Deep Learning with Augmented Kalman Filter for Single-Channel Speech Enhancement

Sujan Kumar Roy, Aaron Nicolson, Kuldip K. Paliwal
Signal Processing Laboratory, School of Engineering
Griffith University, Brisbane, QLD, Australia

Presented By:
Sujan Kumar Roy

Paper ID: 1662

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

OUTLINES

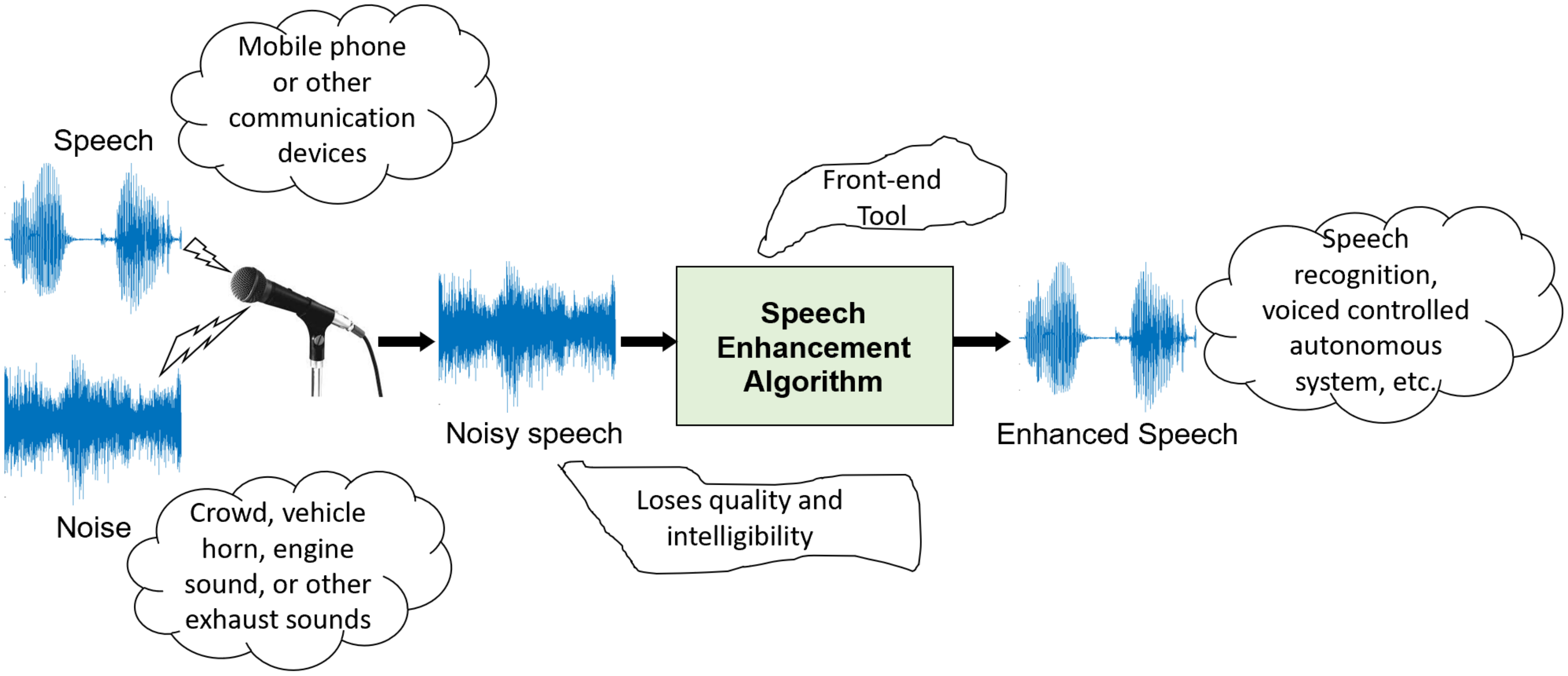

Fig. 1: A typical single channel speech enhancement system.

Literature Review

Existing SEAs

- Spectral Subtraction (SS) (Boll. 1979) [1]-[2]
  - Highly depends on noise estimation
  - Under/over estimation of noise causes ***musical*** noise and ***distortion*** in the enhanced speech
- MMSE (Ephraim and Malah 1984, 1985) [3]-[4] Wiener Filter [5]-[6]
  - Suffers from *a priori* SNR estimation accurately in noisy conditions
  - The enhanced speech suffers from ***distortion*** and ***musical*** noise
- Kalman Filter (KF) (Paliwal and Basu 1987) [7]
  - Introduced for enhancing white noise corrupted speech
  - Suffers from clean speech LPC parameter estimation in practice
- Augmented Kalman Filter (AKF) (Gibson et al. 1991) [10]
  - Introduced for speech enhancement in colored noise conditions
  - Suffers from speech and noise LPC parameter estimation in practice

Objective

- We focused on improving the AKF performance for speech enhancement in various noise conditions by incorporating machine learning technique
- Specifically, we utilize the LPC estimates of speech and noise signal for the AKF using Deep Learning Technique

AUGMENTED KALMAN FILTER FOR SPEECH ENHANCEMENT (1/2)

Signal Model

$$y(n) = s(n) + v(n) \qquad (1)$$

where colored noise $v(n)$ is assumed to be additive and uncorrelated with clean speech $s(n)$, $y(n)$ is noisy speech, and n is sample index)

Autoregressive Process of Speech and Noise Signal [16]

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + w(n), \qquad (2)$$

$$v(n) = -\sum_{k=1}^{q} b_k v(n-k) + u(n), \qquad (3)$$

- a_i's and b_k's are p^{th} and q^{th} order LPCs of speech and noise
- $w(n)$ and $v(n)$ are assumed to be zero mean and white noise with variances, σ_w^2 and σ_u^2

Augmented State-Space Model of AKF [10]

$$\boldsymbol{x}(n) = \boldsymbol{\Phi}\boldsymbol{x}(n-1) + \boldsymbol{d}\boldsymbol{z}(n), \qquad (4)$$

$$y(n) = \boldsymbol{c}^\top \boldsymbol{x}(n), \qquad (5)$$

In the above ASSM,

1) $\boldsymbol{x}(n) = [s(n) \ \ldots s(n-p+1)\ v(n) \ \ldots \ v(n-q+1)]^T$ is a $(p+q)\times 1$ state-vector,

2) $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_s & 0 \\ 0 & \boldsymbol{\Phi}_v \end{bmatrix}$ is a $(p+q)\times(p+q)$ state-transition matrix with:

$$\boldsymbol{\Phi}_s = \begin{bmatrix} -a_1 & -a_2 & \ldots & a_{p-1} & a_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \qquad \boldsymbol{\Phi}_v = \begin{bmatrix} -b_1 & -b_2 & \ldots & b_{q-1} & b_q \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}$$

3) $\boldsymbol{d} = \begin{bmatrix} \boldsymbol{d}_s & 0 \\ 0 & \boldsymbol{d}_v \end{bmatrix}$, where $\boldsymbol{d}_s = [1 \quad 0 \quad \ldots \quad 0]^\top$, $\boldsymbol{d}_v = [1 \quad 0 \quad \ldots \quad 0]^\top$,

4) $\boldsymbol{z}(n) = \begin{bmatrix} w(n) \\ u(n) \end{bmatrix}$,

5) $\boldsymbol{c}^T = [\boldsymbol{c}_s^T \quad \boldsymbol{c}_v^T]$, where $\boldsymbol{c}_s = [1 \quad 0 \quad \ldots \quad 0]^T$ and $\boldsymbol{c}_v = [1 \quad 0 \quad \ldots \quad 0]^T$ are $p\times 1$ and $q \times 1$ vectors,

6) $y(n)$ is the noisy measurement at sample n.

AUGMENTED KALMAN FILTER FOR SPEECH ENHANCEMENT (2/2)

Augmented Kalman filter based SEA (Contd.)

- Firstly, *y(n)* is converted to ono-overlapped frames (20 ms). For a frame, AKF recursively computes an unbiased linear MMSE estimate $\hat{\boldsymbol{x}}(n|n)$ given noisy speech $y(n)$ by using the following equations [10].

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{\Phi}\hat{\boldsymbol{x}}(n-1|n-1), \quad (6)$$

$$\boldsymbol{\Psi}(n|n-1) = \boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^\top + \boldsymbol{d}\boldsymbol{Q}\boldsymbol{d}^\top, \quad (7)$$

$$\boldsymbol{K}(n) = \boldsymbol{\Psi}(n|n-1)\boldsymbol{c}(\boldsymbol{c}^\top\boldsymbol{\Psi}(n|n-1)\boldsymbol{c})^{-1}, \quad (8)$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^\top\hat{\boldsymbol{x}}(n|n-1)], \quad (9)$$

$$\boldsymbol{\Psi}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^\top]\boldsymbol{\Psi}(n|n-1), \quad (10)$$

where $\boldsymbol{Q} = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix}$ is the process noise covariance.

- Estimated speech at sample *n* [13]:

$$\hat{s}(n|n) = [1 - K_0(n)]\hat{s}(n|n-1) + K_0(n)[y(n) - \hat{v}(n|n-1)] \quad (11)$$

where $K_0(n)$ is the 1^{st} component of $\boldsymbol{K}(n)$ given by [13]:

$$K_0(n) = \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \beta^2(n) + \sigma_u^2}, \quad (12)$$

$\alpha^2(n)$ and $\beta^2(n)$ are the transmission of a posteriori error variance [13]

Unknown Parameters: $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$

PROPOSED SPEECH ENHANCEMENT ALGORITHM (1/3)

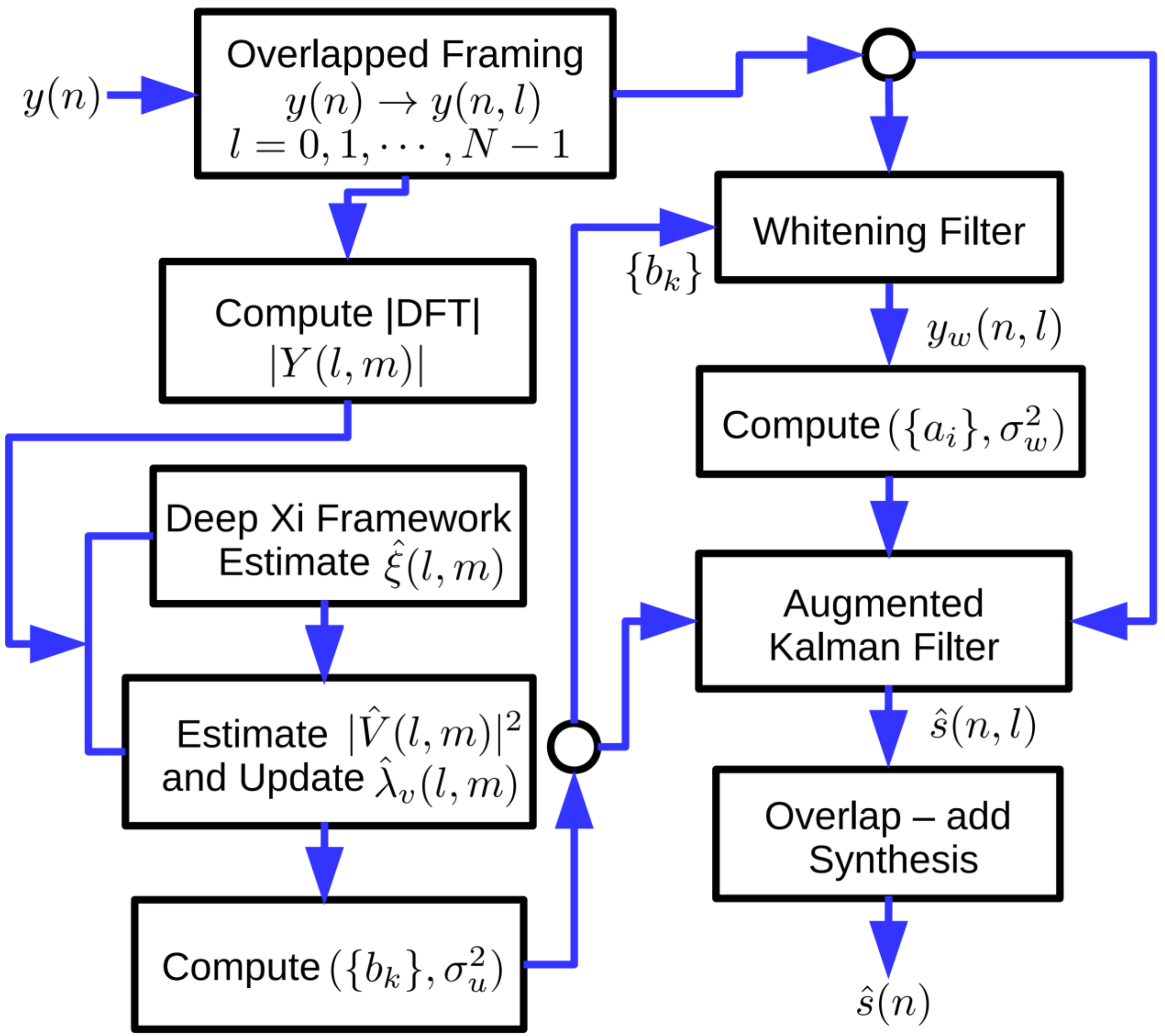

Fig2. Block-diagram of the proposed SEA.

Time-domain Framing

- $y(n,l) = s(n,l) + v(n,l)$ (using rectangular window with 50% overlap)
- l is frame index and N is the number of samples in each frame

STFT-Analysis of Noisy Speech

- $Y(l,m) = S(l,m) + V(l,m),$ (13) (STFT with 50% overlap Hamming window, m is frequency bin index)

It is assumed that $S(l,m)$ and $V(l,m)$ follow a Gaussian distribution with zero-mean and variances $E\{|S(l,m)|^2\} = \lambda_s(l,m)$, and $E\{|V(l,m)|^2\} = \lambda_v(l,m)$, where $E\{\cdot\}$ represents the statistical expectation operator.

PROPOSED SPEECH ENHANCEMENT ALGORITHM (2/3)

Proposed $(\{a_i\}, \sigma_w^2)$, and $(\{b_k\}, \sigma_u^2)$ Estimation

- Compute $(\{b_k\}, \sigma_u^2)$ from estimated noise PSD, $\widehat{\lambda}_v(l,m)$
- Estimate noise power, $|\widehat{V}(l,m)|^2$ using simplified MMSE method (setting $\hat{\gamma}(l,m) = \xi(l,m) + 1$, where $\xi(l,m) = \frac{\lambda_s(l,m)}{\lambda_v(l,m)}$ and $\hat{\gamma}(l,m)$ are *a priori* and *a posteriori* SNRs [17]-[18].
- The |IDFT| of $\widehat{\lambda}_v(l,m)$ yields the estimated auto-correlation coefficients, $\widehat{R}_{vv}(\tau)$, where τ is the autocorrelation lag
- By solving $\widehat{R}_{vv}(\tau)$ using Levinson-Durbin recursion [16], gives, $(\{b_k\}, \sigma_u^2)$ $(q = 40)$
- By employing whitening filter, $H_w(z)$ eq. (17) to $y(n,l)$, yielding a pre-whitened speech, $y_w(n,l)$ [13], [16]:
- Then compute $(\{a_i\}, \sigma_w^2)$ from $y_w(n,l)$ by using autocorrelation method [16].

$$|\widehat{V}(l,m)|^2 = \left(\frac{1}{1+\xi(l,m)}\right)|Y(l,m)|^2, \tag{14}$$

$$\xi(l,m) = \frac{\lambda_s(l,m)}{\lambda_v(l,m)}, \tag{15}$$

$$\widehat{\lambda}_v(l,m) = \eta\widehat{\lambda}_v(l-1,m) + (1-\eta)|\widehat{V}(l,m)|^2. \tag{16}$$

where η is a smoothing constant and set to 0.9.

$$H_w(z) = 1 + \sum_{k=1}^{q} b_k z^{-k}. \tag{17}$$

Unknown parameter, $\hat{\xi}(l,m)$ is estimated using the Deep Xi Framework [19]

Proposed Speech Enhancement Algorithm (3/3)

Proposed $\hat{\bar{\xi}}(l,m)$ Estimation using Deep Xi Framework constructed with ResNet []

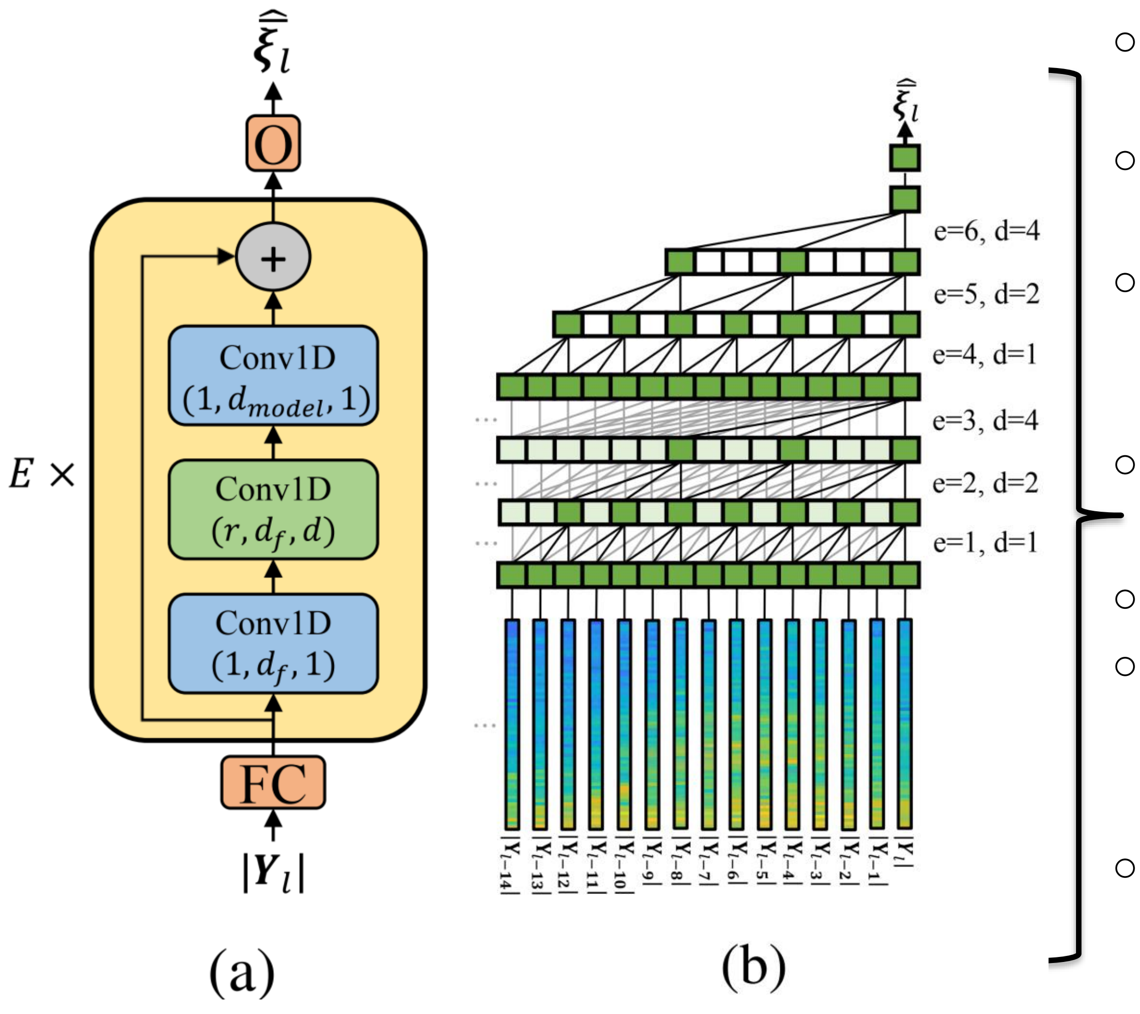

Fig. 3: Deep Xi-ResNet and (b) example of the contextual field of Deep Xi-ResNet with *D* = *4*, *E* = *6*, and *r* = *3*.

- It takes $\boldsymbol{Y}_l = \{Y_l(0), Y_l(1), \cdots, Y_l(m-1)\}$ as input, yields $\hat{\bar{\xi}}_l$. $\boldsymbol{Y}_l$ is passed through **FC**, followed by *E* =40 blocks, where *e* is the block index
- Each block contains three Convolutional unit (CU), with (***kernel_size***, ***output_size***, ***dilation_rate***) as (r, d_f, d)
- CU1 and CU2 have $d_f = 64$, while $d_{model} = 256$ for CU3. The CU1 and CU3 have *r=1*, *d=1*, while $r = 3$ cyclic $d = 2^{(e-1 \, mod \, (log_2 \, (D))+1)}$ is used for CU2 with *D=16*
- An example with *D* = *4*, and *E* = *6 in* Fig. 3 (b) shows that the DR is reset after block three, which increases the contextual field.
- The last, *e=40* is passed through O (output) followed by sigmoidal unit.
- As training target, mapped *a priori* SNR is used:

$$\bar{\xi}(l,m) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\xi_{\text{dB}}(l,m) - \mu_m}{\sigma_m\sqrt{2}}\right)\right]. \qquad (18)$$

- During inference, $\hat{\xi}(l,m)$ is found from $\hat{\xi}_{\text{dB}}(l,m)$

$$\hat{\xi}(l,m) = 10^{(\hat{\xi}_{\text{dB}}(l,m)/10)}, \qquad (19)$$

where the $\hat{\xi}_{\text{dB}}(l,m)$ is computed from $\hat{\bar{\xi}}(l,m)$ as follows:

$$\hat{\xi}_{\text{dB}}(l,m) = \sigma_m\sqrt{2}\text{erf}^{-1}\left(2\hat{\bar{\xi}}(l,m) - 1\right) + \mu_m. \qquad (20)$$

EXPERIMENTAL SETUP (1/2)

Training Strategy

- 74, 250 clean speech recordings are used in the training set [19]
- 2, 382 noise recordings are used as the noise training set [19]
- 5% of clean speech and noise signals are used as validation set
- All speech and noise are single-channel with sampling frequency 16 kHz
- Cross-entropy as the loss function.
- The Adam algorithm with default hyper-parameters is used for gradient descent optimization [28].
- Gradients are clipped between $[-1, 1]$.
- 175 epochs are used to train the ResNet.
- The noisy signals are created as follows: each randomly selected clean speech for the mini-batch (size=10) is mixed with a randomly selected noise at a randomly selected SNR level (-10 to 20 dB, in 1 dB increments)

EXPERIMENTAL SETUP (2/2)

Test Set
[9, *Chapter 12*]

a. 30 sentences (15+15 male & female) from NOIZEUS corpus [9]
b. *Babble* and *factory2* noises from NOISEX- 92 database [29].
c. All test datasets are single-channel with sampling frequency 16 kHz
d. **Stimuli set:** Corrupt (a) with (b) (-5dB to 15 dB SNRs)

Performance Evaluation Methods

- Perceptual evaluation of speech quality (PESQ) [0.5-4.5] (**Object quality**) [30]
- Quasi-stationary speech transmission index (QSTI) (**Object intelligibility~%**) [31]
- Spectrogram analysis (**Object quality**)
- AB Listening test [1-5] (**Subjective evaluation**) [32]
  - Participates 5 English speaking listeners
  - (sp05 corrupted with 5 dB *babble* noise)

Specifications of competitive SEAs

- **AKF-Ideal:** Parameters are computed in ideal case.
- **MMSE-STSA (**Ephraim and Malah**,** IEEE Trans. of A. S. S. P, **1984)**
- **AKF-IT** (Gibson et al. IEEE Trans. on S. P., **1991**)
- **RMBT-AKF (**George et al., Speech Communication, **2018)**
- **Proposed Method**

EXPERIMENTAL RESULTS & DISCUSSION (1/3)

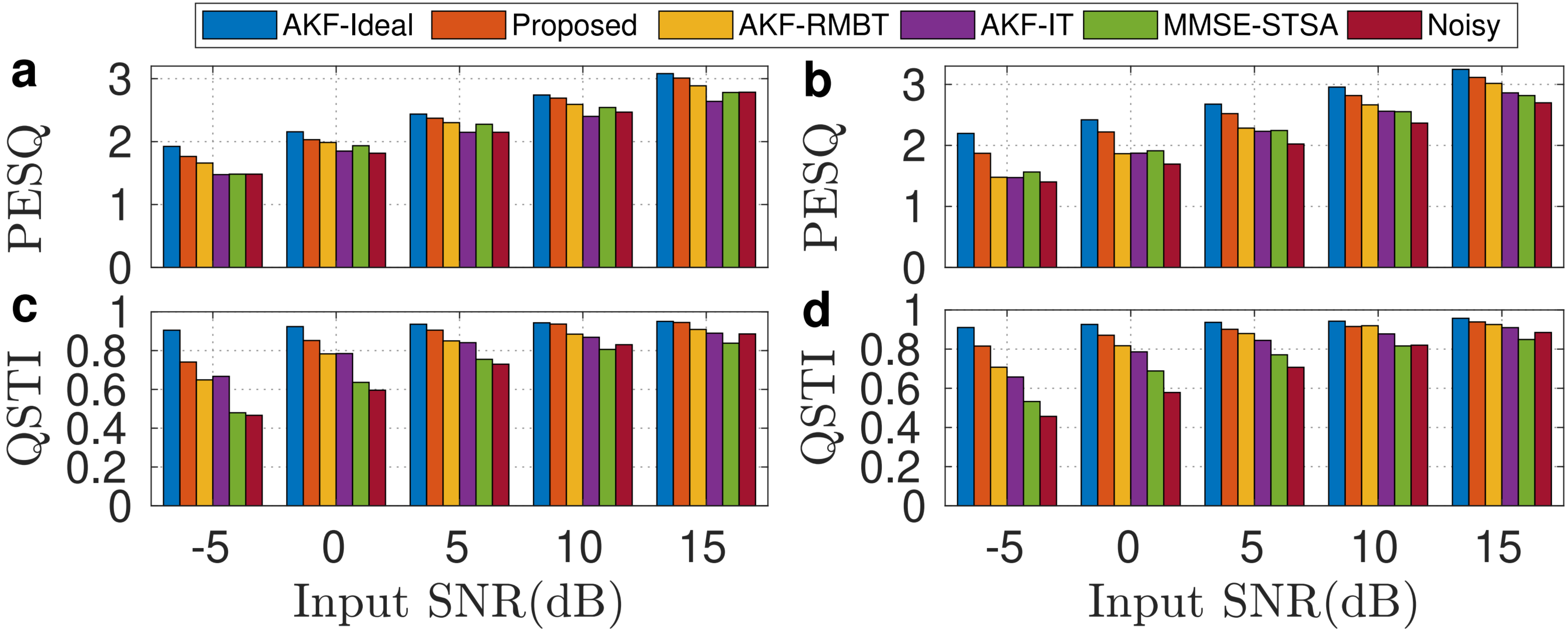

Fig. 4. Performance comparison of the SEAs in terms of average: PESQ; (a) *babble*, (b) *factory2* and QSTI; (c) *babble*, (d) *factory2* noise conditions.

EXPERIMENTAL RESULTS & DISCUSSION (2/3)

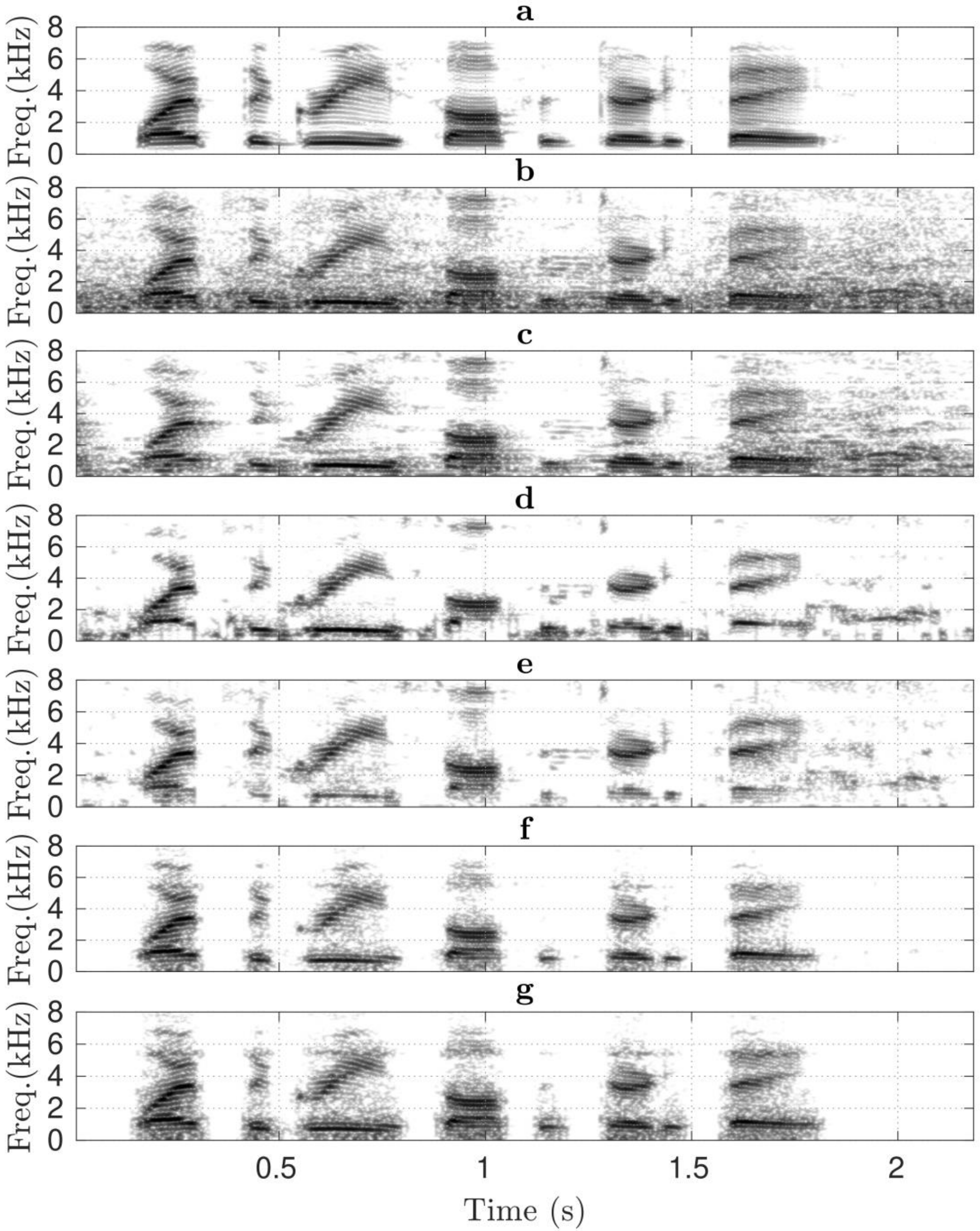

Fig. 5. (a) Clean speech, (b) noisy speech (sp05 is corrupted with 5 dB babble noise), the enhanced speech spectrograms produced by the: (c) MMSE-STSA, (d) AKF-IT, (e) AKF-RMBT, (f) proposed, and (g) AKF-Ideal methods.

EXPERIMENTAL RESULTS & DISCUSSION (3/3)

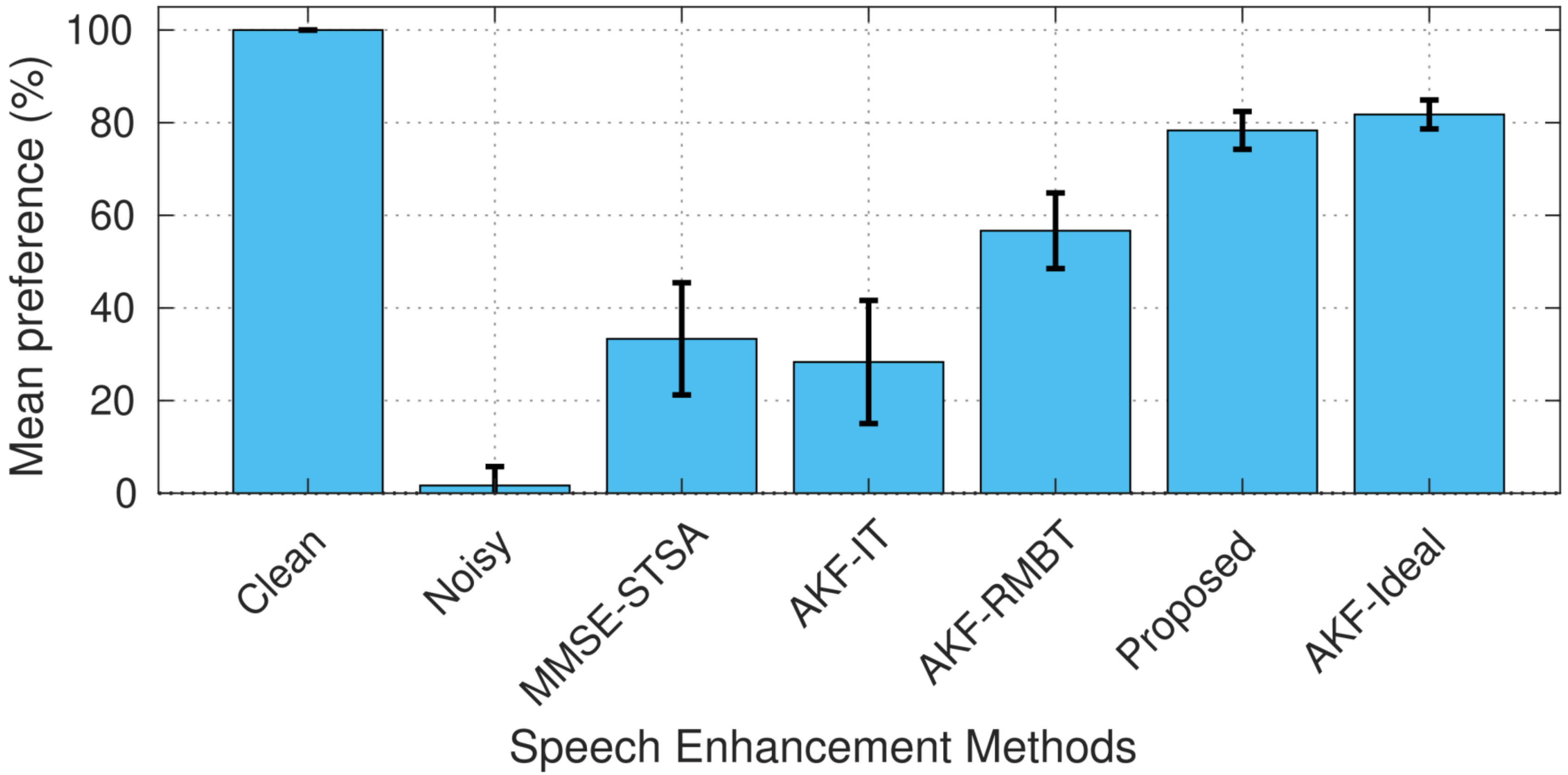

Fig. 6. The mean preference score (%) for each SEA on sp05 corrupted with 5 dB *babble* noise.

CONCLUSIONS

- We investigate a deep learning-based augmented Kalman filter for speech enhancement.
- A Deep Xi-ResNet-based noise PSD estimator is used to compute the noise LPC parameters.
- A whitening filter constructed with the noise LPCs is used to pre-whiten the noisy speech prior to speech LPC parameter estimation.
- The improved speech and noise LPCs enable the AKF to minimize the *residual* noise and *distortion* in the enhanced speech.
- Objective and subjective testing confirms that the proposed method outperforms the benchmark methods in various noise conditions.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208–211, April 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

[5] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 629–632, May 1996.

[6] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 289–292, May 2004.

[7] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 177–180, April 1987.

[8] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Procedia Computer Science*, vol. 54, pp. 574 – 584, 2015.

[9] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[10] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, August 1991.

[11] S. So, A. E. W. George, R. Ghosh, and K. K. Paliwal, "A non-iterative Kalman filtering algorithm with dynamic gain adjustment for single-channel speech enhancement," *International Journal of Signal Processing Systems*, vol. 4, pp. 263–268, August 2016.

[12] S. So, A. E. W. George, R. Ghosh, and K. K. Paliwal, "Kalman filter with sensitivity tuning for improved noise reduction in speech," *Circuits, Systems, and Signal Processing*, vol. 36, no. 4, pp. 1476–1492, April 2017.

[13] A. E. W. George, S. So, R. Ghosh, and K. K. Paliwal, "Robustness metric-based tuning of the augmented Kalman filter for the enhancement of speech corrupted with coloured noise," *Speech Communication*, vol. 105, pp. 62 – 76, December 2018.

[14] H. Yu, Z. Ouyang, W. Zhu, B. Champagne, and Y. Ji, "A deep neural network based Kalman filter for time domain speech enhancement," *IEEE International Symposium on Circuits and Systems*, pp. 1–5, May 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, June 2016.

[16] S. V. Vaseghi, "Linear prediction models," in *Advanced Digital Signal Processing and Noise Reduction*, chapter 8, pp. 227–262. John Wiley & Sons, 2009.

[17] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4266–4269, March 2010.

[18] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, December 2012.

[19] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, August 2019.

[20] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *ArXiv*, vol. abs/1803.01271, 2018.

[21] J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *ArXiv*, vol. abs/1607.06450, 2016.

[22] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *27th International Conference on Machine Learning*, pp. 807–814, June 2010.

[23] Y. Luo and N. Mesgarani, "Tasnet: Surpassing ideal time-frequency masking for speech separation," *ArXiv*, vol. abs/1809.07454, 2018.

[24] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. V. D. Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," *ArXiv*, vol. abs/1610.10099, 2016.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, April 2015.

[26] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ArXiv*, vol. abs/1412.6980, 2014.

[29] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, May 2001.

[31] B. Schwerin and K. K. Paliwal, "An improved speech transmission index for intelligibility prediction," *Speech Communication*, vol. 65, pp. 9–19, December 2014.

[32] K. K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010.

Thanks for Your Attention