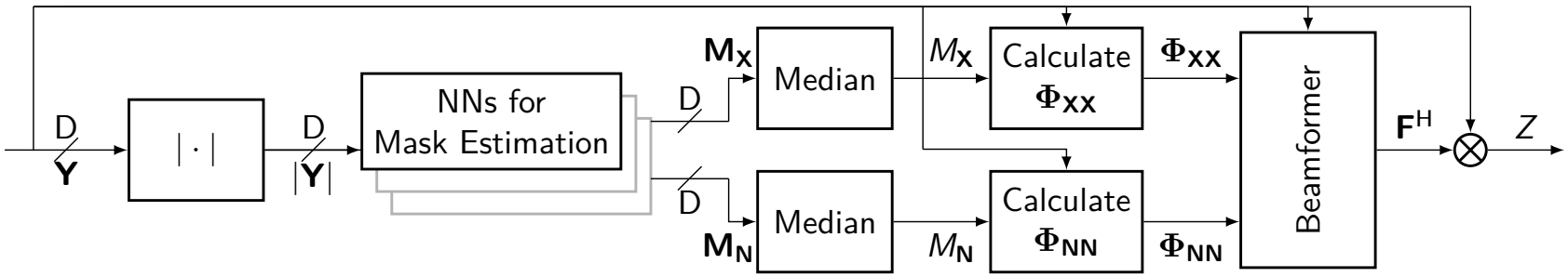
Smoothing along Frequency in Online Neural Network Supported Acoustic Beamforming

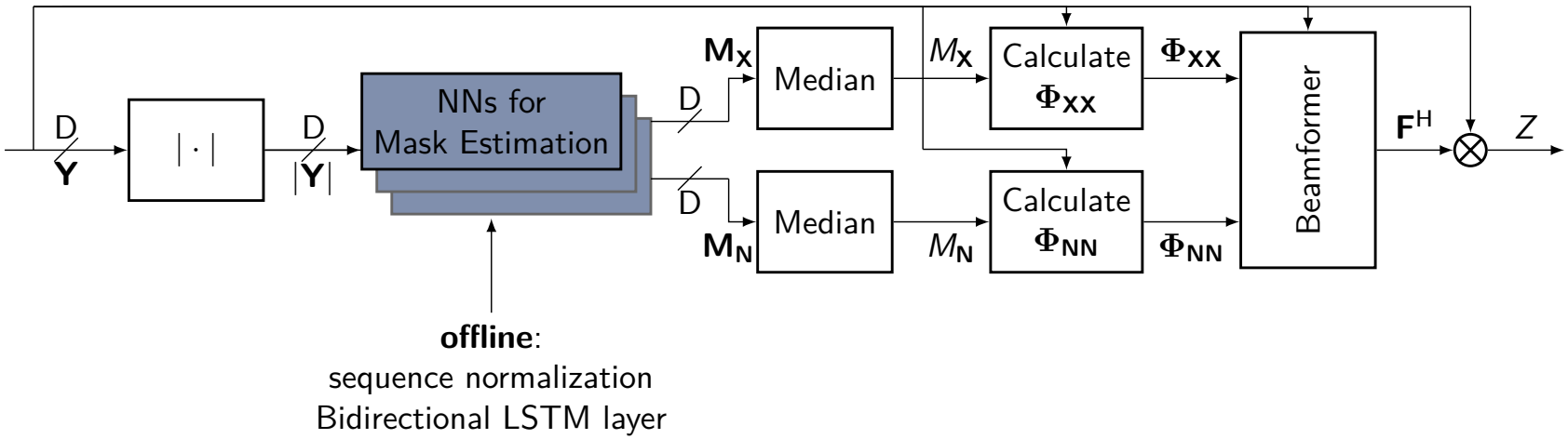
Jens Heitkaemper, Jahn Heymann, Reinhold Haeb-Umbach

Department of Communications Engineering - Paderborn University
Prof. Dr.-Ing. Reinhold Haeb-Umbach
12.10.2018

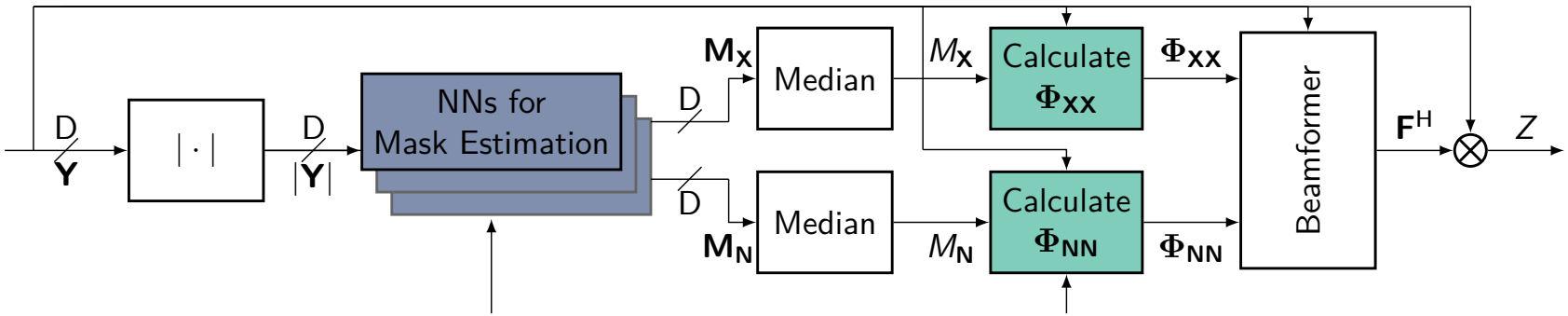
Neural Network Supported Acoustic Beamforming



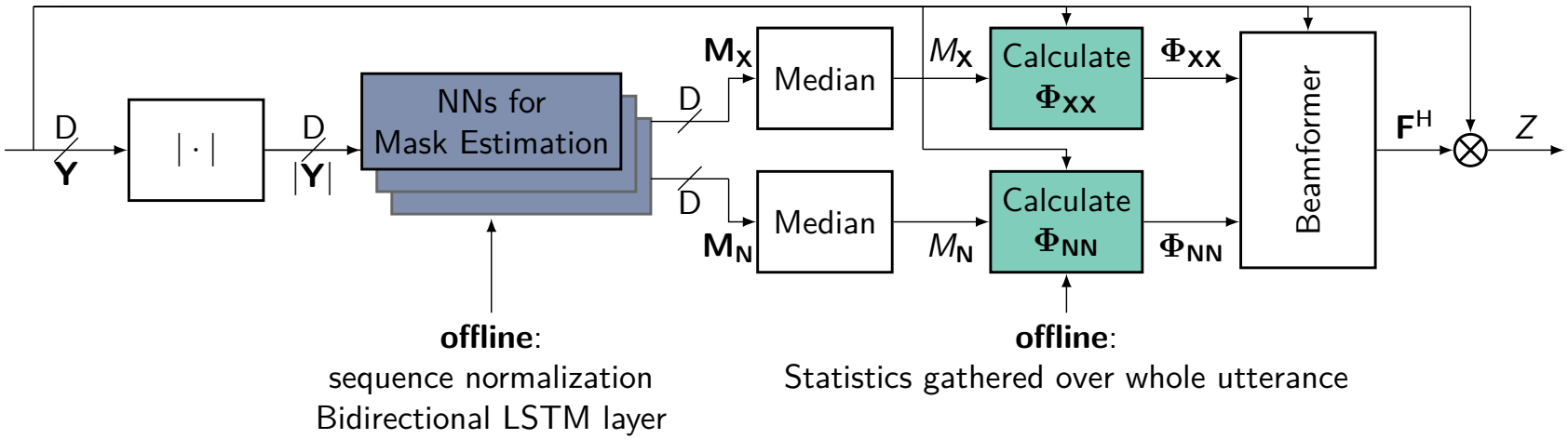
Neural Network Supported Acoustic Beamforming



Neural Network Supported Acoustic Beamforming



Neural Network Supported Acoustic Beamforming

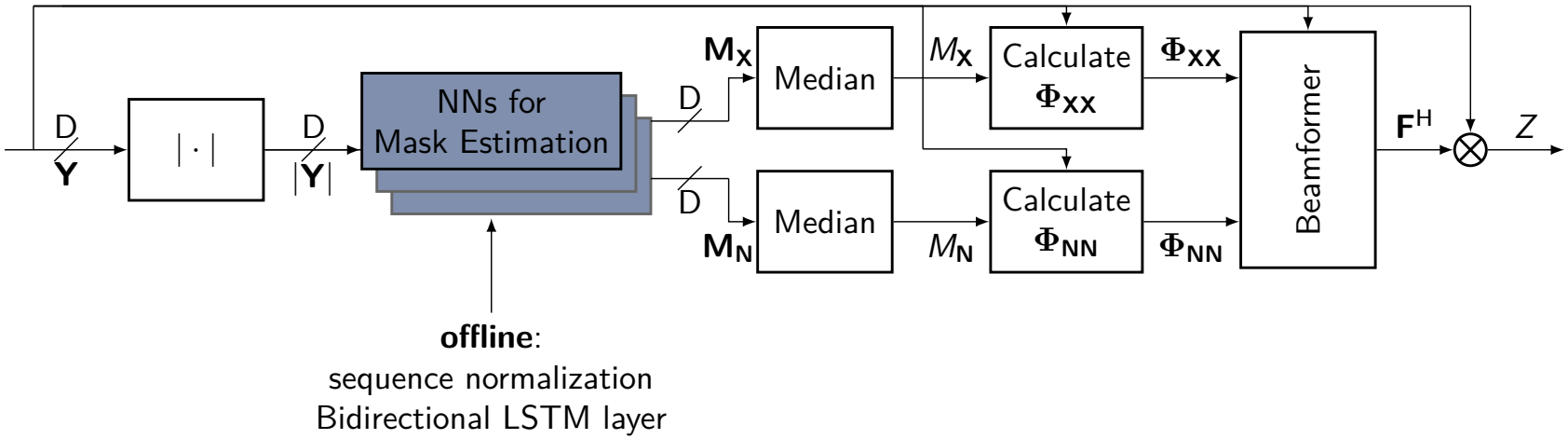


Offline ⇒ Online

- Offline: Utterance wise beamforming vector estimation
- Online: Low latency beamforming vector estimation

Table of contents

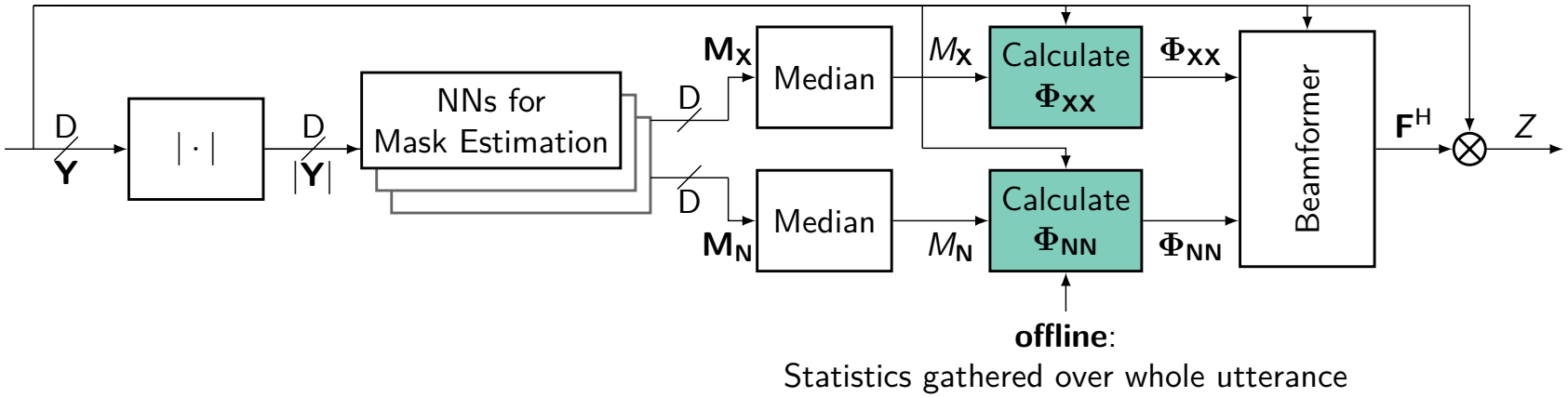
Online Mask Estimation



Low latency mask estimation

1. Bidirectional LSTM ⇒ LSTM
2. Sequence normalization ⇒ Recursive mean normalization

Online Beamforming



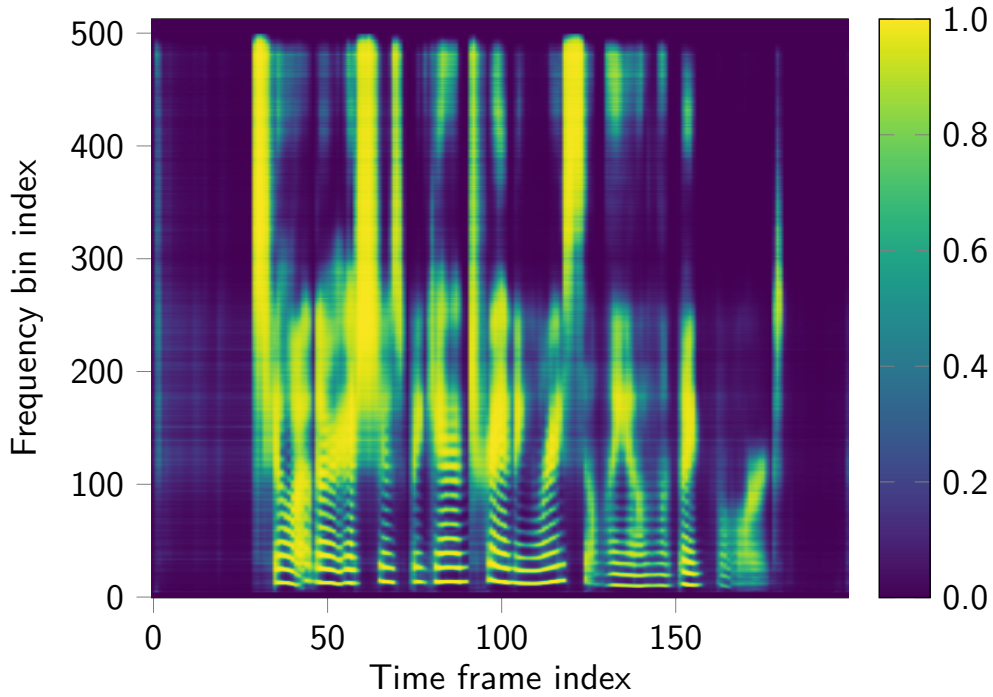
Recursive spatial convariance matrix estimation:

$$\mathbf{\Phi}_{\nu\nu}(nN) = \alpha_\nu \mathbf{\Phi}_{\nu\nu}((n-1)N) + (1-\alpha_\nu)\,\hat{\mathbf{\Phi}}_{\nu\nu}(nN)$$

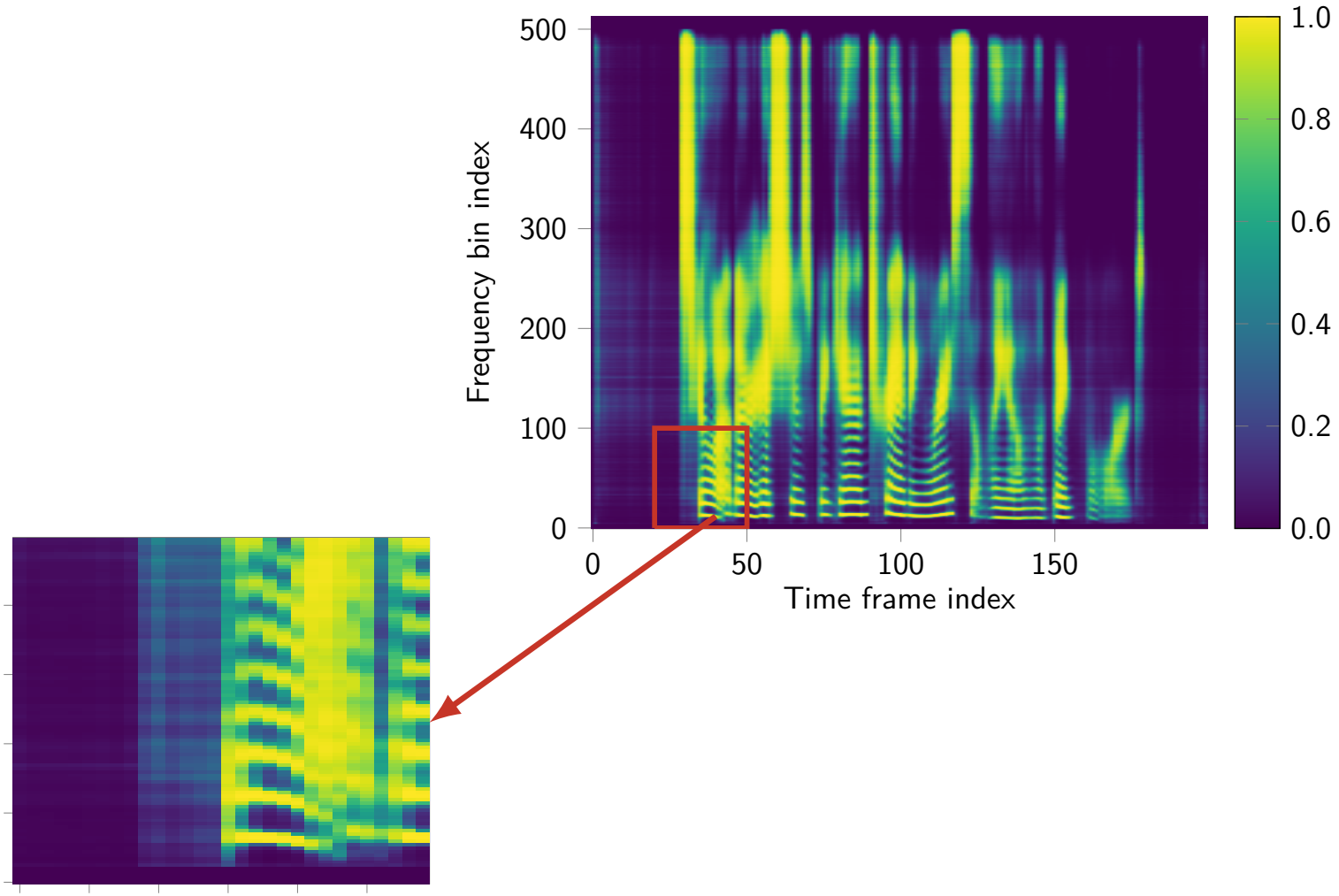
$$\hat{\mathbf{\Phi}}_{\nu\nu}(nN) = \sum_{\ell=0}^{N-1} M_\nu(nN-\ell)\mathbf{Y}(nN-\ell)\mathbf{Y}^{\mathsf{H}}(nN-\ell) \text{ with } \nu \in [\mathbf{N}, \mathbf{X}]$$

Table of contents

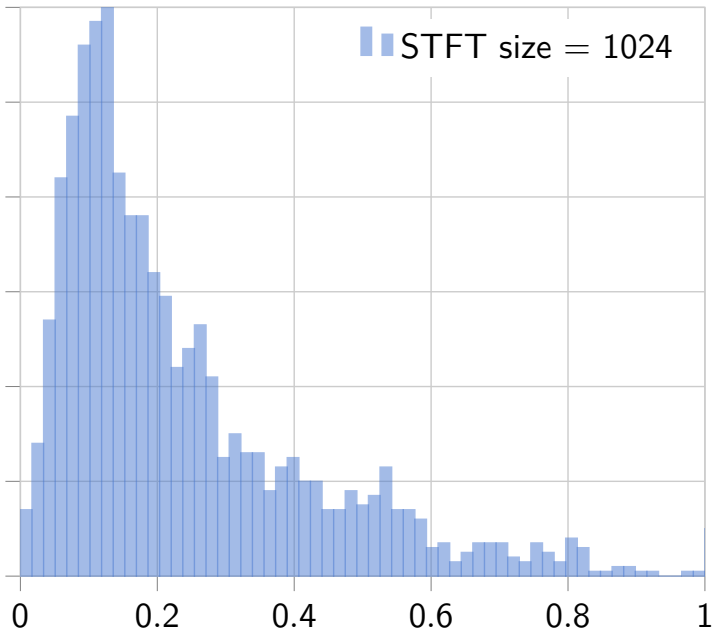
Estimated masks



Estimated masks

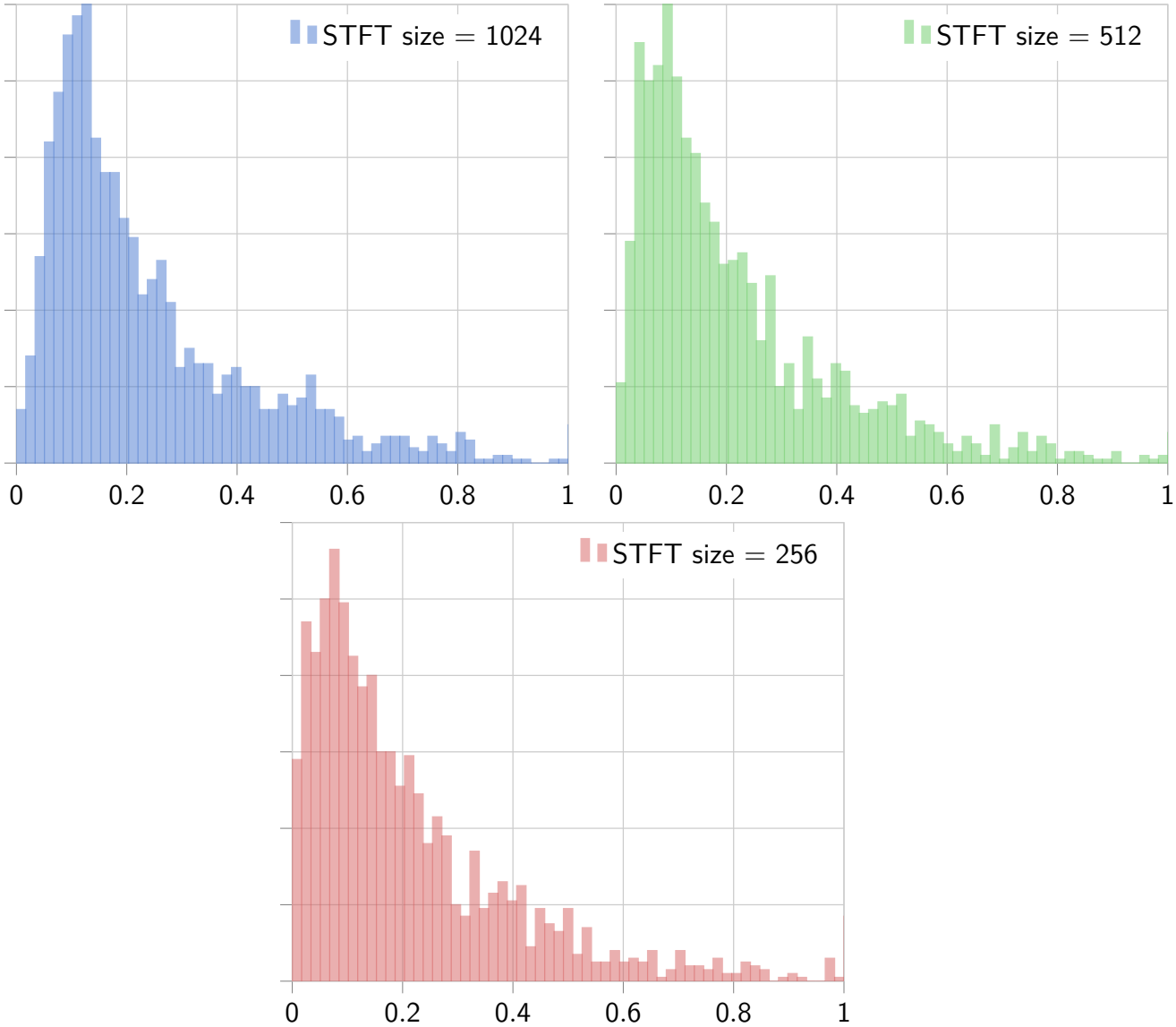


Speech activity in the first frames



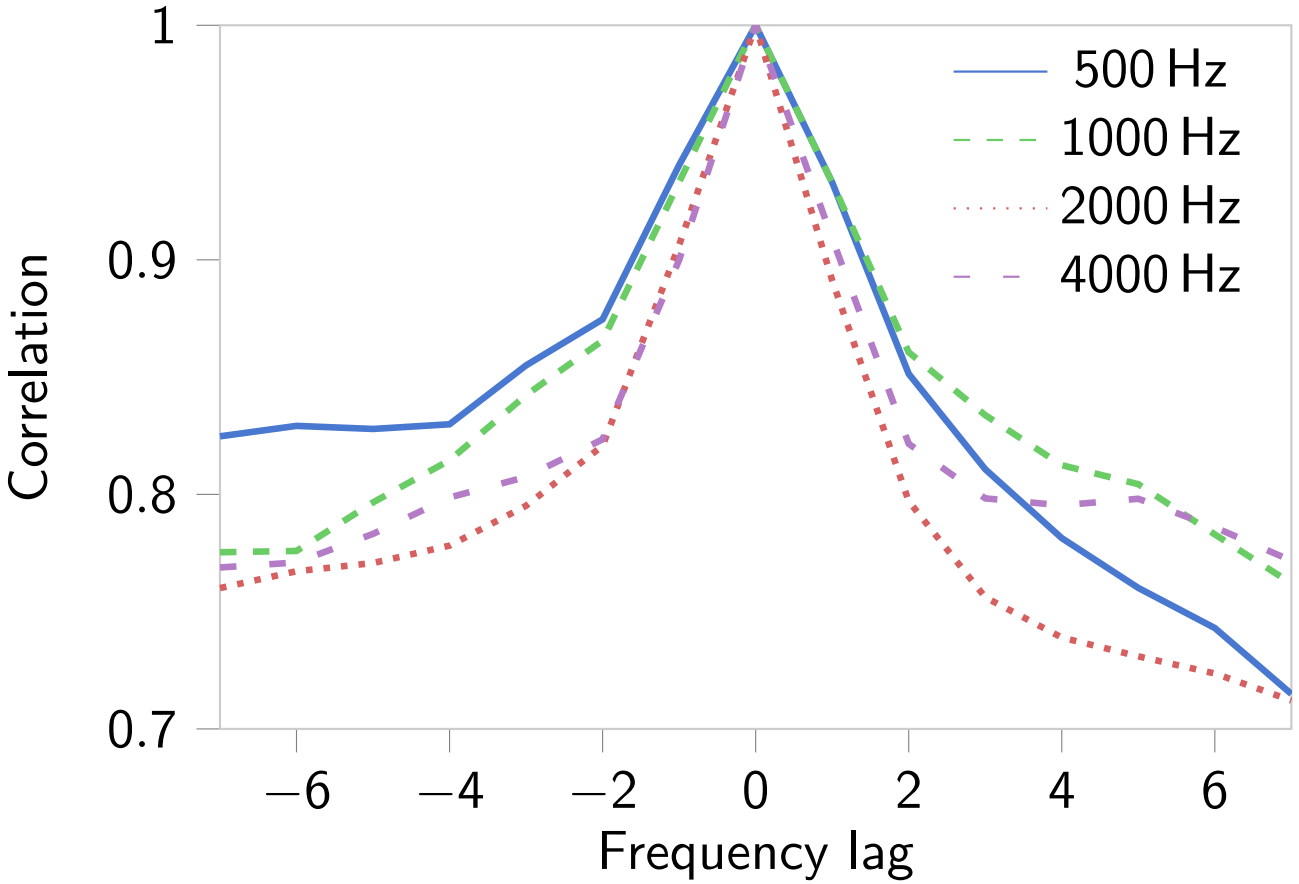
Number of seconds until the first observed speech activity per fbin averaged over all fbins

Speech activity in the first frames

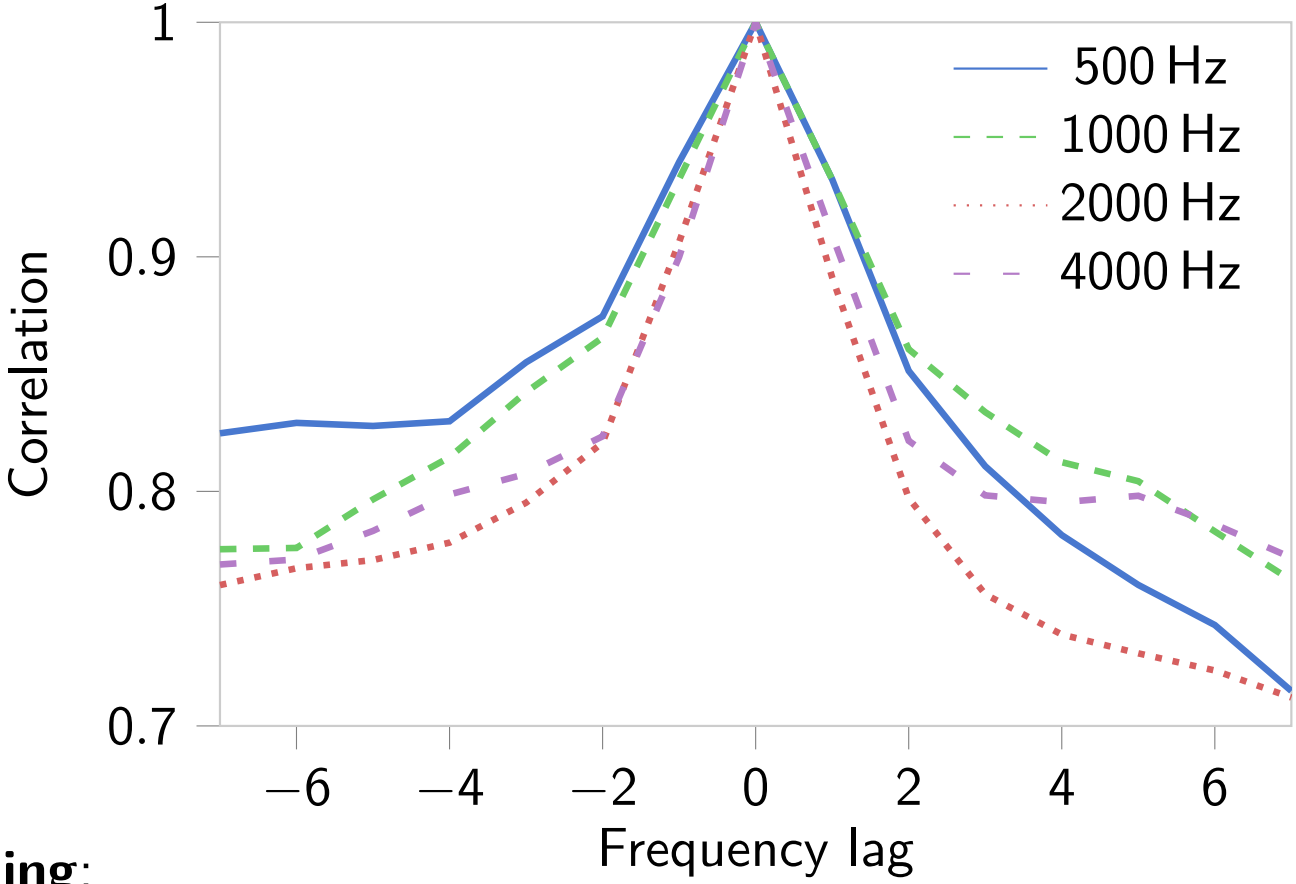


Number of seconds until the first observed speech activity per fbin averaged over all fbins

Frequency Smoothing



Frequency Smoothing



Smoothing:

$$\tilde{\mathbf{F}}(\ell,k) = \frac{\sum_{i=-\lfloor \frac{\kappa}{2} \rfloor}^{\lfloor \frac{\kappa}{2} \rfloor} \tilde{M}_X(\ell,k+i)\mathbf{F}(\ell,k+i)}{\sum_{i=-\lfloor \frac{\kappa}{2} \rfloor}^{\lfloor \frac{\kappa}{2} \rfloor} \tilde{M}_X(\ell,k+i)} \text{ with } \tilde{M}_X = (\ell,k) = \sum_{\iota=0}^{\ell} M_X(\iota,k)$$

Table of contents

Databases

CHiME-4

- Simulated and real data recordings
- 6 channels
- 4 types of real background noise

WSJ-VoiceHome

- Simulated data: WSJ speech data convolved with RIR from VoiceHome
- 8 channels
- dynamic and realistic background noise between 0 dB and 10 dB (Includes speech like noise for example TV, french Dialog etc.)

CHiME-4 STFT sizes

Method	STFT size	WER simu	WER real
channel #5	–	13.36	18.60
Offline	1024	8.39	10.01
Online	1024	9.37	13.32
	512	8.70	12.66
	256	**8.54**	**11.71**
	128	9.57	12.14

- As expected: online beamforming slightly worse than offline
- WER results extremly dependent on STFT size

CHiME-4 frequency smoothing

Method	STFT size	Smoothing	WER simu	WER real
channel #5	–	–	13.36	18.60
Offline	1024	–	8.39	10.01
Online	1024	–	9.37	13.32
		✓	8.70	12.01
	256	–	8.54	**11.71**
		✓	**8.29**	11.89

Spectral smoothing :

- reduces the effect for higher STFT size
- Only small effects on lower STFT sizes

WSJ-VoiceHome

Method	STFT size	Smooth.	noise all	noise other	noise sp-like
ch #1	–	–	25.73	18.42	28.30
Offline	1024	–	13.30	12.70	13.52
Online	1024	–	17.70	17.01	17.94
		✓	**16.85**	**15.10**	**17.47**
	256	–	17.96	16.85	18.36

- Small STFT sizes seem not to work with speech-like noise
- Spectral smoothing works on speech-like noise

Conclusion

- Block online neural network supported beamforming depends on stable initialization
- Problem may be reduced by spectral smoothing or reducing the STFT size
- Optimal STFT dependent on the noise scenario

Next steps

- Evaluating the system on databases with moving speakers

Method	STFT size	WER	
		simu	real
channel #5	–	13.36	18.60
Offline	1024	8.39	10.01
	256	8.05	10.22
Online	1024	9.37	13.32
	1024, 256 window	8.67	12.52
	512	8.70	12.66
	256	8.54	11.71
	128	9.57	12.14

Method	STFT size	Smoothing	WER [%] simu	WER [%] real
Offline	1024	–	8.39	10.01
	1024	✓	8.39	10.11
Online	1024	–	9.37	13.32
	256	–	8.54	11.71
	1024	✓	8.70	12.01

Method	STFT size	Smoothing	WER	
			near	far
ch #1			20.82	23.6
Offline	1024	–	12.78	14.15
Online	1024	–	18.40	18.84
	256	–	16.64	17.02
	1024	✓	17.08	17.37