INTERSPEECH 2019

Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks

Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, Yoshua Bengio

santi.pascual@upc.edu

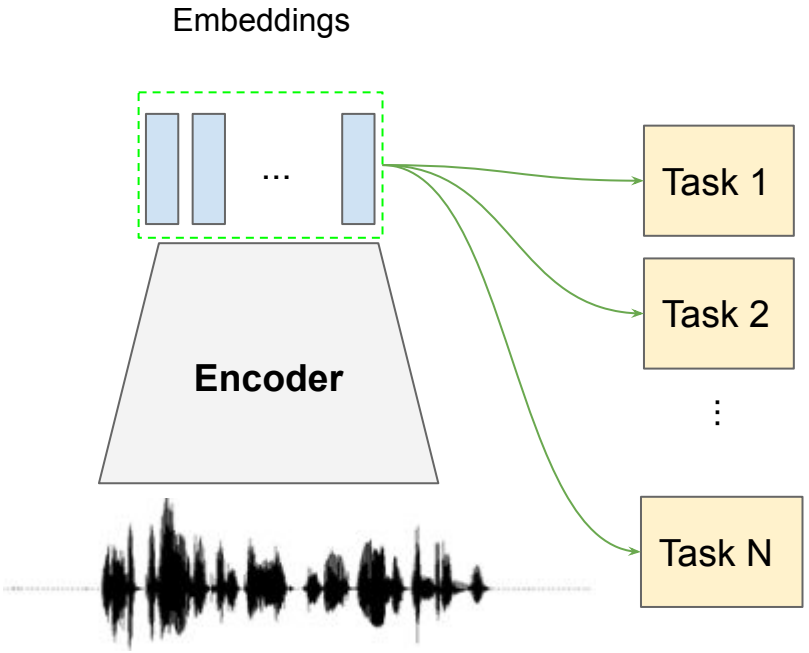
Universitat Politècnica de Catalunya
Telefónica Research
MILA - Université de Montréal

Outline

- Introduction
  - Motivation and contributions
  - Self-supervised learning
- Problem-Agnostic Speech Encoder
  - Architecture
  - Worker Learning
- Evaluation Methodology
- Results
- Conclusions

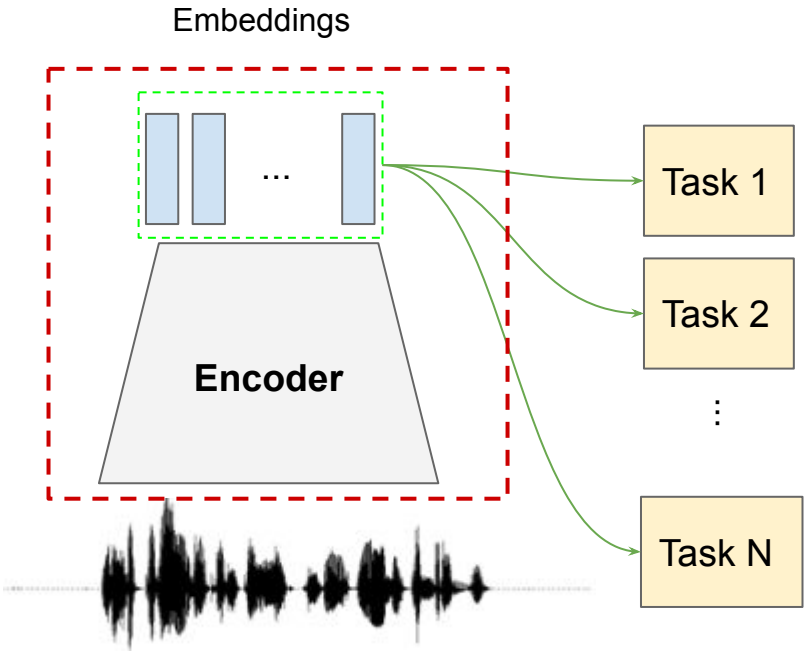
Introduction

Motivation



A generalizable speech encoder that yields compact representations which are problem-agnostic and can be used for any task.

Motivation



A generalizable speech encoder that yields compact representations which are problem-agnostic and can be used for any task.

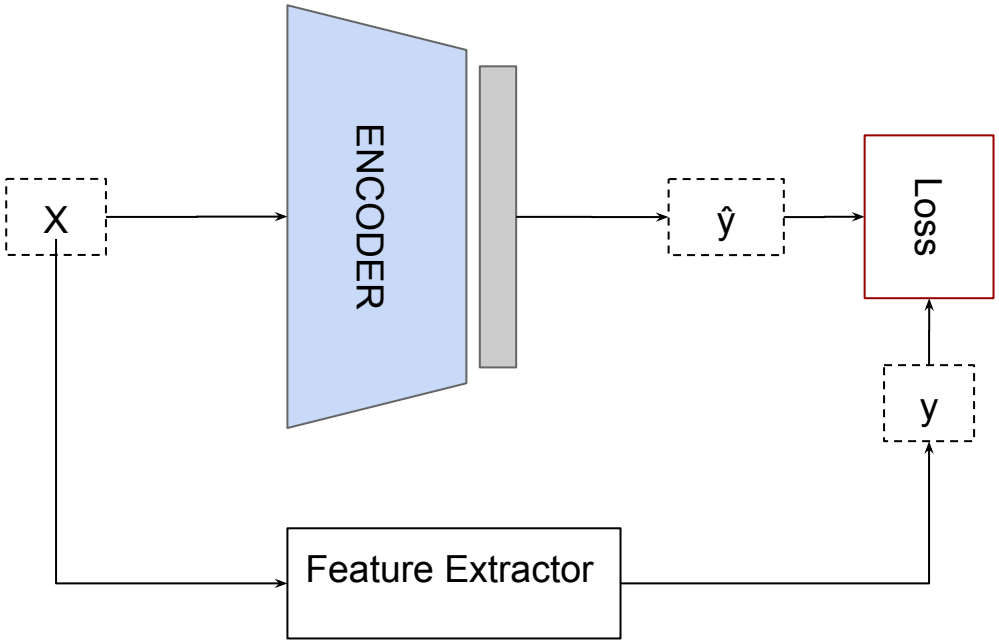
BUT, HOW SHOULD WE TRAIN THIS?

Contributions

1. **Problem-Agnostic Speech Encoder (PASE)**
   A fully convolutional front-end that processes speech waveforms and embeds them, emulating a sliding window.

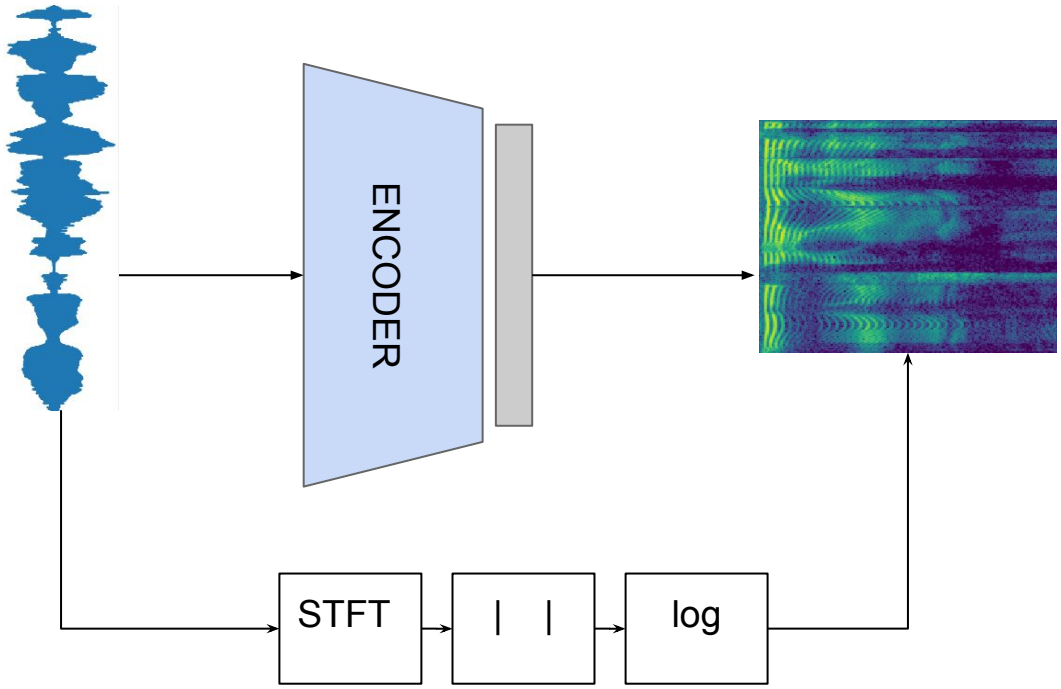
2. **Multi-worker Learning**
   Self-supervised learning with multiple sub-networks (workers), predicting different levels of speech features.

Self-supervised Learning



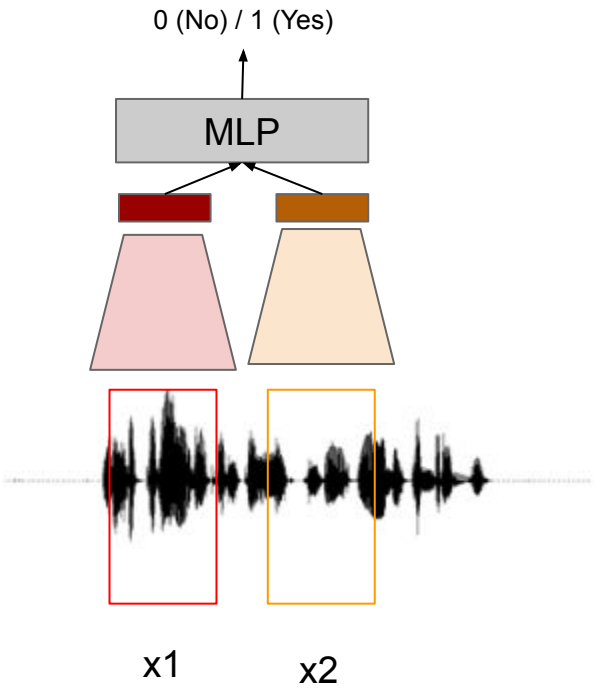
Self-supervised Learning

Example: Predict log-power spectrum



Self-supervised Learning

Example: Detect if waveform chunks are neighbours



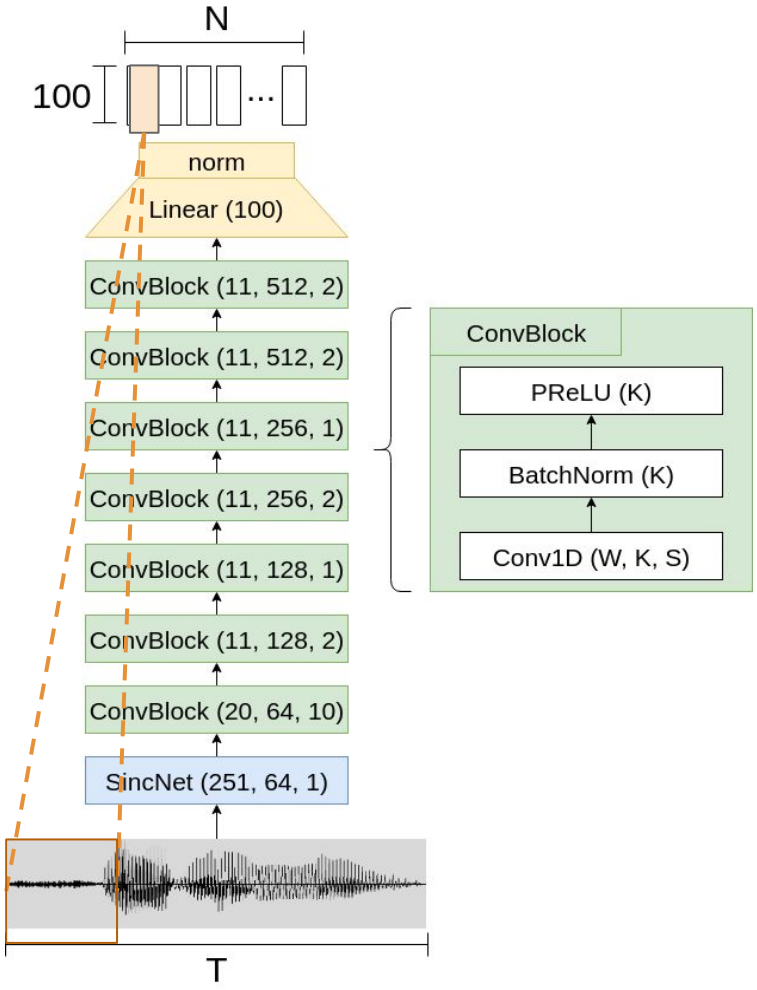
Problem-Agnostic Speech Encoder

Architecture

Designing a specialized front-end for speech.

Deep structure total stride:
TS = 10x2x1x2x1x2x2 = 160 = 10ms@16kHz

Similar to typical STFT processing pipelines.

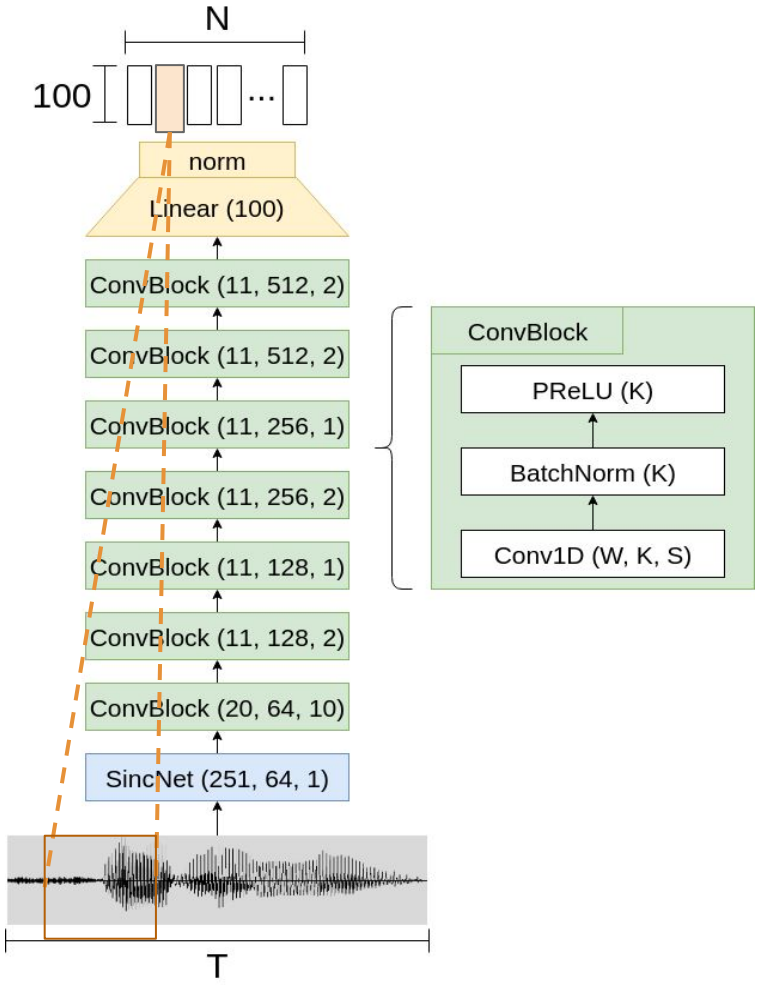


Architecture

Designing a specialized front-end for speech.

Deep structure total stride:
TS = 10x2x1x2x1x2x2 = 160 = 10ms@16kHz

Similar to typical STFT processing pipelines.

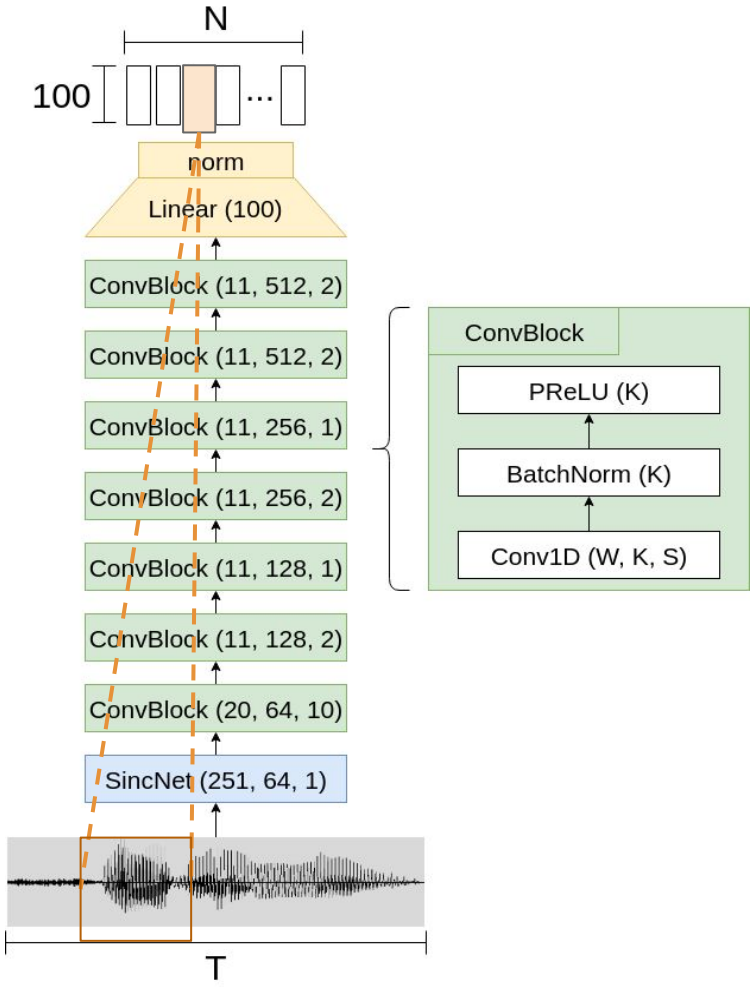


Architecture

Designing a specialized front-end for speech.

Deep structure total stride:
TS = 10x2x1x2x1x2x2 = 160 = 10ms@16kHz

Similar to typical STFT processing pipelines.

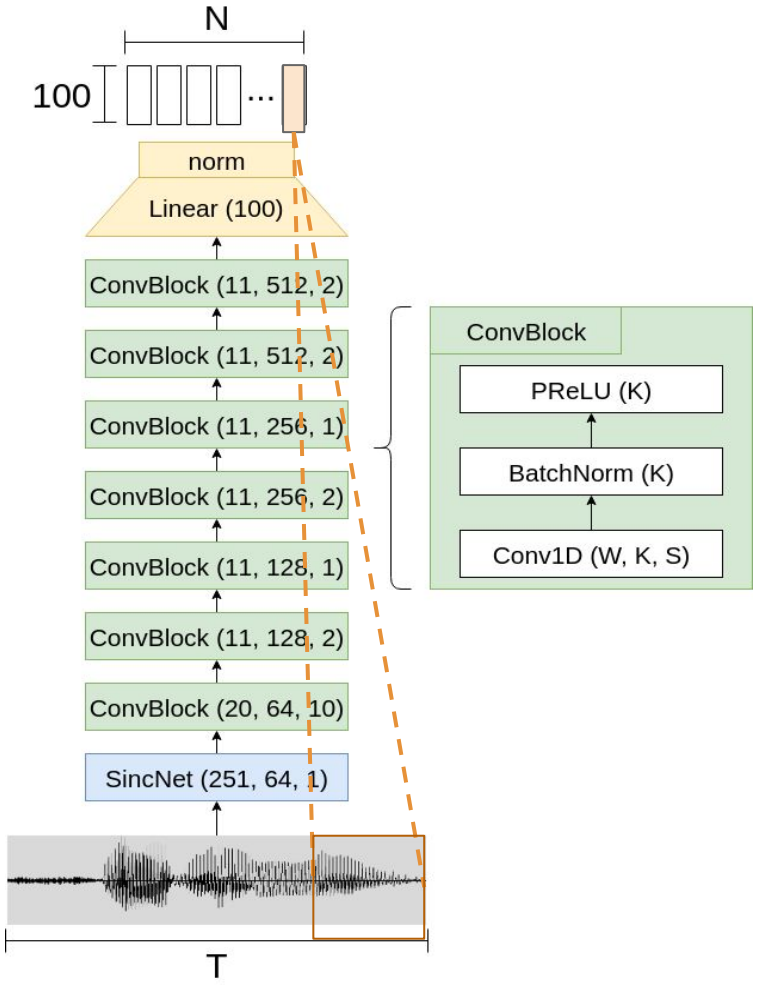


Architecture

Designing a specialized front-end for speech.

Deep structure total stride:
TS = 10x2x1x2x1x2x2 = 160 = 10ms@16kHz

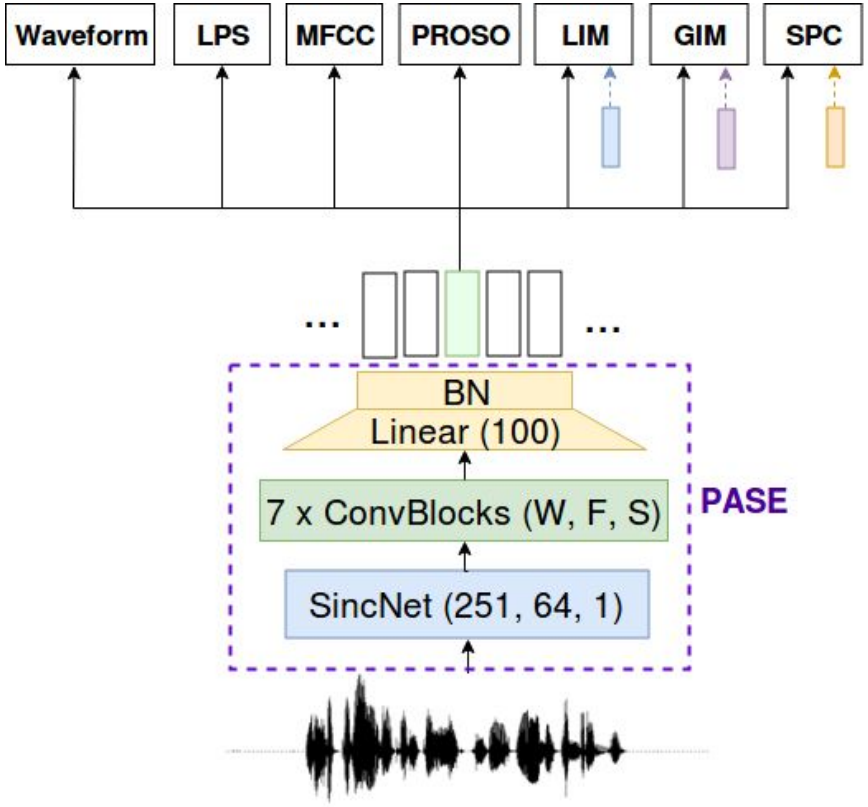
Similar to typical STFT processing pipelines.



Worker Learning

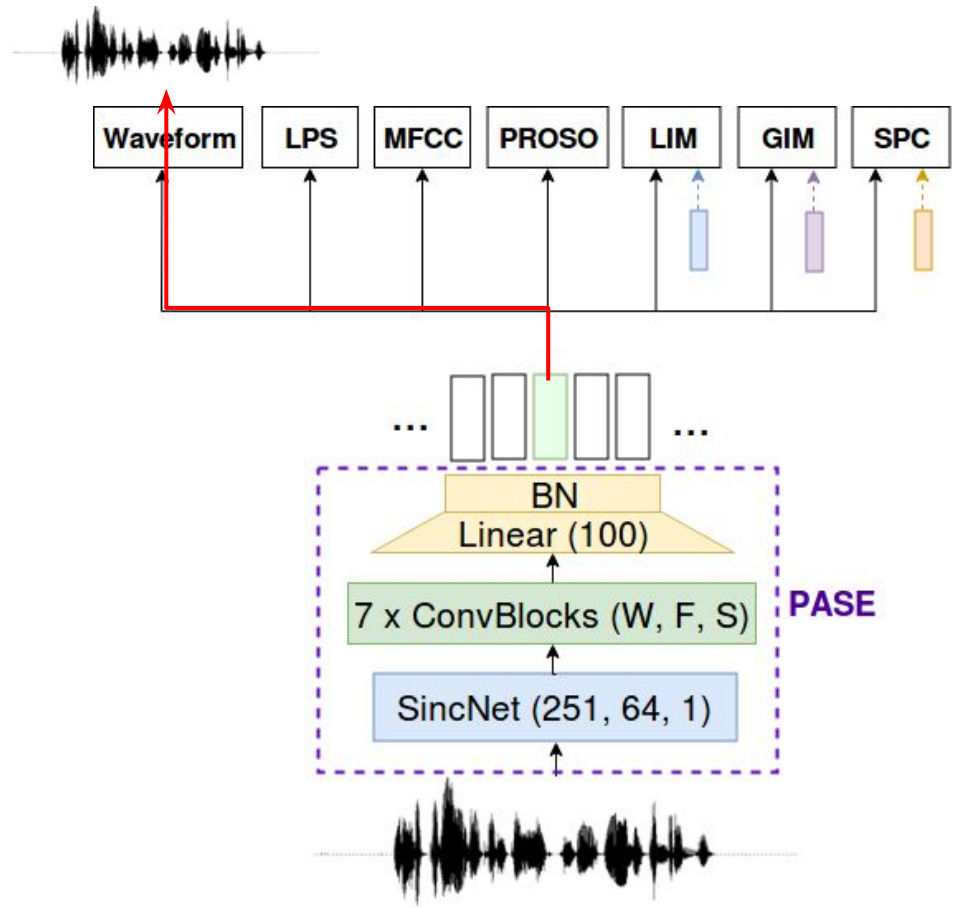
A consensus of workers with different tasks:

- One simplistic worker per task.
- Same input per worker, the PASE embeddings.
- Sum of worker losses.

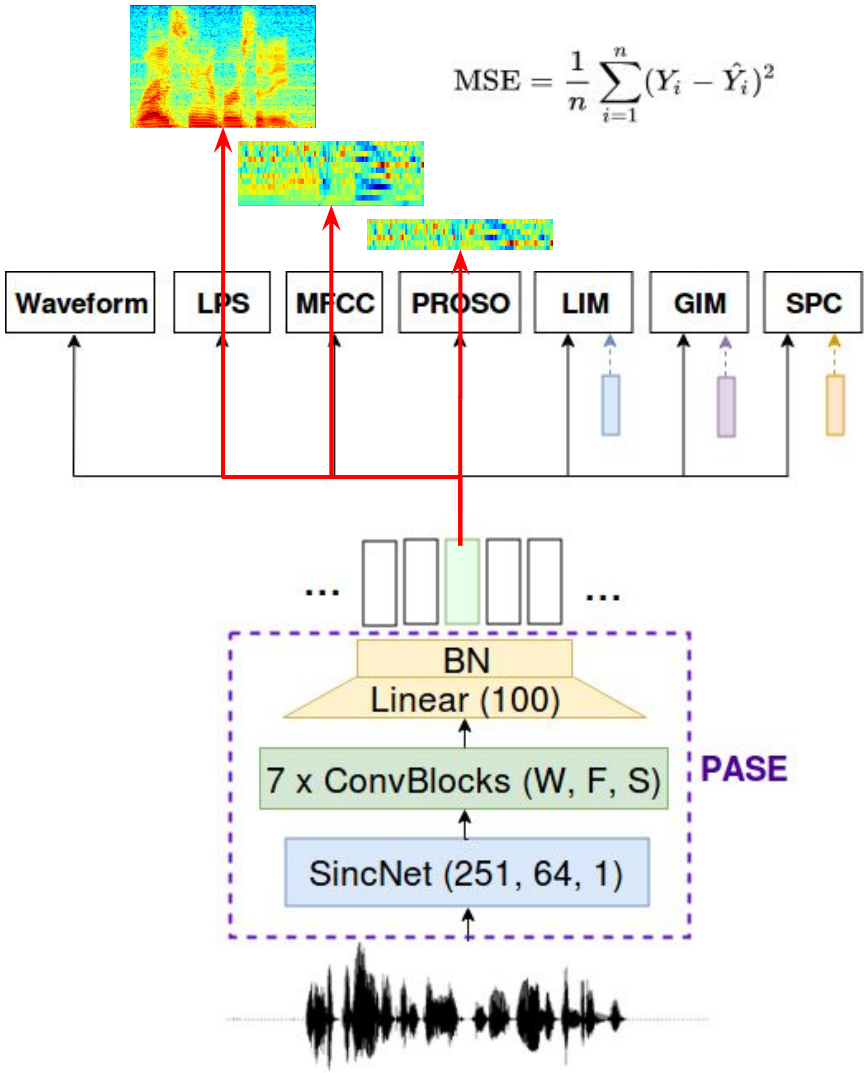


Workers: Waveform

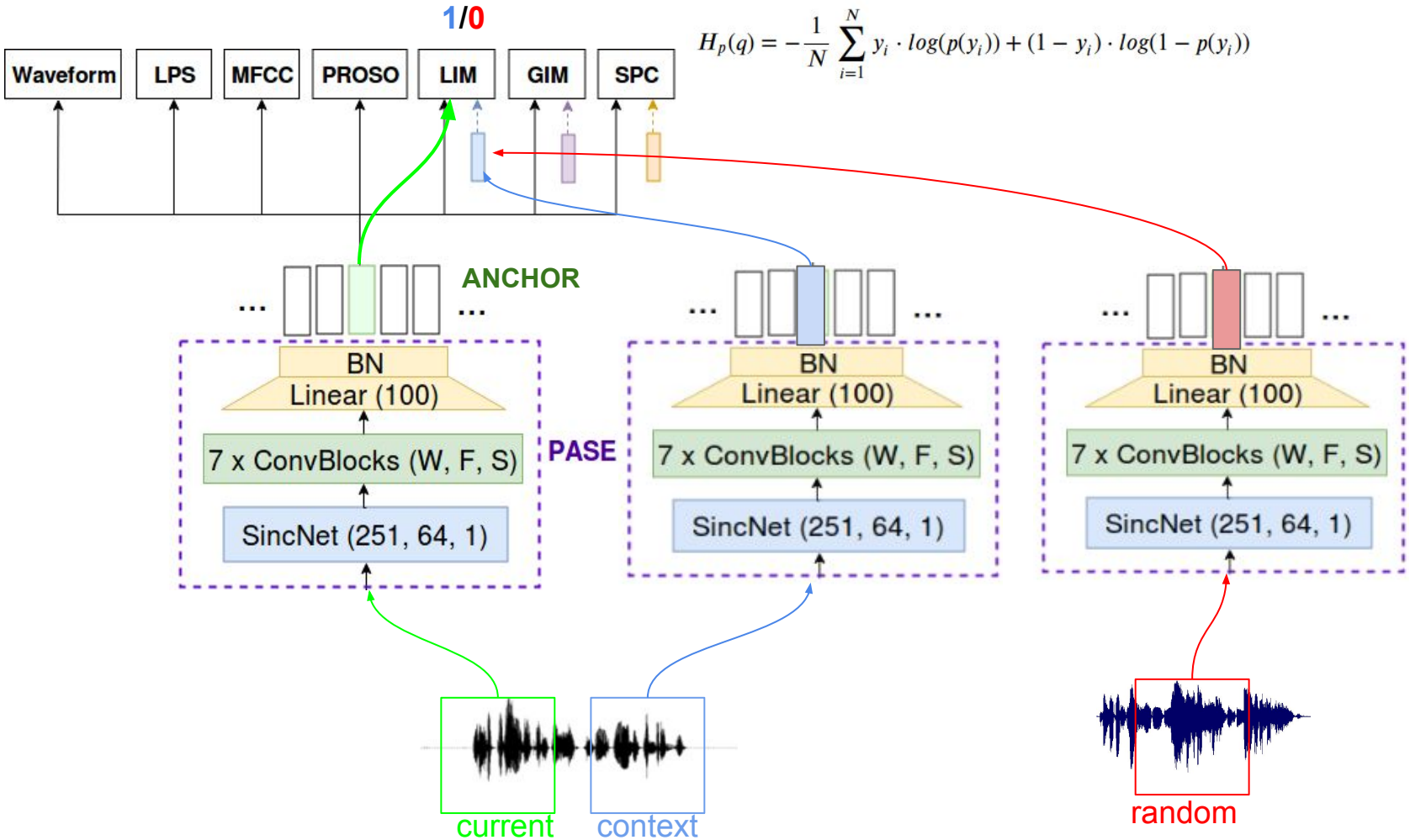
$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$



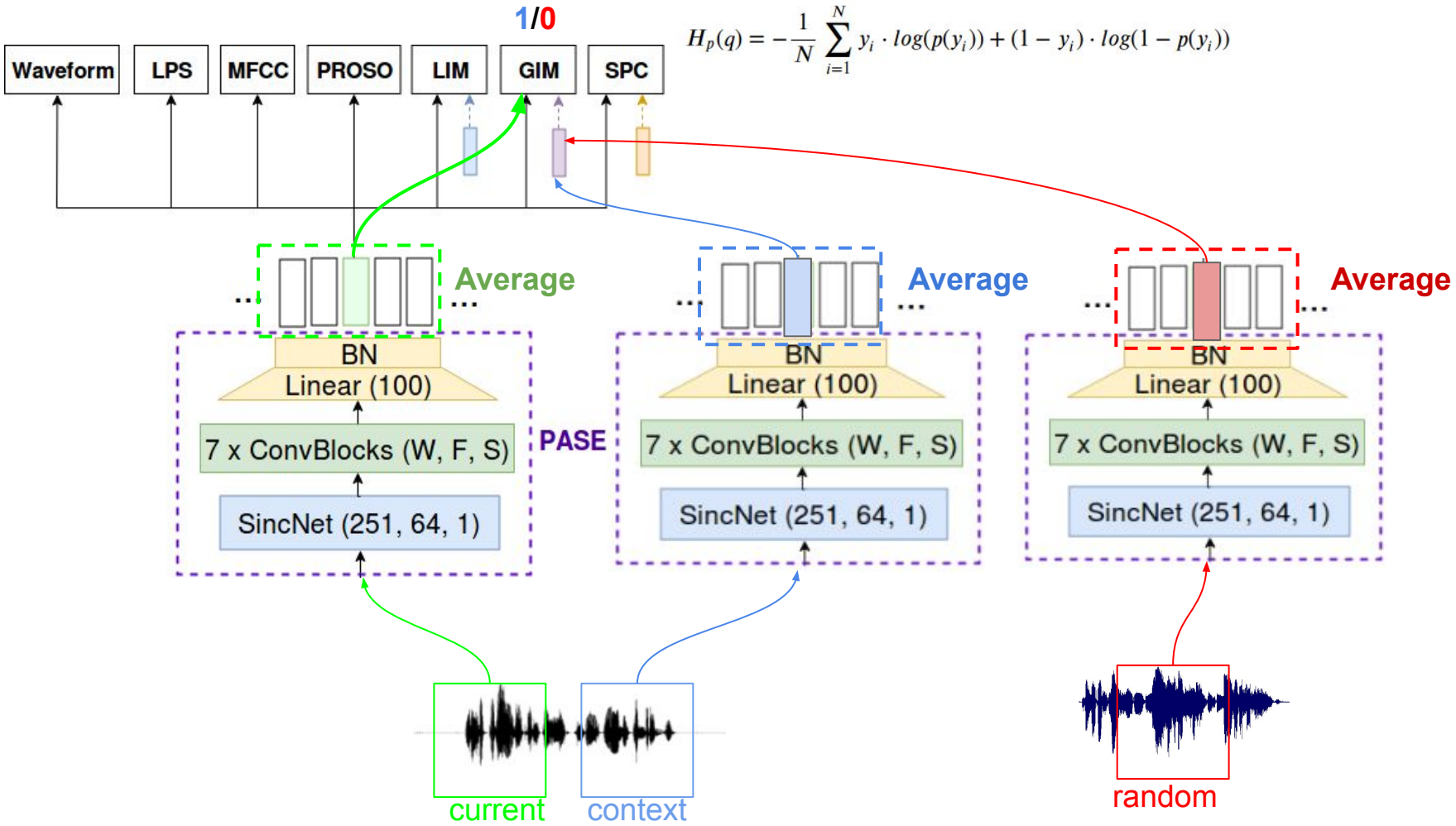
Workers: Regressors



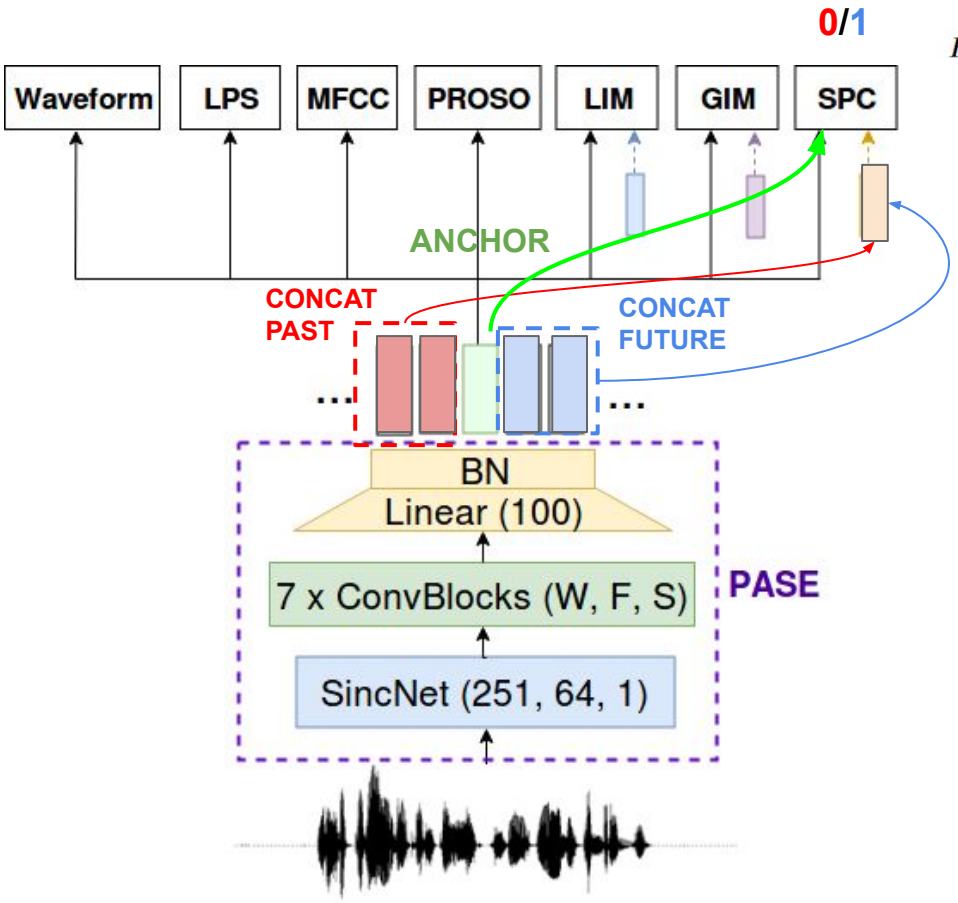
Workers: Local InfoMax



Workers: Global InfoMax

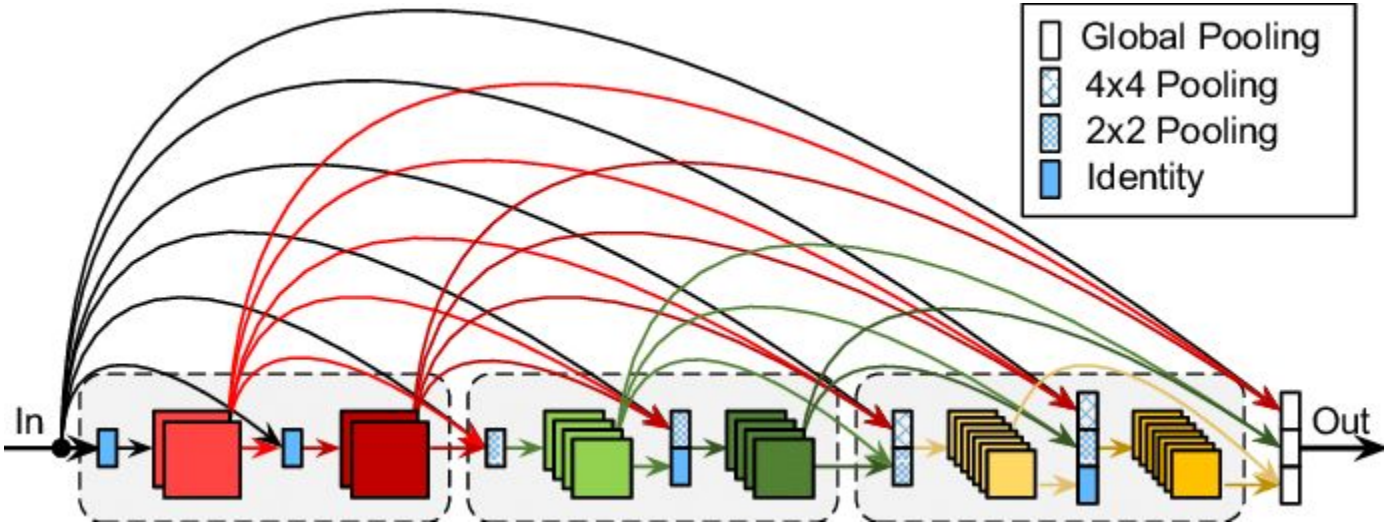


Workers: Sequence Predictive Code



$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Workers Design



Workers Design

Evaluation Methodology

Evaluation Methodology

Training Data (Self-supervised)

LibriSpeech 10 hours (12 sec. x 2484 speakers)

Test Tasks (Supervised)

1. Speaker recognition
2. Speech recognition
3. Emotion recognition

Evaluation Metrics

Classification accuracy (%)

Word error rate (%)

Evaluation Methodology

Training Data (Self-supervised)

LibriSpeech 10 hours (12 sec. x 2484 speakers)

Test Tasks (Supervised)

1. Speaker recognition: VCTK
2. Speech recognition: TIMIT (clean) and DIRHA (noisy)
3. Emotion recognition: Interface

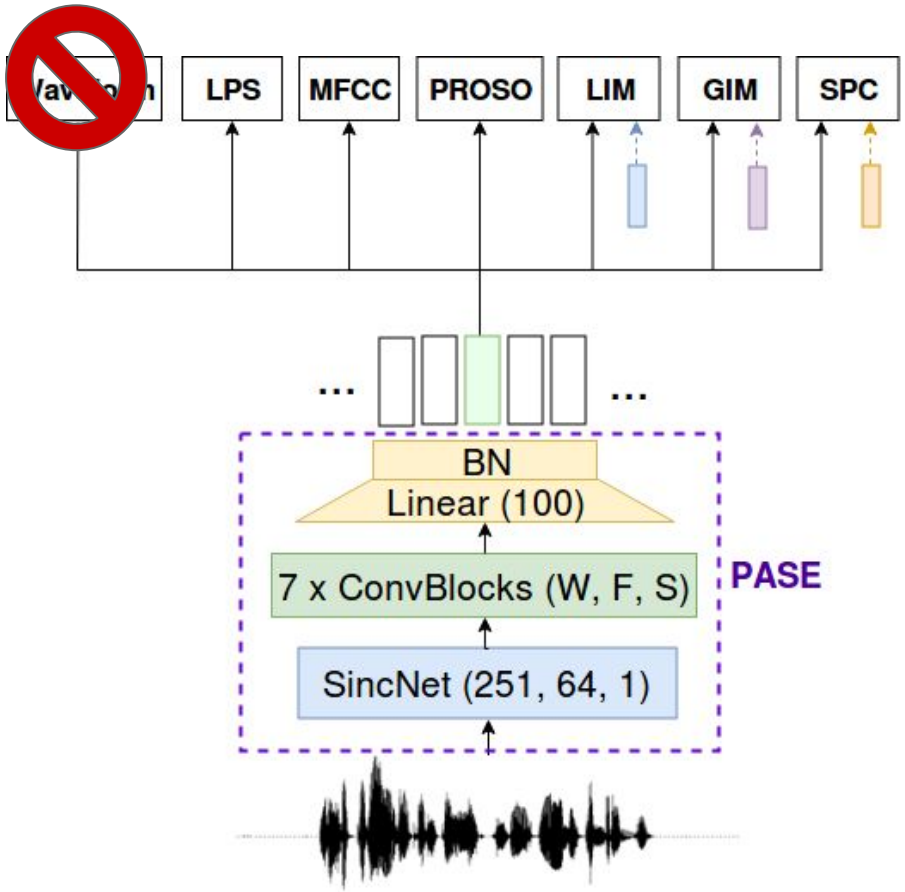
Evaluation Metrics

Classification accuracy (%)

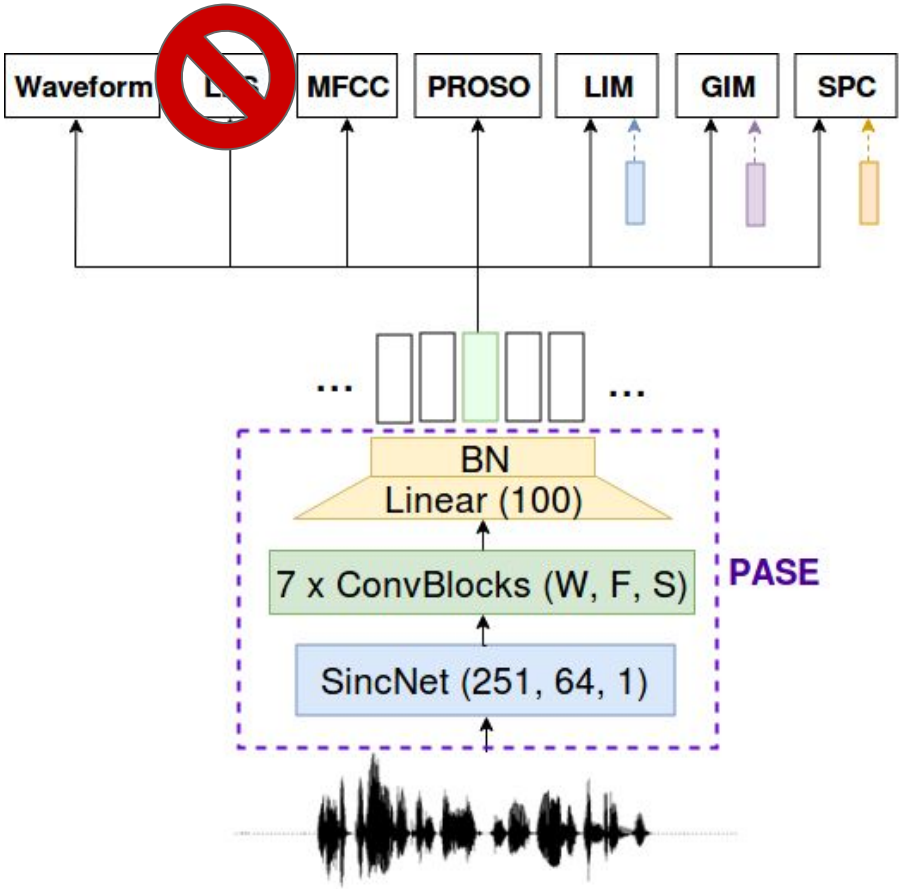
Word error rate (%)

Results

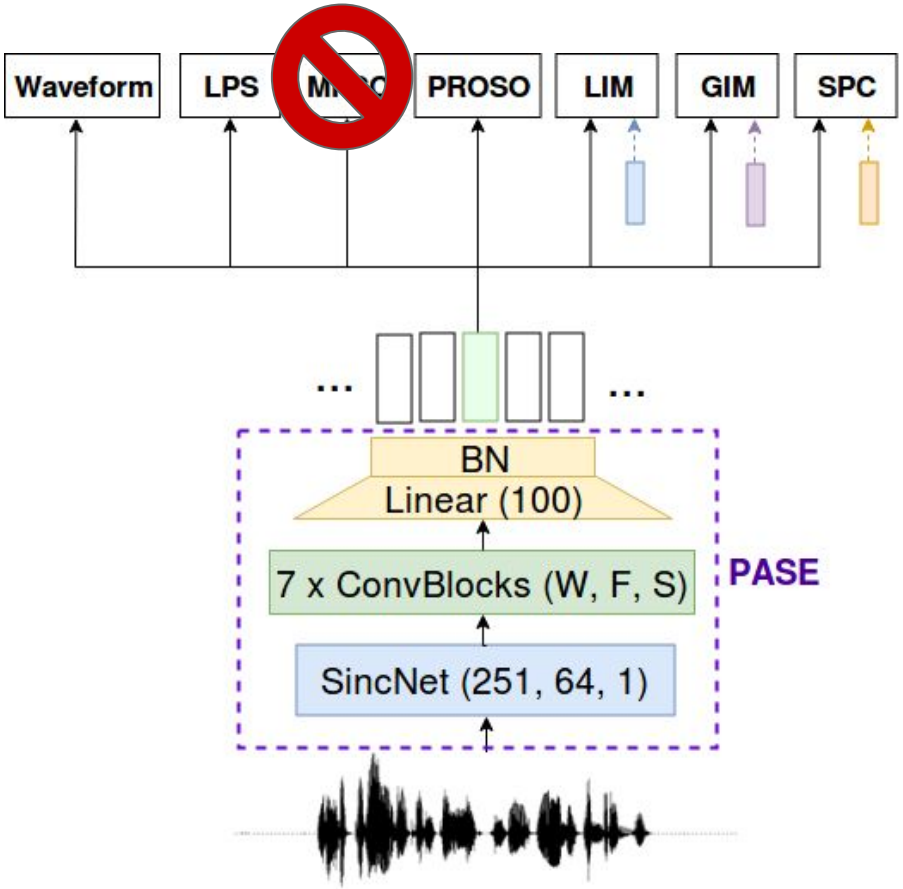
Ablation Study



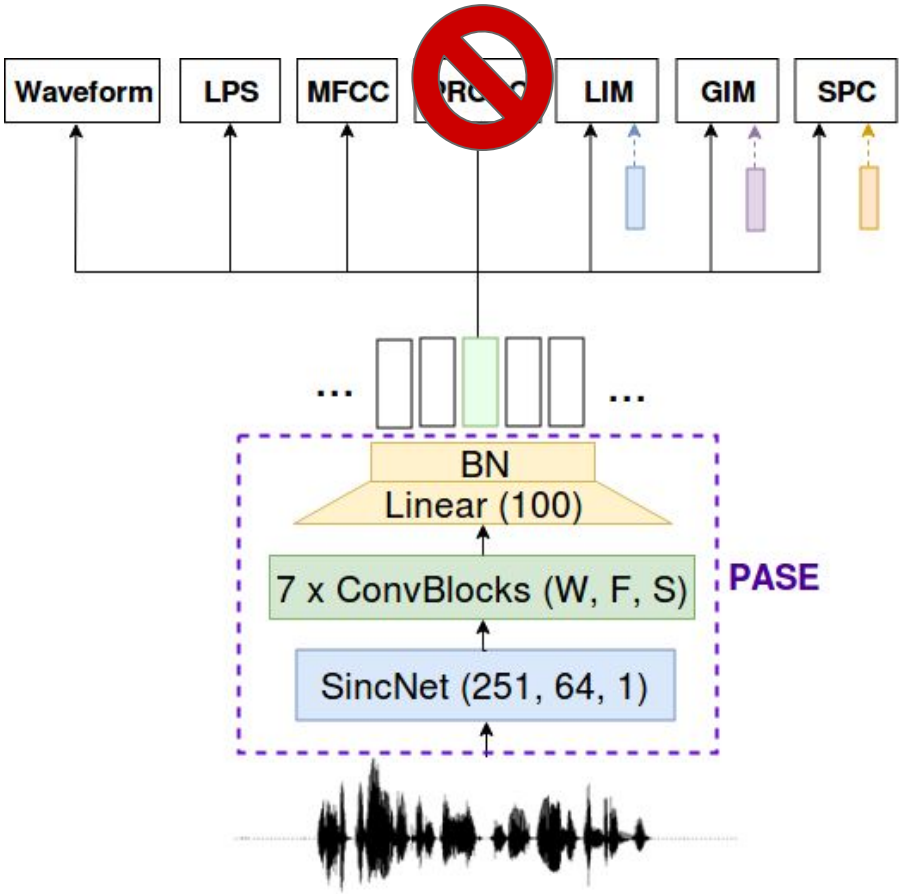
Ablation Study



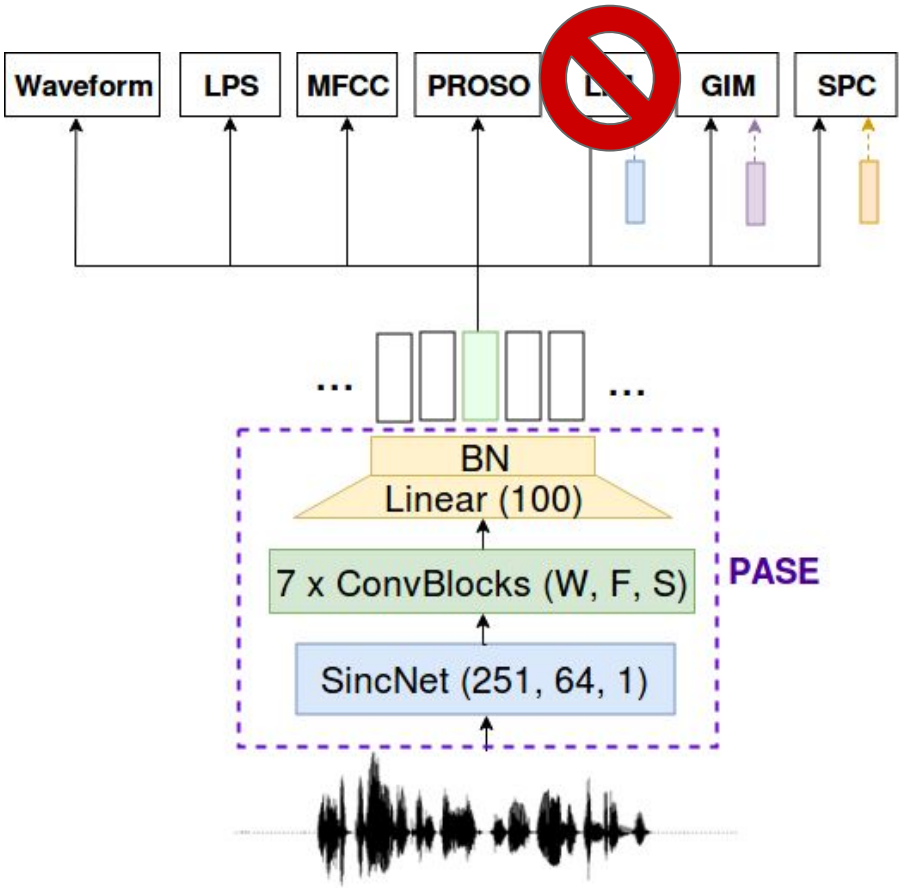
Ablation Study



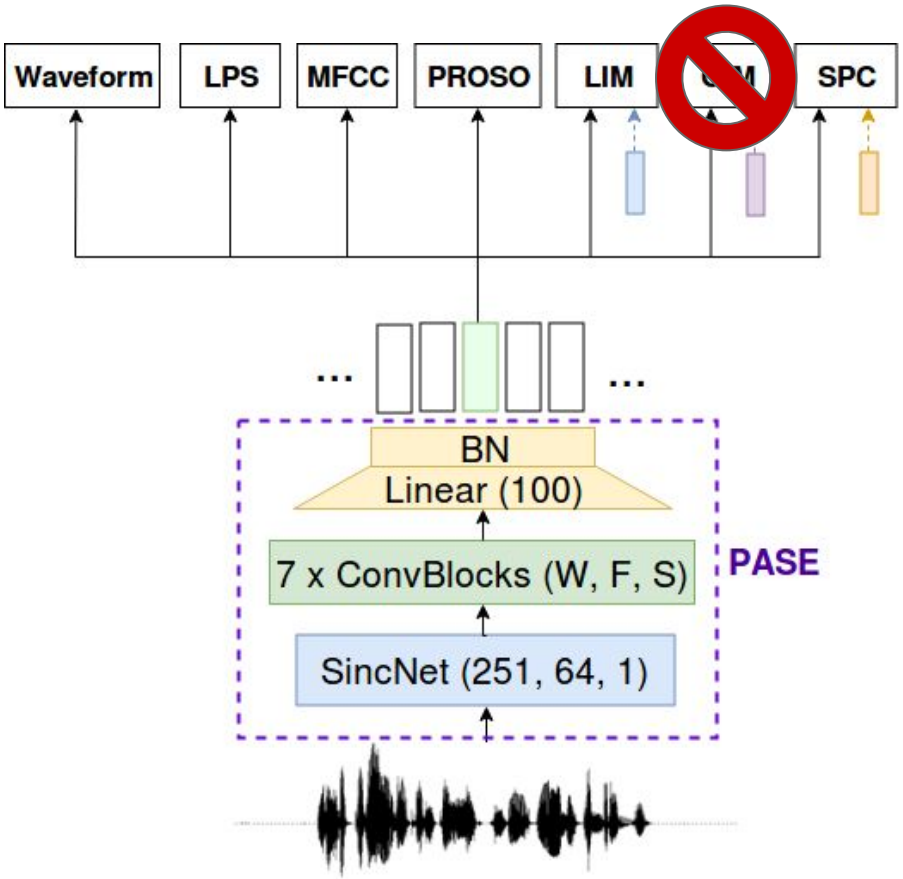
Ablation Study



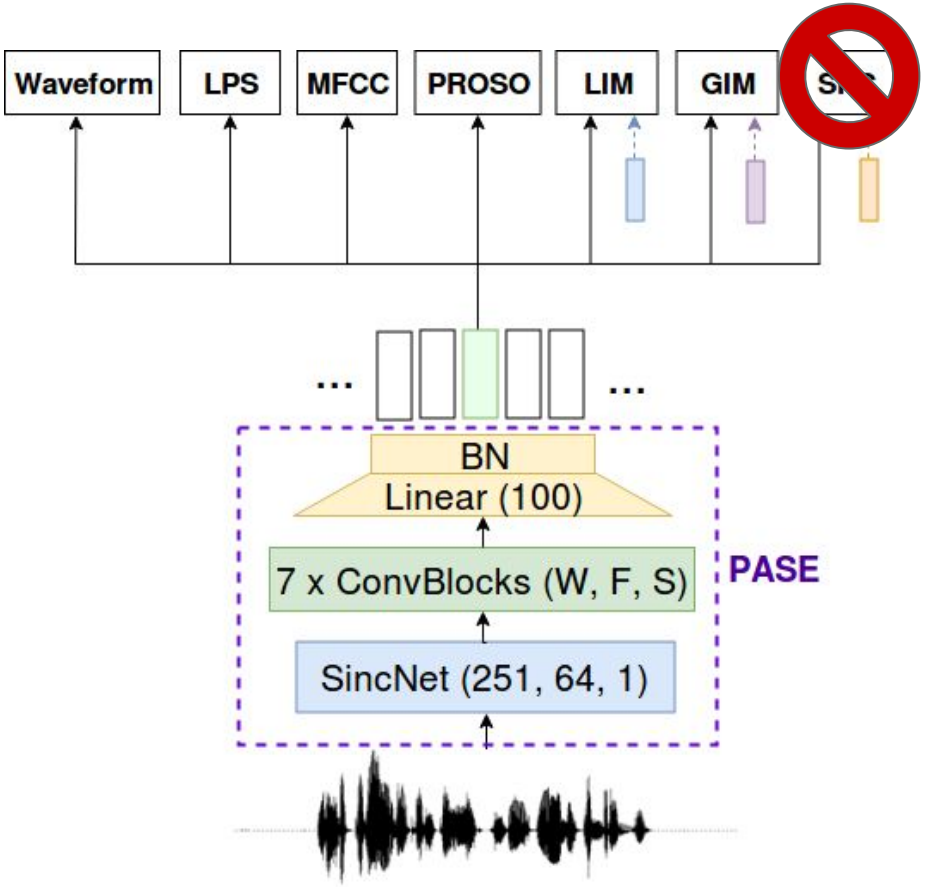
Ablation Study



Ablation Study



Ablation Study



Ablation Study

We always lose performance throughout the supervised tasks when a minion was removed during self-supervision, or at most we lose nothing (not detrimental either).

Model	Classification accuracy [%]		
	Speaker-ID *(VCTK)*	Emotion *(INTERFACE)*	ASR *(TIMIT)*
PASE (All workers)	97.5	88.3	81.1
– Waveform	−1.3	−3.9	−0.3
– LPS	−1.5	−5.3	−0.5
– MFCC	−2.4	−3.2	−0.7
– Prosody	−0.5	−5.3	−0.1
– LIM	−0.8	−1.3	−0.0
– GIM	−0.6	−0.5	−0.3
– SPC	−0.4	−1.6	−0.0

Transferred Classification Accuracy

Model	Classification accuracy [%]					
	Speaker-ID *(VCTK)*		Emotion *(INTERFACE)*		ASR *(TIMIT)*	
	MLP	RNN	MLP	RNN	MLP	RNN
MFCC	96.9	72.3	90.8	91.1	81.1	84.8
FBANK	98.4	75.1	94.1	92.8	80.9	85.1
PASE-Supervised	97.0	80.5	93.8	92.8	82.1	84.7
PASE-Frozen	97.3	82.5	91.5	92.8	81.4	84.7
PASE-FineTuned	**99.3**	**97.2**	**97.7**	**97.0**	**82.9**	**85.3**

Word Error Rate

Word error rate (WER) obtained on the DIRHA corpus.

	WER [%]
MFCC	35.8
FBANK	34.0
PASE-Supervised	33.5
PASE-Frozen	32.5
PASE-FineTuned	**29.8**

Word Error Rate

Word error rate (WER) obtained on the DIRHA corpus.

	WER [%]
MFCC	35.8
FBANK	34.0
PASE-Supervised	33.5
PASE-Frozen	32.5
PASE-FineTuned	**29.8**

Conclusions

Conclusions

- Multi-task self-supervised approach to learn speech representations.
- Effective and exportable speech encoder that converts waveforms into latent embeddings.
- The learned embeddings carry useful speech information, related to (at least): speaker ID, phonemes, and emotional cues.
- Embeddings also show their potential for transferability to new datasets, tasks, and acoustic conditions.

Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks

Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, Yoshua Bengio

santi.pascual@upc.edu

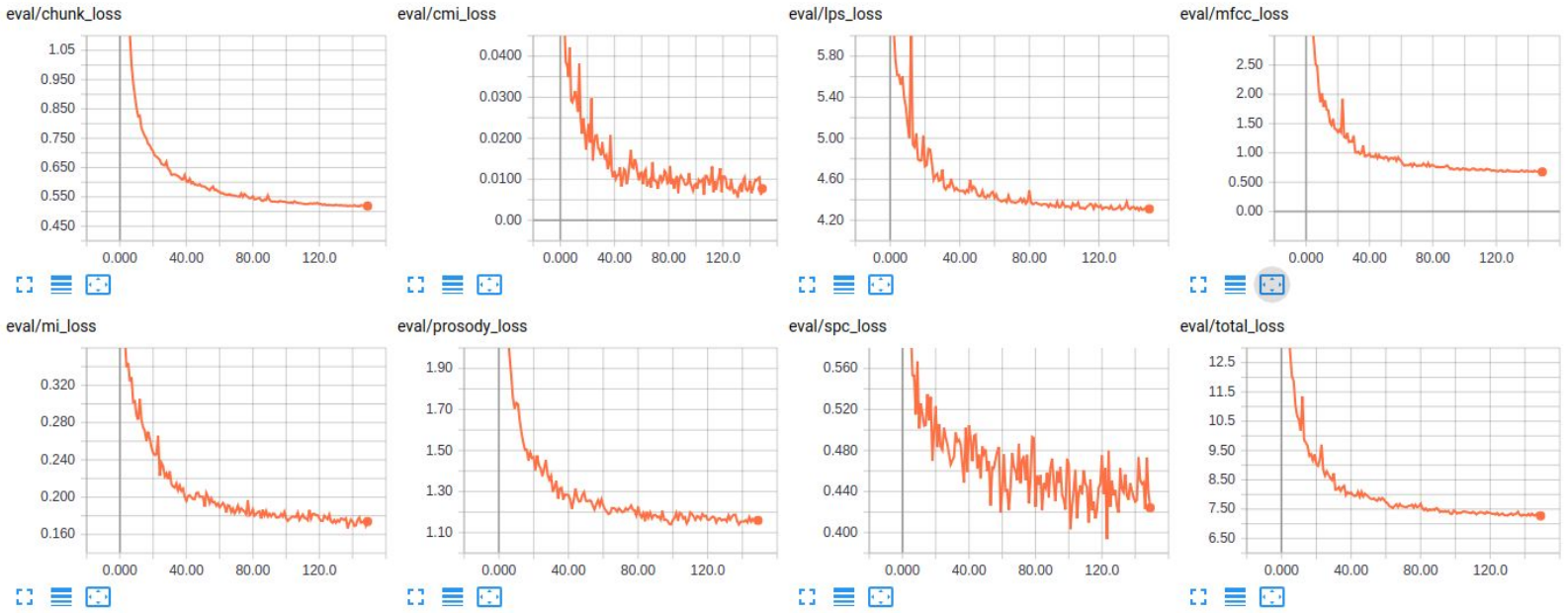
Universitat Politècnica de Catalunya
Telefónica Research
MILA - Université de Montréal

Post-Credits

Summation convergence on validation

We can observe the convergence trend throughout all workers with a validation set.

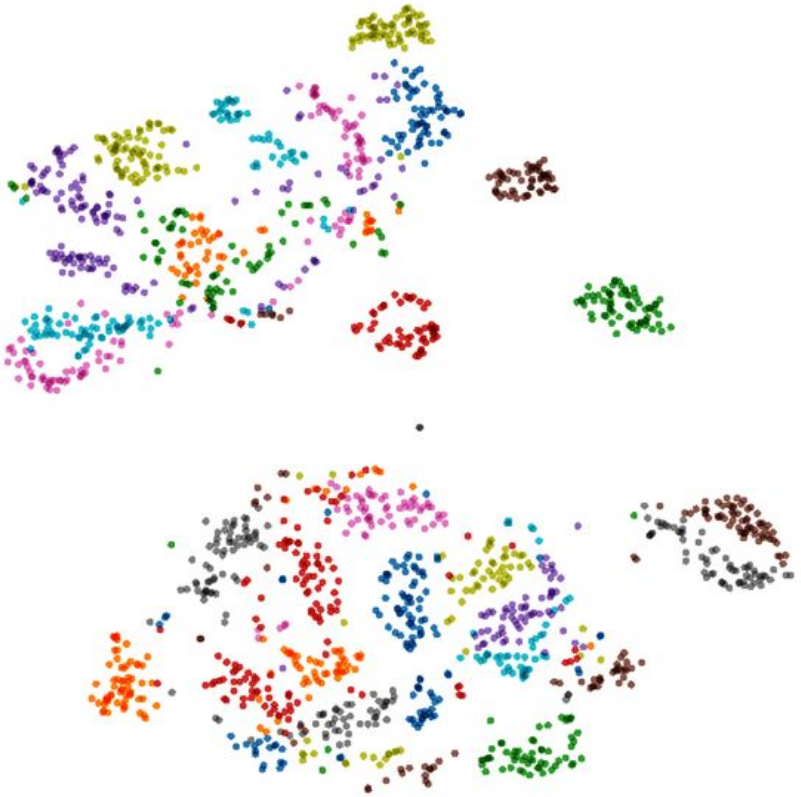
NOTE: Each output is normalized to be zero-mean unit-variance.



Global PASE-QRNN Feature: SpkID

Experiment to visualize the meaning of "utterance average PASE vector":

- Take 30 speakers from VCTK Dataset and extract PASE features from it.
- Apply t-SNE projection.



Observations:

1. We can observe clusters of speaker identities.
2. We can observe two big groups of clusters: genders.

Global PASE-QRNN Feature: SpkID

Experiment to visualize the meaning of "utterance average PASE vector":

- Take 30 speakers from VCTK Dataset and extract PASE features from it.
- Apply t-SNE projection.

Observations:

1. We can observe clusters of speaker identities.
2. We can observe two big groups of clusters: genders.

Global PASE-QRNN Feature: SpkID

Experiment to visualize the meaning of "utterance average PASE vector":

- Take 30 speakers from VCTK Dataset and extract PASE features from it.
- Apply t-SNE projection.

Observations:

1. We can observe clusters of speaker identities.
2. We can observe two big groups of clusters: genders.

Global PASE-QRNN Feature: SpkID

Experiment to visualize the meaning of "utterance average PASE vector":

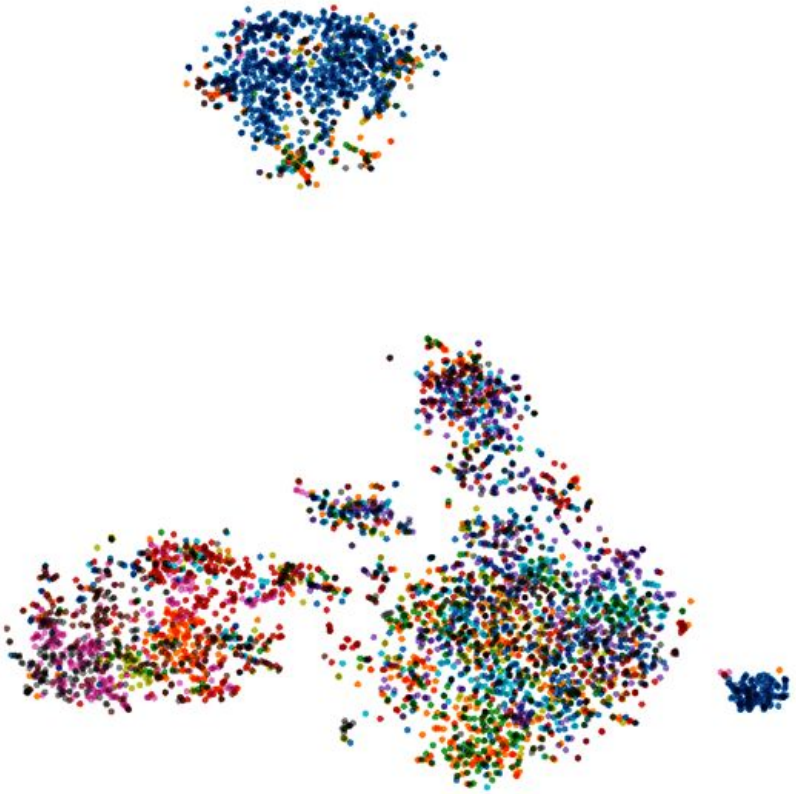
- Take 30 speakers from VCTK Dataset and extract PASE features from it.
- Apply t-SNE projection.

Observations:

1. We can observe clusters of speaker identities.
2. We can observe two big groups of clusters: genders.

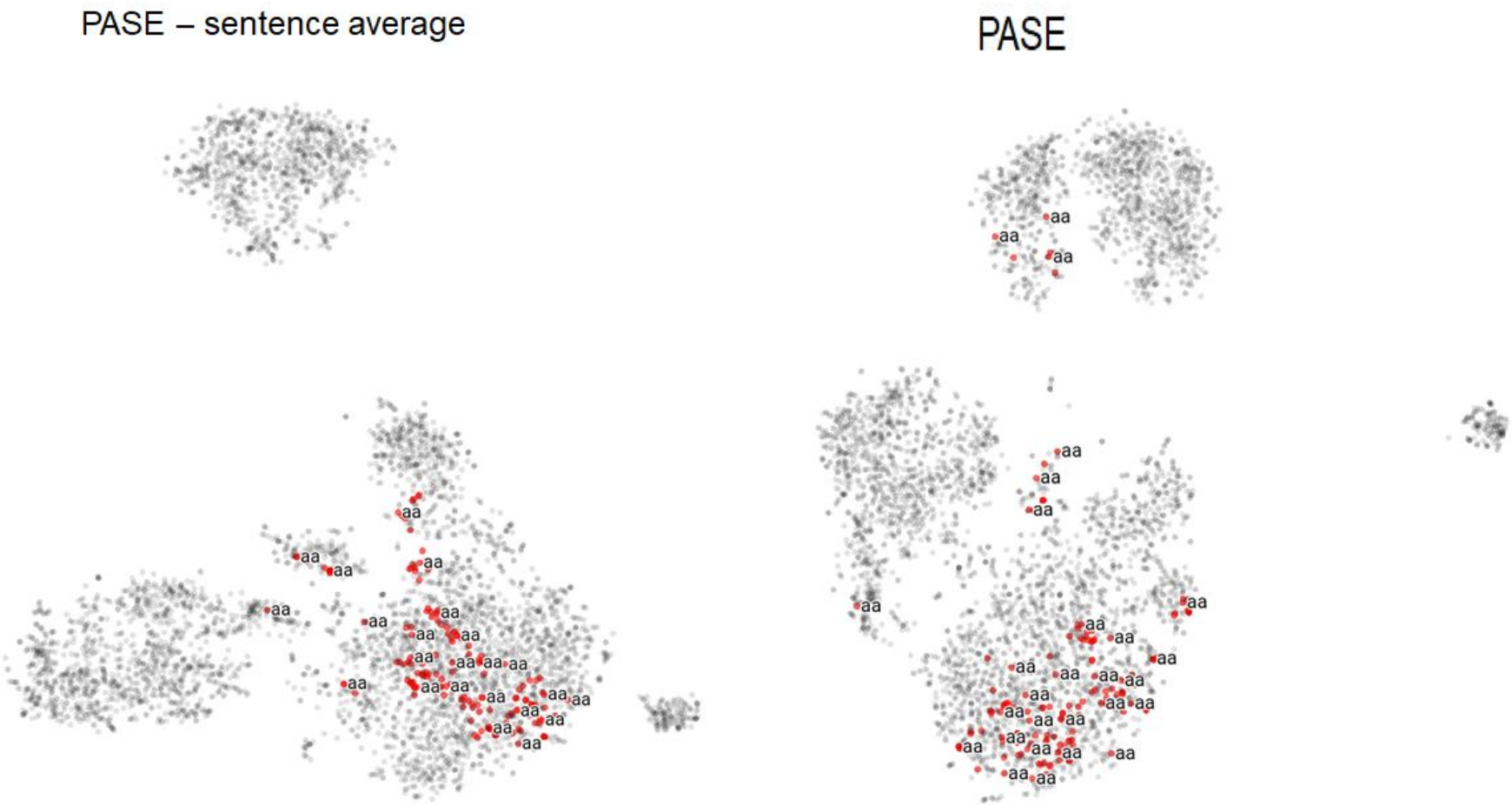
Local PASE-QRNN Feature: Phonemes

PASE – sentence average

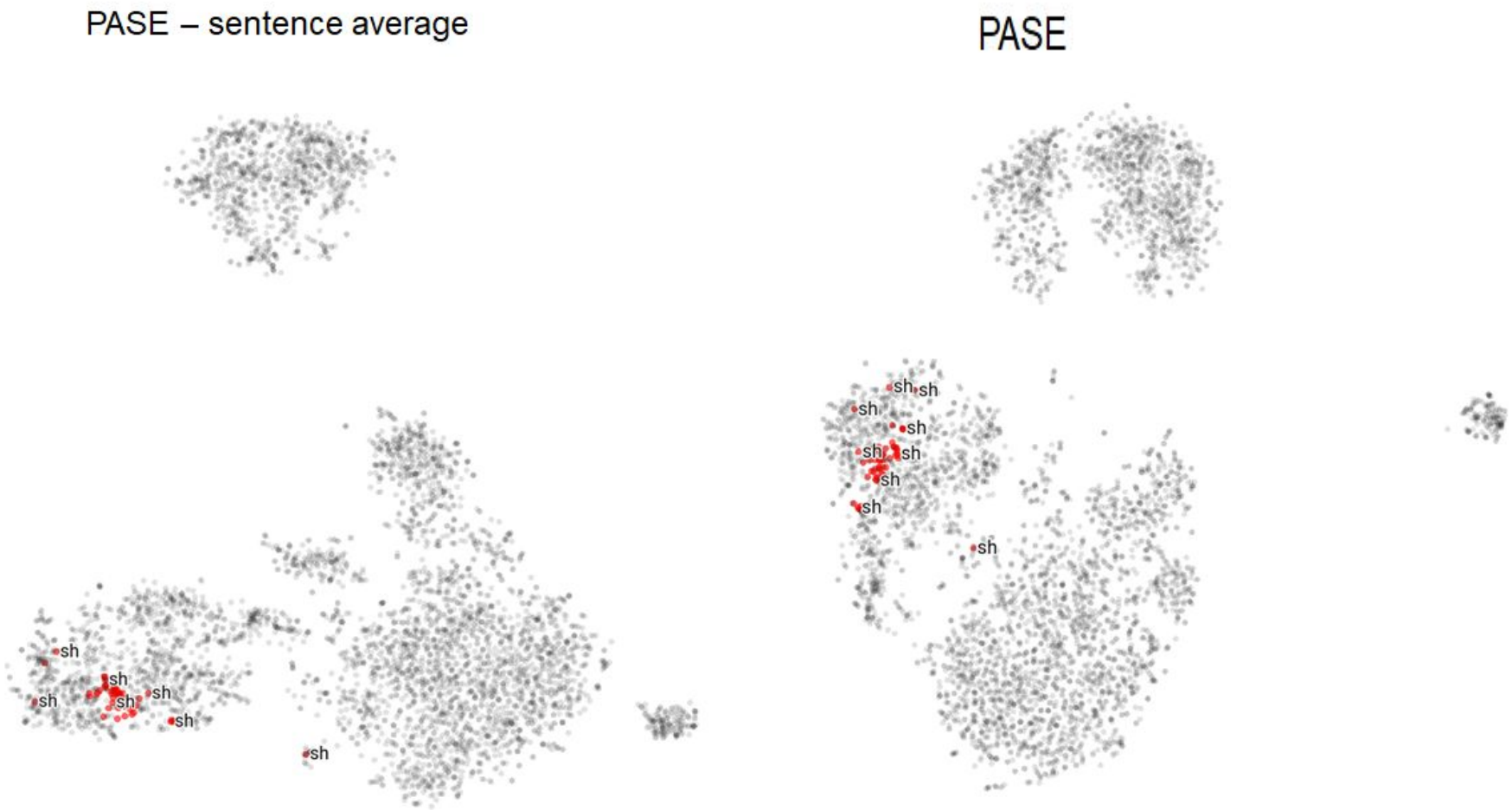
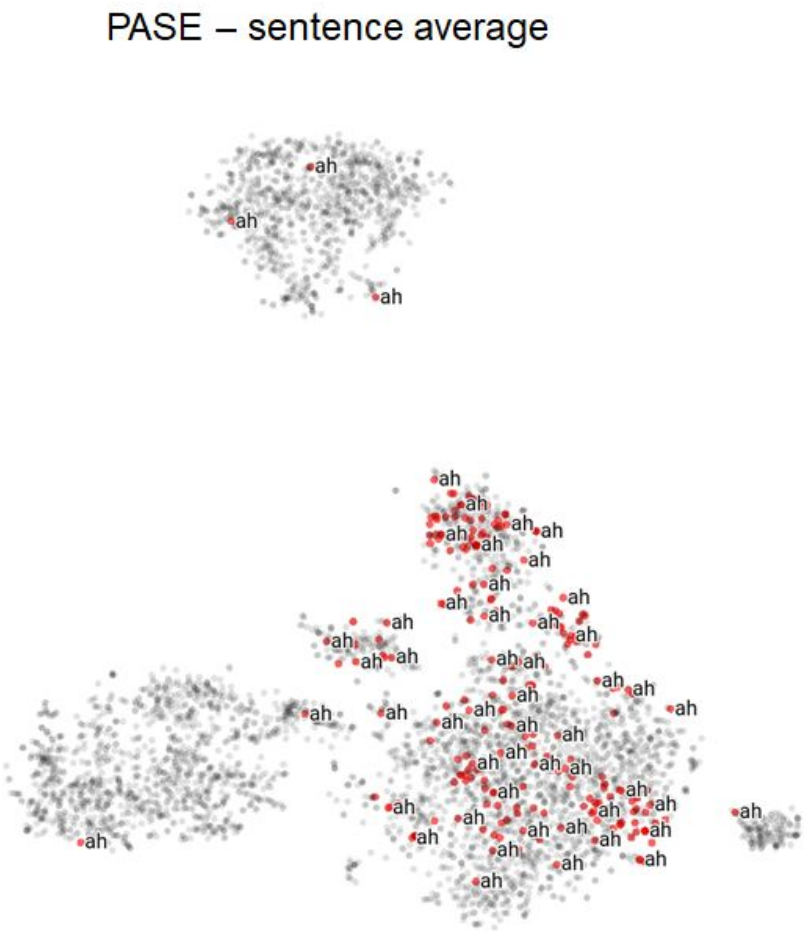
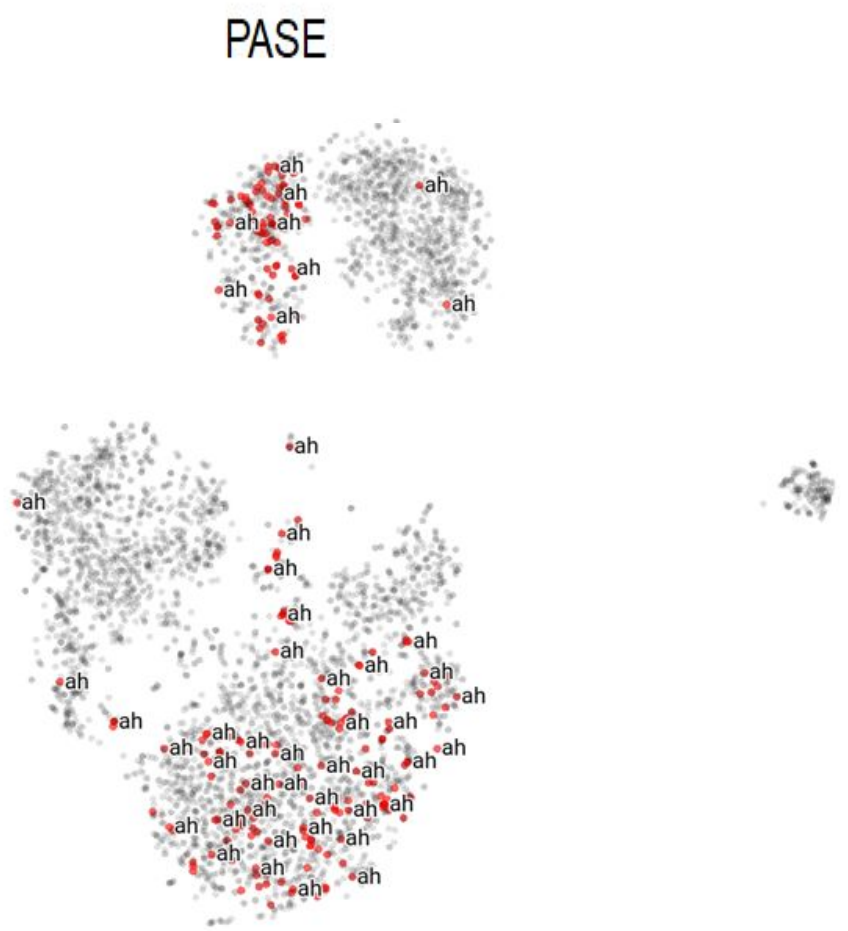


PASE

Local PASE-QRNN Feature: Phonemes



Local PASE-QRNN Feature: Phonemes



Local PASE-QRNN Feature: Phonemes