Vol.2017-MUS-115 No.54
2017/6/18

録音環境に頑健な授業音声認識のための音声コーデックとその活用の検討

南條 浩輝[1,a)]　西崎 博光[2,b)]　高橋 徹[3,c)]

概要：授業音声の利活用のための音声認識の研究を行う．これまでの授業の音声認識は大学などの収録設備が整えられた教室で録音された音声を対象として研究が行われていた．我々は，初等・中等教育での授業の音声認識とそれを用いた学習・教育支援の研究を推進してきており，収録設備が整えられていない教室で簡易的な録音デバイスで収録された音声を扱うことの重要性を指摘している．本稿では録音環境に頑健な授業音声認識のための検討を行ったのでその結果について報告する．

1. はじめに

音声に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であり [1][2][3][4][5][6][7][8][9][10][11][12][13]，我々はこれらを用いた授業音声の利活用を研究している．

これまでの授業音声の利活用の研究 [14][15][16] は，主に十分な収音設備が整った教室で収録された音声を対象としたものであり，小・中学校高等学校などの一般の教室において簡易デバイスを用いて収録した授業音声を対象とした研究は充分でない．

このような背景に基づき，本研究では，多様な録音環境で収録された授業音声を利活用するための基礎的技術を研究する．本稿では，様々な音声コーデックで収録された授業音声の扱いを検討する．具体的には，我々が有する小学校授業音声データベースの授業音声に様々な音声コーデックで変換を施して擬似的な種々の収録環境音声を生成し，それらの扱いを検討する．具体的には，授業音声の利活用の基礎技術である音声認識について検討を行う．

2. 授業音声の特徴とモデル化の難しさ

我々は，これまでに小学校での授業音声を扱ってきた [17][18][19][20][21][22]．小学校での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 [17][22]．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 [22]．さらに，少量ではあるものの小学校授業（子供向け発話）で学習した音響モデルを用いたほうが，大人向け発話で学習した音響モデルを用いるよりも性能が高いことも確認している [23][24]．ただし，そのような授業音声の音響モデルの作成にあたっては，学習データが少ないこと，および，学習データと異なる収録環境で収録された音声に対する頑健性の点で問題があると考えられる．

本研究では，既に有する小学校授業音声データベースの授業音声に様々な音声コーデックで変換を施して擬似的な種々の収録環境音声を生成し，それらを用いた音響モデル学習の効果を検討する．

3. 初等教育授業データ

3.1 授業データの内訳

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．

収録した授業音声の内訳は表 1 のとおりである．合計 44 授業，26.3 時間である．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

[1] 京都大学学術情報メディアセンター Academic Center for Computing and Media Studies, Kyoto University
[2] 山梨大学 University of Yamanashi
[3] 大阪産業大学 Osaka Sangyo Univ.
[a)] nanjo@media.kyoto-u.ac.jp
[b)] hnishi@yamanashi.ac.jp
[c)] takahashi@ise.osaka-sandai.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　擬似的に作成した音声データ

original	16kHz, 16bit wav （以下 wav）
aac1_128	wav → aac 128kbps → wav
aac1_64	wav → aac 64kbps → wav
aac1_32	wav → aac 32kbps → wav
aac1_16	wav → aac 16kbps → wav
aac2_128	wav → aac 128kbps → wav → aac 128kbps → wav
aac2_64	wav → aac 64kbps → wav → aac 64kbps → wav
aac2_32	wav → aac 32kbps → wav → aac 32kbps → wav
aac2_16	wav → aac 16kbps → wav → aac 16kbps → wav
vorbis1_64	wav → vorbis 64kbps → wav
vorbis1_32	wav → vorbis 32kbps → wav
vorbis1_16	wav → vorbis 16kbps → wav
vorbis2_64	wav → vorbis 64kbps → wav → vorbis 64kbps → wav
vorbis2_32	wav → vorbis 32kbps → wav → vorbis 32kbps → wav
vorbis2_16	wav → vorbis 16kbps → wav → vorbis 16kbps → wav

表 3　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

まれており，実際に教師発話が含まれる部分は 12.8 時間分である．本研究では，この 12.8 時間分の発話音声を用いて種々の音響モデルの学習の実験を行う．

4. 様々な音声コーデックによる擬似的な多環境音声の作成

音声コーデックとして，AAC（Advanced Audio Coding）と Vorbis を使用した．これらは，ともに非可逆圧縮手法であり，聴覚心理に基づいて音声データ中の人間の聞こえに影響しない成分を符号化しないものである．録音機では長時間録音のために非可逆圧縮アルゴリズムを採用していることがある．録音環境に頑健な音声認識を実現するために，このような圧縮音声の扱いを考える必要がある．音声圧縮を適用することで，人間には聞こえないが，音声認識には影響を及ぼすような雑音成分が取り除かれることも期待できる．

さらに，この圧縮音声を作成することで，元の音声ファイルと内容が同じであるが音声信号が異なるデータを得ることができるため，音声認識の音響モデルの学習データを増やすことができると考えられる．

表 4　音声認識に用いた音響モデル詳細

モデル化単位	音節 (mono syllable)
音節種類数	117 種類 + ショートポーズ
各音節の状態数	母音・撥音・促音は 3 状態，その他は 5 状態
特徴量	MFCC: MFCC(13)+Δ+ΔΔ LDA+MLLT: MFCC(13)x9 フレーム（117 次元）を LDA で 40 次元に圧縮し, MLLT さらに fMLLR を適用した上で，sMBR 学習を 2 回繰り返し
隠れ層	6 層
隠れ層ノード数	1905
入力層ノード数	1400（40 次元 fMLLR x 35 フレーム）

4.1 実験環境

ffmepg[*1]を用いて音声圧縮を行ったのち，その圧縮音声を 16kHz, 16bit の wav に戻して新たな学習データとした．さらに，その圧縮音声に再度音声圧縮と復元を施して，新たな学習データとした．AAC では libfdk-aac を，Vorbis では libvorbis を用いた．用いた音声コーデックとビットレートの一覧を表 2 に示す．

テストデータは，表 3 に示す 4 つの授業であり，それぞれの授業音声認識の音響モデルは，当該授業を除いた全てのデータで学習した．

音声認識システムには Kaldi Speech Recognition Toolkit[25] を使用した．音響モデルの詳細は表 4 のとおりである．

4.2 種々の音響モデルとそれを用いた音声認識

収録音声オリジナルの wav ファイルだけで学習した音響モデル，各圧縮音声（を wav に戻したもの）だけで学習した音響モデル，それら全てで学習したマルチコンディション音響モデルを作成し，種々の音声コーデックで圧縮した音声の認識を行った．言語モデルには音節 3-gram[23] を用いた．

はじめに収録音声オリジナルの wav ファイルだけで学習した音響モデルおよび各圧縮音声だけで学習した音響モデルで，それぞれ一致する条件の音声を音声認識した結果を表 5 に示す．学習データとテストデータともに音声圧縮（64kbps, 128kbps）を行ったのちに音声認識システムを構築することで，音声圧縮を行わない場合に比べて誤りが増加する悪影響はほとんどなく，誤りが減少するケース（図中太字）があることがわかった．人間の聴覚に作用しない背景雑音などのデータを除去した効果が示唆された．また，音声圧縮のビットレートを 32kbps にしてもほとんど

[*1] https://ffmpeg.org/

表 5　テストデータと学習データ一致条件での音節認識結果（%Syllable error rate）

	テストセット ID			
	F1	F2	F3	M1
original	27.70	18.35	16.84	55.26
aac1_128	**27.61**	18.82	**16.54**	**54.68**
aac1_64	**27.21**	18.37	16.88	55.96
aac1_32	28.10	19.65	17.55	56.04
aac1_16	30.93	20.76	19.35	58.80
aac2_128	27.76	18.70	**16.70**	56.29
aac2_64	**27.24**	18.93	**16.80**	**54.97**
aac2_32	28.53	19.25	17.81	57.36
aac2_16	31.38	21.52	19.82	60.54
vorbis1_64	27.83	19.23	**16.48**	**55.01**
vorbis1_32	28.41	19.79	17.23	55.79
vorbis1_16	30.38	21.03	19.80	58.76
vorbis2_64	27.80	19.19	17.41	**55.05**
vorbis2_32	29.25	20.39	18.54	56.78
vorbis2_16	32.90	23.09	17.41	61.07

表 6　マルチコンディション学習音響モデルでの音節認識結果（%Syllable error rate）

	テストセット ID			
	F1	F2	F3	M1
original	26.27	16.85	15.81	52.33
aac1_128	26.25	16.76	15.73	52.04
aac1_64	26.43	16.68	15.73	52.37
aac1_32	27.40	16.82	16.30	54.10
aac1_16	30.83	20.35	21.26	62.89
aac2_128	26.24	16.64	15.77	51.96
aac2_64	26.37	16.68	15.71	52.33
aac2_32	27.50	17.17	16.52	54.85
aac2_16	32.40	21.40	22.15	65.77
vorbis1_64	26.33	16.82	15.97	52.58
vorbis1_32	26.61	17.11	16.13	53.40
vorbis1_16	29.63	19.17	18.42	58.85
vorbis2_64	26.52	16.93	16.07	52.95
vorbis2_32	27.08	18.39	16.74	55.34
vorbis2_16	32.18	21.13	21.28	62.27

表 7　環境オープン音声（オリジナル音声+SS）の音節認識結果（%Syllable error rate）

	テストセット ID	
学習データ	F2	F3
original	18.82	18.70
multi cond.	17.55	18.24

音声認識に悪影響がないが，16kbps では顕著に誤りが増加することもわかった．音声収録にあたってはこれらに留意する必要を指摘できた．

次に，マルチコンディション音響モデルで音声認識した結果を表 6 に示す．音響モデルは全ての圧縮音声で学習し，各テストデータを認識した．16kbps 以外の圧縮率の圧縮音声の認識でマルチコンディション学習の効果が見られ，収録時の音声コーデックの違いに頑健な音声認識システムを学習できた．

さらに，圧縮しない元の音声の音声認識誤りも減っていることがわかる．このことは，種々の音声コーデックを施した音声データを作成して，それらを用いて音響モデルを学習することの有効性を示唆している．本実験では，約 12.8 時間の音声データ[*2]を 15 倍（およそ 190 時間）にして学習しており，学習データ量を補った効果と考えられる．なお，GMM-HMM の学習ではこの学習データを増加させる学習法に効果は見られなかった．sMBR を適用しない DNN の学習でも顕著な効果は見られなかった．

最後に，学習環境にマッチしない音声，具体的には，オリジナル音声に雑音抑圧技術である SS（スペクトルサブトラクション）を適用した音声の認識を行った．オリジナル wav ファイルだけで学習した音響モデルとマルチコンディション学習した音響モデルを用いて認識を行った．結果は表 7 に示されている．ここでもマルチコンディション学習した音響モデルにより認識誤りが減っており，学習データ増加の効果と多様な環境の音声への頑健性が高まったことが示唆されている．

5. おわりに

種々の音声コーデックを施して擬似的な種々の収録環境音声を生成し，それらを用いて音響モデルを学習する効果を調べた．元の音声と圧縮音声の全てを用いた音響モデルの学習を行い，収録時の音声コーデックの違いに頑健な音声認識システムを構築できること，sMBR を行う DNN 音響モデルの学習において効果が見られることがわかった．また，音声圧縮による人間の聴覚に作用しないデータを除去することで音声認識が向上できる可能性を示した．音声収録時において 16kbps に音声圧縮するのは悪影響が大きいこともわかった．

謝辞　本研究は科研費「15K00254」の助成を受けた．

参考文献

[1] Akiba, T., Nishizaki, H., Nanjo, H. and Jones, G. J. F.: Overview of the NTCIR-12 SpokenQuery&Doc-2 Task, *Proceedings of the 12th NTCIR Conference*, pp. 167–179 (2016).
[2] Akiba, T., Nishizaki, H., Nanjo, H. and Jones, G. J. F.: Overview of the NTCIR-11 SpokenQuery&Doc Task, *NTCIR-11 Workshop Meeting*, pp. 350–364 (2014).
[3] Akiba, T., Nishizki, H., Aikawa, K., Hu, X., Itoh, Y., Kawahara, T., Nakagawa, S., Nanjo, H. and Yamashita, Y.: Overview of the NTCIR-10 SpokenDoc-2 Task, *NTCIR-10 Workshop Meeting*, pp. 573–587 (2013).
[4] Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T. and

[*2] 認識対象の授業は学習データから除いている

Matsui, T.: Overview of the IR for Spoken Documents Task, *NTCIR-9 Workshop Meeting*, pp. 223–235 (2011).
[5] 小嶋和徳，紺野和磨，田中和世，李　時旭，伊藤慶明：音声中の検索語検出における同文書内の高順位候補を利用したリスコアリング方式，電子情報通信学会論文誌 D，No. 1, pp. 70–80 (2017).
[6] 小田原一成，山下洋一：音声中の検索語検出における単語共起情報の利用，情報処理学会研究報告，2016-SLP-110, pp. 1–6 (2016).
[7] 大野哲平，秋葉友良：音節継続時間を利用した直線検出に基づく音声検索語検出，情報処理学会論文誌，Vol. 54, No. 2, pp. 484–494 (2013).
[8] 森田直樹，南條浩輝，山本凌紀，馬　青：音声ドキュメントを検索対象とした用語検索，情報処理学会論文誌，Vol. 58, No. 3, pp. 762–767 (2017).
[9] Sawada, N. and Nishizaki, H.: Re-Ranking Approach of Spoken Term Detection using Conditional Random Fields-based Triphone Detection, *IEICE Trans. on Information & Systems*, Vol. E99-D, No. 10, pp. 2518–2527 (2016).
[10] Domoto, K., Utsuro, T., Sawada, N. and Nishizaki, H.: Spoken Term Detection using SVM-based Classifier Trained with Pre-indexed Keywords, *IEICE Trans. on Information & Systems*, Vol. E99-D, No. 10, pp. 2528–2538 (2016).
[11] 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).
[12] 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).
[13] 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).
[14] 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).
[15] 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).
[16] 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).
[17] 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).
[18] 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).
[19] 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).
[20] 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).
[21] 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).
[22] 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).
[23] 南條浩輝，西崎博光：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，情報処理学会研究報告，2015-SLP-106, No. 2 (2015).
[24] 南條浩輝，高橋　徹，西崎博光：初等教育授業音声の利活用のためのアーカイブ技術の基礎的検討，日本音響学会研究発表会講演論文集，3-P-17 春季 (2016).
[25] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K.: The Kaldi Speech Recognition Toolkit, *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (2011).