人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-066-04

話者情報の半教師あり学習を用いたオフライン話者ダイアライゼーション

Offline Speaker Diarization with Semi-Supervised Learning using Speaker Information

阿坂 脩平[1*]　Benjamin Yen[1]　糸山 克寿[2]　中臺 一博[1]
Shuhei Asaka[1]　Benjamin Yen[1]　Katsutoshi Itoyama[2]　Kazuhiro Nakadai[1]

[1] 東京科学大学
[1] Institute of Science Tokyo
[2] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[2] Honda Research Institute Japan, Co. Ltd.

Abstract: 本稿では，高性能なオフライン話者ダイアライゼーションの実現を目指す．そのために，以下の 2 点からなるオフライン話者ダイアライゼーションモデルを提案する: 1) 一部のアノテーション済みデータからでも適切に特徴量を抽出できるように, 話者基盤モデル (WavLM) を導入, 2) アノテーションのコストを抑えるための, 半教師あり学習に基づくデータ拡張を導入.
会議録コーパスである AMI コーパスを用いて，提案手法を評価したところ, 半教師あり学習の導入によって, DER で最大 3.54 ポイント改善した．これは, pyannote に実装されている End-to-End のダイアライゼーションモデルよりも高い性能であり, 提案手法の有効性を示すことができた.

1　はじめに

話者ダイアライゼーション（Speaker diarization）は，会議等の複数人話者の音声収録データから,「いつ」「誰が」発話したかを推定する技術である．音声認識と組み合わせて用いることで，自動で議事録の作成が可能となり，また，対話分析技術をもとに，教育分野などへの応用も期待できる．話者ダイアライゼーションには，発話区間検出 (VAD)，話者特徴量の抽出，話者識別の 3 つの要素技術が含まれている．従来は，各要素に対応する処理を順番に行うカスケード処理 [2] が研究されてきたが，近年では，これらの処理を同時に行う EEND [1] などの End-to-End 話者ダイアライゼーションの研究が盛んであり，従来より高い性能をもつモデルが報告されている．これらのモデルのトレーニングには膨大な学習データが必要だが, アノテーション済みの複数人音声コーパスは一般的に限られており, 多くの研究ではシミュレーションデータを使って大規模な学習を行っている．しかし, 実際の会議ケースを考えると, 過去の会議音声データを事前に取得できる場合が多く, また, オフライン処理では会議の冒頭の一部にアノテーションを行うことも可能である．そこから話者情報を抽出することで, その会議や発話者に適応できるモデルを作成できる可能性がある．本稿では, 入力データのシーンに適応できる高性能なオフライン話者ダイアライゼーションの実現を目指す．そのために, 一部のアノテーション済みデータからでも適切に特徴量を抽出できる, 話者基盤モデル (WavLM) を導入する．また, アノテーションのコストを抑えるために, 半教師あり学習に基づくデータ拡張を導入する.

*連絡先： 東京科学大学
152-8552 東京都目黒区大岡山 2-12-1
E-mail: asaka@ra.sc.e.titech.ac.jp

2　関連研究

話者ダイアライゼーションの実シーンへの適用における改善点を指摘し, 提案手法につながる関連研究を挙げる.

2.1　モジュールベース手法

古典的な話者ダイアライゼーションは, タスクを複数の操作に分割し, 各モジュールを連結することで機能をなす, 図 1 に示すモジュールベース手法 [2] で実現されている．詳細には, まず音声処理の性能向上のために, 音声強調, 残響除去などの前処理が施される．次に, 発話区間検知 (VAD) を適用することで, 複数の発話を非発話区間から分離する．その後, 分離された音声信号を,

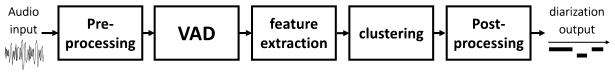

図 1: Classical speaker diarization method

特徴量抽出器によって話者特徴量に変換する．さらに，クラスタリングによって，話者のクラスタを形成し，各発話に話者クラスを振り分ける．最後に，後処理で出力の時系列を考慮して，話者ダイアライゼーション結果を生成する．しかし，このような複数のモジュールからなる手法は，モジュール単位で最適化されるため，システム全体として最適化されているとは限らない．また，一般に VAD は，重複発話を扱うことができないという課題もある．

2.2 End-to-END 手法

モジュールベース手法の課題の解決のため，EEND [1] に代表される End-to-End 話者ダイアライゼーションが提案されている．これは，話者ダイアライゼーションタスクを，各フレームの多ラベル分類問題として扱うことで重複発話の対応を図るものである．具体的には，図 2 のように，特徴量抽出層，Multi-head self-attention 層，線形層を合わせた End-to-END モデルとして学習される．このモデルは長い発話の取り扱いが難しいという問題があり，その解決のため，Chunk 単位の処理が導入されている．また，話者数が事前に必要であるという課題もあり，この課題に対しては，話者識別が可能となるようクラスタリングの枠組みを取り込んだ研究 [3][4] が報告されている．このように EEND からなる End-to-END の話者ダイアライゼーションモデルは，広く検討されており高い性能が報告されている．また，これらのモデルは学習に時間がかかるうえ，多くのデータを必要とする．そのため，複数人話者音声を単一話者音声から作成し，シミュレーションデータで学習量を補っている研究もある [5]．

しかし，これらのオンライン手法では，対象とする音声の話者情報は直接学習に使われていない．実シーンでの自動議事録生成を考えると，過去の会議音声データを事前に取得できる場合が多いうえに，本稿で提案するオフライン処理を前提とすれば，会議の冒頭の挨拶や自己紹介などの，各話者の発話区間が明確な部分に，一部アノテーションを行い，それを学習に利用することも容易である．そこから話者特徴量をうまく抽出できれば，その状況に適応したモデルを作ることが可能であると考えられる．

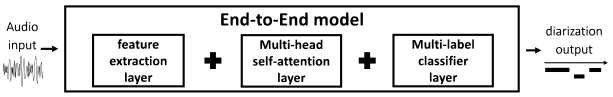

図 2: End-to-End speaker diarization method

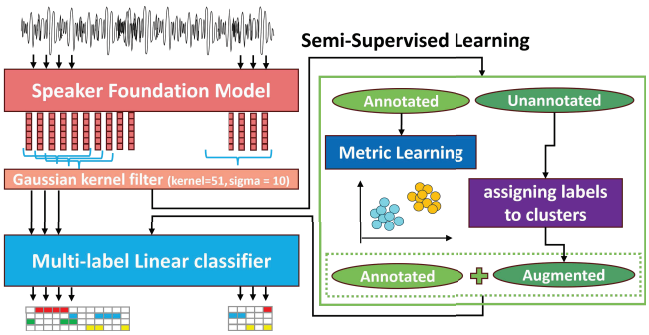

図 3: Proposed Method

2.3 話者基盤モデルの活用

話者ダイアライゼーションタスクの性能向上において，特徴量抽出器として，近年注目されているのが wav2vec2 [6]，WavLM [7] などの話者基盤モデルである．これらは，大量の音声データを用いて学習した音声認識モデルであるが，多数話者データで学習しているため，話者識別の基盤モデルとして利用できる．事前学習された特徴量抽出器として活用すれば，より音響環境に頑健かつ，表現が豊かな特徴量を得ることが可能である．実際に，話者基盤モデルは話者ダイアライゼーション，話者識別にも活用されており，各タスクの性能向上に寄与していることが報告 [8][9] されている．また，話者基盤モデルを活用するにあたって，一般的に，入力に近い浅い層では，局所的な音響情報を表し，深い層ではより抽象的，言語的な情報を表す [10][11] と言われている．しかし，話者ダイアライゼーションタスクにおいて，話者基盤モデルのどの隠れ層の出力が有用であるかの検討はなされていない．

3 提案手法

図 3 に本稿の提案手法を示す．本手法は，オフライン処理を前提としており，話者ダイアライゼーションを行う音声の冒頭部分や過去の会議音声に対し，一部アノテーションを行う．そこから話者特徴量を話者基盤モデルを用いて抽出し，多ラベル識別器の学習を行うことで話者ダイアライゼーションモデルを実現する．また，必要なラベル付きデータ量を低減するために，半教師あり学習に基づくデータ拡張も導入する．

以下に，各手法の詳細を示す．

表 1: Speaker foundation models

	Params	Layers
wav2vec 2.0 [6]	95.04M	12
HuBERT [12]	94.68M	12
Unispeech-SAT [13]	94.68M	12
WavLM [7]	94.70M	12

3.1 話者基盤モデルによるタスクに適した特徴量抽出

前節で述べた通り, より音響環境に頑健かつ, 表現が豊かな特徴量を得るために, 話者特徴量抽出器として話者基盤モデルを活用し, 話者ダイアライゼーションタスクに適した活用法を検討する. 本稿で利用するモデルは表 1 にある 4 種類のモデルである. これらは, 主に音声認識モデルとして学習されているが, そのエンコーダ部分を用いることで, 事前学習済みの話者特徴量抽出器 [1]として利用することが可能である. 例えば, wav2vec2.0 への入力は音声時間波形であり，この入力に対して，まず，stride が 20 ms，カーネルサイズが 25 ms の一次元 CNN (convolutional neural network) を適用する. 次に Transformer 層で，時間的な前後関係も考慮して，話者情報を含む音声特徴量に変換される. Transformer 層の隠れ層出力を得ることで，シングルチャネルの音声から, 時間波形 20 ms ごとの 768 次元の音声特徴量を取得する. 他のモデルも, 事前学習における学習戦略は異なるものの, 同様の構造を踏襲している. よって, 本稿では異なるパラメータを持つ, 入出力形状は同じ特徴量抽出器として扱う. 各モデル, 各層の出力を用いて話者ダイアライゼーションタスクを行い, その中で性能の高い層を最終的に提案手法として用いる. またそれによって, 話者ダイアライゼーションタスクに適した話者基盤モデルの活用法を示す.

加えて, 各フレームごとの話者特徴量は, ある発話の一部分の特徴量表現となっているため, 同じ話者の発話であったとしても, 大きく異なった特徴量を示すことがある. これを防ぐために, kernelsize = 51, sigma = 10 のガウシアンフィルタを適用する. これはあるフレームの前後 0.5 秒における特徴量の値を, 時間が近いものほどより反映するように処理するフィルタであり, これによって, 話者特徴量の統一化を行いつつ, 出力の発話セグメントが細かく分割されることを防ぐ.

3.2 多ラベル話者識別器

識別器の出力が直接, 重複発話を表現可能な話者ダイアライゼーション出力となるように, EEND [1] の多ラベル線形識別器の構造を取り入れる. 識別器は図 4 に示す通り, 4 層の線形層からなっており, 入力をフレーム数× 768 次元として, 出力をフレーム数×話者クラス数とする. 本稿では扱うコーパスの都合上, 話者数を 4 と決定している. モデルの出力に sigmoid 関数を適用することで, 1 フレーム当たりの出力は話者クラスそれぞれの確率となる. それを閾値を 0.5 として, 各話者の発話の有無を判断し, フレーム毎の多ラベル話者識別結果を得る. これにより, 各フレームでどの話者が発話しているかを, 重複発話を含めて表現することが可能となり, 出力を直接話者ダイアライゼーション結果として扱うことができる. また学習の損失関数に関しては, 多ラベル識別タスクを扱うため, Binary Cross Entropy に sigmoid 関数を統合した BCEwithLogitsLoss を用いる. これにより, 同一フレームに対して話者クラスが複数割り当てられることを許容する学習が可能となる. また, すべてのフレームにおいて誰かが発話していると概算すると, ラベルが 0 である個数と 1 である個数は, 3:1 であるため, この損失関数のパラメータである pos_weight を 3 に設定している. これにより, 多ラベルの性質上, ラベルが 1 となる可能性が低い傾向にあるものの, 適切に学習を行うことができる.

$$\mathrm{BCEWithLogitsLoss}(y, z) = \frac{1}{N}\sum_{i=1}^{N}\Big[-y_i \log\left(\sigma(z_i)\right) - (1-y_i)\log\left(1-\sigma(z_i)\right)\Big] \tag{1}$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{2}$$

3.3 半教師あり学習によるデータ拡張

学習に必要なアノテーション済みデータの量を低減させるため, 距離学習を用いた半教師あり学習に基づいたデータ拡張を行う. 本手法は, オフライン処理を前提としており, 話者ダイアライゼーションを行う音声の冒頭部分や過去の会議音声に対し, アノテーションを行い, 学習に利用することができる. しかし, 音声データすべてにアノテーションを行うことは現実的ではないので，本稿では，一部のラベル付きデータとと, その他ラベルなしデータを用いて, 半教師あり学習を行う.

半教師あり学習では, ラベルなしデータで学習を行うために疑似ラベルを割り当てる操作が必要となるが, 本稿では距離学習の一種である Siamese ネットワーク [14] を用いることでその疑似ラベルの品質を向上させる. 距離学習はクラスが異なるデータ間の距離を大きく，同じクラスのデータの間の距離を小さくするように学習する手法であり，クラス識別が容易な, 話者情報をより表現した特徴量を得ることができる. Siamese ネットワークの特徴はデータセットの構成と損失関数にある. データセットには，wav2vec2.0 の出力から 2 つのサンプルを抽出し，そのペアが同じクラス同士なら

[1] https://s3prl.github.io/s3prl/tutorial/upstream_collection.html

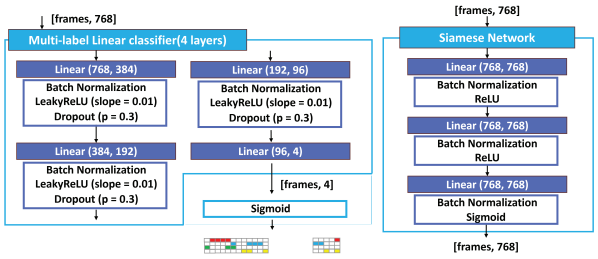

図 4: Multi-label classifier & Siamese Network

Positive ペアとして 1 を，異なるクラスなら Negative ペアとして 0 をラベルとして付与して作成する．サンプルペアとラベルを入力として 3 層の線形層にて学習を行う．損失関数は以下の Contrastive Loss を用いる．

$$\mathcal{L}_{contrastive} = \frac{1}{2}YD^2 + \frac{1}{2}(1-Y)\max(margin - D, 0)^2 \quad (3)$$

D は特徴量ベクトルのユークリッド距離，Y はサンプルペアのラベルであり，Negative ペアなら 0，Positive ペアなら 1 が入る．mergin はハイパーパラメータであり Negative ペアを遠ざける最低距離を定める．Siamese ネットワークは図 4 に示されるように, 3 層の線形層とし，入出力はともに 768 次元である．本稿では，話者クラスタを明確化するために，ラベル付きデータのうち, 単一話者の区間のデータ（各区間の両端 5 フレームは除外）から作られる Positive，Negative ペアを学習データの 50 倍の個数ずつ用意し，それを用いて学習を行った．以上のような距離学習によって, 少ないデータから抽出された特徴量をもとに, 話者のクラスタがはっきりと確認できる特徴量に変換するための特徴量変換器を作成することができる．

これを用いて, ラベルなしデータを距離学習モデルに入力し埋め込みを行い, ラベル付きデータ分布と重ねることで, クラスタリングを行う．クラスタの重心からのユークリッド距離が閾値以内であるデータの中から, 距離が近いデータを順に取得し疑似ラベルを付加する．この際, 学習データの偏りが生まれないように, 各クラスにつき同数ずつ, 閾値内のデータを取得する．以上から得られた, 疑似ラベル付きのデータも併せて学習を行うことで, 少ないラベル付きデータでも, データ量を確保した学習を行うことが可能となる．

4 評価実験

4.1 データセット

使用するデータセットは，AMI corpus [15] である．本コーパスは実際の複数人会議音声を収録したものであり, 本稿ではシングルチャネルの音声データを用いる．

表 2: AMI corpus meeting ID

	speaker		audio duration			
	male	female	a	b	c	d
ES2006	1	3	1284.3	2183.1	2181.7	1967.4
ES2008	1	3	1043.4	2231.7	2102.6	2625.8
ES2009	3	1	1402.2	1435.3	1956.9	2114.9
ES2015	0	4	1146.6	2294.9	2135.9	1931.6

利用する会議 ID とその属性は, 表 2 の通りである．これらは, 各話者の発話量がある程度均等であること, 話者の交代が単純でないことを基準として選定した．各会議 ID の中には a から d の四つの音声が含まれているが, それらは同じグループによる会議である．また, 学習で用いるアノテーションデータは Lamdini らの研究 [16] を元に取得した．これらは事前に話者基盤モデルを通して, フレーム数× 768 次元の特徴量に変換されているものとする．

4.2 評価指標

話者ダイアライゼーションの性能評価の指標として, 本稿では, DER (diarization error rate) [17] を利用する.DER の算出方法は以下のとおりである．

$$\text{DER} = \frac{\text{false alarm+mis-detection+confusion}}{\text{total}} \quad (4)$$

false alarm は非発話が誤って発話として分類された区間，**mis-detection** は発話が誤って非発話として分類された区間，**confusion** は発話区間のうち話者 ID が誤って識別された区間，**total** は全話者の正解発話区間の合計時間である．他の研究では, DER の算出において, 簡単のために重複発話区間を除外する, 話者交代時の誤差を許容するカラーを設ける等の操作がとられることもあるが, 本稿では, 例外は設けず最も厳しい条件での DER の算出 [16] を行う．

4.3 実験 1 ：モデル, 隠れ層の性能比較

本稿の話者ダイアライゼーションタスクにおいて, どの基盤モデルのどの隠れ層が有効であるかを検証する．簡単のために, 半教師あり学習に基づくデータ拡張を行わない場合の提案モデルにおいて比較を行う．学習データは, 各会議 ID の a から c までの音声特徴量を合わせて, ガウシアンフィルタを適用したものを利用し, 検証データはデータの 20% をランダムに抽出する．評価データは, 各会議 ID の d の音声特徴量にフィルタを適用したものを利用する．パラメータは, batch_size = 64, epoch = 200, オプティマイザは Adam, 初期学習率は 1e-4, スケジューラは cosineannealingLR (T_max=20 , eta_min = 1e-7), に設定し学習を行い, 検証損失最低時のモデルを取得した．

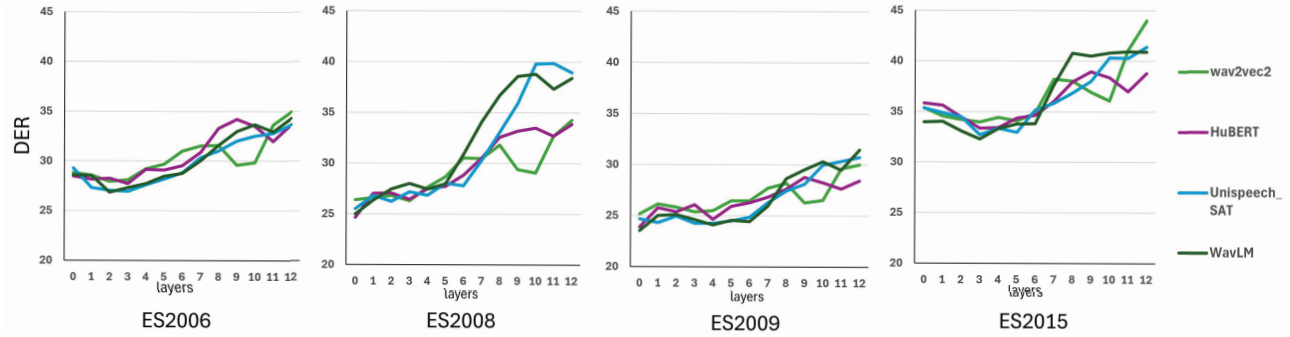

図 5: DER using one layer of speaker foundation models

4.4 結果 1 ：モデル, 隠れ層の性能比較

図 5 に各会議 ID における, モデルとその出力を取得する層を変化させた場合の DER の推移を示す. 結果からわかるように, どのモデルもグラフも, 浅い層では DER が低く, 深い隠れ層になるほど DER の値が上昇する傾向にある. DER が低い値をとるほど, モデルの性能は良いため, 話者ダイアライゼーションタスクでは, 話者基盤モデルの浅い層で特徴量を抽出することが効果的であることがわかる. しかし, 会議 ID によって最適な層は異なり, ES2008 では 0 層目が最も性能が高いが, 他の会議 ID では, 3 から 6 層目の出力も効果的であることが判明した. また, パラメータ数がほぼ等しいこの 4 種類のモデルにも性能の違いが見られた. Unispeech-SAT, WavLM は似た推移を見せ, 他二つのモデルに比べ, 前半層において高い性能を発揮し, wav2vec2 は, 9,10 層目において, 話者ダイアライゼーションタスクへの適性を確認することができた. 以上より, 話者ダイアライゼーションタスクに適した, 話者基盤モデルとその隠れ層は, WavLM, Unispeech-SAT の前半層であるとし, 以下の実験を続ける.

4.5 実験 2 ：半教師あり学習の有効性検証

本稿の話者ダイアライゼーションモデルにおいて, 半教師あり学習に基づくデータ拡張が, モデルの性能に寄与するか検証する. ラベル付きの学習データは, 表 3 に示されるように, 各会議 ID のの a から c までの音声において 600 秒間の, 各話者がある程度均等かつ, 長く発話している部分を用いる. 例として, ES2006 のラベル付きの区間を図 6 に示す. ラベルなしの学習データは, 各会議 ID の a から c までの音声のうちラベル付きとして利用していない区間を用いる. 評価データは, 各会議 ID の d の音声を利用する. 特徴量抽出器として, 実験 1 の結果を参考に, WavLM の 0 から 6 層目を利用する. Siamese ネットワークのパラメータは, batch_size = 64, epoch = 5, オプティマイザは Adam,

表 3: labeled data for Ex.2

ID	target segment(s)
ES2006	ES2006a: [200, 800]
ES2008	ES2008a: [0 , 600]
ES2009	ES2009b: [0 , 600]
ES2015	ES2015a: [500, 1100]

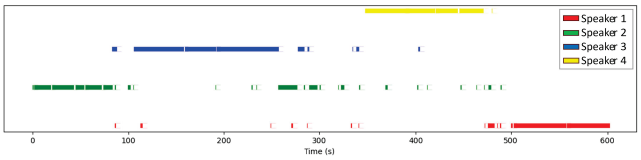

図 6: Labeled segments of ES2009 (600 seconds)

初期学習率は 1e-6 Contractive loss は margin = 2.0 に設定し, 学習を行う. また, 疑似ラベルを付加するためのクラスタの重心からの距離の閾値は 0.5 とした. 多ラベル識別器のパラメータは, 実験 1 と同様に設定し, 学習を行い, 検証損失最低時のモデルを取得した.

有効性評価のために, 半教師あり学習に基づくデータ拡張を行わない場合の性能との比較を行う. また, 実用に足るかという観点で, 利用しやすい形で提供されている, pyannote の End-to-End 話者ダイアライゼーションモデル [19][18] の性能との比較も行う. ただし, このモデルは提案手法のように, 分析する対象データの一部で学習が行われているわけではなく, 話者が 4 人であるという情報のみが渡されている.

4.6 結果 2 ：半教師あり学習の有効性検証

学習の流れを可視化するために, 半教師あり学習に基づくデータ拡張による性能向上が確認された, ES2009 の WavLM 第 4 層の出力をもとにした話者ダイアライゼーションの様子を表す. 図 7 には, 距離学習による, 学習データの話者クラスタの生成の様子を UMAP [20] を用いて示す. これを見るに, 距離学習によって, 話者のクラスタをより表現する特徴量変換器を作成できて

いることが確認できる．また，紐状の特徴量が確認できるが，これはガウシアンフィルタを適用したことによる，特徴量の時間連続性の表れである．図 8 には，半教師あり学習における疑似話者ラベル付与の様子を表す．灰色の分布は，距離学習によって得られたモデルに通したラベルなし特徴量を表し，青の×印は各クラスタの重心であり，ラベル付きデータからなる色のついた各クラスタ上のピンク色の点は，疑似ラベルを貼ったデータを表す．以上のように，ラベルなしデータのクラスタリングを行うことで，疑似ラベル付きのデータを作成していることが確認できる．

表 4 は，半教師あり学習 (Semi-Supervised Learning) に基づくデータ拡張の有無による，各層の性能比較を表す．DER は値が小さいほど，性能が高いことを示す．半教師あり学習を適用しない場合は，0 層目を用いる場合が最も性能が高くなっているが，半教師あり学習を適用した場合は，0 層目と 4 層目周辺で性能が高くなっている．全体としては，半教師あり学習の導入による，大幅な性能向上は見られない．これは，600 秒という少ないラベル付きデータでは，普遍的な話者情報を抽出することが困難であり，データ拡張をうまく適用できなかったため，また，ラベルなしデータの量に限りがあったため，信頼度の低い疑似データを付与せざるを得なかったためであると考えられる．しかし，ES2006, 2009 の 4 層目周辺で半教師あり学習によって性能向上がみられ，これらは半教師あり学習を用いない場合の最高性能と比べても高い性能である．これは，ラベル付きデータとして与えた音声が，話者区間が明瞭であり扱いやすかったためであると考えられる．加えて，少ない情報から，学習を行わなければならない今回のタスクにおいて，半教師あり学習を導入したことにより，0 層目から得られる直接的な音素情報ではなく，データ拡張により，4 層目から得られる抽象的なメタ情報を含んだ音声特徴量を扱うことが可能であると考えられる．また，表 5 は，各手法における最高性能をまとめたものであり，評価の基準となる pyannote の話者ダイアライゼーションモデル [19][18] の性能も示した．半教師あり学習によるデータ拡張の導入によって，ES2006 では 0.87 ポイント，ES2009 では 3.54 ポイント性能が向上した．また，End-to-End の比較モデルに比べて，ES2006 では 2.94 ポイント，ES2009 では 0.77 ポイントの性能向上を確認した．加えて，図 9 には，ES2009d の正解データと，提案手法による話者ダイアライゼーションの予測結果を表す．これらより，少量のラベル付きデータだけでも，モデル，パラメータを適切に選ぶことができれば，提案手法を用いて，End-to-End 話者ダイアライゼーションモデルよりも高い性能の話者ダイアライゼーションを行うことが可能であることがわかる．

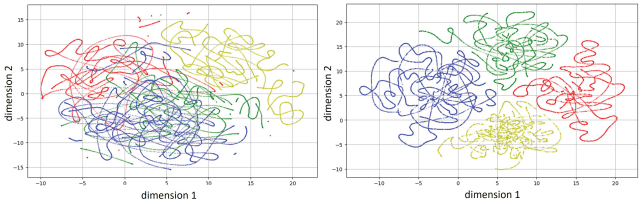

図 7: Speaker clusters in training data before and after metric learning

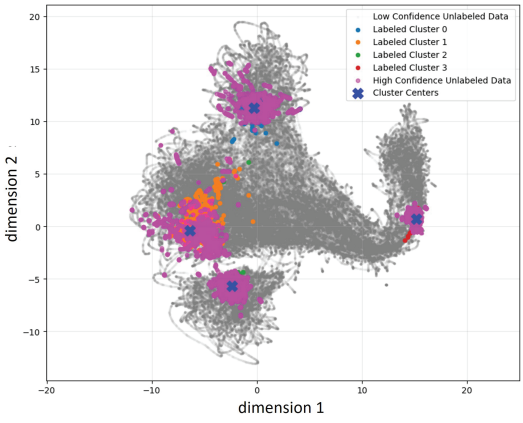

図 8: Pseudo-label generation based on clustering

5 おわりに

本稿では，入力データのシーンに適応できる高性能な話者ダイアライゼーションを実現することを目指し，以下の 2 手法を導入したオフライン話者ダイアライゼーションモデルを提案した:

1) 一部のアノテーション済みデータからでも適切に特徴量を抽出できるように，話者基盤モデル (WavLM)，2) アノテーションのコストを抑えるための，半教師あり学習に基づくデータ拡張．

複数話者対話の会議録コーパスである AMI コーパスを用いて，提案手法を評価したところ，半教師あり学習の導入によって，DER で最大 3.54 ポイント改善した．これは，pyannote に実装されている End-to-End のダイアライゼーションモデルよりも高い性能であり，提案手法の有効性を示すことができた．今後の課題として，話者人数に柔軟な手法の検討が挙げられる．

謝辞

議論でフィードバックをいただいた HRI-JP の周藤唯博士に感謝する．

表 4: Comparison of DER with and without data expansion based on semi-supervised learning

	Without Semi-Supervised Learning				With Semi-Supervised Learning			
layer	ES2006	ES2008	ES2009	ES2015	ES2006	ES2008	ES2009	ES2015
0	**37.86**	**35.62**	**36.05**	**41.34**	38.76	**36.66**	35.46	51.11
1	44.87	41.4	40.45	51.22	44.45	52.89	45.27	58.78
2	42.53	52.52	42.92	55.87	43.03	58.65	43.48	67.12
3	41.2	49.93	40.79	49.89	39.46	49.68	39.04	**51.03**
4	42.15	59.25	37.57	46.78	**36.99**	53.06	**32.51**	56.6
5	44.17	57.12	40.55	48.31	37.87	59.54	38.02	60.35
6	49.52	59.25	40.84	48.6	39.46	65.06	35.54	58.04

表 5: Comparison of minimum DER for each method

	ES2006	ES2008	ES2009	ES2015
pyannote	39.93	44.92	33.28	**39.56**
without SSL	37.86	**35.62**	36.05	41.34
with SSL	**36.99**	36.66	**32.51**	51.03

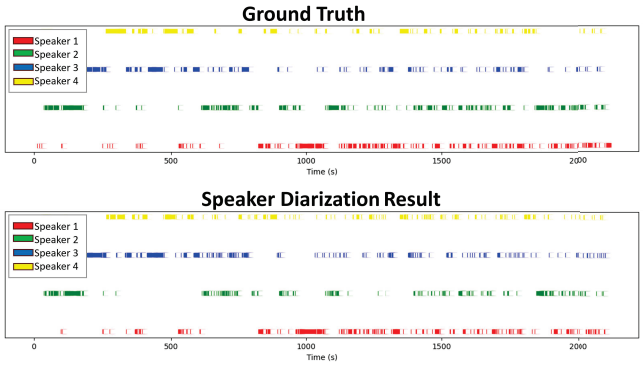

図 9: Speaker diarization output of the proposed method (ES2009d)

参考文献

[1] Fujita, Y., et al,: End-to-End Neural Speaker Diarization with Self-attention, *ASRU 2019*, pp.296–303 (2019)

[2] Ryant, N., et al,: The Second DIHARD Diarization Challenge: Dataset, task, and baselines, *arXiv preprint arXiv:1906.07839* (2019)

[3] Kinoshita, K., Delcroix, M., Tawara, N.,: Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds, *ICASSP 2021*, pp.7198–7202 (2021)

[4] Horiguchi, S., et al,: Encoder-decoder based attractors for end-to-end neural diarization *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.30, pp.1493–1507 (2022)

[5] Yamashita, N., Horiguchi, S., Homma, T.,: Improving the Naturalness of Simulated Conversations for End-to-End Neural Diarization, *Odyssey 2022*, pp.133–140 (2022)

[6] Baevski, A., et al,: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, *Advances in neural information processing systems*, Vol.33, pp.12449–12460 (2020)

[7] Chen, S., et al,: WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing *IEEE Journal of Selected Topics in Signal Processing*, Vol.16, No.6, pp.1505–1518 (2022)

[8] Vaessen, N., Leeuwen, D.,: Fine-tuning wav2vec2 for speaker recognition, *ICASSP 2022* pp.7967-7971 (2022)

[9] Yang, S., et al,: SUPERB: Speech processing Universal PERformance Benchmark, *arXiv preprint arXiv:2105.01051*, (2021)

[10] Chung, Y., and Hsu, W., Tang, H., Glass, J.,: An unsupervised autoregressive model for speech representation learning, *INTERSPEECH 2019*, pp.146–150 (2019)

[11] Ashihara, T., et al,: What Do Self-Supervised Speech and Speaker Models Learn? New Findings from a Cross Model Layer-Wise Analysis, *ICASSP 2024*, pp.10166–10170 (2024)

[12] Hsu, W., et al.,: HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, *IEEE/ACM transactions on audio, speech, and language processing*, Vol. 29, pp.3451–3460(2021)

[13] Chen, S., et al.,: UniSpeech-SAT: Universal Speech Representation Learning with Speaker Aware Pre-Training, *ICASSP 2022*, pp.6152-6156(2022)

[14] Koch, G., Zemel, R., Salakhutdinov, R.,: Siamese neural networks for one-shot image recognition *ICML deep learning workshop* Vol. 2. No. 1. pp1–30 (2015)

[15] Carletta, J., et al.,: The AMI meeting corpus: A pre-announcement, *International workshop on machine learning for multimodal interaction*, pp.28–39 (2006)

[16] Landini, F., Profant, J., Diez, M., Burget, L.,: Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks, *Computer Speech & Language*, Vol.71, pp.101254 (2022)

[17] Bredin, H.,: Pyannote.metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems, *INTERSPEECH 2017,* pp. 3587–3591 (2017), http://pyannote.github.io/pyannote-metrics

[18] Plaquet, A., Bredin, H.: Powerset multi-class cross entropy loss for neural speaker diarization, *arXiv preprint arXiv:2310.13025*, (2023)

[19] Bredin, H.,: pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe, *INTERSPEECH 2023*, pp.1983–1987 (2023)

[20] McInnes, L., Healy, J., James Melville, J.,: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *arXiv preprint arXiv:1802.03426*, (2018)