Multiple-input neural network-based residual echo suppression

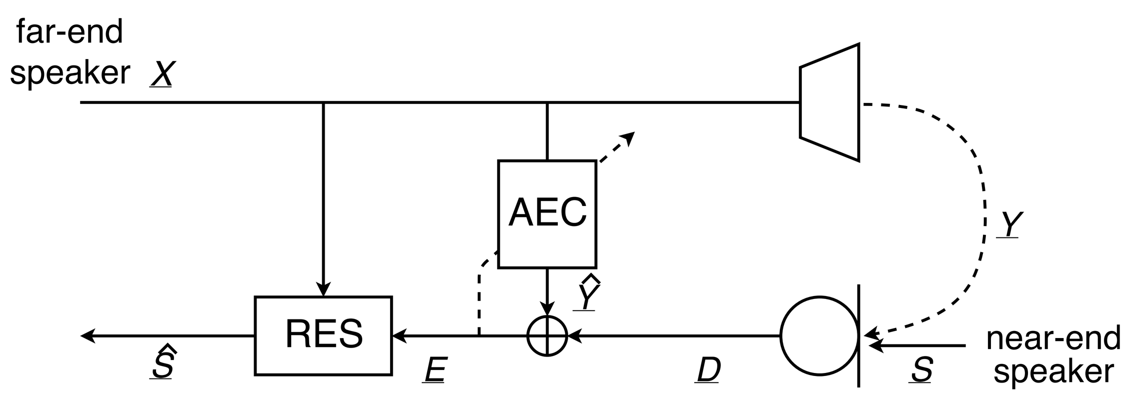
Guillaume Carbajal[†⋆] **Romain Serizel**[⋆] **Emmanuel Vincent** [⋆] **Éric Humbert** [†]

[†]Invoxia SAS, 2 Rue Maurice Hartmann, 92130 Issy-les-Moulineaux, France
[⋆] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

Acoustic echo cancellation

Goal: eliminate echo $\underline{Y}$, do not distort near-end $\underline{S}$.

Figure 1: *General setting for an acoustic echo canceller (AEC) and a residual echo suppressor (RES).*



$$\underline{E} = \underline{S} + \underbrace{\underline{Y} - \widehat{\underline{Y}}}_{\underline{Z}}$$

Residual echo suppression

Principle: suppress residual echo $\underline{Z}$

1. Estimate $\widehat{M}$ from a target mask M
2. $\widehat{\underline{S}} = \widehat{M}\underline{E}$

$$= \underbrace{\underline{S}_{\text{RES}}}_{\text{distorted near-end}} + \underbrace{\underline{Z}_{\text{RES}}}_{\text{post-residual echo}}$$

Single-input vs. multiple-input methods

- **Single-input methods**
  - ▷ $\widehat{M}$ using single signal X [1] or $\widehat{Y}$ [2]
- **Multiple-input methods**
  - ▷ $\widehat{M}$ using multiple signals together (e.g. D & X) [3]

Spectral-based vs. mask-based methods

- **Spectral-based methods: 2 steps**
  1. compute $\widehat{Z}$
     - ▷ *linear models*: $\widehat{Z} = \lambda X$ [4] or $\widehat{Z} = \lambda\widehat{Y}$ [2]
     - ▷ *nonlinear models*: $\widehat{Z}$ using a neural network [1]
  2. derive $\widehat{M}$ from $\widehat{Z}$

$$\widehat{M} = \max\left(M_{\text{min}}, 1 - \mu\frac{\widehat{Z}^2}{E^2}\right)$$

- **Mask-based methods: 1 step**
  - ▷ Derive $\widehat{M}$ using a neural network [3]

Table 1: *Example target masks.*

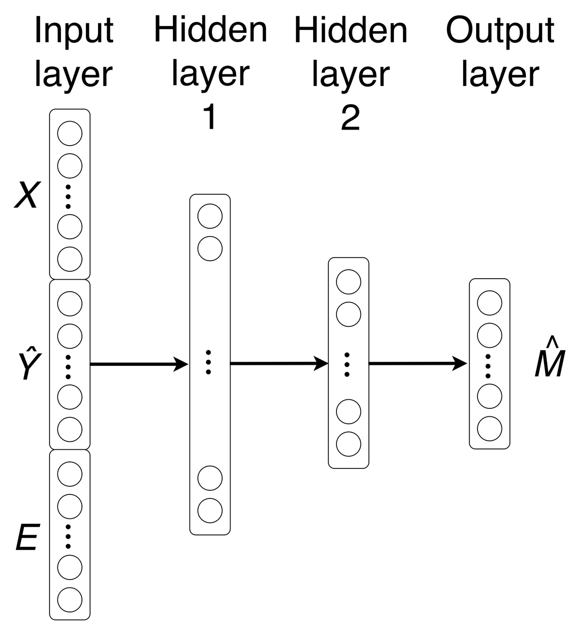
Ideal ratio mask (IRM)	$M = \frac{S}{\sqrt{S^2+Z^2}}$
Ideal amplitude mask (IAM)	$M = \frac{S}{E}$
Phase-sensitive filter (PSF)	$M = \frac{S}{E}\cos(\theta_S - \theta_E)$

Multiple-input NN-based RES

Proposed RES

- **Multiple inputs** E, X, and/or $\widehat{Y}$
- **Target mask** M PSF
- **NN structure** MLP with 2 hidden layers

Figure 2: *Example multiple-input NN-based RES.*



Experiments

- Proposed vs single-input spectral-based RES [1, 2]

	Train	Test
Data	real $\underline{Y}$ simulated $\underline{S}$	real $\underline{Y}$ real $\underline{S}$
Room size	$3 \times 3 \times 3$ m	$7 \times 7 \times 3$ m
Reverberation time	0.2 s	0.5 s

Table 3: *Metrics.*

Echo Return Loss Enhancement (ERLE)	echo reduction
Signal-to-Distortion Ratio (SDR)	distortion of $\widehat{\underline{S}}$
Signal-to-Artifacts Ratio (SAR)	distortion of $\widehat{\underline{S}}_{\text{RES}}$

Figure 3: *Experimental settings.*

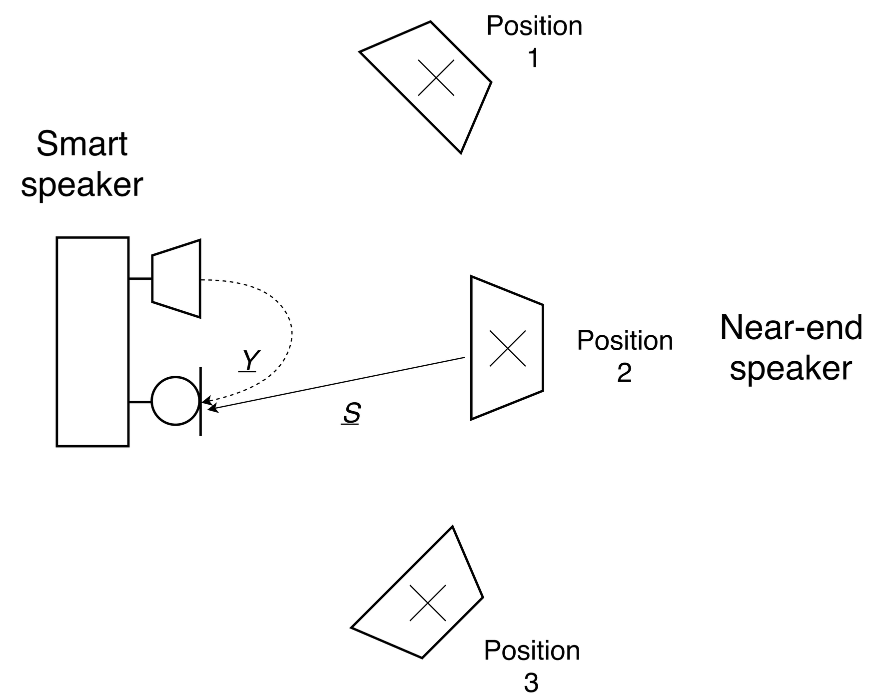


Table 4: *Average ERLE (dB) and SDR (dB) of the proposed RES.*

	Double-talk	NN inputs: E	E, X	$E, \widehat{Y}$	$E, X, \widehat{Y}$
ERLE	Yes	10.8	19.3	16.5	$\mathbf{20.3}$
	No	12.3	22.6	18.5	$\mathbf{23.5}$
SDR	Yes	-2.7	3.6	1.0	$\mathbf{4.1}$

(a) *With various NN inputs.*

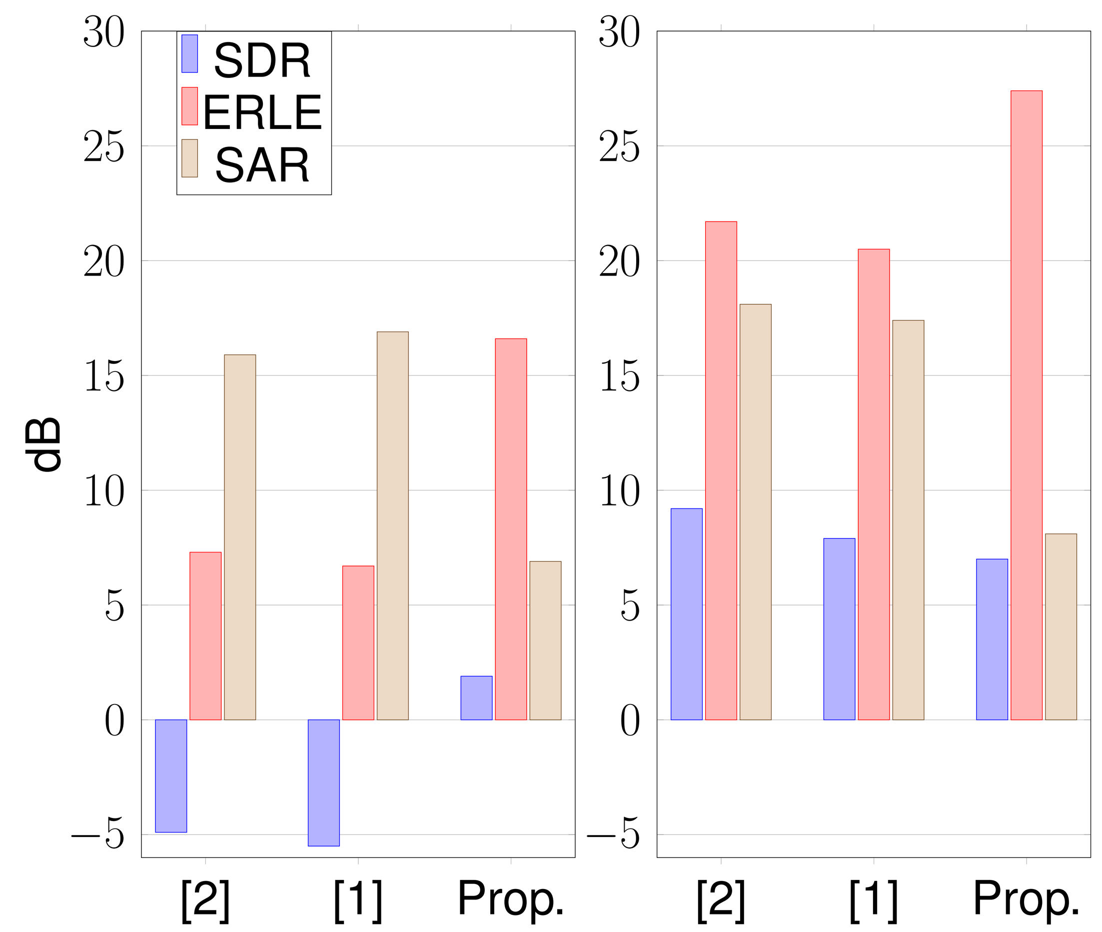
	Double-talk	Target mask: IRM	IAM	PSF
ERLE	Yes	14.8	16.7	$\mathbf{17.8}$
	No	16.1	18.7	$\mathbf{20.2}$
SDR	Yes	0.2	1.7	$\mathbf{2.5}$

(b) *With various target masks.*

	Double-talk	AEC	AEC+RES: Valin [2]	Schwarz [1]	Prop. RES
ERLE	Yes	10.6	12.5	11.8	$\mathbf{21.2}$
	No	12.2	13.8	13.3	$\mathbf{24.4}$
SDR	Yes	-1.1	0.4	-0.2	$\mathbf{4.9}$

(c) *Compared to other RES and to AEC only.*

Figure 4: *Detailed analysis during double-talk.*



(a) Before AEC convergence (b) After AEC convergence

Conclusion

- **Multiple inputs**
  - ▷ greater residual echo reduction than single-input
- **Target mask** M
  - ▷ best performance with PSF
- **Prop. RES vs single-input spectral-based RES**
  - ▷ robust to train/test room mismatch
  - ▷ robust to different scenarios

References

[1] A. Schwarz, C. Hofmann and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression," in *Proc. WASPAA*, 2013.

[2] J. M. Valin, "On adjusting the learning rate in frequency domain echo cancellation with double-talk," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[3] J. Madrid Portillo, "Deep learning applied to acoustic echo cancellation," Master's Thesis, Aalborg University, 2017.

[4] S. Gustafsson, R. Martin and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony state of the art and perspectives," in *Proc. EUSIPCO*, 1996.