Neural Homomorphic Vocoder

Zhijun LIU

zhijunliu@sjtu.edu.cn

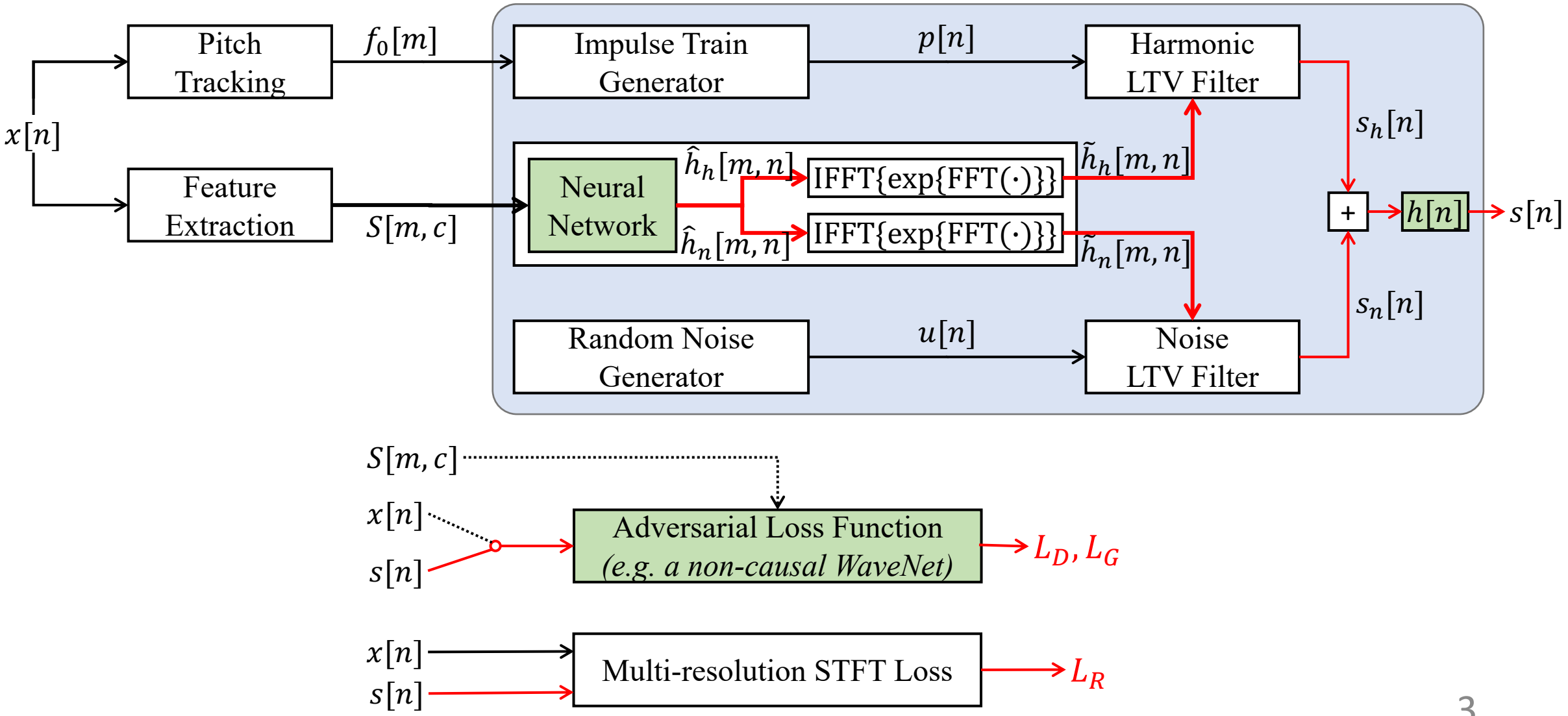
@SpeechLab

Shanghai Jiao Tong University, China

Content of this talk

- Review and compare NHV with related works
- Phase structure in Speech and Phase Perception
- A walkthrough of NHV
- Experiments and Results

Overview of Neural Homomorphic Vocoder

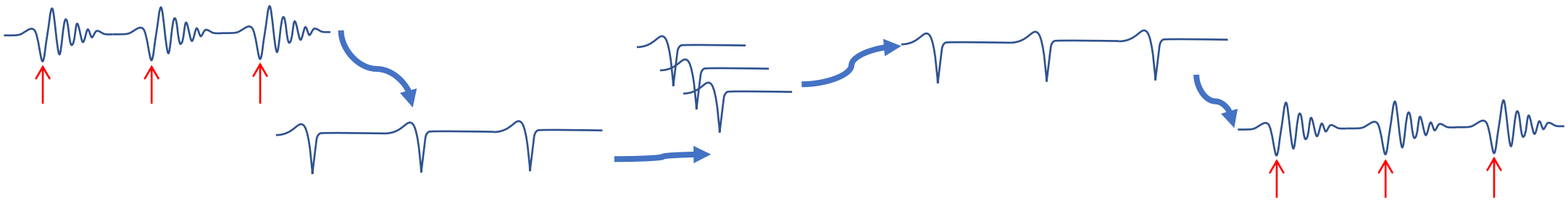


Pitch Synchronous Excitation Models

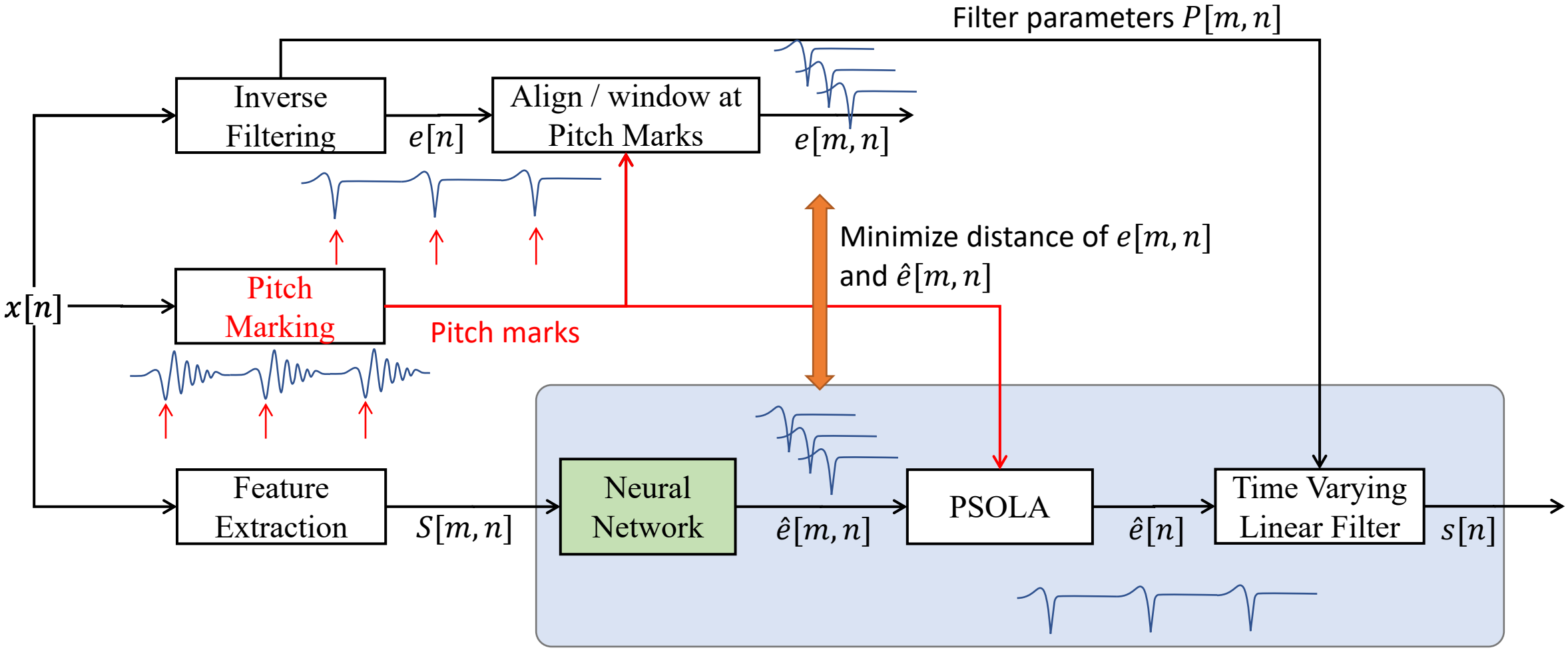
- Many variants under the same framework:
  - *M. Airaksinen,* GlottDNN -- A Full-Band Glottal Vocoder for Statistical Parametric Speech Synthesis.
  - *M. Hwang,* A Unified Framework for the Generation of Glottal Signals in Deep Learning-based Parametric Speech Synthesis Systems.
  - *L. Juvela,* Speech Waveform Synthesis from MFCC Sequences with Generative Adversarial Networks.
  - *L. Juvela,* Waveform Generation for Text-to-speech Synthesis Using Pitch-synchronous Multi-scale Generative Adversarial Networks.
- Recommend reading: *L. Juvela,* Neural waveform generation for source-filter vocoding in speech synthesis.
- All these models follow a similar pattern.

Pitch Synchronous Excitation Models

- All these methods follow the same pattern:
  - Speech is inverse filtered to obtain the time varying filter and residual signal.
  - Each period in residual is identified, for example, with estimated glottal closure instant.
  - Neural network models the excitation waveform.
  - Generated excitation put back together into full residual signal with PSOLA.
  - Residual signal is filtered by a time varying filter to obtain the full signal.

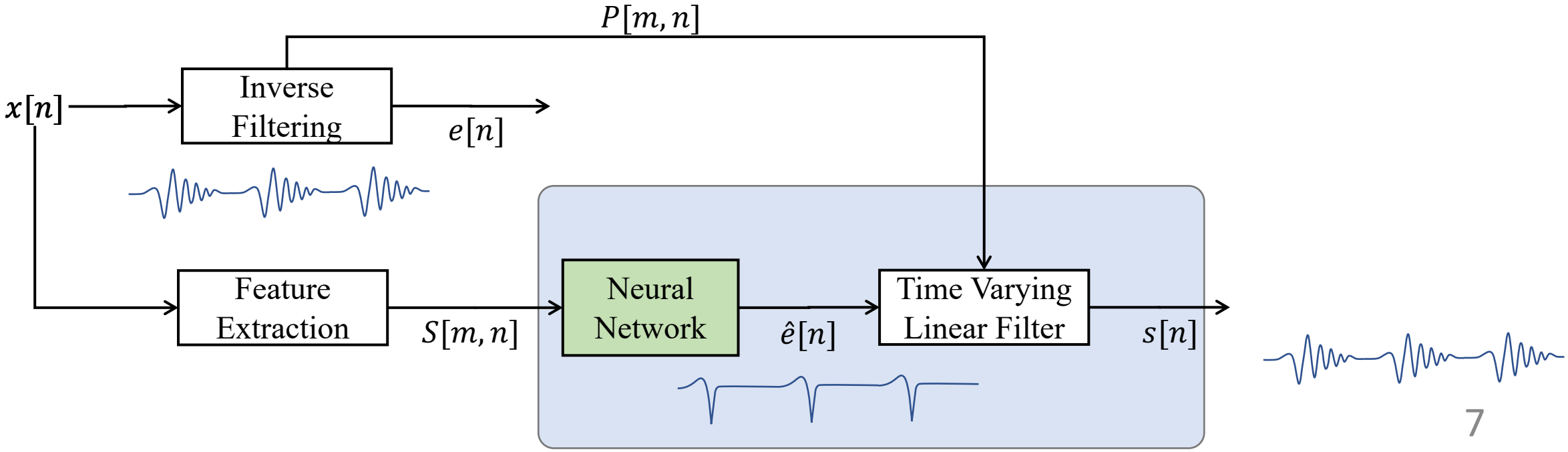


Pitch Synchronous Excitation Models



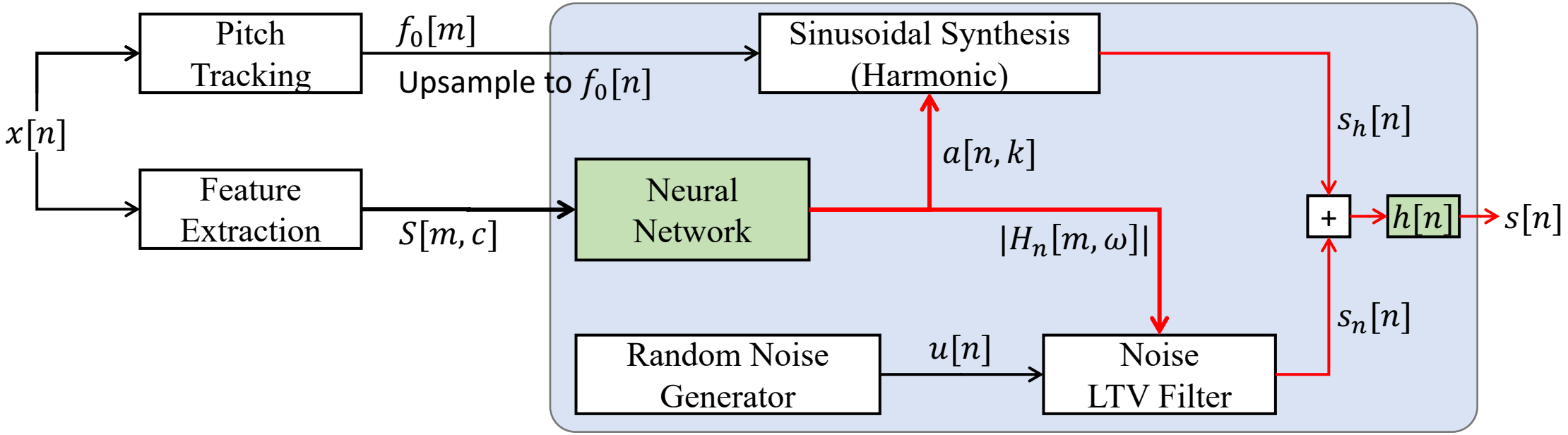
Pitch Asynchronous Excitation Models

- The asynchronous methods does not require pitch synchronous analysis. The neural network directly outputs the residual signal asynchronously.
- Autoregressive generation: GlotNet[L. Juvela 2019], LPCNet[J-M Valin 2019], ExcitNet[E. Song 2019]; Parallel generation: GELP[L. Juvela 2019].
- Asynchronous at the cost of much higher computational complexity.



DDSP

$$s_h[n] = \sum_{k=1}^{f_s/(2f_0)} a[n,k] \sin\left(2\pi \sum_{s=0}^{n} k f_0[s] + \phi_{0,k}\right)$$

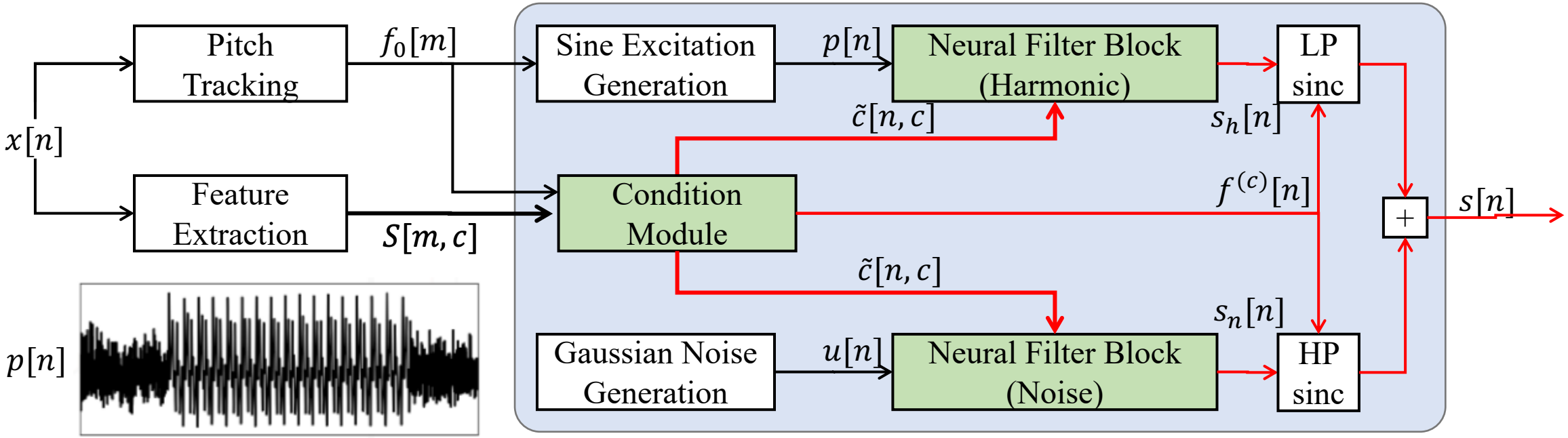


- In DDSP, a neural network controlled HNM trained with STFT Loss is proposed. However relative phase shift of sinusoids is not modeled (fixed or random) in DDSP.
- NHV can be viewed as a DDSP + Source/filter phase model.

Engel J. (2020) DDSP: Differentiable Digital Signal Processing.
Stylianou Y. (2005) Modeling Speech Based on Harmonic Plus Noise Models.
Quatieri, T. (2001). Discrete-Time Speech Signal Processing: Principles and Practice. p.460

Neural Source Filter Models

- There are several variants. We illustrate the structure of hn-sinc-NSF.

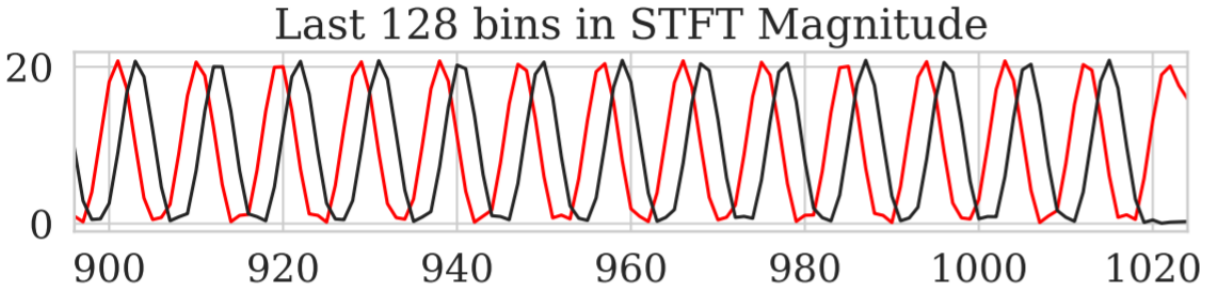


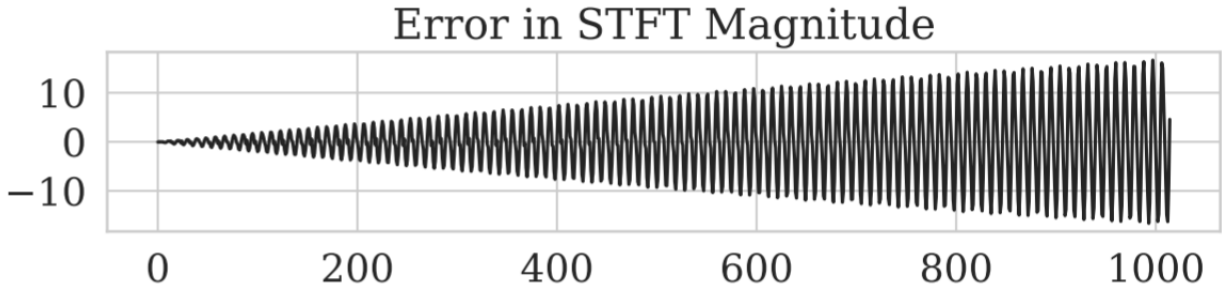
- Same as DDSP, NSF Models are all trained with STFT magnitude loss.

Wang, X., & Yamagishi, J. (2019). Neural Harmonic-plus-Noise Waveform Model with Trainable Maximum Voice Frequency for Text-to-Speech Synthesis.

Problems with multi-resolution STFT Loss

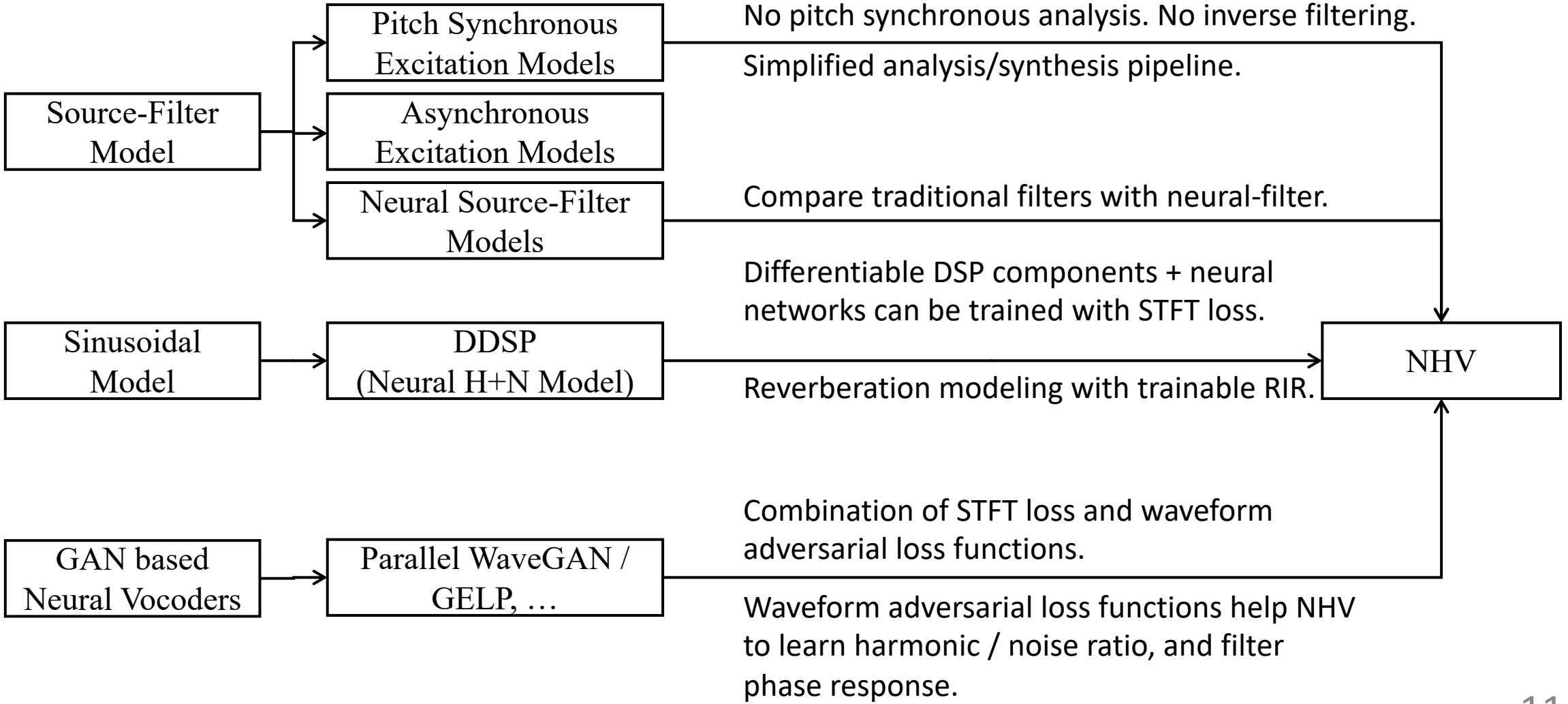
- Pitch error sensitive: harmonic mismatch occurs even with small error in pitch estimation. (STFT magnitude of 200 Hz and 200.5 Hz impulse train is plotted.)





- Phase error tolerant: Although theoretically, STFT magnitude encodes all phase information. In practice, STFT Loss failed to guarantee correct relative phase shifts.

Related works



Why should we care about phase?

Phase structure in speech

- Consider the ideal model of vowel production, where an constant period impulse train $u(t)$ is filtered by an constant linear filter $h(t)$. The output signal is $s(t) = u(t) * h(t)$.

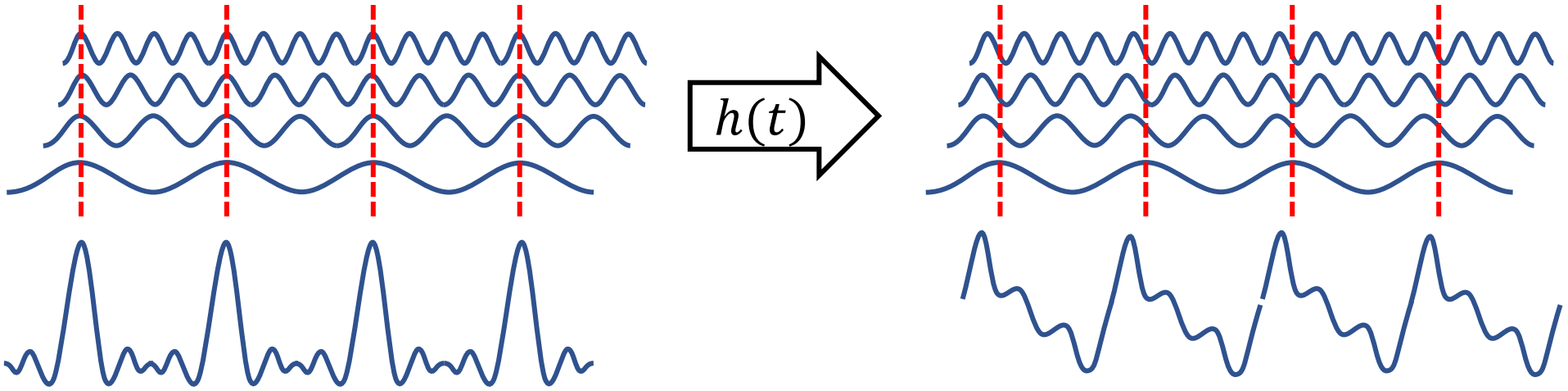
$$u(t) = \mathrm{Re} \sum_{k=1}^{K} \exp\left[jtk\Omega_0\right] \qquad H(\Omega) = M(\Omega)\exp[j\Phi(\Omega)]$$

$$s(t) = \mathrm{Re} \sum_{k=1}^{K} M\left(k\Omega_o\right) \exp\left[j\left(k\Omega_o t + \Phi\left[k\Omega_o\right]\right)\right]$$

- The output consists of sinusoids keeping the original angular velocity, but different phase shift.

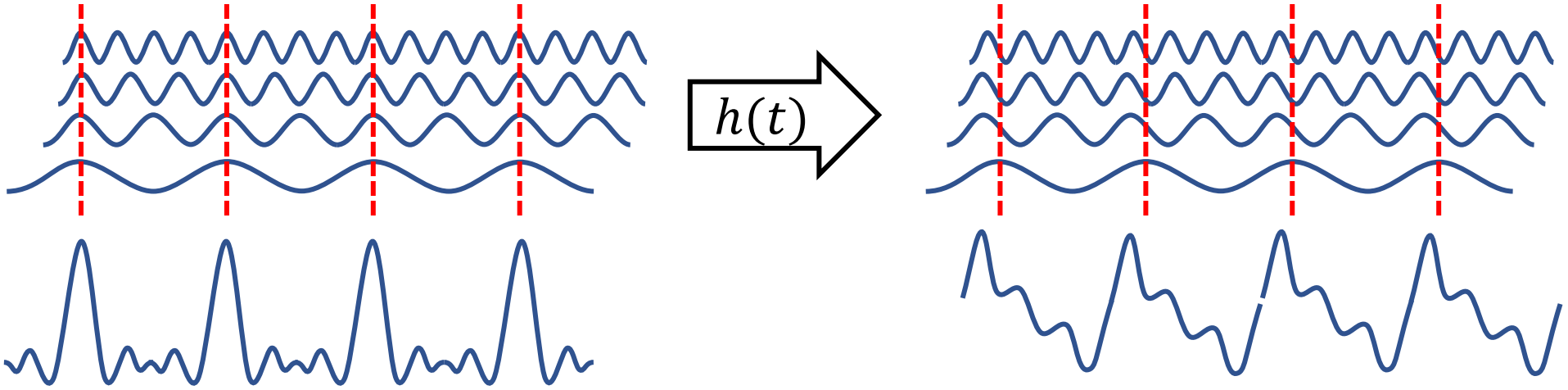
Phase structure in speech: Relative phase shift

- The shift of sinusoids causes the change in waveform shape in each period.
- Suppose the phase of k th sinusoid at time t is $\varphi_k(t)$. A simple derivation shows that the value $\varphi_k(t) - k\varphi_1(t)$ is a constant. We call this value the relative phase shift.



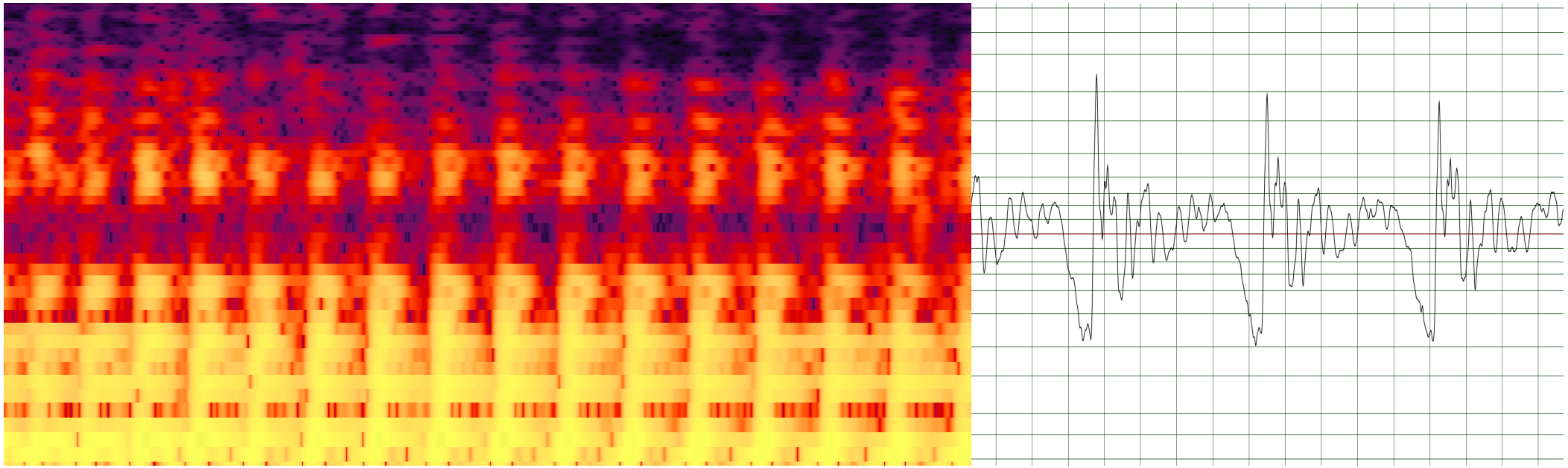
Phase structure in speech: Relative phase shift

- The phase can be split into two component: the phase of the base sinusoid (determines the pitch), and the relative phase shifts.
- A vocoder system should reconstruct both.
  - In classical vocoders for SPSS, minimum phase assumption of filters is often adopted, due to the difficulty of modeling phase.
  - However, realistic reconstruction requires mixed phase.



Phase structure in speech: Relative phase shift

- The relative phase shift can be observed in many ways.
  - The simplest way is to observe the shape of speech waveform.
  - Phase response also influence the distribution of energy in different frequency in each period, which can be observed on broadband spectrogram.



Phase structure in speech

- The relative phase shift can be observed in many ways.
  - The simplest way is to observe the shape of speech waveform.
  - Phase response also influence the distribution of energy in different frequency in each period, which can be observed via broadband spectrogram.
- There are many methods for phase representation
  - Recommended reading: P. Mowlaee (2016). Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice.
  - See Chapter 02. Code can be found online.

Phase perception: Are we phase deaf?

- We can not hear some types of phase changes, such as polarity inversion.
- But we can hear relative phase shifts to a certain degree.
- Auditory processing mechanisms offers possible explanations.

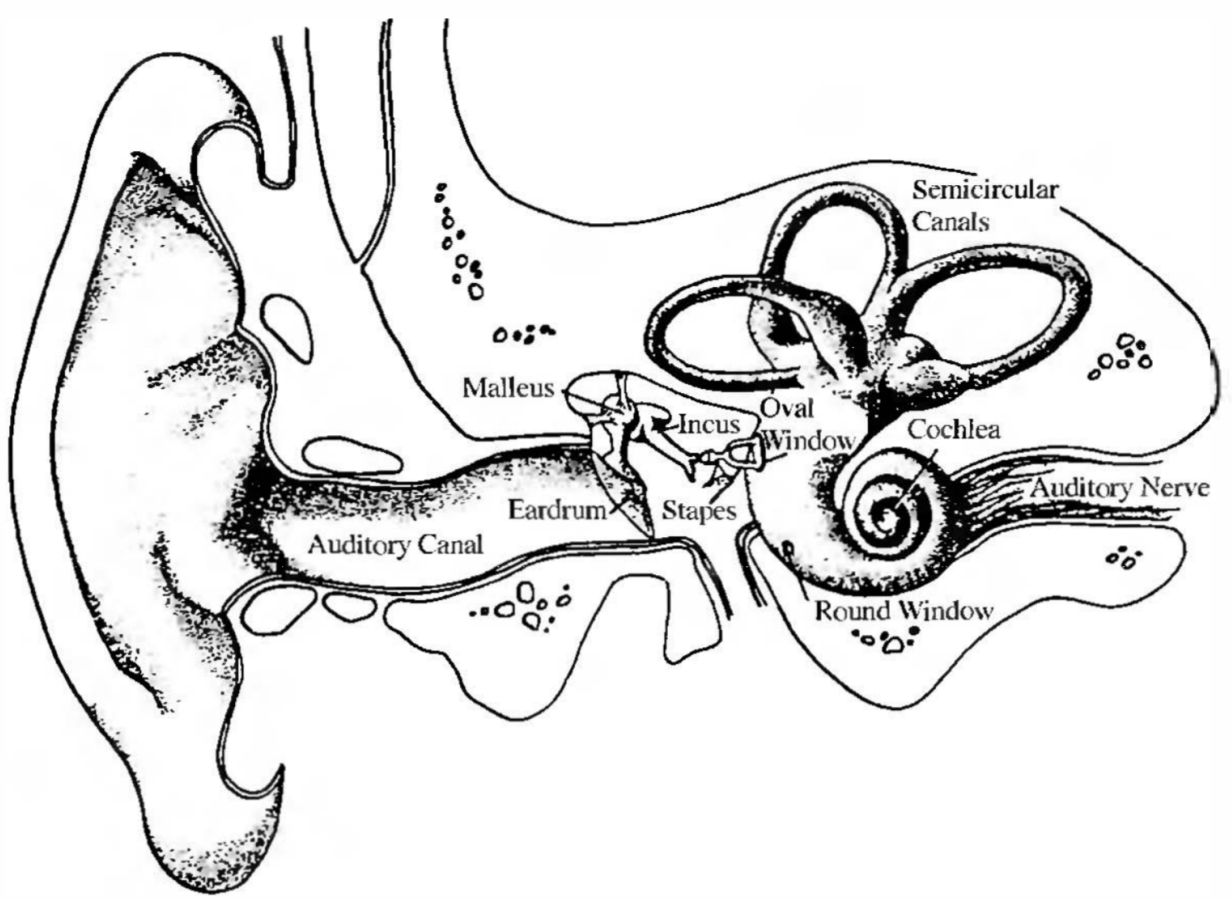
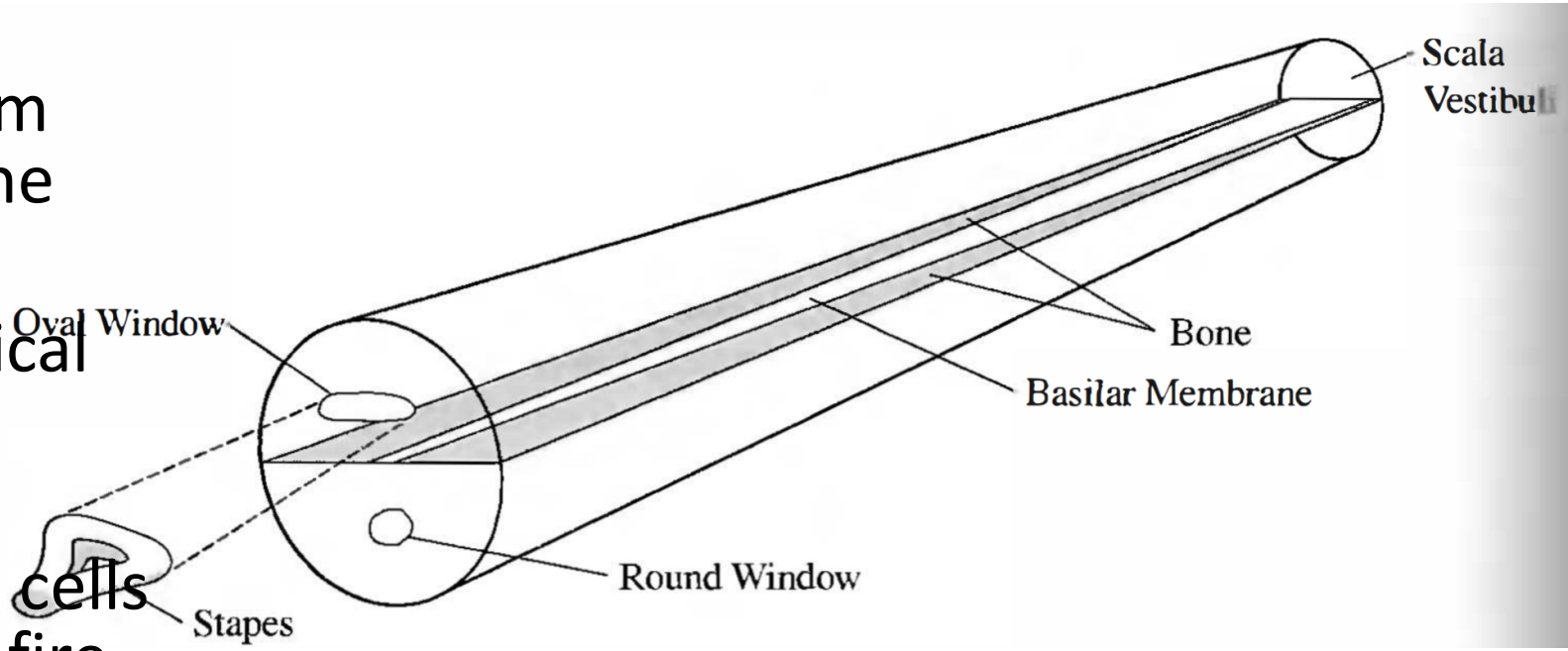


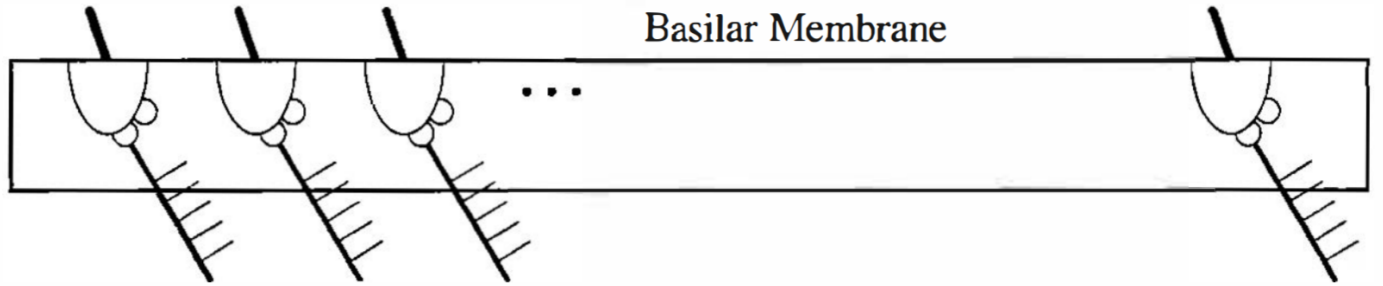
Figure 8.24 Primary anatomical components of the peripheral auditory system are the outer, middle, and inner ear.

Quatieri, T. (2001). Discrete-Time Speech Signal Processing: Principles and Practice. Chapter 8.

Phase perception: Are we phase deaf?

- Vibrations of the eardrum result in movement of the oval window.
- This wave, causes a vertical vibration of the basilar membrane.
- The vibration causes the cells on basilar membrane to fire.
- Neural impulses are conducted to the auditory nerves.

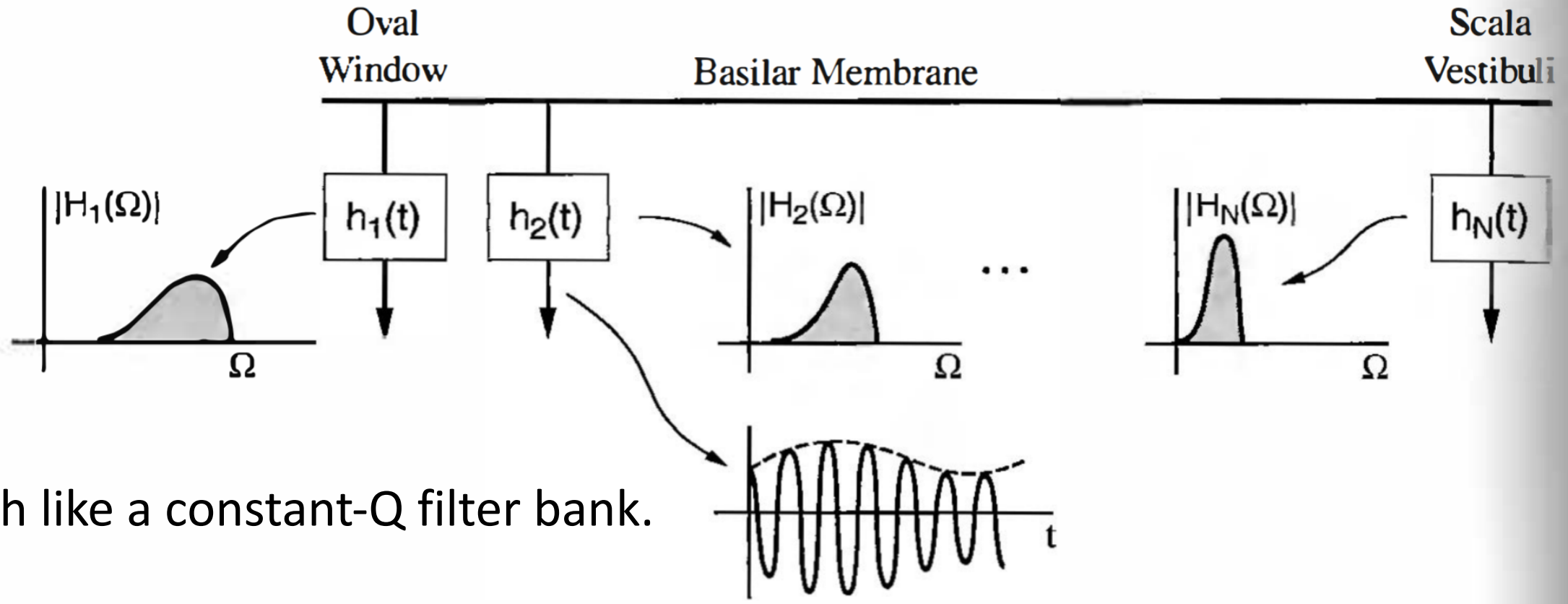




Quatieri, T. (2001). Discrete-Time Speech Signal Processing: Principles and Practice. Chapter 8.

Cochlear Filters

- Different regions on the basilar membrane acts like a filter bank.

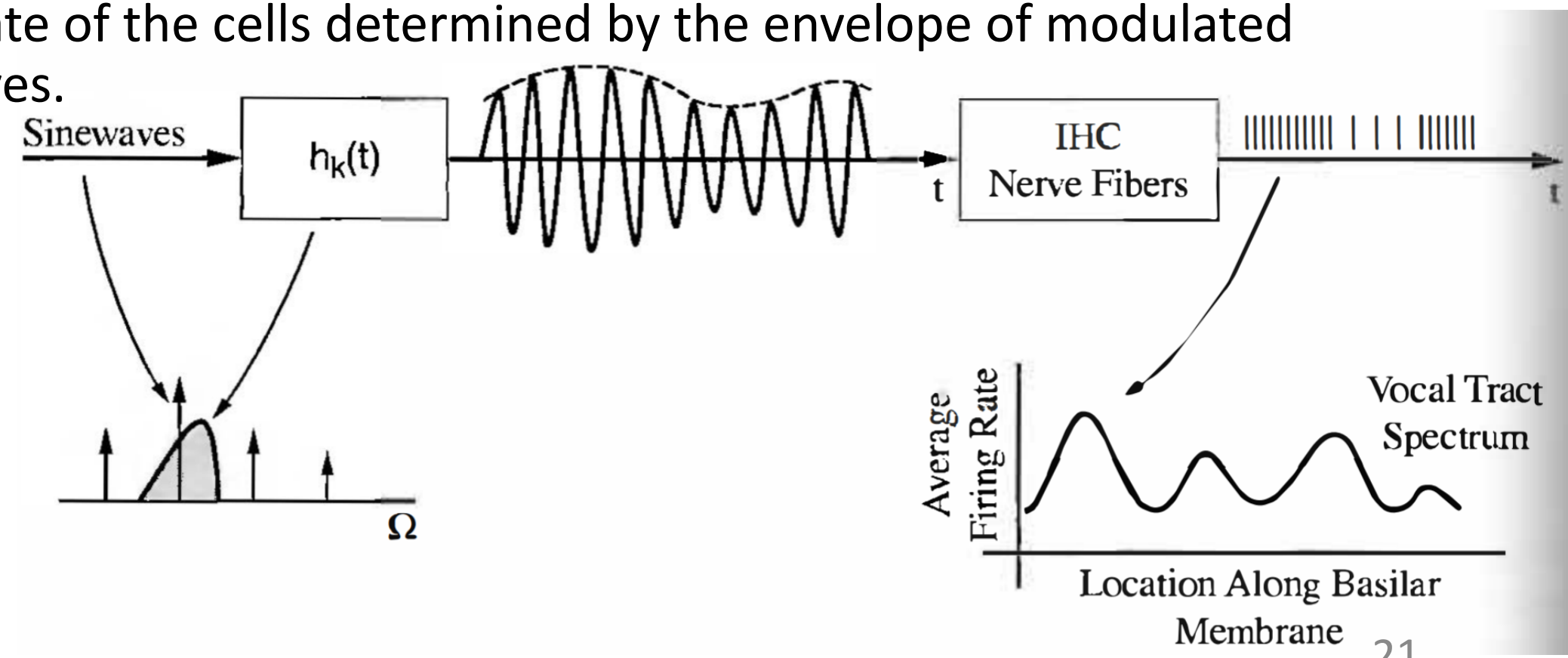


- Much like a constant-Q filter bank.

Quatieri, T. (2001). Discrete-Time Speech Signal Processing: Principles and Practice. Chapter 8.

AM Sinewaves from Cochlear Filter

- Filtered speech waveform forms modulated sine.
- Firing rate of the cells determined by the envelope of modulated sinewaves.



Quatieri, T. (2001). Discrete-Time Speech Signal Processing: Principles and Practice. Chapter 8.

Interference between sinusoids

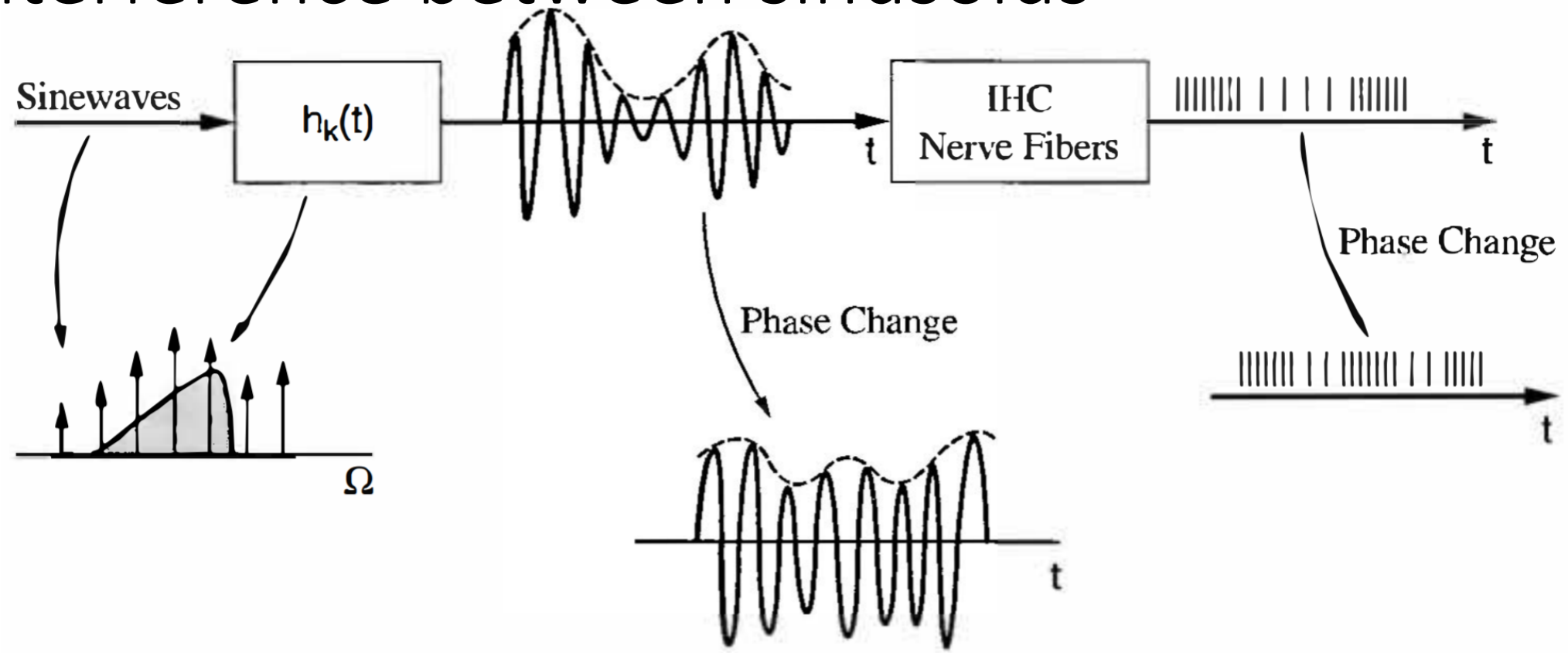
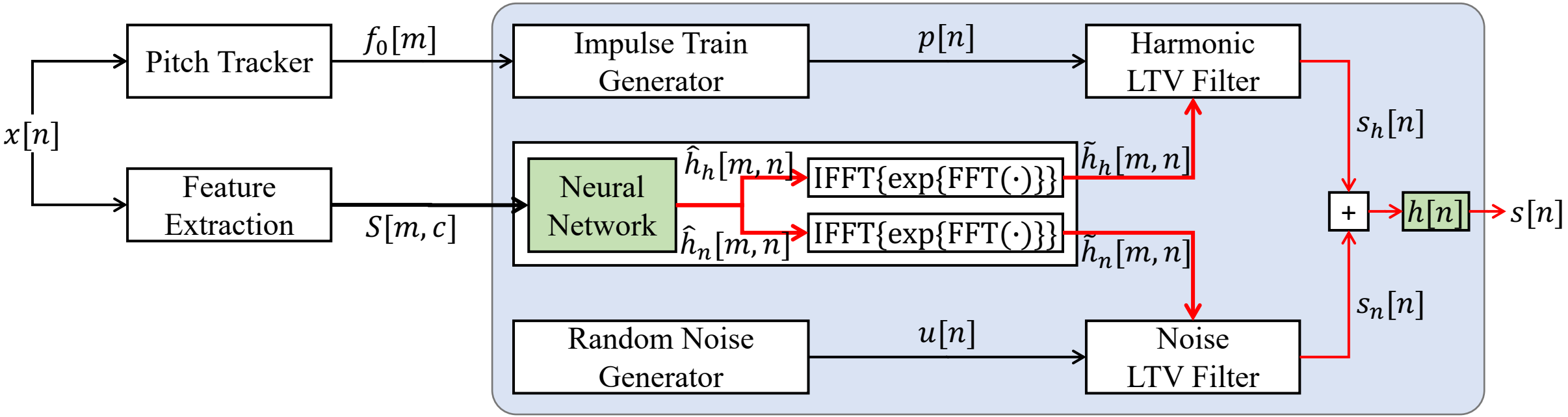


Figure 8.27 Auditory processing of a multiple of slowly varying AM-FM sinewaves as input to a cochlear filter, likely to occur with low pitch and at cochlear filters of high characteristic frequency. In this case, change in the phase relations of the input can alter the envelope shape and firing patterns of inner hair cell (IHC) nerve fibers, and thus, perhaps, perception of the sound.

Quatieri, T. (2001). Discrete-Time Speech Signal Processing: Principles and Practice. Chapter 8.

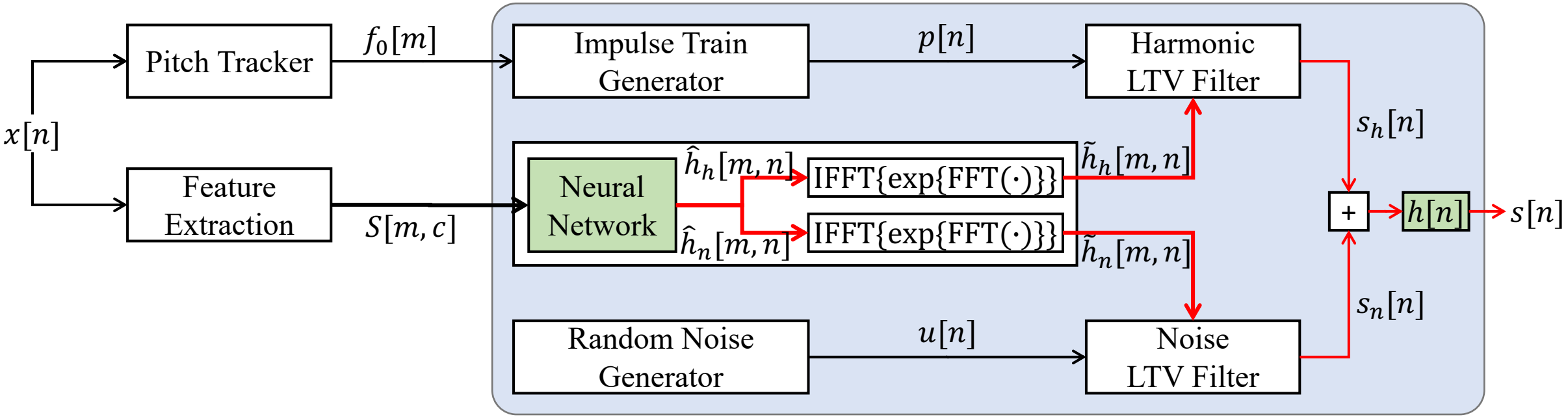
A Walkthrough of NHV

Neural Homomorphic Vocoder: Generation



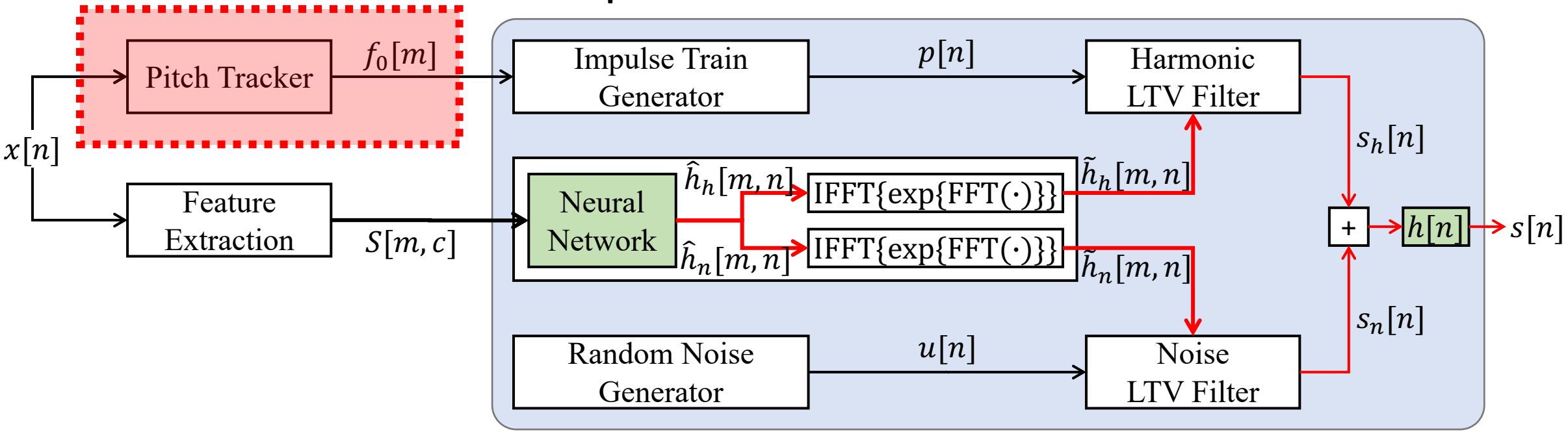
- n stands for the discrete time index. The range of n is different in different signals.
  - In waveform signals x, p, u, s_h, s_n, $0 \le n < N$, where N is the total number of samples.
  - In complex cepstrums $\hat{h}_h, \hat{h}_n$ and impulse responses$\tilde{h}_h, \tilde{h}_n$, n is in a short range symmetric about zero, e.g. $-512 \le n < 512$, when FFT size is 1024.
  - In causal finite impulse response $h[n]$, $0 \le n < N_h$, where N_h is the FIR length.

Neural Homomorphic Vocoder: Generation

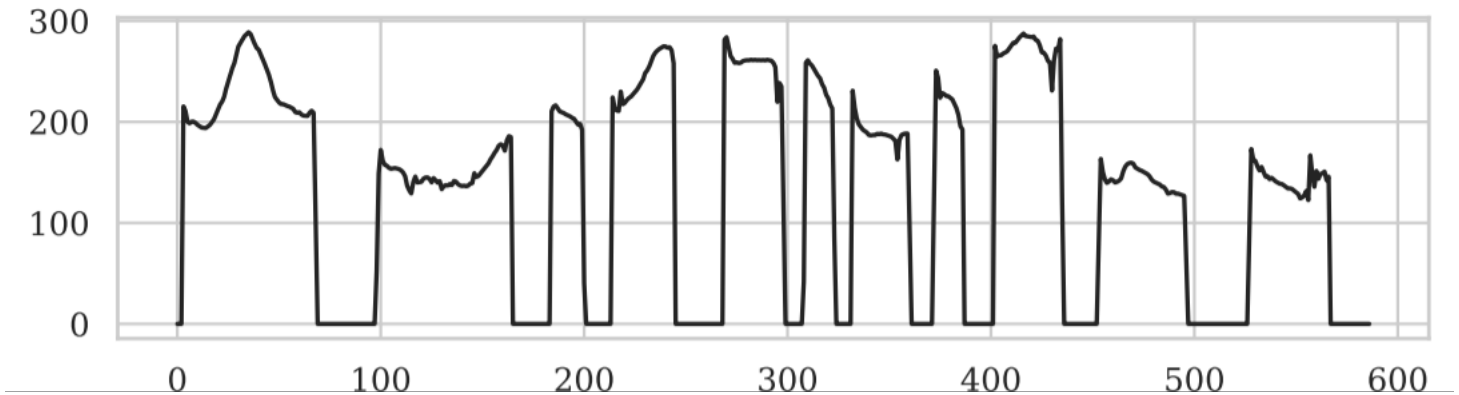


- m stands for the frame index. $0 \leq m < M$, where M is the total number of frames.
- The total number of samples and the total number of frames are related by $N = M \times L$, where L is the frame length.
- c stands for the index of the feature dimension. $0 \leq c < C$.

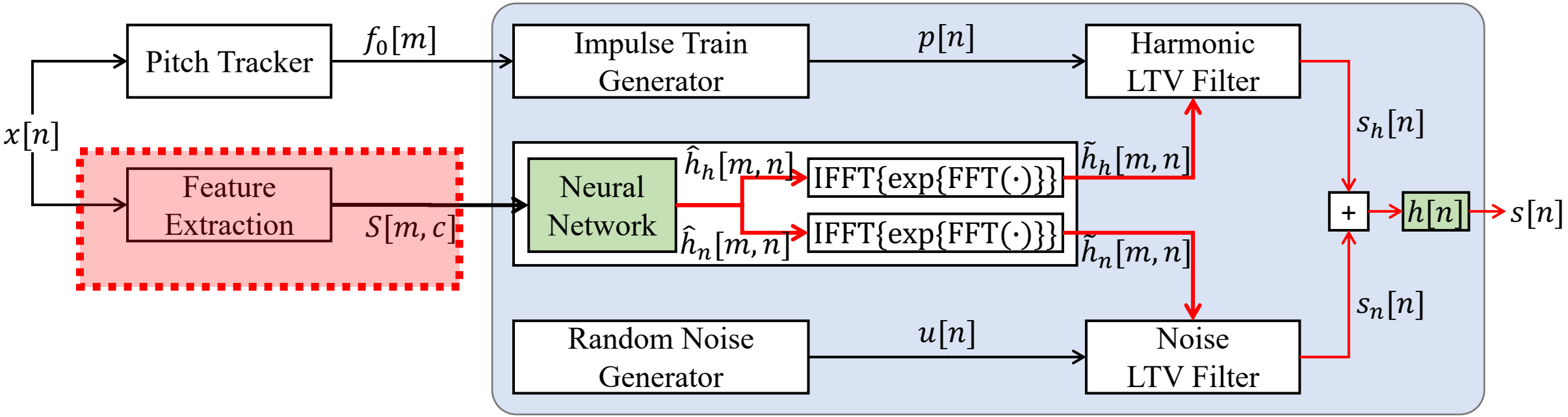
Neural Homomorphic Vocoder: Generation



$f_0[m]$

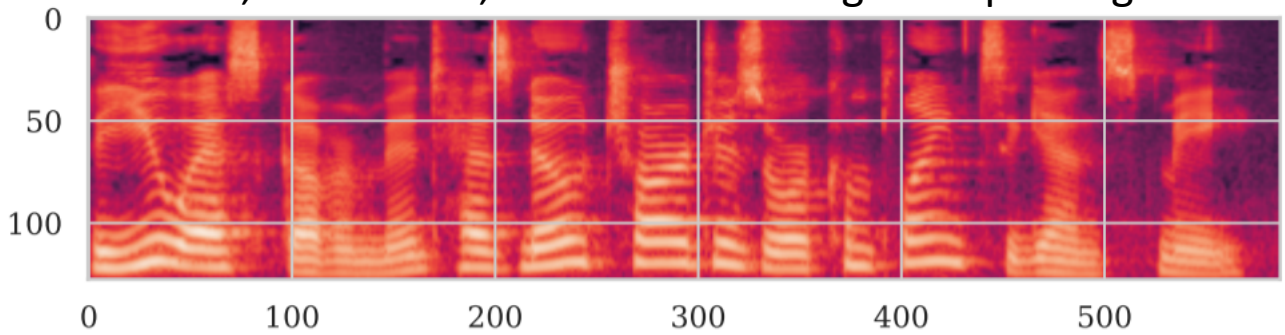


Neural Homomorphic Vocoder: Generation

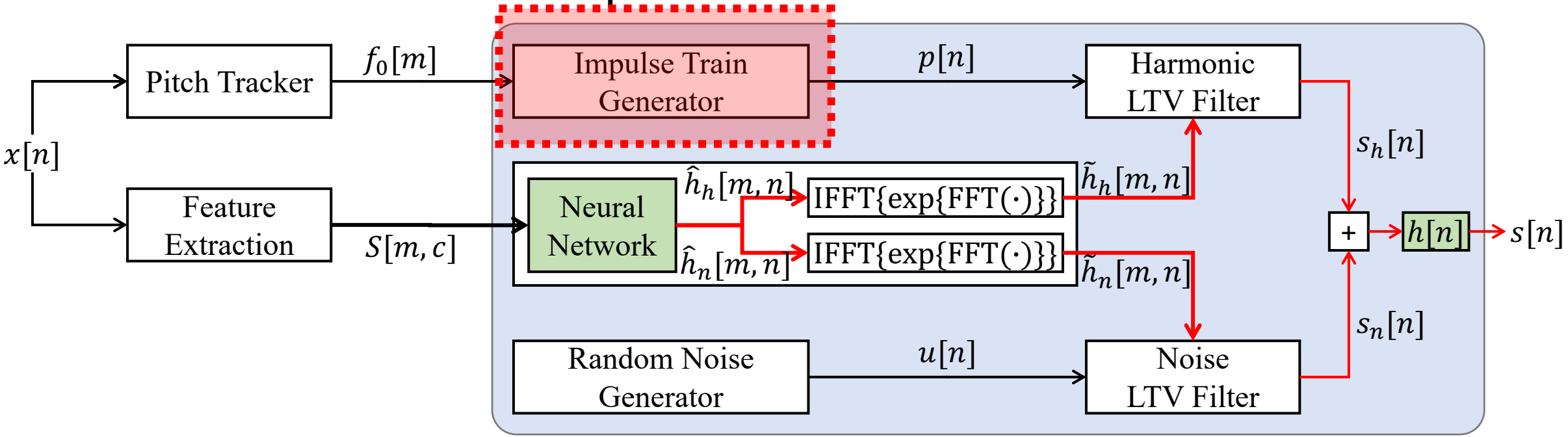


The acoustic feature is also extracted in each frame, in this case, the feature is a log Mel Spectrogram.

$S[m,c]$



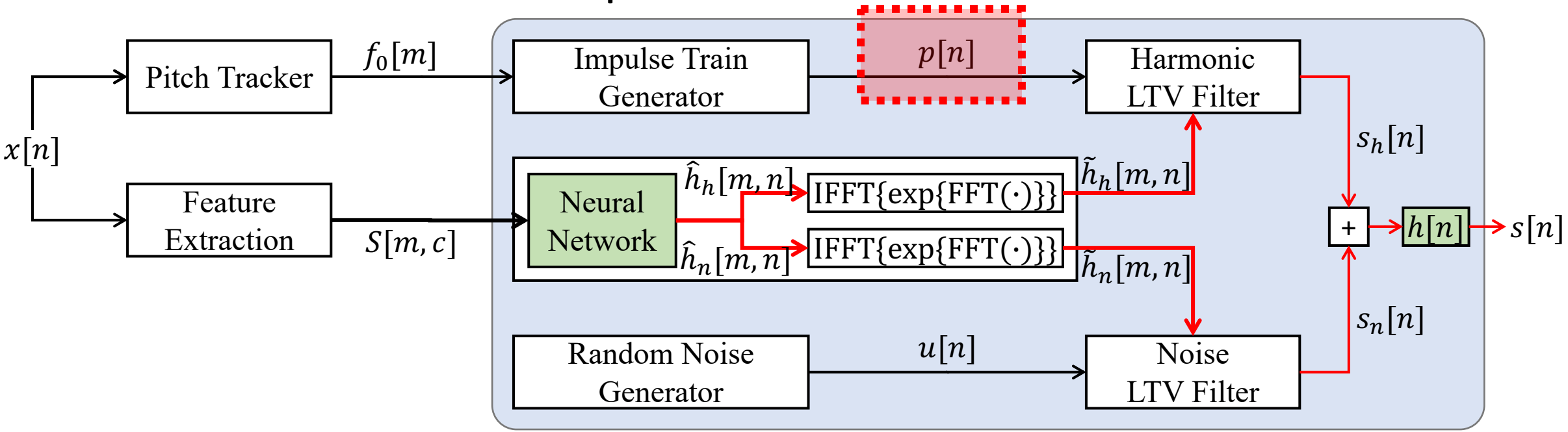
Neural Homomorphic Vocoder: Generation

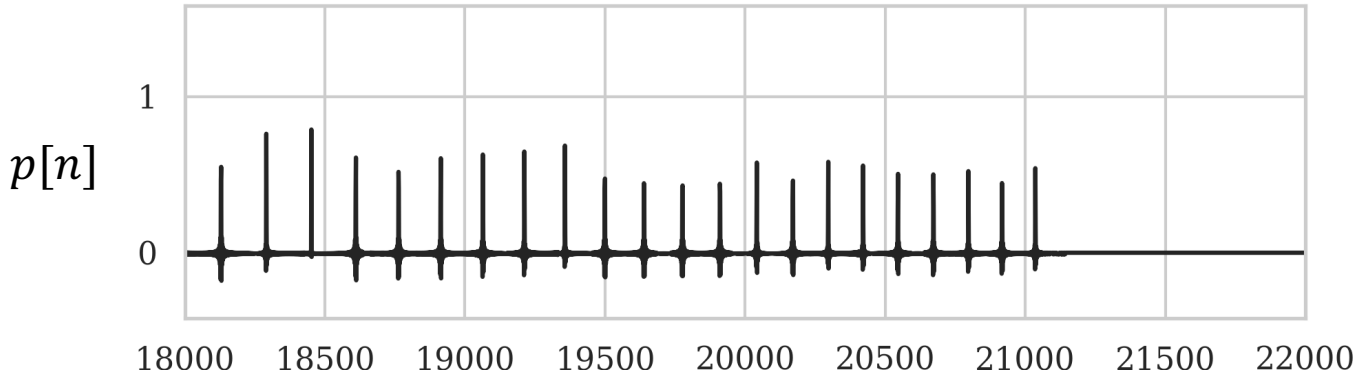


The model generates the impulse train according to the input f0. The alias free low-passed impulse train is synthesized by adding cosines together, according to the given formula. Cosines with frequencies above half the sampling rate are abandoned.

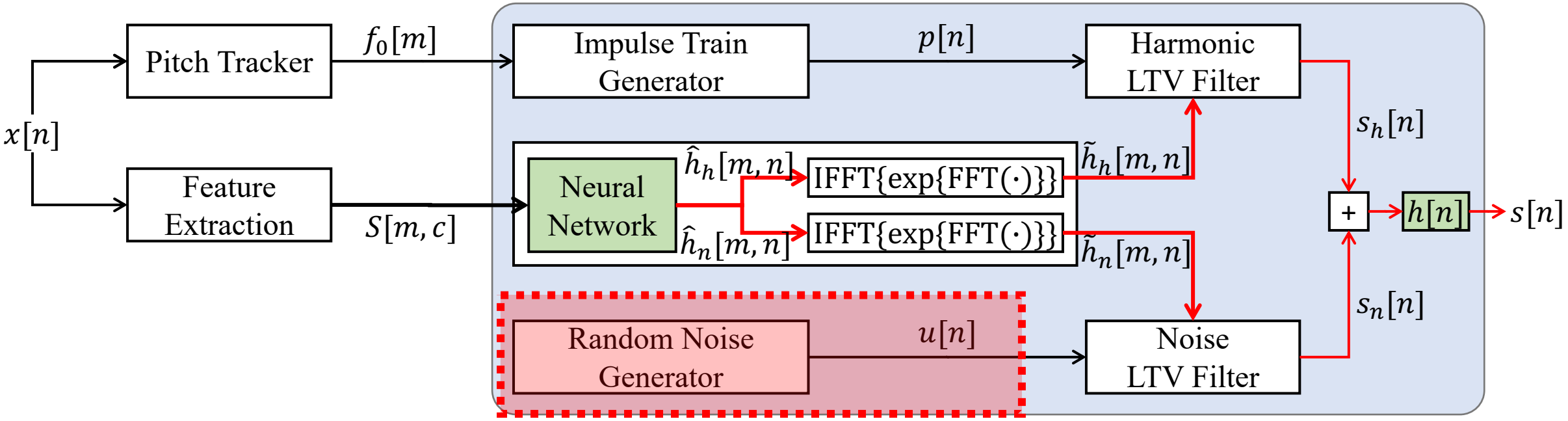
$$p[n] = \sum_{k=1}^{f_s/(2f_0)} \mathbf{1}(f_0[n] > 0) \cos\left(2\pi \sum_{s=0}^{n} k f_0[s]\right)$$

Neural Homomorphic Vocoder: Generation

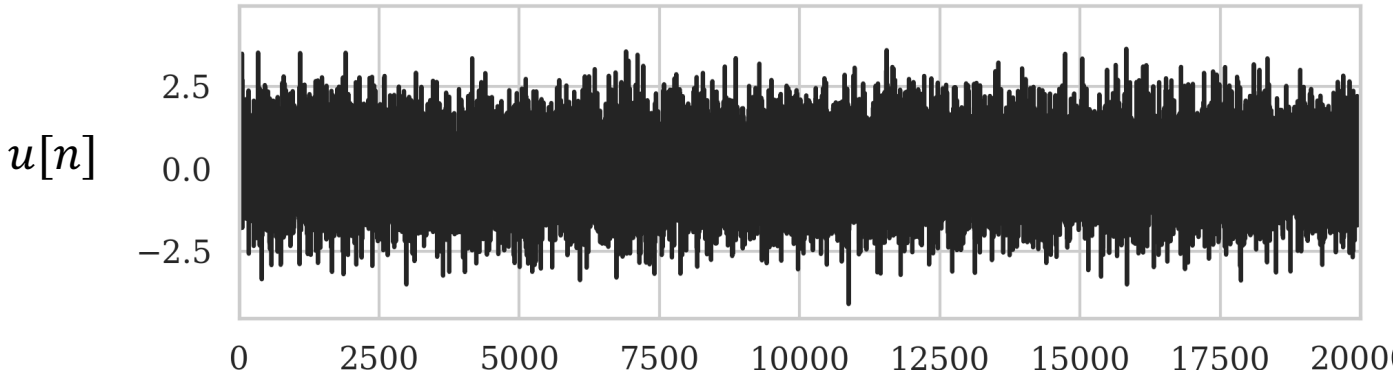




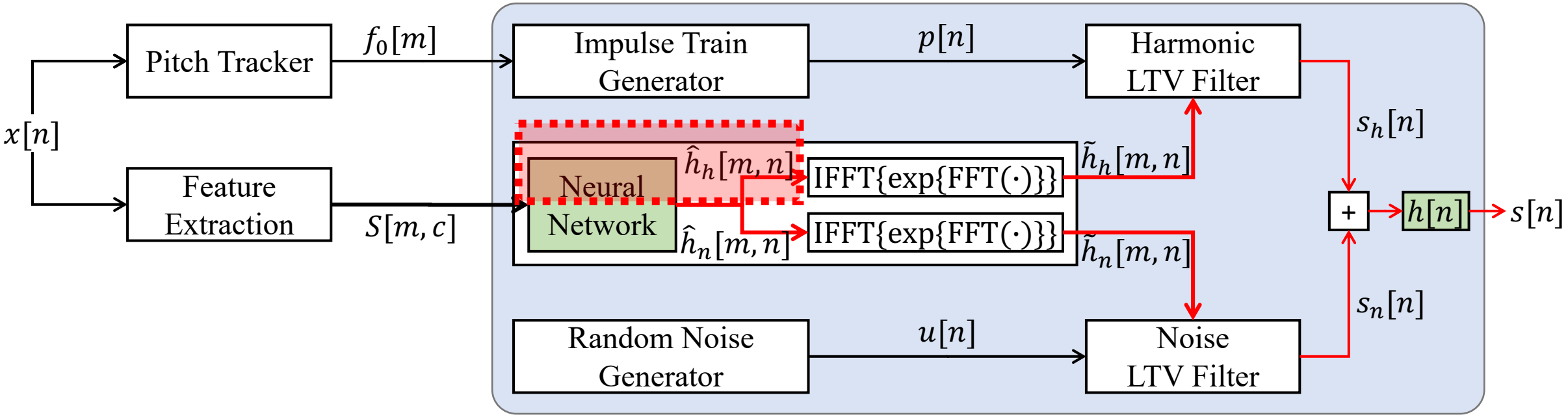
Neural Homomorphic Vocoder: Generation



The noise generator generates noise signal of the same length as the impulse train.

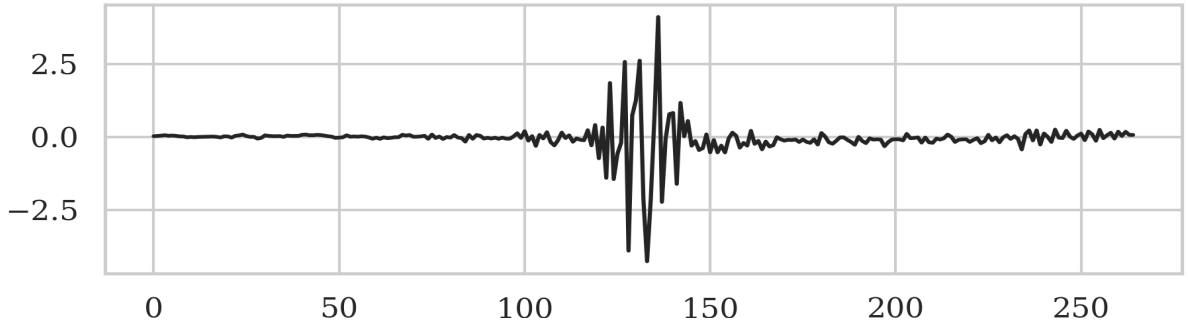


Neural Homomorphic Vocoder: Generation



Next, the neural work generates complex cepstrum according to the input features. The network runs at the frame rate rather than at the waveform sampling rate. The plot shows a complex cepstrum sampled from the harmonic branch network output.

$\hat{h}_h[200,n]$



Why complex cepstrum based filter parameterization?

- Complex cepstrum is just one of many solutions. We have other methods for filter parameterization:
  - Direct parameterization: Directly generating full impulse response in speech is challenging for neural networks.
  - Log amplitude + phase spectrum.
  - Complex amplitude.
  - …
- Even though the work is named Neural Homomorphic Vocoder, the key is the training procedure rather than filter parameterization with complex cepstrums.
- Complex cepstrum seems to be an elegant solution.

A crash course to complex cepstrum

- Recall DTFT and its inverse (IDTFT)

$$X\left(e^{j\omega}\right) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}$$

$$x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X\left(e^{j\omega}\right)e^{j\omega n}d\omega$$

- Recall DFT and its inverse (IDFT)

$$X[k] = X\left(e^{i2\pi k/N}\right) = \sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn}, \qquad 0 \le k \le N-1$$

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{j\frac{2\pi}{N}kn}, \qquad 0 \le n \le N-1$$

Rabiner, Lawrence, and Ronald Schafer. Theory and applications of digital speech processing. Prentice Hall Press, 2010.

Definition of the complex cepstrum

- $\hat{x}[n]$ is the complex cepstrum of $x[n]$. It is defined as the following:

$$\hat{X}\left(e^{j\omega}\right) = \log\left\{X\left(e^{j\omega}\right)\right\} = \log\left|X\left(e^{j\omega}\right)\right| + j\arg\left\{X\left(e^{j\omega}\right)\right\}$$

$$\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left\{X\left(e^{j\omega}\right)\right\} e^{j\omega n} d\omega$$

- Complex logarithm is a multivalued function. We define the imaginary part of $\hat{X}(e^{j\omega})$ to be **odd and continuous** branch of the complex log.
- Since $x[n]$ is a **real** signal, $X\left(e^{j\omega}\right)$ and $\hat{X}\left(e^{j\omega}\right)$ are **conjugate symmetric**. This implies $\hat{x}[n]$ is a **real** signal.

Rabiner, Lawrence, and Ronald Schafer. Theory and applications of digital speech processing. Prentice Hall Press, 2010.

Convolution becomes additive in complex cepstrum

- Suppose we have convolved two signal, $x[n] = x_1[n] * x_2[n]$. The complex cepstrum of $x[n]$ is the sum of $x_1[n]$ and $x_2[n]$.

$$X\left(e^{j\omega}\right) = X_1\left(e^{j\omega}\right) \cdot X_2\left(e^{j\omega}\right)$$

$$\log\left\{X_1\left(e^{j\omega}\right) X_2\left(e^{j\omega}\right)\right\} = \log\left\{X_1\left(e^{j\omega}\right)\right\} + \log\left\{X_2\left(e^{j\omega}\right)\right\}$$

$$\hat{X}\left(e^{j\omega}\right) = \hat{X}_1\left(e^{j\omega}\right) + \hat{X}_2\left(e^{j\omega}\right)$$

$$\hat{x}[n] = \hat{x}_1[n] + \hat{x}_2[n]$$

Rabiner, Lawrence, and Ronald Schafer. Theory and applications of digital speech processing. Prentice Hall Press, 2010.

The odd and even component of a complex cepstrum

- Any given complex cepstrum can be decomposed into even and odd components. $\hat{x}[n] = c[n] + d[n]$.
- The even component $c[n] = {\hat{x}[n]+\hat{x}[-n]}/{2}$ corresponds to a system with zero-phase response (an linear-phase system).
- The odd component $\mathrm{d}[n] = {\hat{x}[n]-\hat{x}[-n]}/{2}$ corresponds to a system with zero-amplitude response (an all-pass system).

Rabiner, Lawrence, and Ronald Schafer. Theory and applications of digital speech processing. Prentice Hall Press, 2010.

Complex cepstrum of a rational LTI system

- Consider a LTI system with the following Z-transform.

$$X(z) = Az^{-r} \frac{\prod_{k=1}^{M_i} \left(1 - a_k z^{-1}\right) \prod_{k=1}^{M_o} \left(1 - b_k z\right)}{\prod_{k=1}^{N_i} \left(1 - c_k z^{-1}\right) \prod_{k=1}^{N_o} \left(1 - d_k z\right)}$$

$$|a_k|, |b_k|, |c_k|, |d_k| < 1, \quad r \in \mathbf{Z}$$

- $1/b_k$ and $1/d_k$ correspond to zeros and poles outside the unit circle.
- a_k and c_k correspond to zeros and poles inside the unit circle.
- z^{-r} corresponds to zeros and poles at 0 and ∞.

Rabiner, Lawrence, and Ronald Schafer. Theory and applications of digital speech processing. Prentice Hall Press, 2010.

Complex cepstrum of a rational LTI system

- The complex cepstrum not considering $\log(z^{-r})$ is given by:

$$\hat{x}[n] = \begin{cases} \log A & n = 0 \\ \sum_{k=1}^{N_i} \frac{c_k^n}{n} - \sum_{k=1}^{M_i} \frac{a_k^n}{n} & n > 0 \\ \sum_{k=1}^{M_o} \frac{b_k^{-n}}{n} - \sum_{k=1}^{N_o} \frac{d_k^{-n}}{n} & n < 0 \end{cases}$$

- The "time shift" term's contribution: $Z^{-1}\left[\log\left(z^{-r}\right)\right] = r\frac{\cos(\pi n)}{n} \quad n \neq 0$
- For any rational LTI system, the system is minimum phase if and only if its complex cepstrum is causal.
- For any rational LTI system, its complex cepstrum falls as fast as $1/|n|$.
- This is also true for non-rational LTI systems. (But I could not find a proof for this.)

Rabiner, Lawrence, and Ronald Schafer. Theory and applications of digital speech processing. Prentice Hall Press, 2010.

Complex cepstrum inversion with DFT approximation

- The filter can be parameterized by complex cepstrums. The inverse is given by: (No phase unwrapping in synthesis!)

$$\hat{X}\left(e^{j\omega}\right) = \sum_{n=-\infty}^{\infty} \hat{x}[n]e^{-j\omega n}, \quad X\left(e^{j\omega}\right) = \exp \hat{X}\left(e^{j\omega}\right), \quad x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X\left(e^{j\omega}\right)$$

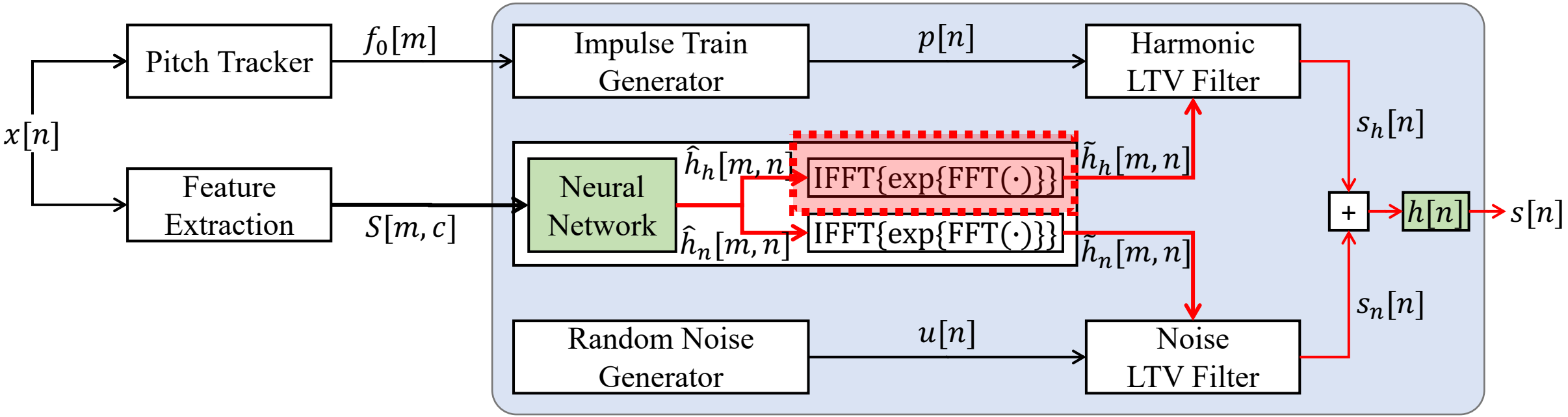
- For implementation, we use the DFT approximation of above:

$$\tilde{\hat{X}}[k] = \hat{X}\left(e^{j2\pi k/N}\right) = \sum_{n=0}^{N-1} \hat{x}[n]e^{-j\frac{2\pi k}{N}n}, \qquad 0 \le k \le N-1$$

$$\tilde{X}[k] = \exp\{\tilde{\hat{X}}[k]\}, \qquad 0 \le k \le N-1$$

$$\tilde{x}[n] = \sum_{r=-\infty}^{\infty} x[n+rN] \approx x[n], \qquad 0 \le n \le N-1$$

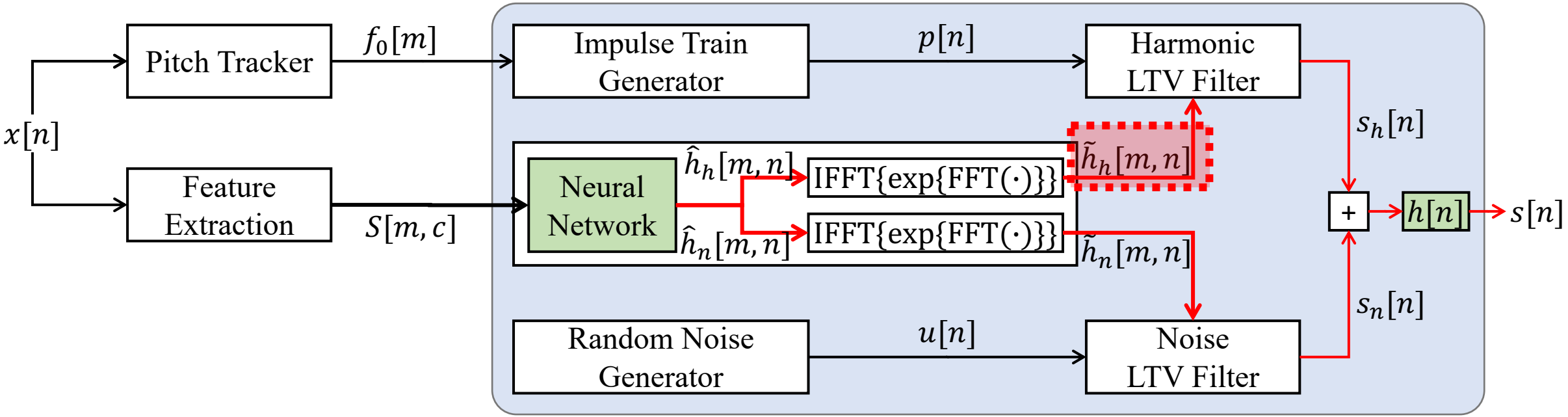
Neural Homomorphic Vocoder: Generation



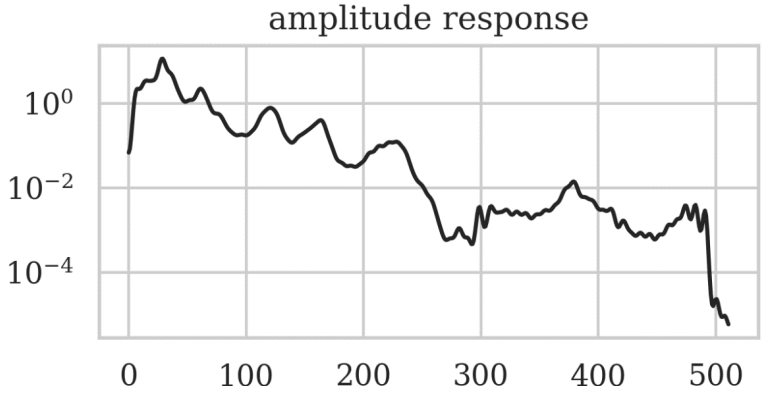
The finite complex cepstrum estimated by the neural network corresponds to infinitely long impulse responses. The FFT size should be made large enough to avoid serious aliasing. 1024 is good enough in practice.

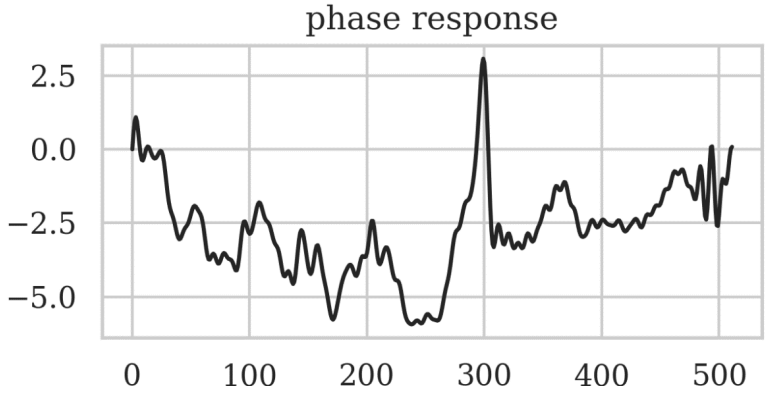
$$\tilde{h_h}[n] = \sum_{r=-\infty}^{\infty} h_h[n + rN] \approx h_h[n], \qquad 0 \le n \le N-1$$

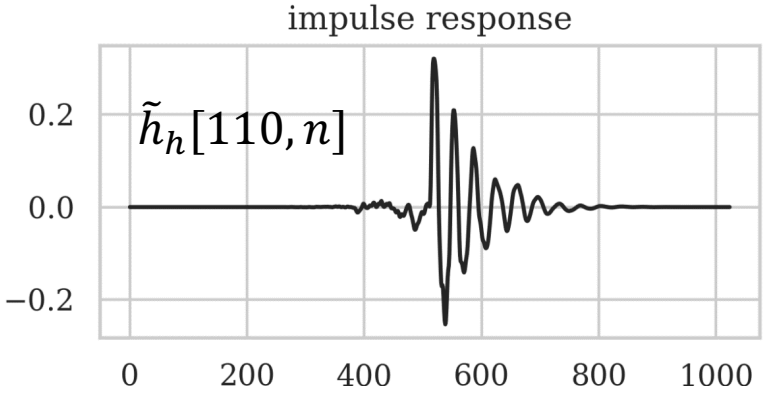
Neural Homomorphic Vocoder: Generation



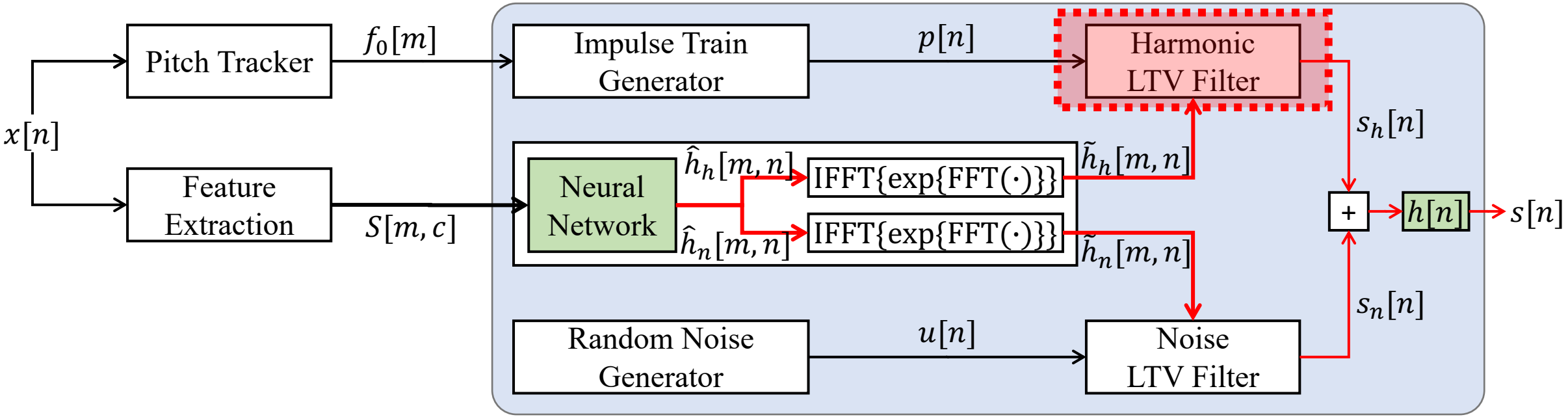
By using complex cepstrums, the neural network predicts the magnitude and phase response of the filters simultaneously.





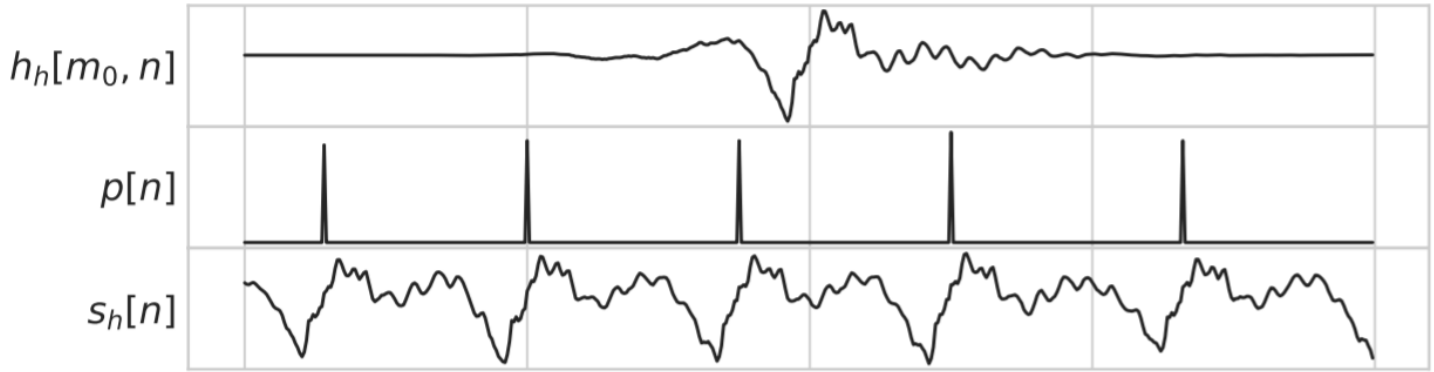


Neural Homomorphic Vocoder: Generation

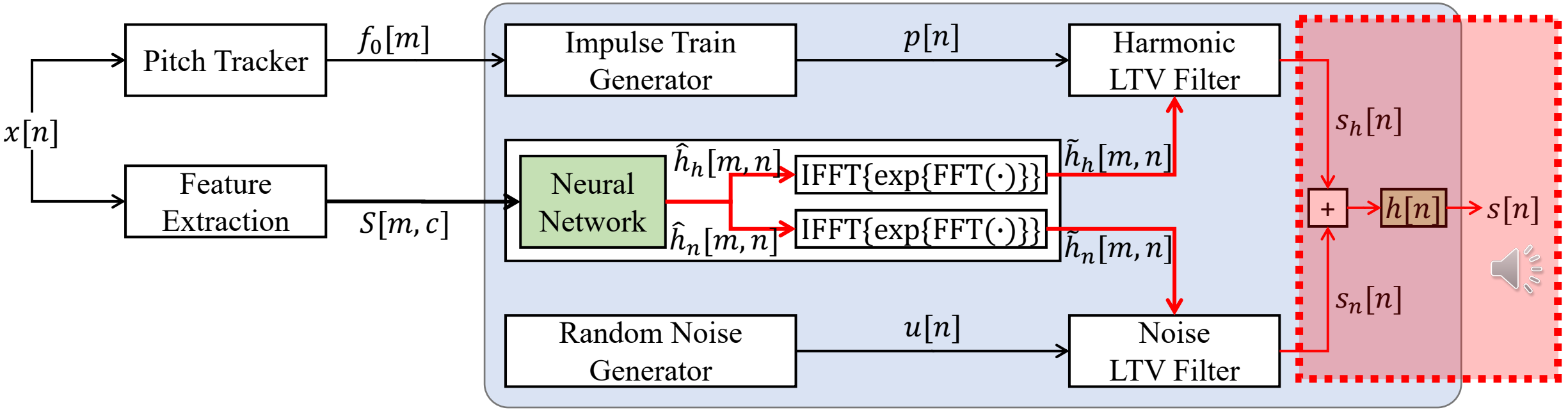


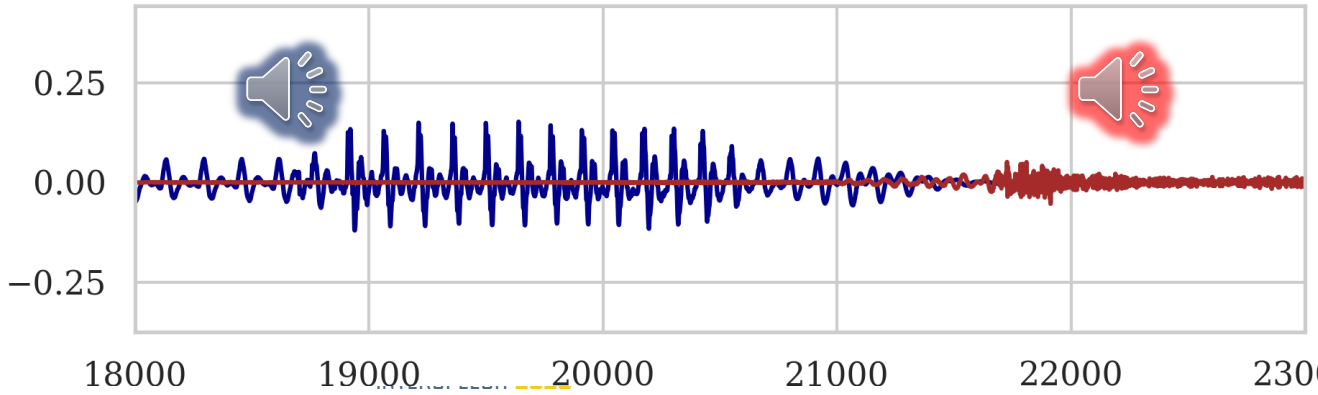
$$w_L[n] \triangleq \begin{cases} 1, & 0 \le n \le L-1 \\ 0, & \text{otherwise} \end{cases}$$

$$s_h[n] = \sum_{m=0}^{m<M} (w_L[n-mL]p[n]) * h_h[m,n]$$

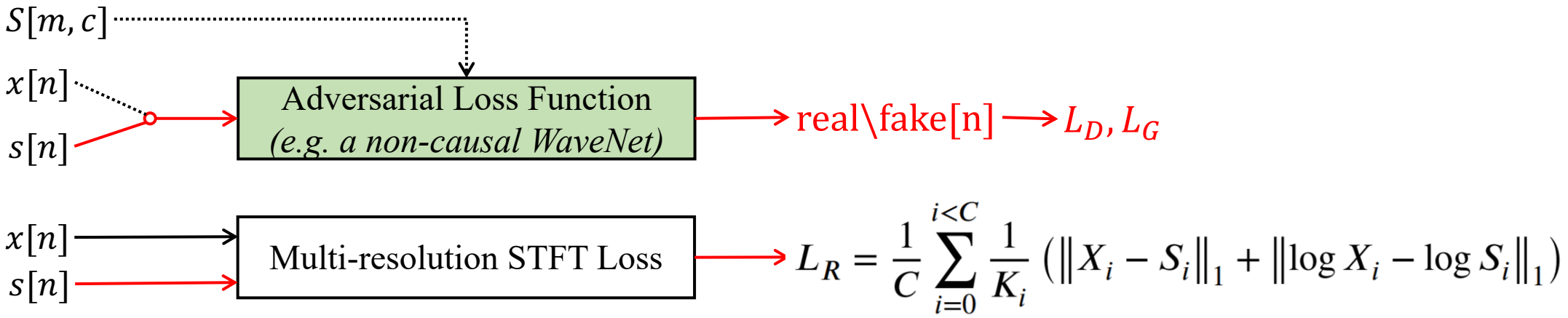


Neural Homomorphic Vocoder: Generation



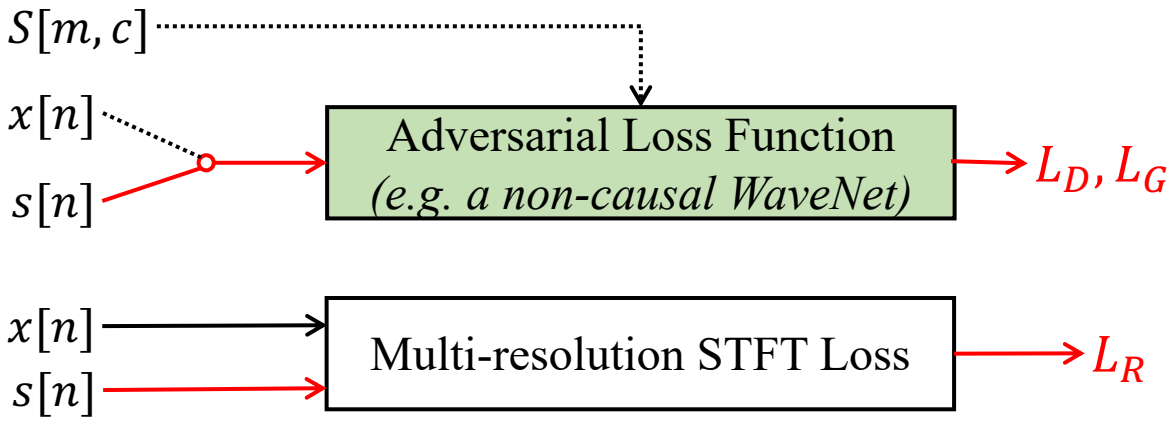


Neural Homomorphic Vocoder: Training



- In NHV, training with waveform loss such as the L1 or L2 loss is unfeasible, due to its sensitivity to linear phase shifts.
- STFT magnitude Loss is not sensitive to linear phase shifts.
- We find using many different scales leads to better performance.

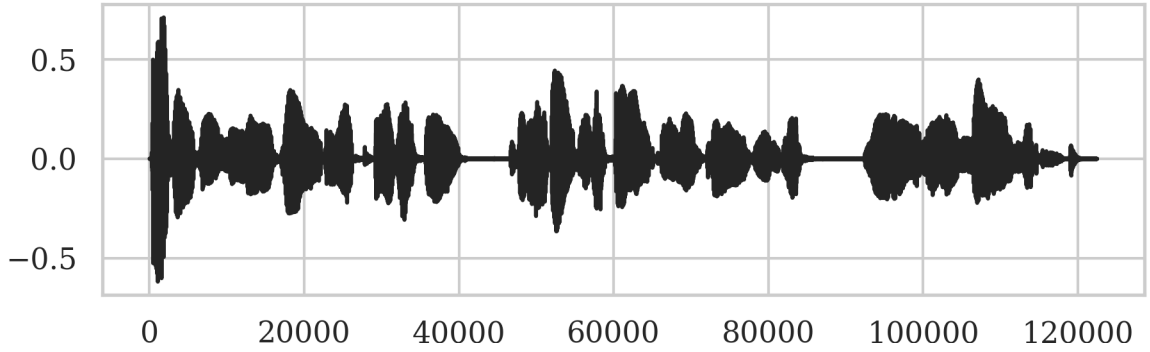
Neural Homomorphic Vocoder: Training

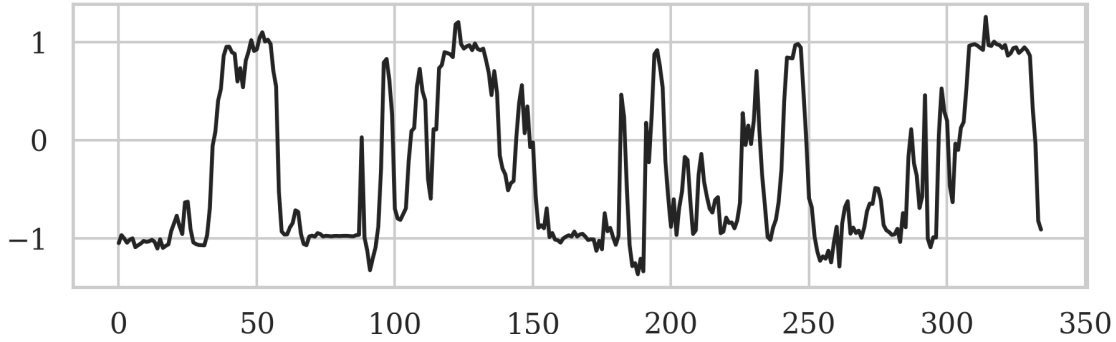


$$L_D = \mathbb{E}_{x,S}\left[\max\left(0, 1 - D(x, S)\right)\right] + \mathbb{E}_{f_0,S}\left[\max\left(0, 1 + D(G(f_0, S), S)\right)\right]$$

$$L_G = \mathbb{E}_{f_0,S}\left[-D(G(f_0, S), S)\right]$$

- A waveform input adversarial loss function is necessary for high quality reconstruction. It mostly affects the phase response of filters.





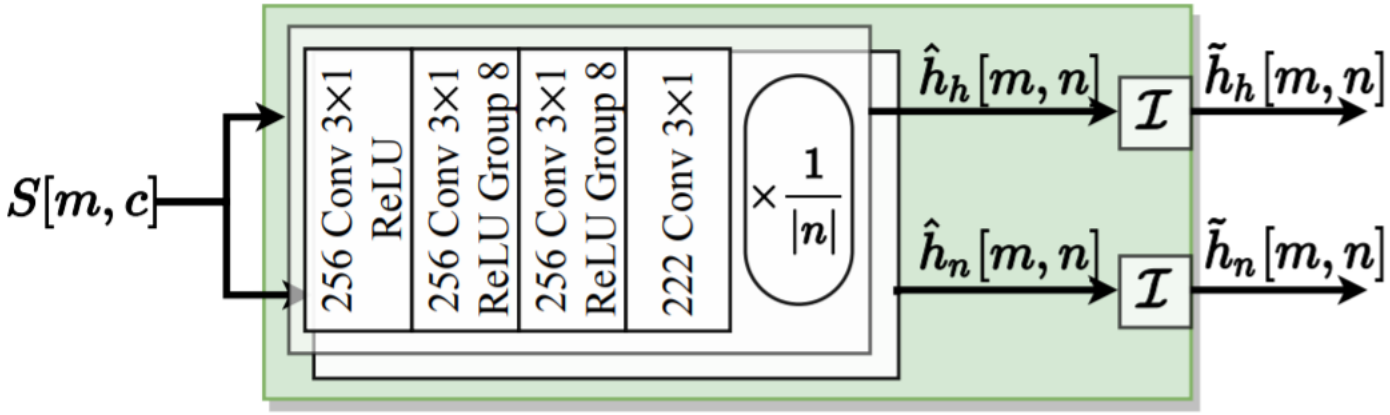
Dataset and Task

- We tested NHV on the task of speech synthesis based on mel-spectrograms.
- The input features to NHV are 80 dimensional band-passed log mel spectrograms and fundamental frequencies.
- The STFT window length was set to 512 for Mel analysis. The frame shift was set to about 6 milliseconds, or 128 sampling points.
- We used the open sourced *Chinese Standard Mandarin Speech Copus* (CSMSC) for evaluation. The 48000 Hz recordings were down-sampled to 22050 Hz.
- There are 10,000 utterances (about 12 hours) in the dataset. The last 100 utterances were reserved for testing.

Models in Comparison

- **NSF(GAN)**: b-NSF augmented with GAN.
- **hn-sinc-NSF**: Official open sourced implementation of hn-sinc-NSF.
- NHV(GAN): NHV trained with GAN.
- **NHV(cGAN)**: NHV trained with conditional GAN.
- NHV(cGAN, Minimum Phase): NHV(cGAN) with phase of filter constrained to MP.
- NHV(cGAN, Zero Phase): NHV(cGAN) with phase of filter constrained to ZP.
- **(MoL) WaveNet**: ESPNet open sourced pretrained mixture of logistic WaveNet.
- **Parallel WaveGAN**: Our reproduction of Parallel WaveGAN.
- **NHV-noadv**: NHV trained with multi-resolution STFT Loss only.
- DDSP: Our reproduction of DDSP.
- DDSP(cGAN): DDSP augmented with conditional GAN.
- WORLD

NHV Generator Structure



- Two 4 layer 1D CNNs were used for complex cepstrum estimation. One for the harmonic and the other for the noise component.
- The entire model contains about 0.6 million parameters.
- FFT size in complex cepstrum inversion was set to 1,024.
- The length of the complex cepstrums $\hat{h}_*[m, n]$ was set to 222.
- The final FIR is set to 1,024 points long.

Discriminator Structure

- All discriminators took waveform samples as input.
- All cGAN based models shared the same discriminator network structure.
  - Non-causal WaveNet with 14 layers and 2 dilation cycle. (1, 2, …, 128, 1, 2, …, 128).
  - Kernel size = 3.
  - # of skip channels = # of residual channels = 64.
  - This non-causal WaveNet has a receptive field of 1024 sampling points.
- All GAN based models shared the same discriminator network structure. (Except for Parallel WaveGAN)
  - A 10 layer 1D CNN
  - Kernel size = 3.
  - Strides = (2 ,2, 4, 2, 2, 2, 1, 1, 1, 1).
  - Each layer followed by leaky ReLU activation with negative slope 0.2. No conditioning used.
  - This network has a receptive field of 1263 and a stride of 128 sampling points.

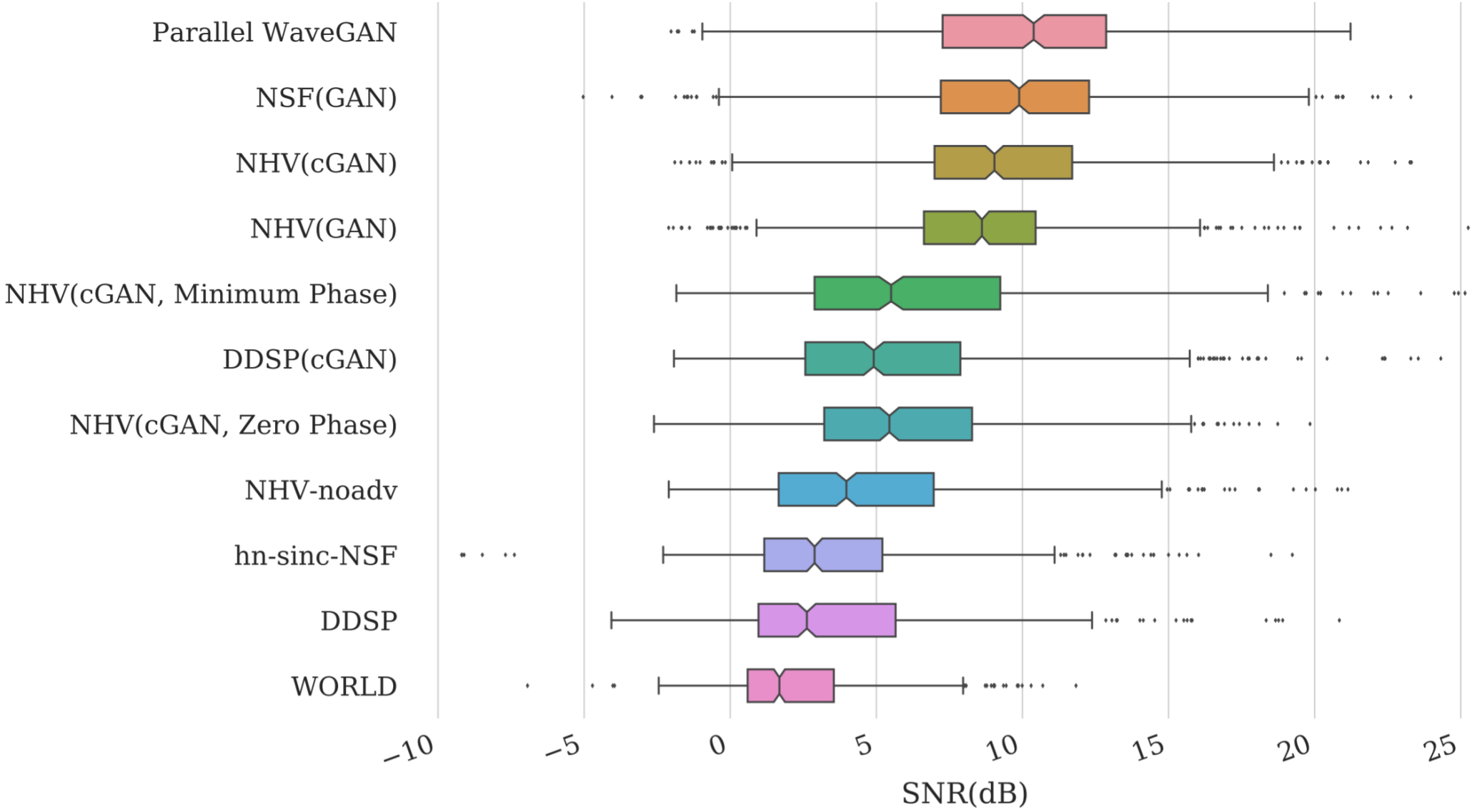
Signal to Noise Ratio Evaluation

- Signal to noise ratio evaluates magnitude and phase reconstruction simultaneously.
- Compute SNR for voiced region with following formula.

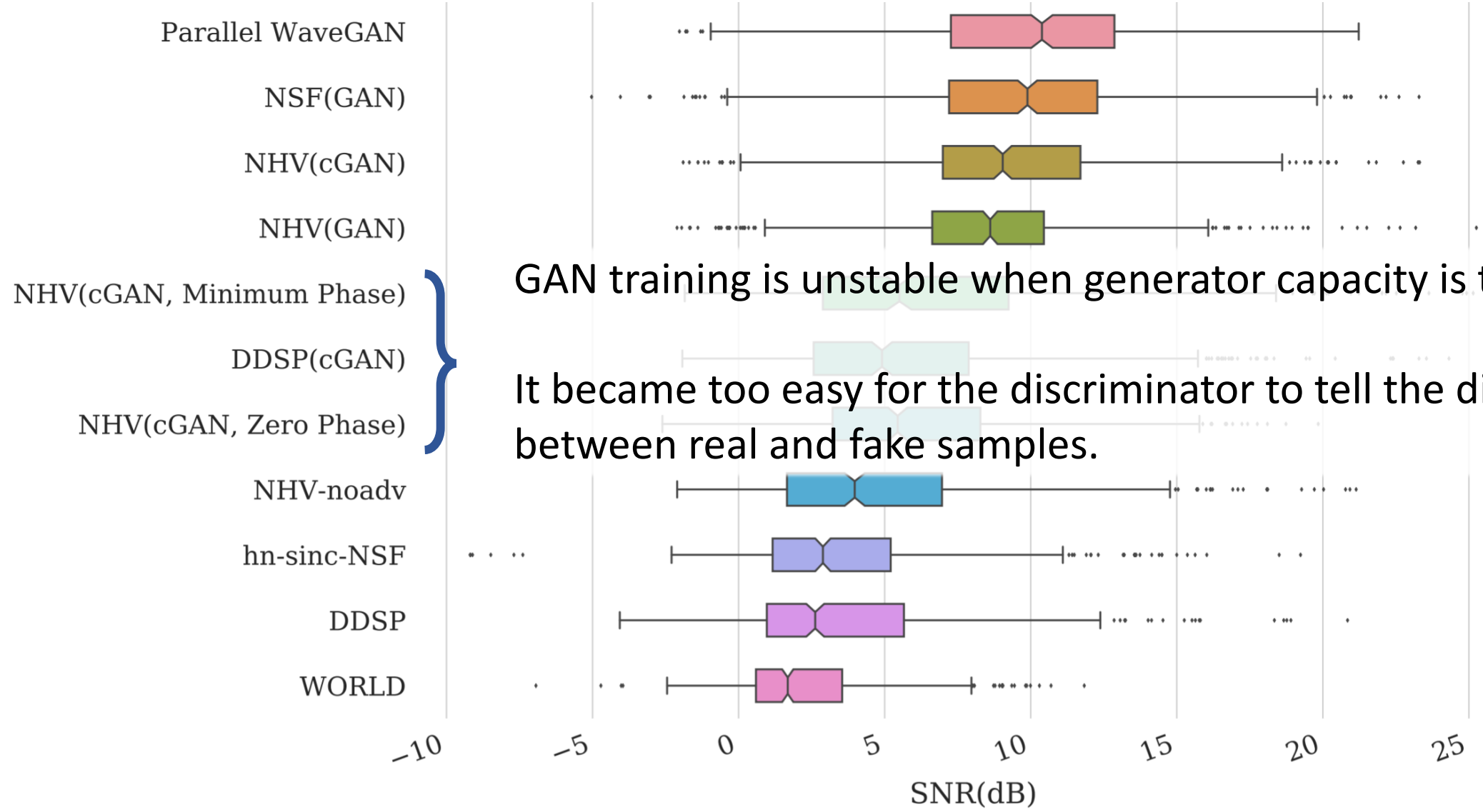
$$\mathrm{SNR}(x[n], s[n]) = 10 \log_{10} \left(\frac{\sum_{n=1}^{N} s[n]^2}{\sum_{n=1}^{N} (x[n] - s[n])^2} \right) \mathrm{dB}$$

- Frame length was set to 512, SNR is calculated for each frame. Linear phase shift compensated by moving $s[n]$ left and right at most 256 sampling points. The SNR is computed with the best possible match.

Signal to Noise Ratio Evaluation



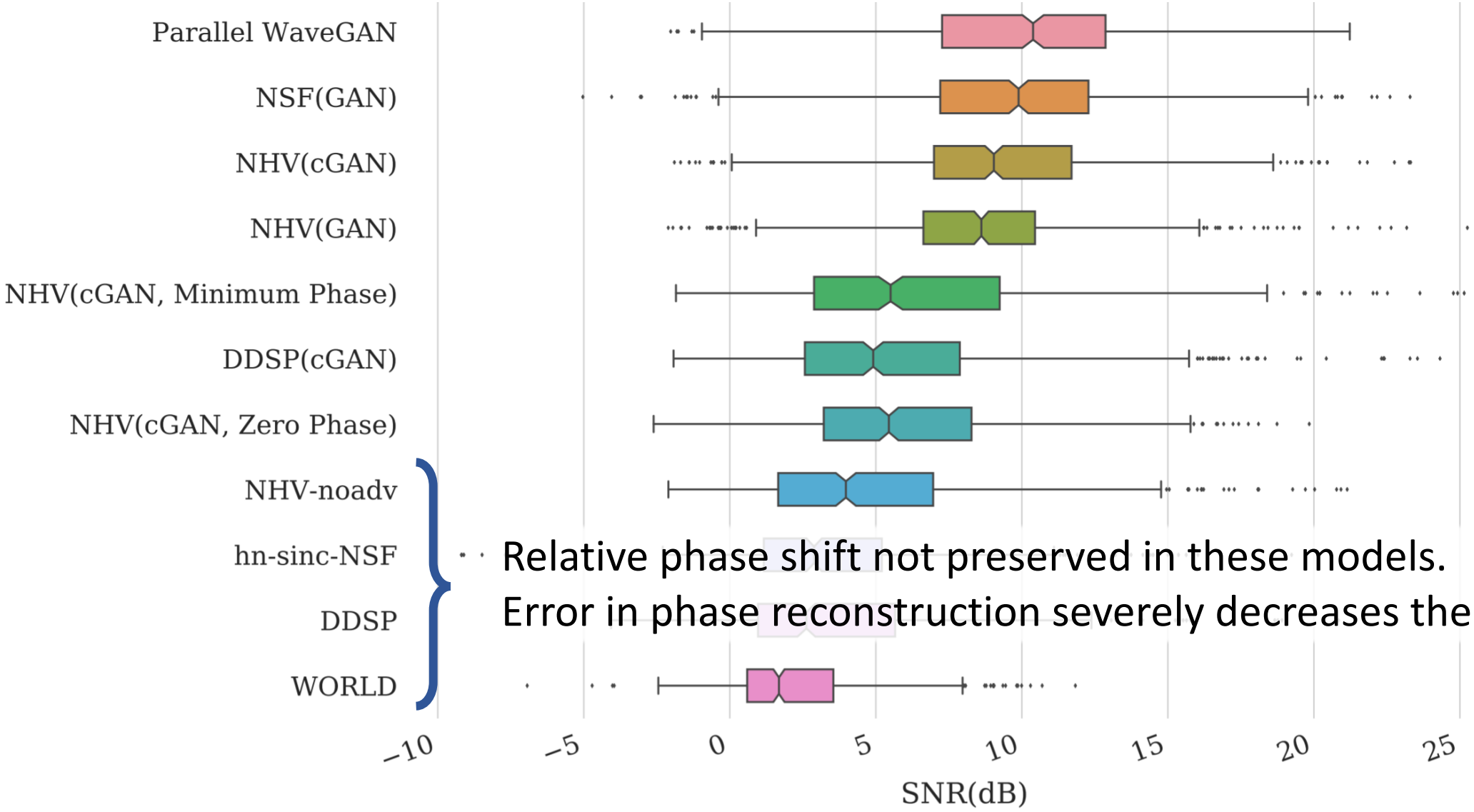
Signal to Noise Ratio Evaluation



GAN training is unstable when generator capacity is too low.

It became too easy for the discriminator to tell the difference between real and fake samples.

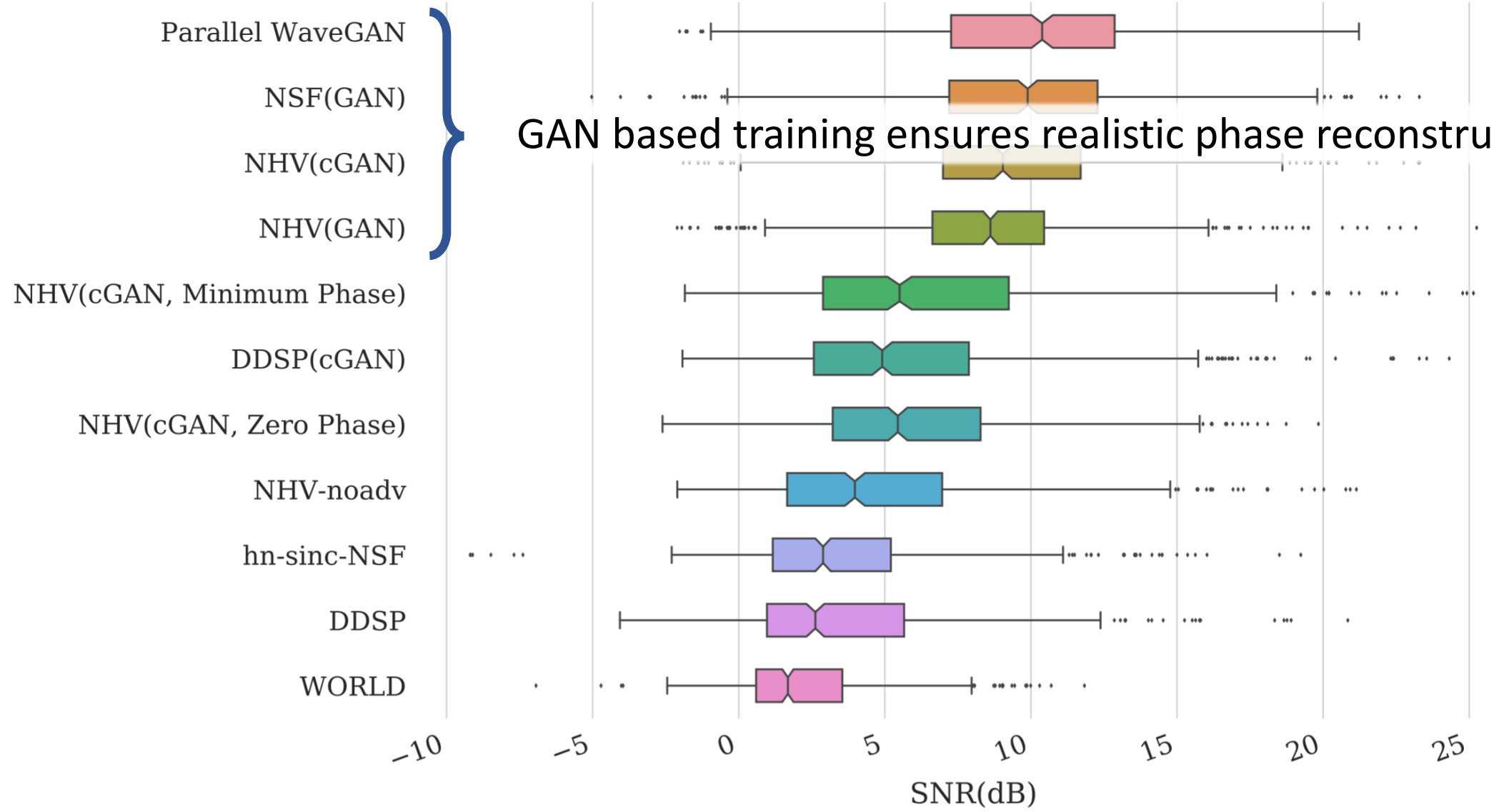
Signal to Noise Ratio Evaluation



Relative phase shift not preserved in these models.

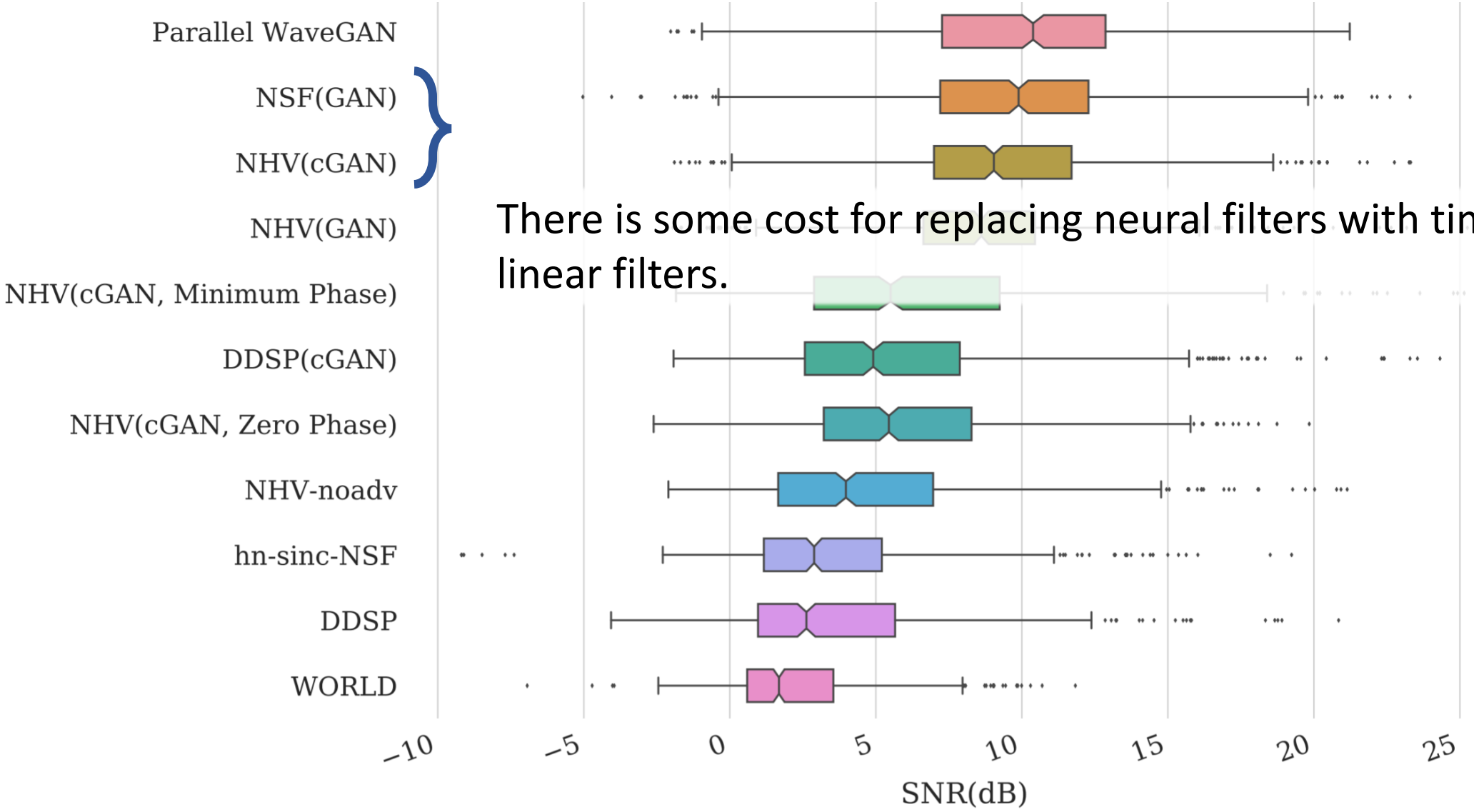
Error in phase reconstruction severely decreases the SNR.

Signal to Noise Ratio Evaluation

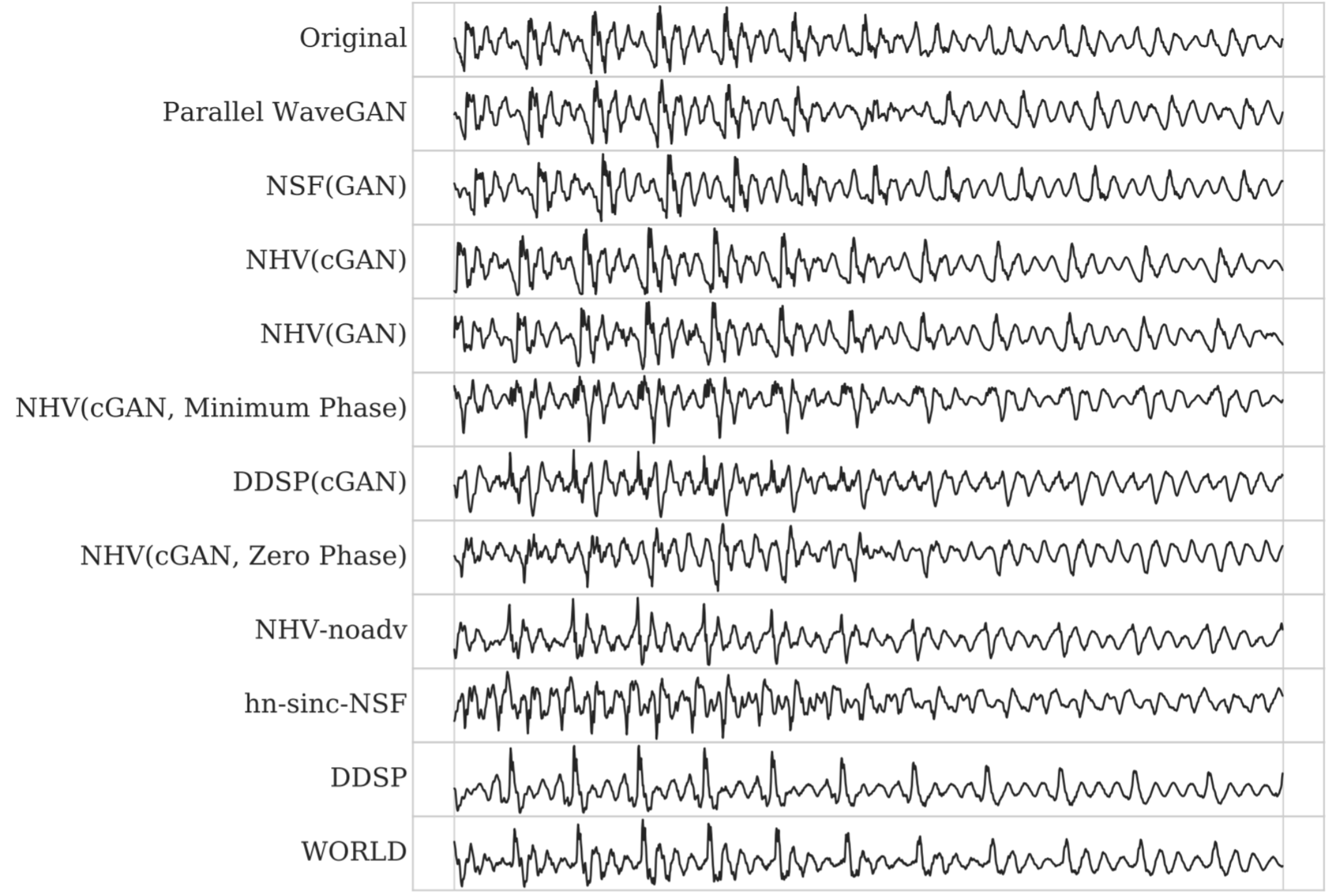


GAN based training ensures realistic phase reconstruction.

Signal to Noise Ratio Evaluation

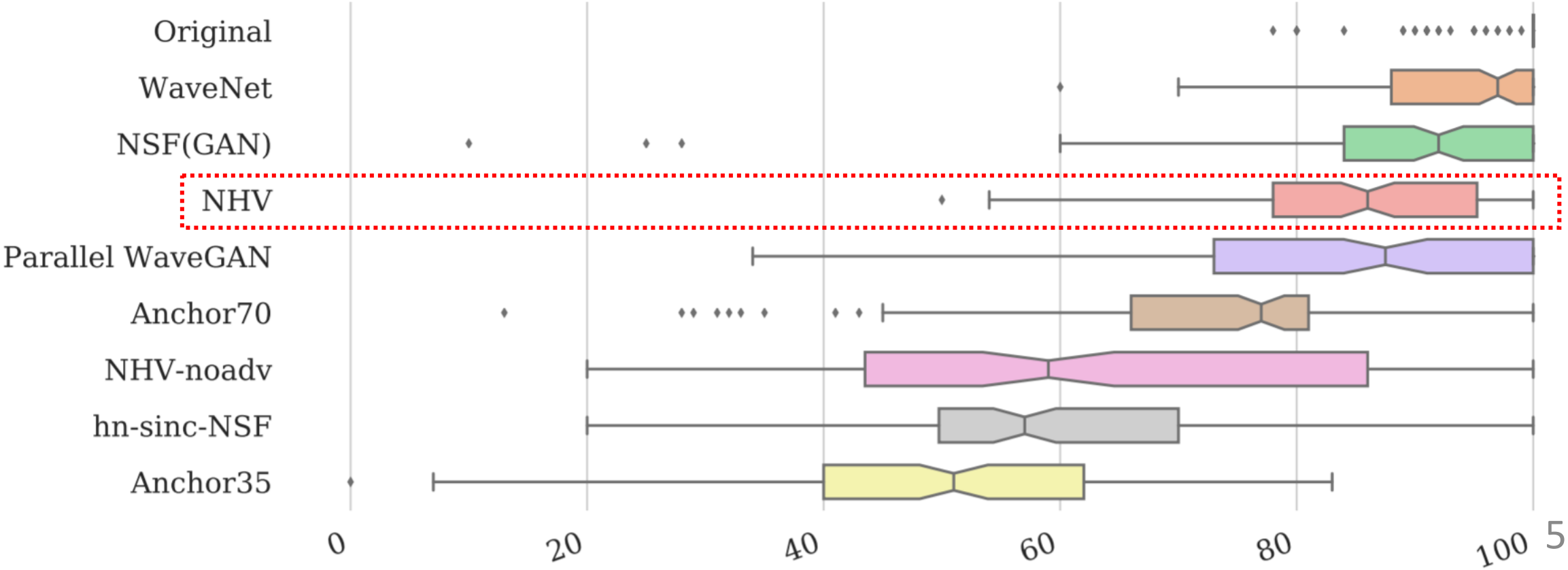


There is some cost for replacing neural filters with time varying linear filters.


Original
Parallel WaveGAN
NSF(GAN)
NHV(cGAN)
NHV(GAN)
NHV(cGAN, Minimum Phase)
DDSP(cGAN)
NHV(cGAN, Zero Phase)
NHV-noadv
hn-sinc-NSF
DDSP
WORLD

Copy Synthesis: MUSHRA Results

- A MUSHRA test was carried out to evaluate performance in copy synthesis. 24 Listeners participated in the test. Standard anchors were used.



Text-to-Speech: MOS

- For text-to-speech a Tacotron2 model was trained to predict V/UV Flag, F0, and log Mel spectrogram.
- Performance of NHV is comparable to other baseline models.

Model	MOS Score
Original	4.71 ± 0.07
Tacotron2 + hn-sinc-NSF	2.83 ± 0.11
Tacotron2 + NSF(GAN)	3.76 ± 0.10
Tacotron2 + Parallel WaveGAN	3.76 ± 0.12
Tacotron2 + NHV	3.83 ± 0.09

Computational Complexity

- We report the number of floating point operations required for generating a sample of audio in NHV and other existing models.
- FLOPs count is low as the network is shallow and runs only at the frame rate.
- The neural network accounts for approximately half of the total computational complexity.

Model	FLOPs/sample
b-NSF	$4. \times 10^6$
Parallel WaveGAN	$2. \times 10^6$
Gaussian WaveNet	$2. \times 10^6$
MelGAN	$4. \times 10^5$
LPCNet	1.4×10^5
NHV	1.5×10^4

Thank you for listening!
Q&A Time

For further information, visit the online supplement of this paper at
https://zjlww.github.io/is2020/