Documents

Saleem, N.[a b] , Gunawan, T.S.[b c] , Shafi, M.[d] , Bourouis, S.[e] , Trigui, A.[f]

Multi-Attention Bottleneck for Gated Convolutional Encoder-Decoder-Based Speech Enhancement
(2023) *IEEE Access*, 11, pp. 114172-114186.

DOI: 10.1109/ACCESS.2023.3324210

[a] Gomal University, Faculty of Engineering and Technology, Department of Electrical Engineering, Dera Ismail Khan, 29050, Pakistan
[b] International Islamic University Malaysia (IIUM), Information Systems Department, Kuala Lumpur, 53100, Malaysia
[c] Telkom University, School of Electrical Engineering, Bandung, 40257, Indonesia
[d] Sohar University, Faculty of Computing and Information Technology, Sohar, 311, Oman
[e] Taif University, College of Computers and Information Technology, Department of Information Technology, Taif, 21944, Saudi Arabia
[f] King Khalid University, College of Computer Science, Department of Computer Science, Abha, 61421, Saudi Arabia

Abstract
Convolutional encoder-decoder (CED) has emerged as a powerful architecture, particularly in speech enhancement (SE), which aims to improve the intelligibility and quality and intelligibility of noise-contaminated speech. This architecture leverages the strength of the convolutional neural networks (CNNs) in capturing high-level features. Usually, the CED architectures use the gated recurrent unit (GRU) or long-short-term memory (LSTM) as a bottleneck to capture temporal dependencies, enabling a SE model to effectively learn the dynamics and long-term temporal dependencies in the speech signal. However, Transformers neural networks with self-attention effectively capture long-term temporal dependencies. This study proposes a multi-attention bottleneck (MAB) comprised of a self-attention Transformer powered by a time-frequency attention (TFA) module followed by a channel attention module (CAM) to focus on the important features. The proposed bottleneck (MAB) is integrated into a CED architecture and named MAB-CED. The MAB-CED uses an encoder-decoder structure including a shared encoder and two decoders, where one decoder is dedicated to spectral masking and the other is used for spectral mapping. Convolutional Gated Linear Units (ConvGLU) and Deconvolutional Gated Linear Units (DeconvGLU) are used to construct the encoder-decoder framework. The outputs of two decoders are coupled by applying coherent averaging to synthesize the enhanced speech signal. The proposed speech enhancement is examined using two databases, VoiceBank+DEMAND and LibriSpeech. The results show that the proposed speech enhancement outperforms the benchmarks in terms of intelligibility and quality at various input SNRs. This indicates the performance of the proposed MAB-CED at improving the average PESQ by 0.55 (22.85%) with VoiceBank+DEMAND and by 0.58 (23.79%) with LibriSpeech. The average STOI is improved by 9.63% (VoiceBank+DEMAND) and 9.78% (LibriSpeech) over the noisy mixtures.

Author Keywords
channel attention; gated convolutional encoder-decoder; Multi-attention; speech enhancement; time-frequency attention; transformer

Index Keywords
Convolution, Decoding, Long short-term memory, Memory architecture, Network architecture, Network coding, Quality control, Speech communication, Speech enhancement; Channel attention, Convolutional encoders, Convolutional neural network, Decoding, Encoder-decoder, Encodings, Gated convolutional encoder-decoder, Multi-attention, Noise measurements, Time frequency, Time-frequency Analysis, Time-frequency attention, Transformer; Benchmarking

References
- Gupta, M., Singh, R.K., Singh, S.
  Analysis of optimized spectral subtraction method for single channel speech enhancement
  (2023) *Wireless Pers. Commun*, 128 (3), pp. 2203-2215.
  Feb

- Boll, S.
  Suppression of acoustic noise in speech using spectral subtraction
  (1979) *IEEE Trans. Acoust., Speech, Signal Process*, ASSP-27 (2), pp. 113-120.
  Apr

- Jannu, C., Vanambathina, S.D.
  Weibull and Nakagami speech priors based regularized NMF with adaptive Wiener

filter for speech enhancement
(2023) *Int. J. Speech Technol*, 26 (1), pp. 197-209.
Mar

- Ephraim, Y., Malah, D.
Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator
(1984) *IEEE Trans. Acoust., Speech, Signal Process*, ASSP-32 (6), pp. 1109-1121.
Dec

- Mukhutdinov, D., Alex, A., Cavallaro, A., Wang, L.
Deep learning models for single-channel speech enhancement on drones
(2023) *IEEE Access*, 11, pp. 22993-23007.

- Rosenbaum, T., Cohen, I., Winebrand, E., Gabso, O.
Differentiable mean opinion score regularization for perceptual speech enhancement
(2023) *Pattern Recognit. Lett*, 166, pp. 159-163.
Feb

- Lu, X., Tsao, Y., Matsuda, S., Hori, C.
Speech enhancement based on deep denoising autoencoder
(2013) *Proc. Interspeech*, pp. 436-440.
Aug

- Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.
An experimental study on speech enhancement based on deep neural networks
(2014) *IEEE Signal Process. Lett*, 21 (1), pp. 65-68.
Jan

- Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.
A regression approach to speech enhancement based on deep neural networks
(2015) *IEEE/ACM Trans. Audio, Speech, Language Process*, 23 (1), pp. 7-19.
Jan

- Chakrabarty, S., Habets, E.A.P.
Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks
(2019) *IEEE J. Sel. Topics Signal Process*, 13 (4), pp. 787-799.
Aug

- Saleem, N., Khattak, M.I., Al-Hasan, M., Qazi, A.B.
On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks
(2020) *IEEE Access*, 8, pp. 160581-160595.

- Jiang, Y., Zhou, H., Feng, Z.
Performance analysis of ideal binary masks in speech enhancement
(2011) *Proc. 4th Int. Congr. Image Signal Process*, 5, pp. 2422-2425.
Oct

- Bao, F., Abdulla, W.H.
A new ratio mask representation for CASAbased speech enhancement
(2019) *IEEE/ACM Trans. Audio, Speech, Language Process*, 27 (1), pp. 7-19.
Jan

- Williamson, D.S., Wang, Y., Wang, D.
Complex ratio masking for monaural speech separation
(2016) *IEEE/ACM Trans. Audio, Speech, Language Process*, 24 (3), pp. 483-492.
Mar

- Wang, Y., Narayanan, A., Wang, D.
  On training targets for supervised speech separation
  (2014) *IEEE/ACM Trans. Audio, Speech, Language Process*, 22 (12), pp. 1849-1858.
  Dec

- Saleem, N., Khattak, M.I.
  Deep neural networks for speech enhancement in complex-noisy environments
  (2020) *Int. J. Interact. Multimedia Artif. Intell*, 6 (1), pp. 84-91.

- Saleem, N., Khattak, M.I., Qazi, A.B.
  Supervised speech enhancement based on deep neural network
  (2019) *J. Intell. Fuzzy Syst*, 37 (4), pp. 5187-5201.
  Oct

- Soleymanpour, R., Soleymanpour, M., Brammer, A.J., Johnson, M.T., Kim, I.
  Speech enhancement algorithm based on a convolutional neural network reconstruction of the temporal envelope of speech in noisy environments
  (2023) *IEEE Access*, 11, pp. 5328-5336.

- Girirajan, S., Pandian, A.
  Real-time speech enhancement based on convolutional recurrent neural network
  (2023) *Intell. Autom. Soft Comput*, 35 (2), pp. 1987-2001.

- Xia, Y., Wang, J.
  Low-dimensional recurrent neural networkbased Kalman filter for speech enhancement
  (2015) *Neural Netw*, 67, pp. 131-139.
  Jul

- Saleem, N., Gao, J., Khattak, M.I., Rauf, H.T., Kadry, S., Shafi, M.
  DeepResGRU: Residual gated recurrent neural network-augmented Kalman filtering for speech enhancement and recognition
  (2022) *Knowl.-Based Syst*, 238.
  Feb

- Fu, S.-W., Yu, C., Hsieh, T.-A., Plantinga, P., Ravanelli, M., Lu, X., Tsao, Y.
  (2021) *MetricGAN+: An improved version of MetricGAN for speech enhancement*, arXiv:2104.03538

- Pascual, S., Bonafonte, A., Serra, J.
  SEGAN: Speech enhancement generative adversarial network
  (2017) *Proc. Interspeech*, pp. 3642-3646.

- Abdulatif, S., Cao, R., Yang, B.
  (2022) *CMGAN: Conformer-based metric-GAN for monaural speech enhancement*, arXiv:2209.11112

- Kim, E., Seo, H.
  SE-conformer: Time-domain speech enhancement using conformer
  (2021) *Proc. Interspeech*, pp. 2736-2740.

- Koizumi, Y., Karita, S., Wisdom, S., Erdogan, H., Hershey, J.R., Jones, L., Bacchiani, M.
  DF-Conformer: Integrated architecture of Conv-TasNet and Conformer using linear complexity self-attention for speech enhancement
  (2021) *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, pp. 161-165.
  Oct

- Tan, K., Wang, D.
  A convolutional recurrent neural network for realtime speech enhancement
  (2018) *Proc. Interspeech*, pp. 3229-3233.

- Karthik, A., MazherIqbal, J.L.
  Efficient speech enhancement using recurrent convolution encoder and decoder
  (2021) *Wireless Pers. Commun*, 119 (3), pp. 1959-1973.
  Aug

- Strake, M., Defraene, B., Fluyt, K., Tirry, W., Fingscheidt, T.
  Fully convolutional recurrent networks for speech enhancement
  (2020) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 6674-6678.
  May

- Xian, Y., Sun, Y., Wang, W., Naqvi, S.M.
  Multi-scale residual convolutional encoder decoder with bidirectional long short-term memory for single channel speech enhancement
  (2021) *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, pp. 431-435.
  Jan

- Xian, Y., Sun, Y., Wang, W., Naqvi, S.M.
  Convolutional fusion network for monaural speech enhancement
  (2021) *Neural Netw*, 143, pp. 97-107.
  Nov

- Roy, S.K., Paliwal, K.K.
  Causal convolutional encoder decoder-based augmented Kalman filter for speech enhancement
  (2020) *Proc. 14th Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*, pp. 1-7.
  Dec

- Wang, Z., Zhang, T., Shao, Y., Ding, B.
  LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement
  (2021) *Appl. Acoust*, 172.
  Jan

- Tan, K., Wang, D.
  Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement
  (2019) *IEEE/ACM Trans. Audio, Speech, Language Process*, 28, pp. 380-390.

- Pandey, A., Wang, D.
  TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain
  (2019) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 6875-6879.
  May

- Zhao, H., Zarar, S., Tashev, I., Lee, C.-H.
  Convolutional-recurrent neural networks for speech enhancement
  (2018) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 2401-2405.
  Apr

- Li, A., Yuan, M., Zheng, C., Li, X.
  Speech enhancement using progressive learning-based convolutional recurrent neural network
  (2020) *Appl. Acoust*, 166.
  Sep

- Tan, K., Wang, D.
  Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement
  (2019) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 6865-6869.
  May

- Hsieh, T.-A., Wang, H.-M., Lu, X., Tsao, Y.
  WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement
  (2020) *IEEE Signal Process. Lett*, 27, pp. 2149-2153.

- Ullah, R., Wuttisittikulkij, L., Chaudhary, S., Parnianifard, A., Shah, S., Ibrar, M., Wahab, F.-E.
  End-to-end deep convolutional recurrent models for noise robust waveform speech enhancement
  (2022) *Sensors*, 22 (20), p. 7782.
  Oct

- Braun, S., Gamper, H., Reddy, C.K.A., Tashev, I.
  Towards efficient models for real-time deep noise suppression
  (2021) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 656-660.
  Jun

- Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Xie, L.
  DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement
  (2020) *Proc. Interspeech*, pp. 2472-2476.
  Oct

- Valentini, C.
  (2016) *Noisy speech database for training speech enhancement algorithms and TTS models*,
  School Inform., Centre Speech Res., Univ. Edinburgh, Tech. Rep

- Panayotov, V., Chen, G., Povey, D., Khudanpur, S.
  Librispeech: An ASR corpus based on public domain audio books
  (2015) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 5206-5210.
  Apr

- Lakew, S.M., Cettolo, M., Federico, M.
  A comparison of transformer and recurrent neural networks on multilingual neural machine translation
  (2018) *Proc. 27th Int. Conf. Comput. Linguistics*, pp. 641-652.

- Zhang, Q., Song, Q., Nicolson, A., Lan, T., Li, H.
  Temporal convolutional network with frequency dimension adaptive attention for speech enhancement
  (2021) *Proc. Interspeech*, pp. 166-170.
  Aug

- Zhang, Q., Qian, X., Ni, Z., Nicolson, A., Ambikairajah, E., Li, H.
  A time-frequency attention module for neural speech enhancement
  (2022) *IEEE/ACM Trans. Audio, Speech, Language Process*, 31, pp. 462-475.

- Dean, D., Sridharan, S., Vogt, R., Mason, M.
  The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms
  (2010) *Proc. Interspeech*, pp. 3110-3113.
  Sep

- Wang, Y., Wang, D.
  Towards scaling up classification-based speech separation
  (2013) *IEEE Trans. Audio, Speech, Language Process*, 21 (7), pp. 1381-1390.
  Jul

- Steeneken, H.J., Geurtsen, F.W.
  Description of the RSG-10 noise database
  (1988) *Tech. Rep., 3*,

- Varga, A., Steeneken, H.J.M.
  Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems
  (1993) *Speech Commun*, 12 (3), pp. 247-251.
  Jul

- Salamon, J., Jacoby, C., Bello, J.P.
  A dataset and taxonomy for urban sound research
  (2014) *Proc. 22nd ACM Int. Conf. Multimedia*, pp. 1041-1044.
  Nov

- Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.
  A shorttime objective intelligibility measure for time-frequency weighted noisy speech
  (2010) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 4214-4217.
  Mar

- Andersen, A.H., De Haan, J.M., Tan, Z.-H., Jensen, J.
  A non-intrusive short-time objective intelligibility measure
  (2017) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 5085-5089.
  Mar

- Rix, A.W., Hollier, M.P., Hekstra, A.P., Beerends, J.G.
  Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment. Part I-Time-delay compensation
  (2002) *J. Audio Eng. Soc*, 50 (10), pp. 755-764.

- Hu, Y., Loizou, P.C.
  Evaluation of objective quality measures for speech enhancement
  (2008) *IEEE Trans. Audio, Speech, Language Process*, 16 (1), pp. 229-238.
  Jan

- Hasannezhad, M., Ouyang, Z., Zhu, W.-P., Champagne, B.
  An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement
  (2020) *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, pp. 764-768.
  Dec

- Westhausen, N.L., Meyer, B.T.
  (2020) *Dual-signal transformation LSTM network for real-time noise suppression*, arXiv:2005.07551

- Kim, J., El-Khamy, M., Lee, J.
  T-GSA: Transformer with Gaussianweighted self-attention for speech enhancement
  (2020) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 6649-6653.
  May

- Zhang, Q., Nicolson, A., Wang, M., Paliwal, K.K., Wang, C.
  DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation
  (2020) *IEEE/ACM Trans. Audio, Speech, Language Process*, 28, pp. 1404-1415.

- Shah, N., Patil, H.A., Soni, M.H.
  Time-frequency mask-based speech enhancement using convolutional generative adversarial network
  (2018) *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA*

ASC), pp. 1246-1251.
Nov

- Chen, J., Wang, Y., Yoho, S.E., Wang, D., Healy, E.W.
Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises
(2016) *J. Acoust. Soc. Amer*, 139 (5), pp. 2604-2612.
May

- Fu, S.-W., Hu, T.-Y., Tsao, Y., Lu, X.
Complex spectrogram enhancement by convolutional neural network with multi-metrics learning
(2017) *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, pp. 1-6.
Sep

- Rothauser, E.H.
IEEE recommended practice for speech quality measurements
(1969) *IEEE Trans. Audio Electroacoust*, AE-17 (3), pp. 225-246.
Sep

- Li, A., Zheng, C., Peng, R., Li, X.
On the importance of power compression and phase estimation in monaural speech dereverberation
(2021) *JASA Exp. Lett*, 1 (1).
Jan

- Li, A., Liu, W., Zheng, C., Fan, C., Li, X.
Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement
(2021) *IEEE/ACM Trans. Audio, Speech, Language Process*, 29, pp. 1829-1843.

- Li, A., Zheng, C., Zhang, L., Li, X.
Glance and gaze: A collaborative learning framework for single-channel speech enhancement
(2022) *Appl. Acoust*, 187.
Feb

- Nikzad, M., Nicolson, A., Gao, Y., Zhou, J., Paliwal, K.K., Shang, F.
Deep residual-dense lattice network for speech enhancement
(2020) *Proc. AAAI Conf. Artif. Intell*, 34 (5), pp. 8552-8559.

- Defossez, A., Synnaeve, G., Adi, Y.
(2020) *Real time speech enhancement in the waveform domain*,
arXiv:2006.12847

- Wang, K., He, B., Zhu, W.-P.
TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain
(2021) *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 7098-7102.
Jun

- Fan, C., Yi, J., Tao, J., Tian, Z., Liu, B., Wen, Z.
Gated recurrent fusion with joint training framework for robust end-to-end speech recognition
(2021) *IEEE/ACM Trans. Audio, Speech, Language Process*, 29, pp. 198-209.

Correspondence Address
Saleem N.; Gomal University, Pakistan; email: nasirsaleem@gu.edu.pk

Publisher: Institute of Electrical and Electronics Engineers Inc.

ISSN: 21693536
Language of Original Document: English

Abbreviated Source Title: IEEE Access
2-s2.0-85174826114
Document Type: Article
Publication Stage: Final
Source: Scopus