Graduate School of Fundamental Science and Engineering
Waseda University

博　士　論　文　概　要
Doctoral Dissertation Synopsis

論　文　題　目
Dissertation Title

A Study on Speaker Diarization based on End-to-end Optimization

エンドツーエンド最適化に基づく話者ダイアライゼーションに関する研究

申　請　者
(Applicant Name)
Yusuke FUJITA
藤田　雄介

Department of Computer Science and Communications Engineering, Research on Perceptual Computing

December, 2023

本研究は、複数人の会話録音を話者ごとの区間に分割するタスクである話者ダイアライゼーションに取り組む。話者ダイアライゼーションの従来法は、複数のモデルを複雑に組み合わせたシステム構成であるため、システム全体の最適化が難しかった。そこで本研究は、話者ダイアライゼーションへエンドツーエンド最適化を導入することを提案する。

本論文は、以下のように構成される。第 1 章では、話者ダイアライゼーションの評価方法、従来システムおよび最新システムの構成を概説し、これらの問題点を明らかにする。第 2 章においてこれらの問題点を解決するエンドツーエンド最適化に基づくモデルを提案し、従来法と比較して顕著に精度が向上することを示す。第 3 章および第 4 章では、エンドツーエンド最適化モデルの精度をさらに向上させる条件付き推定手法について検討する。第 3 章では、話者ごとに出力系列を分解して条件付き推定を行う方法を提案する。これにより、可変の話者数にも対応しながら精度向上を実現する。第 4 章は、中間予測を条件付けに利用する方法を提案する。中間予測を上位層にフィードバックする構成により、層ごとに話者ダイアライゼーション出力を反復的に改善する。現在最先端の精度を示す自己教師付き事前学習モデル（WavLM）が、大規模な学習データと多くのパラメータを使用する一方で、提案法は、従来と変わらない学習データと学習効率で、同等の精度を達成する。最後に第 5 章では、本研究の貢献をまとめるとともに、提案法を拡張した最新研究を紹介し、将来の研究の方向性を議論する。

話者ダイアライゼーションは、話者の登録プロセスなしで、入力音声内の複数の話者をそれぞれ識別する問題と考えることができる。そのため、長年にわたり話者識別問題の一種として研究され、大規模データで学習された話者識別モデルの利用が不可欠と考えられてきた。従来システムは、話者ダイアライゼーション問題を分割し、音声区間検出、話者識別モデルによる話者埋め込みの抽出、クラスタリングの少なくとも 3 つのモジュールを備える。これらはそれぞれ異なる目的関数（音声／非音声の分類精度、話者識別の精度、クラスターの純度）に基づいて独立に最適化される。2019 年以降の研究では、ダイアライゼーション出力（話者ラベル）の教師データを用いて最適化するモデルが提案されてきた。しかし、話者識別モデルを事前に学習しておくなど、複数の独立した最適化モジュールのパイプラインが必要となる点では変わっていない。また、ほとんどの従来法は重複音声を取り扱うように構成されておらず、重複音声検知のためには追加のモジュールを組み合わせる必要がある。このような複数モジュールに分割された構成のために、システム全体としての最適化が困難となっていることが、本研究が着目した課題である。

そこで本研究では、複数モジュールへの分割を行わない話者ダイアライゼーション問題の新しい解法と定式化を提案する。従来法は、「単一系列」の話者インデックス推定問題として定式化される。一方、提案法は、複数話者の音声活動検出を

並べたものを話者ラベルとする「複数系列」推定問題として定式化される。推定された話者ラベル出力には、音声区間検出、話者識別、重複音声検出等の結果が全て同時に含まれており、話者ダイアライゼーション問題を単一モデルのエンドツーエンド最適化に帰着出来る。この提案法を End-to-End Neural Diarization (EEND)と呼ぶ。EEND モデルを最適化するためには、複数系列の出現順序に依存しないパーミュテーションフリー学習法が必要となる。我々は、双方向長短期記憶（BLSTM）に基づく EEND モデルをシミュレーションデータで学習する実験を行い、パーミュテーションフリー学習が不可欠であること示した。さらに、BLSTM の代わりに自己注意機構を備えた EEND（SA-EEND）モデルは、実電話音声（CALLHOME）の評価セットを用いた実験で顕著な精度向上を示した。また、SA-EEND における自己注意メカニズムの振る舞いを可視化し、自己注意機構が時系列全体に分散されたグローバルな話者特徴を捉え、フレーム単位の話者の存在に応じてグローバルな話者特徴を各フレームに反映させていることを明らかにした。

EEND モデルは従来のシステムに比べて精度が良いが、いくつか実用上の弱点がある。EEND モデルは、対象ドメインに適合した大規模な学習データを必要とし、学習データ以外の事前知識を精度向上のために利用することができない。例えば、部分的に与えられた話者ラベルをコンテキスト情報として利用して精度を向上させるような条件付き推定の手法が使えない。そこで、話者ラベルをコンテキスト情報として導入するための条件付けに関する 2 つの手法を提案する。

1 つ目に、話者ラベル内部の依存関係を利用する条件付け手法「Speaker-wise chain rule」を提案する。提案法は、出力話者ラベルを話者ごとの系列に分解し、話者ごとの推定器のチェーンを形成する。先に推定された話者の系列を入力に加えながら、話者ラベル系列を反復的に生成することにより、提案法は他の話者の存在を事前知識として利用できる。提案法は、精度の向上に加えて、チェーンをいつ停止するかを学習することにより、話者数が事前に分からない音声にも適用できる。CALLHOME 評価セットの実験では、2 話者固定の条件でベースラインの EEND モデルを上回るだけでなく、話者数を不定にした実験でも、従来システムを上回る話者数カウントの精度を示した。さらに、音声区間検出および重複音声検出のサブタスクを利用した条件付け方法も提案する。Speaker-wise chain rule を拡張し、異なるタスクからの事前条件付け入力を可能とする。実験では、サブタスクを先に行って条件付けするモデルが Speaker-wise chain rule の精度を向上することを示した。

2 つ目に我々が提案する条件付け手法は、「中間予測を通じた自己条件付け (Self-conditioning via Intermediate Prediction) 」と呼ぶ。提案する自己条件付けモデルでは、ニューラルネットワークの中間層から生成された話者ラベルを上位層のネットワークにフィードバックする。提案法は、層ごとの中間予測を通

じて話者ラベルを反復的に改善できる。最先端の方法と比較するために、我々は、エンコーダーデコーダアトラクター（EDA）モデルに自己条件付けを実装した。しかし、自己条件付けを使用した場合、EDAの自己回帰構造がボトルネックとなることが分かった。そこで我々は、EDAの自己回帰構造を、非自己回帰の注意機構で置き換える非自己回帰アトラクターを提案する。実験では、提案法が元のEDAと比較して精度と学習効率の両方を向上させることを示した。自己条件付けを使用すると、中間層の出力がそのまま話者ダイアライゼーション出力として評価できる。モデルの各層の出力を評価すると、層を重ねるごとに精度が向上することが確認された。得られた精度を様々な既存手法と比較した結果、提案法より大規模なトレーニングデータと多くのパラメータを使用して学習された最新の自己教師付き事前学習モデル（WavLM）にも提案法が比肩することを示している。

EENDの導入は、話者ダイアライゼーション研究の焦点を、話者識別モデルとクラスタリングの改良から、話者ラベルを直接学習に活用するモデル開発へ大きく変革した。多くの論文が現在、EENDモデルの拡張に取り組んでいる。本研究のまとめとして、EENDのコンセプトに基づいて提案された最近の研究を例示しながら、将来の方向性を議論する。EENDの改良として最も注目されているのはEDAであり、従来クラスタリングによって行われきた話者埋め込みの集約を、明示的にEENDのネットワークの中で行うようにした点が高い性能に寄与している。また、最近の研究では、話者埋め込みの抽出をマルチタスク学習の枠組みで行うことでEENDと話者埋め込みのクラスタリングを組み合わせる手法や、事前にクラスタリングを通じて抽出した話者埋め込みを入力に加えてEENDを改善する方法が提案されている。これらは、高い精度が報告されている一方で、エンドツーエンド最適化から再びパイプライン構成に戻っている。エンドツーエンド最適化のシンプルさを生かしながら、従来法が活用してきた話者埋め込みのエッセンスをEENDに注入することは、未だ課題として残る。また、EENDが考慮してこなかった要素に「言語依存性」がある。特定のフレーズが話者の交替を示唆することがあるため、音声認識と組み合わせた手法で話者ダイアライゼーションの精度を向上する研究はいくつか報告されている。ただし、初めから使用言語を固定して音声認識と組み合わせて学習することは、学習効率が悪いと考えられる。既存の言語非依存なEENDモデルを、少量の学習データで言語依存モデルに適応させるような研究も期待される。このように話者ダイアライゼーションが音声認識等の上位タスクと統合される形で最適化される際、今後重要になるのは、自己教師付き事前学習モデルの活用であると考えられる。WavLMはその端緒であるが、GPT-3等の大規模言語モデルや、AV-HuBERTのような音声と画像のマルチモーダル事前学習の活用もEENDのさらなる拡張として期待される。

List of research achievements for application of Doctor of Engineering, Waseda University

Full Name： 藤田　雄介　　seal or signature

Date Submitted(yyyy/mm/dd): 2023/12/07

種類別 (By Type)	題名、 発表・発行掲載誌名、 発表・発行年月、 連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
Journal	○Yusuke Fujita, Tetsuji Ogawa, Tetsunori Kobayashi, Self-conditioning via Intermediate Predictions for End-to-end Neural Speaker Diarization, IEEE Access, 2023 (Accepted), DOI: 10.1109/ACCESS.2023.3340307
Journal	Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, Paola Garcia, "Encoder-decoder based attractors for end-to-end neural diarization," IEEE/ACM Transaction on Audio, Speech, and Language Processing, 2022
Conference	Robin Scheibler, Takuya Hasumi, Yusuke Fujita, Tatsuya Komatsu, Ryuichi Yamamoto, Kentaro Tachibana, "Foley Sound Synthesis with a Class-Conditioned Latent Diffusion Model," The 8th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), 2023
Conference	Aoi Ito, Tatsuya Komatsu, Yusuke Fujita, Yusuke Kida, "Target Vocabulary Recognition Based on Multi-Task Learning with Decomposed Teacher Sequences," Proc. Interspeech 2023
Conference	○Yusuke Fujita, Tatsuya Komatsu, Robin Scheibler, Yusuke Kida, Tetsuji Ogawa, "Neural Diarization with Non-Autoregressive Intermediate Attractors, " Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2023, DOI: 10.1109/ICASSP49357.2023.10094824
Conference	Yusuke Fujita, Tatsuya Komatsu, Yusuke Kida, "Alternate Intermediate Conditioning with Syllable-Level and Character-Level Targets for Japanese ASR," Proc. IEEE Spoken Language Technology Workshop (SLT), 2022
Conference	Tatsuya Komatsu, Yusuke Fujita, "Interdecoder: using Attention Decoders as Intermediate Regularization for CTC-Based Speech Recognition," Proc. IEEE Spoken Language Technology Workshop (SLT), 2022
Conference	Robin Scheibler, Tatsuya Komatsu, Yusuke Fujita, Michael Hentschel, "On Sorting and Padding Multiple Targets for Sound Event Localization and Detection with Permutation Invariant and Location-based Training," Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2022.
Conference	Robin Scheibler, Tatsuya Komatsu, Yusuke Fujita, Michael Hentschel, "Sound event localization and detection with pre-trained audio spectrogram transformer and multichannel separation network," Proc. The 7th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), 2022
Conference	Tatsuya Komatsu, Yusuke Fujita, Jaesong Lee, Lukas Lee, Shinji Watanabe, Yusuke Kida, "Better Intermediates Improve CTC Inference," Proc. Interspeech 2022
Conference	Yu Nakagome, Tatsuya Komatsu, Yusuke Fujita, Shuta Ichimura, Yusuke Kida, "InterAug: Augmenting Noisy Intermediate Predictions for CTC-based ASR," Proc. Interspeech 2022
Conference	Yawen Xue, Shota Horiguchi, Yusuke Fujita, Yuki Takashima, Shinji Watanabe, Leibny Paola Garcia Perera, Kenji Namagatsu, "Online Streaming End-to-End Neural Diarization Handling Overlapping Speech and Flexible Numbers of Speakers," Proc. Interspeech 2021
Conference	Yuki Takashima, Yusuke Fujita, Shota Horiguchi, Shinji Watanabe, Leibny Paola Garcia Perera and Kenji Nagamatsu, "Semi-Supervised Training with Pseudo-Labeling for End-to-End Neural Diarization," Proc. Interspeech 2021
Conference	Shota Horiguchi, Paola Garcia, Yusuke Fujita, Shinji Watanabe, Kenji Nagamatsu, "End-to-End Speaker Diarization as Post-Processing," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021

List of research achievements for application of Doctor of Engineering, Waseda University

Full Name： 藤田　雄介　　seal or signature

Date Submitted(yyyy/mm/dd): 2023/12/07

種類別 (By Type)	題名、 発表・発行掲載誌名、 発表・発行年月、 連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
Conference	Shota Horiguchi, Nelson Yalta, Paola Garcia, Yuki Takashima, Yawen Xue, Desh Raj, Zili Huang, Yusuke Fujita, Shinji Watanabe, Sanjeev Khudanpur, "The Hitachi-JHU DIHARD III System: Competitive End-to-End Neural Diarization and X-Vector Clustering Systems Combined by DOVER-Lap," The Third DIHARD Speech Diarization Challenge Workshop, 2021
Conference	○Yuki Takashima, Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Paola Garcia, Kenji Nagamatsu, “End-to-End Speaker Diarization Conditioned on Speech Activity and Overlap Detection,” Proc. IEEE Spoken Language Technology Workshop (SLT), 2021, DOI: 10.1109/SLT48900.2021.9383555
Conference	Yawen Xue, Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Paola Garcia, Kenji Nagamatsu, “Online End-to-End Neural Diarization with Speaker-Tracing Buffer,” Proc. IEEE Spoken Language Technology Workshop (SLT), 2021
Conference	Shota Horiguchi, Yusuke Fujita, Kenji Nagamatsu, “Block-Online Guided Source Separation,” Proc. IEEE Spoken Language Technology Workshop (SLT), 2021
Conference	Jing Shi, Xuankai Chang, Pengcheng Guo, Shinji Watanabe, Yusuke Fujita, Jiaming Xu, Bo Xu, Lei Xie, “Sequence to Multi-Sequence Learning via Conditional Chain Mapping for Mixture Signals,” Proc.Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS), 2020
Conference	Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, Kenji Nagamatsu, “End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors,” Proc. Interspeech 2020
Conference	Shota Horiguchi, Yusuke Fujita, Kenji Nagamatsu, “Utterance-Wise Meeting Transcription System Using Asynchronous Distributed Microphones,” Proc. Interspeech 2020
Conference	Jing Shi, Jiaming Xu, Yusuke Fujita, Shinji Watanabe, Bo Xu, “Speaker-Conditional Chain Model for Speech Separation and Extraction,” Proc Interspeech 2020
Conference	Zili Huang, Shinji Watanabe, Yusuke Fujita, Paola Garcia, Yiwen Shao, Daniel Povey, Sanjeev Khudanpur, "Speaker Diarization with Region Proposal Network," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020
Conference	○Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, Shinji Watanabe, "End-to-End Neural Speaker Diarization with Self-attention," Proc. Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 296-303, 2019, DOI: 10.1109/ASRU46091.2019.9003959
Conference	Naoyuki Kanda, Shota Horiguchi, Yusuke Fujita, Yawen Xue, Kenji Nagamatsu, Shinji Watanabe, "Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models," Proc. Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 31-38, 2019
Conference	Matthew Maciejewski, Gregory Sell, Yusuke Fujita, Leibny Paola Garcia-Perera, Shinji Watanabe, Sanjeev Khudanpur, "Analysis of robustness of deep single-channel speech separation using corpora constructed from multiple domains," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 165-169, 2019
Conference	○Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu and Shinji Watanabe, "End-to-End Neural Speaker Diarization with Permutation-Free Objectives," Proc. Interspeech, pp. 4300-4304, 2019, DOI: 10.21437/Interspeech.2019-2899
Conference	Naoyuki Kanda, Shota Horiguchi, Ryoichi Takashima, Yusuke Fujita, Kenji Nagamatsu and Shinji Watanabe, "Auxiliary Interference Speaker Loss for Target-Speaker Speech Recognition," Proc. Interspeech, pp. 236-240, 2019

List of research achievements for application of Doctor of Engineering, Waseda University

Full Name : 藤田　雄介　　seal or signature

Date Submitted(yyyy/mm/dd): 2023/12/07

種類別 (By Type)	題名、 発表・発行掲載誌名、　　発表・発行年月、　　連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
Conference	Naoyuki Kanda, Christoph Boeddeker, Jens Heitkaemper, Yusuke Fujita, Shota Horiguchi, Kenji Nagamatsu, Reinhold Haeb-Umbach, "Guided Source Separation Meets a Strong ASR Backend: Hitachi/Paderborn University Joint Investigation for Dinner Party ASR," Proc. Interspeech, pp. 1248-1252, 2019
Conference	Naoyuki Kanda, Yusuke Fujita, Shota Horiguchi, Rintaro Ikeshita, Kenji Nagamatsu, Shinji Watanabe, "Acoustic Modeling for Distant Multi-talker Speech Recognition with Single-and Multi-channel Branches", Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 6630-6634, 2019
Conference	Vimal Manohar, Szu-Jui Chen, Zhiqi Wang, Yusuke Fujita, Shinji Watanabe, Sanjeev Khudanpur, "Acoustic Modeling for Overlapping Speech Recognition: JHU Chime-5 Challenge System", Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.6665-6669, 2019
Conference	Naoyuki Kanda, Rintaro Ikeshita, Shota Horiguchi, Yusuke Fujita, Kenji Nagamatsu, Xiaofei Wang, Vimal Manohar, Nelson Enrique Yalta Soplin, Matthew Maciejewski, Szu-Jui Chen, Aswin Shanmugam Subramanian, Ruizhi Li, Zhiqi Wang, Jason Naradowsky, L. Paola Garcia-Perera, Gregory Sell, "The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays" , Proc. 5th International Workshop on Speech Processing in Everyday Environments (CHiME), 2018
Conference	Naoyuki Kanda, Yusuke Fujita, Kenji Nagamatsu, "Lattice-free State-level Minimum Bayes Risk Training of Acoustic Models", Proc. INTERSPEECH, 2018
Conference	Naoyuki Kanda, Yusuke Fujita, Kenji Nagamatsu, "Sequence distillation for purely sequence trained acoustic models", Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5964-5968, 2018
Conference	Naoyuki Kanda, Yusuke Fujita, Kenji Nagamatsu, "Investigation of lattice-free maximum mutual information-based acoustic models with sequence-level Kullback-Leibler divergence.", Automatic Speech Recognition and Understanding Workshop (ASRU), 2017, pp. 69-76, 2017
Conference	Rintaro Ikeshita, Yohei Kawaguchi, Masahito Togami, Yusuke Fujita, Kenji Nagamatsu, "Independent vector analysis with frequency range division and prior switching.", Proc. European Signal Processing Conference (EUSIPCO), pp. 2329-2333, 2017
Conference	Rintaro Ikeshita, Masahito Togami, Yohei Kawaguchi, Yusuke Fujita, Kenji Nagamatsu, Local Gaussian model with source-set constraints in audio source separation. IEEE International Workshop on Machine Learning for Signal Processing, 2017
Conference	Yusuke Fujita, Takeshi Homma, Masahito Togami, "Unsupervised network adaptation and phonetically-oriented system combination for the CHiME-4 challenge," Proc. The 4th International Workshop on Speech Processing in Everyday Environments (CHiME), pp. 49–51, 2016.
Conference	Yusuke Fujita, Ryoichi Takashima, Takeshi Homma, Masahito Togami, "Data Augmentation Using Multi-Input Multi-Output Source Separation for Deep Neural Network Based Acoustic Modeling," Proc. Interspeech 2016, pp. 3818-3822, 2016
	他, 国際会議3件