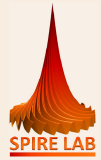
SFNet: A computationally efficient source filter model based neural speech synthesis

Achuth Rao M V
Advisor: Prasanta Kumar Ghosh

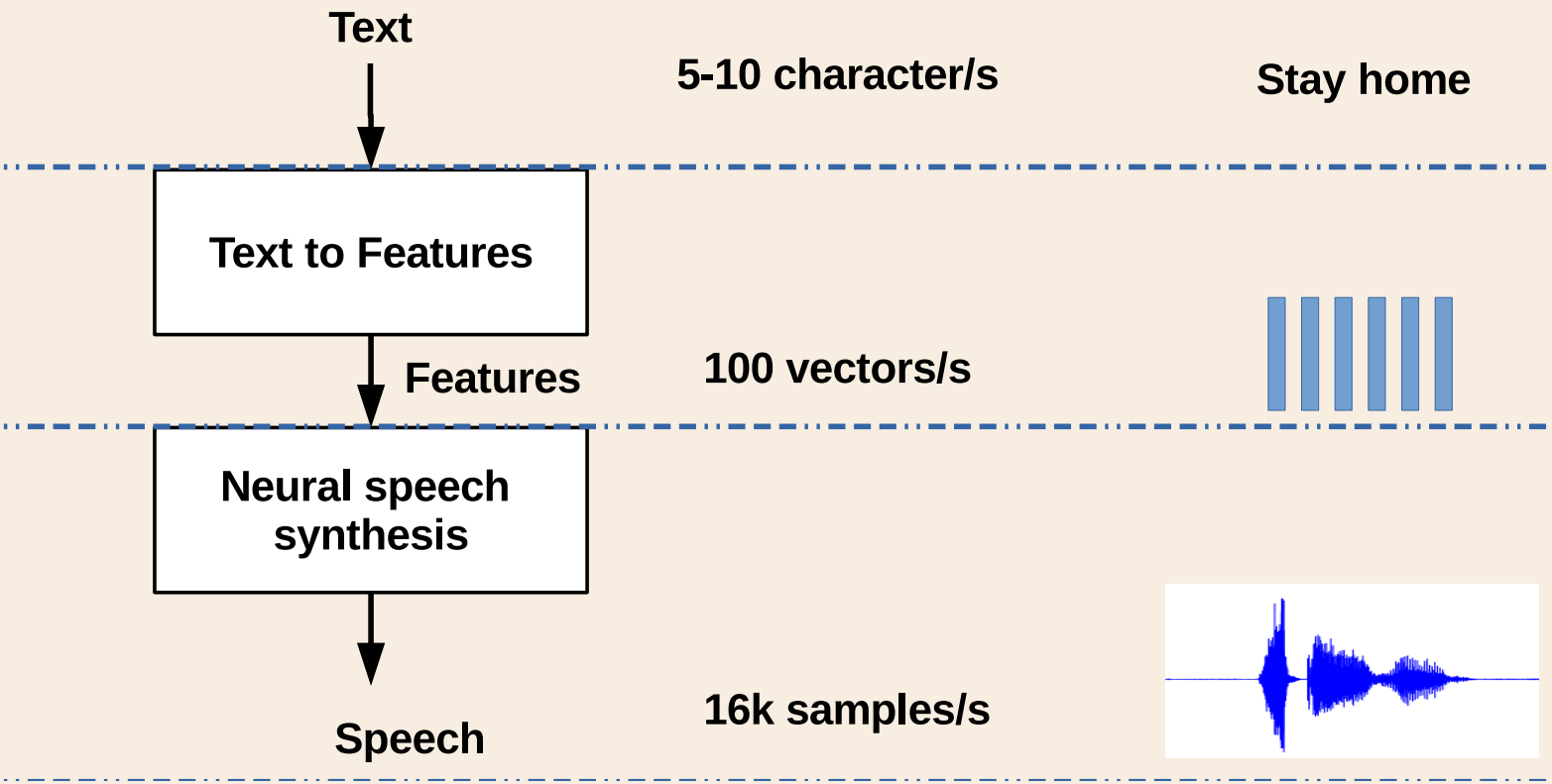
Electrical Engineering, IISc, Bengaluru, India



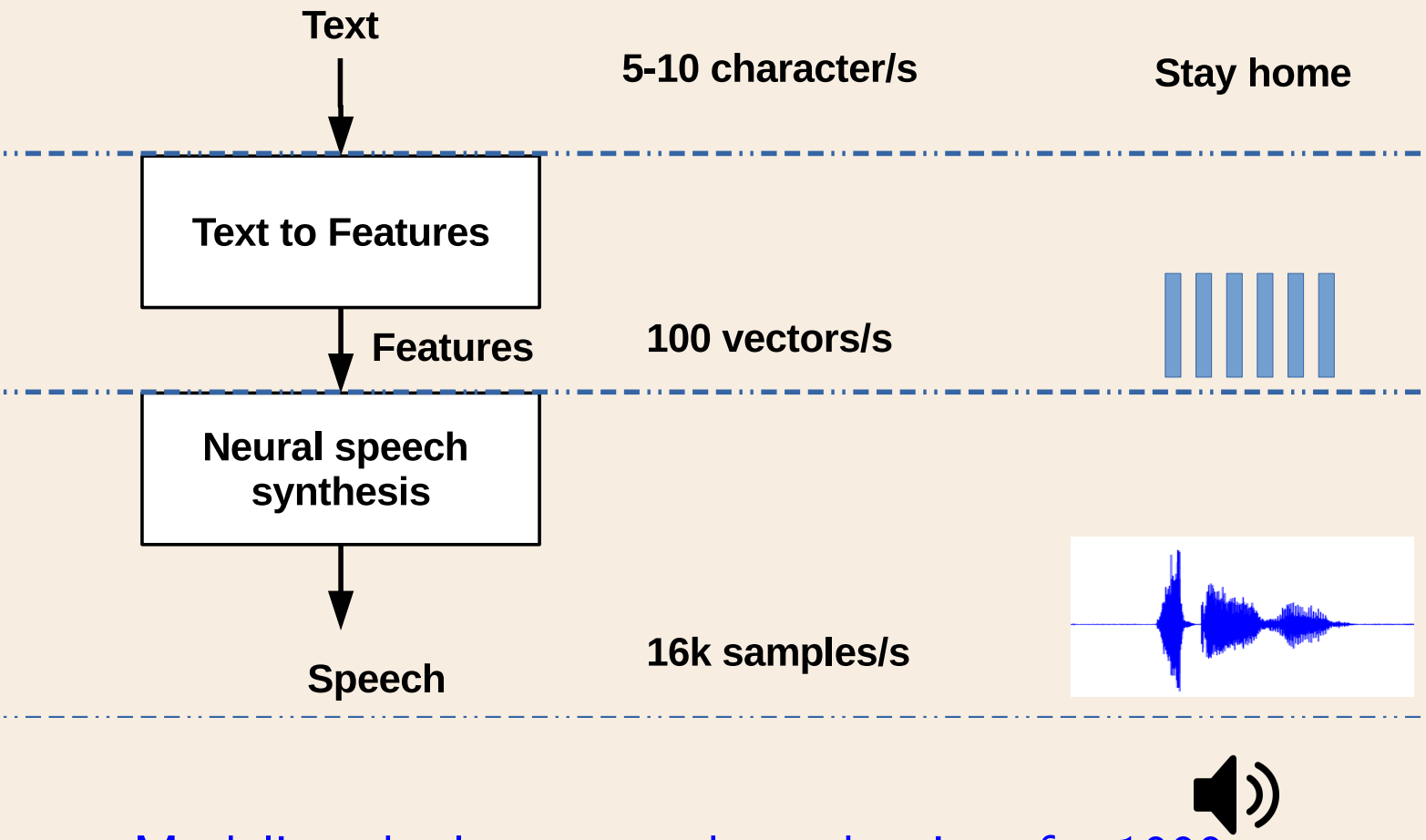
Overview

Overview

Text to speech synthesis



Text to speech synthesis



Challenge: Modeling the long term dependencies of >1000 steps.

Overview

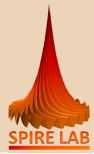
Related work

- Wavenet [1][2]:
  - uses dilated convolution to model the long term dependencies.
  - Synthesizes the speech in an autoregressive (AR) manner.
  - The complexity is ∼50GFLOPs.

[1] Oord et al., "Wavenet: A generative model for raw audio", 2016

[2] Oord et al., "Parallel WaveNet: Fast High-Fidelity Speech Synthesis", 2018

Related work



- Structured neural network: [1][2]:
  - Uses some kind of structured network with sparse weights to reduce the complexity.
  - The real-time synthesis is only possible in GPUs.

[1] Kalchbrenner et al., "Efficient Neural Audio Synthesis", 2018

[2] Prenger, Valle, and Catanzaro, "Waveglow: A flow-based generative network for speech synthesis", 2019

Related work

- Linear prediction based: [3][4][5]:
  - Complexity is reduced by modeling the spectrum envelope using linear prediction (LP). where

$$X(z) = \frac{E(z)}{A(z)}$$

    where $X(z)$ is the z-transform of the speech $x[n]$, $E(z)$ is the z-transform of $e[n]$, $\frac{1}{A(z)} = \frac{1}{1+\sum_{k=1}^{M} a_k z^{-k}}$ is the AR filter and $\mathbf{a} = [a_1, a_2, \ldots, a_M]$.
  - the excitation signal $e[n]$ is generated using neural networks.

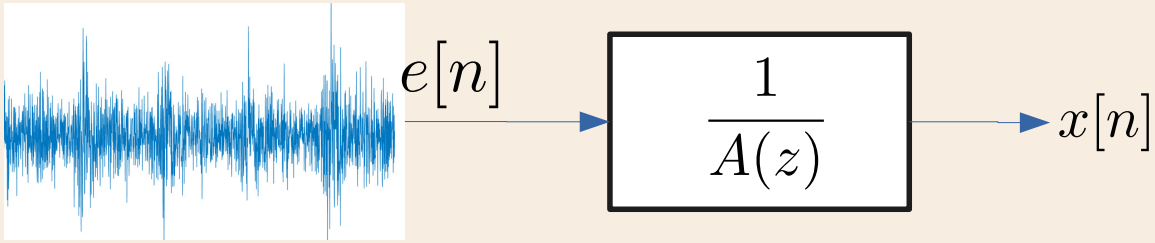
[3] Juvela et al., "Glotnet – A raw waveform model for the glottal excitation in statistical parametric speech synthesis", 2019

[4] Juvela et al., "GELP: GAN-Excited Linear Prediction for Speech Synthesis from Mel-Spectrogram", 2019

[5] Valin and Skoglund, "LPCNET: Improving Neural Speech Synthesis through Linear Prediction", 2019

Related work

- Linear prediction based: [3][4][5]:
  - Illustration:



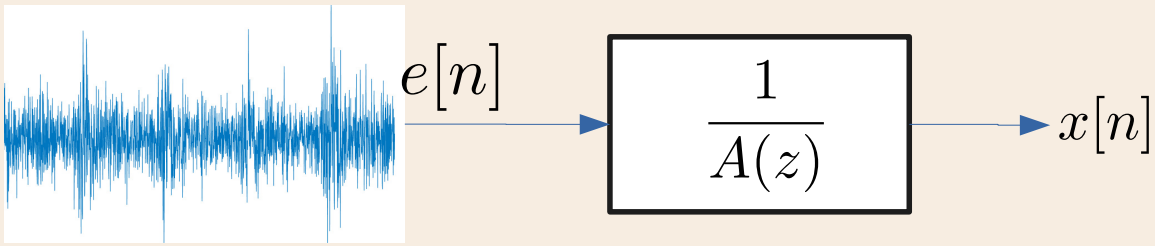
[3] Juvela et al., "Glotnet – A raw waveform model for the glottal excitation in statistical parametric speech synthesis", 2019

[4] Juvela et al., "GELP: GAN-Excited Linear Prediction for Speech Synthesis from Mel-Spectrogram", 2019

[5] Valin and Skoglund, "LPCNET: Improving Neural Speech Synthesis through Linear Prediction", 2019

Related work

- Linear prediction based: [3][4][5]:
  - Illustration:



  - Only LPCnet is shown to synthesize speech in real-time using CPU only. Still complexity is ∼2.5GFLOPs.

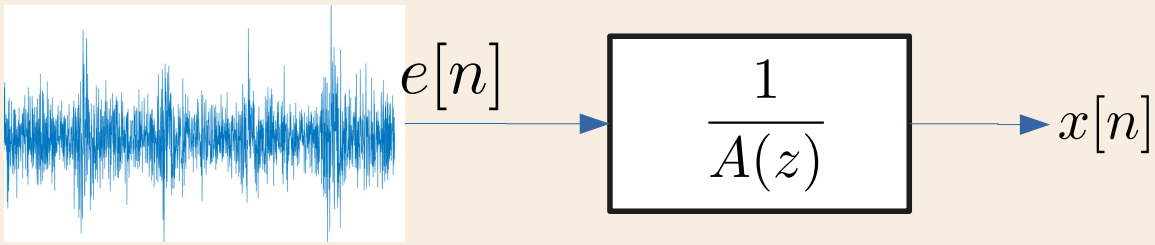
[3] Juvela et al., "Glotnet – A raw waveform model for the glottal excitation in statistical parametric speech synthesis", 2019

[4] Juvela et al., "GELP: GAN-Excited Linear Prediction for Speech Synthesis from Mel-Spectrogram", 2019

[5] Valin and Skoglund, "LPCNET: Improving Neural Speech Synthesis through Linear Prediction", 2019

Related work

- Linear prediction based: [3][4][5]:
  - Illustration:



  - Only LPCnet is shown to synthesize speech in real-time using CPU only. Still complexity is $\sim$2.5GFLOPs.
  - Can we break down the $E(z)$ into further components?

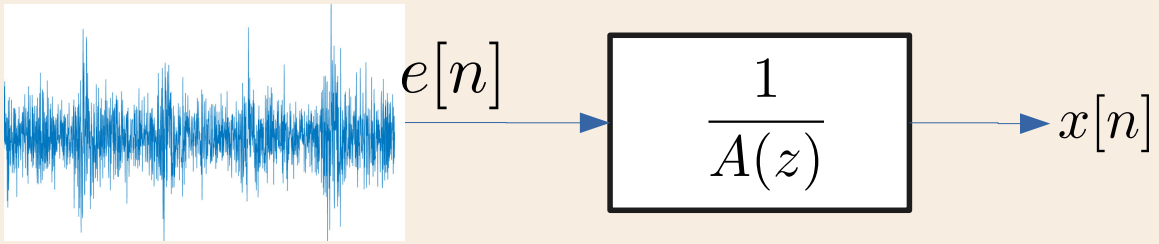
[3] Juvela et al., "Glotnet – A raw waveform model for the glottal excitation in statistical parametric speech synthesis", 2019

[4] Juvela et al., "GELP: GAN-Excited Linear Prediction for Speech Synthesis from Mel-Spectrogram", 2019

[5] Valin and Skoglund, "LPCNET: Improving Neural Speech Synthesis through Linear Prediction", 2019

Overview

Reformulation of source filter model.



- We further break-down excitation signal of the AR model as follows:

$$X^p(z) = \frac{1}{A(z)}\left(\alpha I(z)C(z) + W(z)D(z)\right) = \frac{E(z)}{A(z)}$$

- where $\frac{1}{A(z)}$ is the AR filter, $I(z)$ is the z-transform of the impulse train, $C(z)$ is fixed FIR filter, $W(z)$ is the white noise $(\mathcal{N}(0,1))$ and $D(z) = \sum_{r=1}^{Q} \sigma_r^2 \sum_{k=0}^{K-1} h_{rk} z^{-k}$. $D(z)$ models the noise in different sub-bands.

Reformulation of source filter model.

- We further break-down excitation signal of the AR model as follows:

$$X^p(z) = \frac{1}{A(z)}\left(\alpha I(z)C(z) + W(z)D(z)\right) = \frac{E(z)}{A(z)}$$

- where $\frac{1}{A(z)}$ is the AR filter, $I(z)$ is the z-transform of the impulse train, $C(z)$ is fixed FIR filter, $W(z)$ is the white noise ($\mathcal{N}(0,1)$) and $D(z) = \sum_{r=1}^{Q} \sigma_r^2 \sum_{k=0}^{K-1} h_{rk} z^{-k}$. $D(z)$ models the noise in different sub-bands.
- The parameters of the models for each frame are $\theta = \{\mathbf{a}, \{\sigma_r^2\}_{r=1}^{Q}, \alpha\}$ and a set of common parameters across frames $\zeta = \{H, \mathbf{c}\}$.

Reformulation of source filter model.

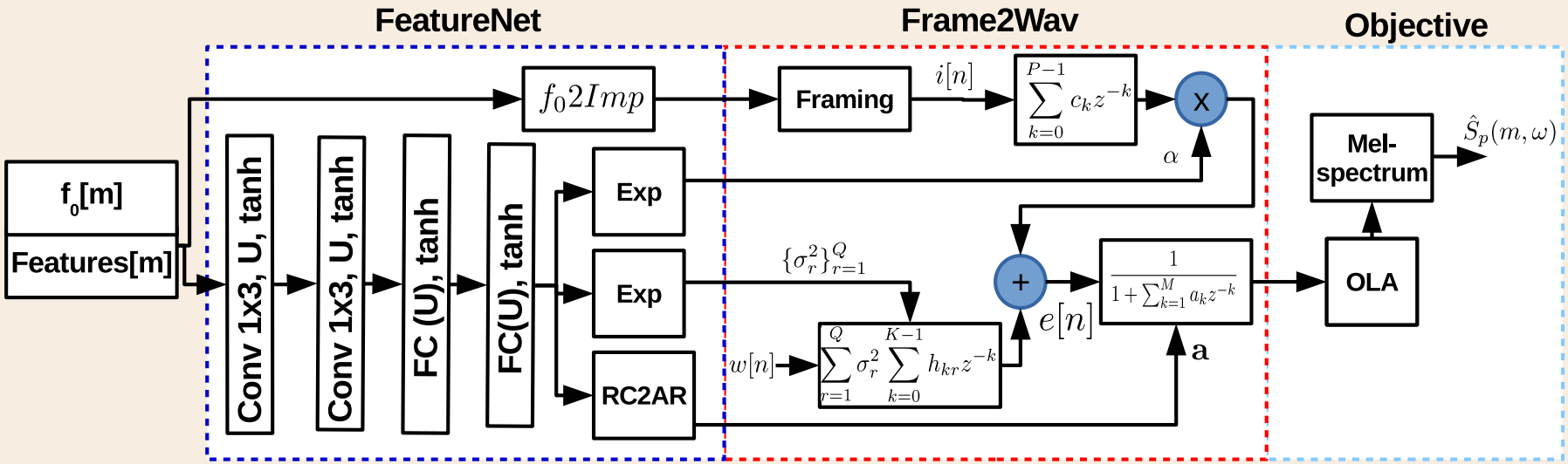
- We further break-down excitation signal of the AR model as follows:

$$X^p(z) = \frac{1}{A(z)}\left(\alpha I(z)C(z) + W(z)D(z)\right) = \frac{E(z)}{A(z)}$$

- where $\frac{1}{A(z)}$ is the AR filter, $I(z)$ is the z-transform of the impulse train, $C(z)$ is fixed FIR filter, $W(z)$ is the white noise ($\mathcal{N}(0,1)$) and $D(z) = \sum_{r=1}^{Q} \sigma_r^2 \sum_{k=0}^{K-1} h_{rk} z^{-k}$. $D(z)$ models the noise in different sub-bands.
- The parameters of the models for each frame are $\theta = \{\mathbf{a}, \{\sigma_r^2\}_{r=1}^{Q}, \alpha\}$ and a set of common parameters across frames $\zeta = \{H, \mathbf{c}\}$.
- We propose a data-driven neural network architecture, called SFNet, to predict the parameters (θ) of the source and the filter.

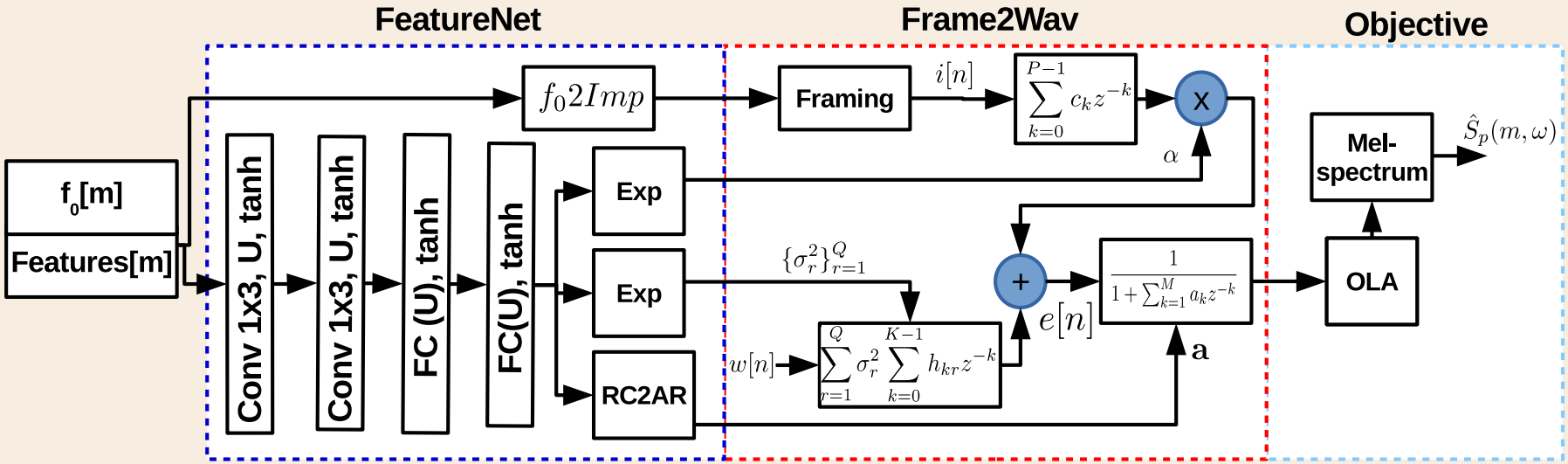
Overview

Architecture



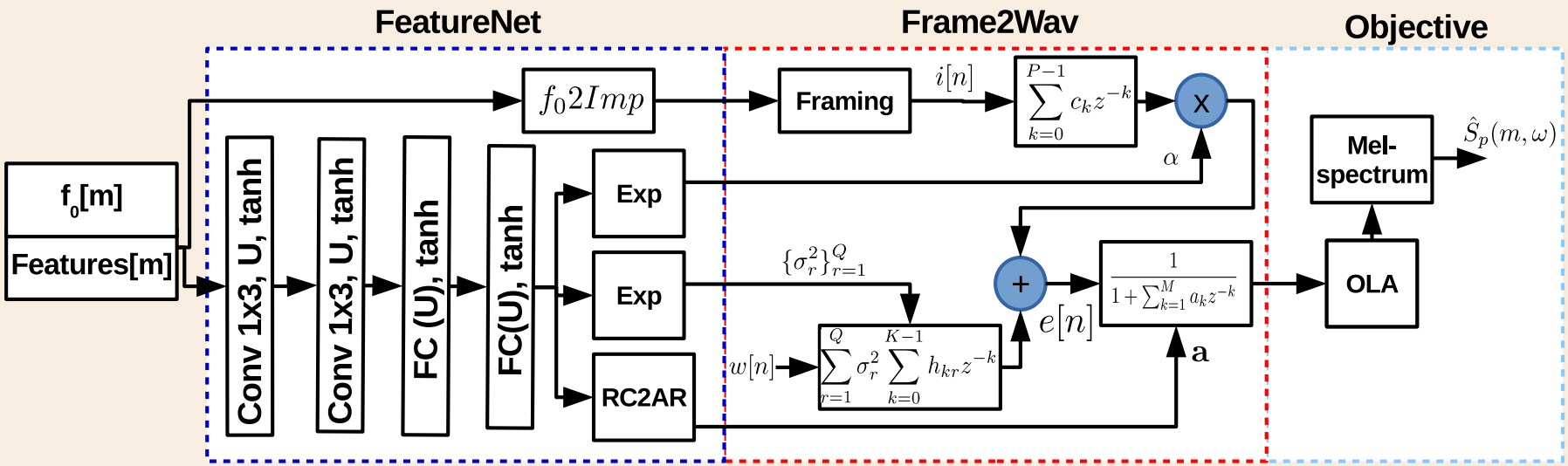
- There are three main parts:
  - FeatureNet
  - Frm2Wav
  - Objective function

Architecture



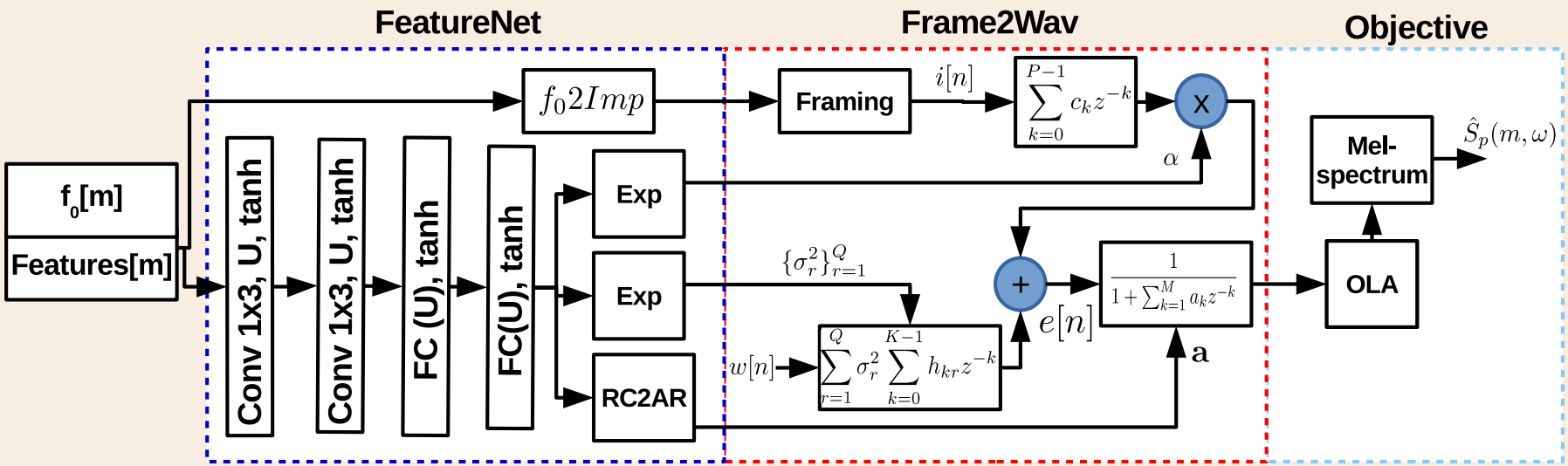
- FeatureNet
  - We use a 2 layer of Conv1D with U filters of kernel size 3 and tanh activation.
  - We use a two fully connected layer with tanh activation to get a inter-mediate representation.
  - RC2AR to guarantee the stability of the AR filter.
  - Exp function to guarantee the non-negative value.

Architecture



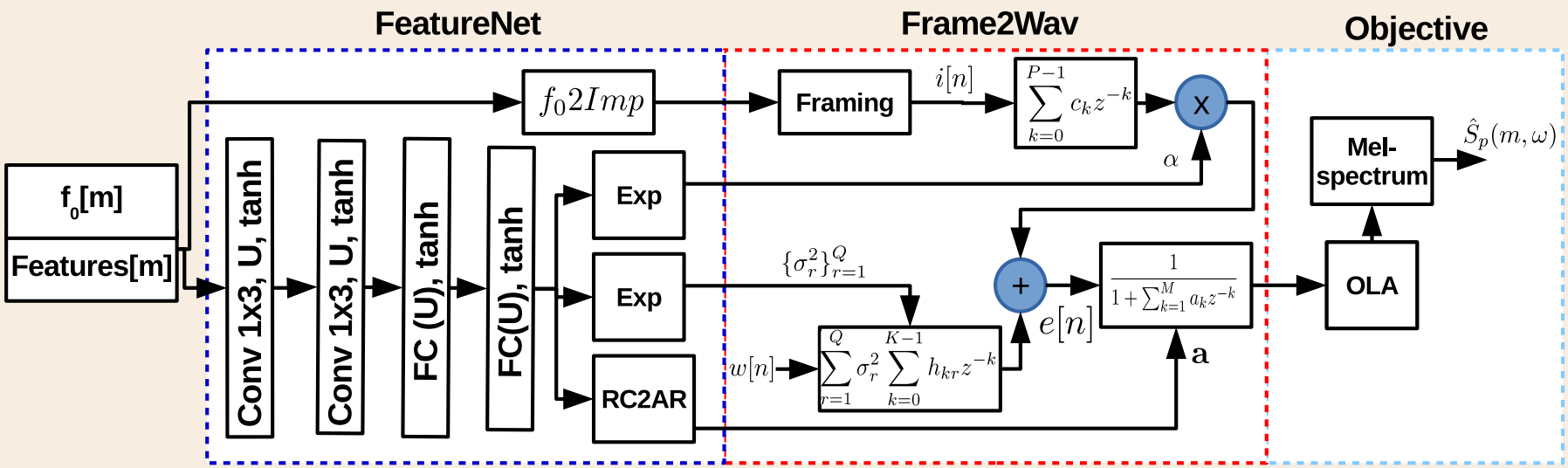
- Frame2Wav
  - We construct the impulse train by finding discontinuity in the instantaneous phase constructed using pitch contour.
  - We use the reformulated source filter model to reconstruct the waveform.

Architecture



- Objective function
  - The exact phase the signal is lost because of the white noise sampling and the impulse construction.
  - We use l_1 error between the ground truth melspec and predicted mel spec (window 20ms, shift=10ms and order 80).

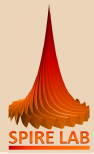
Architecture



- Complexity:
  - In the proposed network the complexity arises only from the three filtering operations and interpolation.
  - The complexity is given by $2 \times (P + 2 \times K \times Q + M) \times fs$, where fs is the sampling rate.

Overview

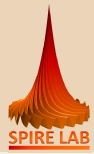
Database



- We use TSP database [6].
- approximately 4 hours of data (at a sampling rate of 16kHz)
- A total of 1400 English utterances spoken by 24 speakers (12 male, 12 female).

[6] Kabal, "TSP speech database", 2002

Experimental setup



- Speaker independent setup.
- 18 dimensional mel cepstrum, pitch correlation and pitch value is used as input.
- We augment the data to construct a 14hrs of data.
- We use a LP order 25, 40 filter of order 60 for noise filterbank.
- We experiment with $U = \{64, 32\}$.
- The noise filter is learned as a part of the training SFNet-U-L and a fixed gamma tone filter.
- We use Perceptual Evaluation of Speech Quality (PESQ) [7] and MUSHRA test[8].

[7] Hu and Loizou, "Evaluation of objective quality measures for speech enhancement", 2007

[8] Assembly, ITU Radiocommunication, "Methods for the subjective assessment of small impairments in audio systems including Multichannel Sound Systems", 1994

Results

model	#params	GFLOPs	PESQ (E)	Quality (E)
Reference	-	-	-	98.19
LPCnet-192	602k	1.2	3.25(0.25)	71.95(20.31)
SFNet-64	50k	0.2	3.82(0.15)	79.81(17.37)
SFNet-32	11k	0.2	3.80(0.14)	79.04 (18.22)
SFNet-32-L	11k	0.2	**3.85(0.14)**	**82.04(16.95)**

- PESQ/Quality score using SFNet is better than that using LPCnet-192 for unseen speakers case.

Results

model	#params	GFLOPs	PESQ (E)	Quality (E)
Reference	-	-	-	98.19
LPCnet-192	602k	1.2	3.25(0.25)	71.95(20.31)
SFNet-64	50k	0.2	3.82(0.15)	79.81(17.37)
SFNet-32	11k	0.2	3.80(0.14)	79.04 (18.22)
SFNet-32-L	11k	0.2	**3.85(0.14)**	**82.04(16.95)**

- It can be observed from the table that the learned filter-bank is better than the fixed Gammatone filter bank

Results

model	#params	GFLOPs	PESQ (E)	Quality (E)
Reference	-	-	-	98.19
LPCnet-192	602k	1.2	3.25(0.25)	71.95(20.31)
SFNet-64	50k	0.2	3.82(0.15)	79.81(17.37)
SFNet-32	11k	0.2	3.80(0.14)	79.04 (18.22)
SFNet-32-L	11k	0.2	**3.85(0.14)**	**82.04(16.95)**

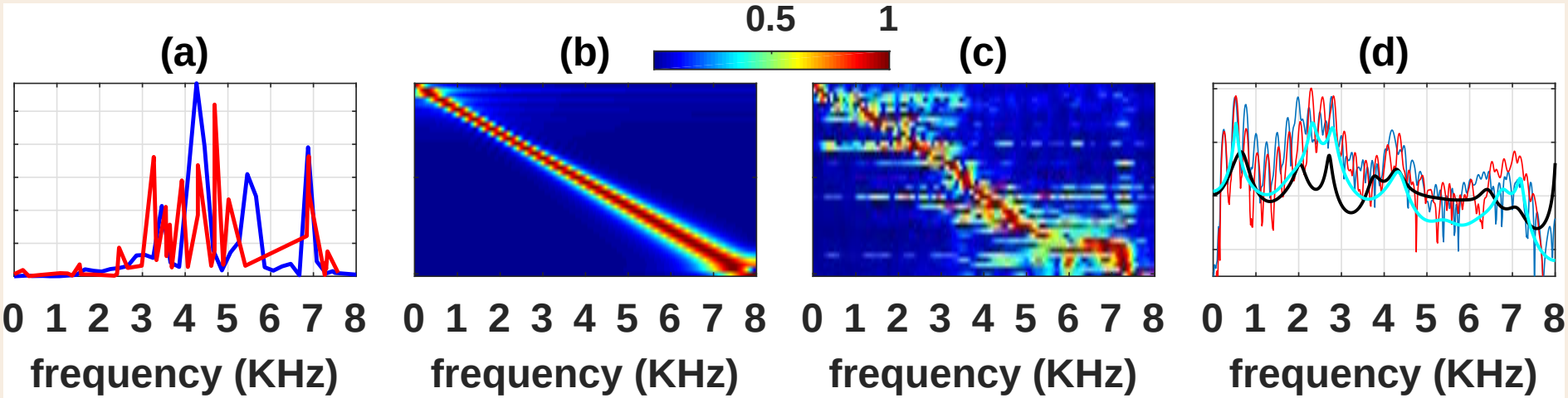
- The complexity of the SFNet is reduced significantly compared to LPCNet.

Results

model	#params	GFLOPs	PESQ (E)	Quality (E)
Reference	-	-	-	98.19
LPCnet-192	602k	1.2	3.25(0.25)	71.95(20.31)
SFNet-64	50k	0.2	3.82(0.15)	79.81(17.37)
SFNet-32	11k	0.2	3.80(0.14)	79.04 (18.22)
SFNet-32-L	11k	0.2	**3.85(0.14)**	**82.04(16.95)**

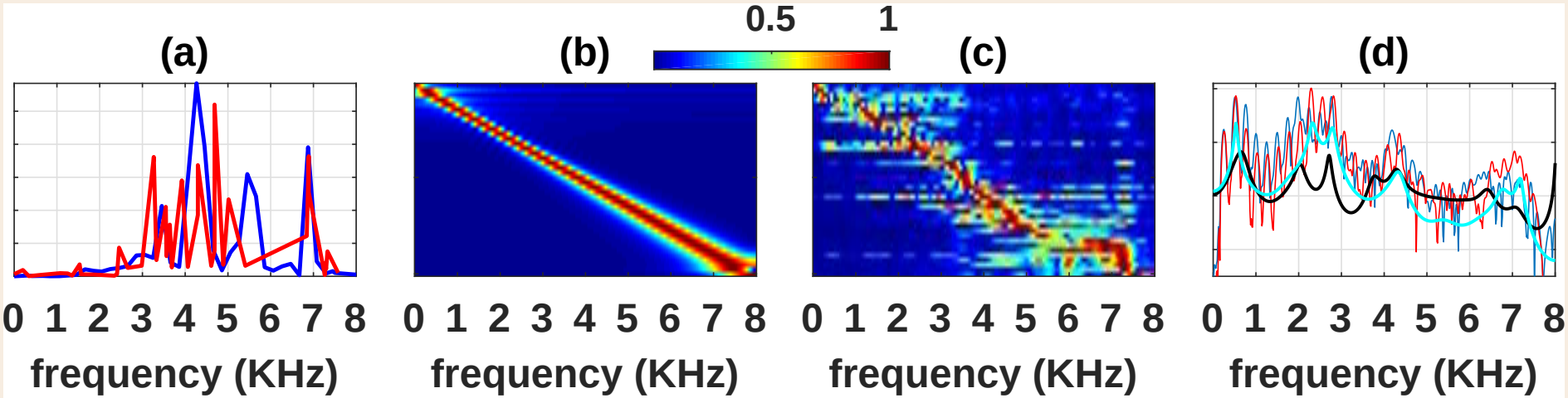
- The number of parameters are also reduced by 98% compared LPCNet.

Analysis of estimated parameters



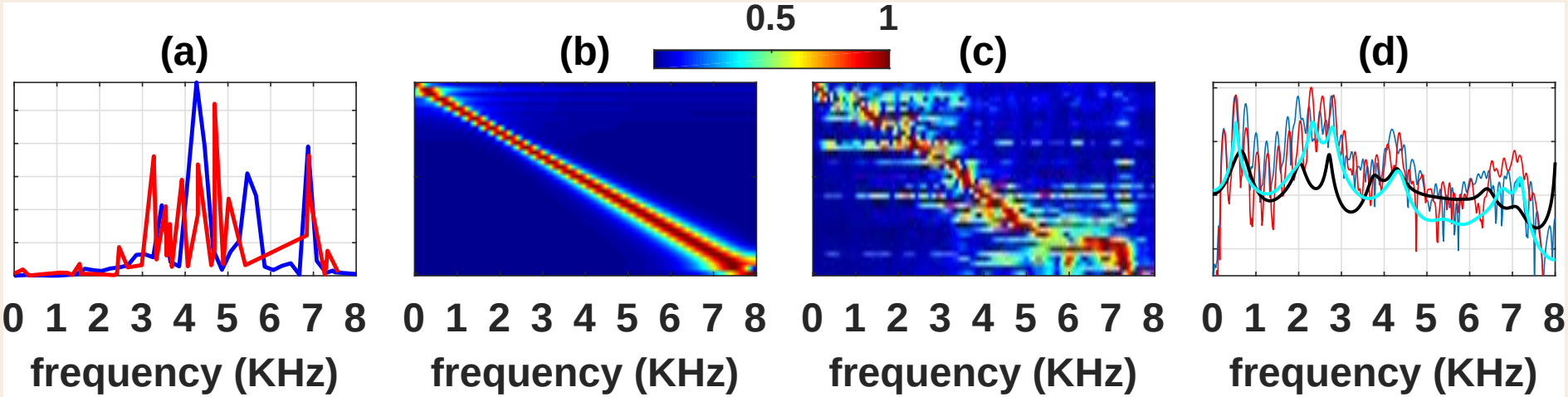
(a) The noise variance predicted at different sub-bands for SFNet-U and SFNet-U-L in the voiced regions.

Analysis of estimated parameters



(b-c) The Magnitude of frequency response of the 40 filters sorted by center frequency for SFNet-32 and SFNet-32-L.

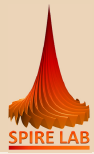
Analysis of estimated parameters



(d) Sample spectrum of a frame (blue) with spectrum of the predicted frame (red) along with the frequency response of the estimated LP filter using the SFNet (black) and Autocorrelation method (cyan)

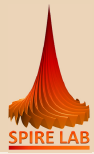
Overview

Conclusion & Future works



- Source-filter model is reformulated where the source signal is further modeled as addition of two source signals.
- We propose a neural network architecture called SFNet to predict the model parameters given a 20- dimensional speech feature vector.
- We plan to use the proposed method as a vocoder in a text-to-speech system and speech compression systems.

Acknowledgement



Authors thank the Department of Science and Technology, Govt of India for their support in this work.

Questions

Samples Demo: https://araomv.github.io/SFNet/
If you have any questions, please feel free to drop an email to **spirelab.ee@iisc.ac.in**.

THANK YOU