3%	25.3%	9.4%
M1	CSJ+WEB	47.3%	31.5%	21.2%
	School	37.9%	27.5%	34.6%
	CSJ+WEB+School	50.6%	29.6%	19.9%

語）の授業であるため，日本語テキストにはほとんど出現しない表現が含まれており，本コーパスからのモデル化は難しかった．

次に，3 つの言語モデルによるテストセットの n-gram カバレージも比較した．結果を表 12 に示す．CSJ+WEB+School_LM が，どの授業データに対しても最も高い 3-gram のヒット率を示している．School_LM 単体では，APP の値は小さいものの 3-gram, 2-gram ヒット率は低いことがわかる．

これらことは，小学校授業の音声認識のための言語モデル学習には，実際の小学校授業の書き起こしを用いることが有効であること，実書き起こしデータが大量に得られないときは，CSJ や子供向け WEB サイトのテキストなどを併用することが有効であることを示唆している．

6. 小学校授業音声の音声認識

最後に，2013 授業音声コーパスを用いて作成した音響モデルおよび言語モデルを用いて大語彙連続音声認識を行い，小学校授業音声認識システムおよび各種モデルの評価を行う．

音響モデルには，4 節で述べた 2013 授業音声コーパスから学習した DNN-HMM 音響モデル（話者オープンモデルと話者クローズドモデル）を用いる．言語モデルには，5 節で述べた CSJ+WEB 言語モデルと CSJ+WEB+2013 授業音声コーパスから学習した言語モデルを用いる．デコーダには WFST 版 SPOJUS を用いる．なお，CSJ+WEB+School_LM を用いる場合の認識用辞書には授業コーパス 2013 中の固有名詞が含まれており，CSJ+WEB_LM を用いる場合の認識用辞書には含まれていない．

音声認識結果を表 13 と表 14 に示す．CSJ+WEB だけで学習した言語モデルよりも，School_LM を併用するほうが高い認識率を得られることがわかる．ただし，辞書中の固有名詞が異なるため，この結果は純粋な言語モデルの性能比較とはなっていない．

表 13 連続単語音声認識結果（音響モデルオープン）

	CSJ+WEB		CSJ+WEB+School	
	Corr.	Acc.	Corr.	Acc.
F1	45.0	37.6	50.3	43.0
F2	44.8	36.6	52.6	45.0
F3	56.0	43.7	59.5	46.7
M1	13.3	4.1	13.7	5.0

表 14 連続単語音声認識結果（音響モデルクローズ）

	CSJ+WEB		CSJ+WEB+School	
	Corr.	Acc.	Corr.	Acc.
F1	53.3	46.7	59.4	53.4
F2	57.0	50.1	63.2	56.6
F3	67.7	59.9	72.9	64.4
M1	18.2	11.6	20.1	13.3

話者クローズな音響モデルと CSJ+WEB+School で学習した言語モデルにより，テストセット F1，F2，F3 に対して 60%程度の認識精度が得られた．テストセット M1 については，英語を扱う授業であることに起因する言語・音響両モデルのミスマッチと背景音の混入が大きな悪影響を及ぼしており，低い認識率となったと考えられる．

今後は，様々な授業データでの音声認識実験と誤り分析を行い，音声認識を困難にしている諸要因を明らかにしていく予定である．また学年や教科ごとの差異などの分析も行っていく予定である．

7. おわりに

初等教育（小学校）での授業の音声認識の研究を行った．実際の小学校授業音声を 44 件収録し，それを用いて音声認識のモデルの学習を試みた．小学校授業音声は大人向け発話やテキストだけでのモデル化は難しいこと，および実際の授業データを用いて学習することの有効性を確認できた．今後は，小学校授業音声の特徴や授業音声の認識を困難にしている諸要因を明らかにしていく予定である．

謝辞 WFST 版 SPOJUS は Google Inc. の藤井康寿氏に提供しただいた．深く感謝します．本研究は科研費「24500225」「15K00254」の助成を受けた．

参考文献

[1] Ferdiansyah, V. and Nakagawa, S.: Automatic Speech Recognition and Machine Translation System for MIT English Lectures using MIT and TED Corpus, 第 8 回音声ドキュメント処理ワークショップ，SDPWS2014-01 http://www.cl.ics.tut.ac.jp/~sdpwg/sdpws2014_proceedings/ (2014).

[2] 今井　亨，小林彰夫，佐藤庄衛，本間真一，奥　貴裕，都木　徹：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講演論文集，pp. 113–120 (2008).

[3] 西崎博光，杉本樹世貴，関口芳廣：音声ドキュメント内容検索のための WEB を用いたドキュメント拡張，情報処理学会論文誌， Vol. 52, No. 12, pp. 3461–3470 (2011).

[4] 西尾友宏，南條浩輝，吉見毅彦：講演音声ドキュメント

検索のための擬似適合性フィードバック，情報処理学会論文誌， Vol. 55, No. 15, pp. 1573–1584 (2014).
[5] 桑原暢弘，秋田祐哉，河原達也：音声認識結果の有用性の自動判定に基づく講義のリアルタイム字幕付与システム，第 8 回音声ドキュメント処理ワークショップ, SDPWS2014-02 http://www.cl.ics.tut.ac.jp/~sdpwg/sdpws2014_proceedings/ (2014).
[6] 勝浦広大, 桂田浩一, 入部百合絵, 森本容介, 辻　靖彦, 青木久美子，新田恒雄：放送大学の講義音声を対象とした高速キーワード検索の性能評価，第 6 回音声ドキュメント処理ワークショップ，SDPWS2012-05 http://www.cl.ics.tut.ac.jp/~sdpwg/sdpws2012_proceedings/ (2012).
[7] 中川聖一，富樫慎吾，山口　優，藤井康寿，北岡教英：講義音声ドキュメントのコンテンツ化と視聴システム，電子情報通信学会論文誌， Vol. J91-D, No. 2, pp. 238–249 (2008).
[8] 穂坂圭一，伊藤信義，西崎博光，関口芳廣：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，第 4 回音声ドキュメント処理ワークショップ, SDPWS2010-02 http://www.cl.ics.tut.ac.jp/~sdpwg/sdpws2010_proceedings/ (2010).
[9] 文部科学省：教育の情報化, http://jouhouka.mext.go.jp/school/.
[10] 久木一平，南條浩輝：小学校授業の音声認識のための児童向けサイトを用いた言語モデルの構築，日本音響学会研究発表会講演論文集，1-10-17 秋季 (2011).
[11] 南條浩輝，久木一平，和田祐樹：初等中等教育における授業音声認識のための言語モデルの検討，電子情報通信学会技術研究報告，SP2011-54 (WIT2011-36), pp. 13–18 (2011).
[12] 南條浩輝，久木一平，和田祐樹：初等中等教育の授業音声認識のための子供向け表現の抽出と言語モデル学習，日本音響学会研究発表会講演論文集，3-P-19 秋季 (2012).
[13] 南條浩輝，谷奥大喜：初等中等教育授業における教師発話の言語的特徴のモデル化のための学習データ選択方法の検討，第 12 回情報科学技術フォーラム（FIT2013），E-031, pp. 257–258 (2013).
[14] 南條浩輝，堀　智織：初等中等教育の授業を対象とした音声認識の基礎的分析，日本音響学会研究発表会講演論文集，2-P-32 秋季 (2013).
[15] K.Maekawa: Corpus of Spontaneous Japanese: Its Design and Evaluation, *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12 (2003).
[16] 小磯花絵，前川喜久雄：『日本語話し言葉コーパス』の設計の概要と書き起こし基準について，情報処理学会研究報告，NL-143, pp. 41–48 (2001).
[17] 村瀬俊樹，小椋たみ子，山下由紀恵：養育者における育児語使用傾向の構造と育児語使用を規定する要因，社会文化論集：島根大学法文学部紀要社会文化学科編， Vol. 4, pp. 17–30 (2007).
[18] 関　博史，中川聖一：音節単位 DNN-HMM による音声認識の検討，情報処理学会研究報告，2013-SLP-99, No. 4 (2013).
[19] 長友健太郎，西村竜一，小松久美子，黒田由香，李　晃伸，猿渡　洋，鹿野清宏：相補的バックオフを用いた言語モデル融合ツールの構築，情報処理学会論文誌， Vol. 43, No. 9, pp. 2884–2893 (2002).