Non-Uniform Speaker Disentanglement for Depression Detection from Raw Speech Signals

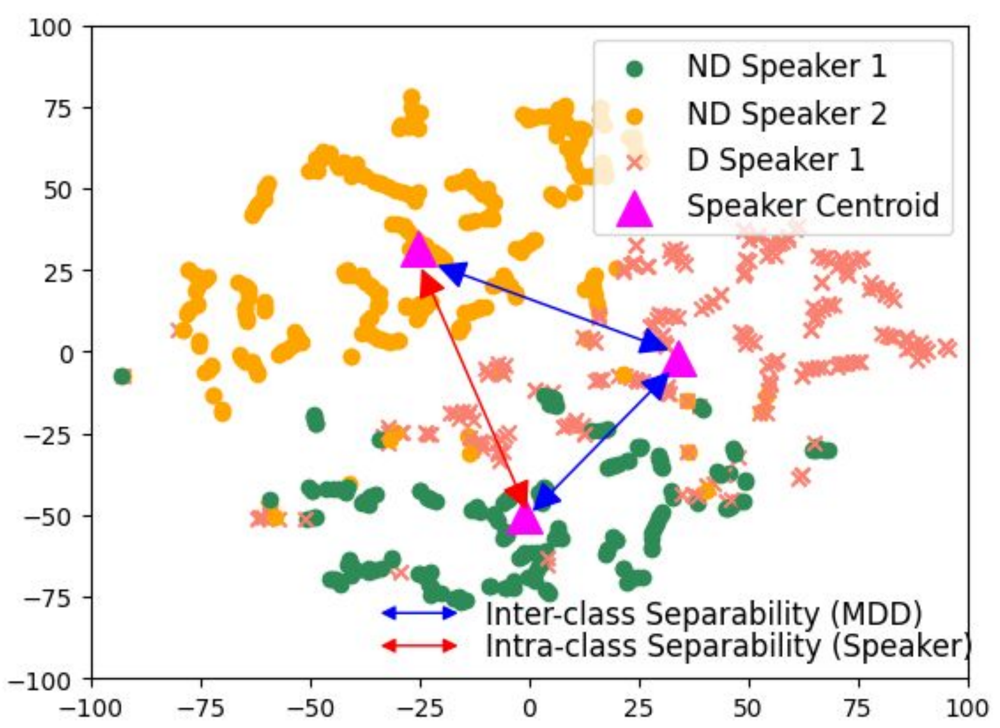
Jinhan Wang, Vijay Ravi, Abeer Alwan
Dept. Electrical and Computer Engineering, University of California, Los Angeles, USA

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

I. Introduction

- Speech signals -> effective biomarkers in Major Depressive Disorder (MDD) detection.
  - Speaker-identity-related features (x-vector [21] , speaker-embeddings [23]) have been used and result in good performance.
- Problem: Over-reliance on speaker-identity-related features raises privacy-preservation concern.
- **Solution: A novel speaker-disentanglement method for depression detection from raw speech signals**
  - **Outperforms audio-only SOTA for MDD detection.**
  - **Reduced speaker ID accuracy.**

II. Privacy Concern and Speaker Bias



A tSNE plot of embeddings for three female speakers taken from the DepAudioNet model. ND means 'non-depressed' class and D means 'depressed' class. Each point is a segment from target speaker's utterance.

- Higher intra-class separability than average inter-class separability.
- Model may tend to discriminate speakers instead of depression states.

III. Uniform Speaker Disentanglement

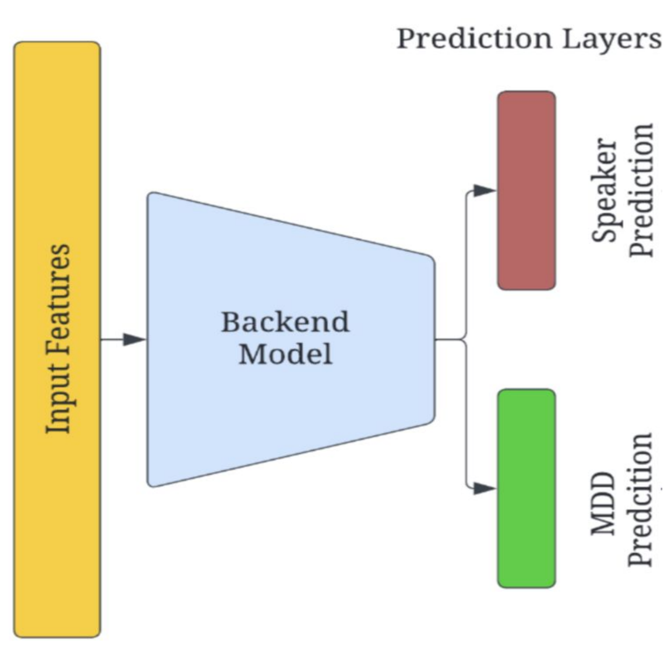
- Uniform Speaker Disentanglement (USD[30]) minimizes the MDD prediction loss and maximize the speaker identification (SID) loss.
- USD loss is defined as:

$$L_{USD} = L_{MDD} - \lambda(L_{SPK})$$

L_{MDD} : Depression Detection Loss
L_{SPK} : SID Loss
λ : Adversarial Loss Factor

- During training: θ_{ALL} : model parameters

$$\theta_{ALL} = \theta_{ALL} + \alpha\left(\frac{\partial L_{SPK}}{\partial \theta_{ALL}} - \frac{\partial L_{MDD}}{\partial \theta_{ALL}}\right)$$



- Better control over adversarial disentanglement applied to different layers might improve performance.

IV. Non-uniform USD (NUSD)

- Separate loss gradients of the auxiliary task (SID) into multiple components based on model layers.
- Applying different loss maximization strength on different components allows varying levels of disentanglement.
- As a preliminary study, models are split into:
  - Feature Extraction (FE) : initial layers
  - Feature Processing (FP): final layers
- During Training:

$$\frac{\partial L_{SPK(NUSD)}}{\partial \theta_{ALL}} = \left[\frac{\partial(\lambda_1 L_{SPK})}{\partial \theta_{FE}}, \frac{\partial(\lambda_2 L_{SPK})}{\partial \theta_{FP}}\right]$$

θ_{FE} : FE component parameters λ_1 : FE adversarial loss factor
θ_{FP} : FP component parameters λ_2 : FP adversarial loss factor

$$\beta = \lambda_1/\lambda_2 \quad (\beta = 1 \rightarrow \text{USD})$$

V. Experiments

- Dataset: DAIC-WoZ
  - 189 Speakers
- Input:
  - Raw Audio
  - 3.84s sample length
- Training and Evaluation:
  - Random sampling
  - 5 model averaging
- Models:
  - DepAudioNet
    - FE: 2x conv1d layers
    - FP: 2x LSTM layers
  - ECAPA-TDNN
    - FE: Input and 3x SE-Res2 Blocks
    - FP: Aggregation ~ prediction layers.

Metrics: MDD Classification F1-Score & SID Accuracy

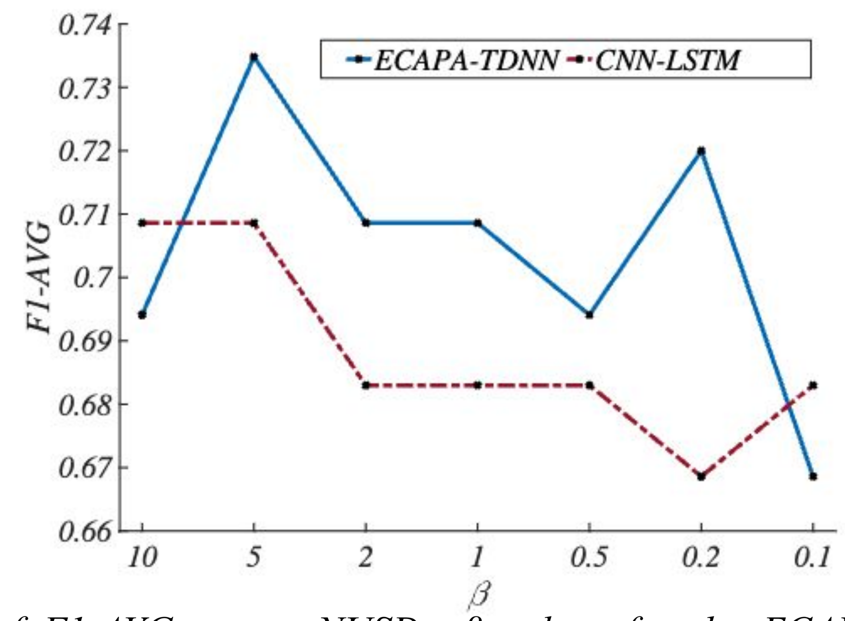
VI. Results

Depression detection performance for DepAudioNet and ECAPA-TDNN based on F1-AVG and Speaker ID accuracy using the DAIC-WoZ dataset. The symbols '↑' and '↓' indicate a higher or lower value is better, respectively. Best results are highlighted in bold.

Model Architecture	Disentanglement Method	Model Parameters	F1-AVG ↑	SID Accuracy ↓
DepAudioNet [37] ($D1$)	None	445k	0.6259	10.04%
DepAudioNet [30] ($D2$)	USD	459k	0.6830	8.91%
DepAudioNet ($D3$)	NUSD	459k	0.7086	8.05%
Δ ($D3$ vs $D2$) in %	-	-	3.75	-9.65
ECAPA-TDNN ($E1$)	None	595k	0.6329	42.33%
ECAPA-TDNN ($E2$)	USD	609k	0.7086	9.38%
ECAPA-TDNN ($E3$)	NUSD	609k	**0.7349**	**4.68%**
Δ ($E3$ vs $E2$) in %	-	-	3.70	-50.11

- NUSD achieves better MDD average F1 score and lower speaker accuracy than USD on two systems (D and E).

VII. Effect of β



A plot of F1-AVG versus NUSD β values for the ECAPA-TDNN and the DepAudioNet CNN-LSTM model.

- Higher weights on FE layers leads to better performance.
- Observation holds true for both ECAPA-TDNN and DepAudioNet models using RawAudio as input.

VIII. Conclusion

- NUSD shows promising results by utilizing a non-uniform mechanism of adversarial SID loss maximization.
- NUSD chieves an F1-Score of 0.7349 on the publicly available DAIC-WoZ dataset without any data augmentation, pre-training, or handcrafted features.
- Future Work Directions:
  - Examining the effect of number of speakers in the training set.
  - More fine-grained variants of NUSD.
  - Extension to other domains.

- Acknowledgement
  - Work funded by NIH award number R01MH122569

wang7875@g.ucla.edu