Adversarial-Free Speaker Identity-Invariant Representation Learning for Automatic Dysarthric Speech Classification

Parvaneh Janbakhshi and Ina Kodrasi

Idiap Research Institute
Speech and Audio Processing Group

Virtual INTERSPEECH 2022

September 2022

Outline

Automatic dysarthria speech classification

- Speech dysarthria due to Parkinson's disease (PD) $\rightarrow$ disturbances of muscular control on speech production system
  - Imprecise articulation, abnormal speech rhythm, reduced stress, breathiness

Automatic dysarthria speech classification

- Speech dysarthria due to Parkinson's disease (PD) $\rightarrow$ disturbances of muscular control on speech production system
  - Imprecise articulation, abnormal speech rhythm, reduced stress, breathiness
- Dysarthric speech classification: discriminating between normal speech and speech from patients with dysarthria (e.g., PD)

Automatic dysarthria speech classification

- » Speech dysarthria due to Parkinson's disease (PD) → disturbances of muscular control on speech production system
  - Imprecise articulation, abnormal speech rhythm, reduced stress, breathiness

- » Dysarthric speech classification: discriminating between normal speech and speech from patients with dysarthria (e.g., PD)

Dysarthric speech classification using:

- » Subjective screening based on judgement of medical practitioners
  - Labor-intensive
  - Inconsistency
  - Difficulties with early diagnosis

- » Automatic and objective method
  - Efficient and economical
  - Repeatable
  - Early diagnosis

Automatic dysarthria speech classification

- Speech dysarthria due to Parkinson's disease (PD) → disturbances of muscular control on speech production system
  - Imprecise articulation, abnormal speech rhythm, reduced stress, breathiness
- Dysarthric speech classification: discriminating between normal speech and speech from patients with dysarthria (e.g., PD)

Dysarthric speech classification using:

- Subjective screening based on judgement of medical practitioners
  - Labor-intensive
  - Inconsistency
  - Difficulties with early diagnosis
- Automatic and objective method
  - Efficient and economical
  - Repeatable
  - Early diagnosis

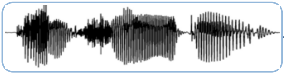
Outline

State-of-the-art automatic dysarthric speech classification systems

- Traditional machine learning approaches
- Deep learning approaches

State-of-the-art automatic dysarthric speech classification systems

» Traditional machine learning approaches (Hegde et al., 2019; Kodrasi and Bourlard, 2020; Hernandez et al., 2020)



Handcrafted feature computation

e.g., MFCCs, sparsity parameter

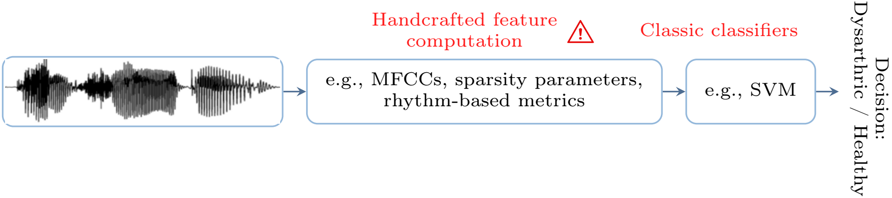
Classic classifiers

e.g., SVM

Decision: Dysarthric / Healthy

State-of-the-art automatic dysarthric speech classification systems

» Traditional machine learning approaches (Hegde et al., 2019; Kodrasi and Bourlard, 2020; Hernandez et al., 2020)



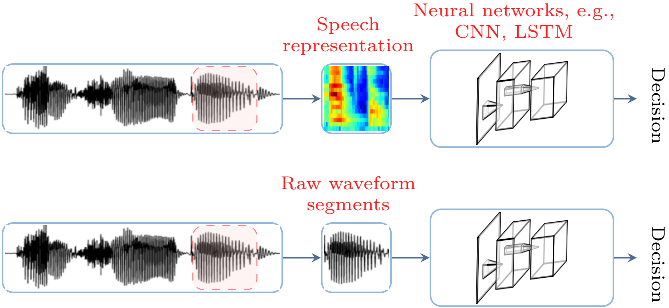
⚠ May fail to adequately capture pathological speech characteristics

⚠ May fail to characterize abstract but important acoustic cues

State-of-the-art automatic dysarthric speech classification systems

» Deep learning approaches → data-driven approaches using no prior knowledge

- Exploit high-level abstract features from low-level speech representations or raw waveforms using neural networks (Vaiciukynas et al., 2017; Mallela et al., 2020; Narendra et al., 2021)



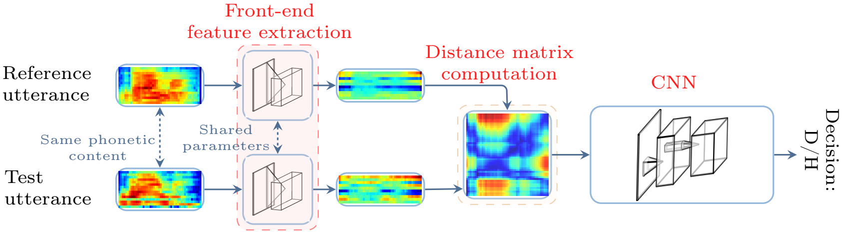
⚠ no explicit attempts to guide the networks to learn robust features

State-of-the-art automatic dysarthric speech classification systems

» Deep learning approaches

- Pairwise training using distance-based CNNs (Janbakhshi et al., 2021)

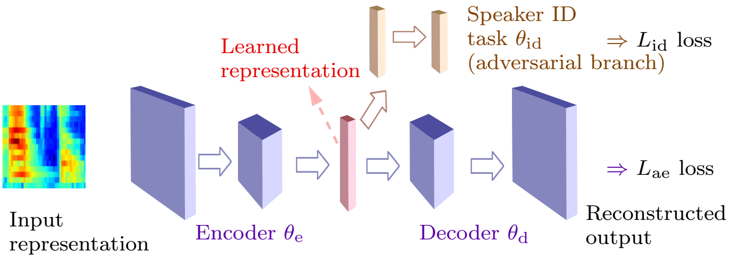
  ⚠ A single network for different but phonetically matched utterances



State-of-the-art automatic dysarthric speech classification systems

» Deep learning approaches

- Supervised speaker identity-invariant representation learning with adversarial training (Janbakhshi and Kodrasi, 2021)

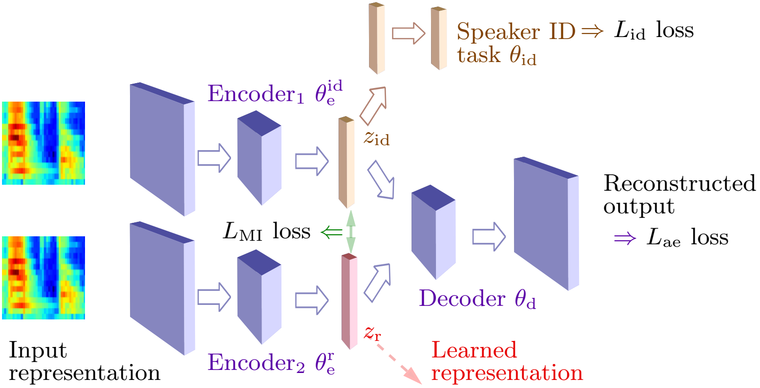


⚠ Difficulty of training the adversarial network for obtaining the speaker identity-invariant representation

Outline

Supervised representation learning via feature separation

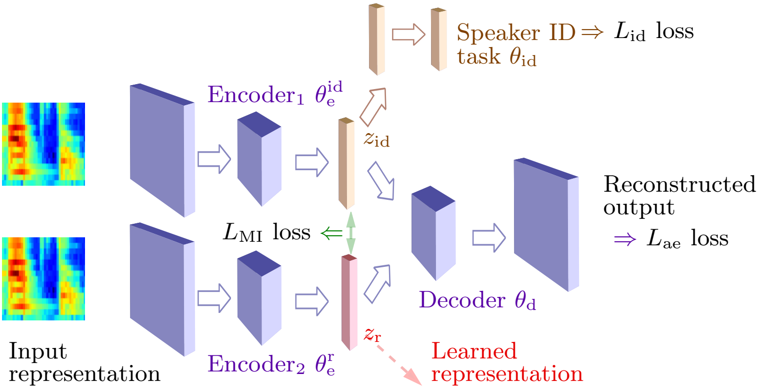
- Speaker identity-invariant representation via feature separation
  - Dual representation learning
  - Jointly <u>minimizing</u> the auto-encoder reconstruction loss and <u>minimizing</u> a speaker ID loss using one of the representations while <u>minimizing</u> a MI criterion between the two representations



Supervised representation learning via feature separation

Speaker identity-invariant representation training → feature separation

- Requires MI approximation, i.e., $I(z_{\mathrm{id}}, z_{\mathrm{r}})$
  - Challenging for high-dimensional variables with unknown probability distributions



Supervised representation learning via feature separation

Speaker identity-invariant representation training → feature separation

- Requires MI approximation, i.e., $I(z_{\mathrm{id}}, z_{\mathrm{r}})$
  - Challenging for high-dimensional variables with unknown probability distributions
  - Variational contrastive log-ratio upper bound (vCLUB) → an upper bound for MI (Cheng et al., 2020)

$$I_{\mathrm{vCLUB}}(z_{\mathrm{id}}, z_{\mathrm{r}}) = \mathbb{E}_{p(z_{\mathrm{id}}, z_{\mathrm{r}})}\left[\log q_{\phi}(z_{\mathrm{id}}|z_{\mathrm{r}})\right] - \mathbb{E}_{p(z_{\mathrm{id}})}\mathbb{E}_{p(z_{\mathrm{r}})}\left[\log q_{\phi}(z_{\mathrm{id}}|z_{\mathrm{r}})\right], \tag{1}$$

  - $q_{\phi}(z_{\mathrm{id}}|z_{\mathrm{r}})$ → variational approximation of $p(z_{\mathrm{id}}|z_{\mathrm{r}})$ → modelled by a neural network with overall parameters of ϕ

Supervised representation learning via feature separation

Speaker identity-invariant representation training $\rightarrow$ feature separation

- Requires MI approximation, i.e., $I(z_{\mathrm{id}}, z_{\mathrm{r}})$
  - Challenging for high-dimensional variables with unknown probability distributions
  - Variational contrastive log-ratio upper bound (vCLUB) $\rightarrow$ an upper bound for MI (Cheng et al., 2020)

$$I_{\mathrm{vCLUB}}(z_{\mathrm{id}}, z_{\mathrm{r}}) = \mathbb{E}_{p(z_{\mathrm{id}}, z_{\mathrm{r}})}\left[\log q_\phi(z_{\mathrm{id}}|z_{\mathrm{r}})\right] - \mathbb{E}_{p(z_{\mathrm{id}})}\mathbb{E}_{p(z_{\mathrm{r}})}\left[\log q_\phi(z_{\mathrm{id}}|z_{\mathrm{r}})\right], \qquad (2)$$

  - $q_\phi(z_{\mathrm{id}}|z_{\mathrm{r}})$ $\rightarrow$ variational approximation of $p(z_{\mathrm{id}}|z_{\mathrm{r}})$ $\rightarrow$ modelled by a neural network with overall parameters of ϕ
  - Approximating ϕ by maximizing the log-likelihood loss, i.e., $L_{\mathrm{ll}}(\phi) = \log q_\phi(z_{\mathrm{id}}|z_{\mathrm{r}})$
  - $L_{\mathrm{MI}}(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}) = I_{\mathrm{vCLUB}}(z_{\mathrm{id}}, z_{\mathrm{r}})$

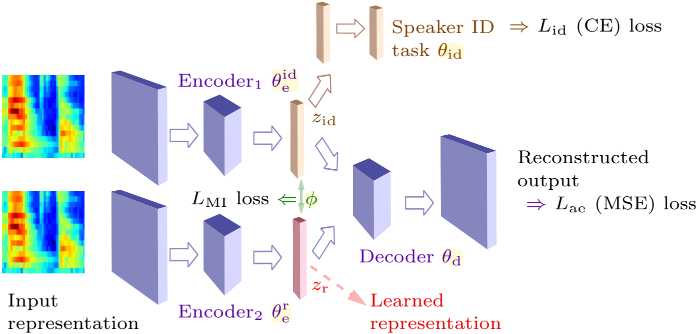
Supervised representation learning via feature separation

Speaker identity-invariant representation training → feature separation

$$(\hat{\theta}_e^{\mathrm{id}}, \hat{\theta}_e^{\mathrm{r}}, \hat{\theta}_{\mathrm{d}}, \hat{\theta}_{\mathrm{id}}) = \arg \min_{\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}} E(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}, \hat{\phi}) \tag{3}$$

$$\hat{\phi} = \arg\min_{\phi} -L_{\mathrm{ll}}(\phi, \hat{\theta}_e^{\mathrm{id}}, \hat{\theta}_e^{\mathrm{r}}) \tag{4}$$

$$E(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}, \hat{\phi}) = L_{\mathrm{ae}}(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \theta_{\mathrm{d}}) + \lambda L_{\mathrm{id}}(\theta_e^{\mathrm{id}}, \theta_{\mathrm{id}}) + \beta L_{\mathrm{MI}}(\theta_e^{\mathrm{id}}, \theta_e^{\mathrm{r}}, \hat{\phi}), \tag{5}$$

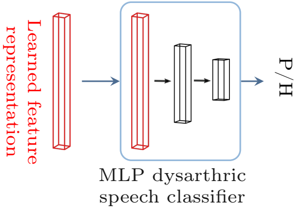


Supervised speech representation learning for pathological detection

» Final dysarthric speech classification
  - Training the final dysarthric speech classifier operating on the learned feature (bottleneck) representation

» Evaluating an unseen test speaker
  - Applying soft voting on the classifier prediction scores for all input Mel spectrograms belonging to that speaker

(Step 2)
final dysarthric speech classification



Outline

Experimental results

- Dataset: Spanish PC-GITA database (Orozco et al., 2014)
  - Discriminating PD patients from healthy speakers
  - Dysarthric speech classification task train/evaluation data → Speaker-independ 10-fold cross-validation framework
  - Speaker ID auxiliary task train/evaluation data → utterance splits of neurotypical speakers in the training set
- Evaluation metrics for dysarthric speech classification and speaker ID
  - Accuracy
  - AUC: area under ROC curve

Experimental results

» Proposed learned representation in the feature separation framework vs. the state-of-the-art representation learned by adversarial training

- Feature separation framework with and without auxiliary modules, i.e., speaker ID ($\lambda \neq 0$) and MI minimizer ($\beta \neq 0$)

λ	β	PD classification ↑		speaker ID ↓	
		Accuracy (%)	AUC	Accuracy (%)	AUC
✗	✗	57.2	0.72	58.3	0.98
✓	✗	61.4	0.75	49.6	0.98
Proposed adversarial-free feature separation framework					
✓	✓	75.2	0.82	5.0	0.67
Adversarial framework from Janbakhshi and Kodrasi (2021)					
–	–	77.0	0.85	5.2	0.67

Experimental results

» Proposed learned representation in the feature separation framework vs. the state-of-the-art representation learned by adversarial training

- Feature separation framework with and without auxiliary modules, i.e., speaker ID ($\lambda \neq 0$) and MI minimizer ($\beta \neq 0$)

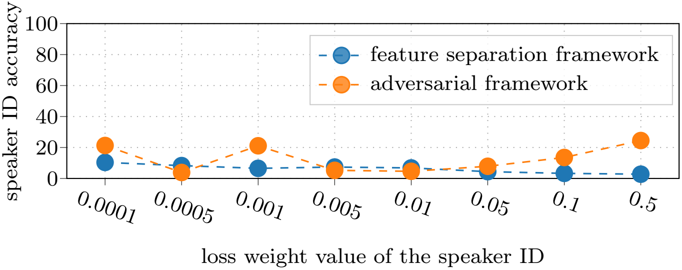
λ	β	PD classification ↑		speaker ID ↓	
		Accuracy (%)	AUC	Accuracy (%)	AUC
✗	✗	57.2	0.72	58.3	0.98
✓	✗	61.4	0.75	49.6	0.98
Proposed adversarial-free feature separation framework					
✓	✓	75.2	0.82	5.0	0.67
Adversarial framework from Janbakhshi and Kodrasi (2021)					
–	–	77.0	0.85	5.2	0.67

- Efficacy of the proposed feature separation framework in suppressing speaker identity cues → robust representation for the dysarthric speech classification
- No statistically significant difference in performance of the feature separation and adversarial framework

Experimental results

» Proposed learned representation in the feature separation framework vs. the state-of-the-art representation learned by adversarial training

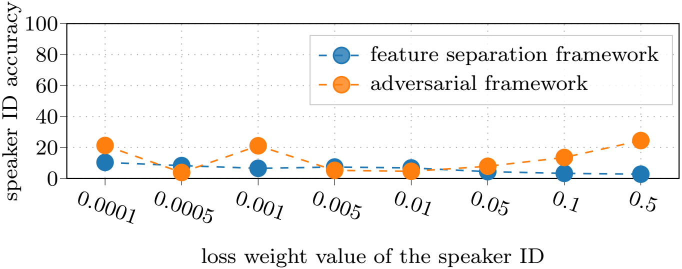
- Sensitivity of speaker-invariant representation learning to the weight values of speaker ID loss (λ)



Experimental results

» Proposed learned representation in the feature separation framework vs. the state-of-the-art representation learned by adversarial training

- Sensitivity of speaker-invariant representation learning to the weight values of speaker ID loss (λ)



- The feature separation framework is less sensitive to the weight values of speaker ID loss compared to adversarial training

Outline

Summary

- Proposing a supervised representation learning framework for dysarthric speech classification

- An adversarial-free feature separation framework to suppress speaker identity cues unrelated to dysarthria in the learned representation
  - Training a dual encoder generating two representations and a single decoder
  - Supervising one of the representations with a speaker ID task while minimizing a MI criterion between the two representations

- Proposed framework is successful in obtaining a speaker identity-invariant representation $\rightarrow$ more informative representations for dysarthric speech classification
  - While is more robust to the choice of some of the training parameters compared to its adversarial counterpart framework

Thank You

Reference

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. (2020). CLUB: A contrastive log-ratio upper bound of mutual information. In *Proc. 37th International Conference on Machine Learning*, volume abs/2006.12013. PMLR.

Hegde, S., Shetty, S., Rai, S., and Dodderi, T. (2019). A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, 33(6):947.e11–947.e33.

Hernandez, A., Yeo, E. J., Kim, S., and Chung, M. (2020). Dysarthria Detection and Severity Assessment Using Rhythm-Based Metrics. In *Proc. 21st Annual Conference of the International Speech Communication Association*, pages 2897–2901, Shanghai, China.

Janbakhshi, P. and Kodrasi, I. (2021). Supervised speech representation learning for Parkinson's disease classification. In *Proc. 14th ITG Conference on speech communication*, pages 1–5, Virtual Conference.

Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2021). Automatic dysarthric speech detection exploiting pairwise distance-based convolutional neural networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 7328–7332, Toronto, Canada.

Kodrasi, I. and Bourlard, H. (2020). Spectro-temporal sparsity characterization for dysarthric speech detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 28(1):1210–1222.

Mallela, J., Illa, A., Belur, Y., Atchayaram, N., Yadav, R., Reddy, P., Gope, D., and Ghosh, P. K. (2020). Raw Speech Waveform Based Classification of Patients with ALS, Parkinson's Disease and Healthy Controls Using CNN-BLSTM. In *Proc. 21st Annual Conference of the International Speech Communication Association*, pages 4586–4590, Shanghai, China.

Narendra, N., Schuller, B., and Alku, P. (2021). The detection of Parkinson's disease from speech using voice source information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1925–1936.

Orozco, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J., González-Rátiva, M., and Noeth, E. (2014). New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *Proc. International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Vaiciukynas, E., Gelzinis, A., Verikas, A., and Bacauskiene, M. (2017). Parkinson's disease detection from speech using convolutional neural networks. In *In Proc. International Conference on Smart Objects and Technologies for Social Good*, pages 206–215, Pisa, Italy. Springer International Publishing.