Graduate School of Fundamental Science and Engineering
Waseda University

博 士 論 文 審 査 報 告 書
Doctoral Dissertation Review Report

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

February, 2024

本論文が研究対象として扱う話者ダイアライゼーションとは，複数の話者が混在する状況において，誰がいつ話したかを特定する技術であり，会議議事録の作成や，会話状態の評価指標算出の前処理として用いられる重要な要素技術である．

話者ダイアライゼーションは，入力データ内の音声セグメントを，それぞれ誰が発声したかを識別する問題として捉えることができることから，長年にわたり話者識別問題の一種として研究が進められてきた．話者識別は，発話区間における話者埋め込みの分布あるいはそれに等価な特徴量を用いて行うことから，話者ダイアライゼーションシステムは，まず音声区間検出を行って，話者識別埋め込みを抽出し，最後にそれらをクラスタリングでまとめあげるという３段階の処理からなるパイプラインによって構成されてきた．時代とともに，それぞれのモジュールが最新技術に置き換わることになるが，このパイプラインに基づく方式が一定の成果をあげてきたこともあってか，これと異なるアプローチを採用する研究は本研究の開始以前においては皆無であった．一方で，前段でセグメンテーション，最終段でクラスタリングを必要とするこの構造において，それぞれのモジュールは，異なる目的関数に基づいて独立に最適化することが避けられず，近年様々な識別問題に導入され成果を挙げているEnd-to-endの枠組みによる全体最適化の方法は，この分野では長い間導入されることはなかった．

本論文は，音源分離に着想を得て，パーミュテーション不変学習法を話者ダイアライゼーション技術に導入することによって，陽なる音声区間検出，クラスタリングを必要としない，End-to-end 最適化に適する新たなシステム構成法を提案したものである．長年主流であった，話者識別モデルを中核に据えたパイプライン構成を捨て，音源分離に似せた定式化を行って当該分野に新たな流れを作ったものであり，極めて画期的な研究ということができる．

本論文は，以下に示す５章より構成される．

第１章では，話者ダイアライゼーションとはどのような問題かを述べ，その評価指標を紹介した後，従来システムの構成を概説し，その問題点を明らかにしている．

第２章では，提案の根幹となる，EEND (End-to-end Neural Diarization) と呼ぶ新たな話者ダイアライゼーションの基本方式について述べている．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していたのに対し，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している．音源分離で開発された，複数系列の出現順序に依存しないパーミュテーション不変学習法を導入することで，話者ラベル出力には，音声検出，話者識別，重複音声検出の結果を全て同時に反映させることが可能になり，話者ダイアライゼーション問題を単一モデルのEnd-to-end最適化に帰着させることに成功している．また，こうした構成をとることで，従来法では困難であった重複音声を取り扱うことも可能にして

いる．標準的なダイアライゼーションの評価セットであるCALLHOMEを用いた実験において，ダイアライゼーションエラー率（DER）を従来手法の11.5%から9.5%に減じるなど，効果を示している．また，シミュレーションで作成した発話重複率が30%を超える難しい評価データセットにおいては，従来手法において28.8%であったDERを4.6%に減じるなど，劇的に性能を改善することを示している．

第３章では，話者ラベル間の依存関係を利用するSpeaker-wise chain ruleを提案し，基本EENDモデルの精緻化と機能拡張について検討している．推定された話者系列を逐次入力に加えながら，話者ラベル系列を反復的に生成することにより，提案モデル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話者推定の連鎖をいつ停止するかも定めることができ，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から6の間で話者数が変動するタスクにおいて，20.5%であったDERを15.8%に減じるとともに，話者数の推定精度を54.6%から74.8%と大幅に向上することが示されている．

第４章では，Self-Conditioningとよぶ，不完全な話者ラベル推定結果で条件づけた話者ラベル推定を繰り返し行うことで推定精度を精緻化する方式を提案している．エンコーダはトランスフォーマであって多層構造を持つ．基本方式ではその最終層の出力をデコードする形で話者ラベルを推定するのに対し，提案法では中間各層においてもその出力をデコードして話者ラベルを推定する．その結果はエンコーダの当該層の出力，すなわちエンコーダ次層の入力に戻され，次層のラベル確率を条件づける形で利用される．このことによって，各中間層における話者ラベルの推定結果が層を重ねるごとに徐々に改善する形になっている．CALLHOME 評価セットの実験において，EEND基本方式において8.7%であったDERは，Self-conditioningにより7.8%に改善するなど，効果が示されている．

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ながら，将来の研究の方向性を議論している．短期間に数多くの研究が，本研究を発展させる形で提案されていることが紹介されており，話者ダイアライゼーションにおける本研究の重要性が理解できる．

以上，これを要するに，本論文は，話者ダイアライゼーションと呼ぶ複数人の会話音声データを話者ごとの区間に分割するタスクにおいて，音源分離における基本技術として考案されたパーミュテーション不変学習を導入することで初めてEnd-to-end学習を適用可能とするとともに，これを基本として様々な精度改善の方法を提案することで，顕著なダイアライゼーション性能を達成したものである．本研究を契機に，多くの関連方式が派生しており，本研究は当該研究分野に新たな潮流を作ったものとして高く評価できる．また，提案手法は，話者ダイアライゼーションへの適用に留まらず，　音環境理解をはじめ識別対象が重複しながら入れ替わり観測される一次元系列デ

ータの分類タスクに対して広く適用可能であり，その工学的価値は高い．よって，本論文は，博士（工学）（早稲田大学）の学位を授与するに相応しいものと認める．

2024年　2月

審査員

主査	早稲田大学　教授	工学博士（早稲田大学）	小　林　哲　則
副査	早稲田大学　教授	博士（情報学）（京都大学）	河　原　大　輔
	早稲田大学　教授	博士（工学）（早稲田大学）	小　川　哲　司
	カーネギーメロン大学　准教授	博士（工学）（早稲田大学）	渡　部　晋　治