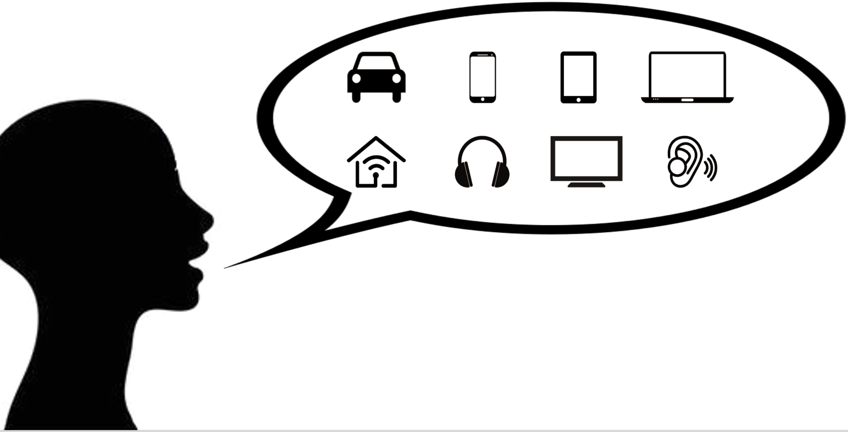
Complex Neural Beamforming

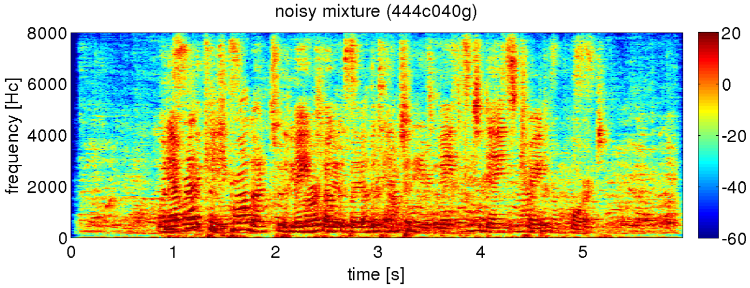
Lukas Pfeifenberger, Matthias Zöhrer, and Franz Pernkopf
Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria

Speech recognition



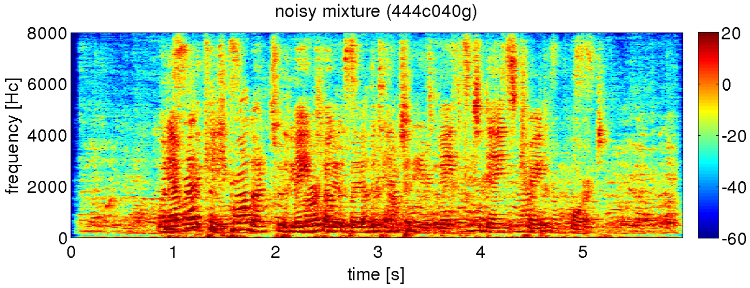
Speech recognition

...is still a challenging task in adverse environments



Speech recognition

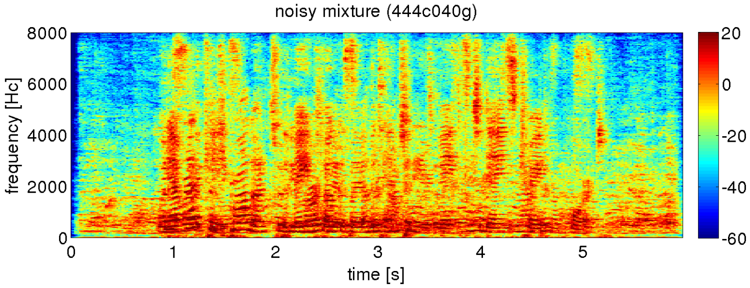
...is still a challenging task in adverse environments



TRANSCRIPTION:
"Whatever the case the main focus of attention remains today's trade report."
"He said such products would be marketed by other companies with experience in that business."

Speech recognition

...is still a challenging task in adverse environments

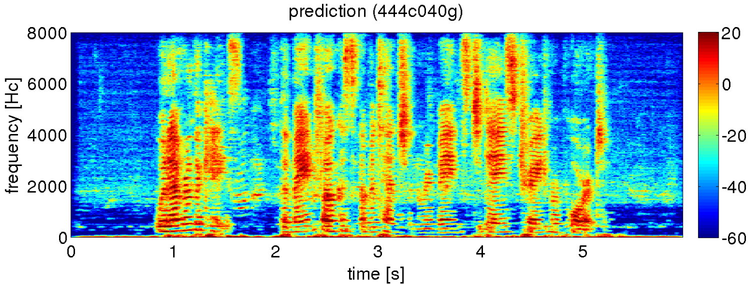


TRANSCRIPTION:
"Whatever the case the main focus of attention remains today's trade report."
"He said such products would be marketed by other companies with experience in that business."

CHiME5: Kaldi (optimized AM/LM): 46.6% WER [Du et al., 2018]

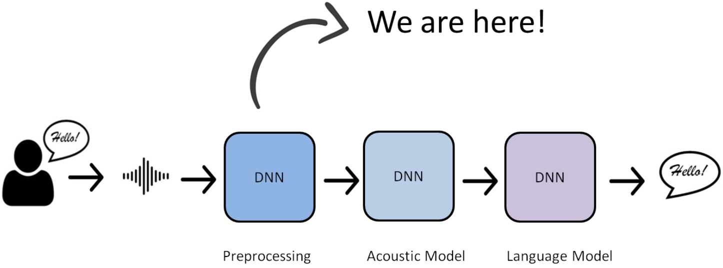
Our Contribution:

Complex Neural Beamforming



Main idea:
Spatially select sources using complex neural networks

ASR pipeline



- End-to-End training
- Acoustic front-end

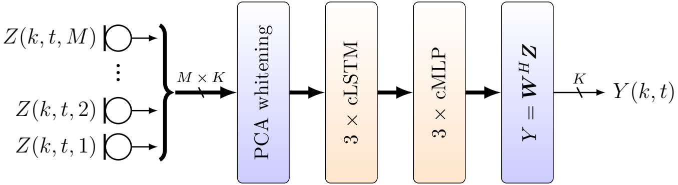
Source Separation

- Single-channel
  - Deep Clustering [Hershey et al., 2016]
  - Attractor Networks [Chen et al., 2016]
  - Attention Models [Kinoshita et al., 2018]
- Multi-channel
  - Statistical models (CGMM-EM) [Higuchi et al., 2016]
  - Mask-based beamforming [Erdogan et al., 2016]
  - Eigenvector beamforming [Pfeifenberger et al., 2017]

Limitations

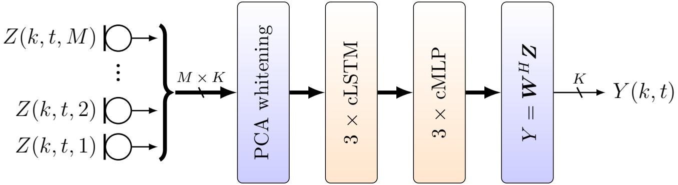
- Mask-based beamforming
  - Cannot separate multiple speakers (exception: Eigenvector features [Pfeifenberger et al., 2017])
  - Performance drops if speaker is moving
  - Limited to block processing
- Attractor Networks / Attention Models
  - Additional clustering step required (block processing)
  - Speaker re-identification/tracking only partially solved
  - No spatial exclusion (background noise)
  - Block permutation problem (PIT)

Complex Neural Beamforming



- Input signal: $\boldsymbol{Z}(k,t) = \sum_{c=1}^{C} \boldsymbol{S}_c(k,t)$
- PCA whitening: $\bar{\boldsymbol{Z}} = \boldsymbol{U}_{PCA}\boldsymbol{Z} \in \mathbb{C}^{K \times T \times M}$ [Kuttruff, 2009]
- Weight estimation: $\boldsymbol{W} = f_{\Theta}(\bar{\boldsymbol{Z}}) \in \mathbb{C}^{K \times T \times M}$

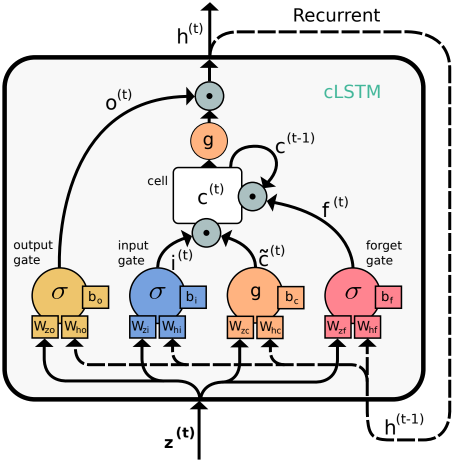
Complex Neural Beamforming



Layer #	type	activation	shape	# of parameters
1*	cMLP	cTanh	$K(M \times M)$	18,468
2	cLSTM	cTanh	$K(M \times M)$	147,744
3	cMLP	cTanh	$M(K \times K)$	1,579,014
4*	cLSTM	cTanh	$K(2M \times 2M)$	590,976
5	cLSTM	cTanh	$K(2M \times M)$	295,488
6	cMLP	cNorm	$K(M \times M)$	18,468

*Reduction to 4 layers is possible

Complex LSTM cell



$$\mathbf{i}^{(t)} = \sigma\left(\text{Re}\left\{\mathbf{W}_{zi}\mathbf{z}^{(t)} + \mathbf{W}_{hi}\mathbf{h}^{(t-1)} + \mathbf{b}_i\right\}\right)$$

$$\mathbf{f}^{(t)} = \sigma\left(\text{Re}\left\{\mathbf{W}_{zf}\mathbf{z}^{(t)} + \mathbf{W}_{hf}\mathbf{h}^{(t-1)} + \mathbf{b}_f\right\}\right)$$

$$\mathbf{o}^{(t)} = \sigma\left(\text{Re}\left\{\mathbf{W}_{zo}\mathbf{z}^{(t)} + \mathbf{W}_{ho}\mathbf{h}^{(t-1)} + \mathbf{b}_o\right\}\right)$$

$$\tilde{\mathbf{c}}^{(t)} = g(\mathbf{W}_{zc}\mathbf{z}^{(t)} + \mathbf{W}_{hc}\mathbf{h}^{(t-1)} + \mathbf{b}_c)$$

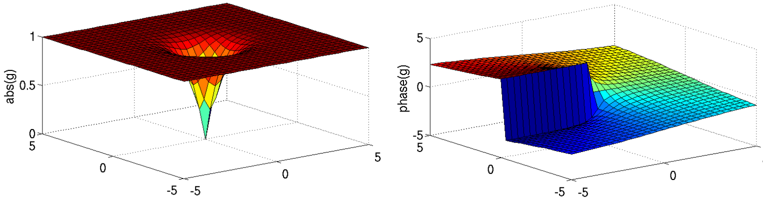
$$\mathbf{c}^{(t)} = \mathbf{f}^{(t)} \odot \mathbf{c}^{(t-1)} + \mathbf{i}^{(t)} \odot \tilde{\mathbf{c}}^{(t)}$$

$$\mathbf{h}^{(t)} = \mathbf{o}^{(t)} \odot g(\mathbf{c}^{(t)})$$

Complex activations

Non-holomorphic functions required for neural beamforming:

- Applying BF weights: $\boldsymbol{W}^H\vec{\mathbf{z}}$
- Magnitude normalization: $\frac{\vec{\mathbf{z}}}{|\vec{\mathbf{z}}|_2}$
- Phase normalization: $\mathbf{z} \odot e^{-j\varphi_{\mathbf{z}}}$
- Sigmoid activation function: $\sigma(\mathrm{Re}\{\mathbf{z}\})$
- tanh activation function: $\tanh(|\mathbf{z}|) \odot \frac{\mathbf{z}}{|\mathbf{z}|}$



Complex Gradients

- Many non-holomorphic functions are partially differentiable in their real and imaginary parts:
- Separate $\mathbf{z} \in \mathbb{C}$ into $\mathbf{z} = \mathbf{x} + j\mathbf{y}$
- Redefine $g(\mathbf{z})$ to $g(\mathbf{z}, \mathbf{z}^*)$
- Basis for partial derivatives:
  [Wirtinger, 1927, Bouboulis and Theodoridis, 2011]

$$\frac{\partial g}{\partial \mathbf{z}} = \frac{1}{2}\left(\frac{\partial g}{\partial \mathbf{x}} - j\frac{\partial g}{\partial \mathbf{y}}\right)$$

$$\frac{\partial g}{\partial \mathbf{z}^*} = \frac{1}{2}\left(\frac{\partial g}{\partial \mathbf{x}} + j\frac{\partial g}{\partial \mathbf{y}}\right)$$

- Chain rule: $\nabla_{\mathbf{z}^*} = \left(\nabla_{g^*}\right)^* \frac{\partial g}{\partial \mathbf{z}^*} + \nabla_{g^*}\left(\frac{\partial g}{\partial \mathbf{z}}\right)^*$
- For a real-valued cost function: $\nabla_{\mathbf{z}} = \left(\nabla_{\mathbf{z}^*}\right)^*$

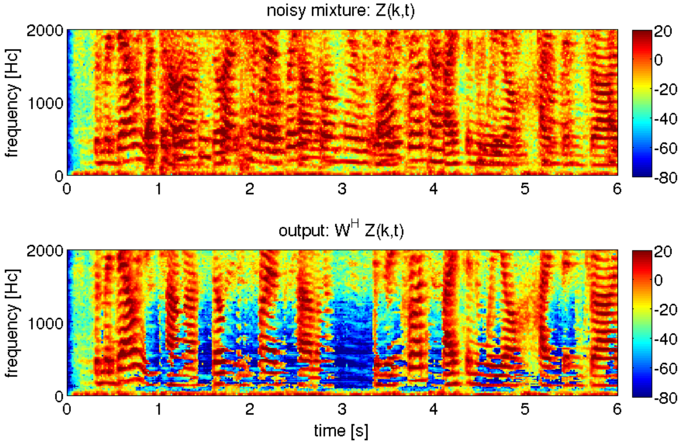
Cost function

Maximize the ΔSNR: $$10log_{10}\frac{|\boldsymbol{W}^H\boldsymbol{S}_1|^2}{|\boldsymbol{W}^H\boldsymbol{S}_{2\ldots N}|^2} - 10log_{10}\frac{||\boldsymbol{S}_1||_2^2}{||\boldsymbol{S}_{2\ldots N}||_2^2}$$

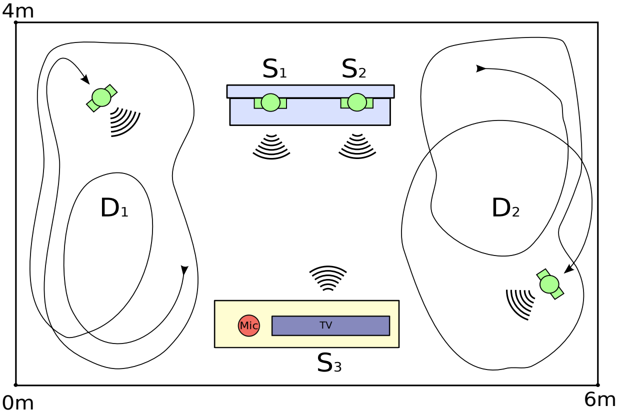
- complex neural beamformer $\boldsymbol{W} = f_\Theta(\bar{\boldsymbol{Z}})$
  - estimates a new set of BF weights for each time-freuency bin
  - instantaneous adaption to isotropic noise or moving speakers
- statistical beamformer (i.e. MVDR)
  - requires a block T of data to estimate BF weights
  - spatial characteristics must not change during T

Cost function

Maximize the ΔSNR: $$10log_{10}\frac{|\boldsymbol{W}^H\boldsymbol{S}_1|^2}{|\boldsymbol{W}^H\boldsymbol{S}_{2\ldots N}|^2} - 10log_{10}\frac{||\boldsymbol{S}_1||_2^2}{||\boldsymbol{S}_{2\ldots N}||_2^2}$$



Experiment 1: Simulated RIRs



Simulated living room scenario with multiple moving speakers from WSJ0, and a 6-channel microphone array.

Experiment 2: Real RIRs

Recording setup for 1792 real 6-channel RIRs.

Results

WER* for the WSJ0 si_et_05 set + simulated RIRs:

Scenario	BeamformIt	MBF**	CN-BF
dynamic1 vs. dynamic2	76.7%	46.1%	21.1%
dynamic1 vs. isotropic	17.7%	32.8%	9.0%
static1 vs. isotropic	17.9%	18.5%	6.1%
static1 vs. static3	43.2%	45.6%	13.4%
static2 vs. dynamic1, static3	88.3%	58.3%	33.7%

WER* for the WSJ0 si_et_05 set + real RIRs:

Scenario	BeamformIt	MBF**	CN-BF
static1 vs. isotropic	22.8%	21.8%	7.9%
static1 vs. static3	84.7%	73.1%	14.5%

*Google Speech-to-Text API: https://pypi.org/project/SpeechRecognition/
**Mask-based beamforming with block-online processing [Böddeker et al., 2018]

Conclusion

- CN-BF optimizes BF weights for each T-F bin
- Outperforms statistical beamformers
- Real-time capability down to 1 frame delay
- Further research:
  - Overlapping speaker paths
  - Speaker (re-)identification
  - Dependency on trained room acoustics

Conclusion

- CN-BF optimizes BF weights for each T-F bin
- Outperforms statistical beamformers
- Real-time capability down to 1 frame delay
- Further research:
  - Overlapping speaker paths
  - Speaker (re-)identification
  - Dependency on trained room acoustics

Thank you for your attention!

References

[Böddeker et al., 2018] Böddeker, C., Erdogan, H., Yoshioka, T., and Haeb-Umbach, R. (2018).
Exploring practical aspects of neural mask-based beamforming for far-field speech recognition.
In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6697–6701.

[Bouboulis and Theodoridis, 2011] Bouboulis, P. and Theodoridis, S. (2011).
Extension of wirtinger's calculus to reproducing kernel hilbert spaces and the complex kernel lms.
Trans. Sig. Proc., 59(3):964–978.

[Chen et al., 2016] Chen, Z., Luo, Y., and Mesgarani, N. (2016).
Deep attractor network for single-microphone speaker separation.
CoRR, abs/1611.08930.

[Du et al., 2018] Du, J., Gao, T., Sun, L., Ma, F., Fang, Y., Liu, D.-Y., Zhang, Q., Zhang, X., Wang, H.-K., Pan, J., Gao, J.-Q., Lee, C.-H., and Chen, J.-D. (2018).
The ustc-iflytek systems for chime-5 challenge.
pages 11–15.

[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*.

[Hershey et al., 2016] Hershey, J. R., Chen, Z., Roux, J. L., and Watanabe, S. (2016).
Deep clustering: Discriminative embeddings for segmentation and separation.
IEEE International Conference on Acoustics, Speech, and Signal Processing (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.

[Kinoshita et al., 2018] Kinoshita, K., Drude, L., Delcroix, M., and Nakatani, T. (2018).
Listening to each speaker one by one with recurrent selective hearing networks.
pages 5064–5068.

[Kuttruff, 2009] Kuttruff, H. (2009).
Room Acoustics.
Spoon Press, London–New York, 5th edition.

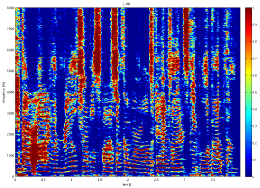
[Pfeifenberger et al., 2017] Pfeifenberger, L., Zöhrer, M., and Pernkopf, F. (2017).
Dnn-based speech mask estimation for eigenvector beamforming.
In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 66–70.

[Scheibler et al., 2017] Scheibler, R., Bezzam, E., and Dokmanic, I. (2017).
Pyroomacoustics: A python package for audio room simulations and array processing algorithms.
CoRR, abs/1710.04196.

[Wirtinger, 1927] Wirtinger, W. (1927).
Zur formalen theorie der funktionen von mehr komplexen veränderlichen.
Math. Ann., 97:357–375.

Mask-based-BF vs. CN-BF

- Mask-based-BF
  - $p(k,t) = f_{\Theta}(|Z(k,t,m)|)$
  - $\hat{\boldsymbol{\Phi}}_{SS}(k) = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Z}(k,t)\boldsymbol{Z}^H(k,t)p(k,t)$
  - $\boldsymbol{W}_{MVDR}(k) = \frac{\hat{\boldsymbol{\Phi}}_{NN}^{-1}(k)\boldsymbol{v}_S(k)}{\boldsymbol{v}_S^H(k)\hat{\boldsymbol{\Phi}}_{NN}^{-1}(k)\boldsymbol{v}_S(k)}$

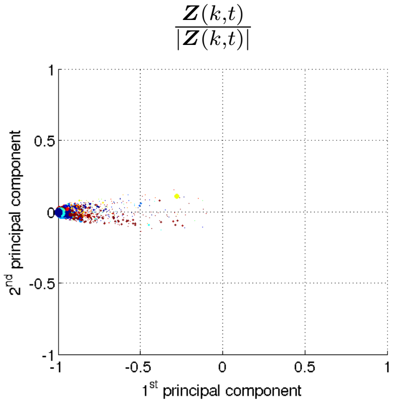


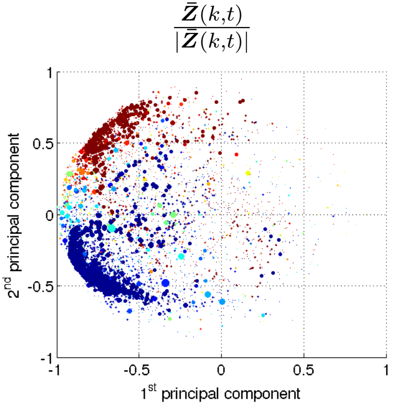
- CN-BF
  - $\boldsymbol{W}(k,t) = f_{\Theta}(\bar{\boldsymbol{Z}}(k,t))$

PCA whitening

additive mixture: $\boldsymbol{Z}(k,t) = \boldsymbol{S_1}(k,t) + \boldsymbol{S_2}(k,t)$

whitening: $\bar{\boldsymbol{Z}}(k,t) = \boldsymbol{U}_{PCA}(k,t)\boldsymbol{Z}(k,t)$





Alternatives to CN-BF

- Stacking: $g(\mathbf{z}) = \tanh\left(\begin{bmatrix}\mathrm{Re}\{\mathbf{z}\} \\ \mathrm{Im}\{\mathbf{z}\}\end{bmatrix}\right)$
  - complex properties are lost (i.e. rotation)
- Individual gradients: $g(\mathbf{z}) = \tanh(\mathrm{Re}\{\mathbf{z}\}) + i\tanh(\mathrm{Im}\{\mathbf{z}\})$
  - complex phase gets distorted
  - recurrent structures become unstable

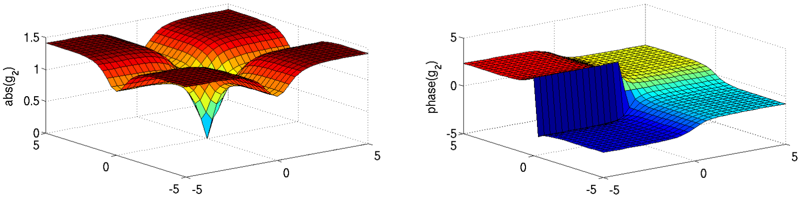


Image Source Method (ISM)

$$h_{\boldsymbol{m},\boldsymbol{s}}(n) = \sum_{\boldsymbol{x} \in \nu_m(\boldsymbol{s})} \frac{(1-\beta)^{\mathrm{order}(\boldsymbol{x})}}{4\pi||\boldsymbol{m}-\boldsymbol{x}||} \mathrm{sinc}\left(n - f_s \frac{||\boldsymbol{m}-\boldsymbol{x}||}{c}\right)$$

[Scheibler et al., 2017]