Workshop

”Teaching AI to Hear Like We Do: Psychoacoustics in Machine Learning”

Gerald Schuller

Ilmenau University of Technology

Ilmenau, Germany

- Organized with the AES Technical Committee on Machine Learning and Artifical Intelligence (TC-MLAI)
- Format: Short presentations with Questions and Answers and possibly panel discussion

Technical Committee on Machine Learning and Artifical Intelligence

- focuses on applications of machine learning and artificial intelligence in audio
- Founded in 2021
- https://www.aes.org/technical/mlai/ (https://www.aes.org/technical/mlai/)

Panelists, Topics

- Gordon Wichern, MERL: High level perceptual loss functions, phase and magnitude
- Renato Profeta, Gerald Schuller, Ilmenau University of Technology: Perceptual loss functions: psycho acoustic models and loss functions
- Stefan Goetze, George Close, University of Sheffield: GAN-based perceptual metric prediction for speech enhancement
- Martin Strauss, Bernd Edler, AudioLabs Erlangen: Perceptually motivated conditional input for Flow-based speech enhancement

MITSUBISHI ELECTRIC
Changes for the Better

Audio Loss Functions in the Time and Frequency Domain

Gordon Wichern

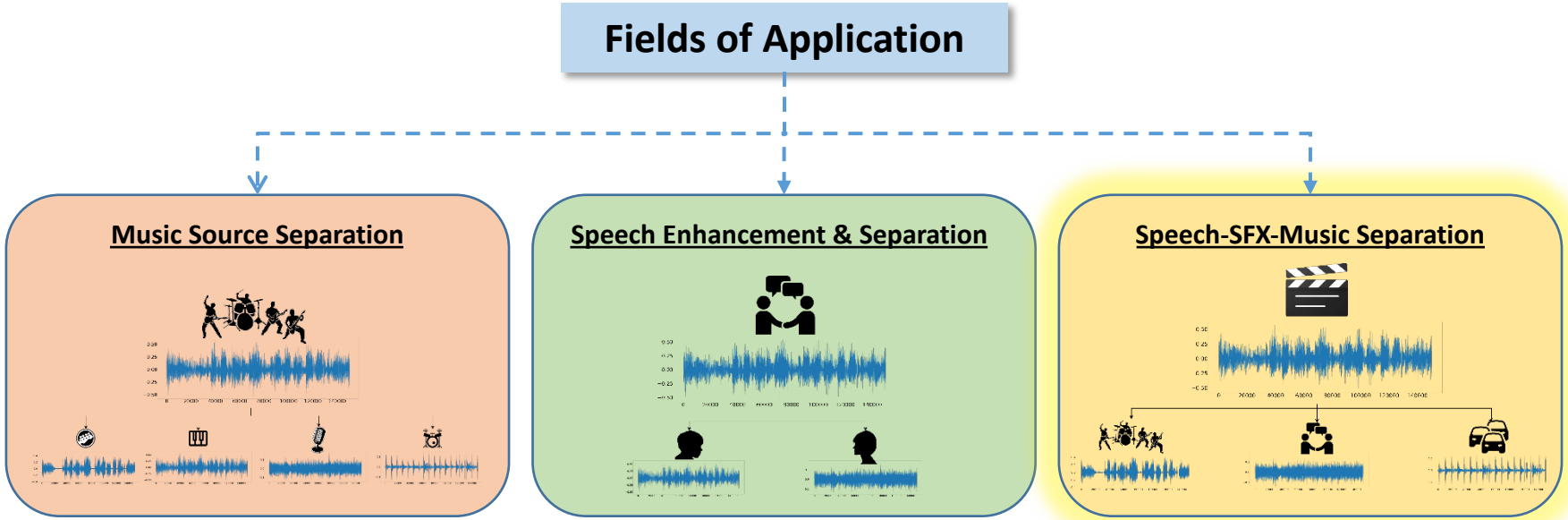
153rd Audio Engineering Convention - NYC

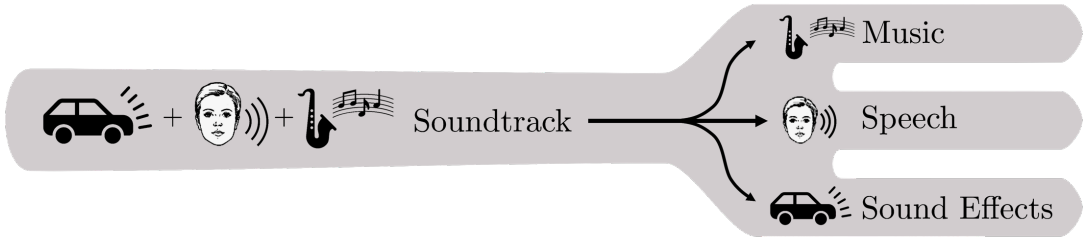
October 20, 2022

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Audio Source Separation:

- Separate a **mixed signal** into its **components**
- Unlike labeling tasks, we **listen** to the output

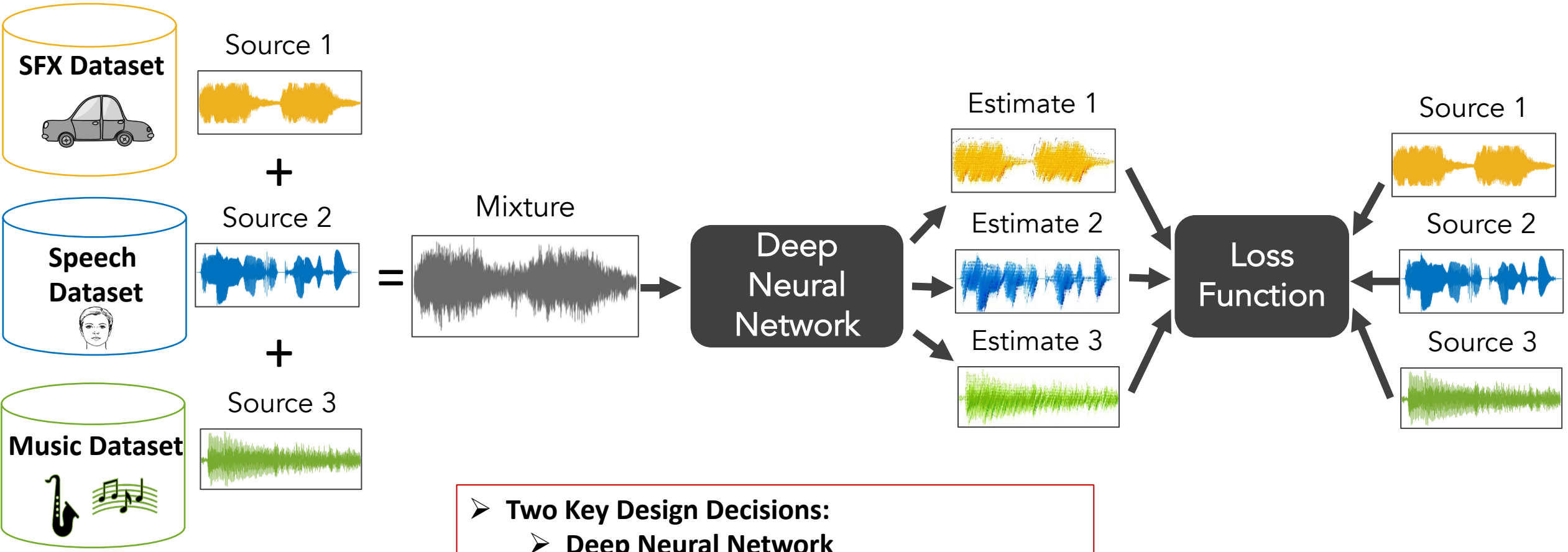




Petermann, D., Wichern, G., Wang, Z.-Q., Le Roux, J., **"The Cocktail Fork Problem: Three-Stem Audio Separation for Real-World Soundtracks"**, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), April 2022, pp. 526-530.

Demo video:
https://cocktail-fork.github.io/

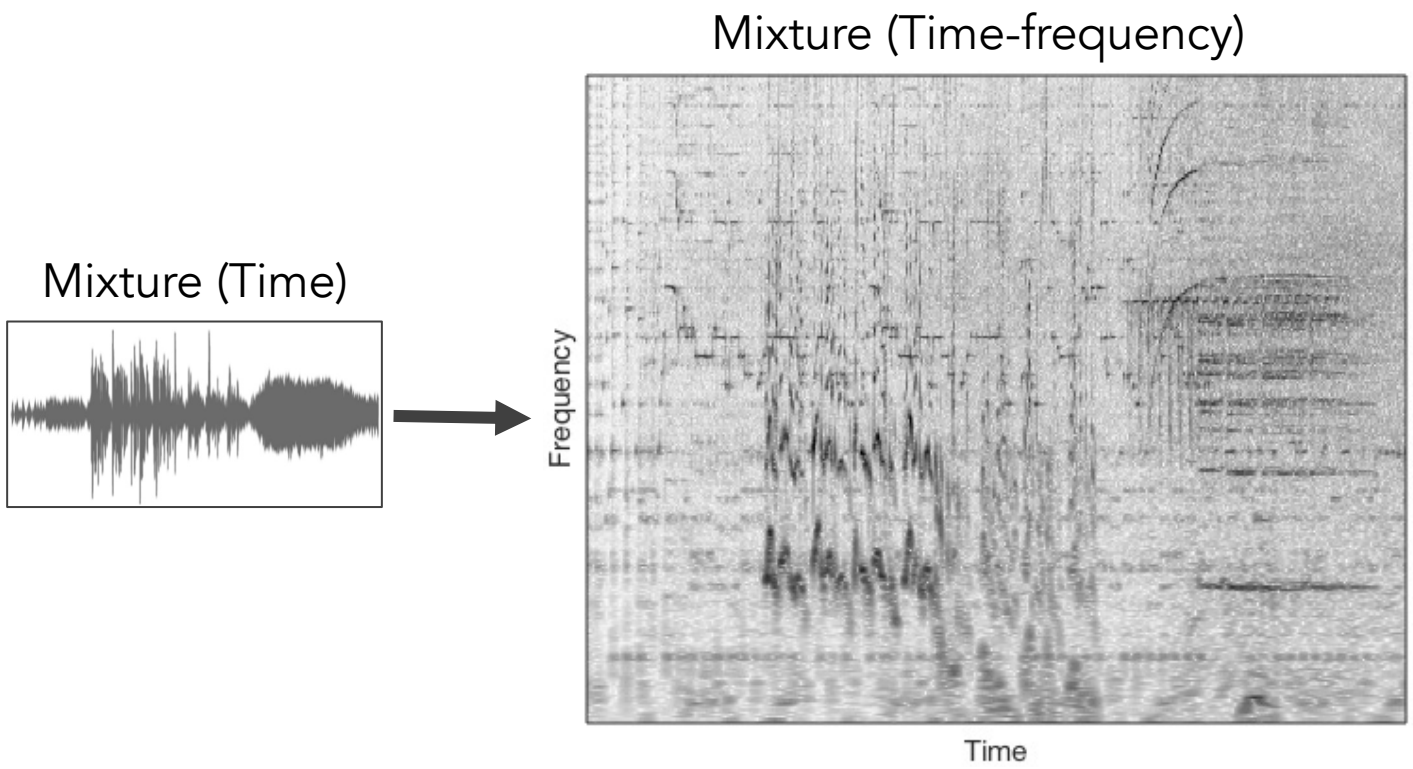
Deep Learning-based Source Separation Pipeline



- **Two Key Design Decisions:**
  - **Deep Neural Network**
    - **Operate on waveforms**
    - **Operate on spectrograms**
  - **Loss Function**
    - **Time domain (waveform)**
    - **Frequency domain (spectrogram)**

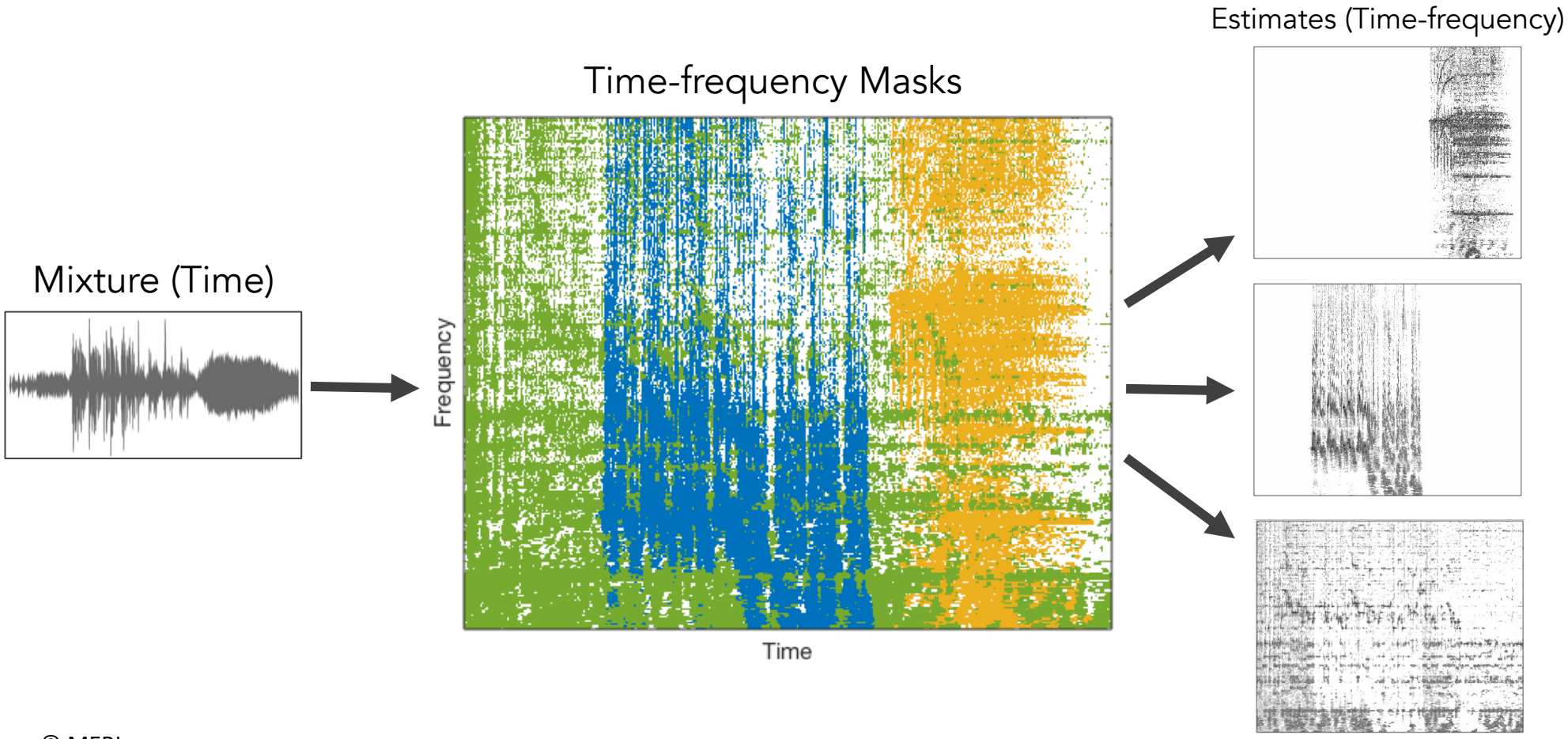
Frequency Domain (Magnitude Spectrogram)-based Loss Functions

- Deep learning revolution began with images (e.g., ImageNet)
- Magnitude spectrograms are an image
- Magnitude correlates strongly with human perception
- We can weight magnitude based on psychoacoustic models

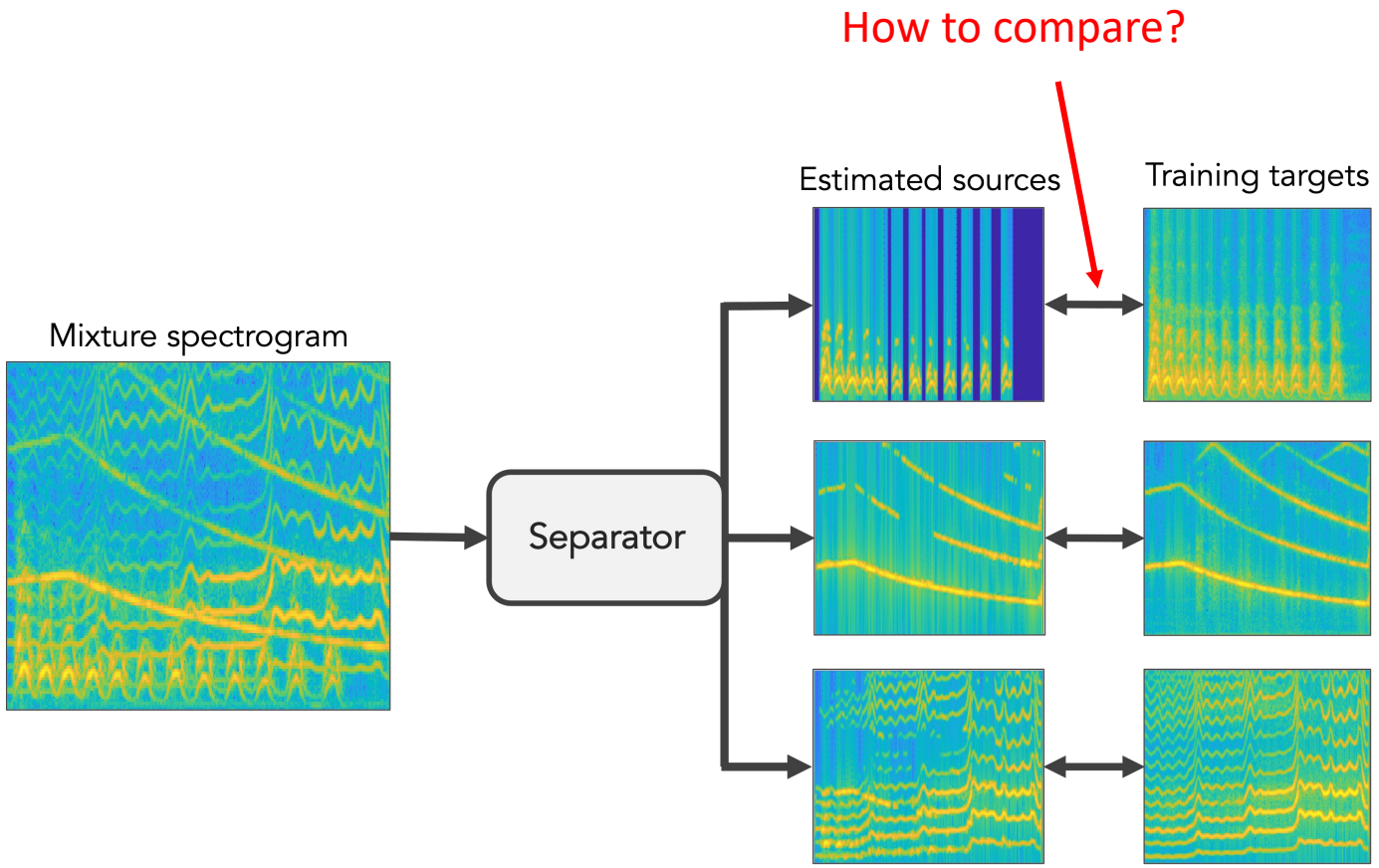


Masking-based audio source separation

- Classify the source each TF-bin belongs to
- Estimated source = mask * mixture



Frequency Domain (Magnitude Spectrogram)-based Loss Functions

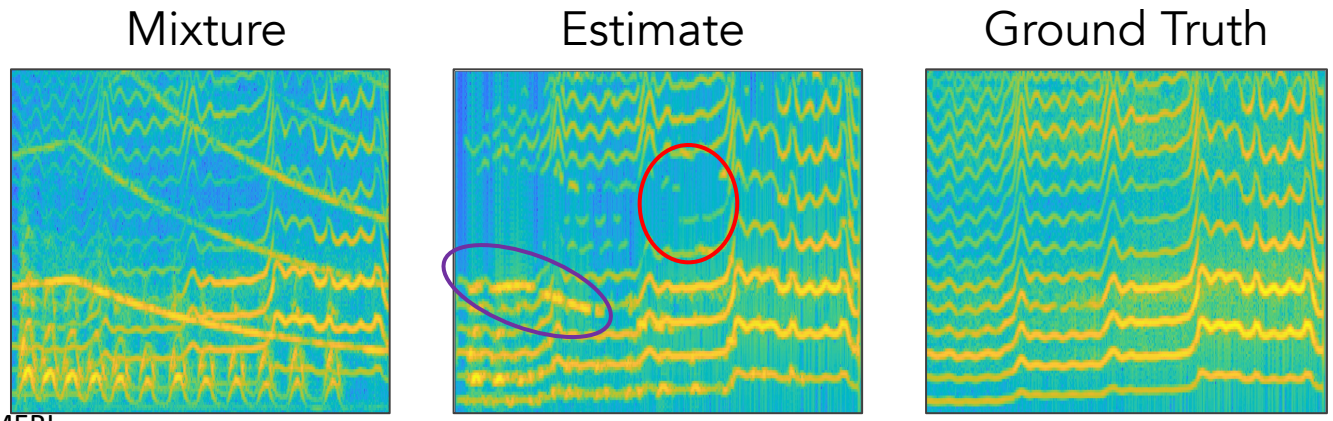


- Mean square error/Mean absolute error on magnitude spectrum
- Mask-based classification loss
- Typically we use noisy phase, only estimate magnitude
- Complex spectrogram loss
  - Complex numbers now fully supported in most deep learning toolkits
  - Similar to waveform losses
- Weight spectrum based on perceptual knowledge

MITSUBISHI ELECTRIC
Changes for the Better

Bias Models for Interference and Artifacts

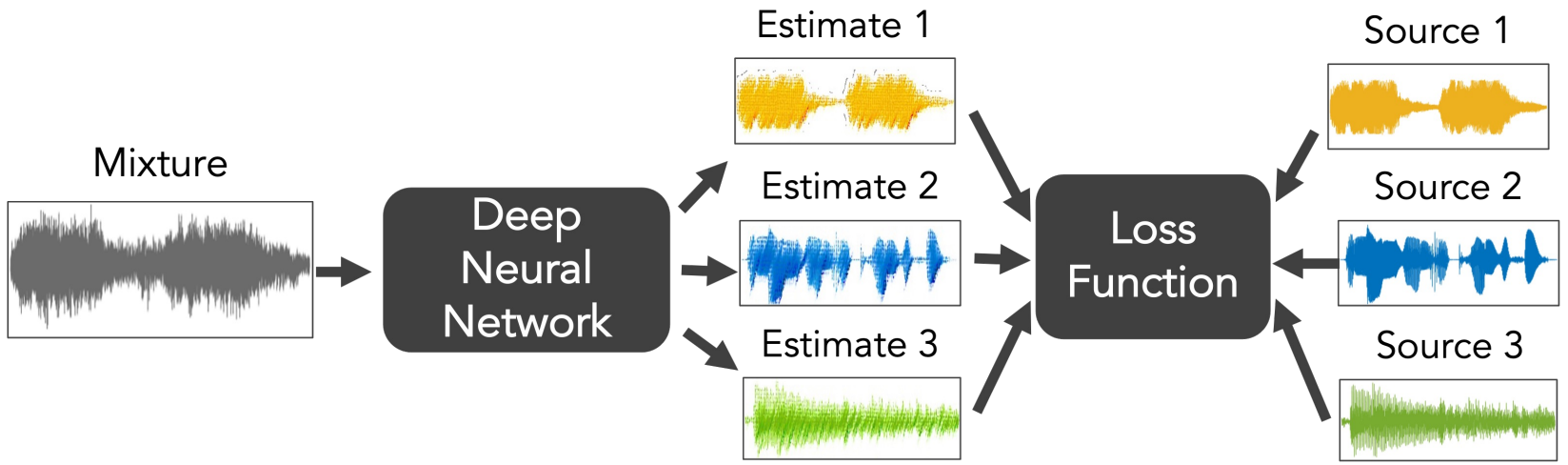
- Fundamental trade-off in source separation
  - Interference reduction
  - Artifact introduction
- In listening tests, artifacts tend to be more objectionable
- Incorporate bias in magnitude spectrogram loss
  - Estimate less than ground truth (high weight)
  - Estimate greater than ground truth (low weight)



MITSUBISHI ELECTRIC
Changes for the Better

Time Domain (Waveform) Loss Functions

- Spectrograms have many parameters we must hand-select
- Major contribution of deep learning revolution is we can learn features
- STFT features often require long window-sizes (high latency)
- Time-domain models work well with short windows (low-latency)
- Time domain loss functions can preserve spatial cues
- Mean-square (absolute) error on waveforms
- SNR-based loss functions
  - Scale invariant
  - Shift-invariant
  - Filter-invariant

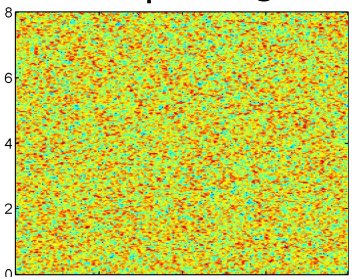


The Phase Compensation Problem

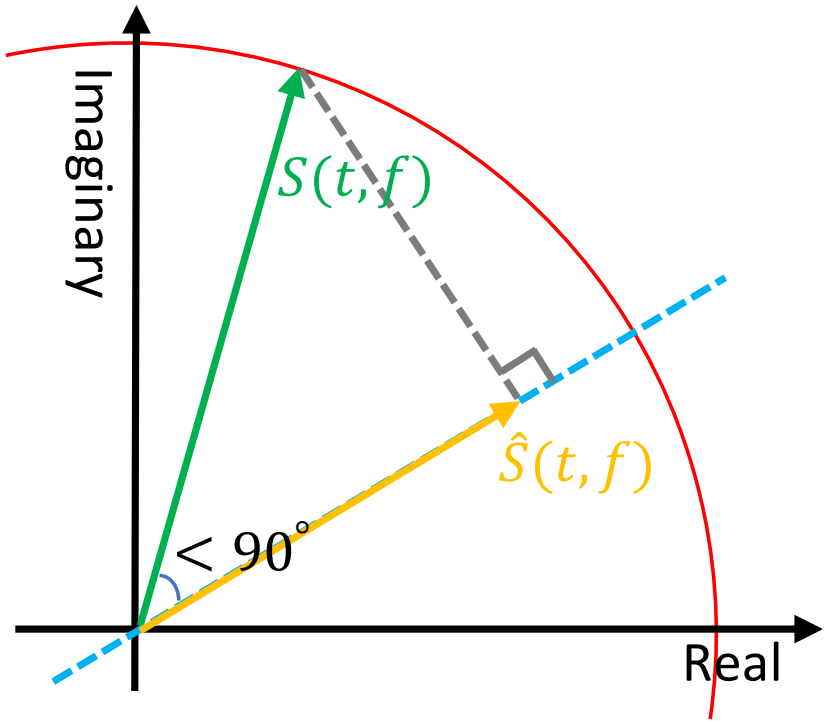
Wang, Z.-Q., Wichern, G., Le Roux, J., **"On The Compensation Between Magnitude and Phase in Speech Separation"**, IEEE Signal Processing Letters, DOI: 10.1109/LSP.2021.3116502, Vol. 28, pp. 2018-2022, November 2021.

- Waveform and complex spectrogram models must estimate phase
- Estimating phase is hard
- Bad phase estimates will hurt magnitude and cause artifacts
  - Especially at low SNR

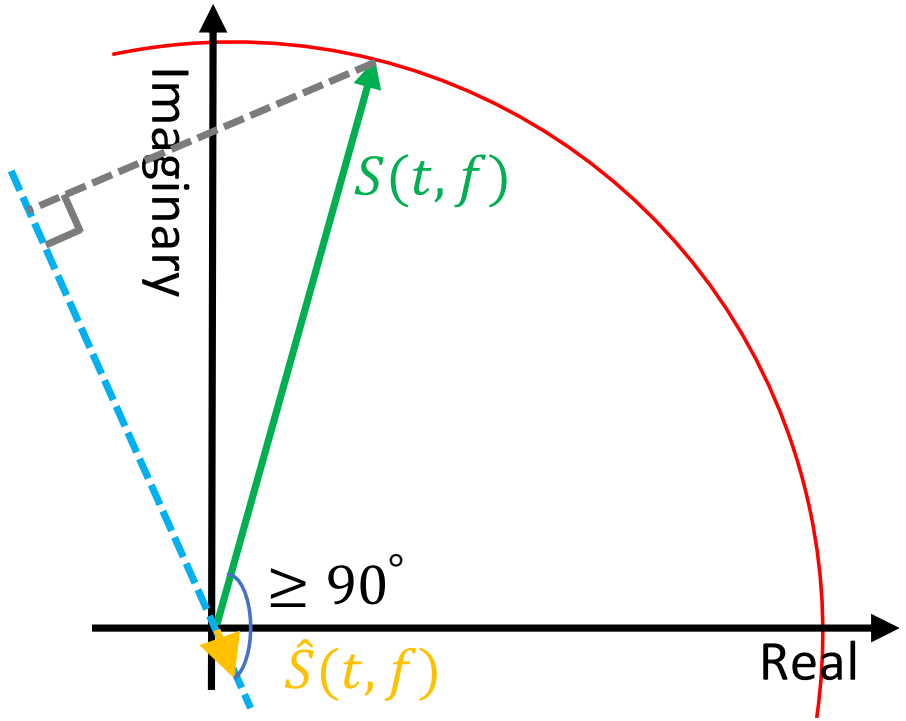
Phase Spectrogram



Phase Error less than 90 deg.

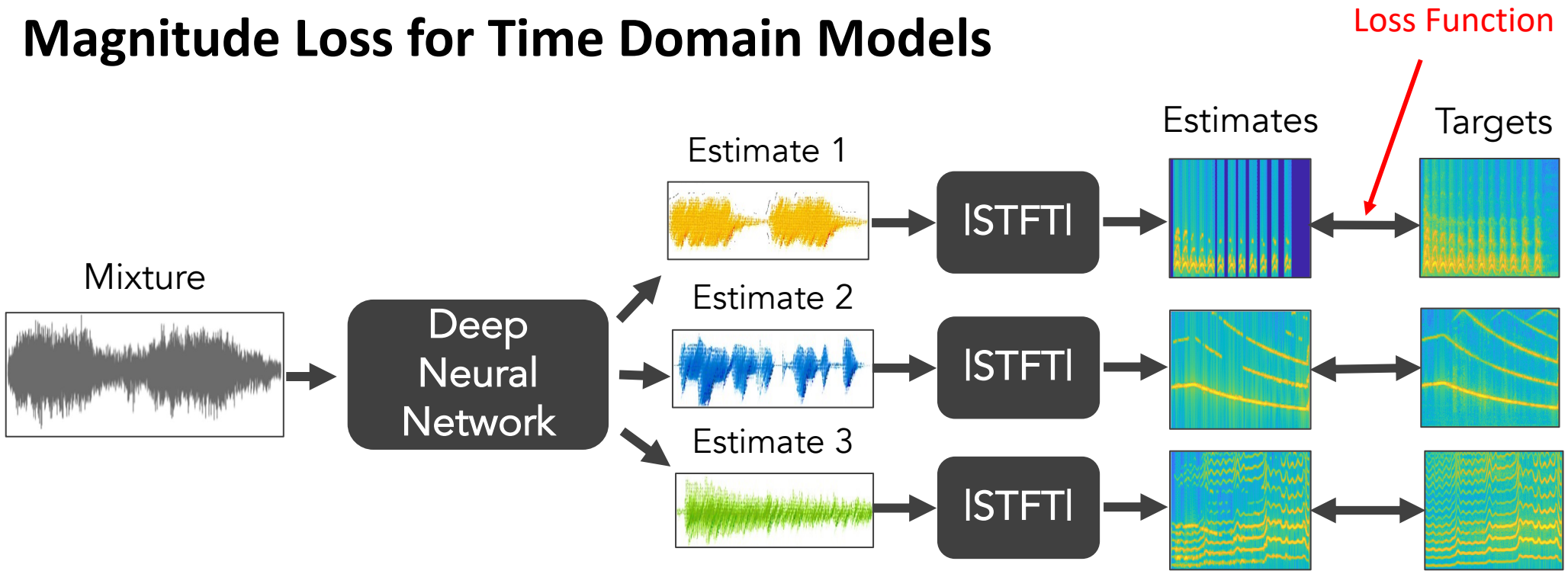


Phase Error greater than 90 deg.



MITSUBISHI ELECTRIC
Changes for the Better

Magnitude Loss for Time Domain Models



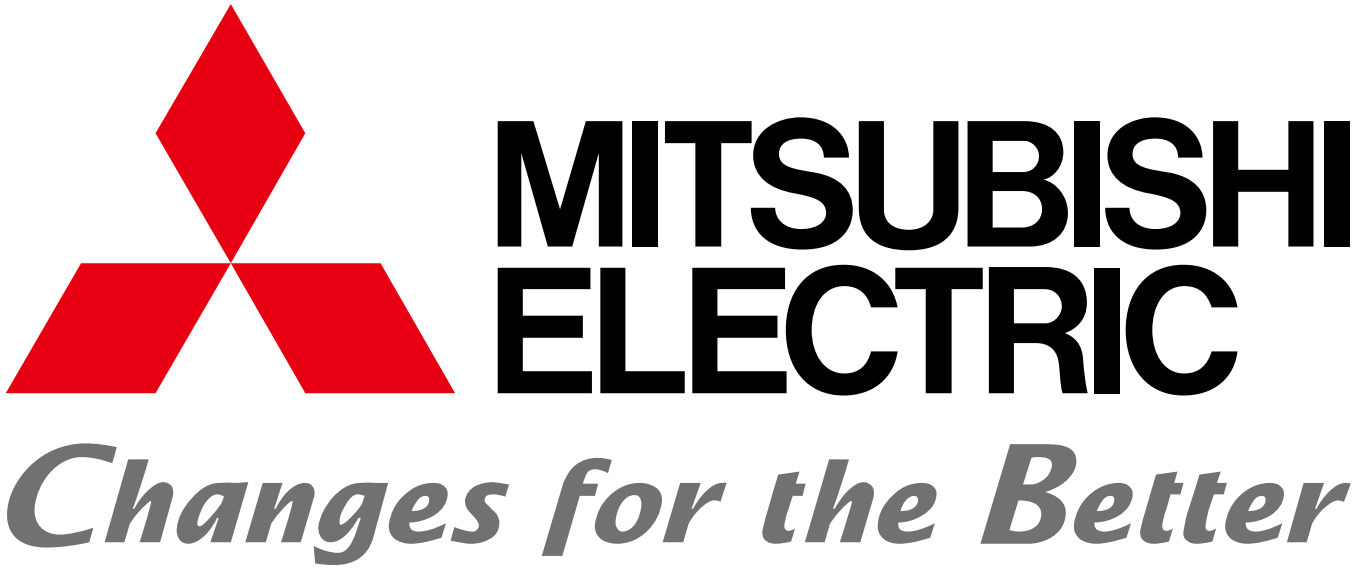
- Perceptual evaluation of speech quality (PESQ)
  - Based on spectrogram magnitude
- Scale-invariant source to distortion ratio (SI-SDR)
  - Compares time domain signals
  - Requires phase alignment

	Noisy Speech Separation	
	SI-SDR	**PESQ**
Noisy	-4.6	1.36
Waveform loss	**6.5**	1.61
Magnitude Loss	-9.24	**1.9**
Waveform + Magnitude Loss	**6.5**	1.80

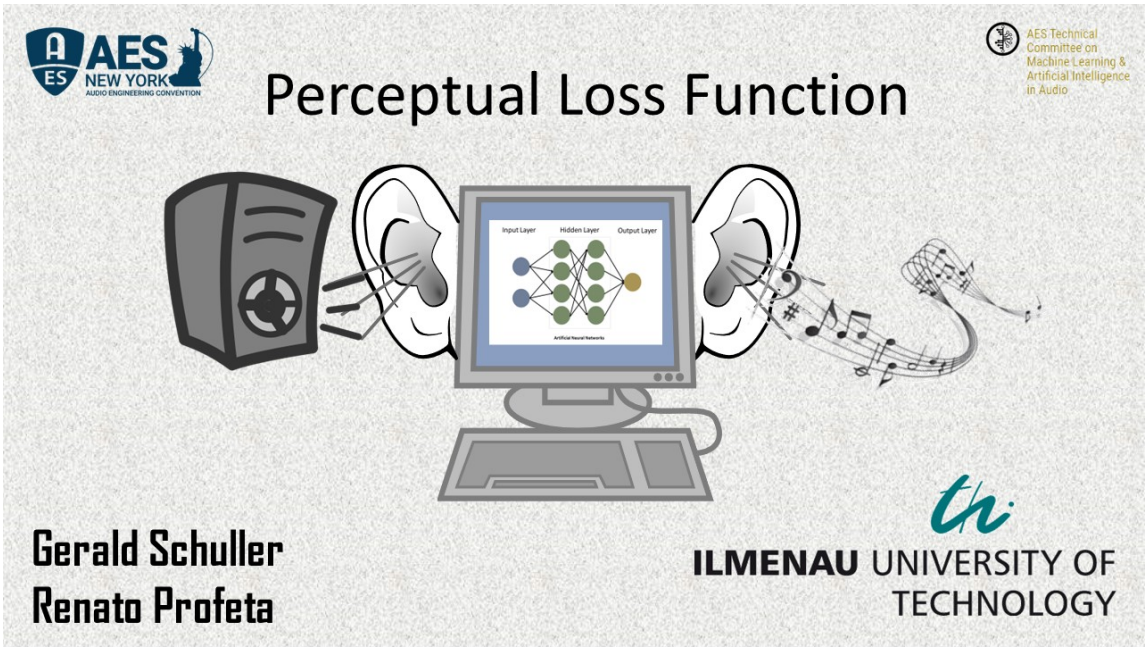
MITSUBISHI ELECTRIC
Changes for the Better

Take-Aways

- Time-domain (waveform) neural networks have multiple advantages
  - Strong performance
  - Low latency
  - No feature engineering
- Waveform loss functions
  - Preserve spatial cues
  - Artifacts due to poor phase estimates
- Magnitude spectrogram loss functions
  - Correlate with human perception
  - Perceptual weighting
  - Can trade-off artifact/Interference
  - No time-alignment when estimating phase
- **Use waveform+magnitude loss**
  - Relative weights depend on application


MITSUBISHI
ELECTRIC
Changes for the Better

Link to the following presentation including soundfiles:
https://tuilmenauams.github.io/PsychoacousticLoss/



Goal: Loss function which favors what sounds better to the human ear

- Depending on the temporal/spectral shape of the noise, the ear favours one over the other, depsite the same noise power

Problem: Popular and for gradient descent effective loss functions like MSE don't reflect these ear preferences

- Examples
- Sound with spectrally flat noise from PCM quantization.
- Sound with psychoacoustically spectrally shaped noise, with even higher noise power.

Evalutation with MSE

- Evaluating these example with the Mean Square Error (MSE) loss favors the first, noisy, example, wrongly!

Perceptual Loss Function using a Psycho-Acoustic Prefilter

- A psycho-acoustic prefilter uses a linear time-varying filter to normalize an audio signal to its psycho-acostic masking threshold.
- This is generated by a psycho-acoustic model, similar to what is used in audio coders.
- After this prefilter, we have a new signal or domain and we apply the MSE loss function there.

Psycho-Acoustic Prefilter Example

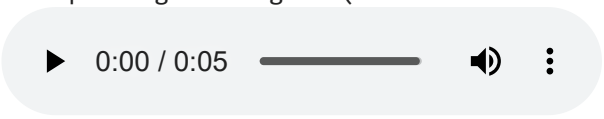
Musical exerpt: Slash - Anastasia, Released: 2012, Album: Apocalyptic Love

In [1]:
```
# Imports
import torch
import torchaudio
import IPython.display as ipd
```

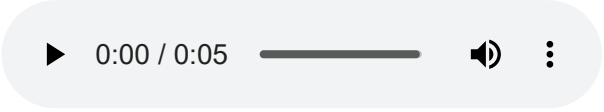
In [2]:
```
# Load audio files
audio_wav, sr_wav = torchaudio.load('audio_original.wav')
audio_mp3, sr_mp3 = torchaudio.load('audio_mp3_128k.wav')
audio_quantized, sr_wav = torchaudio.load('audio_quantized.wav')
```

In [3]:
```
# Playback
print('Example Signal Orignal (PCM 16-bit 44.1kHz)')
display(ipd.Audio(audio_wav,rate=sr_wav))
print('Example Signal MP3 128k')
display(ipd.Audio(audio_mp3,rate=sr_mp3))
print('Example Signal Quantized (choosen quantization factor)')
display(ipd.Audio(audio_quantized,rate=sr_wav))
```

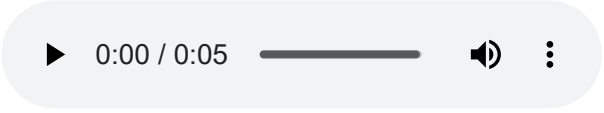
```
Example Signal Orignal (PCM 16-bit 44.1kHz)
```




```
Example Signal MP3 128k
```




```
Example Signal Quantized (choosen quantization factor)
```



Mean Squared Error (MSE) Loss

- One of the most common loss functions, widely used in many different applications.
- It assesses the average squared difference between the observed and predicted values.

Reference:

https://pytorch.org/docs/stable/generated/torch.nn.MSELoss.html

In [4]:
```
# MSE Loss
loss_mse = torch.nn.MSELoss()
mse_mp3_original = loss_mse(audio_mp3,audio_wav)
print('MSE Loss (mp3 and original):', mse_mp3_original*100)
mse_quant_original = loss_mse(audio_quantized,audio_wav)
print('MSE Loss (quanitzed and original):', mse_quant_original*100)
```

```
MSE Loss (mp3 and original): tensor(5.8766)
MSE Loss (quanitzed and original): tensor(1.3066)
```

Observe:

- The MSE loss of the mp3 is significantly greatar than the MSE loss of the quantized audio even though the perceived hearing quality of the mp3 is significantly superior.

Psycho-Acoustic Pre-Filtering + MSE

- A Psycho-Acoustic Model is used to generate filters that are applied to each block in the time-frequency domain.
- Computational expensive.

Reference:

Schuller, G. (2020). Filter Banks and Audio Coding. Springer International Publishing. https://doi.org/10.1007/978-3-030-51249-1

In [5]:
```
# Load pre-filtered audio files
audio_wav_pref, sr_wav = torchaudio.load('audio_originalpref.wav')
audio_mp3_pref, sr_mp3 = torchaudio.load('audio_mp3_128kpref.wav')
audio_quantized_pref, sr_wav = torchaudio.load('audio_quantizedpref.wav')
```

In [6]:
```
# Pre-Filtering + MSE Loss
loss_mse = torch.nn.MSELoss()
mse_mp3_original = loss_mse(audio_mp3_pref[0,:],audio_wav_pref[0,:])
print('Pre-Filtering + MSE Loss mp3:', mse_mp3_original.numpy()*10000)
mse_quant_original = loss_mse(audio_quantized_pref,audio_wav_pref)
print('Pre-Filtering + MSE Loss Quanitzed:', mse_quant_original.numpy()*10000)
```

```
Pre-Filtering + MSE Loss mp3: 1.00347948318813
Pre-Filtering + MSE Loss Quanitzed: 1.4562977594323456
```

Observe:

- Now, calculating the same MSE Loss in the psycho-acoustic pre-filtered domain, the MSE Loss for the mp3 audio is smaller than the MSE loss of que quantized sound.

Log Spectral Difference

- A distance measure (in dB) between log magnitudes of the spectra.
- Much less computational expensive.
- Can work well in certain applications.

Reference:

Rabiner, L. and Juang, B., 1993. Fundamentals of speech recognition. Englewood Cliffs, N.J.: PTR Prentice Hall.

In [7]:
```
from lsd_loss import LSDLoss
loss_lsd = LSDLoss()
lsd_mp3_original = loss_lsd(audio_mp3[0,:],audio_wav[0,:])
print('LSD Loss mp3:', lsd_mp3_original)
lsd_quant_original = loss_lsd(audio_quantized[0,:],audio_wav[0,:])
print('LSD Loss Quanitzed:', lsd_quant_original)
```

```
LSD Loss mp3: tensor(0.9744)
LSD Loss Quanitzed: tensor(1.9903)
```

Observe:

- The MSE loss of the quantized audio is also greater than the one for the mp3 audio, favouring the best sounding audio.

Multi Scale Spectral Loss

- More computational expensive than the LSD but less than the psycho-acoustic pre-filtering.
- Given two audio files, we compute their (magnitude) spectrogram Si and $\hat{S}_i$, respectively, with a given FFT size i, and define the loss as the sum of the L1 difference between Si and $\hat{S}_i$ as well as the L1 difference between log Si and log $\hat{S}_i$. The total reconstruction loss is then the sum of all the spectral losses with different FFT sizes.

Reference:
Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, & Adam Roberts (2020). DDSP: Differentiable Digital Signal Processing. In International Conference on Learning Representations.

In [8]:
```
from asteroid.losses import SingleSrcMultiScaleSpectral
loss_multiScaleSpectral = SingleSrcMultiScaleSpectral()
multiScale_mp3_original = loss_multiScaleSpectral(audio_mp3_pref,audio_wav_pref)
print('Multi Scale Spectral Loss mp3:', multiScale_mp3_original.numpy()/1000000)
multiScale_quant_original = loss_multiScaleSpectral(audio_quantized_pref,audio_wav_
print('Multi Scale Spectral Loss Quanitzed:', multiScale_quant_original.numpy()/100
```

```
Multi Scale Spectral Loss mp3: [1.14763687]
Multi Scale Spectral Loss Quanitzed: [2.28411725]
```

Observe:

- The MSE loss of the quantized audio is also greated than the one for the mp3 audio, favouring the best sounding audio.

Results

- Some losses favor the better sounding audio, while others don't.
- The psycho-acoustic filtering makes use of psycho-acoustic effects of the human hearing system and can be used in combination with a loss funcion in the design of a psycho-acoustically perceptual loss function.
- We can also transform the audio to a 'psycho-acoustic pre-filter domain' and perform different processing in this domain.

GAN-based Perceptual Metric Prediction for Speech Enhancement

Stefan Goetze, Geoge L. Close

{s.goetze, glclose1}@sheffield.ac.uk

153rd AES Convention

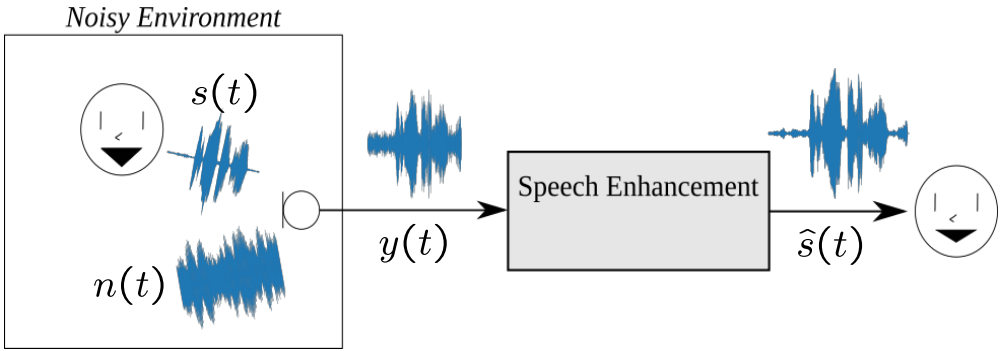
Teaching AI to hear like we do: psychoacoustics in machine learning

Oct 20th 2022, New York

Motivation

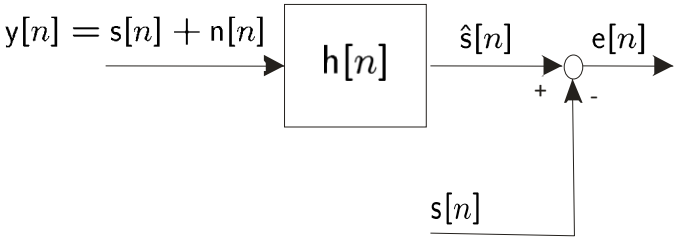
- Single channel speech enhancement still an active research area
- Mean Squared error (MSE) distance based loss functions between enhanced and clean reference speech **do not** consider human perception
- There exist many **metrics** which are designed with human perception in mind:
  - STOI - intrusive measure of speech intelligibility
  - PESQ - intrusive measure of speech quality
- PESQ (and other metrics) may be unsuitable as loss functions due to non linearities
- Design loss functions **derived from metrics** which incorporate a model of human perception

Problem Definition – Signal Enhancement



Goal is to recover the clean speech from the noisy mixture from the microphone

Wiener weighting rule



- **Wiener filter**

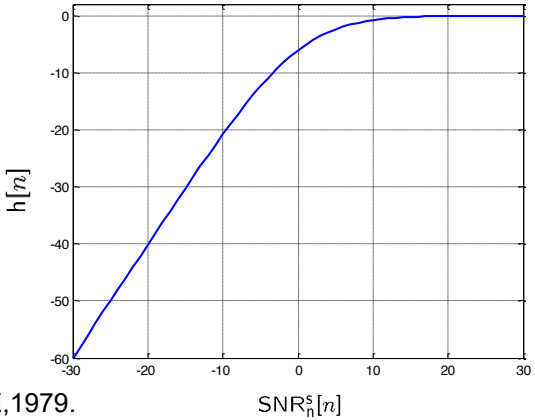
$$h_{opt}[n] = \frac{E\{y^*[n] \cdot s[n]\}}{E\{|y[n]|^2\}} = \frac{\Phi_{sy}[n]}{\Phi_{yy}[n]}$$

  - Problem: Dependency of unknown clean speech signal $s[n]$
  - Reformulation in terms of auto-power spectral densities possible

$$h[n] = \frac{\Phi_{ss}[n]}{\Phi_{ss}[n] + \Phi_{nn}[n]}$$

- **Spectral subtraction**

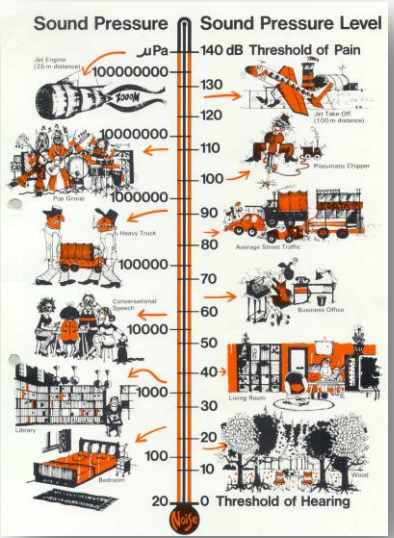
$$h[n] = \frac{\Phi_{yy}[n] - \Phi_{nn}[n]}{\Phi_{yy}[n]}$$



[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proc. IEEE,1979.
[2] *S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. ASSP, 1979*

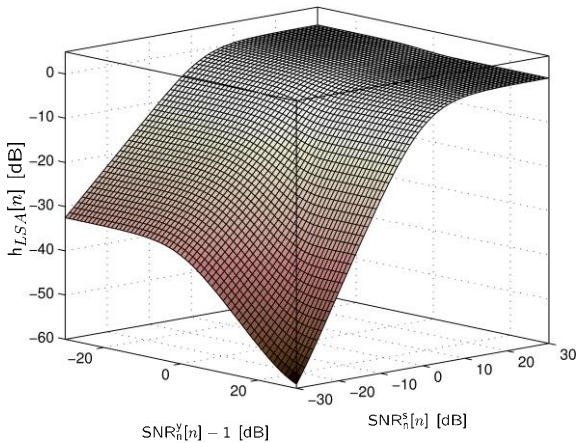
Weighting Function of Ephraim & Malah

- Approach: Minimise the logarithmic error/loss between clear speech and filter output



$$h_{LSA}[n] = \frac{\mathrm{SNR}_n^s[n]}{\mathrm{SNR}_n^s[n]+1} \exp\left(\frac{1}{2}\int_{\Psi[n]}^{\infty} \frac{e^{-\tau}}{\tau} d\tau\right)$$

$$\Psi[n] = \frac{\mathrm{SNR}_n^s[n]}{\mathrm{SNR}_n^s[n]+1}\mathrm{SNR}_n^y[n]$$



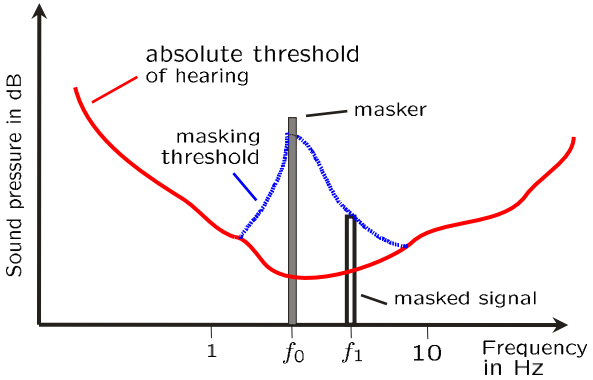
[1] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean–square error log-spectral amplitude estimator,” IEEE Trans. ASSP, 1985.

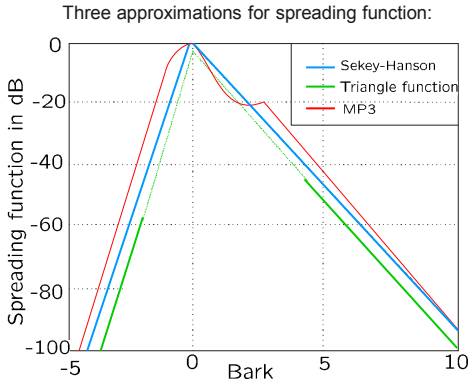
Use of Psychoacoustically Motivated Targets

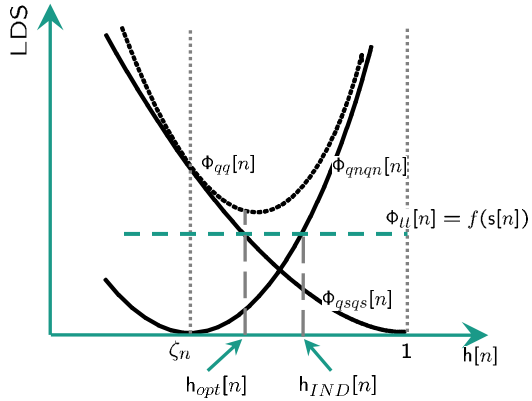
- Approach: Hide noise distortion under the masking threshold

$$h_{IND}[n] = \min\left(\sqrt{\frac{\Phi_{tt}[n]}{\Phi_{nn}[n]}} + \zeta_n,\ 1\right)$$

IND: inaudiable noise distortion





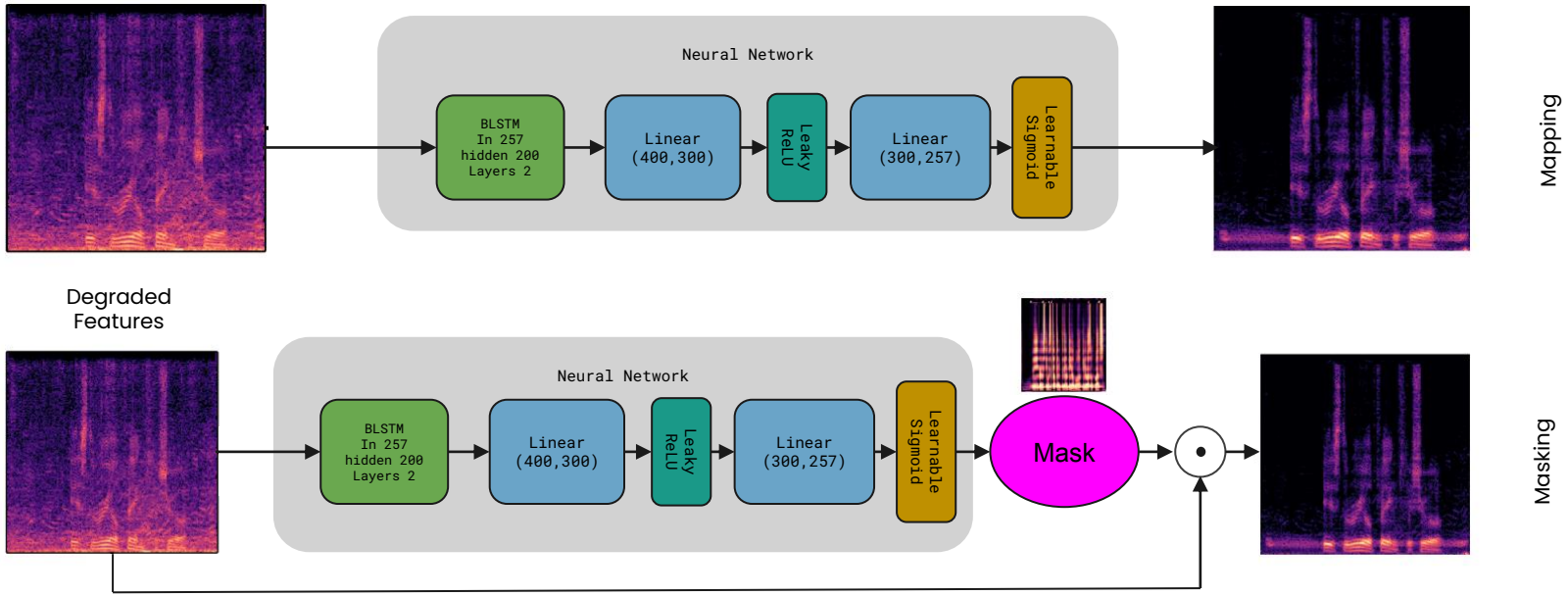


[1] S. Gustafsson, “Enhancement of Audio Signals by Combined Acoustic Echo Cancellation and Noise Reduction”, PhD thesis, Aachen, 1999

[2] *S. Goetze, V.* Mildner, KD Kammeyer*: “*A Psychoacoustic Noise Reduction Approach for Stereo Hands-Free Systems*”, AES 120th Convention, 2006*

Neural Network-based Signal Enhancement

- NN-based approaches are able to deal with non-stationary disturbances



Use of Psychoacoustically Motivated Targets

- Various different quality metrics exist
- Approach: Directly optimise for quality

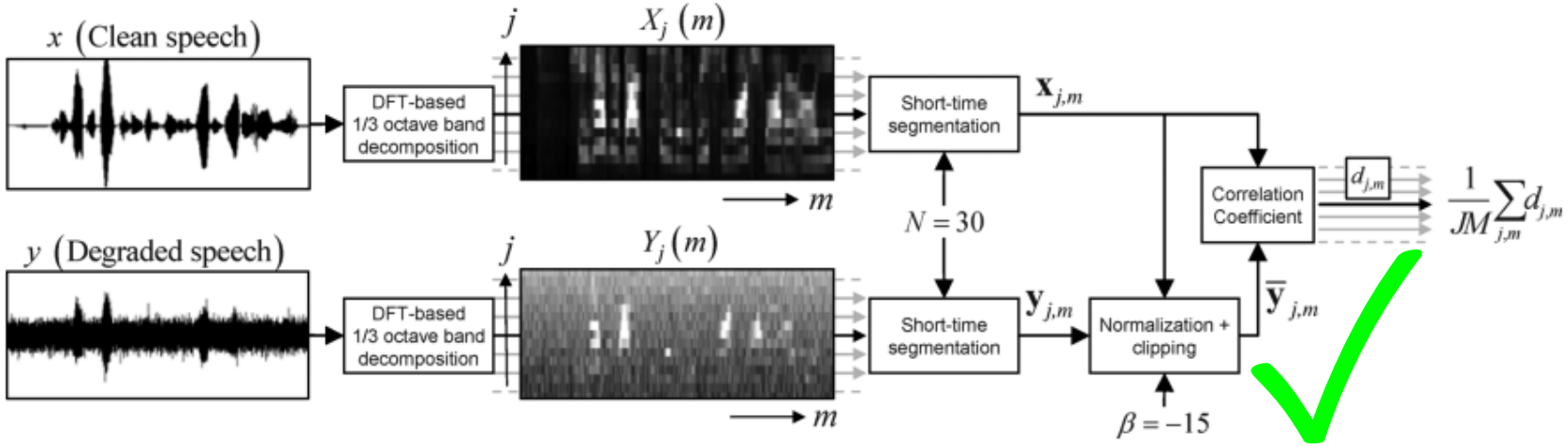
Signal-based quality measured	
Acronym	**Measure**
SSRR	Segmental Signal to Reverberation Ratio
SRRE	Signal to Reverberation Ratio Enhancement
FWSSRR	Frequency Weighted SSRR
WSS	Weighted Spectral Slope
OMCR	Objective Measure of Colouration in Reverberation
d_{IS} , d_{CEP}	Itakura-Saito-Measure, Cepstral Distance
d_{LAR} , d_{LLR}	Log Area Ratio, Log Likelihood Ratio
LSD	Log Spectral Distortion

Psychoacoustically motivated measures	
Acronym	**Measure**
BSD	Bark Spectral Distortion
R_{DT}	Reberberation Dacay Tail Measure
PSM	Perceptual Similarity Measure
PSM_t	Perceptual Similarity Measure (focus on low correlations)
ΔPSM	PSM „enhancement“
ΔPSM_t	PSM „enhancement“
PESQ	Perceptual Evaluation of Speech Quality
SRMR	Speech to Reverberation Modulation Energy Ratio

- Are these metrics differentiable to be used as a loss?

Short-Term Speech Intelligibility (STOI)

- Only few metrics are simple enough to be differentiable



[1] C. H. Taal, et. al, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech”, IEEE Trans. ASLP, 2011.
[2] *Fu et al, “End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks”,TASLP, 2018*

Perceptual Evaluation of Quality (PESQ) Metric

- Commonly used metric PESQ is not differentiable.

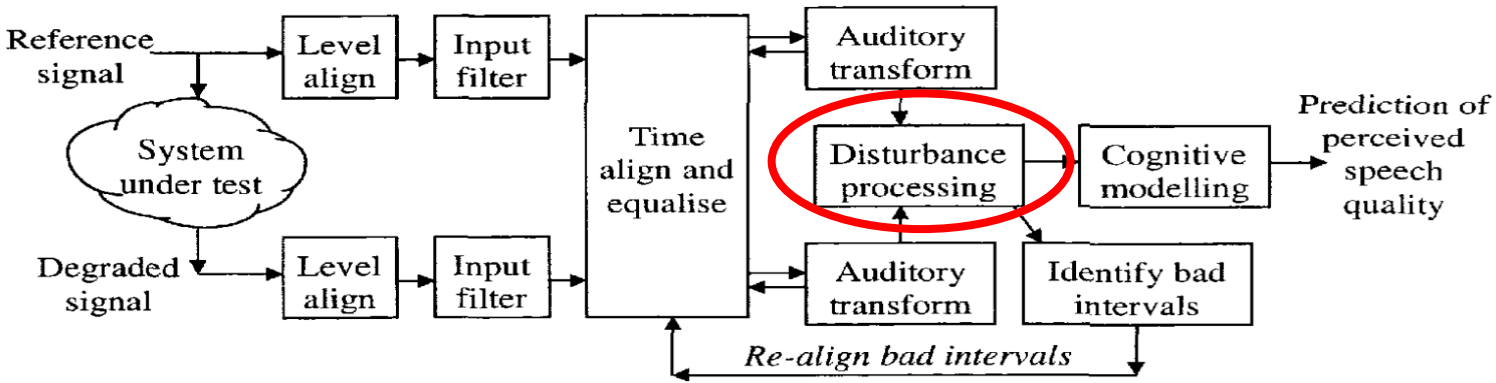


Figure 1: Structure of perceptual evaluation of speech quality (PESQ) model.

[1] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in ICASSP’01, 2001.

Differentiable PESQ?

- Approaches exist, but are only approximations

10.2.9 Calculation of the disturbance density

The signed difference between the distorted and original loudness density is computed. When this difference is positive, components such as noise have been added. When this difference is negative, components have been omitted from the original signal. This difference array is called the raw disturbance density.

The minimum of the original and degraded loudness density is computed for each time-frequency cell. These minima are multiplied by 0.25. The corresponding two-dimensional array is called the mask array. The following rules are applied in each time-frequency cell:

- If the raw disturbance density is positive and larger than the mask value, the mask value is subtracted from the raw disturbance.
- If the raw disturbance density lies in between plus and minus the magnitude of the mask value, the disturbance density is set to zero.
- If the raw disturbance density is more negative than minus the mask value, the mask value is added to the raw disturbance density.

The net effect is that the raw disturbance densities are pulled towards zero. This represents a dead zone before an actual time frequency cell is perceived as distorted. This models the process of small differences being inaudible in the presence of loud signals (masking) in each time-frequency cell. The result is a disturbance density as a function of time (window number n) and frequency, $D(f)_n$.

PESQ Specification

B. Symmetrical and Asymmetrical Disturbances Computation

Here we simplify the computation of the symmetrical disturbance vector proposed in PESQ by applying a center-clipping operator over the absolute difference between the loudness spectra as,

$$\mathbf{d}_t^{(s)} = \max\left(|\hat{\mathbf{s}}_t - \mathbf{s}_t| - \mathbf{m}_t, \mathbf{0}\right), \quad (5)$$

with a clipping factor

$$\mathbf{m}_t = 0.25 \cdot \min\left(\hat{\mathbf{s}}_t, \mathbf{s}_t\right), \quad (6)$$

where $|\cdot|$, $\min(\cdot)$ and $\max(\cdot)$ are applied element-wise and $\mathbf{0}$ is a zero-filled vector of length Q (note that, although non-derivable at singular points, previous operators allow to compute a sub-gradient for backpropagation at these points). This way, the psychoacoustic process by which small spectra differences are inaudible when loud signals are present is accounted for [23].

We obtain the asymmetrical disturbance vector as $\mathbf{d}_t^{(a)} = \mathbf{d}_t^{(s)} \odot \mathbf{r}_t$, where $\odot$ indicates an element-wise multiplication and $\mathbf{r}_t$ is a vector of asymmetry ratios whose components are computed from the Bark spectra as,

$$R_{t,q} = \left(\frac{\widehat{B}_{t,q} + \epsilon}{B_{t,q} + \epsilon}\right)^{\lambda}. \quad (7)$$

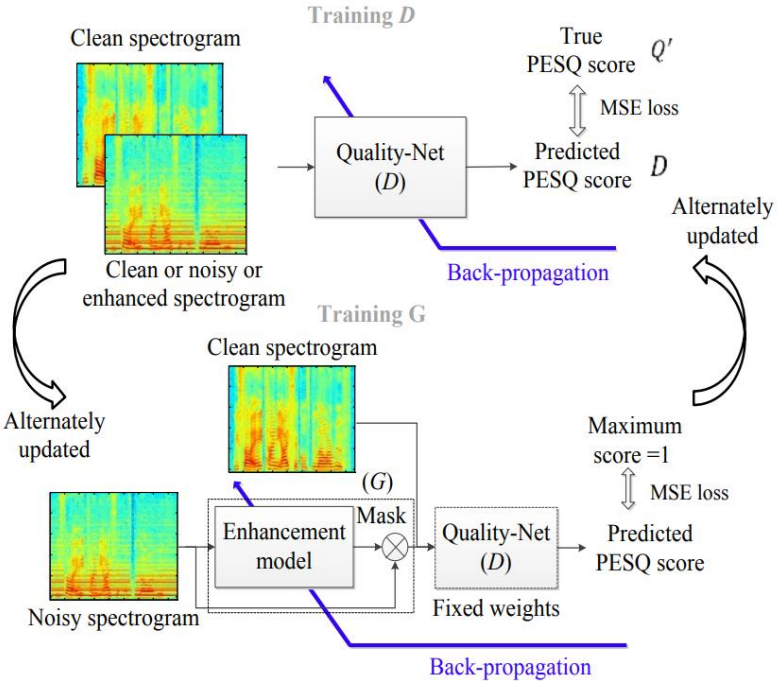
The asymmetry ratio accounts for positive ($\widehat{B}_{t,q} > B_{t,q}$) and negative ($\widehat{B}_{t,q} < B_{t,q}$) differences between the enhanced and the target spectra in the Bark domain by correspondingly applying a gain or an attenuation to the symmetrical disturbance. The constants ϵ and λ, set to 50 and 1.2 respectively (see next subsection), stabilize the ratio against very small Bark spectrum values and magnify the effect of the resulting ratio, respectively. Prior to the element-wise multiplication, asymmetry ratios $R_{t,q}$ are upper-bounded by a maximum value of 12, while those lower than 3 are set to 0, as in [23].

Finally, we can obtain the symmetrical and asymmetrical disturbance terms in a vectorized way, for each frame, as weighted norms of the symmetrical and asymmetrical vectors,

Martin-Donas et al "A deep learning loss function based on the perceptual evaluation of the speech quality", 2018

Baseline System: MetricGAN+ [1]

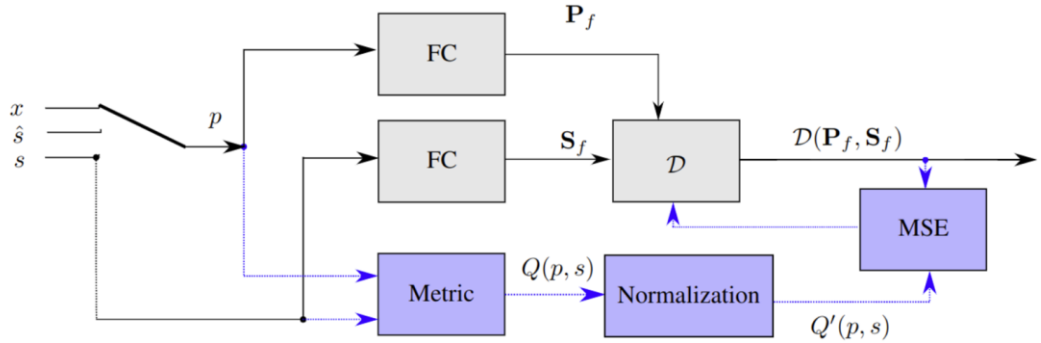
- Train an **additional** network to predict the behaviour of the metric
- Take inference of this predictor to form loss function for speech enhancement network
- The speech enhancement network and this metric prediction network are trained **adversarially** as Generator and Discriminator in a **GAN**



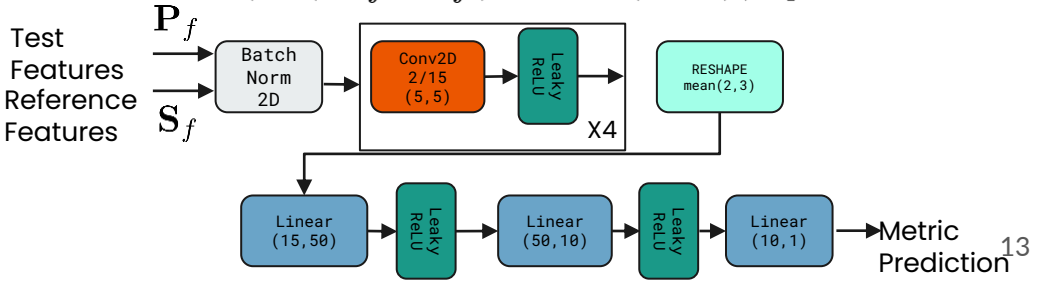
[1] *"MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement" - Fu et al, Interspeech,2021*

Metric Prediction Discriminator

- Neural model tasked with predicting the target metric
- The signals tested are:
  - Clean speech s
  - Enhanced speech (output of Generator) $\hat{s}$
  - Noisy speech x
- Loss function is MSE between true and predicted score

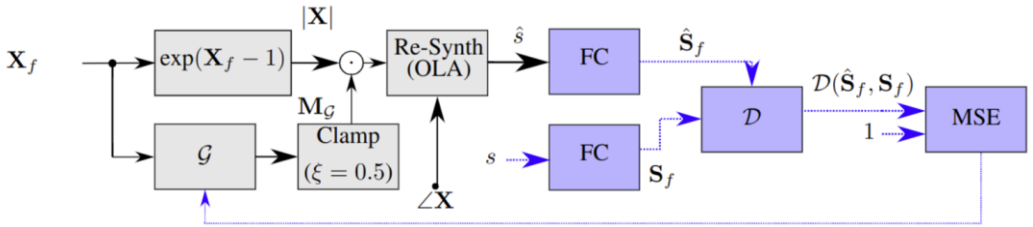


$$L_{D,MG+} = \mathbb{E}[(\mathcal{D}(\mathbf{S}_f, \mathbf{S}_f) - 1)^2 + (\mathcal{D}(\hat{\mathbf{S}}_\mathbf{f}, \mathbf{S}_f) - Q'(\hat{s}, s))^2 + (\mathcal{D}(\mathbf{X}_f, \mathbf{S}_f) - Q'(x, s))^2]$$



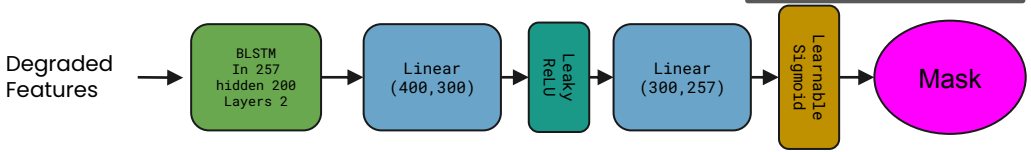
Speech Enhancement Generator

- Neural model tasked with outputting a magnitude ‘mask’ to be multiplied by the noisy features
- Loss function is based **entirely** on inference of Discriminator network
  - Goal is to produce outputs with ‘perfect’ score of 1

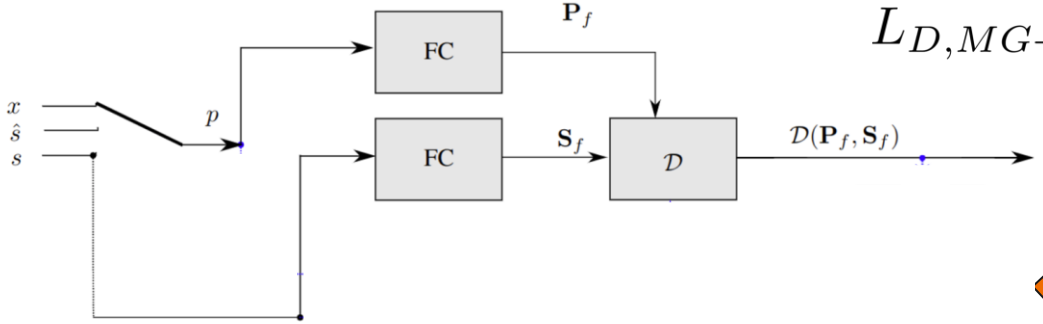


$$L_{\mathcal{G},\mathrm{MG+}} = \mathbb{E}[(\mathcal{D}(\hat{\mathbf{S}}_f, \mathbf{S}_f) - 1)^2]$$

Normalised optimum PESQ

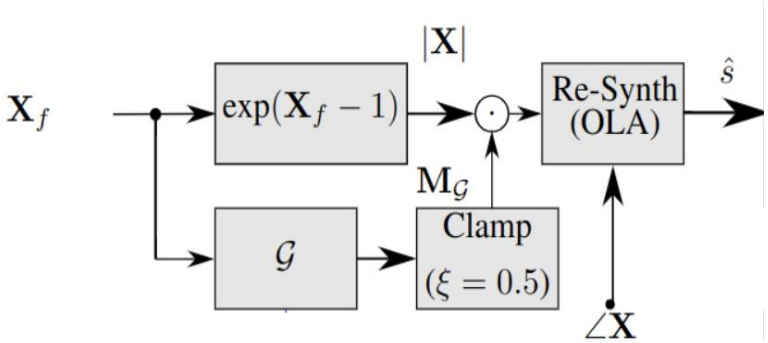


MetricGAN+



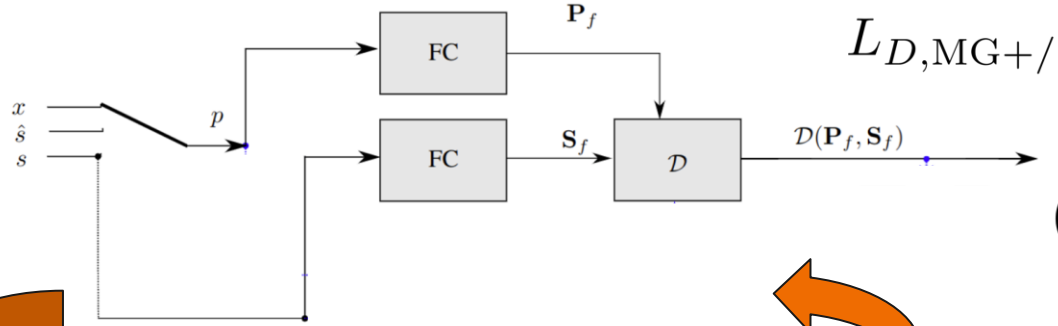
$$L_{D,MG+} = \mathbb{E}[(\mathcal{D}(\mathbf{S}_f, \mathbf{S}_f) - 1)^2 + (\mathcal{D}(\hat{\mathbf{S}}_\mathbf{f}, \mathbf{S}_f) - Q'(\hat{s}, s))^2 + (\mathcal{D}(\mathbf{X}_f, \mathbf{S}_f) - Q'(x, s))^2]$$

Alternately trained every epoch



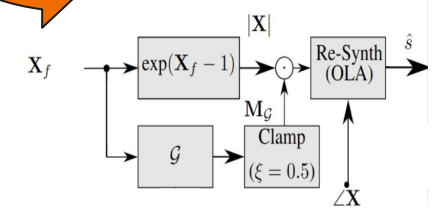
$$L_{\mathcal{G},\mathrm{MG+}} = \mathbb{E}[(\mathcal{D}(\hat{\mathbf{S}}_f, \mathbf{S}_f) - 1)^2]$$

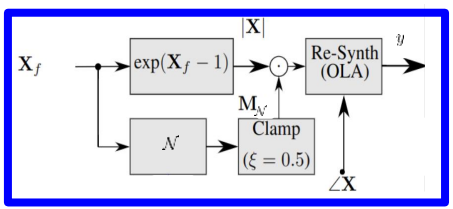
MetricGAN+/-



$$L_{D,\mathrm{MG+/-}} = \mathbb{E}[(\mathcal{D}(\mathbf{S}_f, \mathbf{S}_f) - 1)^2 + (\mathcal{D}(\hat{\mathbf{S}}_f, \mathbf{S}_f) - Q'(\hat{s}, s))^2 + (\mathcal{D}(\mathbf{X}_f, \mathbf{S}_f) - Q'(x, s))^2 + (\mathcal{D}(\mathbf{Y}_f, \mathbf{S}_f) - Q'(y, s))^2]$$

Alternately trained every epoch





$$L_{\mathcal{G},\mathrm{MG+}} = \mathbb{E}[(\mathcal{D}(\hat{\mathbf{S}}_f, \mathbf{S}_f) - 1)^2]$$

Hyperparameter value

$$L_{\mathcal{N},\mathrm{MG+/-}} = \mathbb{E}[(\mathcal{D}(\mathbf{Y}_f, \mathbf{S}_f) - w)^2], \quad \text{for } 0 < w < 1,$$

[1] G. Close, T. Hain, and S. Goetze: "MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data", EUSIPCO22, 2022

Results: VoiceBank-DEMAND Testset

Model	PESQ	STOI	Csig	Cbak	Covl
Noisy	*1.97*	*92.0*	*3.35*	*2.44*	*2.63*
SEGAN	2.42	92.5	3.61	2.61	3.01
MetricGAN+ (PESQ)	3.05	**93.0**	4.03	2.87	3.52
MetricGAN+/- (PESQ)	**3.22**	91.3	**4.05**	**2.94**	**3.62**

Improves over baseline in terms of PESQ, Composite measure and comparative in STOI

[1] G. Close, T. Hain, and S. Goetze: “MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data“, EUSIPCO22, 2022

Results: CHiME3 Testset

Model	PESQ	STOI	Csig	Cbak	Covl
Noisy	*1.32*	*65.0*	*2.79*	*1.40*	*1.99*
MetricGAN+ (PESQ)	1.84	**66.6**	2.86	2.12	2.27
MetricGAN+/- (PESQ)	**1.97**	65.3	**2.89**	**2.15**	**2.33**

Improves over baseline in terms of PESQ, Composite measure and comparative in STOI

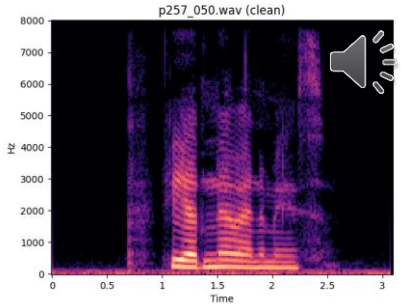
[1] G. Close, T. Hain, and S. Goetze: “MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data“, EUSIPCO22, 2022

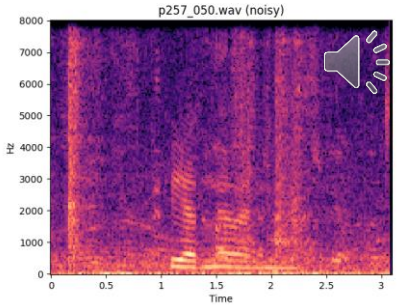
SLT Centre for Doctoral Training

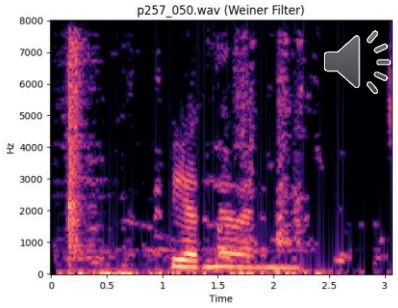
The University Of Sheffield.

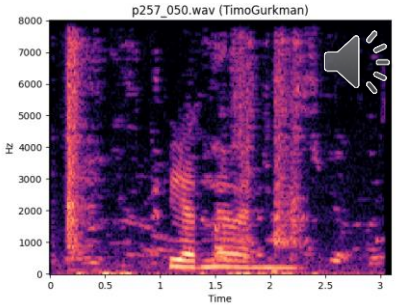
UK Research and Innovation

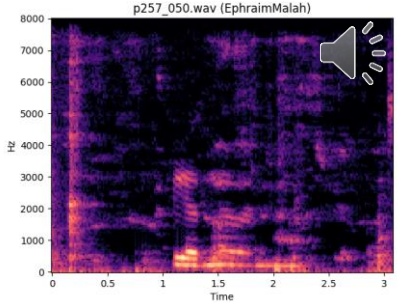
Some Sound Examples

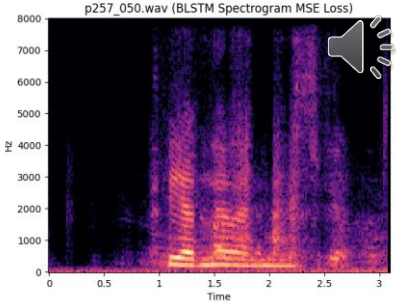


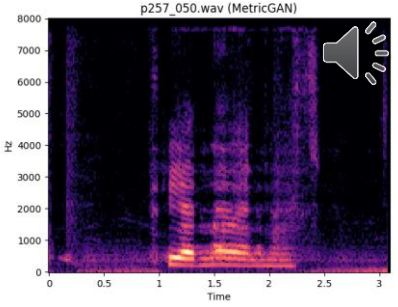


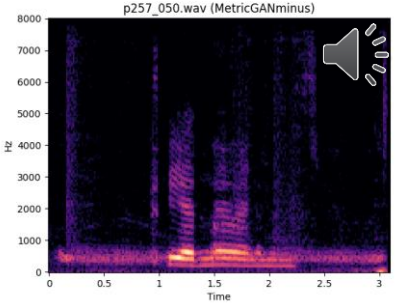










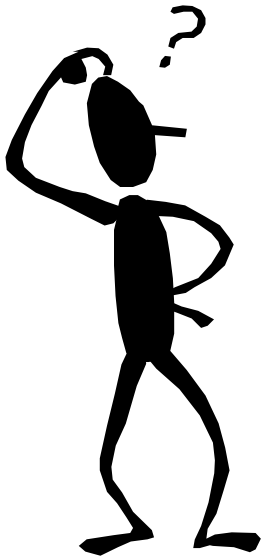


Conclusions

- Psychoacoustically motivated metrics beneficial in speech enhancement loss
- Neural networks outperform signal-processing-based approaches
- Incorporating perceptual knowledge into speech enhancement is possible
- Non-differentiable metrics can be used in losses using GAN structures
- Networks have to be exposed to a wide range of signal quality
    - Additional predictor networks can be used: MetricGAN+/-

Thanks for listening

- Any questions?



Stefan Goetze, Geoge L. Close
{s.goetze, glclose1}@sheffield.ac.uk

Perceptually Motivated Conditional Input for Flow-Based Speech Enhancement

AES Workshop
Teaching AI to hear like we do: psychoacoustics in machine learning

Martin Strauss and Bernd Edler
martin.strauss@audiolabs-erlangen.de
bernd.edler@audiolabs-erlangen.de

20.10.2022

Friedrich-Alexander-Universität Erlangen-Nürnberg
Fraunhofer IIS

Introduction

Speech enhancement (SE)

- Improve (perceptual) quality of speech degraded by background noise
- Predominant approach: Deep Neural Networks (DNN)
- Often application of a separation mask

Introduction

Generative models for SE

- Outline probabilistic process
- **Examples**:
  - Generative Adversarial Networks [1]
  - Diffusion probabilistic models [2]
  - **Normalizing flows** [5]

Normalizing flows
Fundamentals

- Invertible neural networks

$$\mathbf{x} = f(\mathbf{z}), \qquad \mathbf{z} = f^{-1}(\mathbf{x}), \qquad \mathbf{x}, \mathbf{z} \in \mathbb{R}^D$$

- Change of variables

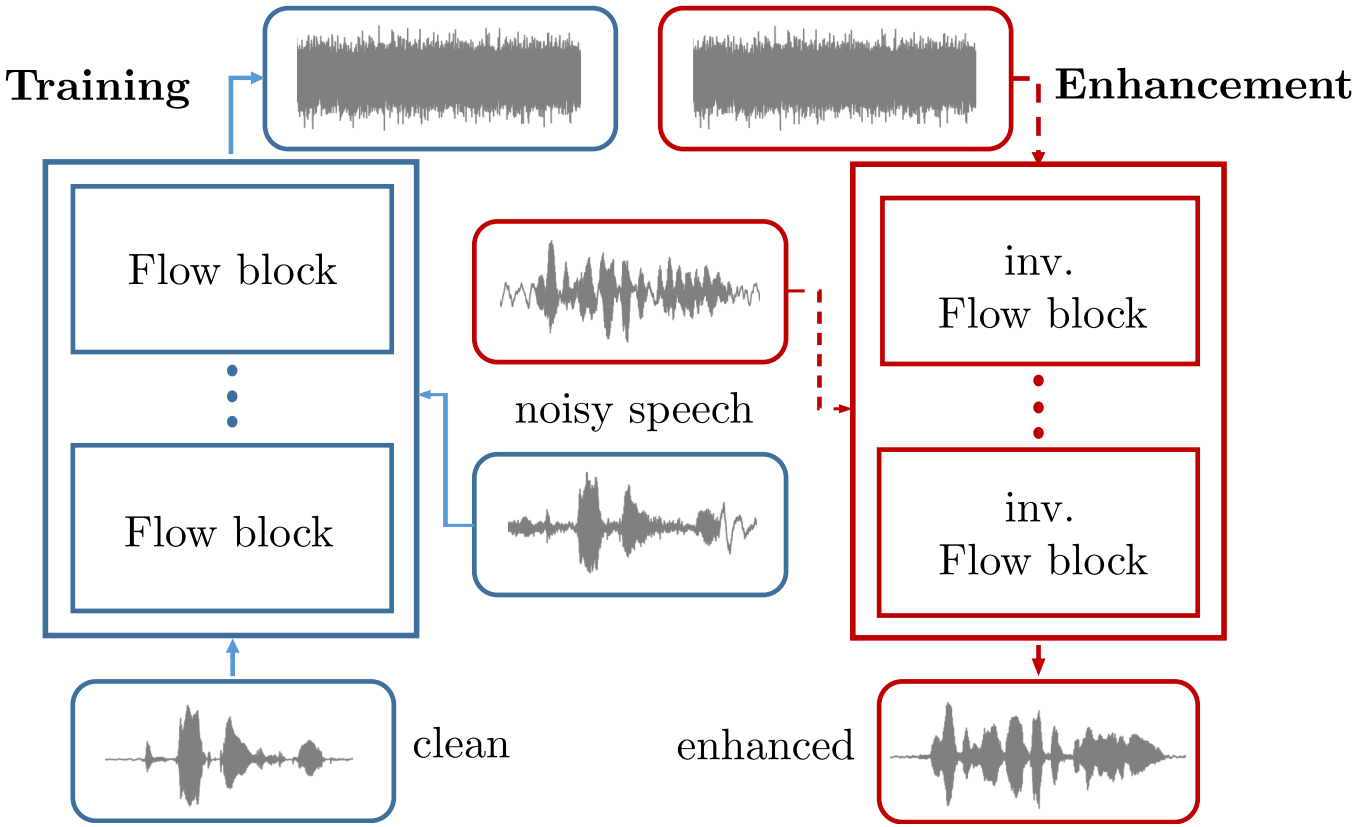
$$p_x(\mathbf{x}) = p_z(\mathbf{z}) \left|\det\left(J(\mathbf{x})\right)\right|$$

- Optimization of maximum likelihood

$\Rightarrow$ Integration of perceptual loss criteria difficult

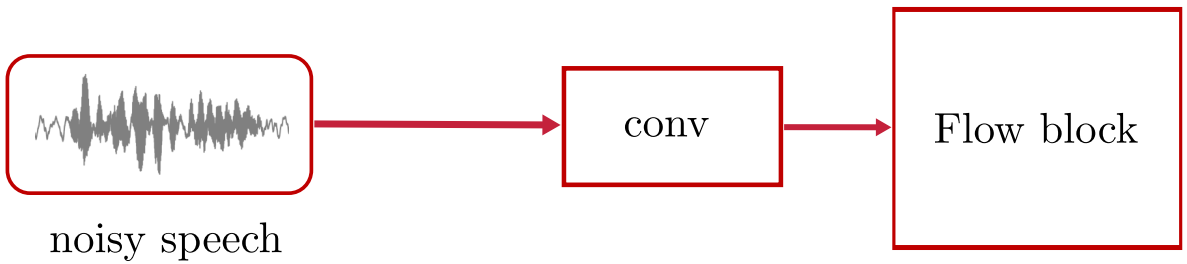
Normalizing flows

Architecture [5]



Conditional input
Time domain

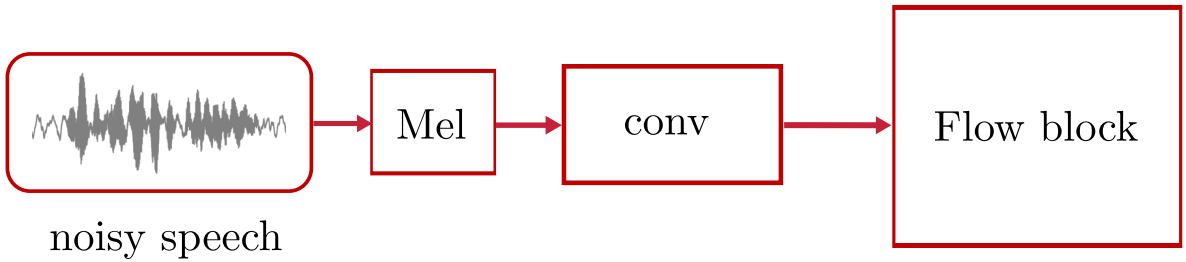
Time domain input processed by a conv layer



- Initial experiments show promising results
- Approach for further improvement: perceptually motivated pre-processing

Conditional input
Mel spectrogram

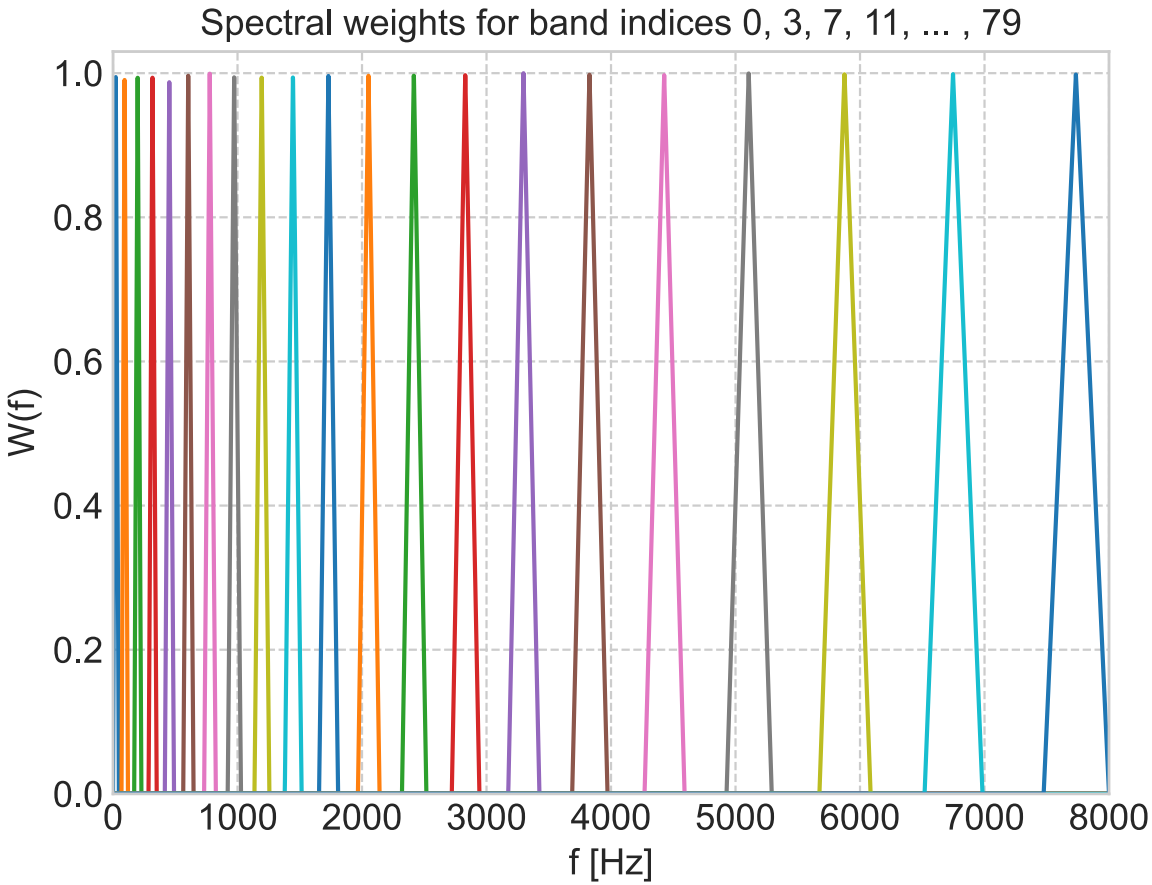
- Frequent choice or conditional input representation: Mel spectrogram (e.g. in neural vocoders [4, 3])



- Resulting quality ⇓
- Possible reason: Low time resolution due to weighted summation of STFT output magnitudes

Conditional input

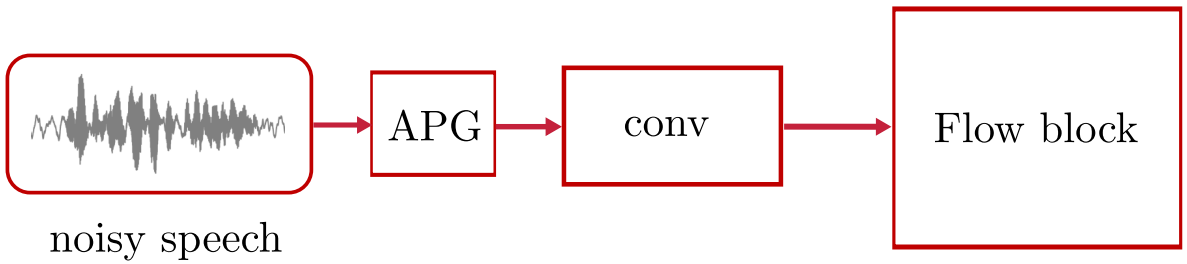
Mel spectrogram



Conditional input

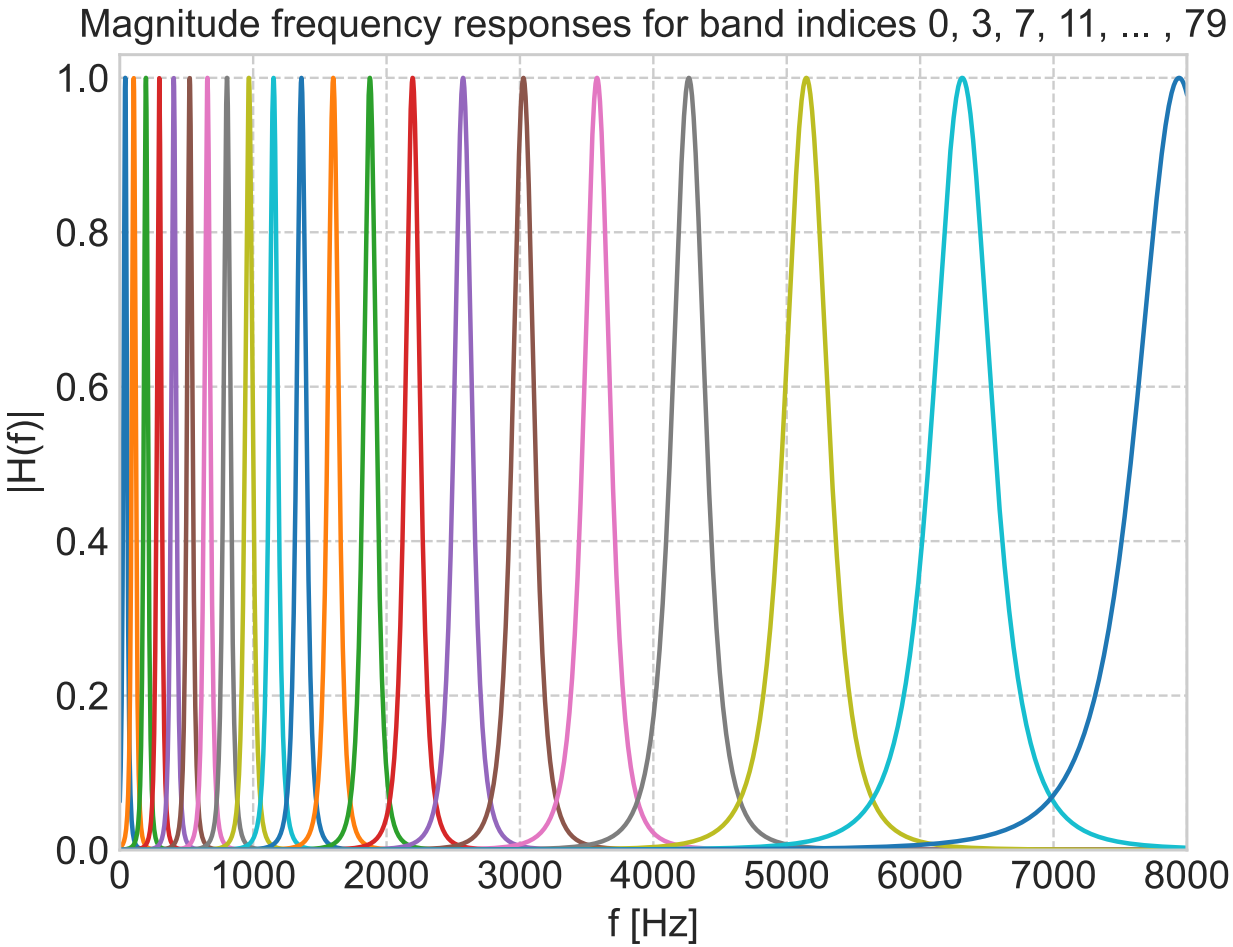
All-pole gammatone filterbank (APG) [6]

- Bark spaced all-pole gammatone filterbank (APG)
- Conditional input representation: output magnitudes of complex valued IIR filters



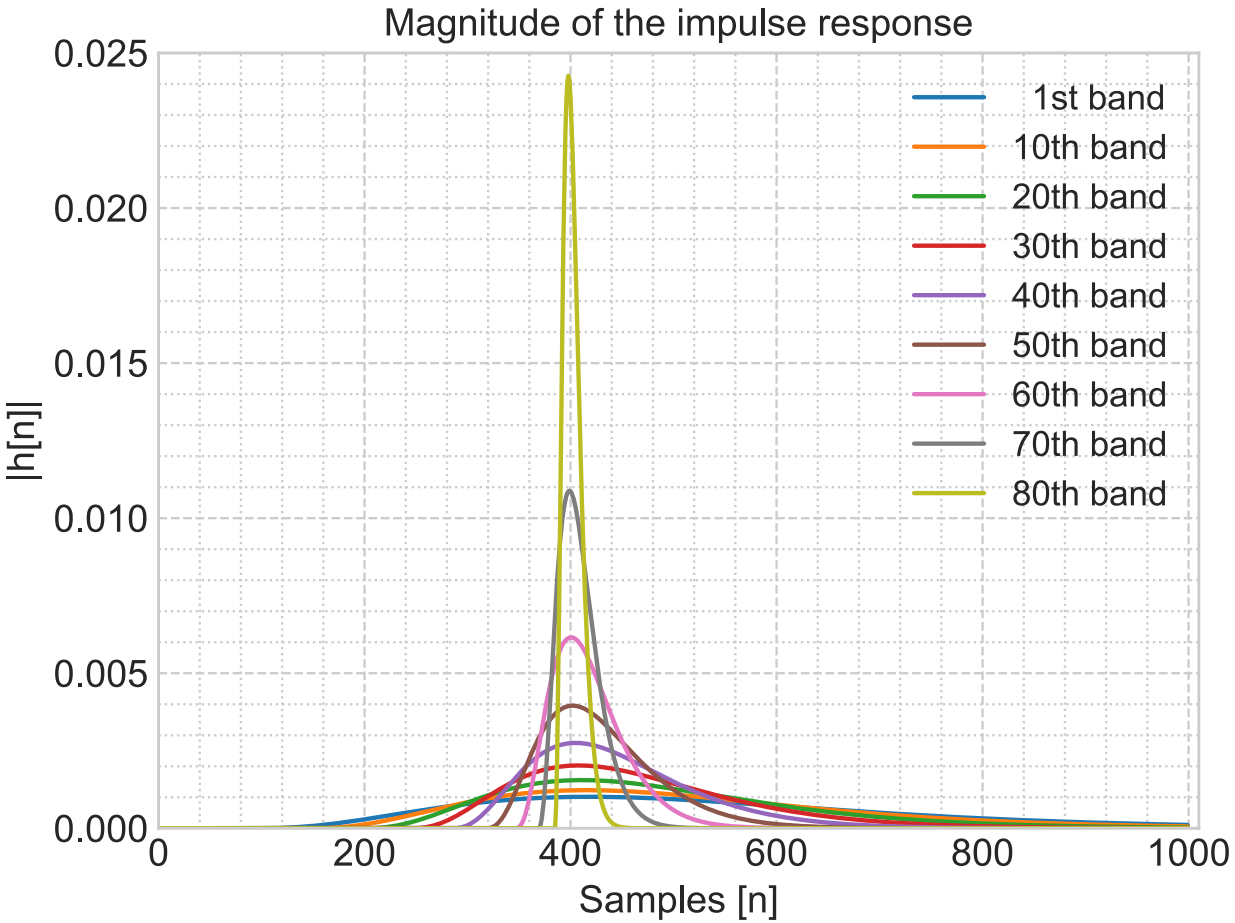
Conditional input

APG



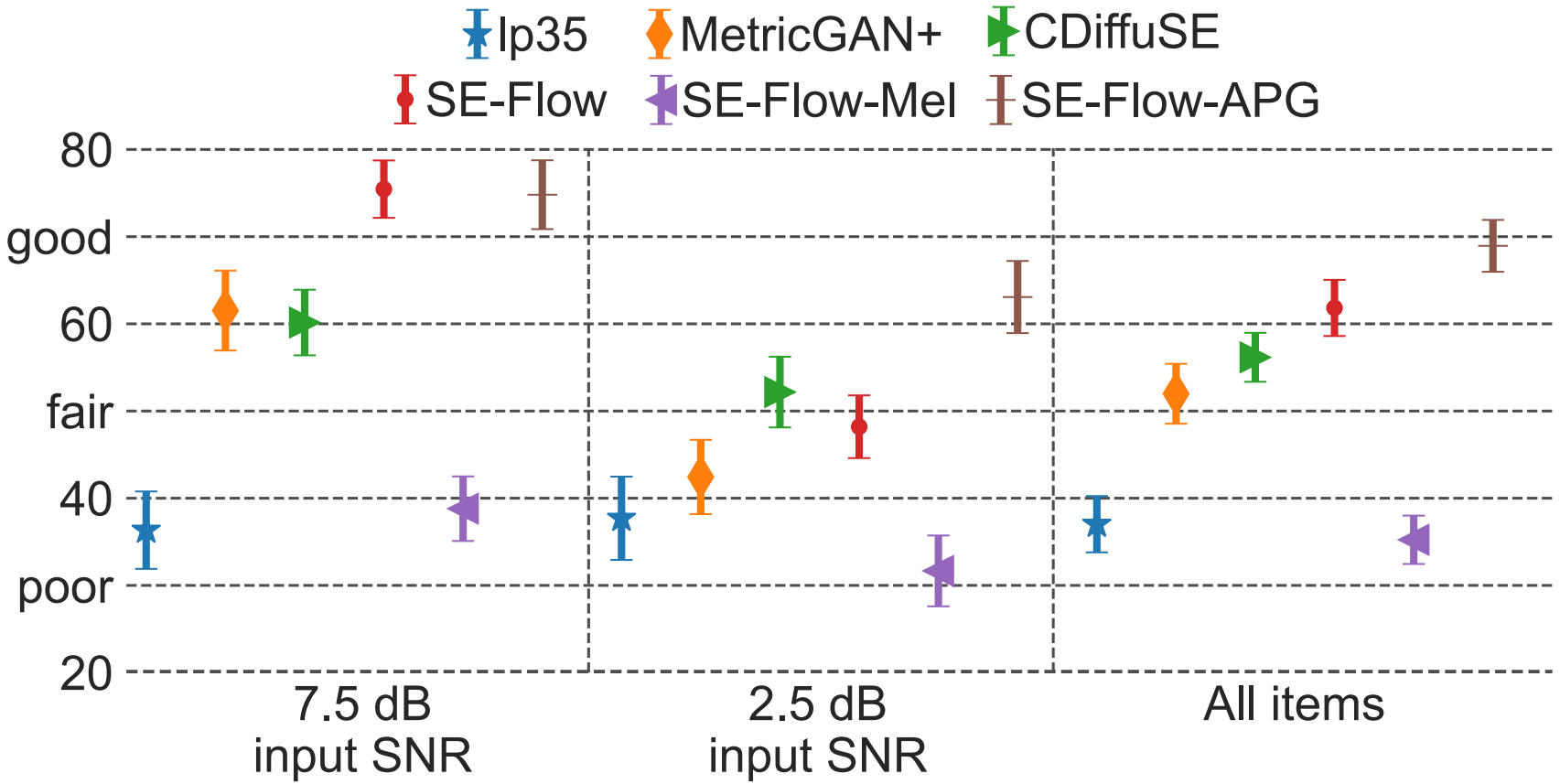
Conditional input

APG



Results

Listening test [6]



Results

Computational metrics [6]

Table: Computational evaluation results for the test set (VoiceBank-DEMAND). Mean values. Best results in bold.

Method	PESQ	CSIG	CBAK	COVL	STOI	2f-model
Noisy	1.97	3.35	2.45	2.63	0.92	31.70
CDiffuSE [2]	2.52	3.72	2.91	3.10	0.91	33.65
MetricGAN+ [1]	**3.13**	**4.08**	**3.16**	**3.60**	**0.93**	34.36
SE-Flow	2.41	3.79	3.11	3.09	**0.93**	**46.92**
SE-Flow-Mel	1.63	2.84	2.00	2.19	0.85	35.63
SE-Flow-APG	2.05	3.30	2.51	2.65	0.89	40.16

Listening examples

Noisy

Reference

SE-Flow

SE-Flow-Mel

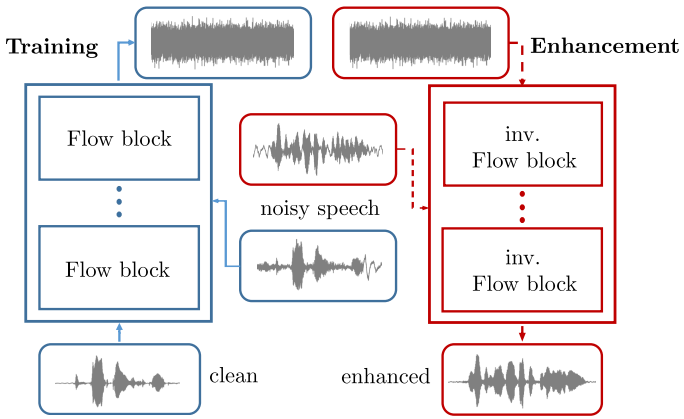
SE-Flow-APG

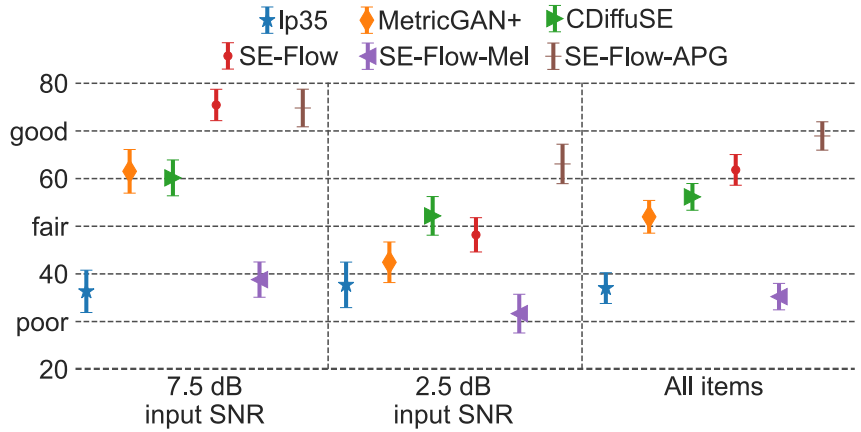
ed from the VoiceBank-DEMAND testset [7]

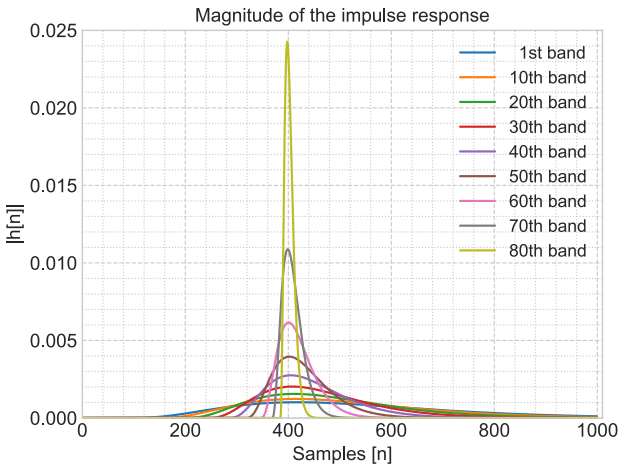
Conclusions

- SE-Flow and SE-Flow-APG show favourable perceptual performance
- SE-Flow-APG overcomes Mel induced problems
- SE-Flow-APG remains stable in lower SNR
- Listening test results not reflected in computational metrics

Thank you!







Method	PESQ	CSIG	CBAK	COVL	STOI	2f-model
Noisy	1.97	3.35	2.45	2.63	0.92	31.70
CDiffuSE [2]	2.52	3.72	2.91	3.10	0.91	33.65
MetricGAN+ [1]	**3.13**	**4.08**	**3.16**	**3.60**	**0.93**	34.36
SE-Flow	2.41	3.79	3.11	3.09	**0.93**	**46.92**
SE-Flow-Mel	1.63	2.84	2.00	2.19	0.85	35.63
SE-Flow-APG	2.05	3.30	2.51	2.65	0.89	40.16

References I

[1] S.-W. Fu, C. Yu, et al., *MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement*, in Proc. Interspeech Conf., 2021, pp. 201–205.

[2] Y.-J. Lu, Z.-Q. Wang, et al., *Conditional Diffusion Probabilistic Model for Speech Enhancement*, in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7402–7406.

References II

[3] A. Mustafa, N. Pia, and G. Fuchs, *StyleMelGAN: An Efficient High-Fidelity Adversarial Vocoder with Temporal Adaptive Normalization*, in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6034–6038.

[4] R. Prenger, R. Valle, and B. Catanzaro, *Waveglow: A Flow-based Generative Network for Speech Synthesis*, in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 3617–3621.

References III

[5] M. Strauss and B. Edler, *A flow-based neural network for time domain speech enhancement*, in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 5754–5758.

[6] M. Strauss, M. Torcoli, and B. Edler, *Improved normalizing flow-based speech enhancement using an all-pole gammatone filterbank for conditional input representation*. accepted at IEEE Spoken Language Technology Workshop (SLT), 2023.

References IV

[7] C. Valentini-Botinhao, X. Wang, et al., *Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks*, 2016, pp. 352–356.