Documents

Dhahbi, S.[a] , Saleem, N.[b] , Gunawan, T.S.[c] , Bourouis, S.[d] , Ali, I.[e] , Trigui, A.[f] , Algarni, A.D.[g]

Lightweight Real-Time Recurrent Models for Speech Enhancement and Automatic Speech Recognition

(2024) *International Journal of Interactive Multimedia and Artificial Intelligence*, 8 (6), pp. 74-85.

DOI: 10.9781/ijimai.2024.04.003

[a] Department of Computer science, College of science and art at Mahayil, King Khalid University, Muhayil Aseer62529, Saudi Arabia
[b] Department of Electrical Engineering, FET, Gomal University, KPK, D.I. Khan, 29050, Pakistan
[c] Electrical and Computer Engineering Department, Islamic International University Malaysia, Kuala Lumpur, Malaysia
[d] Department of Information Technology, College of Computers and Information Technology, Taif University, Taif, 21944, Saudi Arabia
[e] Department of Computer Science, University of Swat, Swat, Pakistan
[f] Department of Computer Science, College of Computer Science, King Khalid University, Abha, Saudi Arabia
[g] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia

Abstract
Traditional recurrent neural networks (RNNs) encounter difficulty in capturing long-term temporal dependencies. However, lightweight recurrent models for speech enhancement are important to improve noisy speech, while being computationally efficient and able to capture long-term temporal dependencies efficiently. This study proposes a lightweight hourglass-shaped model for speech enhancement (SE) and automatic speech recognition (ASR). Simple recurrent units (SRU) with skip connections are implemented where attention gates are added to the skip connections, highlighting the important features and spectral regions. The model operates without relying on future information that is well-suited for real-time processing. Combined acoustic features and two training objectives are estimated. Experimental evaluations using the short time speech intelligibility (STOI), perceptual evaluation of speech quality (PESQ), and word error rates (WERs) indicate better intelligibility, perceptual quality, and word recognition rates. The composite measures further confirm the performance of residual noise and speech distortion. With the TIMIT database, the proposed model improves the STOI and PESQ by 16.21% and 0.69 (31.1%) whereas with the LibriSpeech database, the model improves STOI by 16.41% and PESQ by 0.71 (32.9%) over the noisy speech. Further, our model outperforms other deep neural networks (DNNs) in seen and unseen conditions. The ASR performance is measured using the Kaldi toolkit and achieves 15.13% WERs in noisy backgrounds.

Author Keywords
Real-Time Speech; Simple Recurrent Unit (SRU); Speech Enhancement; Speech Processing; Speech Quality

Funding details
Princess Nourah Bint Abdulrahman UniversityPNU
Deanship of Scientific Research, King Khalid UniversityRGP2/383/44
Deanship of Scientific Research, King Khalid University
PNURSP2024R51

The current work was supported by the Deanship of Scientific Research at King Khalid University through large group Research Project under grant number RGP2/383/44. Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R51), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

The current work was supported by the Deanship of Scientific Research at King Khalid University through large group Research Project under grant number RGP2/383/44.

References

- Boll, S.
  Suppression of acoustic noise in speech using spectral subtraction
  (1979) *IEEE Transactions on acoustics, speech, and signal processing*, 27 (2), pp. 113-120.
- Nasir, S., Sher, A., Usman, K., Farman, U.
  Speech enhancement with geometric advent of spectral subtraction using connected time-frequency regions noise estimation
  (2013) *Research Journal of Applied Sciences, Engineering and Technology*, 6 (6), pp. 1081-1087.
- Lim, J., Oppenheim, A.
  All-pole modeling of degraded speech
  (1978) *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26 (3), pp. 197-210.

- Ephraim, Y., Malah, D.
  Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator
  (1984) *IEEE Transactions on acoustics, speech, and signal processing*, 32 (6), pp. 1109-1121.

- Ephraim, Y., Malah, D.
  Speech enhancement using a minimum mean-square error log-spectral amplitude estimator
  (1985) *IEEE transactions on acoustics, speech, and signal processing*, 33 (2), pp. 443-445.

- Mohammadiha, N., Smaragdis, P., Leijon, A.
  Supervised and unsupervised speech enhancement using nonnegative matrix factorization
  (2013) *IEEE Transactions on Audio, Speech, and Language Processing*, 21 (10), pp. 2140-2151.

- Tashev, I., Slaney, M.
  Data driven suppression rule for speech enhancement
  2013 Information Theory and Applications Workshop,

- Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.
  An experimental study on speech enhancement based on deep neural networks
  (2013) *IEEE Signal processing letters*, 21 (1), pp. 65-68.

- Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.
  A regression approach to speech enhancement based on deep neural networks
  (2014) *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23 (1), pp. 7-19.

- Kolbæk, M., Tan, Z.-H., Jensen, J.
  Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems
  (2016) *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25 (1), pp. 153-167.

- Hochreiter, S., Schmidhuber, J.
  Long short-term memory
  (1997) *Neural computation*, 9 (8), pp. 1735-1780.

- Wang, Y., Narayanan, A., Wang, D.
  On training targets for supervised speech separation
  (2014) *IEEE/ACM transactions on audio, speech, and language processing*, 22 (12), pp. 1849-1858.

- Saleem, N., Khattak, M.I.
  Deep neural networks for speech enhancement in complex-noisy environments
  (2020) *International Journal of Interactive Multimedia and Artificial Intelligence*, 6 (1), pp. 84-91.

- 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.

- Saleem, N., Khattak, M.I.
  Multi-scale decomposition based supervised single channel deep speech enhancement
  (2020) *Applied Soft Computing*, 95, p. 106666.

- Saleem, N., Khattak, M.I., Al-Hasan, M., Jan, A.
  Multi-objective long-short term memory recurrent neural networks for speech enhancement
  (2021) *Journal of Ambient Intelligence and Humanized Computing*, 12 (10), pp. 9037-9052.

- Samui, S., Chakrabarti, I., Ghosh, S.K.
  Time–frequency masking based supervised speech enhancement framework using fuzzy

deep belief net-work
(2019) *Applied Soft Computing*, 74, pp. 583-602.

- Soni, M.H., Shah, N., Patil, H.A.
Time-frequency masking-based speech enhancement using generative adversarial network
(2018) *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5039-5043.
IEEE

- Shah, N., Patil, H.A., Soni, M.H.
Time-frequency mask-based speech enhancement using convolutional generative adversarial network
(2018) *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1246-1251.
IEEE

- Yu, W., Zhou, J., Wang, H., Tao, L.
Setransformer: speech enhancement transformer
(2022) *Cognitive Computation*, 14, pp. 1152-1158.

- Cadore, J., Valverde-Albacete, F.J., Gallardo-Antolín, A., Peláez-Moreno, C.
Auditory-inspired morphological processing of speech spectrograms: Applications in automatic speech recognition and speech enhancement
(2013) *Cognitive computation*, 5, pp. 426-441.

- Sutskever, I., Vinyals, O., Le, Q.V.
Sequence to sequence learning with neural networks
(2014) *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2, pp. 3104-3112.

- Serban, I., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.
Building end-to-end dialogue systems using generative hierarchical neural network models
(2016) *Proceedings of the AAAI Conference on Artificial Intelligence*, 30 (1).

- Zarzycki, K., Ławryńczuk, M.
LSTM and GRU neural networks as models of dynamical processes used in predictive control: A comparison of models developed for two chemical reactors
(2021) *Sensors*, 21 (16), p. 5625.

- Chen, J., Wang, D.
Long short-term memory for speaker generalization in supervised speech separation
(2017) *The Journal of the Acoustical Society of America*, 141 (6), pp. 4705-4714.

- Sundermeyer, M., Ney, H., Schlüter, R.
From feedforward to recurrent lstm neural networks for language modeling
(2015) *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23 (3), pp. 517-529.

- Fernández-Díaz, M., Gallardo-Antolín, A.
An attention long short-term memory based system for automatic classification of speech intelligibility
(2020) *Engineering Applications of Artificial Intelligence*, 96, p. 103976.

- Gandapur, M. Q., Verdú, E.
ConvGRU-CNN: Spatiotemporal Deep Learning for Real-World Anomaly Detection in Video Surveillance System
(2023) *International Journal of Interactive Multimedia & Artificial Intelligence*, 8 (4).

- 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) *Knowledge-Based Systems*, 238, p. 107914.

- Ali Reshi, J., Ali, R.
  An Efficient Fake News Detection System Using Contextualized Embeddings and Recurrent Neural Network
  International Journal of Interactive Multimedia and Artificial Intelligence, pp. 1-13.

- Chang, B., Meng, L., Haber, E., Tung, F., Begert, D.
  (2017) *Multi-level residual networks from dynamical systems view*,
  arXiv preprint arXiv:1710.10348

- Shao, Y., Srinivasan, S., Jin, Z., Wang, D.
  A computational auditory scene analysis system for speech segregation and robust speech recognition
  (2010) *Computer Speech & Language*, 24 (1), pp. 77-93.

- Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S.
  Darpa timit acoustic-phonetic continous speech corpus cd-rom
  (1993) *NIST speech disc 1-1.1. NASA STI/Recon technical report*, (93), p. 27403.

- Panayotov, V., Chen, G., Povey, D., Khudanpur, S.
  Librispeech: an ASR corpus based on public domain audio books
  (2015) *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206-5210.
  IEEE

- Pearce, D., Picone, J.
  Aurora working group: DSR front end LVCSR evaluation AU/384/02
  (2002) *Inst. for Signal & Inform. Process*,
  Mississippi State Univ., Tech. Rep

- Varga, A., Steeneken, H.J.
  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 communication*, 12 (3), pp. 247-251.

- Damayanti, T. F., Wanto, A., Tambunan, H.S.
  Prediction of Palm Oil Seed Stock Production Results with the Back-propagation Algorithm
  (2023) *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, 2 (2), pp. 105-112.

- Song, Q., Wu, Y., Soh, Y.C.
  Robust adaptive gradient-descent training algorithm for recurrent neural networks in discrete time domain
  (2008) *IEEE Transactions on Neural networks*, 19 (11), pp. 1841-1853.

- Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.
  (2012) *Improving neural networks by preventing co-adaptation of feature detectors*,
  arXiv preprint arXiv:1207.0580

- 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) *Journal of the Audio Engineering Society*, 50 (10), pp. 755-764.

- Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.
  A short-time objective intelligibility measure for time-frequency weighted noisy speech
  (2010) *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214-4217.
  IEEE

- Hu, Y., Loizou, P.C.
  Evaluation of objective measures for speech enhancement
  (2006) *Ninth International Conference on Spoken Language Processing*,

- Kounovsky, T., Malek, J.
  Single channel speech enhancement using convolutional neural network
  (2017) *2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and Their Application to Mechatronics (ECMSM)*, pp. 1-5.
  IEEE

- Sun, P., Qin, J.
  Low-rank and sparsity analysis applied to speech enhancement via online estimated dictionary
  (2016) *IEEE Signal Processing Letters*, 23 (12), pp. 1862-1866.

- Shi, W., Zhang, X., Zou, X., Han, W., Min, G.
  Auditory mask estimation by RPCA for monaural speech enhancement
  (2017) *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 179-184.
  IEEE

- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Schwarz, P.
  The kaldi speech recognition toolkit
  (2011) *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*,
  IEEE Signal Processing Society

- Tachioka, Y., Watanabe, S., Le Roux, J., Hershey, J.R.
  Discriminative methods for noise robust speech recognition: A chime challenge benchmark
  (2013) *The 2nd International Workshop on Machine Listening in Multisource Environments*, pp. 19-24.

- Shewalkar, A., Nyavanandi, D., Ludwig, S.A.
  Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU
  (2019) *Journal of Artificial Intelligence and Soft Computing Research*, 9 (4), pp. 235-245.

- Li, A., Zheng, C., Zhang, L., Li, X.
  Glance and gaze: A collaborative learning framework for single-channel speech enhancement
  (2022) *Applied Acoustics*, 187, p. 108499.

- Pascual, S., Serra, J., Bonafonte, A.
  Time-domain speech enhancement using generative adversarial networks
  (2019) *Speech communication*, 114, pp. 10-21.

- Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Xie, L.
  (2020) *DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement*,
  arXiv preprint arXiv:2008.00264

- Nikzad, M., Nicolson, A., Gao, Y., Zhou, J., Paliwal, K.K., Shang, F.
  Deep residual-dense lattice network for speech enhancement
  (2020) *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, pp. 8552-8559.
  05

- Defossez, A., Synnaeve, G., Adi, Y.
  (2020) *Real time speech enhancement in the waveform domain*,
  arXiv preprint 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) *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7098-7102.
  IEEE

- Kim, E., Seo, H.
  SE-Conformer: Time-Domain Speech Enhancement Using Conformer
  (2021) *Interspeech*, pp. 2736-2740.

- Ye, Z., Saleem, N., Ali, H.
  Efficient Gated Convolutional Recurrent Neural Networks for Real-Time Speech Enhancement
  (2023) *International Journal of Interactive Multimedia and Artificial Intelligence*,

- Khattak, M.I., Jan, A., Saleem, N., Verdú, E., Khurshid, N.
  Automated detection of COVID-19 using chest X-ray images and CT scans through multilayer-spatial convolutional neural networks
  (2021) *International Journal of Interactive Multimedia and Artificial Intelligence*, 6 (6), pp. 15-24.

- Yao, G., Wang, C., Wu, Y., Wang, Y.
  Pyramid fully residual network for single image de-raining
  (2021) *Neurocomputing*, 456, pp. 168-178.

- Yue, H., Duo, W., Peng, X., Yang, J.
  Reference-based speech enhancement via feature alignment and fusion network
  (2022) *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (10), pp. 11648-11656.

- Saleem, N., Gunawan, T.S., Shafi, M., Bourouis, S., Trigui, A.
  Multi-Attention Bottleneck for Gated Convolutional Encoder-Decoder-Based Speech Enhancement
  (2023) *IEEE Access*, 11, pp. 114172-114186.

Correspondence Address
Saleem N.; Department of Electrical Engineering, KPK, Pakistan; email: nasirsaleem@gu.edu.pk

Publisher: Universidad Internacional de la Rioja

ISSN: 19891660
Language of Original Document: English
Abbreviated Source Title: Int. J. Interact. Multimed. Artif. Intell.
2-s2.0-85197316656
Document Type: Article
Publication Stage: Final
Source: Scopus