Modeling Of Pre-trained Neural Network Embeddings Learned From Raw Waveform For COVID-19 Infection Detection

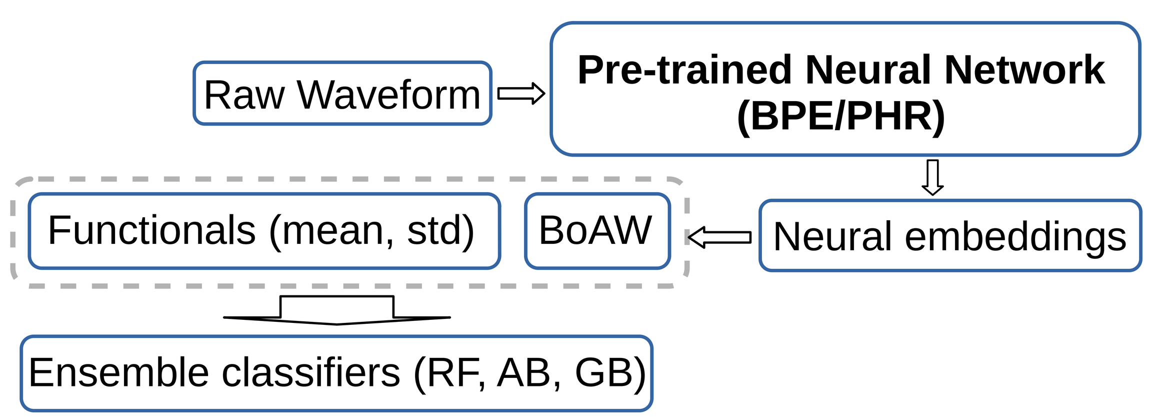
Zohreh Mostaani[1,2] RaviShankar Prasad[1] Bogdan Vlasenko[1] Mathew Magimai-Doss[1]

[1]Idiap Research Institute, Switzerland [2]École polytechnique fédérale de Lausanne, Switzerland

Introduction

- COVID-19 is a respiratory disease.
- Cough sounds and speech based diagnosis of COVID-19 has gained interest.
- Interspeech 2021 ComParE and DiCOVA challenges have propelled the research in this direction.
- DiCOVA II:
  - Breathing (4.6 hrs), Cough (1.7 hours), and Speech (3.9 hours).
  - Total: 965, COVID-19 positive: 172, COVID-19 negative: 793.

Proposed Method



Acoustic features representations

- ComParE LLDs:
  - Functionals: 6373 dimensional vector (CMP_F)
  - BoaW: two sets of codebooks with size 50 for LLDs and ΔLLDs (CMP_L)
- Phoneme Recognition: 1024 dimensional embedding
  - Mean, std: $f_{\mu\sigma}(PHR)$
  - BoaW: one codebook with size 100 $BoAW(PHR)$
- Breathing pattern estimation: 10 dimensional embedding
  - Mean, std: $f_{\mu\sigma}(BPE)$
  - BoaW: one codebook with size 100 $BoAW(BPE)$

Classification

- Ensemble classifiers, grid search and AUC as optimization criterion:
  - Random Forest (RF)
  - Ada Boost (AB)
  - Gradient Boost (GB)
- Fusion:
  - Early fusion (EF): Feature level combination
  - Late fusion (LF): Aggregating (unweighted) posteriors of several classifiers

Results

- Track 1: breathing; Track 2: cough; Track 3: speech; Track 4: Fusion
- The results are expressed in AUC metric and the sensitivity is given for specificity 95% on the Test set
- PHR neural embeddings can yield better systems than hand-crafted LLD-based systems and BPE embedding-based systems
- BPE embedding-based system yields slightly lower performance than LLD-based system but considerably better sensitivity.
- PHR neural embeddings consistently yield better system than BPE neural embeddings (Also look at the ROC plot). One of the reason could be that the effects of COVID-19 for participants could be more discriminatory at articulatory level in comparison to BPE embedding level.

System: Feature	System: Classifier	Dev (%)	Test (%)	Sensitivity (%)
Track 1				
CMP_F	RF	77.83	76.78	30.0
$BoAW(CMP_L)$	RF	73.58	74.52	31.67
CMP_F , $BoAW(CMP_L)$	LF $^{[I]}$	77.56	**78.05**	**43.33**
BASELINE	BLSTM	77.25	84.50	31.67
Track 2				
$BoAW$(PHR)	RF	70.06	74.19	30.0
$f_{\mu\sigma}$(PHR)	RF	70.54	72.87	26.67
CMP_L	RF	66.09	66.68	16.67
$f_{\mu\sigma}$(PHR), $BoAW$(PHR)	LF $^{[II]}$	71.32	**74.63**	**31.67**
BASELINE	BLSTM	75.21	74.89	36.67
Track 3				
$BoAW$(PHR)	RF $^{[III]}$	77.37	80.08	**41.67**
$f_{\mu\sigma}$(PHR)	RF	76.33	79.3	26.67
$BoAW$(BPE)	RF	68.93	73.49	21.67
$f_{\mu\sigma}$(BPE)	RF	68.44	—	—
$BoAW(CMP_L)$	RF	70.38	75.59	15.0
EF($f_{\mu\sigma}$(PHR), $f_{\mu\sigma}$(BPE))	RF $^{[IV]}$	76.67	79.1	28.33
EF($BoAW$(PHR), $BoAW$(BPE) , $BoAW(CMP_L)$)	RF	77.47	79.95	33.33
$f_{\mu\sigma}$(PHR), $BoAW$(PHR)	LF	77.59	**80.64**	36.67
BASELINE	BLSTM	80.16	84.26	43.33
Track 4				
III, IV	LF	77.79	**80.51**	40.0
I, IV	LF	80.09	78.05	**43.33**
I, II	LF	77.93	78.05	**43.33**
BASELINE	LF	81.67	84.70	55.0

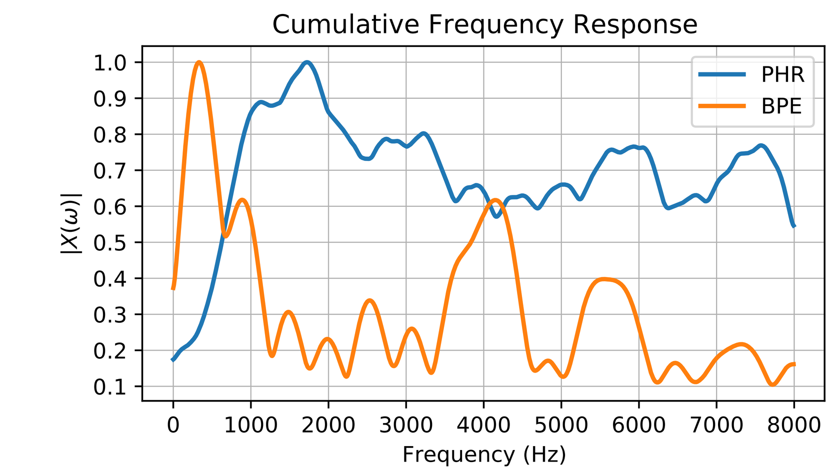
The most discriminating LLDs and functionals

- All Tracks: The auditory spectra coefficients obtained using RASTA filtering and their deltas.
- Track 1: coefficients obtained as the third quartile of these features.
- Track 2: an extended list of functionals prove significant with features capturing primarily the spectral shape.
- Track 3: speech specific features such as MFCC and spectral band energy.

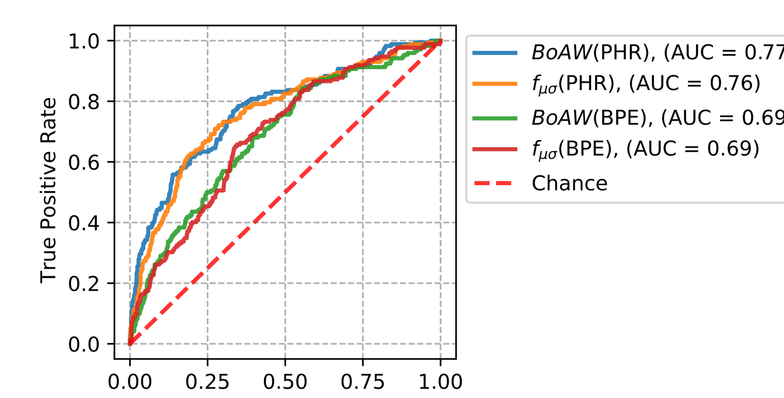
LLDs	functional
Track 1	
Δ audSpec_Rfilt	3^{rd} quartile
voicing parameters	LP–gain
magnitude spectra	RollOff
Δ magnitude spectra	variance
Track 2	
audSpec_Rfilt	regression coefficients, centroid, 2^{nd} quartile
Δ Pitch contour	regression coefficients
Δ RMSenergy	extremums
band energy magnitude spectra	extremums
magnitude spectral slope	regression coefficients
Track 3	
audSpec_Rfilt	regression coefficients, 1^{st} quartile
mfcc	peak behavior , percentiles
Δ audSpec_Rfilt	peak behavior
Δ magnitude spectra	moments

Cumulative frequency response + ROC

The PHR network emphasizes around the formant frequency regions in speech, while the emphasis of the BPE network is significantly towards the lower frequency region.



The cumulative frequency response of the kernels for the first convolution layer of the CNN models: PHR and BPE.



ROC plot for systems trained using PHR embeddings and BPE embeddings on the Dev set of Track 3.

Conclusion

Our studies demonstrate that modeling neural embeddings from neural networks trained on auxiliary or other speech tasks for COVID-19 infection detection is a promising direction and can replace hand-crafted features.

Acknowledgements

This work was partially funded by the Swiss National Science Foundation through the project Towards Integrated processing of Physiological and Speech signals (TIPS), grant no. 200021_188754.