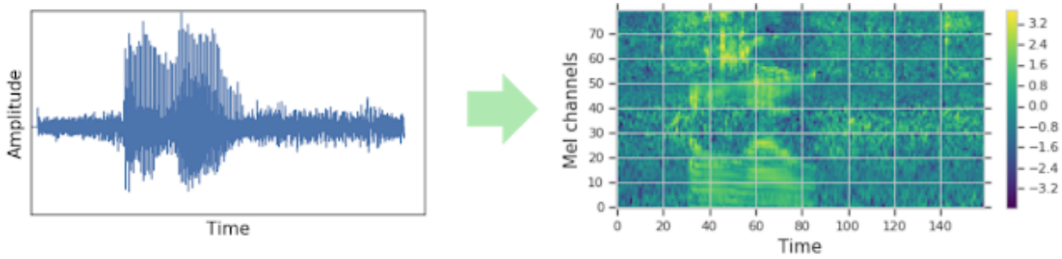
Speech Inpainting

Lukianov Pavel
MSU

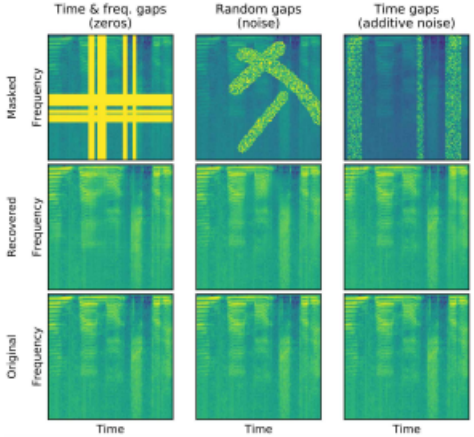
18 декабря 2020 г.

Audio in DL


Amplitude
Time
Mel channels
Time
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
70
3.2
2.4
1.6
0.8
0.0
−0.8
−1.6
−2.4
−3.2

Speech Inpainting task

Examples of Speech Inpainting:



SpeechVgg

Deep Feature Extractor for Speech Processing

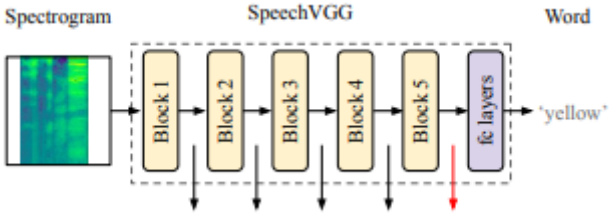


Рис. 1: SpeechVGG architecture

Architecture

Deep Feature Extractor for Speech Processing

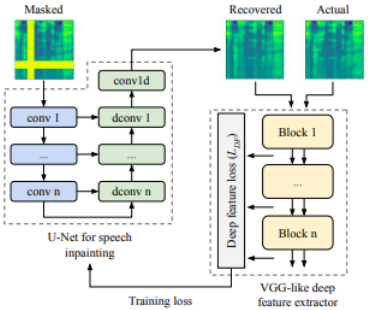


Рис. 2: The speech inpainting framework

Deep feature loss training

The reconstructed (Y_0) and the original (Y) signal were processed through the speechVGG extractor. Activation of all five of the extractor's pooling layers E, one at the end of each block were obtained for the reconstructed $E(Y_0)$ and actual $E(Y)$ samples. Deep feature loss L_{DF} was computed as the L_1 loss between the two representations:
$L_{DF} = L_1(E(Y), E(Y_0))$

Implementation details

- The speechVGG was pre-trained using cross-entropy loss for 50 epochs with a fixed learning rate set to $5 \cdot 10^{-3}$.
- Each considered configuration of the U-Net for speech inpainting was trained for 30 epochs using either deep feature loss with a fixed learning rate of $2 \cdot 10^{-4}$.
- ADAM optimizer was used in all the training routines.

Experiments

The results of experiments are on the table.

		Gaps		Noise		LPC		noVGG		imgVGG		speechVGG	
Intrusion	Size	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ
Time	10%	0.893	2.561	0.901	2.802	0.921	2.798	0.926	3.118	0.917	3.117	**0.938**	**3.240**
	20%	0.772	1.872	0.800	2.260	0.842	2.483	0.879	2.755	0.860	2.713	**0.887**	**2.809**
	30%	0.641	1.476	0.688	1.919	0.750	2.233	0.805	2.428	0.779	2.382	**0.811**	**2.450**
	40%	0.536	1.154	0.598	1.665	0.669	2.015	0.724	2.171	0.695	2.109	**0.730**	**2.179**
Time + Freq.	10%	0.869	2.423	0.873	2.575	0.887	2.692	0.905	2.921	0.899	2.915	**0.920**	**3.034**
	20%	0.729	1.790	0.746	2.010	0.780	2.378	0.845	2.518	0.829	2.491	**0.853**	**2.566**
	30%	0.598	1.391	0.629	1.653	0.668	2.128	0.765	2.158	0.743	2.134	**0.772**	**2.178**
	40%	0.484	1.053	0.520	1.329	0.566	**1.907**	0.672	1.840	0.644	1.809	**0.680**	1.845
Random	10%	0.880	2.842	0.892	3.063	N/A		0.941	3.477	0.927	3.399	**0.944**	**3.496**
	20%	0.809	2.233	0.830	2.543			0.912	3.079	0.897	3.040	**0.918**	**3.114**
	30%	0.713	1.690	0.745	2.085			0.872	2.702	0.856	2.680	**0.878**	**2.735**
	40%	0.644	1.355	0.682	1.802			0.837	2.443	0.823	2.422	**0.846**	**2.479**

Рис. 3: Informed speech inpainting

References

- SpeechVGG: A Deep Feature Extractor for Speech Processing
  Pierre Beckmann, Mikolaj Kegler, Hugues Saltini, Milos Cernak
  https://arxiv.org/pdf/1910.09909.pdf
- Deep Speech Inpainting of Time-frequency Masks
  Mikolaj Kegler, Pierre Beckmann, Milos Cernak
  https://www.isca-speech.org/archive/Interspeech_2020/pdfs/1532.pdf