# DNN-based Speech Mask Estimation for Eigenvector Beamforming

**Lukas Pfeifenberger, Matthias Zöhrer, and Franz Pernkopf**
Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria

# Motivation

- Boost beamforming performance using NNs
- Replace Direction-Of-Arrival estimate by a speech mask
- Use speech mask to construct the MVDR, GSC and GEV Beamformers, and a Postfilter
- Speech mask can be learned from eigenvector features




Sources:
www.maloyalaser.com
GeorgHH via Wikimedia Commons
www.polycom.com
www.amazon.com

# Why use a speech mask?

- Direction-Of-Arrival estimate:
  - Direct-path steering vector
  - Target leakage may occur

# Why use a speech mask?

- Direction-Of-Arrival estimate:
  - Direct-path steering vector
  - Target leakage may occur

- Speech mask:
  - Multi-path steering vector (models reverberation)
  - Sufficient to construct Beamformer + Postfilter
  - Existing estimation approaches: use magnitude features [Erdogan et al., 2016] and [Heymann et al., 2016]

# Why use a speech mask?

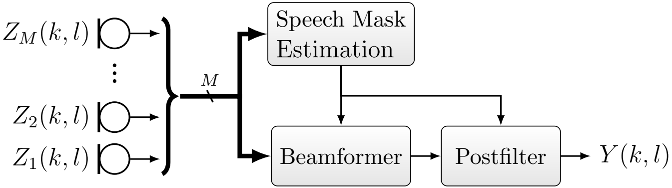
- Direction-Of-Arrival estimate:
  - Direct-path steering vector
  - Target leakage may occur
- Speech mask:
  - Multi-path steering vector (models reverberation)
  - Sufficient to construct Beamformer + Postfilter
  - Existing estimation approaches: use magnitude features [Erdogan et al., 2016] and [Heymann et al., 2016]
- Our idea:
  - Use eigenvector features
  - Exploit spatial information
  - Independent from array geometry and signal energy

# Outline

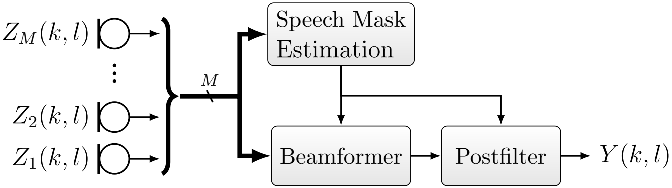
1. System Model
2. Super-directive Beamforming: MVDR, GSC, GEV
3. Speech Mask Estimation
4. Experiments

# System Model

# System Model



# System Model




- Single speech source: $S(k,l)$
- Multi-path ATF: $\boldsymbol{A}(k,l)$
- Unknown noise: $\boldsymbol{N}(k,l)$
- System model: $\boldsymbol{Z}(k,l) = S(k,l)\boldsymbol{A}(k,l) + \boldsymbol{N}(k,l)$
- PSDs: $\boldsymbol{\Phi}_{ZZ} = \boldsymbol{\Phi}_{SS} + \boldsymbol{\Phi}_{NN} = \boldsymbol{A}\boldsymbol{A}^H \Phi_S + \boldsymbol{\Phi}_{NN}$

# Super-directive Beamforming: MVDR, GSC, GEV

# MVDR

- Optimal MWF = MVDR + Wiener Postfilter: [Vary and Martin, 2006]
  - $\boldsymbol{W}_{OPT} = \boldsymbol{\Phi}_{ZZ}^{-1}\boldsymbol{A}\Phi_S = \underbrace{\frac{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}{\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}}_{\boldsymbol{w}_{MVDR}} \cdot \underbrace{\frac{\Phi_S}{\Phi_S + \left[\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}\right]^{-1}}}_{G=\frac{\xi}{1+\xi}}$

# MVDR

- Optimal MWF = MVDR + Wiener Postfilter: [Vary and Martin, 2006]
  - $\boldsymbol{W}_{OPT} = \boldsymbol{\Phi}_{ZZ}^{-1}\boldsymbol{A}\Phi_S = \underbrace{\frac{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}{\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}}_{\boldsymbol{w}_{MVDR}} \cdot \underbrace{\frac{\Phi_S}{\Phi_S + \left[\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}\right]^{-1}}}_{G=\frac{\xi}{1+\xi}}$
- Substitute ATF $\boldsymbol{A}$ by steering vector $\boldsymbol{F}$:
  - use dominant Eigenvector: $\boldsymbol{F} \rightarrow \boldsymbol{v}_{S_1}$ [Pfeifenberger et al., 2016]
  - EVD of the speech PSD: $\boldsymbol{\Phi}_{SS} = \sum_{m=1}^{M} \boldsymbol{v}_{S_m}\boldsymbol{v}_{S_m}^H \lambda_{S_m}$
  - includes multi-path propagation: $\boldsymbol{F} = \boldsymbol{A}\left[\frac{\phi_S}{\lambda_{S_1}\boldsymbol{A}^H\boldsymbol{v}_{S_1}}\right]$

# MVDR

- Optimal MWF = MVDR + Wiener Postfilter: [Vary and Martin, 2006]
  - $\boldsymbol{W}_{OPT} = \boldsymbol{\Phi}_{ZZ}^{-1}\boldsymbol{A}\Phi_S = \underbrace{\frac{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}{\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}}_{\boldsymbol{W}_{MVDR}} \cdot \underbrace{\frac{\Phi_S}{\Phi_S + \left[\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}\right]^{-1}}}_{G=\frac{\xi}{1+\xi}}$
- Substitute ATF $\boldsymbol{A}$ by steering vector $\boldsymbol{F}$:
  - use dominant Eigenvector: $\boldsymbol{F} \rightarrow \boldsymbol{v}_{S_1}$ [Pfeifenberger et al., 2016]
  - EVD of the speech PSD: $\boldsymbol{\Phi}_{SS} = \sum_{m=1}^{M} \boldsymbol{v}_{S_m}\boldsymbol{v}_{S_m}^H \lambda_{S_m}$
  - includes multi-path propagation: $\boldsymbol{F} = \boldsymbol{A}\left[\frac{\phi_S}{\lambda_{S_1}\boldsymbol{A}^H\boldsymbol{v}_{S_1}}\right]$
- Postfilter $G$:
  - uses multi-channel SNR: $\xi = \mathrm{Tr}\left\{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{\Phi}_{SS}\right\}$

# MVDR

- Optimal MWF = MVDR + Wiener Postfilter: [Vary and Martin, 2006]
  - $\boldsymbol{W}_{OPT} = \boldsymbol{\Phi}_{ZZ}^{-1}\boldsymbol{A}\Phi_S = \underbrace{\frac{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}{\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}}}_{\boldsymbol{w}_{MVDR}} \cdot \underbrace{\frac{\Phi_S}{\Phi_S + \left[\boldsymbol{A}^H\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{A}\right]^{-1}}}_{G=\frac{\xi}{1+\xi}}$

- Substitute ATF $\boldsymbol{A}$ by steering vector $\boldsymbol{F}$:
  - use dominant Eigenvector: $\boldsymbol{F} \rightarrow \boldsymbol{v}_{S_1}$ [Pfeifenberger et al., 2016]
  - EVD of the speech PSD: $\boldsymbol{\Phi}_{SS} = \sum_{m=1}^{M} \boldsymbol{v}_{S_m}\boldsymbol{v}_{S_m}^H \lambda_{S_m}$
  - includes multi-path propagation: $\boldsymbol{F} = \boldsymbol{A}\left[\frac{\phi_S}{\lambda_{S_1}\boldsymbol{A}^H\boldsymbol{v}_{S_1}}\right]$

- Postfilter $G$:
  - uses multi-channel SNR: $\xi = \mathrm{Tr}\left\{\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{\Phi}_{SS}\right\}$

Required: $\boldsymbol{\Phi}_{SS}(k,l)$ and $\boldsymbol{\Phi}_{NN}(k,l)$

# GSC

- Split the MVDR into two orthogonal components:
  - $\boldsymbol{W}_{MVDR} \approx \boldsymbol{W}_{GSC} = \boldsymbol{F} - \boldsymbol{B}\boldsymbol{H}_{AIC}$ [Hoshuyama et al., 1999]

# GSC

- Split the MVDR into two orthogonal components:
  - $\boldsymbol{W}_{MVDR} \approx \boldsymbol{W}_{GSC} = \boldsymbol{F} - \boldsymbol{B}\boldsymbol{H}_{AIC}$ [Hoshuyama et al., 1999]
- Steering Vector $\boldsymbol{F}$:
  - distortionless response: $\boldsymbol{F}^H \boldsymbol{A} \stackrel{!}{=} 1$

# GSC

- Split the MVDR into two orthogonal components:
  - $\boldsymbol{W}_{MVDR} \approx \boldsymbol{W}_{GSC} = \boldsymbol{F} - \boldsymbol{B}\boldsymbol{H}_{AIC}$ [Hoshuyama et al., 1999]
- Steering Vector $\boldsymbol{F}$:
  - distortionless response: $\boldsymbol{F}^H \boldsymbol{A} \stackrel{!}{=} 1$
- Blocking Matrix $\boldsymbol{B}$:
  - steers ”nulls” towards speaker: $\boldsymbol{B}^H \boldsymbol{A} \stackrel{!}{=} \mathbf{0}$
  - i.e.: $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}^H$ [Shmulik et al., 2012]

# GSC

- Split the MVDR into two orthogonal components:
  - $\boldsymbol{W}_{MVDR} \approx \boldsymbol{W}_{GSC} = \boldsymbol{F} - \boldsymbol{B}\boldsymbol{H}_{AIC}$ [Hoshuyama et al., 1999]
- Steering Vector $\boldsymbol{F}$:
  - distortionless response: $\boldsymbol{F}^H \boldsymbol{A} \stackrel{!}{=} 1$
- Blocking Matrix $\boldsymbol{B}$:
  - steers ”nulls” towards speaker: $\boldsymbol{B}^H \boldsymbol{A} \stackrel{!}{=} \mathbf{0}$
  - i.e.: $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}^H$ [Shmulik et al., 2012]
- Adaptive Interference Canceller $\boldsymbol{H}_{AIC}$:
  - adapted during speech absence using NLMS

# GSC

- Split the MVDR into two orthogonal components:
  - $\boldsymbol{W}_{MVDR} \approx \boldsymbol{W}_{GSC} = \boldsymbol{F} - \boldsymbol{B}\boldsymbol{H}_{AIC}$ [Hoshuyama et al., 1999]
- Steering Vector $\boldsymbol{F}$:
  - distortionless response: $\boldsymbol{F}^H\boldsymbol{A} \stackrel{!}{=} 1$
- Blocking Matrix $\boldsymbol{B}$:
  - steers "nulls" towards speaker: $\boldsymbol{B}^H\boldsymbol{A} \stackrel{!}{=} \mathbf{0}$
  - i.e.: $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}^H$ [Shmulik et al., 2012]
- Adaptive Interference Canceller $\boldsymbol{H}_{AIC}$:
  - adapted during speech absence using NLMS

Required: $\boldsymbol{\Phi}_{SS}(k, l)$

# GEV

- Maximizes the SNR $\xi$: [Warsitz and Haeb-Umbach, 2007]
  - $\boldsymbol{W}_{SNR} = \arg \underset{\boldsymbol{W}}{max}\, \xi$
  - eigenvalue problem (rank = 1): $\boldsymbol{\Phi}_{NN}^{-1} \boldsymbol{\Phi}_{SS} \boldsymbol{W}_{SNR} = \xi \boldsymbol{W}_{SNR}$

# GEV

- Maximizes the SNR $\xi$: [Warsitz and Haeb-Umbach, 2007]
  - $\boldsymbol{W}_{SNR} = \arg \underset{\boldsymbol{W}}{max}\, \xi$
  - eigenvalue problem (rank = 1): $\boldsymbol{\Phi}_{NN}^{-1}\boldsymbol{\Phi}_{SS}\boldsymbol{W}_{SNR} = \xi \boldsymbol{W}_{SNR}$
- Modification for reduced distortion: [Pfeifenberger et al., 2016]
  - $\boldsymbol{W}_{GEV} = \boldsymbol{P}\boldsymbol{F}$
  - reduced distortions: $\boldsymbol{W}_{GEV}^{H}\boldsymbol{A} \approx 1$
  - projection matrix: $\boldsymbol{P} = \frac{\boldsymbol{\Phi}_{NN}\boldsymbol{W}_{SNR}\boldsymbol{W}_{SNR}^{H}}{\boldsymbol{W}_{SNR}^{H}\boldsymbol{\Phi}_{NN}\boldsymbol{W}_{SNR}}$

# GEV

- Maximizes the SNR $\xi$: [Warsitz and Haeb-Umbach, 2007]
  - $\boldsymbol{W}_{SNR} = \arg \underset{\boldsymbol{W}}{max} \, \xi$
  - eigenvalue problem (rank = 1): $\boldsymbol{\Phi}_{NN}^{-1} \boldsymbol{\Phi}_{SS} \boldsymbol{W}_{SNR} = \xi \boldsymbol{W}_{SNR}$
- Modification for reduced distortion: [Pfeifenberger et al., 2016]
  - $\boldsymbol{W}_{GEV} = \boldsymbol{P}\boldsymbol{F}$
  - reduced distortions: $\boldsymbol{W}_{GEV}^{H} \boldsymbol{A} \approx 1$
  - projection matrix: $\boldsymbol{P} = \frac{\boldsymbol{\Phi}_{NN} \boldsymbol{W}_{SNR} \boldsymbol{W}_{SNR}^{H}}{\boldsymbol{W}_{SNR}^{H} \boldsymbol{\Phi}_{NN} \boldsymbol{W}_{SNR}}$

Required: $\boldsymbol{\Phi}_{SS}(k,l)$ and $\boldsymbol{\Phi}_{NN}(k,l)$

# Speech Mask Estimation

# Speech Mask Estimation

- Speech and noise PSD estimates: [Higuchi et al., 2016]
  - $\hat{\mathbf{\Phi}}_{SS}(k,l) = \frac{\sum_{t=l}^{l+T} \boldsymbol{Z}(k,t)\boldsymbol{Z}^H(k,t)p_{\mathsf{SPP}}(k,t)}{\sum_{t=l}^{l+T} p_{\mathsf{SPP}}(k,t)}$
  - $\hat{\mathbf{\Phi}}_{NN}(k,l) = \frac{\sum_{t=l}^{l+T} \boldsymbol{Z}(k,t)\boldsymbol{Z}^H(k,t)\left(1-p_{\mathsf{SPP}}(k,t)\right)}{\sum_{t=l}^{l+T} \left(1-p_{\mathsf{SPP}}(k,t)\right)}$

# Speech Mask Estimation

- Speech and noise PSD estimates: [Higuchi et al., 2016]
  - $\hat{\mathbf{\Phi}}_{SS}(k,l) = \frac{\sum_{t=l}^{l+T} \boldsymbol{Z}(k,t)\boldsymbol{Z}^H(k,t) p_{\mathsf{SPP}}(k,t)}{\sum_{t=l}^{l+T} p_{\mathsf{SPP}}(k,t)}$
  - $\hat{\mathbf{\Phi}}_{NN}(k,l) = \frac{\sum_{t=l}^{l+T} \boldsymbol{Z}(k,t)\boldsymbol{Z}^H(k,t)\left(1-p_{\mathsf{SPP}}(k,t)\right)}{\sum_{t=l}^{l+T}\left(1-p_{\mathsf{SPP}}(k,t)\right)}$

- Speech presence probability $p_{\mathsf{SPP}}$:
  - ground truth: $p_{\mathsf{SPP,opt}} = \frac{\xi}{1+\xi}$
  - equal to the Wiener postfilter $G$

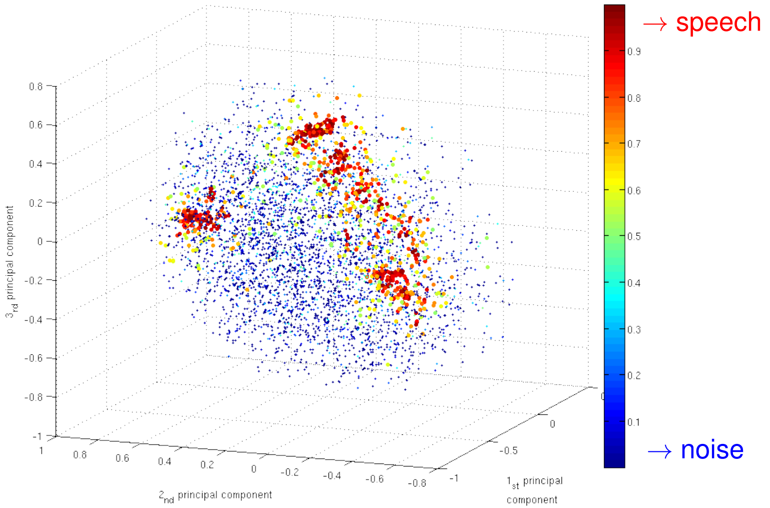
# Speech Mask Estimation

- Speech and noise PSD estimates: [Higuchi et al., 2016]
  - $\hat{\mathbf{\Phi}}_{SS}(k,l) = \frac{\sum_{t=l}^{l+T} \boldsymbol{Z}(k,t)\boldsymbol{Z}^H(k,t) p_{\text{SPP}}(k,t)}{\sum_{t=l}^{l+T} p_{\text{SPP}}(k,t)}$
  - $\hat{\mathbf{\Phi}}_{NN}(k,l) = \frac{\sum_{t=l}^{l+T} \boldsymbol{Z}(k,t)\boldsymbol{Z}^H(k,t) \left(1 - p_{\text{SPP}}(k,t)\right)}{\sum_{t=l}^{l+T} \left(1 - p_{\text{SPP}}(k,t)\right)}$
- Speech presence probability $p_{\text{SPP}}$:
  - ground truth: $p_{\text{SPP,opt}} = \frac{\xi}{1+\xi}$
  - equal to the Wiener postfilter $G$

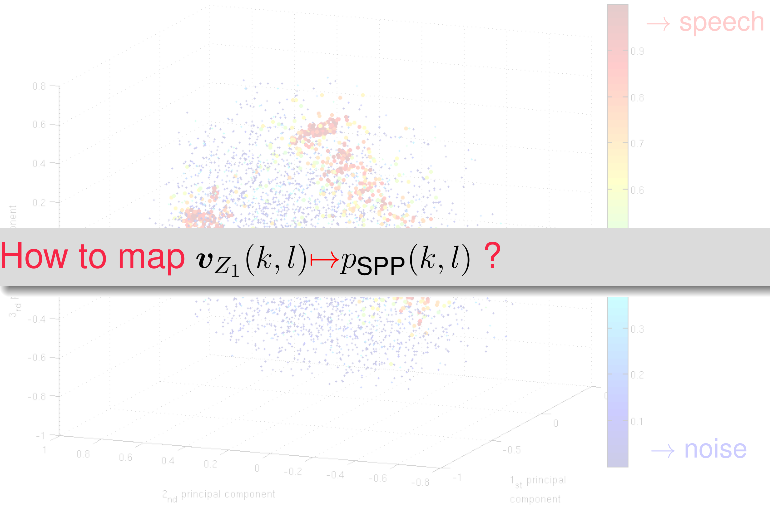
Observation: $p_{\text{SPP,opt}}$ is related to the dominant Eigenvector $\boldsymbol{v}_{Z_1}$

EVD of the noisy speech PSD: $\mathbf{\Phi}_{ZZ} = \sum_{m=1}^{M} \boldsymbol{v}_{Z_m} \boldsymbol{v}_{Z_m}^H \lambda_{Z_m}$

# Distribution of $\boldsymbol{v}_{Z_1}(k,l)$ colored with $p_{\text{SPP,opt}}(k,l)$, for $k \approx 2650Hz$:

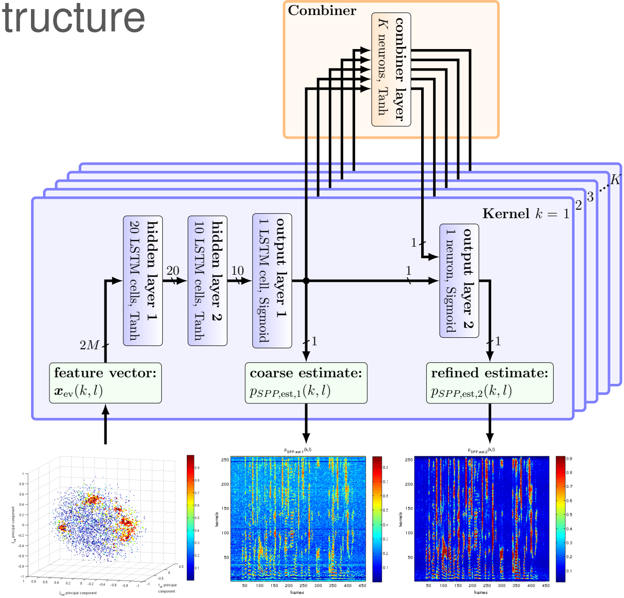


Distribution of $\boldsymbol{v}_{Z_1}(k,l)$ colored with $p_{\text{SPP,opt}}(k,l)$, for $k \approx 2650Hz$:




How to map $\boldsymbol{v}_{Z_1}(k,l) \mapsto p_{\text{SPP}}(k,l)$ ?

# NN structure



# Experiments

# Experiments

- Feature vector variants:
  - Eigenvectors: $\boldsymbol{x}_{\text{ev}}(k,l) = \left[\text{Re}\{\boldsymbol{v}_{Z_1}(k,l)\}^T, \text{Im}\{\boldsymbol{v}_{Z_1}(k,l)\}^T\right]^T$
  - Eigenvector-deltas: $\boldsymbol{x}_{\text{evd}}(k,l) = |\boldsymbol{v}_{Z_1}(k,l)^H \boldsymbol{v}_{Z_1}(k,l+\Delta)|$
  - Energy per channel: $\boldsymbol{x}_{\text{psd}}(k,l,m) = 20log_{10}|Z_m(k,l)|$

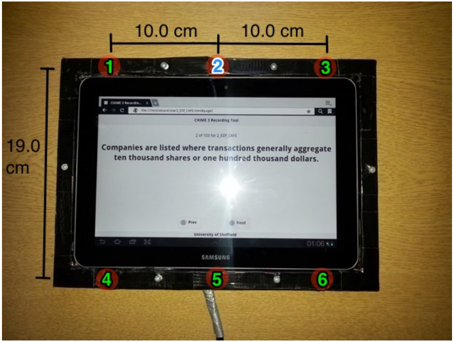
# Experiments

- Feature vector variants:
  - Eigenvectors: $\boldsymbol{x}_{\text{ev}}(k,l) = \left[\text{Re}\{\boldsymbol{v}_{Z_1}(k,l)\}^T, \text{Im}\{\boldsymbol{v}_{Z_1}(k,l)\}^T\right]^T$
  - Eigenvector-deltas: $\boldsymbol{x}_{\text{evd}}(k,l) = |\boldsymbol{v}_{Z_1}(k,l)^H \boldsymbol{v}_{Z_1}(k,l+\Delta)|$
  - Energy per channel: $\boldsymbol{x}_{\text{psd}}(k,l,m) = 20log_{10}|Z_m(k,l)|$

- NN variants:
  - ev_lstm: LSTM cells + $\boldsymbol{x}_{\text{ev}}(k,l)$ features
  - evd_lstm: LSTM cells + $\boldsymbol{x}_{\text{evd}}(k,l)$ features
  - evd_mlp: FF layers + $\boldsymbol{x}_{\text{evd}}(k,l)$ features
  - psd_lstm: LSTM cells + $\boldsymbol{x}_{\text{psd}}(k,l)$ features

# Training data: CHiME4 corpus [Barker et al., 2015]

- 2 and 6 channel data
- 14658 utterances
- 4 background noise types: BUS, STR, PED, CAF
- 12 speakers
- provides ground truth $\xi(k, l)$ for training the NN



# Speech mask prediction error

| *architecture* | $n_\triangle$ | $n_h$ | *prediction error [%]* | | | *# of weights* |
|---|---|---|---|---|---|---|
| | | | *train* | *valid* | *test* | |
| ev_lstm | - | - | 3.375 | 4.568 | 5.166 | 557176 |
| ev_lstm | - | 10 | 2.176 | 3.119 | 3.347 | 799784 |
| ev_lstm | - | 20,10 | 1.889 | 2.685 | 3.003 | 1457704 |
| evd_lstm | 3 | 10 | 2.308 | 2.299 | 2.823 | 614744 |
| evd_lstm | 5 | 10 | 2.251 | 2.244 | 2.689 | 655864 |
| evd_lstm | 7 | - | 2.750 | 2.761 | 3.730 | 546896 |
| evd_lstm | 7 | 10 | 2.281 | 2.267 | 2.690 | 696984 |
| evd_lstm | 7 | 20,10 | 2.184 | 2.183 | **2.520** | 1252104 |
| evd_mlp | 3 | 10 | 2.452 | 2.424 | 3.212 | 76843 |
| evd_mlp | 5 | 10 | 2.405 | 2.372 | 3.069 | 81983 |
| evd_mlp | 7 | - | 2.752 | 2.762 | 3.975 | 68362 |
| evd_mlp | 7 | 10 | 2.384 | 2.376 | 3.156 | 87123 |
| evd_mlp | 7 | 20,10 | 2.349 | 2.285 | 2.825 | 156513 |
| psd_lstm | - | - | 3.489 | 4.391 | 4.603 | 544840 |
| psd_lstm | - | 10 | 2.897 | 3.722 | 3.741 | 676424 |
| psd_lstm | - | 20,10 | 2.711 | 3.415 | 3.489 | 1210984 |

$$\text{error} = \frac{100}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \left| p_{\text{SPP,est,2}}(k,l) - p_{\text{SPP,opt}}(k,l) \right|$$

# Speech mask prediction error

| architecture | $n_\Delta$ | $n_h$ | prediction error [%] train | prediction error [%] valid | prediction error [%] test | # of weights |
|---|---|---|---|---|---|---|
| ev_lstm | - | - | 3.375 | 4.568 | 5.166 | 557176 |
| ev_lstm | - | 10 | 2.176 | 3.119 | 3.347 | 799784 |
| ev_lstm | - | 20,10 | 1.889 | 2.685 | 3.003 | 1457704 |
| evd_lstm | 3 | 10 | 2.308 | 2.299 | 2.823 | 614744 |
| evd_lstm | 5 | 10 | 2.251 | 2.244 | 2.689 | 655864 |
| evd_lstm | 7 | - | 2.750 | 2.761 | 3.730 | 546896 |
| evd_lstm | 7 | 10 | 2.281 | 2.267 | 2.690 | 696984 |
| evd_lstm | 7 | 20,10 | 2.184 | 2.183 | 2.520 | 1252104 |
| evd_mlp | 3 | 10 | 2.452 | 2.424 | **3.212** | 76843 |
| evd_mlp | 5 | 10 | 2.405 | 2.372 | 3.069 | 81983 |
| evd_mlp | 7 | - | 2.752 | 2.762 | 3.975 | 68362 |
| evd_mlp | 7 | 10 | 2.384 | 2.376 | 3.156 | 87123 |
| evd_mlp | 7 | 20,10 | 2.349 | 2.285 | 2.825 | 156513 |
| psd_lstm | - | - | 3.489 | 4.391 | 4.603 | 544840 |
| psd_lstm | - | 10 | 2.897 | 3.722 | 3.741 | 676424 |
| psd_lstm | - | 20,10 | 2.711 | 3.415 | 3.489 | 1210984 |

$$\text{error} = \frac{100}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \left| p_{\text{SPP,est,2}}(k,l) - p_{\text{SPP,opt}}(k,l) \right|$$

# PESQ and PEASS/OPS scores [Emiya et al., 2011]

| architecture | $n_\Delta$ | $n_h$ | PESQ [MOS] | | | OPS [%] | | |
|---|---|---|---|---|---|---|---|---|
| | | | train | valid | test | train | valid | test |
| ev_lstm, MVDR, 6ch | - | 20,10 | 2.204 | 1.850 | 1.788 | 62 | 46 | 39 |
| evd_lstm, MVDR, 6ch | 7 | 20,10 | 1.948 | 1.773 | 1.748 | 53 | 45 | 39 |
| evd_mlp, MVDR, 6ch | 3 | 10 | 1.866 | 1.713 | 1.630 | 50 | 45 | 40 |
| psd_lstm, MVDR, 6ch | - | 20,10 | 1.826 | 1.663 | 1.636 | 54 | 47 | 45 |
| ev_lstm, GSC, 6ch | - | 20,10 | 2.045 | 1.760 | 1.742 | 51 | 41 | 37 |
| evd_lstm, GSC, 6ch | 7 | 20,10 | 1.889 | 1.714 | 1.706 | 46 | 39 | 37 |
| evd_mlp, GSC, 6ch | 3 | 10 | 1.822 | 1.667 | 1.602 | 43 | 38 | 37 |
| psd_lstm, GSC, 6ch | - | 20,10 | 1.783 | 1.620 | 1.622 | 49 | 43 | 43 |
| ev_lstm, GEV, 6ch | - | 20,10 | 2.443 | 2.007 | 1.891 | 72 | 58 | 51 |
| evd_lstm, GEV, 6ch | 7 | 20,10 | 2.226 | 1.969 | **1.874** | 67 | 59 | **52** |
| evd_mlp, GEV, 6ch | 3 | 10 | 2.131 | 1.900 | **1.758** | 65 | 58 | **51** |
| *psd_lstm, GEV, 6ch | - | 20,10 | 1.977 | 1.758 | 1.724 | 63 | 54 | 48 |
| ev_lstm, GEV, 2ch | - | 10,5 | 1.965 | 1.706 | 1.725 | 51 | 44 | 45 |
| evd_mlp, GEV, 2ch | 3 | 5 | 1.980 | 1.778 | 1.774 | 44 | 40 | 40 |
| BeamformIt!, 5ch | - | - | 1.350 | 1.292 | 1.326 | 31 | 36 | 35 |
| **CGMM-EM, 6ch | - | - | 1.635 | 1.483 | 1.468 | 48 | 42 | 38 |

*similar to CHiME4-contributions: [Erdogan et al., 2016] and [Heymann et al., 2016]
**CHiME3 winner: [Higuchi et al., 2016]

# PESQ and PEASS/OPS scores [Emiya et al., 2011]

| *architecture* | $n_\Delta$ | $n_h$ | *PESQ [MOS]* | | | *OPS [%]* | | |
|---|---|---|---|---|---|---|---|---|
| | | | *train* | *valid* | *test* | *train* | *valid* | *test* |
| ev_lstm, MVDR, 6ch | - | 20,10 | 2.204 | 1.850 | 1.788 | 62 | 46 | 39 |
| evd_lstm, MVDR, 6ch | 7 | 20,10 | 1.948 | 1.773 | 1.748 | 53 | 45 | 39 |
| evd_mlp, MVDR, 6ch | 3 | 10 | 1.866 | 1.713 | 1.630 | 50 | 45 | 40 |
| psd_lstm, MVDR, 6ch | - | 20,10 | 1.826 | 1.663 | 1.636 | 54 | 47 | 45 |
| ev_lstm, GSC, 6ch | - | 20,10 | 2.045 | 1.760 | 1.742 | 51 | 41 | 37 |
| evd_lstm, GSC, 6ch | 7 | 20,10 | 1.889 | 1.714 | 1.706 | 46 | 39 | 37 |
| evd_mlp, GSC, 6ch | 3 | 10 | 1.822 | 1.667 | 1.602 | 43 | 38 | 37 |
| psd_lstm, GSC, 6ch | - | 20,10 | 1.783 | 1.620 | 1.622 | 49 | 43 | 43 |
| ev_lstm, GEV, 6ch | - | 20,10 | 2.443 | 2.007 | 1.891 | 72 | 58 | 51 |
| evd_lstm, GEV, 6ch | 7 | 20,10 | 2.226 | 1.969 | 1.874 | 67 | 59 | 52 |
| evd_mlp, GEV, 6ch | 3 | 10 | 2.131 | 1.900 | 1.758 | 65 | 58 | 51 |
| *psd_lstm, GEV, 6ch | - | 20,10 | 1.977 | 1.758 | **1.724** | 63 | 54 | **48** |
| ev_lstm, GEV, 2ch | - | 10,5 | 1.965 | 1.706 | 1.725 | 51 | 44 | 45 |
| evd_mlp, GEV, 2ch | 3 | 5 | 1.980 | 1.778 | 1.774 | 44 | 40 | 40 |
| BeamformIt!, 5ch | - | - | 1.350 | 1.292 | 1.326 | 31 | 36 | 35 |
| **CGMM-EM, 6ch | - | - | 1.635 | 1.483 | 1.468 | 48 | 42 | 38 |

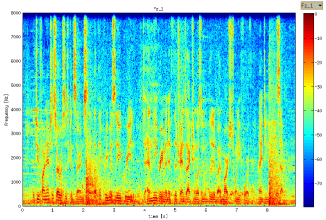
*similar to CHiME4-contributions: [Erdogan et al., 2016] and [Heymann et al., 2016]
**CHiME3 winner: [Higuchi et al., 2016]

# PESQ and PEASS/OPS scores [Emiya et al., 2011]

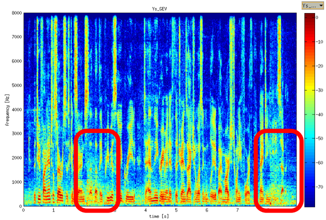
| architecture | $n_\triangle$ | $n_h$ | PESQ [MOS] | | | OPS [%] | | |
|---|---|---|---|---|---|---|---|---|
| | | | train | valid | test | train | valid | test |
| ev_lstm, MVDR, 6ch | - | 20,10 | 2.204 | 1.850 | 1.788 | 62 | 46 | 39 |
| evd_lstm, MVDR, 6ch | 7 | 20,10 | 1.948 | 1.773 | 1.748 | 53 | 45 | 39 |
| evd_mlp, MVDR, 6ch | 3 | 10 | 1.866 | 1.713 | 1.630 | 50 | 45 | 40 |
| psd_lstm, MVDR, 6ch | - | 20,10 | 1.826 | 1.663 | 1.636 | 54 | 47 | 45 |
| ev_lstm, GSC, 6ch | - | 20,10 | 2.045 | 1.760 | 1.742 | 51 | 41 | 37 |
| evd_lstm, GSC, 6ch | 7 | 20,10 | 1.889 | 1.714 | 1.706 | 46 | 39 | 37 |
| evd_mlp, GSC, 6ch | 3 | 10 | 1.822 | 1.667 | 1.602 | 43 | 38 | 37 |
| psd_lstm, GSC, 6ch | - | 20,10 | 1.783 | 1.620 | 1.622 | 49 | 43 | 43 |
| ev_lstm, GEV, 6ch | - | 20,10 | 2.443 | 2.007 | 1.891 | 72 | 58 | 51 |
| evd_lstm, GEV, 6ch | 7 | 20,10 | 2.226 | 1.969 | 1.874 | 67 | 59 | 52 |
| evd_mlp, GEV, 6ch | 3 | 10 | 2.131 | 1.900 | 1.758 | 65 | 58 | 51 |
| *psd_lstm, GEV, 6ch | - | 20,10 | 1.977 | 1.758 | 1.724 | 63 | 54 | 48 |
| ev_lstm, GEV, 2ch | - | 10,5 | 1.965 | 1.706 | 1.725 | 51 | 44 | 45 |
| evd_mlp, GEV, 2ch | 3 | 5 | 1.980 | 1.778 | 1.774 | 44 | 40 | 40 |
| BeamformIt!, 5ch | - | - | 1.350 | 1.292 | 1.326 | 31 | 36 | 35 |
| **CGMM-EM, 6ch | - | - | 1.635 | 1.483 | **1.468** | 48 | 42 | **38** |

*similar to CHiME4-contributions: [Erdogan et al., 2016] and [Heymann et al., 2016]
**CHiME3 winner: [Higuchi et al., 2016]

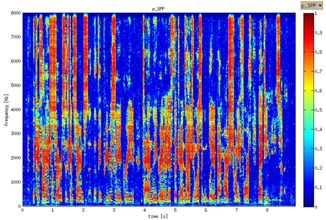
# Example 1: M04_422C0205_CAF



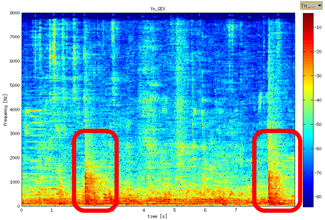

$1^{\text{st}}$ microphone: $Z_1(k,l)$




BF output: $Y_{\text{s,GEV}}(k,l)$

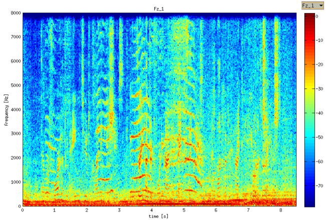



est. speech mask: $p_{\text{SPP,est,2}}(k,l)$

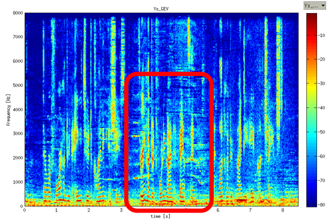



complementary output: $Y_{\text{n,GEV}}(k,l)$

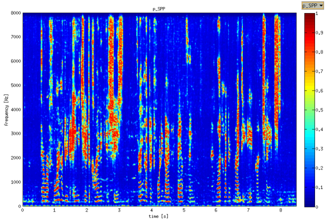
# Example 2: F01_22HC010W_BUS



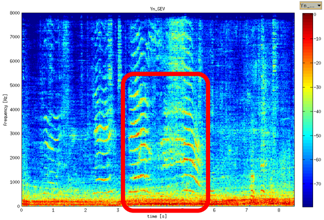

1st microphone: $Z_1(k, l)$




BF output: $Y_{\text{s,GEV}}(k, l)$




est. speech mask: $p_{\text{matlab,est,2}}(k, l)$




complementary output: $Y_{\text{n,GEV}}(k, l)$

# Conclusion

- Take-home message:
  - the MVDR, GSC and GEV Beamformers and the Postfilter solely depend on the speech mask
  - speech mask estimate can be learned from eigenvector features
- Future work:
  - reduce NN size and complexity
  - multiple speakers

# Conclusion

- Take-home message:
  - the MVDR, GSC and GEV Beamformers and the Postfilter solely depend on the speech mask
  - speech mask estimate can be learned from eigenvector features
- Future work:
  - reduce NN size and complexity
  - multiple speakers

**Thank you for your attention!**

# References

[Barker et al., 2015] Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2015).
The third 'chime' speech separation and recognition challenge: Dataset, task and baselines.
In *IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*.

[Emiya et al., 2011] Emiya, V., Vincent, E., Harlander, N., and Hohmann, V. (2011).
Subjective and objective quality assessment of audio source separation.
*IEEE Transactions on Audio, Speech and Language Processing*, 19(7).

[Erdogan et al., 2016] Erdogan, H., Hershey, J., Watanabe, S., Mandel, M., and Roux, J. L. (2016).
Improved mvdr beamforming using single-channel mask prediction networks.
In *Interspeech*.

[Heymann et al., 2016] Heymann, J., Drude, L., and Haeb-Umbach, R. (2016).
Neural network based spectral mask estimation for acoustic beamforming.
In *2016 IEEE ICASSP*.

[Higuchi et al., 2016] Higuchi, T., Ito, N., Yoshioka, T., and Nakatani, T. (2016).
Robust mvdr beamforming using time-frequency masks for online/offline asr in noise.
*IEEE International Conference on Acoustics, Speech, and Signal Processing*, 4:5210–5214.

[Hoshuyama et al., 1999] Hoshuyama, O., Sugiyama, A., and Hirano, A. (1999).
A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters.
*IEEE Transactions on Signal Processing*, 47(10).

[Pfeifenberger et al., 2016] Pfeifenberger, L., Zöhrer, M., and Pernkopf, F. (2016).
Eigenvector-based speech mask estimation for multi-channel speech enhancement.
*IEEE Transactions on Speech and Audio Processing, submitted*.

[Shmulik et al., 2012] Shmulik, M. G., Gannot, S., and Cohen, I. (2012).
A sparse blocking matrix for multiple constraints GSC beamformer.
*IEEE International Conference on Acoustics, Speech and Signal Processing*.

[Vary and Martin, 2006] Vary, P. and Martin, R. (2006).
*Digital Speech Transmission*.
Wiley, West Sussex.

[Warsitz and Haeb-Umbach, 2007] Warsitz, E. and Haeb-Umbach, R. (2007).
Blind acoustic beamforming based on generalized eigenvalue decomposition.
*IEEE Transactions on audio, speech, and language processing*, 15(5).