Vol.2018-SLP-124 No.4
2018/10/10

知識構成型ジグソー法における 中学生発話を対象とした音声認識の試み

長野 徹[1,a)]　東出 紀之[2]　倉田 岳人[1]　立花 隆輝[1]　中山 隆弘[3]　白水 始[3]

概要：知識構成型ジグソー法は，学習者が互いに自分の理解したことを話し合って理解を深めるという協調学習が生じやすい環境を支える枠組みである． この中で，学習者の発話をテキスト化し．理解のプロセスを観察することができれば，学習者の理解度の把握や，それに基づく学習計画の検討が容易になる． 本稿では，本手法に基づいた協調学習における中学生の発話を，音声認識によってテキスト化することで，教員をサポートする試みを紹介する．

キーワード：音声認識, 協調学習, 知識構成型ジグソー法

Trial of Junior High-School Student Speech Recognition in Collaborative Learning, Knowledge Constructive Jigsaw Method

Tohru Nagano[1,a)]　Noriyuki Higashide[2]　Gakuto Kurata[1]　Ryuki Tachibana[1]
Takahiro Nakayama[3]　Hajime Shirouzu[3]

Abstract: Knowledge Constructive Jigsaw Method is a collaborative learning method that has opportunities of engaging both in dialogues to deepen understanding and summative presentations to explain their outcomes. If we succeed in recognizing multi-student speech in timely manner, it helps teachers in real-time monitoring and orchestration of their classes. We conducted three lessons for junior high school students in actual classrooms. The experiments showed that supervised acoustic/linguistic adaptation effieciently improved speech recognition results in accuracy.

Keywords: Speech Recognition, Collaborative Learning, Knowledge Constructive Jigsaw Method

1. はじめに

学習科学は認知科学を基盤として，質の高い学習を導き出そうとする研究分野である．質の高い学習を得るための学習理論の構築には，学習のプロセスを観察し分析することが必要となる．ただ，学習のプロセスは複雑であり，人によってそのプロセスは異なる．同じ事実・教材を用いたとしても，その捉え方は様々である．この理解のプロセスの違いを用いて，各学習者が理解したことを他の学習者と共有し合うことにより，学んだ成果の適用範囲を広げていく学習方法を協調学習とよぶ．知識構成型ジグソー法[*1]は，この協調学習が起きやすい環境を支える授業デザインの枠組みであり，2010 年から全国の教育委員会および学校と連携して，実際の教育現場において授業改善の試みを行っている．授業における学習プロセスの理解のためには，教師や指導員による観察が欠かせないが，教師 1 人による授業中の観察だけでは全ての生徒がどのように理解を進めて

[1] 日本アイ・ビー・エム（株）東京基礎研究所
IBM Reseach - Tokyo
[2] 同社 ソフトウェア＆システム開発研究所
Tokyo Software and Systems Development Lab, IBM Japan
[3] 東京大学 高大接続研究開発センター CoREF ユニット
CoREF, Center for Research and Development on Transition from Secondary to Higher Education, The University of Tokyo
[a)] tohru3@jp.ibm.com

[*1] 東京大学 CoREF 「知識構成型ジグソー法」
http://coref.u-tokyo.ac.jp/archives/5515

いるかを理解することは困難である．各生徒の音声を録音し，音声認識によってその音声をテキスト化することができれば，従来の音声ドキュメント処理技術，自然言語処理技術の適用が容易になり，学習プロセスの理解に役立てることができる．このような背景に基づき，授業音声の音声認識について研究を行う．

授業音声の音声認識としては，学校教育における教師発話の音響モデルの改善 [1]，教師発話の言語モデルの改善 [2] に関する研究，収録音声の詳細な検討に関する研究 [3] およびコーデックによるデータ拡張処理 [4] に関する研究，また子供音声の音声認識としては，音声情報案内システムにおいて収集された音声に含まれる幼児から高学年児童の発話の音声認識の研究 [5] などが行われている．[5] では，「こんにちは」「いま何時ですか」といった比較的単純な発話が多いものの，話者適応の結果，単語認識率が 71.1% と高い精度を達成している．一方，本研究は，授業中の子供音声の発話を対象とする．知識構成型ジグソー法を用いた授業では，子供同士のディスカッションにおける知識の交換に重点が置かれており，通常の授業における生徒の発話（教師の指名による発言，自由な私語）とは異なる授業の内容に関する発話を多く得ることができる．

本稿ではまず，知識構成型ジグソー法について述べる．次に 2017 年に収録された実際の中学校の 3 授業で録音された音声について説明し，このデータを用いた音声認識のための音響・言語モデルの適応，および音声認識の結果について述べる．

2. 知識構成型ジグソー法の実装

知識構成型ジグソー法は，共通の課題に取り組む他の学習者との関わり合いを通じて，単に課題を解決するだけではなく，他の学習者の考え方や学び方自体を学ぶことができる．また知識構成型ジグソー法は，学習の前後で問いに対する回答を二回求めるなどの従来にない特徴を持ち，以下のステップからなる．

(0) 問いを設定する．
(1) 自分のわかっていることを意識化する．
(2) エキスパート活動で専門家になる．
(3) ジグソー活動で交換・統合する．
(4) クロストークで発表し，表現を見つける．
(5) 一人に戻る．

これを実際の授業に当てはめた例を以下に示す．授業内容は中学校社会「関東地方」で，授業時間は 45 分，生徒は 21 人である．

(0) 課題：外国人観光客が関東地方（東京大都市圏）に集まる理由を説明しよう
(1) 個人毎に授業前の答を記述：協調学習を行う前の知識を用いた答を記述（4 分）
記述例

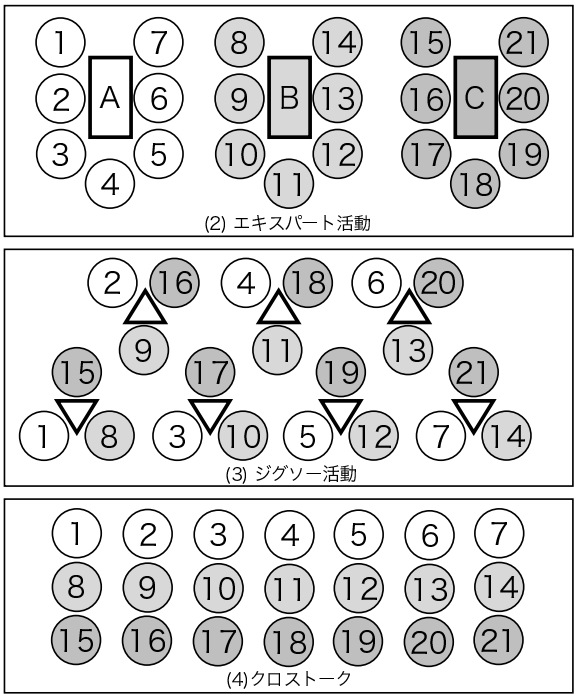

図 1 知識構成型ジグソー法

- 雷門とか秋葉原とか観光地がたくさんあるから．
- 空港が多いから移動しやすい．
- 交通が進んでいるし，サービス業が多いから買い物などが便利．

(2) エキスパート活動（図 1 上）：同じ資料を読み合うグループを作り，その資料に書かれた内容や意味を話し合い，グループで理解を深める．担当する資料に詳しくなる．（10 分：7 人ごとの 3 グループの形式）

A 外国人が利用する主な交通機関の運行本数，路線図をもとに，東京大都市圏と海外および国内地方の交通面のつながりについて考える．

B 宿泊施設，飲食店の数，自動車普及率，大規模小売店，レジャー施設の分布をもとに，外国人観光客にとっての東京大都市圏についての魅力について考える．

C 外国人観光客が日本滞在中に楽しみたいことの統計をもとに，需要と供給を満たす可能性について考える．

(3) ジグソー活動（図 1 中）：違う資料を読んだ人が一人ずついる新しいグループに組み替え，エキスパート活動で理解した内容を説明し合う．この活動により，学習者自身の理解状況を内省したり新たな疑問を持つことが期待される．（20 分：3 人ごとの 7 グループの形式）

(4) クロストーク（図 1 下）：協調学習を行った後に導き出された答を発表（1 分 ×7：ジグソー活動の 7 グループが各グループ毎に発表を行う）

(5) 個人毎に授業後の答えを記述：協調学習の内容を踏まえた答を記述（4 分）

記述例

- 大きな空港が 2 つあり新幹線で地方にも行きやすい．鉄道網が発達しておりホテルも多い．交通網が便利で自分のやりたいことができるところ．
- 新幹線が便利．施設が充実しており買い物も便利．外国人観光客のニーズを満たすことができる．
- 外国人は成田から入国する割合が非常に高く，成田に近い東京は新幹線による地方アクセスも便利．また外国人は日本料理に興味があり，東京には日本料理店が多い．

授業前の答と授業後の答とを比較すると，ジグソー活動でのやりとりを通して，外国人観光客が東京大都市圏に集まる理由には複数の理由があることを理解できたことがわかる．

3. 授業音声の音声認識

3.1 収録環境およびデータの内訳

2017 年度に広島県内の中学校の協力を得て授業音声の収集を行なった．音声収録は複数種類のヘッドセット（エレコム社製）と IC レコーダー (オリンパス社製) を各生徒が装着することで行った．音声録音の開始は生徒による IC レコーダーの録音ボタンの押下によるものとした．マイクと口との距離が一定になるようにヘッドセットのマイクを採用したが，学習の妨げにならないよう単に首にかける形でも良いとした．収録された音声は 16KHz モノラル音声として保存される．最終的に収集できた音声の内訳を表 1 に示す．エキスパート活動（以降 Exp）で収録された音声ファイルの数（人数）・音声ファイルの長さ，ジグソー活動（以降 Jig）およびクロストーク（以降 CrT）における音声ファイルの数と長さ，そして音声ファイルののべ時間を示す．おおよそ 1 回 45 分の授業から 10 時間分の音声が収録され，3 つの授業で合計 29.5 時間の音声データを得ることができた．

表 1 収集した音声の内訳：ファイル数およびファイルあたりの録音時間

	Exp		Jig		CrT		のべ時間
授業	人	(分)	人	(分)	人	(分)	(時間)
数学	22	9	19	24	2	1〜3	10.8
理科	20	9	18	20	7	1〜3	9.1
国語	19	6	20	23	7	1〜4	9.6
計							29.5

全生徒の録音が問題なく行われている場合，Exp 活動と Jig 活動におけるファイルの数は同じ数になり，CrT ステップにおけるファイル数は Exp 活動におけるチームの数（7 チーム）となるが，録音開始の失敗や録音不良によりいくつかのファイルは除外された．この収録された音声ファイルに含まれる発話に対し書き起こしを行った結果，合計 104K 文字の書き起こしを得た．書き起こしの，授業とステップによる分布を表 2 に示す．表 1 にあるとおり，Exp 活動より Jig 活動のほうが倍以上の時間があるため，発話される量も多くなっている．また CrT ステップに関しては各グループの代表のみがプレゼンテーションを行う形式なので，書き起こしの全体に占める割合としては小さくなる．一方，1 分あたりの書き起こし文字数を比較すると，音声ファイル 1 分あたり，Exp 活動 63 文字/分，Jig 活動 94 文字/分，CrT ステップ 101 文字/ 分となっており，理解が進むにつれ，単位時間あたりの発話が多くなっている．

表 2 書き起こしの分布 (文字数%)

授業	Exp	Jig	CrT	計
数学	11.5	23.0	0.2	34.7
理科	9.8	26.3	1.0	37.3
国語	5.1	21.5	1.2	27.9
計	26.5	70.9	2.5	100.0

発話内容に関する詳細な検討は行っていないが，中学校の授業であることから未知語は僅かに生徒同士で呼び合う名前（ニックネーム）程度であった．また CrT ステップを除くと生徒同士の会話であり，発話スタイルとしては非常にくだけた発話が多い．また呟きに近い自己の理解のための発話，チームメンバーへの情報提供や提案などの発話が混在しており，特に呟きに関しては書き起こしが難しいものも多く含まれた．

3.2 連続音声認識による中学生音声の認識

音声認識には CNN-HMM モデルを用いた．音響特徴量は 40 次元の対数メル周波数スペクトル係数に 1 次と 2 次の動的特徴量を付加した 120 次元のベクトルとした [6]．対数メル周波数スペクトルは，フレーム窓長 25msec，フレーム窓長 10msec を単位として抽出する．これら抽出された特徴量について平均・分散正規化を行ったのち，前後 5 フレームを含む合計 11 フレームからなる特徴量を CNN の入力とする．畳み込み層はフィルタ数 128 と 256 の 2 層で構成され，畳み込み層の後にノード数 2048 の全結合層を 4 層追加する．出力層は前後 2 音素のコンテキスト依存決定木に対応する 9300 ノードを持つ．畳み込み層の第 1 層は，前述の特徴量を入力とする 9×9 の畳み込みフィルタから構成され，Max Pooling を行う．第 2 層は 3×4 のフィルタを持ち，第 2 層目の出力は全結合層に接続される．言語モデルには数十万語の一般的な語彙が含まれる 4-gram モデルを用いた．

表 3 に収録音声およびその書き起こしを含まない汎用向け音響モデルと言語モデルを用いた連続音声認識の文字正解率を示す．表 2 に示した書き起こしを正解とし，文字誤り率を計算した．各授業および各ステップにより書き起こし文字数の分布に大きな偏りがあるため，合計にはそれぞれ書き起こしの頻度で重み付けを行った平均を記す．重み付き平均の文字正解率は 34.7%であった．CrT ステップの

高い正解率はプレゼンテーションの発話スタイルがより大人に近いものであったためだと考えられる．

表 3 ベースラインにおける文字正解率 (100-%CER)

授業	Exp	Jig	CrT	計
数学	33.9	31.7	79.6	32.8
理科	26.7	27.3	32.1	27.3
国語	50.4	45.0	71.7	47.3
計	34.4	34.1	56.2	**34.7**

3.3 中学生音声認識のための言語モデル適応

書き起こしをもとに言語モデルの適応を行なった．数学の授業には，理科と国語の書き起こしから作られた言語モデルを作成し，ベースの言語モデルと授業書き起こしを 7:3 の比率で確率の線型補間を行った．理科と国語に関しても同様に，教科オープンの言語モデルを作成しベース言語モデルの線型補間を行った．表 4 に音声認識精度を示す．文字認識率は平均して 2.8%改善し，言語モデル適応の効果が確認された．

表 4 言語モデル適応における文字正解率・オープン (100-%CER)

授業	Exp	Jig	CrT	計
数学	38.4	34.9	78.6	36.4(+3.6)
理科	29.8	29.9	29.7	29.9(+2.6)
国語	52.1	47.1	73.4	49.2(+2.0)
計	37.8	36.7	55.7	**37.5**
	(+3.4)	(+2.6)	(-0.5)	**(+2.8)**

また言語モデルの改善の上限を推定するため，全ての授業書き起こしを用いた場合についても表 5 に結果を示す．全ての授業書き起こしを追加すると，平均して 13.1%改善し 48.8%となる．表 4 では 2.8%の向上しか得られなかったが，別の教科の書き起こしを用いたことの影響も考えられる．また，中学生の授業は同一の内容を学習するための授業が各学級および毎年行われるため，発話の種類は比較的限られていることが予想される．言語モデルの教科依存性や，同じ教科での他の学習項目での書き起こしによる効果など，今後調べていく必要がある．

表 5 言語モデル適応における文字正解率・クローズ (100-%CER)

授業	Exp	Jig	CrT	計
数学	51.6	46.9	89.0	48.7(+15.9)
理科	40.6	41.7	42.5	41.7(+14.4)
国語	61.1	56.7	80.2	58.5(+11.3)
計	49.3	47.9	65.3	**48.8**
	(+14.9)	(+13.8)	(+9.1)	**(+13.1)**

3.4 中学生音声認識のための音響モデル適応

書き起こしとその音声を用いて音響モデルの適応を行った．音響モデルの適応は [7][8] で提案された Weight-Decay に基づくモデル適用を用いた．データの都合上，音響モデル適応の学習データは数学と理科の Jig 活動部分の書き起こしと対応する音声を用いた．対応する音声データ 13.5 時間のうち，有効な発話区間は 2.9 時間である．ファイル長に対して平均約 21%の区間で発話が行われており，3 人で 1 チームであることから，発話衝突がないとすると，おおよそ 30% ～ 40% の時間が思考に当てられていると推測できる．言語モデルはベースラインと同じく汎用の言語モデルを用いた．結果を表 6 に示す．

表 6 音響モデル適応における文字正解率・オープン/話者クローズ ◇/クローズ •(100-%CER)

授業	Exp	Jig	CrT	計
数学	39.2◇	48.2•	81.0◇	
	(+5.3)	(+16.5)	(+1.4)	
理科	41.4◇	50.6•	59.1◇	
	(+14.7)	(+23.3)	(+27.0)	
国語	55.4	47.9	74.2	50.4
	(+5.0)	(+2.9)	(+2.5)	(+3.2)

数学と理科の Jig 活動はテストデータが学習データに含まれたクローズであり，数学と理科のそれ以外の部分は話者クローズの結果となる．学習データの量が違うので単純な比較はできないが，オープンデータである国語の文字正解率を比べると，言語モデル適応の結果 (+2.0%) に比べ音響モデル適応 (+3.2%) のほうが改善の割合が大きかった．さらに言語モデル適応と組み合わせた結果を表 7 に示す．表 4 で用いた言語モデルを利用しており，言語モデルは各教科ごとにオープンになるように実験を行なった．

表 7 音響モデル適応における文字正解率・オープン/話者クローズ ◇/クローズ • + 言語モデル適応・オープン (100-%CER)

授業	Exp	Jig	CrT	計
数学	45.5◇	54.3•	82.6◇	
	(+11.6)	(+22.6)	(+3.0)	
理科	44.0◇	52.1•	57.5◇	
	(+17.3)	(+24.8)	(+25.4)	
国語	57.5	49.8	72.0	52.2
	(+7.1)	(+4.8)	(+0.5)	(+5.0)

実験の結果，言語モデル適用と音響モデル適用，それぞれに効果が出ており，国語の場合に着目すると，各適応の結果：言語モデル適応 (+2.0%) および音響モデル適応 (+3.2%) に対して，両方適応 (+5.0%) とほぼ合計した改善が得られた．

4. おわりに

学習プロセスの理解を目的とした知識構成型ジグソー法による中学生の授業音声の認識を行った．収録した音声を言語モデル・音響モデルの適応に用いたところ，意見を交

換しあうエキスパート活動における発話において，話者オープンで 5%，話者クローズでは 11.6%∼17.3%文字正解率が向上した．また，言語モデルをクローズにし，言語モデルによる改善の上限を調べたところ，11.3%∼15.9%文字正解率が向上した．中学校では同じ問題設定の授業を複数年実施されることから，学習データを蓄積することで，漸次的な音声認識率の向上が期待できる．

一方，中学生音声の音声認識に関しては，必ずしも十分な検討が行われているとは言えない状況にある．成人音声に関しては，人間の聞き取り能力についても様々な分析が行われ [9][10]，一定の条件下では人間の聞き取り能力を超える 認識精度を達成している [11][12] が，子供の音声に対してはこれから調査・検討を行う必要がある．

謝辞 協調学習の実践・データの採取は広島県安芸太田町教育委員会のご協力によるものです．深く感謝致します．本研究は科研費「17H06107」の助成を受けたものです．

参考文献

[1] 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講演論文集，pp. 1–8 (2010).
[2] 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)，pp. 257–258 (2013).
[3] 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音声言語情報処理研究会 (SLP) SLP-106，pp. 1–7 (2015).
[4] 南條浩輝，西崎博光，高橋　徹：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，情報処理学会研究報告音楽情報科学 (MUS) MUS-115(54)，pp. 1–4 (2017).
[5] 鮫島　充，ランディゴメス，李　晃伸，猿渡　洋，鹿野清宏：実環境における子供音声認識のための音韻モデルおよび教師なし話者適応の評価，情報処理学会論文誌，Vol. 47, No. 7, pp. 2295–2304 (2006).
[6] Fukuda, T., Ichikawa, O. and Nishimura, M.: Combining Feature Space Discriminative Training with Long-term Spectro-temporal Features for Noise-robust Speech Recognition, *The 12th conference in the annual series of INTERSPEECH (INTERSPEECH2011)*, pp. 229–232 (2011).
[7] Liao, H.: Speaker Adaptation of Context Dependent Deep Neural Networks, *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013)*, pp. 7947–7951 (2013).
[8] Suzuki, M., Tachibana, R., Thomas, S., Ramabhadran, B. and Saon, G.: Domain Adaptation of CNN based Acoustic Models under Limited Resource Settings, *The 17th conference in the annual series of INTERSPEECH (INTERSPEECH2016)*, pp. 1588–1592 (2016).
[9] Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L. amd Roomi, B. and Hall, P.: English Conversational Telephone Speech Recognition by Humans and Machines, *arXiv preprint*, p. arXiv:1703.02136 (2017).
[10] Stolcke, A. and Droppo, J.: Comparing Human and Machine Errors in Conversational Speech Transcription, *The 18th conference in the annual series of INTERSPEECH (INTERSPEECH2017)*, pp. 137–141 (2017).
[11] Kurata, G., Ramabhadran, B., Saon, G. and Sethy, A.: Language Modeling with Highway LSTM, *IEEE Automatic Speech Recognition and Understanding Workshop 2017 (ASRU2017)* (2017).
[12] Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X. and Stolcke, A.: The Microsoft 2017 Conversational Speech Recognition System, Technical report (2017).