

# Attentive Pooling を用いた話者照合における 話者埋め込みのアンサンブルの検討\*

☆田中康平, 塩田 さやか (東京都立大学)

## 1 はじめに

話者照合 (Automatic speaker verification; ASV) は発話ペアが同一人物によるものか否かを判別するタスクである. 近年, 大規模データセットを用いた自己教師あり学習により得られた事前学習モデル (Pre-trained model; PTM) [1] を ASV モデルのフロントエンドとして用いることで, 従来のフィルタバンクなどの特徴量と比較して, 性能が向上することが示されている [2]. これらの手法では, バックエンドに ECAPA-TDNN[3]などをベースとしたモデルが採用されている.

PTM をフロントエンドとして用いる ASV 手法の一つとして, Attentive pooling のみから構成される軽量のバックエンドである Multi-Head Factorized Attentive Pooling (MHFA) [4] を利用する手法が提案されている. この手法は, ECAPA-TDNN をバックエンドに用いる手法に匹敵する性能を達成することが報告されている. また, バックエンドの ECAPA-TDNN からフレームレベル特徴抽出器を除去する軽量化を行っても同等の性能が維持できることも報告されている [5]. これらの軽量のバックエンドは, 従来の ECAPA-TDNN ベースの手法と比較して学習効率が高く, 少量データでの学習や学習時間の短縮が可能であると報告されている.

PTM の隠れ層では, 層ごとにレベルの異なる特徴抽出が行われていることが報告されており [6], 異なる抽象化レベルの特徴量を利用することで性能向上を図る手法も提案されている [7]. ただし, これらの手法はいずれもバックエンドに ECAPA-TDNN ベースのモデルを利用しており, 軽量バックエンドで PTM の各層の情報を活用する手法は十分に検討されていない. そこで本研究では, 従来法である MHFA による軽量な話者照合において, PTM の各層が持つ異なる抽象化レベルの特徴量を活用するための話者埋め込みのアンサンブル手法について提案する. 提案法では, ネットワーク構造をあまり複雑化することなく話者照合に有効と言われている音声波形に近いレベルの特徴量も個別に活用できることから, 軽量な話者照合モデルの中でも性能の向上が期待できる. 実験の結果, 提案法は従来法と比較して, 約 10% のパラメータ増加のみで約 13% の EER 改善を達成したことを報告する.

## 2 Multi-Head Factorized Attentive Pooling を用いる話者照合

従来法である PTM のバックエンドとして MHFA を用いる話者照合について説明する. MHFA は PTM によって抽出されたフレームレベル特徴表現を音響単位ごとにクラスタ化して個別にプーリングする手法である. MHFA における注意機構を Query-Key-Value の抽象化に基づいて説明する. 式 (1) は, PTM のフレームレベル音声表現を, 音韻情報を含む Key, 話者情報を含む Value に分解するための処理を表している.

$$\mathbf{K} = \left( \sum_{l=1}^L w_l^k \mathbf{Z}_l \right) \mathbf{S}^k, \quad \mathbf{V} = \left( \sum_{l=1}^L w_l^v \mathbf{Z}_l \right) \mathbf{S}^v \quad (1)$$

式 1 により, PTM の各層の隠れ表現  $\mathbf{Z}_l \in \mathbb{R}^{T \times F}$  を Key および Value で独立した重みベクトル  $\mathbf{w}^k \in \mathbb{R}^L$  と  $\mathbf{w}^v \in \mathbb{R}^L$  で重み付けされた加重和により集約する. その後, 集約された隠れ表現を  $\mathbf{S}^k \in \mathbb{R}^{F \times D}$  および  $\mathbf{S}^v \in \mathbb{R}^{F \times D}$  による線形変換で次元削減して,  $\mathbf{K} \in \mathbb{R}^{T \times D}$  および  $\mathbf{V} \in \mathbb{R}^{T \times D}$  を計算する. ここで,  $T$  は総フレーム数,  $F$  は各フレームの特徴次元,  $D$  は削減後の次元数,  $L$  は事前学習モデルの層数を表す. 各アテンションヘッドでは, 学習可能なクエリ行列  $\mathbf{Q} \in \mathbb{R}^{D \times H}$  を用いて以下の処理を行う.

$$\begin{aligned} \mathbf{A} &= \text{softmax}(\mathbf{K}\mathbf{Q}) \\ \mathbf{c}_h &= \sum_{t=1}^T \mathbf{A}_{t,h} \mathbf{V}_t \\ \mathbf{c} &= \text{concat}(\mathbf{c}_1, \dots, \mathbf{c}_H) \end{aligned} \quad (2)$$

ここで,  $H$  はアテンションヘッドの数,  $\mathbf{A}_{t,h}$  は  $h$  番目の head における時刻  $t$  のフレームに対するアテンション重み,  $\mathbf{c}_h \in \mathbb{R}^{1 \times D}$  は各 head の出力を表す. 各 head のクエリが特定の音響単位に対応するフレームを選択的に集約することで, 音韻内容に応じた話者表現  $\mathbf{c}_h$  を得る. 最終的に全ての head の出力を結合した話者表現  $\mathbf{c} \in \mathbb{R}^{1 \times HD}$  を  $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{HD \times d_{\text{spk}}}$  で線形変換して  $d_{\text{spk}}$  次元の話者埋め込み  $\mathbf{e} \in \mathbb{R}^{1 \times d_{\text{spk}}}$  を得る.

$$\mathbf{e} = \mathbf{c} \cdot \mathbf{W}_{\text{emb}} \quad (3)$$

つまり, 従来法では, 話者情報を含む Value 行列  $\mathbf{V}$  を, PTM の各層の隠れ表現の学習可能な重みベクトル

\*Ensemble of Speaker Embeddings in Attentive Pooling-Based Speaker Verification. by TANAKA, Kohei, SHIOTA, Sayaka (Tokyo Metropolitan University)

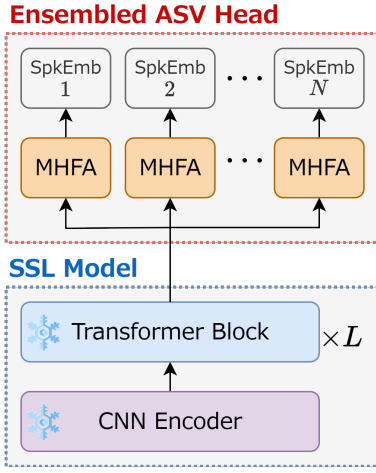


Fig. 1 提案手法の概要

ル  $w^v$  による加重和で計算し、注意機構によるプーリングを行って話者埋込みを得ていた。文献 [4] では、この重みベクトル  $w^v$  が PTM の一部の層に高い重みを割り当てる傾向が示されている。

### 3 提案手法

#### 3.1 MHFA のアンサンブル

本研究では、MHFA を基に PTM の複数の層の情報を効果的に活用するためのアンサンブル手法を提案する。提案法では、Fig. 1 のように複数の MHFA を並列に構成し、各出力をアンサンブルする機構を導入する。Fig.2 のように、各 MHFA モジュールは PTM の異なる層から話者情報を抽出する。Fig.2 の  $K'_n$  と  $V'_n$  は以下の式で計算される PTM 隠れ表現の加重和である。

$$K'_n = \sum_{l=1}^L w_{n,l}^k Z_l, \quad V'_n = \sum_{l=1}^L w_{n,l}^v Z_l \quad (4)$$

ここで、各モジュールが異なる層を重視するように  $w_{n,l}^v$  を学習することにより、PTM のより多くの層から話者情報を収集できる。モジュールの数を  $N$  としたとき、 $N = 1$  の場合は従来の MHFA と等価である。以降の 3.2 節および 3.3 節では、 $w_{n,l}^v$  の多様化を実現する手法を説明する。

#### 3.2 マスク付き重みによる層選択

各 MHFA モジュールが、PTM の異なる層に注目するように強制する制約を課すことで、重みベクトル  $w_{n,l}^v$  の多様化を実現する手法について述べる。 $n$  番目の MHFA モジュールの重みベクトル  $w_n^v \in \mathbb{R}^L$  に対して、一部の要素を 0 にするマスクベクトル  $m_n \in \{0,1\}^L$  を用いる。具体的に  $w_n^v$  は以下のように計算される。

$$w_n^v = m_n \odot \alpha_n \quad (5)$$

ここで、 $\alpha_n \in \mathbb{R}^L$  は学習可能なパラメータであり、 $\odot$  は要素ごとの積を表す。マスクベクトル  $m_n$  の要

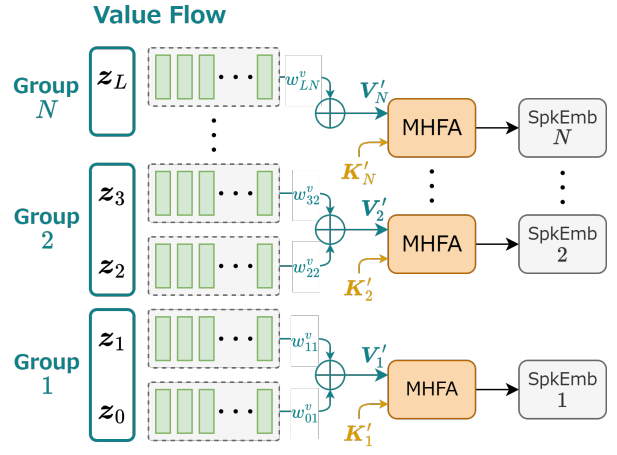


Fig. 2 PTM の多層情報を活用する並列な MHFA 素は、対応する層を使用する場合は 1、使用しない場合は 0 に設定される。この操作により、各 MHFA モジュールは指定された層のみから話者情報を抽出するよう制約される。これにより、グループ毎に均等に話者情報が活用される。Fig. 2 は、PTM の層を連続する 2 層のグループに分割するようにマスクベクトルを設定している。

#### 3.3 多様性ペナルティによる層選択

前節の固定マスクを用いる手法では、データに基づく最適な層の組み合わせの学習ができず、マスク設定に性能が大きく依存する可能性がある。そこで、損失関数に多様性ペナルティ項を追加し、各 MHFA モジュールが学習により自動的に異なる層の組み合わせに注目するよう誘導する手法を提案する。各 MHFA モジュールの重みベクトル  $w_n^v$  について、その絶対値  $|w_n^v|$  を PTM の各層に対する重要度と解釈する。重みの符号ではなく重要度（影響の大きさ）に着目するため絶対値を用いる。この重要度ベクトルをモジュール間で多様化するため、コサイン類似度の総和を最小化するペナルティ項を導入する。ペナルティ項  $\mathcal{L}_d$  を以下のように定義する。

$$\begin{aligned} \hat{w}_n^v &= \frac{|w_n^v|}{\|w_n^v\|_2}, \quad n = 1, 2, \dots, N \\ \hat{W} &= [\hat{w}_1^v, \hat{w}_2^v, \dots, \hat{w}_N^v]^T \in \mathbb{R}^{N \times L} \\ S &= \hat{W} \hat{W}^T \\ \mathcal{L}_d &= \lambda \sum_{i,j} S_{i,j} \cdot (1 - \delta_{i,j}) \end{aligned} \quad (6)$$

ここで、 $\delta_{i,j}$  はクロネッカーのデルタ、 $\lambda$  はペナルティの強度を制御するハイパーパラメータである。 $(1 - \delta_{i,j})$  により異なるモジュール間の類似度のみを対象とする。このペナルティ項を最小化する過程で、各モジュールは他とは異なる層の組み合わせに注目するよう誘導される。

### 3.4 話者埋め込みの分割によるパラメータ効率化

最終的に、 $N$  個の MHFA モジュールから出力される話者埋め込みを連結し、統合された話者埋め込みを得る。各モジュールが独立に  $d_{\text{spk}}$  次元の埋め込みを生成すると、パラメータ数が  $N$  倍に増加する問題がある。そのため、各 MHFA モジュールが  $d_{\text{spk}}$  次元の埋め込みの一部分のみを生成する手法を採用する。具体的には、各モジュールは  $\lceil \frac{d_{\text{spk}}}{N} \rceil$  次元の部分埋め込み（サブ埋め込み）を出力し、これらを連結して  $d_{\text{spk}}$  次元の話者埋め込みを構成する。

$$\mathbf{e}_{\text{final}} = \text{concat}(\mathbf{e}_1, \dots, \mathbf{e}_N) \quad (7)$$

ここで、 $\mathbf{e}_n \in \mathbb{R}^{\lceil d_{\text{spk}}/N \rceil}$  は  $n$  番目の MHFA モジュールが出力する部分埋め込みである。MHFA の主要パラメータである  $\mathbf{S}^k$ ,  $\mathbf{S}^v$ ,  $\mathbf{Q}$  のパラメータ数はそれぞれ  $F \times D$ ,  $F \times D$ ,  $D \times H$  であるのに対し、線形変換  $\mathbf{W}_{\text{emb}}$  のパラメータ数は  $D \times H \times d_{\text{spk}}$  である。 $d_{\text{spk}}$  は通常 128~512 次元であるため、 $\mathbf{W}_{\text{emb}}$  が全パラメータの大部分を占める。提案手法では各モジュールの線形変換が  $\mathbf{W}_{\text{emb},n} \in \mathbb{R}^{H \times \lceil d_{\text{spk}}/N \rceil}$  となるため、この層のパラメータ数を約  $1/N$  に削減でき、アンサンブルによるパラメータ増加を大幅に軽減できる。

## 4 実験

提案するアンサンブル手法の有効性を確認するために、従来の MHFA[4] との比較実験を行った。

### 4.1 実験条件

実験条件は、SUPERB ベンチマーク [8] に準拠している。学習と評価には、VoxCeleb1 データセット [9] を使用し、SUPERB と同様にデータ拡張は行わなかった。検証データは、訓練データからランダムに抽出した 20 話者分の音声を使用した。実験に使用した PTM は WavLM Base+[10] で、12 層の Transformer ブロックから構成される ( $L = 12$ )。PTM の事前学習済みパラメータは固定し、バックエンドのみを学習した。学習には AdamW オプティマイザーを使用し、バッチサイズ 128, 最大 20 エポック学習を行った。損失関数には Additive Angular Margin Loss[11] を使用し、 $\text{scale} = 30.0$ ,  $\text{margin} = 0.2$  に設定した。話者埋め込みの次元は 512 次元である。学習率は  $1e-4$  から  $1e-2$  の範囲で探索し、重みの多様性ペナルティ項の係数  $\lambda$  も 0.1 から 20.0 の範囲で探索した。それぞれ検証データで最良の値を採用した。評価は、異なるシード値で学習した 6 個のモデルで行い、平均と標準偏差を算出した。

### 4.2 ベースラインモデル

以下の 3 つのベースラインモデルを設定し、提案法の評価を行う。

Table 1 提案法のベースラインとの比較

Backend model	Params	EER [%] ↓
Baseline		
MHFA	4.34M	1.83 ± 0.08
MHFA[+hidden256]	8.63M	1.76 ± 0.08
MHFA[+attn128]	8.59M	1.93 ± 0.19
Proposed		
MHFA4[Fixed group]	4.78M	1.63 ± 0.12
MHFA4 ( $\lambda = 0.0$ )	4.78M	1.70 ± 0.10
MHFA4 ( $\lambda = 10.9$ )	4.78M	1.60 ± 0.07
MHFA2 ( $\lambda = 13.0$ )	4.49M	1.76 ± 0.17
MHFA8 ( $\lambda = 7.8$ )	5.37M	1.64 ± 0.10

**MHFA (基本ベースライン)**: 文献 [4] において最良の結果が得られた  $D = 128$ ,  $H = 64$  の設定を使用する。

**MHFA[+hidden256]**: 隠れ表現の次元を  $D = 256$  に増強し、表現能力の向上により性能改善を図る。

**MHFA[+attn128]**: アテンションヘッド数を  $H = 128$  に増強し、より細粒度な音響単位の捕捉による性能改善を図る。

後者 2 つは、提案手法が約 10% のパラメータ増加を伴うため、比較の公平性を確保するために追加した。

### 4.3 アンサンブル構成

実験では、モジュール数を  $N = 2, 4, 8$  として、3.3 節のペナルティ項を使用する提案法をそれぞれ、MHFA2, MHFA4, MHFA8 と記述する。また、MHFA4[Fixed group] は、 $N = 4$  の条件で、3.2 節の手法を使用する条件である。この条件では、各 MHFA モジュールが利用できる層のグループを、 $G_1 = \{0, 1, 2, 3\}$ ,  $G_2 = \{4, 5, 6\}$ ,  $G_3 = \{7, 8, 9\}$ ,  $G_4 = \{10, 11, 12\}$  とする。各 MHFA モジュールはそれぞれ、 $G_1$  から  $G_4$  に含まれる層のみ利用するようにマスクベクトル  $\mathbf{m}_n$  を設定した。上記において、0 は CNN Encoder の出力層、1 から 12 は Transformer ブロックの各層を表す。このグループ設定により、従来法で低い重みが割り当てられていた、前半・後半の層からも個別に話者情報を抽出し、アンサンブルにより性能向上に貢献させることを狙う。

## 4.4 実験結果

### 4.4.1 提案法とベースラインの比較

Table 1 に提案法とベースラインの比較結果を示す。表には、6 つのシード値で学習したモデルの平均値と標準偏差を記載している。丸括弧にペナルティ項の係数  $\lambda$  を示す。最も EER が低かった MHFA ( $\lambda = 10.9$ ) では、ベースラインの MHFA と比較して、約 10% のパラメータ増加に対して約 13% の EER 改善を達成した。隠れ層の次元やアテンションヘッドを増強し

Table 2 MHFA4[Fixed group] におけるサブ埋め込みの EER

Layer Group	EER [%] ↓
$G_1$	$2.05 \pm 0.17$
$G_2$	$1.80 \pm 0.16$
$G_3$	$2.20 \pm 0.22$
$G_4$	$2.24 \pm 0.21$

た MHFA[+hidden256], MHFA[+attn128] に対しても、提案法がより低い EER を示した。ペナルティ項を使用する MHFA4 ( $\lambda = 10.9$ ) と固定グループを使用する MHFA4[Fixed group] では、EER に大きな差はなかった。また、 $\lambda = 0.0$  の条件 (ペナルティなし) と比較すると、ペナルティ項により性能が向上していることが確認できる。モジュール数  $N$  による性能変化については、 $N = 4$  の場合が最も EER が低く、 $N = 2$  から  $N = 4$  に増加すると EER は減少するが、 $N = 4$  から  $N = 8$  に増加しても EER は改善されなかった。これは、512 次元の話者埋め込みを各 MHFA モジュールで分割して生成するため、 $N=8$  の場合には各サブ埋め込みの次元が 64 に減少し、表現能力が制限されることが原因と考えられる。

#### 4.4.2 PTM の層グループ別の性能比較

Table. 2 に MHFA4[Fixed group] において、各モジュールが出力するサブ埋め込みを用いて EER を算出した結果を示す。 $G_2$  (4-6 層) から抽出された話者埋め込みが最も EER が低く、これは文献 [4] において Value の重みベクトル  $w^v$  のピークが 5 層目であることと一致している。PTM の前半層 ( $G_1$ ) から抽出された話者埋め込みも  $G_2$  に次ぐ性能を示し、後半層 ( $G_3, G_4$ ) からの埋め込みも 2% 台の EER を達成している。Table 1 の MHFA4[Fixed group] は、これらのサブ埋め込みを全て結合したアンサンブル埋め込みでの結果であり、各サブ埋め込みよりも低い EER を達成しており、アンサンブルの効果が確認された。特に、 $G_2$  から抽出された話者埋め込みの EER は、Table 1 のベースラインである MHFA よりも低い値を示している。これは、従来法のように全ての層からの隠れ表現を単純な加重和で統合することで、話者識別性能が高い層の情報が他の層の情報と混合して性能が低下する可能性があることを示唆している。提案法では、層を分離してプーリングし、アンサンブルすることでこのような悪影響を回避しながら性能向上を実現していると考えられる。

## 5 むすび

本研究では、軽量バックエンドの MHFA において、PTM の各層が持つ多様な情報を効果的に活用するための話者埋め込みのアンサンブル手法を提案した。

提案法は、層グループごとに独立した MHFA モジュールを用いることで、PTM の多層情報を活用する。VoxCeleb1 での評価において、提案法は約 10% のパラメータ増加のみで約 13% のエラー改善率を達成した。今後の課題としては、PTM ファインチューニング時の有効性検証と、多様なデータセットでの評価があげられる。

## 参考文献

- [1] Alexei Baevski, et al. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proc. NIPS*, 2020.
- [2] Zhengyang Chen, et al. Large-scale self-supervised speech representation learning for automatic speaker verification. In *Proc. ICASSP*, pp. 6147–6151, 2022.
- [3] Brecht Desplanques, et al. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proc. Interspeech*, pp. 3830–3834, 2020.
- [4] Junyi Peng, et al. An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification. In *Proc. SLT*, pp. 555–562, 2023.
- [5] Zakaria Aldeneh, et al. Can you remove the downstream model for speaker recognition with self-supervised speech features? *arXiv preprint arXiv:2402.00340*, 2024.
- [6] Ankita Pasad, et al. Layer-wise analysis of a self-supervised speech representation model. In *Proc. ASRU*, pp. 914–921, 2021.
- [7] Shengyu Peng, et al. Fine-tune pre-trained models with multi-level feature fusion for speaker verification. In *Proc. Interspeech*, pp. 2110–2114, 2024.
- [8] Shu-wen Yang, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.
- [9] Arsha Nagrani, et al. Voxceleb: A large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.
- [10] Sanyuan Chen, 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.
- [11] Jiankang Deng, et al. Arcface: Additive angular margin loss for deep face recognition. In *Proc. CVPR*, pp. 4690–4699, 2019.