基於多重注意力機制的輔助損失函數用於端到端語者標記
Auxiliary Loss to Attention Head for End to End Speaker Diarization

Yi-Ting Yang 楊憶婷, Jiun-Ting Li 李俊廷, Berlin Chen 陳柏琳
國立臺灣師範大學資訊工程學系
Department of Computer Science and Engineering,
National Taiwan Normal University
{61147070s, 60947036s, berlin} @ntnu.edu.tw

摘要

本研究提出新穎的輔助函數用於自注意力端到端語者自動標記模型(SA-EEND)，實現在重疊語音區域進行準確的語者標籤預測。過去的研究缺乏充分利用模型中的語者信息以增強輔助模型訓練的方法，並且未考慮到不同語音活動模式(speech activity patterns)的數量分佈差異。本研究提出了一種新穎的輔助函數，以實現在重疊的語音區域中對語者標籤的預測。通過整體語音活動模式以及不同語者的語音活動模式任務，我們調整了 Transformer 層中的注意力機制(multi-head self-attention)的權重矩陣，並且挑選損失函數能夠加強數量較少的標籤的學習效果，以達到更好的語者辨別效果。本研究在 Mini LibriSpeech 上進行了實驗，雖然成果稍微有限，但仍然取得了一些進展。

Abstract

This study introduces a novel auxiliary function for use in the Self-Attention End-to-End Speaker Diarization (SA-EEND) model, aiming to achieve accurate speaker label prediction within overlapping speech regions. Previous research has lacked effective methods for leveraging speaker information within the model to enhance auxiliary model training and has not taken into account variations in the distribution of different speech activity patterns. This study proposes a novel auxiliary function to facilitate speaker label prediction within overlapping speech regions. By considering both the overall speech activity patterns and the task-specific speech activity patterns for different speakers, we adjust the weight matrices of the multi-head self-attention mechanism in the Transformer layers. We also select loss functions that can improve the learning performance for labels with fewer occurrences, resulting in better speaker discrimination. Experimental evaluations were conducted on Mini LibriSpeech. Although the results exhibited some limitations, there were still notable advancements made.

關鍵詞：語者標記、端到端語者標記、注意力機制、輔助損失函數

Keywords: speaker diarization, end-to-end neural diarization, multi-head attention, auxiliary loss

1 介紹

語者標記(speaker diarization)是一個處理 who speak when 的任務，旨在音訊中標記出同一位語者的片段，語者標記可以應用在許多的場景，例如:廣播採訪、會議(Janin et al., 2003)(Renals et al., 2008)、電話(Kenny et al., 2010)、面試或醫療記錄等，也能夠幫助多個語者情境下的語音辨識。早期的語者標記依賴於手動標註和簡單的基於規則的系統。然而，隨著機器學習技術的出現，該領域取得了顯著的進展，(Imseng et al.,2000)提出了基於高斯混和模型(Gaussian Mixture Model)的語者標註系統。然而基於神經網絡的方法改變了語者標記，(Dehak et al.,2011)提出了 i-vector 框架，提高了辨識準確性。後來(Garcia-Romero et al.,2017)提出了利用深度學習的端到端語者標記方法。

(Fujita et al., 2019)提出了端到端語者自動分段標記(end-to-end neural speaker diarization, EEND)方法，在輸入一個多語者的音頻錄音時，直接輸出每個

時間幀中所有語者的聲音活動信息。具體來說，自注意力機制(self-attention, SA)-EEND 模型(Fujita et al., 2019)由多個 Transformer(Vaswani et al., 2017)層組成，並將每個時間幀的語者後驗概率作為輸出。SA-EEND 模型是使用僅二元交叉熵(binary cross entropy)損失來訓練的，二元交叉熵衡量了從模型最後一層所產生的輸出和真實標籤之間的差異，並通過置換不變訓練(Fujita et al., 2019)來訓練模型。然而，這樣訓練的 SA-EEND 模型在學習過程中並未充分利用語者的資訊，僅僅依賴於最終輸出層的損失優化，沒有適當地引導學習過程(Yu et al., 2022)，導致注意力權重矩陣傾向為單位矩陣(Jeoung et al., 2023)，使得在訓練的過程中無法幫助模型區分不同的說話者與無語音部分。

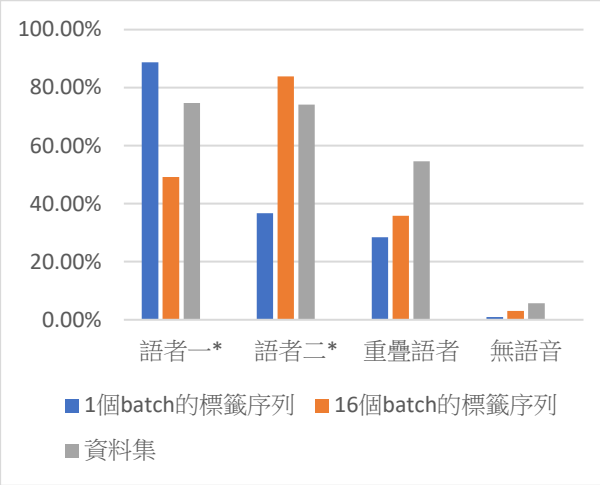
在本研究中，我們關注多語者情況下的單一語者導向的語音活動模式(voice activity pattern)和整體語音模式(overall speech pattern)，這兩者對於達到良好的語者標籤預測是至關重要的。使用真實的單一語者標籤序列與整體語者標籤序列來限制 SA-EEND 模型的注意力權重矩陣，作為輔助損失，並且使用焦點損失(focal loss)(Lin et al., 2017)作為損失函數，以解決在訓練集中存在的語音樣本和無語音樣本不平衡問題，幫助模型更有效地學習中間表示。

2 資料集

LibriSpeech(Panayotov et al., 2015)資料集是一個包含英語朗讀語音的新語料庫，適用於訓練和評估語音分離和語音識別系統。Mini LibriSpeech 是 LibriSpeech 語料庫的一個子語料集，其中 Mini LibriSpeech 語料庫包含 54 位語者的約 2600 個語音片段(Chen et al., 2020)。

Mini LibriSpeech 的構建方法是將 LibriSpeech 的數據分割成訓練集、驗證集和測試集，並在這些集合中均勻選擇不同信噪比的語音混合，這有助於模擬真實世界中的環境噪音，使得 Mini LibriSpeech 成為一個有用的小型語音數據集，特別適合在資源受限的情況下進行語音相關研究和開發。

從 Mini LibriSpeech 資料集產生的標籤序列中，存在不特定的兩位語者情境。如表格 1 中所示，在 16 個批次(batch)中，單一語者分別是整個資料集的 49.2%和 83.9%。在這些語音樣本中，約有 35.8%的樣本涵蓋兩位語者的重疊聲音，同時還有極少數約 3%的樣本是無語音的；在整個資料集



表格 1: Mini Librispeech 的標籤序列在不同批次數(batch)和整個資料集的語者標籤數量。*:表示不特定語者。

中，單一語者分別佔整個資料集的 74.74%和 74.13%。在這些語音樣本中，大約有 54%的樣本包含兩位語者的重疊聲音，同時還有一小部分約 5.74%的樣本是無語音的。

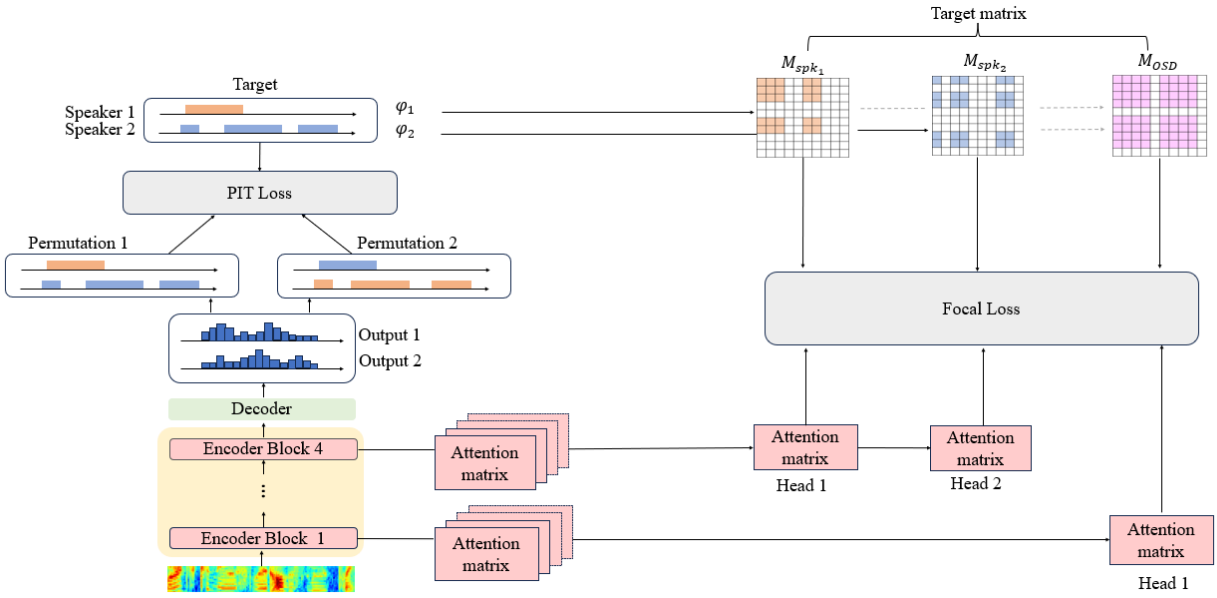
3 方法

3.1 SA-EEND: 回顧

在本節中，我們將簡要敘述 SA-EEND(Fujita et al., 2019)所提出的方法。給定一個長度為T的聲音訊號$X = (x_t \in \mathbb{R}^F | t = 1, \dots, T)$，$x_t$是一個在時間$t$的$F$維特徵向量，語者標記任務可以被歸類為一個多標籤分類問題，將特徵向量序列通過 Transformer 編碼器，當中包括自注意力機制(multi-head self-attention，MHSA)和全連接前向網路(feed forward neural network)，再通過解碼器可以得到在時間幀t上的語者後驗機率。$\hat{y}_t=[\hat{y}_{t,1},\cdots,\hat{y}_{t,S}]$，這些機率值表示屬於不同語者的機率，預測的語者標籤序列與語者標籤進行置換不變訓練(permutation invariant training)，為了訓練 SA-EEND 模型，預測值$\hat{y}_t$和真實標籤y_t之間的損失函數$\mathcal{L}_d$可以如下計算：

$$\mathcal{L}_d = \frac{1}{TS} \min_{\emptyset_1,\dots,\emptyset_S \in \Phi_S} \sum_{t=1}^{T} \sum_{s=1}^{S} BCE(y_{t,s}^{\emptyset_s}, \hat{y}_{t,s}), \quad (1)$$

其中，$BCE(\cdot,\cdot)$代表二元交叉熵(binary cross-entropy)損失。$y_{t,s}$是在時間幀 t 上真實的第 s 個語者的標籤，$y_{\emptyset_s} = [y_{1,s}^{\emptyset_s},\dots,y_{T,s}^{\emptyset_s}] \in \{0,1\}^T$是根據語



圖表 1: 新增輔助損失(auxiliary losses)的 SA-EEND

者排列組合，生成的語者標籤序列。符號$\emptyset_s$代表語者排列組合(speaker permutations)，Φ_s表示所有的排列組合。

3.2 焦點損失: 回顧

焦點損失是一種用於解決類別不平衡問題的損失函數，通常應用於二元分類任務。它在訓練過程中專注於難以區分的樣本，這些樣本可能是少數類別或具有高度困難度的樣本。焦點損失是基於二元交叉熵的一種擴展，它引入了一個額外的可調參數，稱為焦點參數，來加權樣本的損失。

在傳統的二元交叉熵損失中，所有樣本的損失在訓練過程中都被平等地考慮。然而，當存在類別不平衡或難以區分的情況時，這可能導致模型難以有效學習這些困難樣本。然而，當存在類別不平衡或難以區分的情況時，這可能導致模型難以有效學習這些困難樣本。焦樣本的權重，使難以區分的樣本在訓練過程中得到更多的關注。

焦點損失通過引入一個可調參數(稱為焦點因子)來調整損失函數，使模型更關注難以分類的樣本，從而在訓練過程中提高對難樣本的關注度。焦點損失計算方式為：

$$FL(P_t) = -(1-P_t)^{\gamma}\log(P_t), \quad (2)$$

P_t是模型預測的概率，表示樣本屬於正確類別的概率；γ是焦點因子，用於調整難易樣本的權重。當樣本被正確分類(P_t較大)時，焦點損失會減弱交叉熵損失的影響，使模型更專注於難分類的樣本。而當樣本被錯誤分類(P_t較小)時，焦點損失會增強交叉熵損失的影響，強化對該樣本的懲罰，從而更關注難分類的情況。

3.3 語者導向的語音活動檢測損失

為了豐富模型對不同語者語音活動模式的學習，我們引入了一種輔助損失，被稱為「語者導向的語音活動檢測損失」(speaker-wise voice activity detection，SVAD)(Jeoung et al., 2023)。我們通過利用每個語者的標籤序列$y_{\emptyset_s}$來建立一個特殊的單一語者標籤矩陣$M_s = y_{\emptyset_s}^{\top} y_{\emptyset_s}$ ($1 \le s \le S$)(這裡的 s 代表不同的語者編號，$1 \le s \le S$)。這些標籤矩陣將用於微調注意力權重矩陣，以更精準地捕捉不同語者的語音活動情況，具體做法如下：

$$L(M_s, A_s^h) = \frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T} F(m_{ij}, a_{ij}), \quad (3)$$

$$\mathcal{L}_S = \sum_{s=1}^{S} L(M_s, A_s^h), \quad (4)$$

其中$F(\cdot,\cdot)$代表焦點損失。注意力權重矩陣(self-attention weight matrix) A_s^h是由第 h 個注意力機制(self-attention head)進行計算而來。在這個過程中，m_{ij}以及a_{ij}分別代表了標籤矩陣M_s與注意力權重矩陣A_s^h中位於第(i,j)位置的元素。

透過將注意力權重矩陣與語者標籤矩陣通過矩陣相乘的方式，引導模型更加關注每位語者的語音活動，並且藉由使用焦點損失使得模型能夠調整樣本的權重，使模型更有效的處理類別不平衡的情況，並更加關注難以分類的樣本，模型能夠增強對不同語者語音活動模式的感知和區分能力，以幫助模型更好地學習語者的語音活動模式。

3.4 整體語音檢測損失

為了幫助自注意力機制學習整體的語音活動式，我們定義了一個輔助損失，稱為整體語音檢測(overall speech detection, OSD)損失。首先，定義OSD 標籤序列 φ = $[\varphi_1, \cdots, \varphi_T]$，具體定義如下：

$$\varphi_T = \begin{cases} 0 & \text{如果 t 是非語音幀} \\ 1 & \text{如果 t 是單一語者或是重疊語者幀,} \end{cases} \quad (5)$$

整體語者標籤矩陣 $M_{OSD} = \varphi^{\top}\varphi$ 被用來定義 OSD 損失。

$$\mathcal{L}_O = \frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T} F(\omega_{ij}, a_{ij}), \quad (6)$$

在上述描述中，ω_{ij} 表示 M_{OSD} 的(i,j)位置的元素，而 a_{ij} 則是應用於 OSD 損失的對應注意力權重矩陣的元素。通過這種方式，我們期望注意力權重能夠區分有語者說話和沒有語者說話的區域。

3.5 使用提出的輔助損失進行模型訓練

語者標註預測損失 $\mathcal{L}_d$ 和輔助損失 $\mathcal{L}_S$ 以及 $\mathcal{L}_O$ 被一起使用，以幫助自注意力機制不僅區分語音的存在，還區分每個語者的語音。最終，用於訓練我們提出的系統的總損失函數定義如下：

$$\mathcal{L}_{Total} = \mathcal{L}_d + a\mathcal{L}_S + b\mathcal{L}_O, \quad (7)$$

其中 a 和 b 是超參數，用於指示應用每個輔助損失的程度。這樣設計總損失函數的目的是讓模型更好地學習特定任務，能夠進行重要度關注，給予語者的元素較高的注意力權重。通過引入自注意力機制的輔助損失，我們希望模型能夠更好地區分語音的不同特徵，並有效率地處理語者標籤預測任務。

3.6 其他相關研究

過去的研究中，有一篇相關的研究(Jeoung et al., 2023)探討了類似的議題。該研究針對基於 Transformer 的端到端語者標記模型進行了改進，並提出了在 Transformer 層的自注意機制中使用輔助損失，並以二元交叉熵與均方誤差(Mean Square Error)做為損失函數，以增強模型的訓練效果。儘管如此，我們的研究與之不同之處在於我們提出了使用焦點損失取代二元交叉熵與均方誤差解決類別不平均的問題。

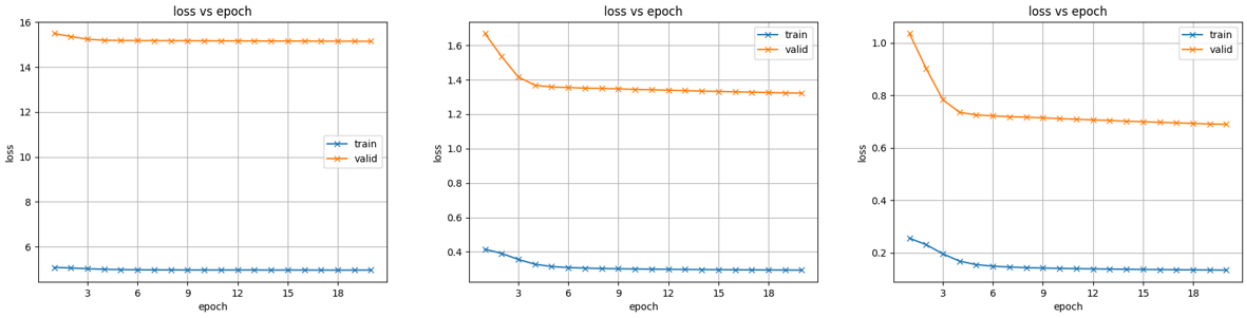
4 實驗設計

基本模型採用了 SA-EEND 模型(Fujita et al., 2019)，其中每個 Transformer 編碼器塊使用了四個 MHSA(multi-head self-attention)。輸入特徵為 23 維的以對數尺度縮放的梅爾濾波器能量，這些特徵是使用 25 毫秒的幀長和 10 毫秒的幀移提取的(Fujita et al., 2019)。每個特徵幀與左側和右側的 7 個幀進行聯接，然後從一個語音片段(utterance)中提取的特徵序列被按照 10 倍的下採樣因子進行時間下採樣(Fujita et al., 2019)，原本的特徵序列時間軸被壓縮，每個時間步長變為原來的 1/10，這意味著在每個時間步長上，只保留了原始特徵序列的十分之一的資訊。所提出的輔助損失被應用於 SA-EEND 模型，如第 3 節所述。在計算 a 和 b(方程式(7))的最優值時，當應用 $\mathcal{L}_S$ 時，a 設置為 1，否則為 0；b 的設置方式相同。SA-EEND 模型使用 Adam 優化器(Kingma et al., 2014)，適應階段的學習率設置為 0.001。最終的語者活動預測使用閾值 0.6 和窗口大小為 11 幀的中值濾波獲取。使用語音解析錯誤率(diarization error rate, DER)(Fiscus et al., 2007)作為評估指標。

5. 實驗結果與分析

5.1 不同損失函數對於總損失($\mathcal{L}_{Total}$)的效果

焦點損失是一種針對稀少類別樣本的重要性進行權重調整的損失函數。這種損失函數使得模型能夠更專注於學習難以區分的類別，並在訓練過程中平衡各個類別的影響。進一步分析我們在第 2 節中資料集的描述，發現靜默標籤(silence label)僅占 5.74%，在 OSD 與 SVAD 標籤序列中，我們只區分了有語音與無語音兩種情況，能夠正確辨識出無語音區域同樣具有重要性。儘管靜默標籤在資料集中僅占 5.74%，但這一部分在我們的研究中同樣具有同等的重要性。由於靜默標籤數量



圖(a): SVAD損失函數:二元交叉熵
OSD損失函數 :均方誤差

圖(b):SVAD損失函數:二元交叉熵
OSD損失函數 :二元交叉熵

圖(c): SVAD損失函數:焦點損失
OSD損失函數 :焦點損失

圖表 2:使用輔助損失的語者標註模型在不同的損失函數的總損失($\mathcal{L}_{Total}$在方程式(7))變化

相對稀少，可能導致模型難以有效辨識這些區域。因此，在訓練過程中適當地處理這些少數樣本將對模型性能的提升至關重要。

圖表 2 是使用輔助損失的語者標註模型在不同的損失函數的總損失($\mathcal{L}_{Total}$在方程式(7))變化，不論是在使用二元交叉熵和均方誤差作為損失函數的模型(Jeoung et al., 2023)(圖(a))或是 OSD 與 SVAD 的損失函數均為二元交叉熵(圖(b))，我們都能觀察到可以看出使用焦點損失(圖(c))作為 OSD 與 SVAD 的損失函數，在驗證數據上呈現出幾個重要的趨勢。

首先，在初始訓練階段，使用焦點損失作為損失函數的模型總損失是使用二元交叉熵和均方誤差作為損失函數的模型總損失的6%，這表明該損失函數有助於模型更快地收斂。此外在訓練過程中，焦點損失所導致的總損失下降速度也相對於使用二元交叉熵和均方誤差所產生的總損失更快。這表示焦點損失能夠引導模型更專注於學習困難的區域，提升模型對於關鍵特徵的捕獲能力。

最終，使用焦點損失的模型在收斂時達到的總損失是在圖表 2 中的三種不同損失函數的模型中最小的。總體而言，圖表 2 呈現出焦點損失作為 OSD 與 SVAD 損失函數的優勢，並且在訓練過程中能夠促使模型更快速地收斂，提升模型的性能。

5.2 輔助損失用於端對端模型效果

我們在 SA-EEND 模型中使用了兩種輔助損失，如方程式(7)所示，其中$a = 1$，$b = 1$。如表格 2 所示，不論是應用 SVAD 損失或是 OSD 損失的 DER 都會比原本 SA-EEND 模型的辨別語者錯誤率減少 1.32%，同時應用 SVAD 損失和 OSD 損失通常會比僅應用其中一種輔助損失的性能更優。

從數據分析中明顯可見，將自注意力機制的資訊融入損失函數，對於訓練模型具有顯著的幫助。這是因為在模型訓練過程中，透過注入更多與語者相關的資訊，無論是單一語者的語音內容，還是兩位語者交互時的語音重疊情況，皆能夠使模型更全面地理解語音場景。

這種將自注意力機制的資訊融入損失函數的方法，能夠使模型更加聚焦於捕捉語者說話的微妙變化，從而提升模型對於語音內容的敏感度。單一語者的說話特徵能夠更深入地被探索和利用，同時在多語者對話中，模型能夠更有效地區分和分離各自的語音輸入，更好地捕捉他們之間的交互細節。

Method	loss		DER
	SVAD	OSD	
SA-EEND †	✘	✘	30.47%
SA-EEND(Focal loss)	Focal Loss	✘	29.15%
SA-EEND(Focal loss)	✘	Focal Loss	29.15%
SA-EEND(Focal loss)	Focal Loss	Focal Loss	29.14%

表格 2: 輔助損失用於端對端模型效果。†: 我們實現的模型

6. 結論

在這項研究中，我們提出使用輔助損失來訓練 SA-EEND 模型，這些輔助損失是通過利用 SVAD 或 OSD 來定義的，這兩者都可以被視為語者標註的重要子任務，為了驗證這一想法，我們提出的輔助損失應用於注意力權重矩陣上。實驗結果表明，所提出的 SVAD 和 OSD 損失都能提升傳統 SA-EEND 模型的性能。

此外我們還探究引入焦點損失對於模型性能的影響，在初始訓練階段，使用焦點損失作為損失函數的模型總損失是使用二元交叉熵和均方誤差作為損失函數的模型總損失的6%，顯示出該損失函數有助於更快的收斂。同時，在訓練過程中，焦點損失所引起的總損失下降速度相對於使用二元交叉熵和均方誤差作為損失函數的模型更快。這表示焦點損失能夠引導模型更專注於學習困難區域，提升模型對關鍵特徵的捕獲能力。

參考文獻

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., ... & Wooters, C. 2003. The ICSI meeting corpus. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP),Vol. 1, pp. I-I.

Imseng, D., & Friedland, G. 2009, November. Robust speaker diarization for short speech recordings. In *IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 432-437).

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. 2010. Front-end factor analysis for speaker verification. In *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788-798.

Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., & McCree, A. 2017. Speaker diarization using deep neural network embeddings. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) (pp. 4930-4934).

Renals, S., Hain, T., & Bourlard, H. 2008. Interpretation of multiparty meetings the AMI and AMIDA projects. In *Proceedings of Hands-Free Speech Communication and Microphone Arrays* (HSCMA), pp. 115-118.

Kenny, P., Reynolds, D., & Castaldo, F. 2010. Diarization of telephone conversations using factor analysis. In *IEEE Journal of Selected Topics in Signal Processing*, *4*(6), 1059-1070

Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., & Watanabe, S. 2019. End-to-end neural speaker diarization with permutation-free objectives. In *Proceedings of Interspeech*.

Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., & Watanabe, S. 2019. End-to-end neural speaker diarization with self-attention. In *IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), pp. 296-303.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. 2017. Attention is all you need. In neural information processing systems (NIPS), 30.

Yu, Y., Park, D., & Kim, H. K. 2022. Auxiliary loss of transformer with residual connection for end-to-end speaker diarization. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp. 8377-8381.

Jeoung, Y. R., Yang, J. Y., Choi, J. H., & Chang, J. H. 2023, June. Improving Transformer-Based End-to-End Speaker Diarization by Assigning Auxiliary Losses to Attention Heads. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), pp. 1-5.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), pp. 2980-2988.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. The ICSI meeting corpus. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), pp. 5206-5210.

Chen, J. J., Mao, Q. R., Qin, Y. C., Qian, S. Q., & Zheng, Z. S. 2020. Latent source-specific generative factor learning for monaural speech separation using weighted-factor autoencoder. In *Frontiers of Information Technology & Electronic Engineering*, 21(11), 1639-1650.

Kingma, D. P., & Ba, J. 2015. Adam: A method for stochastic optimization. In *Proceedings of International Conference for Learning Representations* (ICLR).

Fiscus, J. G., Ajot, J., & Garofolo, J. S. 2007. The rich transcription 2007 meeting recognition evaluation. In *International Evaluation Workshop on Rich Transcription*, pp. 373-389.