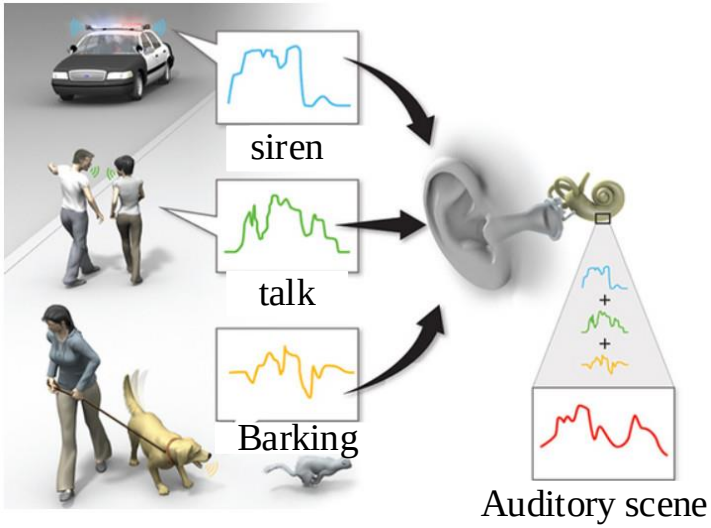
Progressive Learning for Speech Enhancement Based on Nonnegative Matrix Factorization and Deep Neural Network

Wenbo Wang

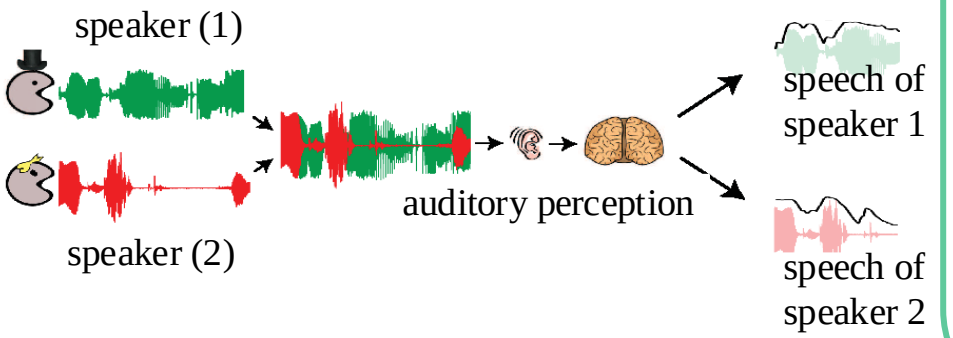
School of Mechatronic Engineering
China University of Mining and Technology

Clarity 2021

Problem Statement



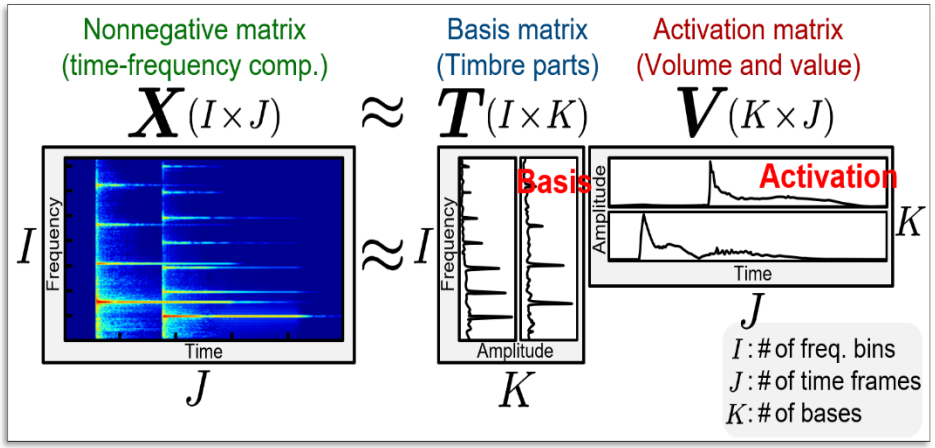
Hearing is a process by which **sound waves** in the air are processed and converted to electrical signal in our ear, then sent to our brain for **sound interpretation**.



However, patients who suffer from sensorineural hearing loss, have less ability to **separate desired from undesired sound**, which leads to a difficulty in hearing even with hearing aid, especially in noisy environments.

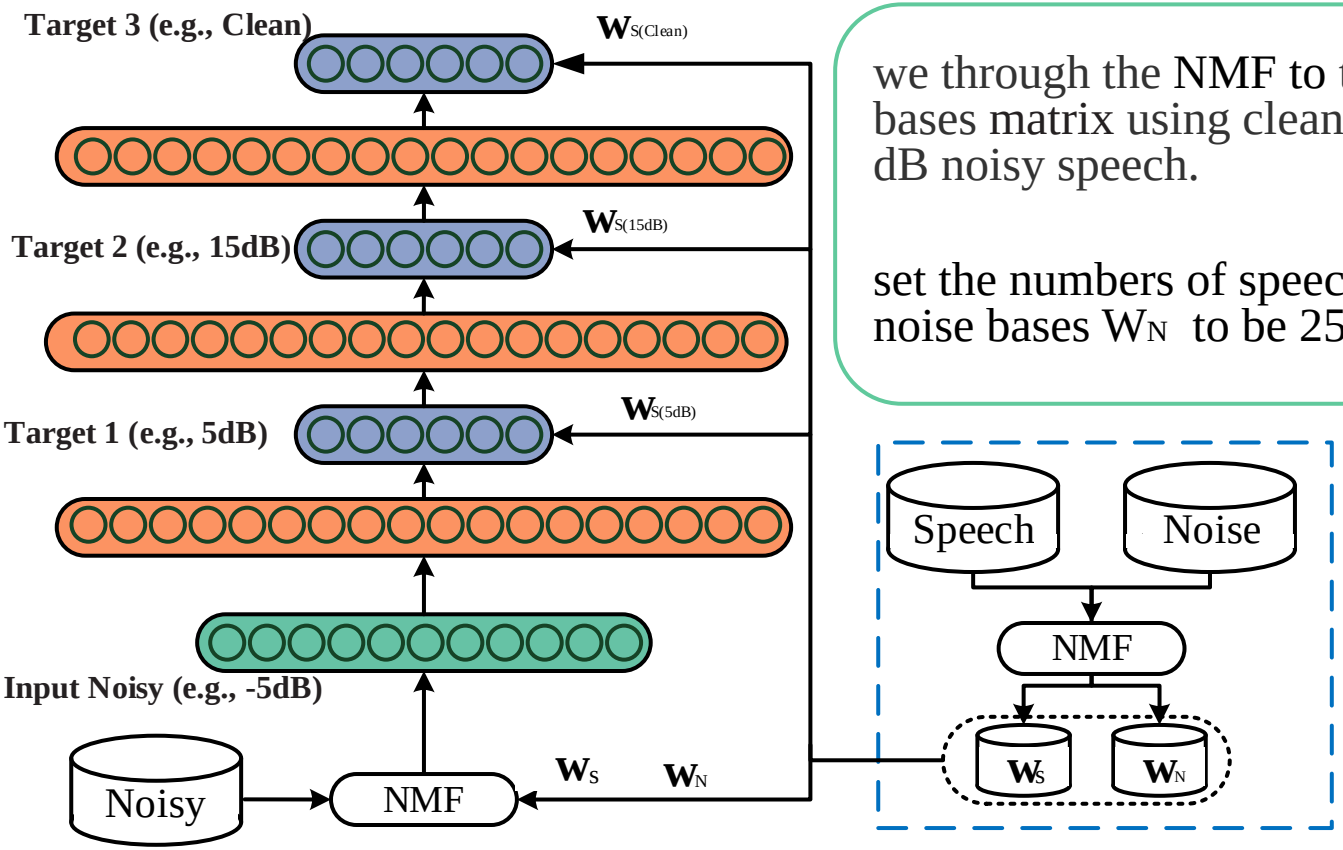
Method introduction

Nonnegative matrix factorization (NMF) is a well-known representation learning technique that is capable of capturing the basic spectral structures. The NMF and the Deep Neural Network techniques have many realistic applications and attract many attentions of speech enhancement community. Therefore, the **combination of deep learning and NMF as an organic whole is a smart strategy.**



At the same time, the **progressive learning** has achieved a good effect in speech enhancement, **but it only used the spectrogram, ignoring the spectral structures of speech.** Therefore, we apply the NMF to progressive learning for speech enhancement.

Method introduction



we through the NMF to train different speech bases matrix using clean speech and 5dB and 15 dB noisy speech.

set the numbers of speech bases matrix Ws and noise bases W_N to be 256.

$$\tilde{\mathbf{y}}_s = \mathbf{B}\hat{\mathbf{a}}_s$$

$$\frac{\partial J}{\partial \hat{\mathbf{a}}_s} = \mathbf{B}^T\left(\tilde{\mathbf{y}}_s - \mathbf{y}_s\right)$$

- **B is speech** basis matrix obtained by matrix factorization.
- $\hat{\mathbf{a}}_s$ **is** activation matrix and **it is also the output of the deep neural network.**

Experiments

Our system is systematically compared to several **speech enhancement(SE) methods** :

- (**Baseline DNN**): a conventional DNN-based SE;
- (**PL DNN**): Basic progressive learning DNN-based SE;
- (**NMF-DNN**): a NMF-based SE with conventional DNN;

Datasets:

- Three noise(**factory, babble, white**) sources from the NOISEX dataset for the training and testing process.
- **1500 randomly utterances** were used from the database to construct the training data at -5, 0dB . And the other **90 utterances** were selected to construct the testing data.

Evaluation criteria

- Perceptual Evaluation of Speech Quality **(PESQ)**
- Short-Time Objective Intelligibility **(STOI)**
- Source to Distortion Ratio **(SDR)**

STOI performance at -5 dB input SNR

	Factory	Babble	White	Average
Unprocessed	0.5751	0.5887	0.6495	0.6044
Baseline DNN	0.5683	0.5723	0.7092	0.6166
PL DNN	0.6126	0.6093	0.7059	0.6426
NMF-DNN	0.6354	0.6319	0.7417	0.6697
Proposed	**0.6505**	**0.6429**	**0.7431**	**0.6788**

- STOI is calculated between the clean speech and the enhanced speech signals has shown highly correlated with **the intelligibility of human speech**.
- The **proposed method has the best effect** on speech intelligibility.
- Compared with the stationary noise(White), the **progressive learning** method has a **better improvement on the non-stationary noise** (Factory, Babble).

STOI performance at 0 dB input SNR

	Factory	Babble	White	Average
Unprocessed	0.6963	0.7013	0.7651	0.7209
Baseline DNN	0.7069	0.6946	0.8152	0.7389
PL DNN	0.7353	0.7140	0.8112	0.7535
NMF-DNN	0.7545	0.7388	0.8369	0.7767
Proposed	**0.7596**	**0.7520**	**0.8391**	**0.7835**

- The **NMF based** method is better than the spectrogram based method.
- The **progressive learning** method has a better effect on **low SNR noise.** This is because progressive learning **gets more speech** features than normal DNN.
- The performance of our proposed method is still the best.

PESQ performance at -5 dB input SNR

	Factory	Babble	White	Average
Unprocessed	1.1779	1.3021	1.1375	1.2058
Baseline DNN	1.3389	1.2724	1.9552	1.5222
SNR-PL DNN	1.6159	1.5664	**2.0011**	1.7278
NMF-DNN	1.6532	1.5803	1.9275	1.7203
Proposed	**1.7344**	**1.6116**	**1.9915**	**1.7792**

- **PESQ** is preferred as an objective quality measure since it showed a strong correlation to **speech quality.**
- The **proposed method has the best effect** on **speech quality**.
- Compared with the **PL DNN** method, the **proposed method** has better performance on **non-stationary noise**.

PESQ performance at 0 dB input SNR

	Factory	Babble	White	Average
Unprocessed	1.4888	1.6301	1.4244	1.5144
Baseline DNN	1.8819	1.7914	**2.3046**	1.9926
PL DNN	2.1298	**2.0239**	**2.3102**	2.1780
NMF-DNN	2.1441	1.9691	2.1705	2.0946
Proposed	**2.2126**	**2.0168**	**2.3171**	**2.1889**

For high SNR, **proposed method doesn't perform very well** mainly because:

1. **The difference between the intermediate target and the noisy is small**, which makes the training effect worse.

2. The PESQ focuses more on **non-speech segments, reduced the effect of spectral structures** and common SE methods also have **good noise reduction effects** on non-speech segments.

SDR performance at -5 dB input SNR

	Factory	Babble	White	Average
Unprocessed	-4.6817	-4.6858	-4.7009	-4.6895
Baseline DNN	1.5829	0.7653	5.9598	2.7693
PL DNN	2.1625	1.7938	5.8185	3.2582
NMF-DNN	3.3963	1.8800	6.7907	4.0223
Proposed	**3.6035**	**2.2178**	**7.0680**	**4.2964**

SDR performance at -0 dB input SNR

	Factory	Babble	White	Average
Unprocessed	0.1289	0.1584	0.1492	0.1455
Baseline DNN	5.2710	4.2812	8.5907	6.0476
PL DNN	5.8480	4.9235	8.6065	6.4593
NMF-DNN	6.7651	5.5344	9.5736	7.2910
Proposed	**6.9769**	**5.9362**	**10.1074**	**7.6735**

Conclusion

- **The proposed method performed well in all evaluation criteria, especially in the case of low signal-to-noise ratio and non-stationary noise.**
- **Speech enhancement based on spectral structures is better than that based on spectrogram.**
- **The proposed method outperform the other competitive speech enhancement methods, especially in speech intelligibility. This could because the STOI do not calculate the silence part of speech.**
- **In the future, we will further explore the influence of the intermediate target on speech enhancement and use speech phase to get better results.**

Thanks for your attention !