IEEE ICASSP 2022, Paper ID: 3663

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

Parvaneh Janbakhshi[1,2], Ina Kodrasi[1]

[1]Idiap Research Institute, Martigny, Switzerland [2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

{parvaneh.janbakhshi, ina.kodrasi}@idiap.ch

Aim

- Automatic discrimination between healthy and dysarthric speech using deep learning approaches

State-of-the-art approaches

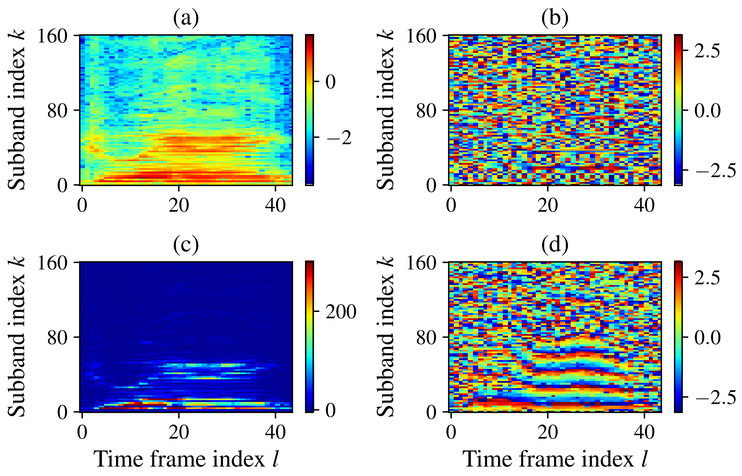
- Mainstream approaches rely on processing the speech magnitude spectrum while ignoring the phase spectrum
- The phase spectrum also contains useful insights

Objectives

- Investigating the applicability of the phase spectrum and alternative phase representations for dysarthric speech detection
- Exploiting phase representations as complementary features to the magnitude spectrum to improve dysarthric speech detection performance

Challenge

- The phase spectrum is discontinuous and exhibits irregular (non-visible) spectro-temporal patterns

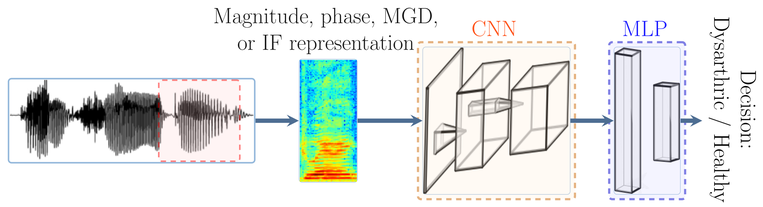


(a) Logarithm of the STFT magnitude **(b)** Phase of the STFT spectrum
(c) Modified group delay (MGD) **(d)** Instantaneous frequency (IF)

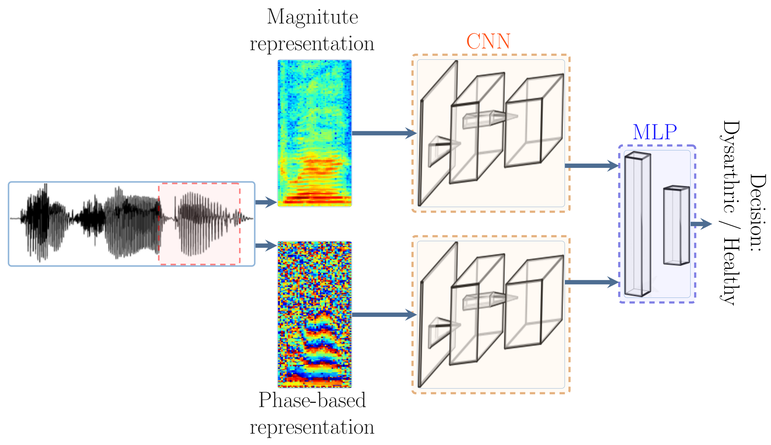
Proposed method

- Exploiting alternative phase representations to reveal spectro-temporal structures hidden in the phase:
  - Modified group delay (MGD) spectrum → reflecting the derivative of phase along the frequency axis
  - Instantaneous frequency (IF) spectrum → reflecting the derivative of phase along the time axis
- Training single input CNN-based frameworks operating a single representation, i.e., magnitude (baseline), unprocessed phase, MGD, or IF
- Training dual input CNN-based frameworks operating on pairs of STFT magnitude spectrum and phase representations, i.e., (magnitude and phase spectra, magnitude and MGD spectra, or magnitude and IF spectra)

- Single input CNN



- Dual input CNN



Evaluation

Database

- 50 Spanish-speaking patients with Parkinson's disease vs. 50 healthy speakers (10-fold speaker-independent cross-validation paradigm)

Performance evaluation

- Detection accuracy and AUC, i.e., area under ROC curve

Results

Performance using the single input CNN operating on the magnitude or phase spectrum

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

Performance using the dual input CNN jointly operating on the magnitude spectrum and different phase representations

Representation	Accuracy [%]	AUC
Magnitude-Phase	87.32	0.93
Magnitude-MGD	80.92	0.90
Magnitude-IF	**93.68**	**0.97**

Conclusion

- Dysarthric cues are present in all considered phase representations
- All phase representations contain complementary cues to the magnitude spectrum, with all dual input CNNs yielding a considerably better performance than their single input counterparts
- Combining the magnitude and IF spectra yields the highest dysarthric speech detection performance