Experimental Investigation on STFT Phase Representations for Deep Learning-based Dysarthric Speech Detection

Parvaneh Janbakhshi and Ina Kodrasi

Idiap Research Institute

Virtual ICASSP 2022

May 2022

Outline

Automatic dysarthric speech detection

- Dysarthria of speech → disturbances of muscular control on speech production system
  - Imprecise articulation, abnormal speech rhythm, pitch variation, breathiness

Automatic dysarthric speech detection

- Dysarthria of speech → disturbances of muscular control on speech production system
  - Imprecise articulation, abnormal speech rhythm, pitch variation, breathiness
- Dysarthric speech detection: discriminating between speech from healthy and dysarthric speakers

Automatic dysarthric speech detection

- Dysarthria of speech → disturbances of muscular control on speech production system
  - Imprecise articulation, abnormal speech rhythm, pitch variation, breathiness

- Dysarthric speech detection: discriminating between speech from healthy and dysarthric speakers

Dysarthric speech detection using:

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

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

Automatic dysarthric speech detection

- Dysarthria of speech → disturbances of muscular control on speech production system
  - Imprecise articulation, abnormal speech rhythm, pitch variation, breathiness

- Dysarthric speech detection: discriminating between speech from healthy and dysarthric speakers

Dysarthric speech detection using:

- Subjective screening based on judgment 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 detection

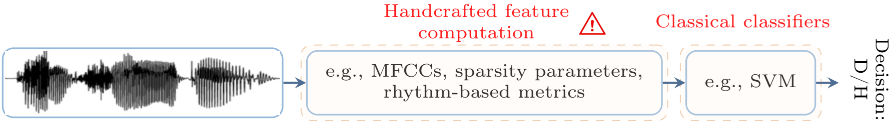
- Traditional machine learning approaches
- Deep learning approaches

State-of-the-art automatic dysarthric speech detection

» Traditional machine learning approaches (Kodrasi and Bourlard, 2020; Hernandez et al., 2020; Solana-Lavalle and Rosas-Romero, 2021)

State-of-the-art automatic dysarthric speech detection

» Traditional machine learning approaches (Kodrasi and Bourlard, 2020; Hernandez et al., 2020; Solana-Lavalle and Rosas-Romero, 2021)



⚠ May fail to adequately capture dysarthric speech characteristics

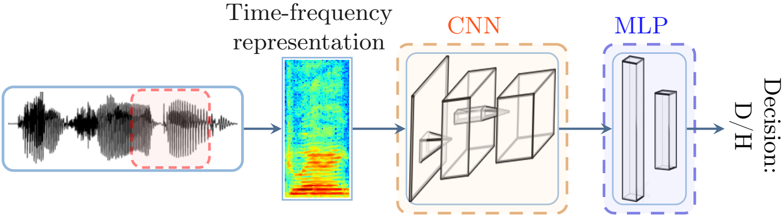
⚠ May fail to characterize abstract but important acoustic cues

State-of-the-art automatic dysarthric speech detection

- Deep learning approaches → data-driven approaches using no prior knowledge
  - Exploit high-level abstract features from low-level time-frequency speech representations or raw waveform using neural networks

State-of-the-art automatic dysarthric speech detection

- Mainstream deep learning approaches
  - Rely on processing magnitude spectrum (or features derived from the magnitude spectrum) (Vaiciukynas et al., 2017; Vasquez et al., 2017; An et al., 2018)

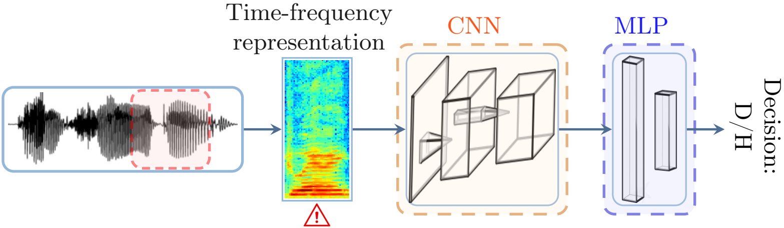


State-of-the-art automatic dysarthric speech detection

» Mainstream deep learning approaches

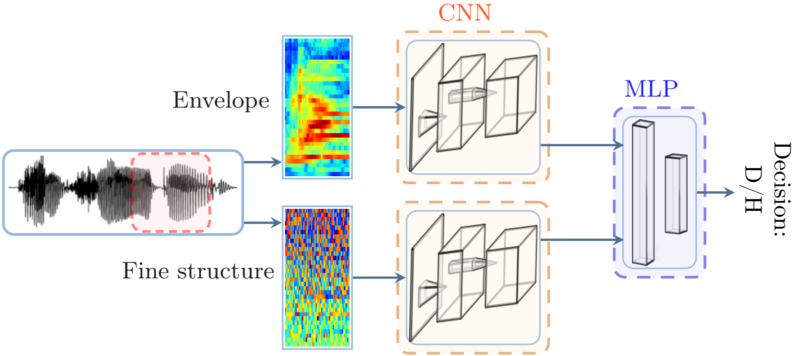
- Rely on processing magnitude spectrum (or features derived from the magnitude spectrum) (Vaiciukynas et al., 2017; Vasquez et al., 2017; An et al., 2018)

⚠ Ignoring complementary acoustic cues in phase spectrum



State-of-the-art automatic dysarthric speech detection

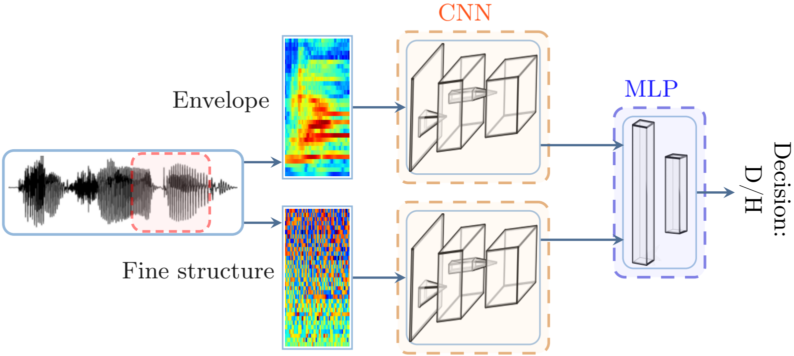
- Deep learning approaches
  - Dual CNN-based framework using temporal envelope and fine structures (Kodrasi, 2021)



State-of-the-art automatic dysarthric speech detection

- Deep learning approaches
  - Dual CNN-based framework using temporal envelope and fine structures (Kodrasi, 2021)

⚠ Not clear if the incorporation of the analytical phase was beneficial



Outline

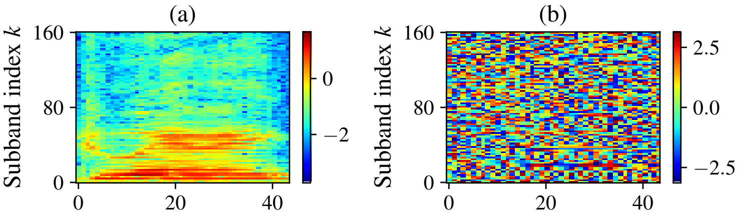
STFT phase representations for dysarthric speech detection

STFT magnitude and phase representations

$$S_{k,l} = |S_{k,l}|e^{j\theta_{k,l}},$$

(a) $|S_{k,l}| \rightarrow$ the magnitude of l-th segment at the k-th subband

(b) $\theta_{k,l} \rightarrow$ the phase of l-th segment at the k-th subband



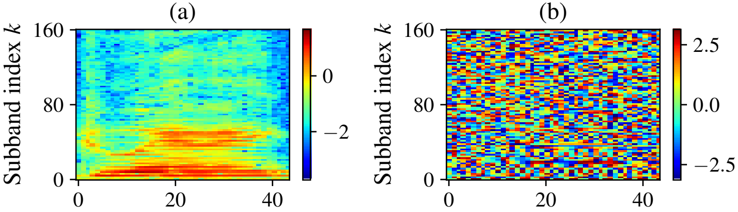
STFT phase representations for dysarthric speech detection

STFT magnitude and phase representations

$$S_{k,l} = |S_{k,l}| e^{j\theta_{k,l}},$$

(a) $|S_{k,l}| \rightarrow$ the magnitude of l-th segment at the k-th subband

(b) $\theta_{k,l} \rightarrow$ the phase of l-th segment at the k-th subband



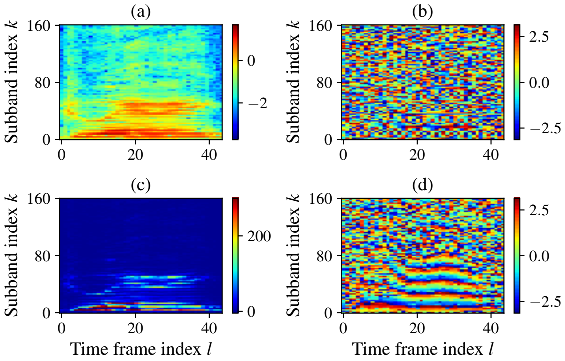
» Expected difficulty in processing the phase spectrum

- Discontinuous and irregular → lack of visible spectro-temporal patterns

STFT phase representations for dysarthric speech detection

» Proposing to use two alternative phase representations revealing spectro-temporal structures

- Modified group delay (MGD) spectrum→ reflecting the cepstrally smoothed derivative of phase along the **frequency** axis (Murthy and Gadde, 2003)
- Instantaneous frequency (IF) spectrum→ reflecting the derivative of phase along the **time axis** (Stark and Paliwal, 2008)



(a) Log of magnitude
(b) Unprocessed phase
(c) MGD
(d) IF

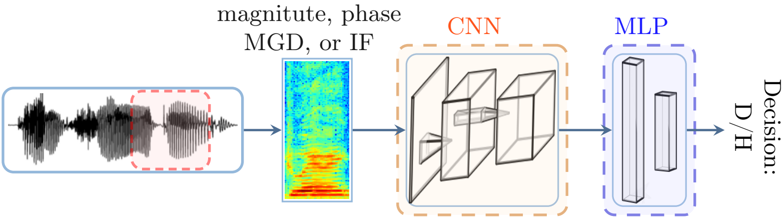
STFT phase representations for dysarthric speech detection

- Proposing to use two alternative phase representations revealing spectro-temporal structures
  - Modified group delay (MGD) spectrum (Murthy and Gadde, 2003)
    - Expected to capture articulation deficiencies and vowel quality changes in dysarthric speech
  - Instantaneous frequency (IF) spectrum (Stark and Paliwal, 2008)
    - Expected to capture pitch variation in dysarthric speech

Dysarthric speech detection experiments via phase representations

» Baseline networks and the proposed network

- Baseline single input CNNs
- Baseline dual input CNN using temporal envelope and fine structure
- Proposed dual input CNNs

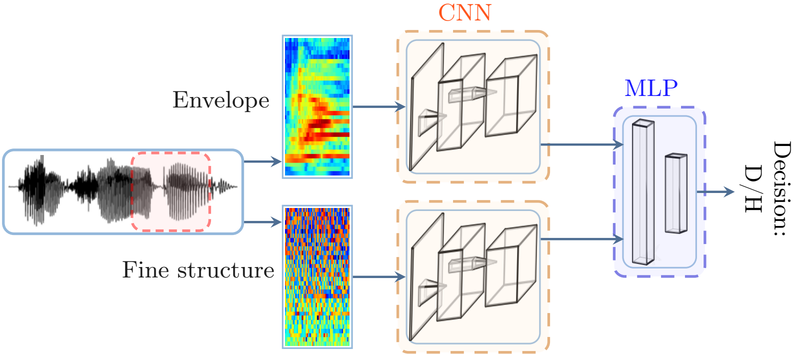


Using either of STFT magnitude, unprocessed phase, MGD, IF spectra as input

Dysarthric speech detection experiments via phase representations

» Baseline networks and the proposed network

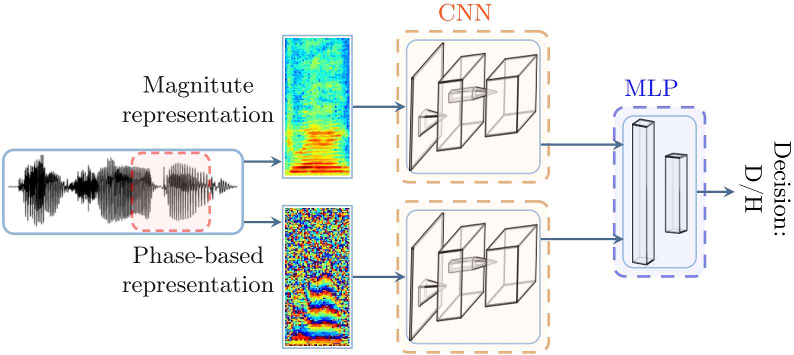
- Baseline single input CNNs
- Baseline dual input CNN using temporal envelope and fine structure (Kodrasi, 2021)
- Proposed dual input CNNs



Dysarthric speech detection experiments via phase representations

» Baseline networks and the proposed network

- Baseline single input CNNs
- Baseline dual input CNN using temporal envelope and fine structure
- Proposed dual input CNNs



Outline

Experimental results

» Dataset: Spanish PC-GITA database (Orozco et al., 2014)

- 50 PD patients vs. 50 healthy speakers
- Speaker-independent 10-fold cross-validation framework

» Evaluation metrics

- Dysarthric speech detection accuracy: percentage of correctly classified neurotypical and PD speakers
- AUC: area under ROC curve

Experimental results

- Dysarthric speech detection performance
  - Using the baseline single input CNNs operating on magnitude and phase representations

Representation	Accuracy [%]	AUC
Magnitude	69.72	0.77
Phase	62.76	0.70
MGD	70.78	**0.79**
IF	**72.64**	**0.79**

Experimental results

- Dysarthric speech detection performance
  - Using the baseline single input CNNs operating on magnitude and phase representations

Representation	Accuracy [%]	AUC
Magnitude	69.72	0.77
Phase	62.76	0.70
MGD	70.78	**0.79**
IF	**72.64**	**0.79**

  - Any of the phase representations similarly to the magnitude spectrum provide useful cues for dysarthric speech detection
  - IF gives the highest performance while the unprocessed phase gives the lowest performance

Experimental results

» Dysarthric speech detection performance

- Using the baseline dual input CNN and proposed dual input CNNs

Representation	Accuracy [%]	AUC
Magnitude-Phase	87.32	0.93
Magnitude-MGD	80.92	0.90
Magnitude-IF	**93.68**	**0.97**
Baseline Envelope-Fine structure	86.04	0.94

Experimental results

» Dysarthric speech detection performance

- Using the baseline dual input CNN and proposed dual input CNNs

Representation	Accuracy [%]	AUC
Magnitude-Phase	87.32	0.93
Magnitude-MGD	80.92	0.90
Magnitude-IF	**93.68**	**0.97**
Baseline Envelope-Fine structure	86.04	0.94

- All dual input CNNs outperform their single input counterparts → all phase representations contain complementary cues to the magnitude spectrum
- Using the magnitude and IF spectra yields the highest performance, outperforming the baseline dual input CNN

Outline

Summary

- » Investigating the applicability of STFT phase representations for dysarthric speech detection
  - Two alternative representations revealing hidden structures of the phase spectrum, i.e., the MGD and IF spectra
- » Using a single input CNN it has been shown that all considered phase representations contain dysarthric cues
- » Using a dual input CNN it has been shown that all phase representations serve as complementary features to the magnitude spectrum
  - Combination of magnitude and IF spectra yields a high performance

Thank You

Reference

An, K., Kim, M., Teplansky, K., Green, J., Campbell, T., Yunusova, Y., Heitzman, D., and Wang, J. (2018). Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. In *Proc. Annual Conference of the International Speech Communication Association*, Hyderabad, India.

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.

Kodrasi, I. (2021). Temporal envelope and fine structure cues for dysarthric speech detection using CNNs. *IEEE Signal Processing Letters*, 28:1853–1857.

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.

Murthy, H. and Gadde, V. (2003). The modified group delay function and its application to phoneme recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I–68–I–71, Hong Kong, China.

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.

Solana-Lavalle, G. and Rosas-Romero, R. (2021). Analysis of voice as an assisting tool for detection of Parkinson's disease and its subsequent clinical interpretation. *Biomedical Signal Processing and Control*, 66:102415.

Stark, A. and Paliwal, K. (2008). Speech analysis using instantaneous frequency deviation. pages 2602–2605, Brisbane, Australia.

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.

Vasquez, J., Orozco, J. R., and Noeth, E. (2017). Convolutional neural network to model articulation impairments in patients with Parkinson's disease. In *In Proc. Annual Conference of the International Speech Communication Association*, pages 314–318, Stockholm, Sweden.