多言語モデルの転移学習による日本人英語音声認識

森 滉介　　HOU, Wenxin　　篠崎 隆宏

東京工業大学 工学院 情報通信系
www.ts.ip.titech.ac.jp

1 はじめに

近年，end-to-end アプローチにより，DNN ベースの単一モデルによる自動音声認識が可能となった [1, 2, 3]．広範なコーパスを学習させたモデルは，人間と同等かそれ以上の音声認識性能を示す．一方，少量の音声データでは，高い認識性能を示すモデルの学習は難しい．非ネイティブ話者の発音は母国語の影響を受けやすいため，非ネイティブ音声認識では，同じ母国語を持つ話者の音声をモデルに学習させることが望ましい．それに伴ってモデルに学習させる非ネイティブ話者の音声データが必要となるが，非ネイティブ話者の数は少ない上に，音声には不自然な発音が多い．非ネイティブ音声の収集とラベル付けは，ネイティブ音声に比べて困難である [4]．

音声データが少ない言語における音声認識手法として転移学習がある．主流となっている手法では，学習データが豊富な言語の音声をモデルに事前学習させ，認識対象言語の音声でモデルを再学習させる [5, 6]．モデルの隠れ層を言語間で共有させることで，言語に依存しない普遍的な音声表現の抽出が可能となる [7, 8]．さらに最近では，適応させる言語知識の範囲を拡大し，多言語の音声を end-to-end ネットワークに事前学習させ，低資源言語に対する音声認識性能を高めている [9, 10]．

非ネイティブ音声を一種の低資源言語と捉え，転移学習を用いて他の言語知識を適応させる手法が提案されている [11, 12]．しかし，これまでの研究では，非ネイティブ音声認識に適応させる知識は 2，３言語に限られている．そこで本稿では，高精度の日本人英語音声認識を目的とし，適応させる知識を多言語に拡張した転移学習アプローチを提案する．

2 関連研究

2.1 低資源言語音声認識

低資源言語における音声認識手法の一つが転移学習である．Bukhar らは，事前学習させた多言語モデルの最終層を転移先言語で再学習させ，ウイグル語やベトナム語などの低資源言語に対する WER を改善した [13]．Cho らは，公開コーパスである BABEL から 10 言語の音声を用いて sequence-to-sequence モデルを学習させ，他の 4 言語に適応させた [8]．Kannan らは，インドにおける 9 言語の音声で学習させたストリーミング多言語 end-to-end モデルを用いて，低資源言語であるカンナダ語とウルドゥー語に対する WER を改善した [14]．Hou らは，42 言語で大規模な多言語 end-to-end モデルを学習させ，14 の低資源言語に転移学習させた [10]．そして，多数の言語を学習させたモデルの適応は，低資源言語の音声認識に対して優れた性能を発揮することを示した．同様に，多言語モデルの活用による低資源言語音声認識の性能改善が報告されている [15, 16, 17]．

2.2 非ネイティブ音声認識

非ネイティブ音声認識は，言語学習者が持つスピーキング能力の自動評価を実現するために必要不可欠である [18]．また，非ネイティブ音声認識は日常業務の中でさまざまな用途で利用されるようになった．国際化した今日の世界では，非ネイティブ音声の認識技術は重要になっている．

非ネイティブ音声認識では，転移学習によって複数言語間で知識を共有させる．Duan らは，共通の隠れ層と独立した出力層からなる DNN にネイティブの英語と日本語を学習させて言語横断的な音響モデルを構築し，日本人英語の音声認識における WER を低減させた [11]．Matassoni らは，イタリア語，ドイツ語，英語のネイティブ音声で学習させた DNN-HMM 音響モデルを，イタリア人ドイツ語，イタリア人英語，ドイツ人英語の非ネイティブ音声に

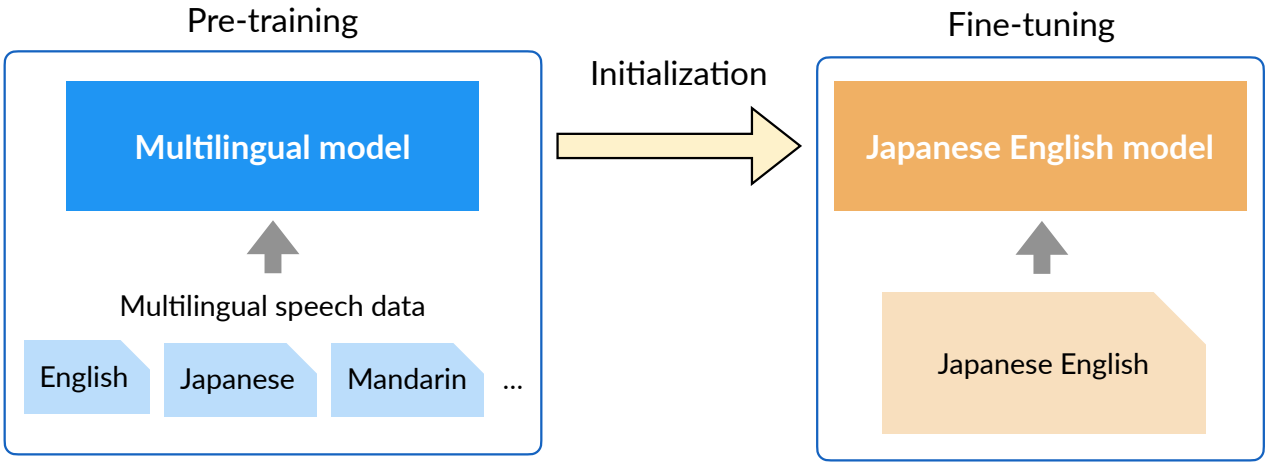

図 1　提案手法の概念図．多数の言語で end-to-end 音声認識モデルを事前学習させ，日本人英語の音声データで fine-tuning する．

転移学習を用いて適応させた [12]．そして，結果として得られた非ネイティブモデルが，非ネイティブ音声を直接学習させたモデルの性能を改善することを示した．

3　提案手法

日本人英語の高精度音声認識を目的として，多数の言語から転移学習させた end-to-end 音声認識モデルを採用し，適応させる言語知識を拡張する．提案手法の概要を図 1 に示す．

3.1　多言語音声認識モデル

多言語音声認識モデルには，end-to-end 音声認識に有効である Conformer [19] と Transformer [20] を用いる．モデルのエンコーダを 12 個の Conformer ブロックで構築し，デコーダを 6 個の Transformer ブロックで構成する．Conformer と Transformer による各ブロックは，256 次元の注意機構ヘッドを 4 つ含み，2048 次元のフィードフォーワードネットワークを持つ．モデルは，Connectionist Temporal Classification (CTC) と注意機構を同時に用いて予測文字をデコードする [21, 9]．デコードするシンボルのセットは，モデルに事前学習させる全ての言語に現れる文字セットを含むよう拡張する．これにより，エンコーダとデコーダのパラメータが全ての言語で共有される多言語モデルの学習が可能となる．モデルは，入力からデコードすべき言語を自動で認識し，予測したテキストを適切な文字セットで出力する．

3.2　日本人英語への転移学習

多言語モデルを日本人英語音声で fine-tuning する．まず，多言語モデルの学習済みパラメータで日本人英語モデルの全パラメータを初期化する．次に，日本人英語音声モデルの出力層に対して，出力次元数を日本人英語音声のシンボル数に変更し，パラメータを乱数で初期化する．そして，日本人英語音声をモデルに学習させる．実験では，10 言語と 42 言語を学習させた 2 種類の多言語モデルを日本人英語音声認識に適応させ，日本人英語音声認識に対して事前学習させる言語数の有効性を調べる．

4　実験

4.1　データセット

10 言語と 42 言語の多言語モデル学習には，11 の公開コーパスを使用した．使用した公開コーパスは，AISHELL [22], Aurora4, Babel, CHiME4, Common Voice [23], Corpus of Spontaneous Japanese (CSJ) [24], Fisher SwitchBoard, Fisher Callhome Spanish, HKUST [25], WSJ, Voxforge である．10 言語モデルの学習には Watanabe らが用いた言語を使用した [9]．また，42 言語モデルには Hou らが用いた言語を学習させた [10]．各多言語モデルの学習に用いた言語を表 1 に示す．また，学習データにおける各言語の発話数を図 2 に示す．

日本人英語音声認識には King-ASR-048 データを用いた．King-ASR-048 は，Android と iOS の携帯電話を同時に用いて収集した 2 チャンネルの日本人英

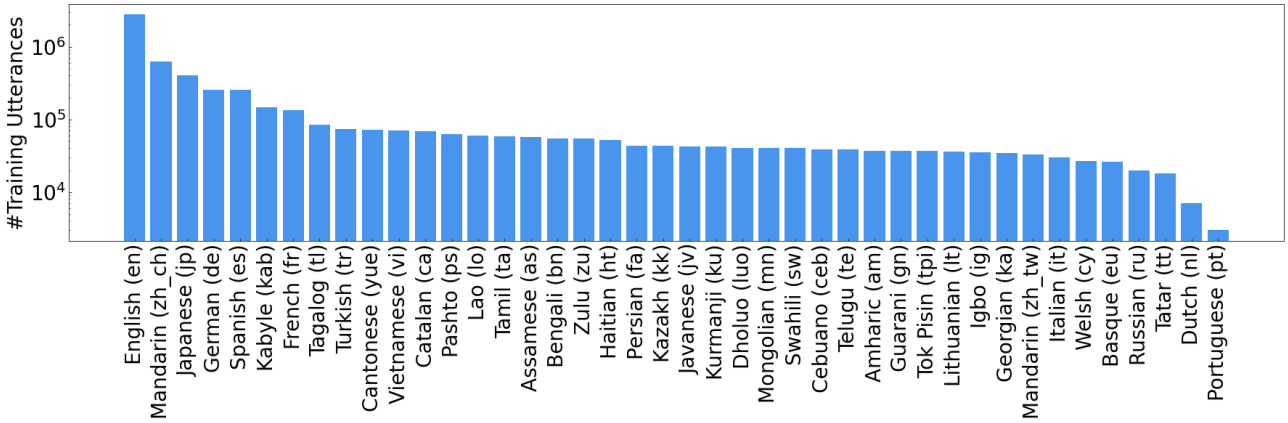

図 2　学習データに含まれている言語と発話数．学習データには，AISHELL [22]，Aurora4，Babel，CHiME4，Common Voice [23]，Corpus of Spontaneous Japanese (CSJ) [24]，Fisher SwitchBoard，Fisher Callhome Spanish，HKUST [25]，WSJ，Voxforge の 11 公開コーパスを用いた．

表 1　多言語モデルに学習させた言語．10 言語モデルの学習には Watanabe らが用いた言語 [9] を使用し，42 言語モデルの学習には Hou らが用いた言語 [10] 使用した．

Model	Languages
10-lingual	de, en, es, fr, it, ja, nl, pt, ru, zh_ch
42-lingual	de, en, es, fr, it, ja, nl, pt, ru, zh_ch, am, as, eu, bn, yue, ca, ceb, luo, ka, gn, ht, ig, jv, kab, kk, ku, lo, lt, zh_tw, mn, ps, fa, sw, tl, ta, tt, te, tpi, tr, vi, cy, zu

語音声データベースである．静かな屋内で 40 人の日本人から収集した 10,983 発話が収録されている．収録されている音声データのサンプリング周波数は 16 kHz であり，1 チャンネルあたりの全録音時間は約 17.9 時間である．実験では，Android で収集したチャンネルを使用した．話者が重複しないようデータを 3 つに分割し，80%を学習，10%を検証，10%をテストに使用した．

4.2　詳細設定

モデルの入力には 83 次元の特徴量を用いた．83 次元特徴量の構成は，80 次元の MFCC フィルタバンクと 3 次元のピッチ特徴量である．特徴量は，25 ms の窓を 10 ms ずつシフトさせて生の音声から抽出した．

モデル学習の最適化アルゴリズムには Adam を使用した．Adam における学習率 lr は，学習率パラメータ k，注意機構の出力次元 d_{model}，学習ステップ数 step，ウォームアップパラメータ warmup_step を用いて，

$$lr = k \cdot d_{\text{model}}^{-0.5} \cdot \min(\text{step}^{-0.5}, \text{step} \cdot \text{warmup\_step}^{-1.5}) \quad (1)$$

で表される．

表 2　多言語モデルの事前学習と fine-tuning におけるパラメータ設定．学習とデコードには ESPnet ツールキット [26] を使用した．

Hyperparameters	Pre-training	Fine-tuning
Training		
Epochs	100	100
Dropout	0.1	0.1
Learning rate factor k	4.5	1.0
Gradient clipping	5	5
Gradient accumulation	1	2
Batch size	1,280	32
Warmup step	25,000	25,000
CTC loss weight α	0.3	0.3
Decoding		
CTC decoding weight λ	0.5	0.5
Beam size	10	10

学習を高速化するため，多言語モデルの学習には TSUBAME 3.0 スーパーコンピュータ [1] を用いた．学習の並列処理に Pytorch のパッケージ [2] を使用し，合計 40 基の NVIDIA TESLA P100 GPU を搭載した 10 台の計算ノード上で多言語モデルを学習させた．多言語モデルの事前学習と fine-tuning におけるパラメータ設定を表 2 に示す．多言語モデルや日本人英語モデルの学習とデコードには ESPnet ツールキット [26] を使用した．

1）https://www.gsic.titech.ac.jp/en/tsubame
2）https://pytorch.org/docs/stable/distributed.html

表 3 日本人英語音声を直接学習させたモデル (Directly trained)，英語を事前学習させたモデル (Monolingual)，英語と日本語を事前学習させたモデル (Bilingual)，多言語を事前学習させたモデル (10-lingual/42-lingual) の各話者における性能比較.

	Speaker No.	Directly trained		Monolingual		Bilingual		10-lingual		42-lingual	
		CER [%]	WER [%]	CER [%]	WER [%]	CER [%]	WER [%]	CER [%]	WER [%]	CER [%]	WER [%]
dev	1	46.4	70.2	41.3	60.6	41.4	61.2	**40.3**	**56.9**	**40.3**	59.7
	2	18.7	47.5	7.3	19.0	6.7	17.9	**6.2**	**16.3**	6.4	17.7
	3	28.7	62.6	9.8	24.5	10.3	24.6	**9.3**	**23.1**	9.4	24.0
	4	31.0	63.2	16.2	33.6	15.7	32.5	**14.9**	**31.3**	15.2	32.1
test	5	27.0	59.7	11.0	25.7	10.4	24.8	**9.1**	**22.3**	9.7	23.2
	6	32.0	64.6	16.0	33.3	15.5	33.4	**14.2**	**31.1**	14.7	31.4
	7	27.6	59.6	7.9	19.4	7.7	18.6	**7.1**	**18.0**	7.8	20.6
	8	47.6	87.6	9.7	26.3	9.4	24.4	8.8	**23.6**	**8.5**	24.6

表 4 検証セット (dev) とテストセット (test) 全体で算出した CER と WER の比較.

Model	CER [%]		WER [%]	
	dev	test	dev	test
Directly trained	31.1	33.5	60.8	67.9
Monolingual	18.5	11.2	34.3	26.2
Bilingual	18.4	10.7	33.9	25.3
10-lingual	**17.5**	**9.8**	**32.4**	**23.8**
42-lingual	17.7	10.2	33.2	25.0

4.3 性能評価

提案手法の有効性を検証するため，日本人英語を直接学習させたモデル，英語を事前学習させたモデル，英語と日本語を事前学習させたモデルと性能を比較した．英語や日本語を用いた事前学習には，多言語モデル学習データにおける英語と日本語のサブセットを用いた．モデル性能の評価指標は CER と WER である．

5 結果

日本人英語音声を直接学習させたモデル，英語を事前学習させたモデル，英語と日本語を事前学習させたモデル，提案手法の各話者に対する CER と WER を表 3 に示す．また，検証セットとテストセット全体における CER と WER の比較を表 4 に示す．事前学習は，日本人英語音声認識におけるモデルの性能を大幅に改善した．日本人英語音声を直接学習させたモデル，英語だけ事前学習させたモデル，英語と日本語を事前学習させたモデルのテストセットにおける CER はそれぞれ 33.5，11.2，10.7 であった．

また，事前学習させたモデルの中で，多言語を学習させた 10 言語モデルと 42 言語モデルは，英語や日本語で事前学習させたモデルの性能をさらに上回った．これにより，多言語から学習した広範な音声表現知識の適応が，日本人英語音声認識に有効であることが分かった．

さらに，10 言語モデルと 42 言語モデルを比較したところ，10 言語モデルが高い認識性能を示した．42 言語モデルが学習した音声表現は，モデルのパラメータ数に対して広範であり，対象言語の音声認識能力を損ねた可能性がある．

6 おわりに

本稿では，多言語モデルを日本人英語音声認識に適応させる転移学習アプローチを提案した．Conformer と Transformer を用いて構築した end-to-end 音声認識モデルに多数言語の音声を事前学習させ，日本人英語音声で fine-tuning した．有効性検証のため，日本人英語を直接学習させたモデル，英語を事前学習させたモデル，英語と日本語を事前学習させたモデルと性能を比較した．その結果，多数の言語を事前学習させたモデルは，日本人英語音声を直接学習させたモデルの性能を大幅に改善した．また，提案手法は，英語だけ事前学習させたモデル，英語と日本語を事前学習させたモデルの性能を押し並べて上回った．これらの結果より，多数の言語から学習させた音声表現の適応が，日本人英語音声認識に有効であることが分かった．今後は，多種類の言語が持つ特性を構造化し，日本人英語音声認識に対して効果的に活用する方法を検討する．

謝辞

本研究は，JSPS 科研費 20H00095 の助成を受けたものである．

参考文献

[1] T. Hori, S. Watanabe, Y. Zhang, and W. Chan. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. In *Proceedings of Interspeech*, pp. 949–953, 2017.

[2] A. Zeyer, K. Irie, R. Schlüter, and H. Ney. Improved training of end-to-end attention models for speech recognition. In *Proceedings of Interspeech*, pp. 7–11, 2018.

[3] C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4774–4778, 2018.

[4] N. F. Chen, D. Wee, R. Tong, B. Ma, and H. Li. Large-scale characterization of non-native mandarin chinese spoken by speakers of european origin: Analysis on icall. *Speech Communication*, Vol. 84, pp. 46–56, 2016.

[5] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7639–7643, 2014.

[6] S. Tong, P. N. Garner, and H. Bourlard. An investigation of deep neural networks for multilingual speech recognition training and adaptation. In *Proceedings of Interspeech*, pp. 714–718, 2017.

[7] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black. Sequence-based multi-lingual low resource speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4909–4913, 2018.

[8] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori. Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 521–527, 2018.

[9] S. Watanabe, T. Hori, and J. R. Hershey. Language independent end-to-end architecture for joint language identification and speech recognition. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 265–271, 2017.

[10] W. Hou, Y. Dong, B. Zhuang, L. Yang, J. Shi, and T. Shinozaki. Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning. In *Proceedings of Interspeech*, pp. 1037–1041, 2020.

[11] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo. Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 391–401, 2020.

[12] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani. Non-native children speech recognition through transfer learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6229–6233, 2018.

[13] D. Bukhari, Y. Wang, and H. Wang. Multilingual convolutional, long short-term memory, deep neural networks for low resource speech recognition. *Procedia Computer Science*, Vol. 107, pp. 842–847, 2017.

[14] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee. Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model. In *Proceedings of Interspeech*, pp. 2130–2134, 2019.

[15] T. Sercu, G. Saon, J. Cui, X. Cui, B. Ramabhadran, B. Kingsbury, and A. Sethy. Network architectures for multilingual speech representation learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5295–5299, 2017.

[16] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao. Multilingual speech recognition with a single end-to-end model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4904–4908, 2018.

[17] A. Datta, B. Ramabhadran, J. Emond, A. Kannan, and B. Roark. Language-agnostic multilingual modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8239–8243, 2020.

[18] K. Knill, M. Gales, K. Kyriakopoulos, A. Malinin, A. Ragni, Y. Wang, and A. Caines. Impact of asr performance on free speaking language assessment. In *Proceedings of Interspeech*, pp. 1641–1645, 2018.

[19] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech*, pp. 5036–5040, 2020.

[20] L. Dong, S. Xu, and B. Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888, 2018.

[21] S. Kim, T. Hori, and S. Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4835–4839, 2017.

[22] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng. Aishell-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pp. 1–5, 2017.

[23] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4218–4222, 2020.

[24] K. Maekawa. Corpus of spontaneous Japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[25] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff. Hkust/mts: A very large scale Mandarin telephone speech corpus. In *International Symposium on Chinese Spoken Language Processing*, pp. 724–735, 2006.

[26] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai. ESPnet: End-to-End Speech Processing Toolkit. In *Proceedings of Interspeech*, pp. 2207–2211, 2018.