A Privacy-Preserving Unsupervised Speaker Disentanglement Method for Depression Detection from Speech

ML4CMH2024
AAAI Workshop
February 26th, 2024
Vancouver, Canada

Vijay Ravi, Jinhan Wang, Jonathan Flint, Abeer Alwan
University of California, Los Angeles, USA

UCLA

vijaysumaravi@g.ucla.edu

Acknowledgement: NIH award number R01MH122569

I. Introduction

- ➔ Automated depression detection using speech can be helpful to clinicians and patients.
- ➔ Challenges: Patient privacy ⚠
  - ◆ Patients can be identifiable if speaker ID information is included (ex. Membership inference attacks can occur)
- ➔ Previous work on privacy-preservation
  - ◆ Adversarial methods - > Unstable loss maximization
  - ◆ Need Speaker labels for training data -> Supervised
  - ◆ Speaker prediction branch needs additional parameters -> Inefficient
- ➔ Solution -> Cosine similarity minimization between speaker and depression embedding spaces
  - ◆ No Speaker labels used: Unsupervised!
  - ◆ No additional parameters: Efficient!

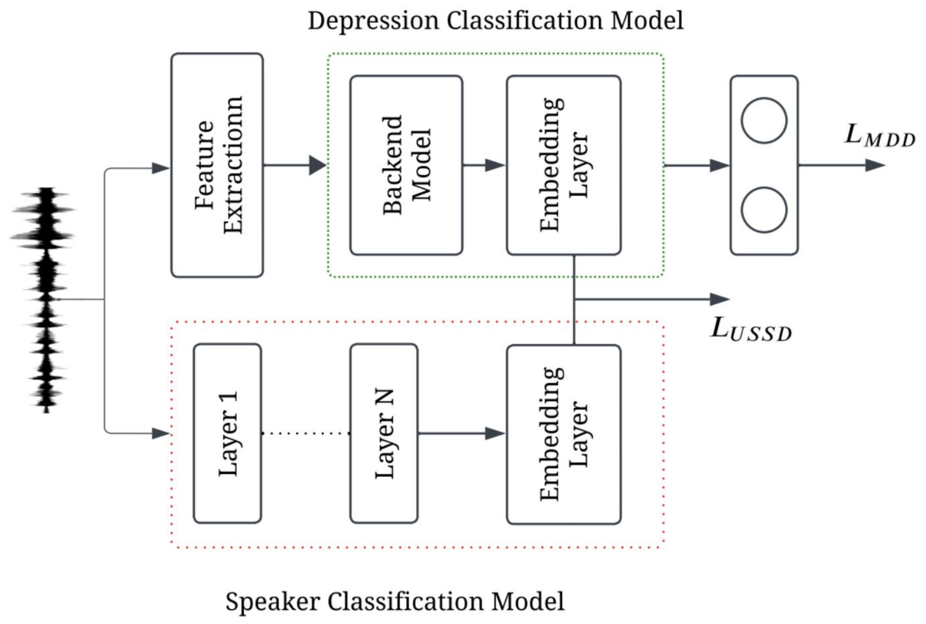
II. Proposed Method

Background: Adversarial Learning (ADV [16])

- ➔ Loss Min. - Max.
  - ◆ Depression prediction Loss L_{MDD} - minimized.
  - ◆ Speaker prediction Loss L_{SPK} - maximized.
  - ◆ Gradient reversal disentangles the speaker identity from depression characteristics.

$$L_{total-ADV} = L_{MDD} - \alpha \cdot L_{SPK-ADV}$$

Unsupervised Speaker Disentanglement (USSD)



Off-the-shelf Speaker Classification Model [35]
(No finetuning or retraining)

Key Idea -

- ➔ For utterance X, get speaker and depression embeddings

$$H_{MDD_X} = \theta_{MDD}(X)$$
$$H_{SPK_X} = \theta_{SPK}(X)$$

- ➔ Minimize Cosine similarity between the two embedding spaces.

$$Y_{pred_{(i,j)}} = \frac{H_{MDD_{X_i}} \cdot H_{SPK_{X_j}}}{||H_{MDD_{X_i}}|| \cdot ||H_{SPK_{X_j}}||}$$

$$Y_{target_{(i,j)}} = 0$$

$$L_{USSD} = MSE(Y_{pred}, Y_{target})$$

$$L_{total-USSD} = L_{MDD} + \alpha \cdot L_{USSD}$$

- ➔ Speaker Classification Model does not need speaker labels as only embeddings are being extracted

III. Experimental Details

- ➔ Dataset: DAIC-WoZ [24]
  - ◆ 189 participants (Male and female)
  - ◆ 42 Depressed (Self-reported)
  - ◆ Audio and text.
- ➔ Input Features
  - ◆ Audio: Mel-spectrograms, Raw-Audio, ComparE16, Wav2Vec2 latent representation [30]
  - ◆ Text: Word2Vec [31]
- ➔ Models
  - ◆ DepAudioNet framework [25] : CNN+LSTM, ECAPA-TDNN [32], LSTM-only
- ➔ Evaluation Metrics
  - ◆ Depression detection: F1-Score (Majority Voting - MV)
  - ◆ Privacy -preservation: DeID Score [37]

De-Identification Score (DeID)

- ❖ Inspired by Voice-privacy literature [38].
- ❖ Evaluates speaker recognizability pre- and post-disentanglement.
- ❖ Scale: 0-100 (100 : Speakers fully unidentifiable post-disentanglement).

Code & Model ➔

IV. Results & Discussion

Baseline vs. ADV vs. USSD for 4 input features and 3 models.

Feature	Model	Disentanglement	Number of Parameters ↓	F1-AVG (MV) ↑	F1-ND ↑	F1-D ↑	*DeID* ↑
Mel-Spectrogram	CNN-LSTM	No	280k	0.658	0.756	0.560	NA
		ADV	293k	0.694	0.773	0.615	14.01%
		USSD	280k	0.683	0.783	0.583	10.29%
	ECAPA-TDNN	No	515k	0.709	0.809	0.609	NA
		ADV	529k	0.746	0.826	0.667	3.69%
		USSD	515k	0.746	0.826	0.667	5.97%
Raw-Audio	CNN-LSTM	No	445k	0.669	0.792	0.546	NA
		ADV	459k	0.709	0.809	0.609	55.83%
		USSD	445k	0.746^{+}	0.826	0.667	45.35%
	ECAPA-TDNN	No	595k	0.694	0.773	0.615	NA
		ADV	609k	**0.790**	0.880	**0.700**	22.32%
		USSD	595k	0.773^{+}	0.851	0.696	19.90%
ComparE16	LSTM-only	No	1.15M	0.694	0.773	0.615	NA
		ADV	1.18M	0.762^{+}	0.857	0.667	68.37%
		USSD	1.15M	0.776	**0.885**	0.667	**92.87%**
Wav2vec2	LSTM-only	No	3.6M	0.683	0.783	0.583	NA
		ADV	3.7M	0.747	0.863	0.632	52.43%
		USSD	3.6M	0.720	0.840	0.600	58.65%

- ➔ Speaker Disentanglement (ADV or USSD) >> Baseline
  - ◆ Avg. improvement in F1-Score - (8.3% for ADV and 8.2% for USSD)
- ➔ F1-USSD ≈ F1- ADV without speaker labels and additional parameters.
- ➔ DeID-USSD > DeID-ADV => USSD has the best speaker disentanglement.

- ➔ ADV
  - ◆ F1-Score: 0.790
  - ◆ DeID: 22.32%
- ➔ USSD
  - ◆ F1-Score: 0.776
  - ◆ DeID: 92.87%

Text-Fusion

Audio-Model	Audio-only	Word2Vec	Audio + Text Fusion	*DeID* (Audio-only)
Raw-Audio ECAPA-TDNN (ADV)	0.790	0.762	0.860	22.32%
ComparE16 LSTM-only (USSD)	0.776	0.762	0.830	92.87%

- ➔ F1 Score: USSD + Text > USSD.
- ➔ Text and USSD may be complimentary.