# Supervised Speech Representation Learning for Parkinson's Disease Classification

Parvaneh Janbakhshi and Ina Kodrasi

Idiap Research Institute

Virtual ITG Conference on Speech Communication

September 2021

# Outline

1. Automatic PD Speech Classification
2. State-of-the-art
3. Proposed Method
4. Experimental Results
5. Summary

# Automatic PD speech classification

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

# Automatic PD speech classification

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

# Automatic PD speech classification

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

- PD speech classification: discriminating between normal speech and speech from patients with PD

PD 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 PD speech classification

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

- PD speech classification: discriminating between normal speech and speech from patients with PD

PD 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

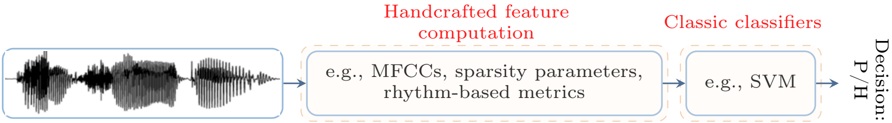
# Outline

# State-of-the-art automatic PD speech classification systems

- Traditional machine learning approaches
- Deep learning approaches

# State-of-the-art automatic PD speech classification systems

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



# State-of-the-art automatic PD 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

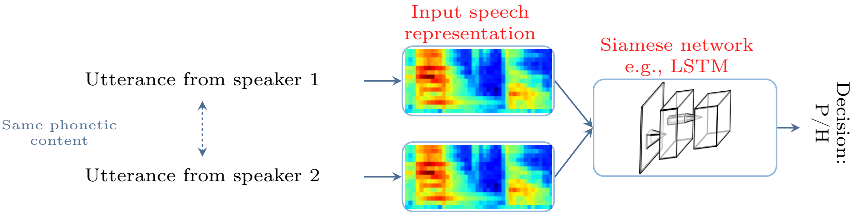
# State-of-the-art automatic PD speech classification systems

- 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
  - Challenge: guiding networks to learn robust and relevant features with limited available pathological training data

# State-of-the-art automatic PD speech classification systems

- Deep learning approaches
  - Pairwise training using LSTM Siamese networks (Bhati et al., 2019)

    ⚠ Different networks for different utterances

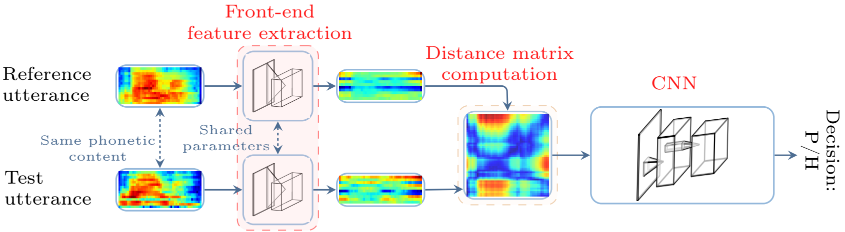


# State-of-the-art automatic PD speech classification systems

» Deep learning approaches

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

⚠ A single network for different but phonetically matched utterances

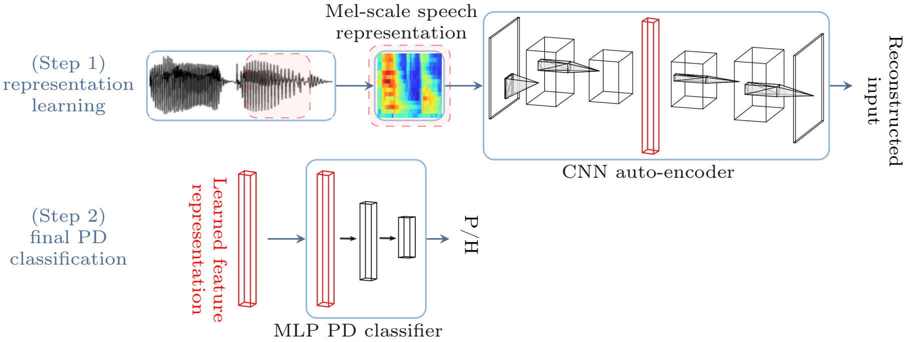


# State-of-the-art automatic PD speech classification systems

» Deep learning approaches

- Unsupervised representation learning (Vasquez-Correa et al., 2020; Karan et al., 2020)

⚠ Learned representations may not be robust to pathology-unrelated cues, e.g., speaker identity and may not be discriminative for pathology detection



# Outline

1. Automatic PD Speech Classification
2. State-of-the-art
3. Proposed Method
4. Experimental Results
5. Summary

# Supervised speech representation learning for PD classification

1. Speaker identity-invariant representation with adversarial training
   - Jointly minimizing the auto-encoder reconstruction loss and minimizing the performance of a (neurotypical) speaker identification (ID) task
   - Improved performance in tasks, e. g., speech emotion classification and phoneme/senone discrimination (Li et al., 2020; Higuchi et al., 2019; Meng et al., 2018)

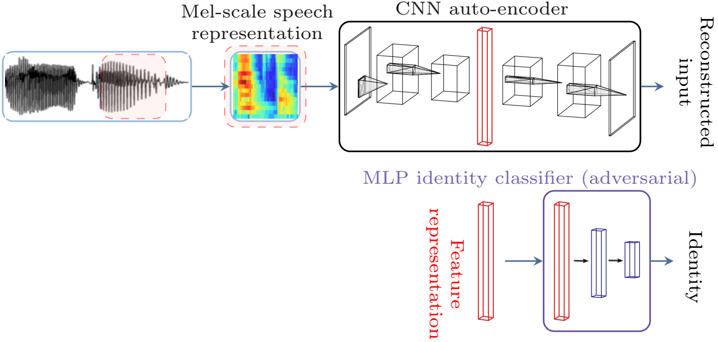
# Supervised speech representation learning for PD classification

1. Speaker identity-invariant representation with adversarial training
   - Jointly minimizing the auto-encoder reconstruction loss and minimizing the performance of a (neurotypical) speaker identification (ID) task
   - Improved performance in tasks, e. g., speech emotion classification and phoneme/senone discrimination (Li et al., 2020; Higuchi et al., 2019; Meng et al., 2018)

2. PD discriminative representation
   - Jointly minimizing the auto-encoder reconstruction loss and maximizing the PD classification performance
   - Such supervision does not harm the performance since the reconstruction loss acts as a regularization method (Le et al., 2018)

# Supervised speech representation learning for PD classification

» Speaker identity-invariant representation (adversarial training)

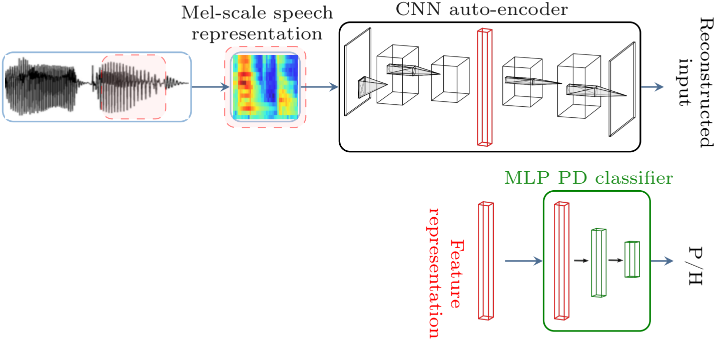



(Step 1)
representation
learning

# Supervised speech representation learning for PD classification

» PD discriminative representation

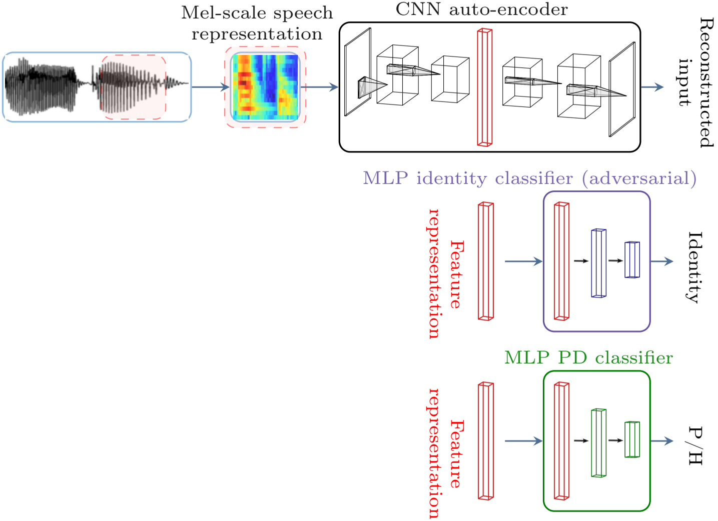
(Step 1) representation learning



# Supervised speech representation learning for PD classification

» Fusion; speaker identity-invariant **and** PD discriminative representation

(Step 1)
representation
learning



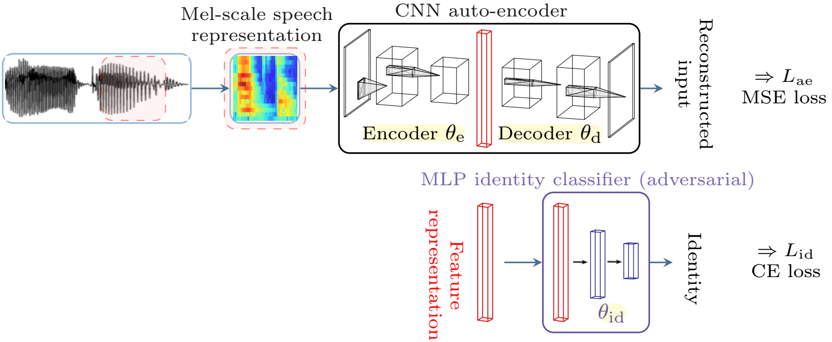
# Supervised speech representation learning for PD classification

» Speaker identity-invariant representation training → min-max objective

$$(\hat{\theta}_{\mathrm{e}}, \hat{\theta}_{\mathrm{d}}, \hat{\theta}_{\mathrm{id}}) = \arg\min_{\theta_{\mathrm{e}}, \theta_{\mathrm{d}}} \arg\max_{\theta_{\mathrm{id}}} E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}), \tag{1}$$

$$E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{id}}) = (1-\lambda) L_{\mathrm{ae}}(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}) - \lambda L_{\mathrm{id}}(\theta_{\mathrm{e}}, \theta_{\mathrm{id}}) \tag{2}$$

- Parameters estimating by alternating training procedure
- Data from neurotypical speakers is used to optimize $L_{\mathrm{id}}$



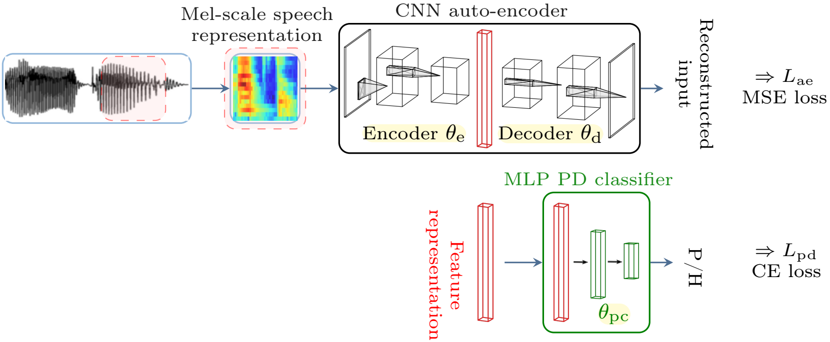
# Supervised speech representation learning for PD classification

» PD discriminative representation training

$$(\hat{\theta}_{\mathrm{e}}, \hat{\theta}_{\mathrm{d}}, \hat{\theta}_{\mathrm{pc}}) = \arg \min_{\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}} E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}), \tag{3}$$

$$E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}) = (1-\alpha) L_{\mathrm{ae}}(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}) + \alpha L_{\mathrm{pc}}(\theta_{\mathrm{e}}, \theta_{\mathrm{pc}}) \tag{4}$$

- Optimal parameters → simultaneously minimizing $L_{\mathrm{ae}}$ and $L_{\mathrm{pd}}$



# Supervised speech representation learning for PD classification

» Fusion

- Jointly learn a speaker identity-invariant and PD discriminative representation → min-max objective

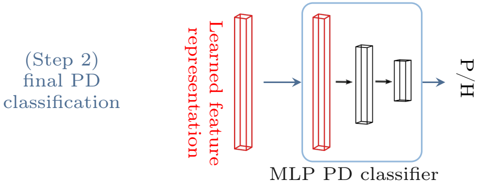
$$(\hat{\theta}_{\mathrm{e}}, \hat{\theta}_{\mathrm{d}}, \hat{\theta}_{\mathrm{pc}}, \hat{\theta}_{\mathrm{id}}) = \arg \min_{\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}} \arg \max_{\theta_{\mathrm{id}}} E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}, \theta_{\mathrm{id}}), \tag{5}$$

$$E(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}, \theta_{\mathrm{pc}}, \theta_{\mathrm{id}}) = (1 - \alpha - \lambda) L_{\mathrm{ae}}(\theta_{\mathrm{e}}, \theta_{\mathrm{d}}) + \alpha L_{\mathrm{pc}}(\theta_{\mathrm{e}}, \theta_{\mathrm{pc}}) - \lambda L_{\mathrm{id}}(\theta_{\mathrm{e}}, \theta_{\mathrm{id}}) \tag{6}$$

# Supervised speech representation learning for PD classification

» Final PD classification
  - Training the final PD 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



# Outline

# Experimental results

- Dataset: Spanish PC-GITA database (Orozco et al., 2014)
  - 50 PD patients vs. 50 healthy speakers each uttering 24 words, 10 sentences, and 1 text
  - PD classification task train/evaluation data → Speaker-independent 10-fold cross-validation framework
  - Speaker ID auxiliary task train/evaluation data → utterance splits of neurotypical speakers in the training set (i.e., 45 speakers)
- Evaluation metrics
  - PD classification accuracy: percentage of correctly classified neurotypical and PD speakers
  - Speaker ID accuracy: percentage of correctly identified speakers
  - AUC: area under ROC curve

# Experimental results

- Tuning hyper-parameters $\lambda$ and $\alpha$ for each speaker identity-invariant and PD discriminative representation training tasks
  - Grid-search (selecting values yielding the highest PD classification accuracy on the development set)
- Hyper-parameters for fusion approach $\rightarrow$ not optimized but set to the values obtained above

# Experimental results

- PD classification performance
  - Baseline (unsupervised) representation learning without auxiliary tasks
  - Supervised representation learnings through auxiliary tasks

| Auxiliary task in representation learning | Accuracy[%] | AUC |
|---|---|---|
| No auxiliary task (baseline) | 66.20 | 0.77 |
| Adversarial identity-invariant training | 72.00 | **0.84** |
| PD discriminative training | 71.00 | 0.78 |
| Fusion (identity-invariant+PD discriminative training) | **75.40** | 0.80 |

# Experimental results

» PD classification performance

- Baseline (unsupervised) representation learning without auxiliary tasks
- Supervised representation learnings through auxiliary tasks

| Auxiliary task in representation learning | Accuracy[%] | AUC |
|---|---|---|
| No auxiliary task (baseline) | 66.20 | 0.77 |
| Adversarial identity-invariant training | 72.00 | **0.84** |
| PD discriminative training | 71.00 | 0.78 |
| Fusion (identity-invariant+PD discriminative training) | **75.40** | 0.80 |

- Any of the proposed auxiliary tasks improves the performance of PD classification compared to the baseline
- Fusing both auxiliary tasks yields a better accuracy

# Experimental results

Investigating the suppression of irrelevant speaker identity information in the learned representations

» Speaker ID performance

| Auxiliary task in representation learning | Accuracy[%] | AUC |
|---|---|---|
| No auxiliary task (baseline) | 34.71 | 0.90 |
| Adversarial identity-invariant training | 2.31 | 0.54 |
| PD discriminative training | 18.15 | 0.76 |
| Fusion (identity-invariant+PD discriminative training) | 2.59 | 0.58 |

# Experimental results

Investigating the suppression of irrelevant speaker identity information in the learned representations

» Speaker ID performance

| Auxiliary task in representation learning | Accuracy[%] | AUC |
|---|---|---|
| No auxiliary task (baseline) | 34.71 | 0.90 |
| Adversarial identity-invariant training | 2.31 | 0.54 |
| PD discriminative training | 18.15 | 0.76 |
| Fusion (identity-invariant+PD discriminative training) | 2.59 | 0.58 |

- Highest speaker ID performance using unsupervised representation → speaker identity cues reduce PD classification performance
- Lowest speaker ID performance using the speaker identity-invariant representation
- Lower speaker ID performance using PD discriminative representation → speaker identity cues are less relevant to the PD classification task

# Outline

# Summary

- Using supervised representation learning frameworks with auxiliary tasks for PD classification
- Reducing irrelevant speaker identity cues in the representation
  - Training an auto-encoder jointly with an adversarial auxiliary speaker ID task
- Obtaining a discriminative representation for PD classification
  - Training an auto-encoder jointly with an auxiliary PD classifier
- Supervised representation learning is advantageous for PD classification, outperforming using representations learned in an unsupervised manner

*Thank You*

# Reference


Bhati, S., Velazquez, L. M., Villalba, J., and Dehak, N. (2019). LSTM siamese network for parkinson's disease detection from speech. In *In Proc. IEEE Global Conference on Signal and Information Processing*, pages 1–5, Ottawa, Canada.

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.

Higuchi, Y., Tawara, N., Kobayashi, T., and Ogawa, T. (2019). Speaker adversarial training of DPGMM-based feature extractor for zero-resource languages. In *Proc. Annual Conference of the International Speech Communication Association*, pages 266–270, Graz, Austria.

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.

Karan, B., Sahu, S. S., and Mahto, K. (2020). Stacked auto-encoder based time-frequency features of speech signal for Parkinson disease prediction. In *Proc. International Conference on Artificial Intelligence and Signal Processing*, pages 1–4, Amaravati, India.

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.

Le, L., Patterson, A., and White, M. (2018). Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *Proc. International Conference on Neural Information Processing Systems*, pages 107–117, Montréal, Canada.

Li, H., Tu, M., Huang, J., Narayanan, S., and Georgiou, P. (2020). Speaker-invariant affective representation learning via adversarial training. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7144–7148, Barcelona, Spain.

Meng, Z., Li, J., Chen, Z., Zhao, Y., Mazalov, V., Gong, Y., and Juang, B.-H. (2018). Speaker-invariant training via adversarial learning. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5969–5973, Calgary, Canada.

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.

Vasquez-Correa, J., Arias-Vergara, T., Schuster, M., Orozco-Arroyave, J., and Nöth, E. (2020). Parallel representation learning for the classification of pathological speech: Studies on Parkinson's disease and cleft lip and palate. *Speech Communication*, 122:56–67.