Dictionary-Based Fusion of Contact and Acoustic Microphones for Wind Noise Reduction

Marvin Tammen[1,2*], Xilin Li[1], Simon Doclo[2], Lalin Theverapperuma[1]
[1]Meta Reality Labs, USA; [2]Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany

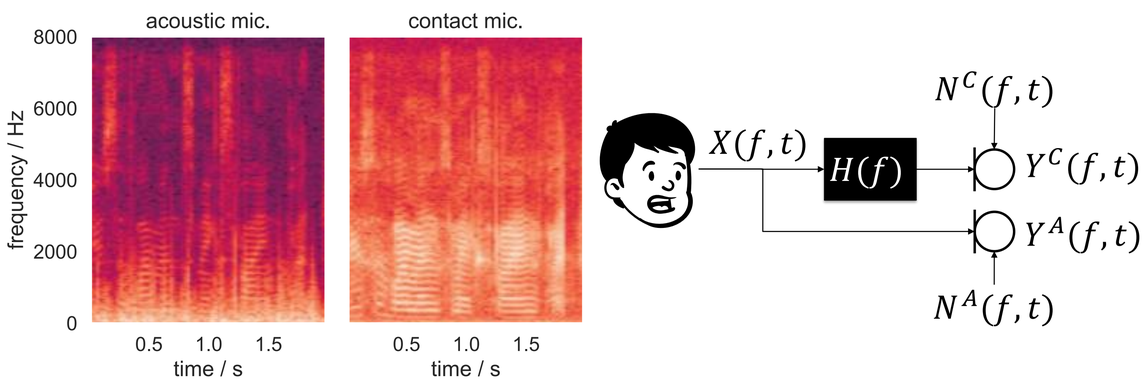
Problem Statement

- wind noise may severely reduce speech intelligibility in mobile communication scenarios
- performance of speech enhancement algorithms tends to degrade in extremely challenging conditions
- **auxiliary information** from sensors such as **contact microphones** can help improve performance, but labeled data is scarce

this poster: **fuse** advantages of **acoustic and contact microphones** for wind noise reduction

Signal Model

- two microphones capturing speech in wind noise
  - acoustic mic. (Y^A): high wind noise level, undistorted speech
  - contact mic. (Y^C): low wind noise level, distorted speech, additional noise



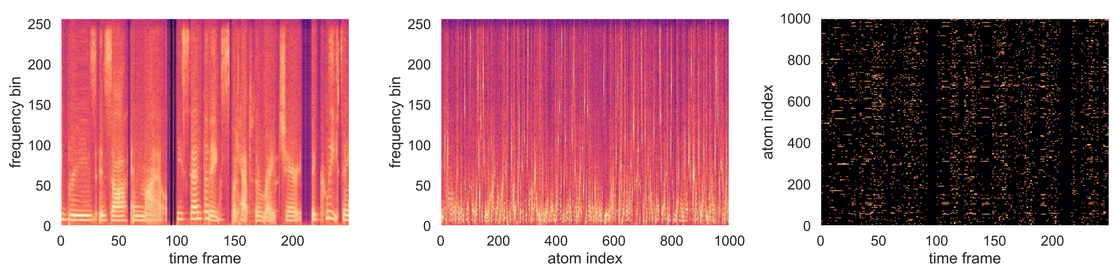
- signal model assumptions:
  - speech components related via **(linear) transfer function** $H(f)$
  - **separate additive noise** components

$$\begin{bmatrix} \boldsymbol{Y}^C \\ \boldsymbol{Y}^A \end{bmatrix} = \begin{bmatrix} \boldsymbol{HX} + \boldsymbol{N}^C \\ \boldsymbol{X} + \boldsymbol{N}^A \end{bmatrix} \in \mathbb{C}^{2F\,x\,T}$$

Dictionary-Based Signal Representation

- **speech contains spectro-temporal structures** (e.g., harmonics)
- these structures can be learned from data ("**dictionary atoms**") [1]
  → represent speech signal $\boldsymbol{X}$ using (complex-valued) learned dictionary $\boldsymbol{D}_X$ and sparse code $\boldsymbol{C}_X$:

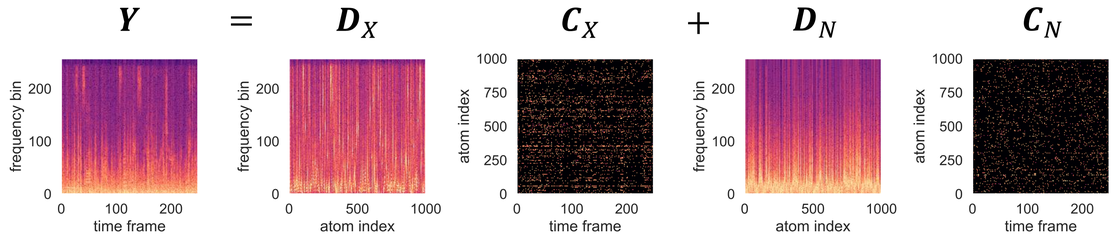
$$\boldsymbol{X} = \boldsymbol{D}_X \quad \boldsymbol{C}_X$$



keeping number of activated atoms low yields noise reduction

- **extension towards noisy signals**:
  - learn both speech and noise dictionaries $\boldsymbol{D}_X$ and $\boldsymbol{D}_N$
  - speech can be estimated using speech dictionary $\boldsymbol{D}_X$ and sparse code $\boldsymbol{C}_X$
  - signal components will be captured according to coherence with dictionaries

$$\boldsymbol{Y} = \boldsymbol{D}_X \quad \boldsymbol{C}_X + \boldsymbol{D}_N \quad \boldsymbol{C}_N$$



- **dictionaries** can be **trained by solving** non-convex optimization problem:

$$\boldsymbol{D}_X, \boldsymbol{C}_X = \underset{\boldsymbol{D},\boldsymbol{C}}{\operatorname{argmin}} \|\boldsymbol{X} - \boldsymbol{DC}\|_2^2 + \lambda \|\boldsymbol{C}\|$$

  - λ trades off more accurate vs. sparser representation
  - optimization alternates between dictionary update and sparse coding step

Dictionary-Based Microphone Fusion

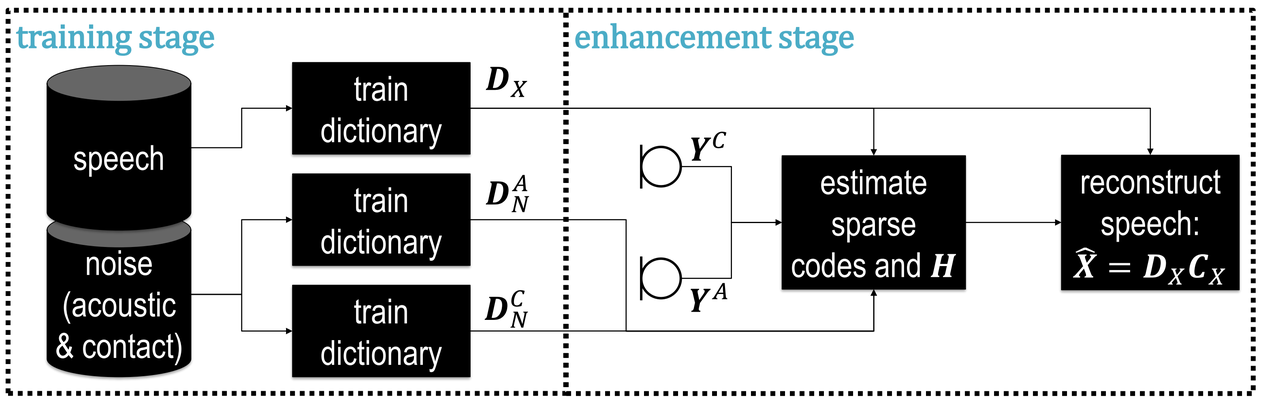
- **integrate signal model for acoustic and contact microphone signals into dictionary training and sparse coding:**
  - model noise components using separate dictionaries $\boldsymbol{D}_N^C$, $\boldsymbol{D}_N^A$
  - model speech components using single dictionary $\boldsymbol{D}_X$ and (unknown) transfer function $\boldsymbol{H}$

$$\begin{bmatrix} \boldsymbol{Y}^C \\ \boldsymbol{Y}^A \end{bmatrix} = \begin{bmatrix} \boldsymbol{HD}_X\boldsymbol{C}_X + \boldsymbol{D}_N^C\boldsymbol{C}_N^C \\ \boldsymbol{D}_X\boldsymbol{C}_X + \boldsymbol{D}_N^A\boldsymbol{C}_N^A \end{bmatrix}$$

- **cost function**:

$$\eta\left\|\boldsymbol{Y}^C - (\boldsymbol{HD}_X\boldsymbol{C}_X + \boldsymbol{D}_N^C\boldsymbol{C}_N^C)\right\|_2^2 + (1-\eta)\left\|\boldsymbol{Y}^A - (\boldsymbol{D}_X\boldsymbol{C}_X + \boldsymbol{D}_N^A\boldsymbol{C}_N^A)\right\|_2^2 + \lambda\|\widetilde{\boldsymbol{C}}\|$$

  - η controls influence of each microphone
  - **alternating optimization**:
    1. (iterative) sparse coding algorithm to estimate $\widetilde{\boldsymbol{C}} = \left[\boldsymbol{C}_X^T \quad \boldsymbol{C}_N^{C,T} \quad \boldsymbol{C}_N^{A,T}\right]^T$
    2. closed-form solution to estimate $\boldsymbol{H}$



Simulations

Dataset

	training	evaluation
clean speech	**LibriTTS** (dev-clean subset) [2]	
noise	**airflow-induced noise**: 3 m/s and 5 m/s (2 min each) **recorded by acoustic and contact microphones**	
SNR	no mixed signals required	{-10, -5, 0, 5} dB

Compared Algorithms

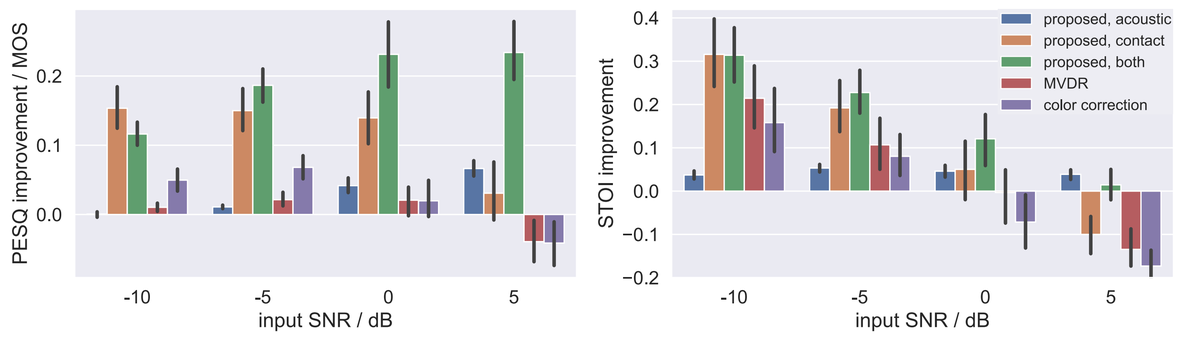
1. **proposed:** only acoustic microphone, only contact microphone, both microphones (fusion)
2. **baselines**: MVDR beamformer, color correction (inversion of $\boldsymbol{H}$)

Settings

- sampling frequency 16 kHz, frame length 32 ms, 50% overlap, $\sqrt{\text{Hann}}$ window
- proposed dictionary-based algorithms:
  - **sparse coding**: accelerated proximal gradient method [3] (1000 iterations)
  - **dictionary update**: method of optimal directions [4] (25 iterations)
  - dictionary size: 1000 atoms each (3000 atoms in total)
  - $\eta = 0.6$, placing **more weight on contact microphone**
  - 5 iterations alternating between sparse coding and estimating transfer function $\boldsymbol{H}$
- (initial) transfer function estimation: **covariance whitening method**

Results

- **performance measures**:
  - perceptual evaluation of speech quality (**PESQ**)
  - short-time objective intelligibility (**STOI**)



fusing both microphones yields **highest improvements** in most conditions

- low SNR: contact mic. > acoustic mic., high SNR: contact mic. < acoustic mic.
- benefit of jointly estimating speech and noise components as well as transfer function vs. only estimating transfer function

Outlook

- **comparison with deep neural network-based approaches** designed for scarce labeled data (e.g., self-supervised pre-training)

References
[1] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech Enhancement Using Generative Dictionary Learning," IEEE Trans. Audio Speech Language Processing, vol. 20, no. 6, pp. 1698–1712, Aug. 2012.
[2] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in Proc. Interspeech, Graz, Austria, Sep. 2019, pp. 1526–1530.
[3] N. Parikh, S. Boyd, "Proximal algorithms," Foundations and trends in Optimization, vol. 1, no. 3, pp. 127–239, 2014.
[4] K. Engan, S. O. Aase, and J. H. Husøy, "Multi-frame compression: theory and design," Signal Processing, vol. 80, no. 10, pp. 2121–2140, Oct. 2000.

*This work was performed while Marvin Tammen was an intern with Meta Reality Labs Research.