Computational Linguistics and Chinese Language Processing
Vol. 28, No.1, June 2023, pp. 53-68

公平語音情緒辨識的雙階段學習策略

A Two-Stage Learning Strategy for Fair Speech Emotion Recognition

簡婉軒 *、李祈均*

Woan-Shiuan Chien, and Chi-Chun Lee

摘要

語音情緒辨識是眾多語音解決方案中的關鍵技術。語音情緒辨識中一個獨特的公平性問題源於評估者提供的數據標註中存在的固有情緒認知偏見。為了讓語音情緒辨識在識別效能與公平性上都提升，我們必須優先解決評估者偏見的問題。在本研究中，我們提出了一種兩階段架構，該架構在第一階段使用公平性約束的對抗性架構來產生去偏見的表徵。隨後，在第二階段的性別感知學習完成後，我們賦予用戶根據需求在特定的性別感知之間自由切換的能力，我們使用兩個重要的公平性指標來評價我們的結果，證明了在性別間的分佈和預測是公平的，進一步的分析指出透過我們的模型在特徵學習空間上有效地避免性別觀點的影響。

關鍵詞：語音情緒辨識、評估者偏見、公平表徵、感知公平性

Abstract

Speech Emotion Recognition (SER) is a key technology within the myriad of speech solutions. A unique fairness issue in SER stems from the inherent emotional perception bias present in the data labels provided by raters. To enhance both recognition performance and fairness in SER, addressing rater bias is paramount. In this study, we propose a two-stage framework. In the first stage, we generate debiased representations using a fairness-constrained adversarial framework. Subsequently, in the second stage, following gender-wise perceptual learning, we

* 國立清華大學電機工程學系
Department of Electrical Engineering, National Tsing Hua University
E-mail: wschien@gapp.nthu.edu.tw; cclee@ee.nthu.edu.tw

empower users with the ability to toggle freely between specific gender-wise perceptions as needed. We utilize two significant fairness metrics to evaluate our results, demonstrating that our distributions and predictions across genders are fair. Further analysis indicates that our model effectively mitigates the influence of gender perspectives in the feature learning space.

Keywords: Speech Emotion Recognition, Rater Bias, Fair Representation, Perceptual Fairness

1. 緒論 (Introduction)

情感人工智能(Emotion AI)，作為當前科技發展的一個重要趨勢，透過導入語音情緒辨識(Speech Emotion Recognition, SER)模組，賦予了機器與人類之間進行智能互動的能力(Kaur & Sharma, 2021; Narayanan & Georgiou, 2013)。然而，儘管這種類型的科技進步帶來了深遠的影響，但也因為情緒感知的自然主觀性以及情緒產生的獨特因素，使得基於數據驅動的機器學習方法所建立的現代語音情感辨識系統，面臨著嚴重的不公平(偏見)問題。情緒作為一種人類基本而可衡量的內在屬性，具有巨大的潛力，並可廣泛地滲透至日常生活中的應用中，從自駕車的安全系統，到醫療保健領域的心理病理診斷，甚至是線上教育中的學習者情緒追蹤，情感 AI 都有著廣泛且深遠的應用前景。然而，當語音情緒辨識系統中挾帶著偏見，不僅會對該系統的可信度造成影響，還可能對使用者帶來不良的社會效應。例如，如果語音情緒辨識系統偏向於辨識特定性別或種族的情緒，那麼該系統可能無法公平地對所有使用者的情緒進行識別。這種偏見可能導致使用者的情緒被忽視或誤讀，從而產生不公平的結果，進而引發使用者的不滿，甚至社會問題。因此，如何配置公平的語音情感辨識系統，已經成為現代社會中的一個重要議題，這不僅包含算法的公平性，也需要兼顧使用者的感知，以確保科技能更有效地服務社會。

公平的語音情感辨識系統是一項涵蓋多個技術面向的複雜挑戰，這不僅涉及到演算法設計與實踐的公平性，還包括對使用情境的深度理解。Chien 等人(Chien & Lee, 2023)指出，我們應該思考一個語音情感辨識系統的開發過程需要含兩個重要角色：技術推動者與服務提供者。技術推動者作為算法的開發者，專注於解決從理論到實踐的各種技術問題，如數據收集與處理、模型設計與訓練以及性能評估與優化等，是推動語音情緒辨識技術實現的關鍵力量。服務提供者則需要從使用者的角度來思考各種可能的使用情境，並將這些情境的具體需求和挑戰反饋給技術推動者，他們不僅需要深入理解和尊重每一位使用者的需求和期望，還需要考慮如何在現實生活中有效地應用及推廣語音情緒辨識技術。對於技術推動者來說，他們的主要任務是開發能夠在不同情境中實現公平決策的算法(Verma, 2019; Angwin *et al.,* 2016)。雖然已有相當多的研究為機器學習應用領域提出了各種公平模型的演算法，但針對公平語音情感辨識的研究卻相對較少。在現有的語音情感辨識研究領域中，大部分聚焦於如何減少主體性偏見。例如，Gorrostieta 等人(Gorrostieta *et al.*,, 2019)提出了一種利用對抗性不變策略去除性別和年齡偏見的語音情緒辨識模型。然而，由於情緒標記（作為基準的真實情緒）的主觀性(Busso *et al.,* 2008;

Mariooryad *et al.,* 2014)，感知公平性（評估者的偏見）成為一個相當重要但受到較少關注的因素。有研究指出(Brody, 1985; Vigil, 2009)性別對情緒的敏感度存在差異，Swerts 等人(Swerts & Krahmer, 2008)也指出，女性對情緒的體驗通常比男性更為強烈。在本研究中，我們將從性別差異，探討其對情緒語音的評價是否存在公平性議題。

再者，大多數為達成公平而提出的策略，其核心概念都圍繞在模型應該產生不易被區分（即去偏）的表示或結果上。這些策略通常透過將模型的訓練數據進行前處理，或者直接在模型訓練的過程中進行調整，以達到減少或消除模型對特定群體的偏見。然而，這些初步的策略雖然具有前瞻性，但並未充分考慮到一個關鍵的問題：使用者的感知。我們認為要使語音情感辨識真正公平，也必須使使用者感受到公平性。換句話說，我們需要從服務提供者的角度來看待這個問題，我們應該提供透明的結果，讓使用者能清楚理解語音情緒辨識是如何根據語音分析對人進行評估，並能在結果間自由切換。這樣不僅能讓使用者對語音情感辨識的工作原理和效果有直觀的了解，還能讓他們在使用語音情緒辨識時有更多的選擇和控制權。透明度在這裡具有兩個重要的含義。首先，透明度可以提供可視化的模型決策過程，使使用者能夠清楚地了解模型是如何根據輸入的語音數據來產生情感識別結果。其次，透明度還可以讓使用者更好地理解模型在特定情境下可能存在的偏見，並根據需要選擇最適合自己的模型設定。且有研究指出，人們將透明度視為公平的象徵，並認為透明度將引導出「更公平」的結果。例如，Grill 等人(Grill & Andalibi, 2022)進行的一項研究就發現，當模型的決策過程被清晰地展現出來，使用者通常會對其結果的公平性有更高的評價。另外，Schoeffer 等人(Schoeffer *et al.,* 2022)的研究也觀察到，提供的資訊量會顯著影響使用者對公平性的感知。因此我們認為，一個公平的語音情緒辨識系統不僅應包含去偏的功能，還需要符合最終使用者對於透明度的期待。在這項研究中，我們也期望使用者能夠在無感知偏見的語音情緒辨識與男性或女性的觀點之間自由切換。

為了實現公平的語音情緒識別系統，我們設定了兩大目標：去除偏見的表示以及性別評估者視角的透明度。我們提出了雙階段的學習模型，在第一階段中，我們作為技術推動者，提供公平的表徵以確保語音情緒辨識模型在處理評估者性別時不受偏見影響。這一步是透過明確地減少不同評估者性別群體間的分佈距離來實現的。在第二階段，作為服務提供者，我們進一步為使用者提供可以在多種透明的輸出結果之間自由切換的能力。這是透過在學習輸出情感標記的過程中，將去偏見的表示引導向特定性別的視角空間來達成的。在評估我們的公平語音情緒辨識系統時，我們採用了兩個重要的公平性指標：統計奇偶校驗(statistical parity)和一致性分數(consistency score)。統計奇偶校驗檢查我們的系統是否對所有群體都給予相同的機會；一致性分數則檢查系統對相似的輸入是否能給出相似的輸出。透過這兩項指標，我們可以審視在追求公平性和識別效能之間的權衡，以確保我們的系統在保證公平性原則的同時，也不損及其情緒辨識的效能。

2. 研究方法 (Research Methodology)

2.1 資料庫 (Database)

IEMOCAP 數據集(Busso *et al.,* 2008)是語音情緒辨識領域的重要數據集之一，並深受研究者們的喜愛。這個數據集包含五組口語交流對話，每組對話由一位男性與一位女性演員進行。在進行情緒評分的過程中，我們有六位評估者參與，其中包括兩位男性和四位女性。這些評估者對每一段話語的主要情緒進行評分，並至少有三位評估者對同一話語進行評分。經過多數原則得出共識標記結果作為真值，也就是該話語的主要情緒。我們的研究重點放在了帶有五種主要情緒標籤的樣本上，這五種情緒包括：中性、快樂、憤怒、悲傷以及沮喪。情緒的選擇與多數傳統的語音情緒辨識研究相吻合，因為這五種情緒是人類情感表達中最常見，也是最基本的情緒類型。我們希望透過這種方式，能夠更深入地探索和理解人類的情感表達，並努力改善語音情感識別技術的準確性和公平性。

2.1.1 評估者的感知 (Rater's Perception)

在這一節中，我們將從性別角度探討情緒標記的公平性問題。我們的研究涵蓋了五種主要的情緒類別，總共收集了 7362 個話語樣本。首先，我們使用所有數據來計算：男性/女性評估者對每個話語的「共識」與投票的真值(Ground Truth Label)之間的匹配百分比，圖 1 顯示了在這個情況下的比例分布。接著，我們將焦點轉向在男性和女性評估者間存在情緒評估差異的樣本，也就是所謂的「非共識數據」，圖 2 展示了在這樣的條件下，真值與不同性別評估者評估結果之間的匹配百分比。這些數據顯示，男性與女性在情感感知上確實存在顯著的差異。與女性評估者相比，男性評估者對挫折這一情緒的真值的認同程度較低，而對其他所有情緒的認同程度則較高。

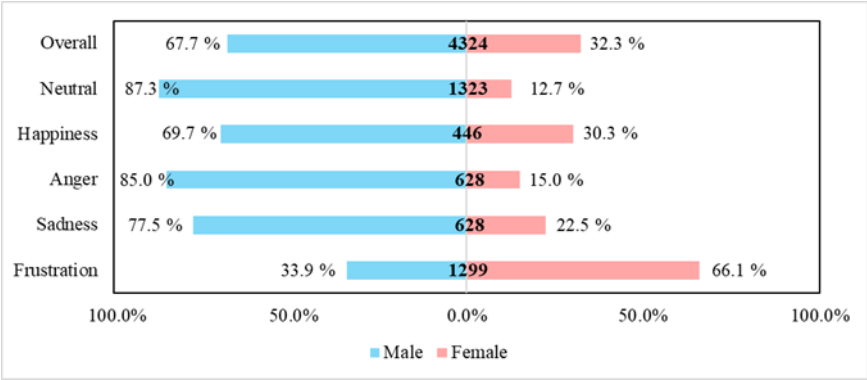

圖 1. 共識數據 – 男性/女性共識的資料集與真值資料匹配百分比的初步分析。

[Figure 1. Consensus Data – Preliminaly analsysis on the matching percetage between male/female consesnsus and ground truth data.]

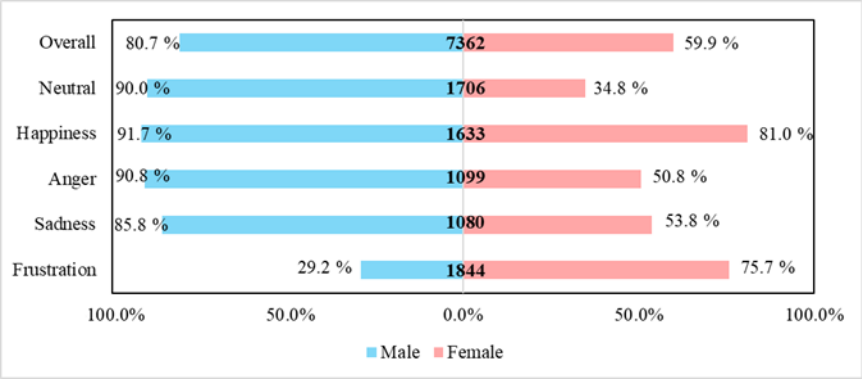

圖 2. 非共識數據 – 男性/女性共識的資料集與真值資料匹配百分比的初步分析。

[Figure 2. Non-Consensus Data – Preliminaly analsysis on the matching percetage between male/female consesnsus and ground truth data.]

基於上述觀察，我們將原始數據集分成兩個子集進行進一步的實驗。第一個子集 S1 為「性別感知無偏見集」，這個子集共有 3038 個樣本，男性和女性評估者對基本事實標籤的情感感知相同。另一個子集 S2 被命名為「性別感知有偏見集」，這個子集共有 4342 個樣本，基本事實標籤會偏向男性或女性的情感感知。透過這樣的數據劃分，我們期望更深入地了解性別角度的感知差異如何影響情感識別的公平性。

2.2 模型架構 (Computational Framework)

在這項研究中，我們設計雙階段模型架構，且確保語音情緒辨識的公平性。首先我們提取語音特徵作為輸入，在第一階段使用模型推導出一種感知公平的表徵(如圖 3 所示)。這種表徵考慮了不同性別評估者對情緒的不同感知，為我們建立公平的語音情緒辨識系統提供了基礎。在這個公平表徵的基礎上，我們進行了第二階段的工作，即按需情緒辨識(如圖 4 所示)。這一階段的目標是能夠讓我們的語音情緒辨識系統在特定性別觀點之間自由切換，從而更好地適應使用者的需求。這種雙階段學習策略不僅增強了語音情緒辨識系統的公平性，也提高了其靈活性和個性化的能力，使其能夠更好地應用於實際場景中。

2.2.1 聲學特徵 (Acoustic Features)

我們使用 fairseq 工具 (Ott *et al.,* 2019) 提取 512 維的潛在 vq-wav2vec 向量作為聲學特徵(Baevski *et al.,* 2019)，以便將原始音訊的訊息嵌入。預先訓練的音頻編碼器可以直接將原始波形投影到潛在空間，所有的特徵都以語者為單位進行 z-標準化。

2.2.2 第一階段：學習公平的表示 (Stage1: Fair Representation Learning)

如圖 3 所示，在我們的研究中，我們首先提出了一個具公平約束的對抗性學習架構，這是實現公平語音情緒辨識的第一階段。我們希望儘可能地消除嵌入的性別訊息，得到的情緒表徵是不受性別視角影響的，以實現真正無偏見的情緒表徵。

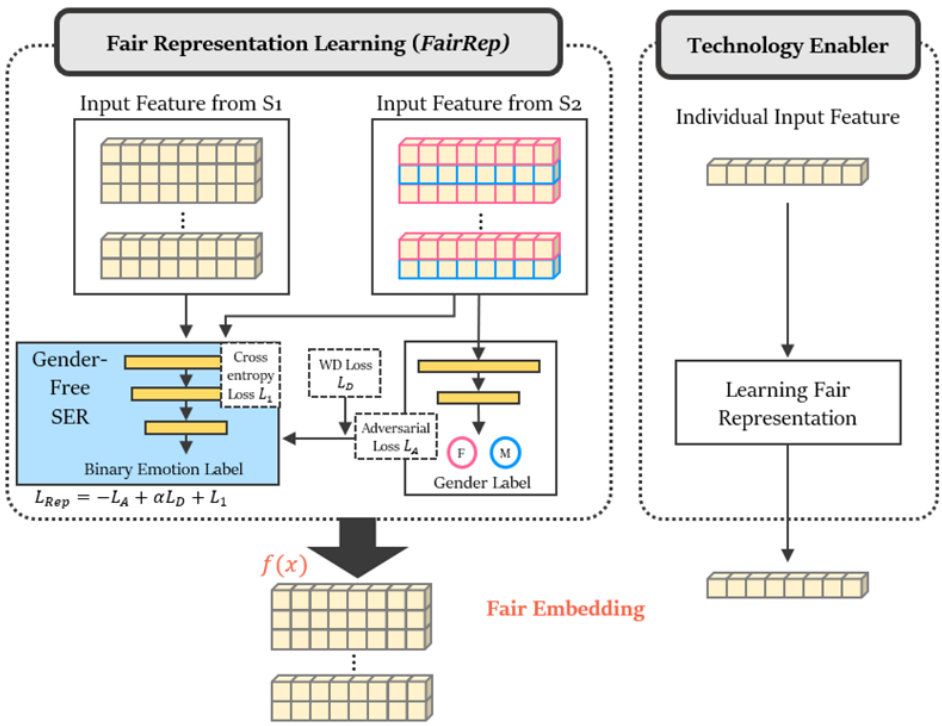

圖 3. 公平的語音情感識別(SER)架構。第一階段用於訓練公平表示 (Fair Representation)，接著利用公平表示進行性別感知預測的訓練。
[Figure 3. Overview of the fair speech emotion recognition (SER) architecture using our proposed first-stage framework. The first stage is used to train for a fair representation, then utilize the fair representation to train for the gender-wise perceptual predictions.]

更具體地說，我們對學習過程施加了一種特殊的約束：要求在特徵空間中，不同性別的條件分佈應當是相同的。這一約束的設定使得我們的學習過程在某種程度上類似於(Ganin *et al.,* 2016)中提出的域不變學習策略，這是一種能夠有效抑制學習過程中不公平現象的方法。我們訓練了專門用於性別分類的模型，並在 S2（偏見集）上使用交叉熵損失作為優化目標。透過這種方法，我們可以有效地識別並減輕模型中的性別感知偏差。然而，我們發現僅僅縮小偏見並不足夠，因此，我們進一步從我們的情緒檢測網路中減去了這部分的損失L_A，以達到更深層次的公平性。

我們進一步對特徵空間中的分布施加公平約束，以確保更好地滿足公平性。Dwork 等人(Dwork *et al.,* 2012)顯示，如果最小化兩個群體之間的 Wasserstein 距離（WD），則可以共同實現兩個重要的公平度量標準：統計平等(Statical Parity)和一致性分數(Consistency Score)。因此，透過明確地將 WD 設計為優化目標，我們可以確保在使用學習特徵進行分類時結果是公平的。為了計算損失L_D，我們測量男性特徵和女性特徵之間的距離D_w (Feng *et al.,* 2019)。最後，這個網路的參數是透過最小化以下損失函數來訓練的：

$$L_{Rep} = L_1 - L_A + \alpha L_D \quad (1)$$

損失函數L_{Rep}由三部分組成：L_1是用於評估 S1 和 S2 上情感分類效能的標準交叉熵損失；L_A是性別訊息損失項，用於評估嵌入中包含多少性別訊息；L_D衡量了評估者性別屬性組之間在潛在空間中的分布接近程度。超參數α控制系統的平衡。在本文接下來的內容中，我們將統稱第一階段的模型為 FairRep。

2.2.3 第二階段：學習性別感知方面的語音情緒辨識 (Stage2: Gender-wise Perceptual SER Learning)

在我們的研究中，三元組損失網路扮演了關鍵的角色，整體架構圖如圖 4 所示。該網路接收一批三元組單元〈x_i^a，x_i^p，x_i^n〉作為輸入，這些三元組單元中的x_i^a，x_i^p代表了同一性別的評估者，而 x_i^a, x_i^n代表了不同的性別。我們用 f(x) 來表示將輸入 x 嵌入到一個 d 維歐幾里德空間的特徵，目標是最小化這些三元組單元的條件〈x_i^a，x_i^p，x_i^n〉三元組損失 L_T，這一概念是根據參考文獻(Veit er al., 2017)中的定義提出的。為了實現此目標，我們在深度神經網路的輸入端設計了一個全連接網路層，這一層的功能是將原始的輸入數據轉換為適合進行三元組損失嵌入的形式，如圖 4 所示。

接著，我們將全連接網路層與深度神經網路結合在一起，形成了一個完整的三元組損失嵌入式網路。我們對這個網路進行優化，使用的是一種總損失函數。這種總損失函數綜合考慮了多個因素，以確保我們的網路在進行性別特定的語音情緒辨識時，能夠達到最佳的性能。

$$L_{Per} = L_2 + \lambda L_T$$

其中 L_2 是目標性別指定情感標記的標準交叉熵，λ 指的是兩個損失之間的權重。在本文接下來的內容中，我們將統稱第一階段的模型為 FairPer。

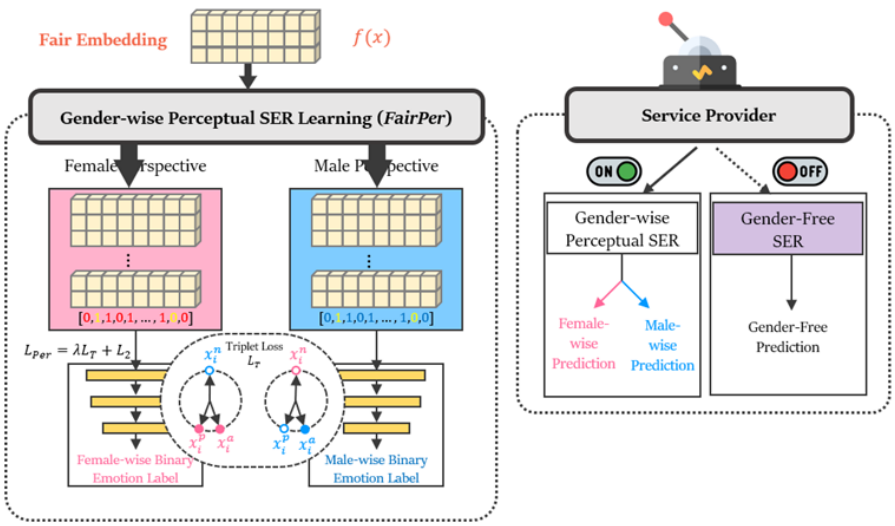

圖 4.這是第二階段。對於相應的應用場景，一旦技術開發者從 FairRep 提供了公平的嵌入，服務提供者就能夠提供使用者按需切換選項的功能。
[Figure 4. Overview of the fair speech emotion recognition (SER) architecture using our proposed second-stage framework. For the corresponding application scenarios, once the technology enabler provides a fair embedding from the FairRep, the service provider provides users the ability to toggle options on-demand.]

3. 實驗設計與結果 (Experimental Setup and Results)

3.1 實驗設置 (Experimental Setup)

我們設計了 2 個實驗來評估我們提出的方法的有效性。 為了進一步理解我們提出的模型在公平性和識別性能之間的權衡，我們提出了識別指標以及為每個實驗選擇的兩個不同的公平性指標。

(1) Exp1：在第一個實驗中，我們的目標情感標記源自真實情緒（ground truth），我們使用加權 F1-score 作為評估指標，衡量我們的模型在 S1 數據集上的情感辨識能力。這種方法讓我們能夠確認在對共識數據進行公平性處理後，識別性能並未受到任何負面影響，也就是說，我們的模型在確保公平性的同時，並未犧牲識別的準確性。至於公平性指標，我們選擇使用統計均等分數（理想值 ＝0）(Yang & Stoyanovich, 2017)。這個指標能夠評估我們的模型在預測不同評估者性別情緒時的公平性。我們在 S2 數據集上進行評估，結果證明我們的模型對性別屬性的任何一個類別都不會有偏好。這種公平性的體現不僅增加了我們模型的可靠性，也使得模型在實際應用中更具公平性。

(2) Exp2：在第二個實驗中，我們的策略是在建立了公平基礎的前提下，進一步提供按需選擇的特定性別感知結果。為此，我們利用了第一階段所學習出的公平性表徵作為所有實驗的基礎。目標情緒標記來自男性評估者或女性評估者的共識。我們透過加權 F1 分數來評估特定性別的情感識別效能，這涵蓋了使用所有男性和女性觀點作為目標標記的情況。在這個階段，我們將公平性定義為同一評估者性別的樣本應該在特徵空間上彼此接近，也就是說，在這種情況下，我們的預測中不應該包含任何不必要的偏向。為了衡量這種公平性，我們選擇使用一致性分數作為公平性指標（理想情況為 1）(Zemel *et al.,* 2013)，該指標衡量了內嵌在 k 最近鄰集（k = 20）內的樣本與其性別屬性之間的一致性。這種評估方式提供了我們對模型性別公平性的深入理解，並能幫助我們繼續優化模型的公平性。

在本研究中，我們選擇了簡單的深度神經網路(DNN) 作為基準線方法，直接使用所有可用特徵來訓練一個包含三個全連接層的分類器。作為比較的對象，我們考慮了兩種現有的公平表示方法：潛在特徵表示(LFR)(Zemel *et al.,* 2013)，該方法將實例直接分配給特定原型作為潛在表示；神經表示學習(NRL)(Feng *et al.,* 2019)則是一種具有公平性約束的最先進框架。我們提出的去除方法 FairRepF 不考慮評分者性別的領域知識，與 NRL 框架相似；而 FairRepG 則不考慮公平性約束。在所有實驗中，我們採用了獨立於會話的交叉驗證方案。學習率和衰減因子設置為 0.001，dropout 率設置為 0.2。我們透過網格搜索在[0.01, 0.001]範圍內對超參數 λ 和 α 進行調整。批量大小固定為 32，最大迭代次數為 500，優化器選擇了 Adam 算法。

3.2 實驗結果 (Experimental Results)

我們對雙階段模型架構的每個階段進行了詳細檢視（即第一階段模型: Exp1 和第二階段模型: Exp2）。首先，表 1 表 2 彙總了第一階段的結果。由於我們在模型中施加了公平性約束，所以相對於未針對公平性進行優化的方法（例如 DNN），我們的方法在整體識別性能上略有下降是可以預見的。然而，我們的 FairRep 方法具有兩大優勢：(1) 相較於未考慮公平性的方法，它更能滿足統計奇偶性分數指標的要求；(2) 在共識數據集 (S1) 上的效能下降最為有限。

除了統計奇偶性分數外(表 2)，還有其他一些重要的觀察點。從表 1 我們可以觀察到，如果沒有公平的機制，語音情緒辨識 (SER) 模型可能會產生明顯的不公平預測，比如偏向某一性別的情緒感知。在我們的消除(ablated)方法 FairRepG 中，性別差異在一定程度上被消除，但由於它並未在學習中考慮公平性約束，因此可能還存在與「性別」相關的其他屬性導致的間接偏差(Zliobaite, 2015)，在檢查統計均等分數時也表現得很明顯。此外，FairRepF 並未直接消除領域知識中的性別差異而保留了性別訊息。總的來說，從兩種消融結果來看，域不變性和距離分佈匹配都是必要的，並且它們都會影響學習模型的性能。我們觀察到，與 DNN 相比，我們提出的方法 FairRep 在 F1-Score 中的中性、幸福、憤怒和沮喪的分數僅分別下降了 1.97%、0.97%、1.88%和 0.79%。然而，它在統計奇偶性分數上展現出了具有競爭力的公平性能力。這些結果表明，我們的 FairRep 方法

能夠實現更公平的情緒識別結果，並能夠有效處理共識數據集 (S1) 和性別偏見數據集 (S2)。

表 1. 第一階段公平表示學習的辨識結果 (F1 分數 (%))。帶底線的數據集(S1)表示我們感興趣的性能，粗體數字表示相應的最佳性能，“All”數據集中的性能表示“無性別 SER”的預測結果，如圖 1 所示。
[Table 1. A summary of the experimental recognition results on Fair Representation Learning by F1 score (%). The underlined datasets indicate the performance we are interested in. The bold numbers represent the corresponding best performance. The performance in 'All' dataset provides the "Gender-Free SER" prediction, which is mentioned in Fig.1.]

Emotion	Dataset	DNN	LFR	NRL	FairRep_G	FairRep_F	FairRep
Neutral	All	77.73	67.89	68.55	66.89	68.82	68.80
	S1	**79.07**	73.64	75.12	77.62	76.46	77.10
	S2	66.12	65.17	66.84	63.58	65.46	63.25
Happiness	All	70.00	66.01	69.46	64.48	66.46	65.14
	S1	**73.75**	70.77	72.10	70.88	71.66	72.78
	S2	62.73	61.22	65.65	64.86	63.84	62.83
Anger	All	76.44	70.92	70.76	76.28	78.26	75.68
	S1	**80.54**	70.56	71.40	78.20	79.00	78.66
	S2	73.28	68.11	69.81	73.90	76.69	70.15
Sadness	All	82.28	76.74	78.74	80.15	78.26	76.84
	S1	**88.54**	80.39	82.01	76.33	79.00	80.00
	S2	78.68	76.28	77.98	79.58	77.44	76.28
Frustration	All	63.54	70.54	64.32	67.54	69.02	70.22
	S1	67.61	60.43	65.02	66.72	66.61	**66.82**
	S2	61.30	62.30	64.88	62.68	63.68	67.43

表 2. 第一階段公平表示學習的公平表現結果 (統計奇偶分數 Statical Parity Score)，在 S2 資料集上的結果。
[Table 2. A summary of the fairness performance by statistical parity score on Fair Representation Learning on S2 subset.]

Emotion	DNN	LFR	NRL	FairRep_G	FairRep_F	FairRep
Neutral	0.650	0.350	0.332	0.342	0.350	**0.350**
Happiness	0.428	0.142	0.134	0.128	0.136	**0.126**
Anger	0.389	0.194	0.197	0.197	0.208	**0.189**
Sadness	0.169	0.069	0.068	0.106	0.104	**0.088**
Frustration	0.626	0.493	0.452	0.472	0.467	**0.448**

進入到第二階段的實驗(Exp2)，我們更進一步分析了我們的方法在處理具有明顯性別差異的情緒識別問題上的效能，表 3 和表 4 分別總結了我們在第二階段的實驗結果。在這個階段，我們針對男性和女性評估者的共識情緒標籤進行了專門的分析，並試圖理解我們的方法如何在確保識別性能的同時，提供具有性別感知的情緒識別選項。我們也觀察到，儘管在實施公平性約束後，我們的模型在部分情緒分數上出現了輕微的下降，但這些下降在大多數情況下都是可接受的，而且遠遠低於未施加公平性約束的模型。更重要的是，我們的模型在提供公平且無偏見的情緒識別結果方面，表現得相當出色。總的來說，這些結果證實我們的雙階段模型架構以及我們提出的 FairRep 方法。我們的模型不僅在保證情緒識別性能的同時，確保了公平性，而且還具有足夠的靈活性，可以對特定性別的情緒識別進行優化。這種平衡性和靈活性使我們的模型成為一種強大的工具，可以用來探索和理解情緒識別中的性別差異，以及如何在確保公平性的同時，提高情緒識別的性能。總的來說，我們提出的 FairPer 方法在為不同性別評估者提供精準性別感知預測方面顯示出優越的性能。這點在我們在所有性別情緒識別任務中取得的優異識別效能中顯得格外明顯。

表 3. 第二階段性別感知 SER 學習的辨識結果 (F1 分數 (%))。M 和 F 代表從男性/女性的角度評估所有帶有男性/女性標記的數據。
[Table 3. A summary of the recognition results on Gender-wise Perceptual SER Learning by F1 (%).. 'M' and 'F' represents evaluating all data with male/female labels from male/female perspectives.]

Emotion	Dataset	DNN	LFR	NRL	FairRep_G	FairRep_F	FairRep	FairPer
Neutral	M	67.68	65.23	67.25	66.32	65.10	68.26	**80.83**
	F	62.18	63.44	65.51	64.04	64.66	75.82	**86.75**
Happiness	M	67.33	64.00	66.12	67.57	64.57	66.73	**73.64**
	F	61.10	60.71	63.02	62.22	63.86	68.15	**75.89**
Anger	M	75.52	69.22	69.98	77.33	76.98	78.36	**80.46**
	F	66.52	68.15	68.33	70.16	78.03	75.54	**88.30**
Sadness	M	82.04	77.06	78.43	73.40	78.80	77.97	**79.16**
	F	75.90	72.41	75.62	75.62	76.69	77.25	**80.96**
Frustration	M	62.17	61.17	64.05	63.00	62.06	75.88	**85.16**
	F	64.96	65.02	63.72	60.77	64.16	77.98	**77.42**

表 4. 第二階段性別感知 SER 學習的公平表現結果 (一致性分數 Consistency Score)。
[Table 4. A summary of the fairness performance by consistency score on Gender-wise Perceptual SER Learning..]

Emotion	DNN	LFR	NRL	$FairRep_G$	$FairRep_F$	FairRep	FairPep
Neutral	0.537	0.685	0.786	0.791	0.788	0.802	**0.853**
Happiness	0.552	0.698	0.783	0.772	0.761	0.777	**0.827**
Anger	0.583	0.722	0.790	0.801	0.762	0.809	**0.822**
Sadness	0.552	0.735	0.762	0.766	0.733	0.739	**0.797**
Frustration	0.568	0.706	0.741	0.743	0.752	0.760	**0.840**

有趣的是，我們發現男性和女性的表現模式反映了圖 1 圖 2 中觀察到的性別感知差異。透過我們的 FairPer 模型，能夠更有效地掌握男性或女性在情緒評估上的分歧，特別是在他們對某些情緒類別的評估上（如圖 2 所示）存在較大差異的情況下，性別對識別結果的影響將更為顯著。以 Frustration 情緒為例，男性評估者的評估與投票的真值不太一致，然而我們的模型卻更能夠辨識男性的觀點， F1 分數高達 85.16%。

我們的模型中高一致性分數的表現，進一步確證了我們的學習表示在第二階段可以將同一性別的樣本更緊密地聚集在一起，從而提供更一致的性別特定觀點。這一點對於提高模型的識別性能和公平性都是十分重要的。總的來說，這些發現證實了我們的模型能夠有效地適應性別的觀點差異，並能夠在需要時靈活地在男性、女性或無性別識別之間進行切換。因此，我們的模型不僅能夠提供更人性化和公平的情緒識別體驗，而且也能夠更好地理解和處理性別在情緒識別中的影響。

3.3 實驗分析 (Experimental Analyses)

在我們的研究中，我們對第一階段是否進行公平操作的模型進行了深入分析。我們比較了 DNN 模型和我們自己提出的 FairRep 模型，並使用 T-SNE 方法將特徵映射到空間中後觀察其分布。

觀察圖 5，我們可以發現在 DNN 模型訓練後，男性觀點和女性觀點的樣本特徵有一定的聚集現象。這種現象表明，該模型可能受到性別觀點的一定影響，並在特徵表示中留下相應的痕跡，可能會對語音情緒辨識的公平性產生潛在影響。然而，當我們觀察圖 6 時，我們的 FairRep 模型有不同的現象，男性觀點和女性觀點的樣本特徵在空間中被平均分散，並無明顯的聚集現象，也就是說我們的模型能夠避免受到性別觀點的影響，確保學習特徵的公平性，這對於實現公平的語音情緒辨識為目標來說至關重要。此分析不僅證明我們模型的有效性，也可驗證語音情緒辨識系統的公平性。

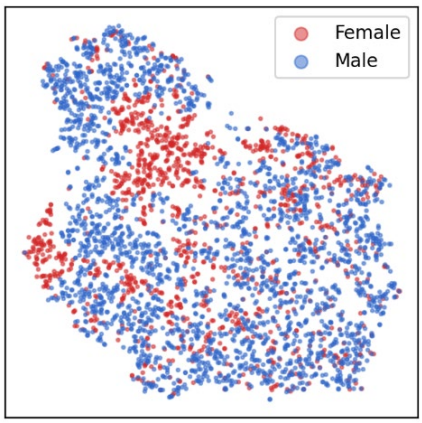

圖 5. 經過 DNN 後的特徵樣本分布 [Figure 5. The t-SNE scatter plot of DNN from Anger detector.]

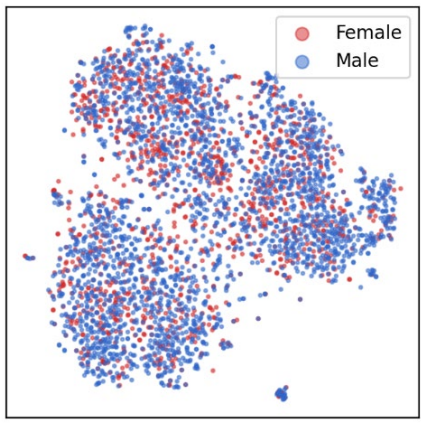

圖 6. 經過 FairRep 後的特徵樣本分布 [Figure 6. The t-SNE scatter plot of FairRep from Anger detector.]

4. 結論 (Conclusion)

在本研究中，我們提出一種創新的雙階段模型架構，該架構的核心在於利用公平約束對抗架構在第一階段產生公平的數據表示，在第二階段允許用戶根據需求靈活地切換到特定的性別感知模式。我們的研究結果發現：(1) 我們提出的模型在面對無偏數據的驗證時，成功地降低了辨識能力的損失；(2) 透過公平性指標的分析，我們確認了我們的模型能夠較好地解決情緒識別中的不公平問題。

5. 未來展望 (Future Work)

為了進一步優化我們的模型並解決情緒識別中的公平性問題，我們首先會在未來的研究中從語者和評估者偏見互相影響的角度，來更全面地理解和解決這一問題，將有助於我們更深入地理解情緒識別中的性別差異，並提供更公平、更人性化的情緒識別服務。而鑒於此領域的研究仍處於初期階段，偏見也不會只存在於性別中，因此我們還對未來的研究提出以下建議，以期進一步拓展和深化這一領域的研究。

- 擴展偏見因素的探討：目前的研究主要集中在性別偏見上，而未來的研究可以將視角擴大到其他可能導致偏見的因素或族群，例如年齡、知識背景、精神狀態等，這將有助於更全面地理解和應對影響語音情緒辨識公平性的各種因素。
- 拓展到其他語音相關任務的應用：目前的方法是否可以擴展到其他與語音相關的任務，如語音增強或自動語音識別，這將是一個值得探討的問題，如果能夠在這些領域中實現公平性，將會大大增強語音技術的應用範圍和深度。

- 探討試驗設計的影響因素：參與者回答問題的順序可能會對研究結果產生影響，因此，研究中應該考慮問題順序對結果的可能影響，Martinez-Lucas 等人(Martinez-Lucas, 2023) 也針對情緒啟動(Affective Priming)對情緒標記的影響做深度的情緒維度分析。這些因素的探討將有助於進一步優化研究設計，提高研究的準確性和可靠性。

本研究在公平的語音情緒辨識領域開啟新的研究方向，不僅對語音情緒辨識技術的發展具有重要的引導意義，也將對人工智慧領域的公平性和倫理性問題提供新的研究視角和思考。我們期待未來的研究也可以在這些建議的基礎上，為公平的語音情緒辨識建立好的開端，為這一個領域帶來更多的創新和突破。

參考文獻(References)

Baevski, A., Schneider, S., & Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. arXiv preprint arXiv:1910.05453.

Brody, L. R. (1985). Gender differences in emotional development: A review of theories and research. *Journal of personality*, *53*(2), 102-149. https://doi.org/10.1111/j.1467-6494.1985.tb00361.x

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335-359.

Chien, W.-S., & Lee, C.-C. (2023). Achieving Fair Speech Emotion Recognition via Perceptual Fairness. I*n Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1-5. https://doi.org/10.1109/ICASSP49357.2023.10094984

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-226. https://doi.org/10.1145/2090236.2090255

Feng, R., Yang, Y., Lyu, Y., Tan, C., Sun, Y., & Wang, C. (2019). Learning fair representations via an adversarial framework. arXiv preprint arXiv:1904.13341.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, *17*(1), 2096-2030.

Gorrostieta, C., Lotfian, R., Taylor, K., Brutti, R., & Kane, J. (2019). Gender De-Biasing in Speech Emotion Recognition. In *Proceedings of Interspeech*, 2823-2827. https://doi.org/10.21437/Interspeech.2019-1708

Grill, G., & Andalibi, N. (2022). Attitudes and Folk Theories of Data Subjects on Transparency and Accuracy in Emotion Recognition. In *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1-35. https://doi.org/10.1145/3512925

Kaur, S., & Sharma, R. (2021). Emotion AI: integrating emotional intelligence with artificial intelligence in the digital workplace. Innovations in Information and Communication Technologies (IICT-2020) *Proceedings of International Conference on ICRIHE-2020*, Delhi, India: IICT-2020, 337-343.

Mariooryad, S., Lotfian, R., & Busso, C. (2014). Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora. In *Proceedings of Fifteenth Annual Conference of the International Speech Communication Association*, 238-242, https://doi.org/10.21437/Interspeech.2014-60

Narayanan, S., & Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, *101*(5), 1203-1233. https://doi.org/10.1109/JPROC.2012.2236291

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. arXiv preprint arXiv:1904.01038.

Schoeffer, J., Kuehl, N., & Machowski, Y. (2022). "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *Proceedings of 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1616-1628. https://doi.org/10.1145/3531146.3533218

Swerts, M., & Krahmer, E. (2008). Gender-related differences in the production and perception of emotion. In *Proceedings of Ninth Annual Conference of the International Speech Communication Association*. https://doi.org/10.21437/Interspeech.2008-99

Veit, A., Belongie, S., & Karaletsos, T. (2017). Conditional Similarity Networks. 830–838. https://openaccess.thecvf.com/content_cvpr_2017/html/Veit_Conditional_Similarity_Networks_CVPR_2017_paper.html

Vigil, J. M. (2009). A socio-relational framework of sex differences in the expression of emotion. *Behavioral and Brain Sciences*, 32(5), 375-390. https://doi.org/10.1017/S0140525X09991075.

Yang, K., & Stoyanovich, J. (2017). Measuring Fairness in Ranked Outputs.In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 1-6. https://doi.org/10.1145/3085504.3085526

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *International conference on machine learning*, *28*(3), 325-333.

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning (arXiv:1511.00148). arXiv. https://doi.org/10.48550/arXiv.1511.00148

Martinez-Lucas, L., Salman, A.N., Leem, S., Upadhyay, S.G., Lee, C., & Busso, C. (2023). Analyzing the effect of affective priming on emotional annotations. In *Proceedings of 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1-7. https://doi.org/10.1109/ACII59096.2023.10388184