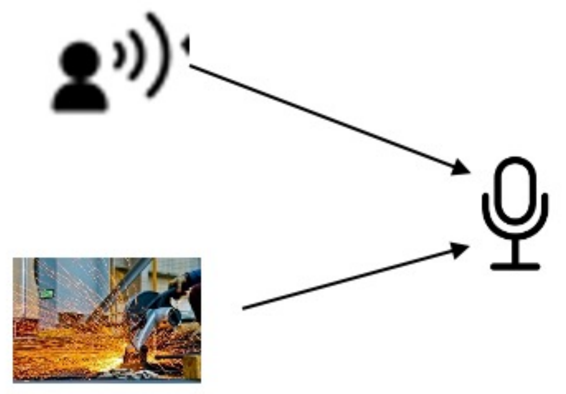
Joint Learning with Shared Latent Space for Self-Supervised Monaural Speech Enhancement

Yi Li, Yang Sun, Wenwu Wang, Syed Mohsen Naqvi

Monaural Speech Enhancement



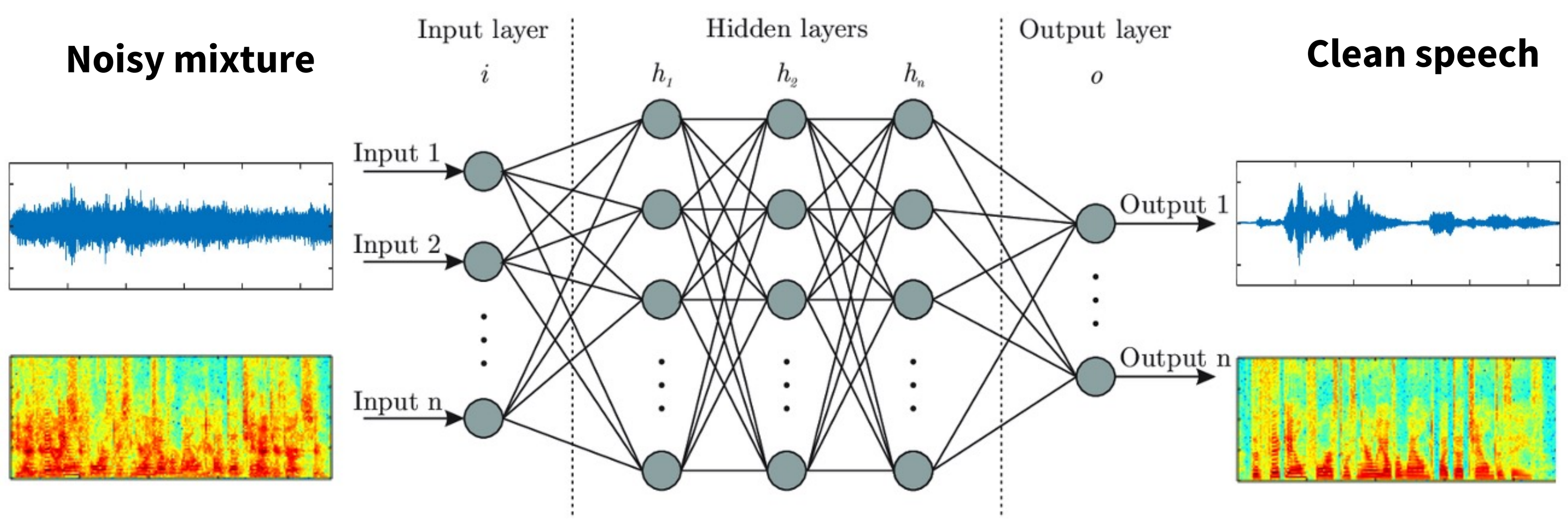
Noisy mixture includes:
- Speech signal
- Background noise
- Potential reverberations if in rooms

Monaural speech enhancement, aiming to improve the quality of the desired speech from noisy mixture recorded by using a single microphone, is a crucial topic of audio signal processing.

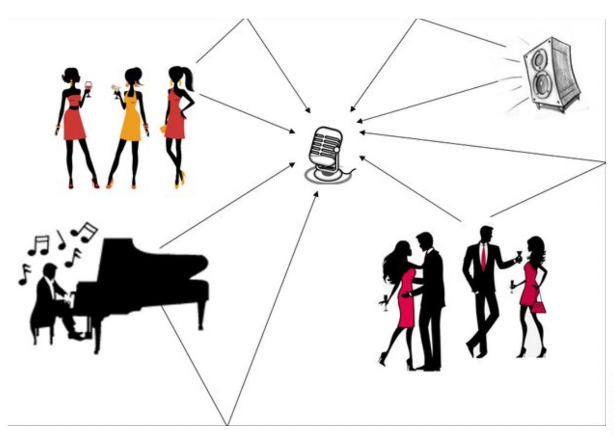
Applications:
- Hearing aids
- Robotics
- Automatic speech recognition (ASR)
- VoIP
- Speaker diarization
- Teleconferencing
- AI assistant

Deep Learning-Based Speech Enhancement

- Recent studies aim to extract the clean speech from noisy mixture by using deep learning-based techniques.



- **Input** of the neural network is noisy mixture signals or spectra depending on the backbone type.
- Network types: deep neural network (DNN), convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM).
- **Learning strategies: Supervised**, unsupervised, semi-supervised, self-supervised
- In recent studies, most of monaural speech enhancement is based on supervised learning setting.
- **Training targets:** Mapping, masking, signal processing.



Cocktail part problem

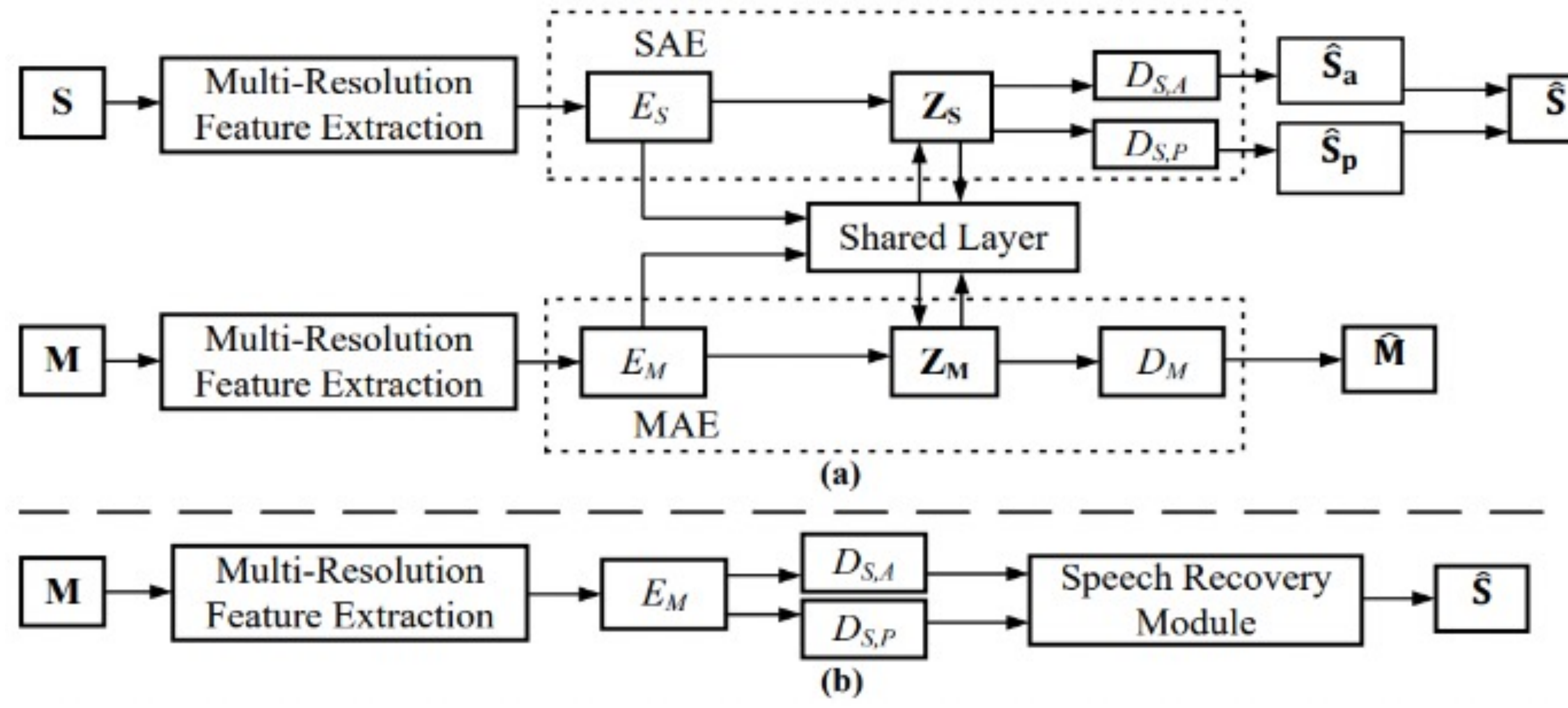
Self-Supervised Learning

What is self-supervised learning (SSL):
- Unlabeled data is processed to obtain useful representations that can help with downstream learning tasks.
- An intermediate form of unsupervised and supervised learning.

Why we need SSL-based monaural speech enhancement?
- Supervised training of the networks requires large sets of labelled paired data. However, these data is difficult or expensive to obtain.
- A trained model may suffer from performance degradation when deployed in previously unseen conditions e.g., a mismatch of room environments between the training and testing datasets.

What do we propose in this work?
- We propose the SAE with two independent decoders to learn the latent representations of both amplitude- and phase-related features.
- We jointly learn a shared latent space between the SAE and the MAE to boost the generalization ability.
- The multi-resolution spectral losses are introduced in the proposed phase-aware SSL enhancement method to further improve the speech enhancement performance.

Framework



- (a) Training; (b) Test
- S = Clean speech spectra
- M = Noisy mixture spectra
- E = Encoder; D = Decoder
- A = Amplitude, P = Phase

- A shared layer from the SAE and MAE is used to obtain a joint latent space of the learned clean speech and noisy mixture representations.
- Multi-resolution feature maps: The feature map is rescaled with the same frame shift (i.e. 32), but with different window sizes (1024, 512, 256, and 128).

Experimental Settings

- Different noises and rooms (RIRs) in training and test stages
- Same dataset in training and test stages
- 140 noisy mixture signals from DAPS and NOISEX datasets for the test stage.
- SAE: 4 1-D convolutional layers.
- MAE: 6 1-D convolutional layers.
- Training datasets: DAPS, NOISEX
- 28 clean speech signals from the DAPS dataset for SAE training
- 392 noisy mixture signals from DAPS and NOISEX datasets for MAE training
- These clean speech signals and noisy mixture signals are unpaired.

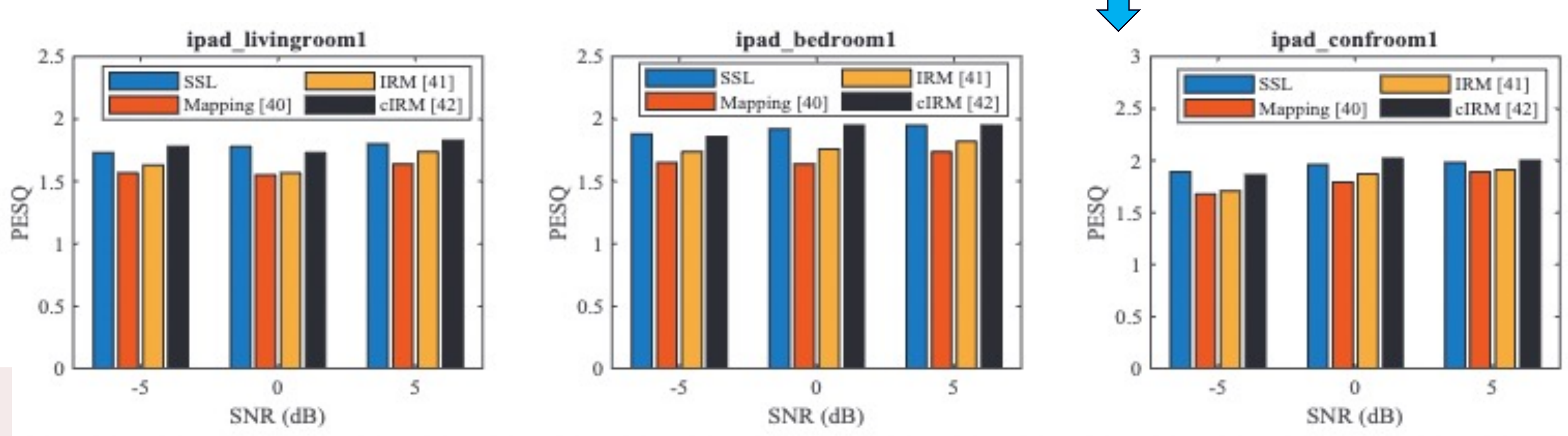
Experimental Results

Speech Enhancement Performance Comparisons to SSL methods

	PESQ			CSIG			CBAK			COVL		
	-5	0	5	-5	0	5	-5	0	5	-5	0	5
SSE	1.32	1.33	1.34	1.97	2.04	2.09	1.74	1.76	1.77	1.59	1.65	1.68
P-VQ	1.68	1.70	1.71	2.24	2.27	2.29	1.76	1.79	1.80	1.72	1.77	1.81
CF	1.71	1.74	1.77	2.29	2.30	2.35	1.80	1.80	1.96	1.76	1.80	1.86
Ours	**1.84**	**1.89**	**1.91**	**2.45**	**2.47**	**2.49**	**1.94**	**1.94**	**2.23**	**1.89**	**1.96**	**2.03**

- SSE, P-VQ, and CF are SOTA self-supervised learning-based speech enhancement algorithms.
- PESQ, CSIG, CBAK, and COVL are commonly used performance measures in speech enhancement tasks to measure the speech quality. The value range is between -0.5 – 4.5. Higher values indicate better performance.

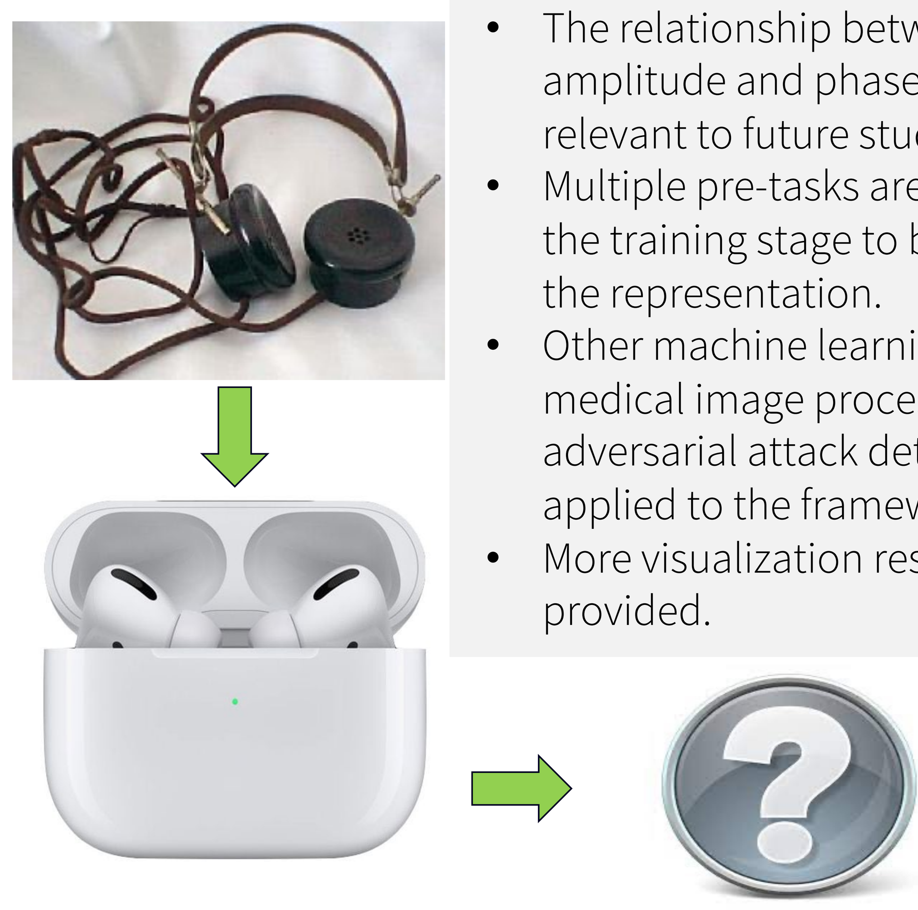
- We further compare the proposed method to supervised learning-based speech enhancement algorithms.
- IRM and cIRM are masking-based methods.
- These supervised learning-based methods suffer a significant performance drop compared to original reported results due to challenging scenarios, i.e., cross-domain setting and high reverberations.



Speech Enhancement Performance Comparisons to SL methods

Conclusions and Future Work

Conclusions:
- Monaural speech enhancement problem is addressed by using a self-supervised learning based method with the complex spectrogram.
- Multi-resolution losses are calculated from feature maps to better extract rich feature information.
- The experiment results confirm the effectiveness of the proposed method.



Future Work:
- The relationship between the amplitude and phase may be relevant to future studies.
- Multiple pre-tasks are added to the training stage to better learn the representation.
- Other machine learning tasks, e.g., medical image processing and adversarial attack detection are applied to the framework.
- More visualization results will be provided.

Lancaster University
UNIVERSITY OF OXFORD
UNIVERSITY OF SURREY
Newcastle University