On Speech Pre-emphasis as a Simple and Inexpensive Method to Boost Speech Enhancement

Iván López-Espejo, Aditya Joglekar, **Antonio M. Peinado**, Jesper Jensen

IBERSPEECH 2024
Aveiro, Portugal

amp@ugr.es

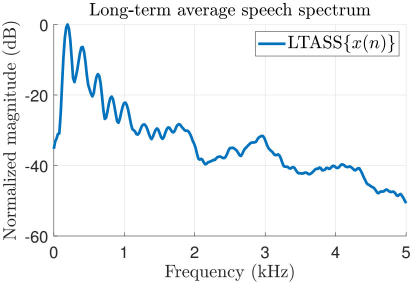
Wednesday 13th November, 2024

Overview

Introduction

- Speech is characterized by a **spectral tilt** stemming from glottal excitation due to vocal fold vibration

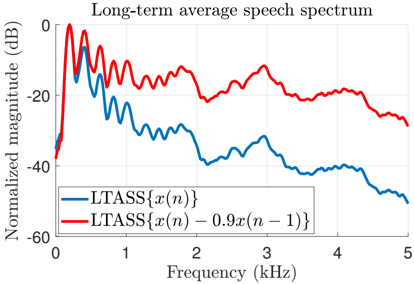
- Spectral tilt may lead to speech processing systems "overlooking" higher frequencies
  - **Perceptually-relevant speech elements such as fricatives, affricates, and some plosives have higher energy at higher frequencies!**



Introduction

- **Pre-emphasis filtering** is a simple yet effective pre-processing step that compensates high-frequency components by flattening the speech spectrum

- Pre-emphasis filtering is a *default consideration in classical ASR and speech coding systems*



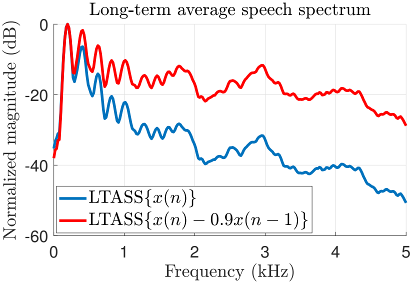
Introduction

- **We study pre-emphasis filtering for DNN-based speech enhancement**

How?

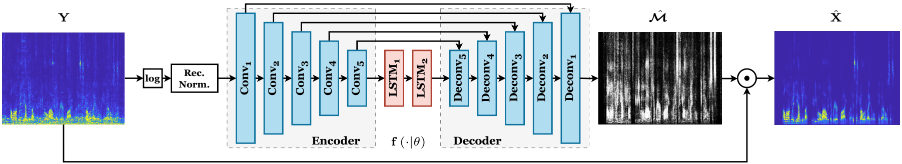
We explore pre-emphasizing the estimated and actual training clean speech during DNN training so that speech is perceptually balanced for loss calculation

- Our expectation is that the contribution of distinct speech frequency components to the total loss better reflects their perceptual importance



Speech Enhancement Framework

- We use a **spectral masking** scheme for speech enhancement purposes
- The mapping function $\mathbf{f}(\cdot|\theta)$ is deployed by a CRNN
- The enhanced waveform is synthesized by using the phase of the noisy signal



MSE loss function

$$\mathcal{L}_{\text{MSE}} = \frac{1}{KT} \sum_{k=0}^{K-1} \sum_{t=0}^{T-1} \left(|\hat{X}(k,t)| - |X(k,t)| \right)^2$$

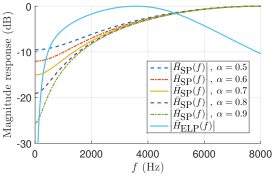
Speech Pre-emphasis Integration

1. We consider **two pre-emphasis variants** to be integrated into the loss function: **standard pre-emphasis** (**SP**) and **equal-loudness pre-emphasis** (**ELP**)
2. **Intensity-to-loudness conversion** (**I2L**) is optionally used to leverage pre-emphasis

---

Standard Speech Pre-emphasis (**SP**)

- First-order high-pass FIR filter
  $|H_{\mathrm{SP}}(f)| = \left|1 - \alpha e^{-j2\pi f/f_s}\right| = \sqrt{\alpha^2 - 2\alpha\cos(2\pi f/f_s) + 1}$
- $|\bar{H}_{\mathrm{SP}}(f)| \in (0, 1]$ is a scaled version of $|H_{\mathrm{SP}}(f)|$
- $|\bar{H}_{\mathrm{SP}}(k)|$ is found by uniform sampling of $|\bar{H}_{\mathrm{SP}}(f)|$



Pre-emphasized MSE loss function

$$\mathcal{L}_{\mathrm{MSE}}^{\mathrm{SP}} = \frac{1}{KT}\sum_{k=0}^{K-1}\sum_{t=0}^{T-1}\left(|\bar{H}_{\mathrm{SP}}(k)| \cdot \left(|\hat{X}(k,t)| - |X(k,t)|\right)\right)^2$$

Speech Pre-emphasis Integration

1. We consider **two pre-emphasis variants** to be integrated into the loss function: **standard pre-emphasis** (**SP**) and **equal-loudness pre-emphasis** (**ELP**)
2. **Intensity-to-loudness conversion** (**I2L**) is optionally used to leverage pre-emphasis

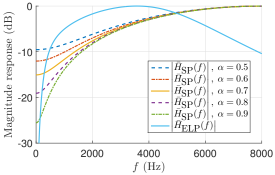
---

Equal-loudness Pre-emphasis (**ELP**)

- ELP approximates the frequency-dependent sensitivity of human hearing at about the 40 dB level:

$$|H_{\mathrm{ELP}}(f)| = \sqrt{\frac{(f^2+\beta_1)f^4}{(f^2+\beta_2)^2(f^2+\beta_3)((2\pi f)^6+\beta_4)}}$$

- $|\bar{H}_{\mathrm{ELP}}(f)| \in [0,1]$ is a scaled version of $|H_{\mathrm{ELP}}(f)|$
- $|\bar{H}_{\mathrm{ELP}}(k)|$ is found by uniform sampling of $|\bar{H}_{\mathrm{ELP}}(f)|$



Pre-emphasized MSE loss function

$$\mathcal{L}_{\mathrm{MSE}}^{\mathrm{ELP}} = \frac{1}{KT}\sum_{k=0}^{K-1}\sum_{t=0}^{T-1}\left(|\bar{H}_{\mathrm{ELP}}(k)|\cdot\left(|\hat{X}(k,t)| - |X(k,t)|\right)\right)^2$$

ELP accounts for the decrease in hearing sensitivity at higher frequencies

Speech Pre-emphasis Integration

1. We consider **two pre-emphasis variants** to be integrated into the loss function: **standard pre-emphasis** (**SP**) and **equal-loudness pre-emphasis** (**ELP**)
2. **Intensity-to-loudness conversion** (**I2L**) is optionally used to leverage pre-emphasis

---

Intensity-to-loudness Conversion (**I2L**)

- **Cubic-root amplitude compression** simulates the *non-linear relationship between the intensity of sound and its perceived loudness*

Pre-emphasized MSE loss function with I2L

$$\mathcal{L}_{\text{MSE}}^{\text{SP/ELP+I2L}} = \frac{1}{KT} \sum_{k=0}^{K-1} \sum_{t=0}^{T-1} \left(\left(|\bar{H}_{\text{SP/ELP}}(k)| \cdot |\hat{X}(k,t)| \right)^{\frac{2}{3}} - \left(|\bar{H}_{\text{SP/ELP}}(k)| \cdot |X(k,t)| \right)^{\frac{2}{3}} \right)^2$$

- Cubic-root amplitude compression can **boost the effect of pre-emphasis** by further reducing the dynamic range of the speech magnitude spectrum

Speech Dataset

- For experimental purposes, we use the **TIMIT-1C** speech dataset comprising clean and simulated noisy signals
- Clean signals were artificially distorted by diverse types of **additive noise**
  - **Training and validation sets:** car, bus station, restaurant, and street (**seen noises**)
  - **Test set:** café, train station, pedestrian street, and bus (**unseen noises**) + *seen noises*
- The training, validation and test sets consider the same discrete set of **SNRs**: $\{-5, 0, 5, 10, 15, 20\}$ dB
- Neither noise realizations nor speakers overlap across sets

Experimental Results

SP Standard pre-emphasis | **ELP** Equal-loudness pre-emphasis | **I2L** Intensity-to-loudness conversion

SNR (dB)	**Metric**	**Seen noises**						**Unseen noises**					
		Noisy	*Processed*					*Noisy*	*Processed*				
			✗	+SP		+ELP			✗	+SP		+ELP	
			✗	✗	+I2L	✗	+I2L		✗	✗	+I2L	✗	+I2L
-5	STOI	0.64	0.74	0.74	0.74	0.74	0.74	0.65	0.73	0.73	0.73	0.73	0.73
	PESQ	1.06	1.57	1.59	**1.62**	1.50	1.58	1.16	1.47	1.47	**1.49**	1.47	1.48
0	STOI	0.73	0.84	0.84	0.84	0.83	0.83	0.75	0.83	0.83	0.83	0.83	0.83
	PESQ	1.11	1.86	1.90	**1.93**	1.81	1.89	1.27	1.76	1.76	**1.81**	1.78	1.78
5	STOI	0.82	0.90	0.90	0.90	0.90	0.90	0.83	0.90	0.90	0.90	0.90	0.90
	PESQ	1.25	2.20	2.26	**2.31**	2.21	2.23	1.51	2.14	2.15	**2.21**	2.20	2.17
10	STOI	0.89	0.94	0.94	0.94	0.94	0.94	0.91	0.94	0.95	0.95	0.94	0.94
	PESQ	1.53	2.61	2.67	**2.72**	2.65	2.64	1.84	2.56	2.59	**2.66**	**2.66**	2.62
15	STOI	0.94	0.97	0.97	0.97	0.97	0.97	0.95	0.97	0.97	0.97	0.97	0.97
	PESQ	1.92	2.94	3.00	**3.08**	3.02	3.01	2.26	2.93	2.96	**3.04**	**3.04**	3.00
20	STOI	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	PESQ	2.45	3.30	3.35	**3.45**	3.38	3.38	2.84	3.32	3.37	**3.44**	3.43	3.40

- Evaluation carried out in terms of *quality* (**PESQ**) and *intelligibility* (**STOI**)
- For standard pre-emphasis, $\alpha = 0.6$ (*limited impact*)

Experimental Results

SP Standard pre-emphasis | **ELP** Equal-loudness pre-emphasis | **I2L** Intensity-to-loudness conversion

SNR (dB)	Metric	Seen noises						Unseen noises					
		Noisy	*Processed*					*Noisy*	*Processed*				
			✗	+SP		+ELP			✗	+SP		+ELP	
			✗	✗	+I2L	✗	+I2L		✗	✗	+I2L	✗	+I2L
-5	STOI	0.64	0.74	0.74	0.74	0.74	0.74	0.65	0.73	0.73	0.73	0.73	0.73
	PESQ	1.06	1.57	1.59	**1.62**	1.50	1.58	1.16	1.47	1.47	**1.49**	1.47	1.48
0	STOI	0.73	0.84	0.84	0.84	0.83	0.83	0.75	0.83	0.83	0.83	0.83	0.83
	PESQ	1.11	1.86	1.90	**1.93**	1.81	1.89	1.27	1.76	1.76	**1.81**	1.78	1.78
5	STOI	0.82	0.90	0.90	0.90	0.90	0.90	0.83	0.90	0.90	0.90	0.90	0.90
	PESQ	1.25	2.20	2.26	**2.31**	2.21	2.23	1.51	2.14	2.15	**2.21**	2.20	2.17
10	STOI	0.89	0.94	0.94	0.94	0.94	0.94	0.91	0.94	0.95	0.95	0.94	0.94
	PESQ	1.53	2.61	2.67	**2.72**	2.65	2.64	1.84	2.56	2.59	**2.66**	**2.66**	2.62
15	STOI	0.94	0.97	0.97	0.97	0.97	0.97	0.95	0.97	0.97	0.97	0.97	0.97
	PESQ	1.92	2.94	3.00	**3.08**	3.02	3.01	2.26	2.93	2.96	**3.04**	**3.04**	3.00
20	STOI	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	PESQ	2.45	3.30	3.35	**3.45**	3.38	3.38	2.84	3.32	3.37	**3.44**	3.43	3.40

- Pre-emphasis filtering has no impact on speech intelligibility
- Best speech quality is achieved by $\mathcal{L}_{\text{MSE}}^{\text{SP+I2L}}$ → PESQ rel. improv. over the baseline of 4.6% (*seen noises*) and 3.4% (*unseen noises*)

- Results indicate that perceptually balancing the estimated and actual clean speech signals prior to loss calculation allows for obtaining supplementary speech quality gains over a conventionally-trained modern speech enhancement system
- Minimal additional computational cost at training time, and *no additional cost at inference time*
- This *simple* and *cheap* methodology may potentially become a **default add-on** for training DNN-based speech enhancement systems

- Investigating the **generalizability** of this pre-emphasis methodology
  1. Different speech enhancement architectures/approaches
  2. Different loss functions

- Running **listening tests** to contrast what is predicted by objective speech quality and intelligibility metrics to strengthen the conclusions drawn

On Speech Pre-emphasis as a Simple and Inexpensive Method to Boost Speech Enhancement

Iván López-Espejo, Aditya Joglekar, **Antonio M. Peinado**, Jesper Jensen

IBERSPEECH 2024
Aveiro, Portugal

amp@ugr.es

Wednesday 13th November, 2024