Documents

Wang, J.[a] , Saleem, N.[b c] , Gunawan, T.S.[c]

Towards Efficient Recurrent Architectures: A Deep LSTM Neural Network Applied to Speech Enhancement and Recognition
(2024) *Cognitive Computation*, 16 (3), pp. 1221-1236.

DOI: 10.1007/s12559-024-10288-y

[a] School of Materials Science and Engineering, Yunnan University, Yunnan Province, Kunming City, China
[b] Department of Electrical Engineering, Faculty of Engineering and Technology, Gomal University, Dera Ismail Khan, 29050, Pakistan
[c] Electrical and Computer Engineering Department, International Islamic University Malaysia (IIUM), Kuala Lumpur, Malaysia

Abstract
Long short-term memory (LSTM) has proven effective in modeling sequential data. However, it may encounter challenges in accurately capturing long-term temporal dependencies. LSTM plays a central role in speech enhancement by effectively modeling and capturing temporal dependencies in speech signals. This paper introduces a variable-neurons-based LSTM designed for capturing long-term temporal dependencies by reducing neuron representation in layers with no loss of data. A skip connection between nonadjacent layers is added to prevent gradient vanishing. An attention mechanism in these connections highlights important features and spectral components. Our LSTM is inherently causal, making it well-suited for real-time processing without relying on future information. Training involves utilizing combined acoustic feature sets for improved performance, and the models estimate two time–frequency masks—the ideal ratio mask (IRM) and the ideal binary mask (IBM). Comprehensive evaluation using perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) showed that the proposed LSTM architecture demonstrates enhanced speech intelligibility and perceptual quality. Composite measures further substantiated performance, considering residual noise distortion (Cbak) and speech distortion (Csig). The proposed model showed a 16.21% improvement in STOI and a 0.69 improvement in PESQ on the TIMIT database. Similarly, with the LibriSpeech database, the STOI and PESQ showed improvements of 16.41% and 0.71 over noisy mixtures. The proposed LSTM architecture outperforms deep neural networks (DNNs) in different stationary and nonstationary background noisy conditions. To train an automatic speech recognition (ASR) system on enhanced speech, the Kaldi toolkit is used for evaluating word error rate (WER). The proposed LSTM at the front-end notably reduced WERs, achieving a notable 15.13% WER across different noisy backgrounds.

Author Keywords
Acoustic features; Attention process; Deep learning; LSTM; Skip connections; Speech enhancement; Speech recognition

Index Keywords
Deep neural networks, Network architecture, Quality control, Speech enhancement, Speech intelligibility, Speech recognition; Acoustic features, Attention process, Deep learning, Neural-networks, Perceptual evaluation of speech qualities, Performance, Sequential data, Skip connection, Speech signals, Word error rate; Long short-term memory

References
- Boll, S.
  Suppression of acoustic noise in speech using spectral subtraction
  (1979) *IEEE Trans Acoust Speech Signal Process*, 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) *Res J Appl Sci Eng Technol*, 6 (6), pp. 1081-1087.

- Lim, J., Oppenheim, A.
  All-pole modeling of degraded speech
  (1978) *IEEE Trans Acoust Speech Signal Process*, 26 (3), pp. 197-210.

- Ephraim, Y., Malah, D.
  Speech enhancement using a minimum-mean square error short-time spectral

amplitude estimator
(1984) *IEEE Trans Acoust Speech Signal Process*, 32 (6), pp. 1109-1121.

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

- Xu, Y., Du, J., Dai, L.-R., Lee, C.-H.
  An experimental study on speech enhancement based on deep neural networks
  (2013) *IEEE Signal Process Lett*, 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 Trans Audio Speech Lang Process*, 23 (1), pp. 7-19.

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

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

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

- 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

- Yu, W., Zhou, J., Wang, H.
  SETransformer: speech enhancement transformer
  (2022) *Cogn Comput*, 14, pp. 1152-1158.

- Sutskever, I., Vinyals, O., Le, Q.V.
  Sequence to sequence learning with neural networks
  (2014) *Adv Neural Inf Process Syst*, p. 27.

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

- Zhu, Q.S., Zhang, J., Zhang, Z.Q., Dai, L.R.
  A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition
  (2023) *IEEE/ACM Trans Audio Speech Lang Process*, 31, pp. 1927-1939.

- 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.

- 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., Schl¨uter, 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.

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

- Fern´andez-D´ıaz, M., Gallardo-Antol´ın, A.
  An attention long short-term memory based system for automatic classification of speech intelligibility
  (2020) *Eng Appl Artif Intell*, 96, p. 103976.

- 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.

- El-Moneim, S.A., Nassar, M., Dessouky, M.I., Ismail, N.A., El-Fishawy, A.S., Abd El-Samie, F.E.
  Text-independent speaker recognition using lstm-rnn and speech enhancement
  (2020) *Multimedia tools and applications*, 79, pp. 24013-24028.

- Chang, B., Meng, L., Haber, E., Tung, F., Begert, D.
  (2017) *Multi-level residual networks from dynamical systems view*,
  arXiv preprint, arXiv:171010348

- Strake, M., Defraene, B., Fluyt, K., Tirry, W., Fingscheidt, T.
  Speech enhancement by lstm-based noise suppression followed by cnn-based speech restoration
  (2020) *EURASIP Journal on Advances in Signal Processing*, 2020, pp. 1-26.

- 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.

- Liang, R., Kong, F., Xie, Y., Tang, G., Cheng, J.
  Real-time speech enhancement algorithm based on attention lstm
  (2020) *IEEE Access*, 8, pp. 48464-48476.

- Li, X., Horaud, R.
  (2020) *Online monaural speech enhancement using delayed subband LSTM. Interspeech*, pp. 2462-2466.
  2005.05037

- Zhang, S., Kong, Y., Lv, S., Hu, Y., Xie, L.
  (2021) *FT-LSTM Based Complex Network for Joint Acoustic Echo Cancellation and Speech Enhancement*,
  arXiv preprint, 2106.07577

- Fedorov, I., Stamenovic, M., Jensen, C., Yang, L.C., Mandell, A., Gan, Y., Mattina, M., Whatmough, P.N.
  (2020) *TinyLSTMs: Efficient neural speech enhancement for hearing aids.*,
  arXiv preprint, 2005.11138

- Li, X., Li, Y., Dong, Y., Xu, S., Zhang, Z., Wang, D., Xiong, S.
(2020) *Bidirectional LSTM network with ordered neurons for speech enhancement*, pp. 2702-2706.
Inter Speech

- Saleem, N., Khattak, M.I., Al-Hasan, M., Jan, A.
Multi-objective long-short term memory recurrent neural networks for speech enhancement
(2021) *J Ambient Intell Humaniz Comput*, 12 (10), pp. 9037-9052.

- Goswami, R.G., Andhavarapu, S., Murty, K.
(2020) *Phase aware speech enhancement using realisation of complex-valued LSTM.*, arXiv preprint, arXiv:2010.14122

- Westhausen, N.L., Meyer, B.T.
Dual-signal transformation LSTM network for real-time noise suppression
Proc. Interspeech; 2020, pp. 2477-2481.
2005.07551

- Garg, A.
Speech enhancement using long short term memory with trained speech features and adaptive wiener filter
(2023) *Multimedia tools and applications*, 82 (3), pp. 3647-3675.

- Yu, J., Luo, Y.
Efficient monaural speech enhancement with universal sample rate band-split rnn
(2023) *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5.
IEEE

- Korkmaz, Y., Boyacı, A.
Hybrid voice activity detection system based on lstm and auditory speech features
(2023) *Biomed Signal Process Control*, 80.

- Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S.
(1993) *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1*, p. 27403.
NASA STI/Recon technical report n

- Panayotov, V., Chen, G., Povey, D., Khudanpur, S.
(2015) *Librispeech: an asr corpus based on public domain audio books*, pp. 5206-5210.
2015 IEEE international conference on acoustics speech and signal processing (ICASSP

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

- Varga, A., Steeneken, H.
Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems
(1993) *Speech Commun*, 12 (3), pp. 247-253.

- 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

- Yi, H.
(2006) *Evaluation of objective measures for speech enhancement*, pp. 1447-1450.
Pittsburgh, Pennsylvania, Interspeech

- 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 Process Lett*, 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

- Tan, K., Wang, D.
A convolutional recurrent neural network for real-time speech enhancement
(2018) *In: Interspeech*, 2018, pp. 3229-3233.

- Zhou, L., Gao, Y., Wang, Z., Li, J., Zhang, W.
(2021) *Complex spectral mapping with attention based convolution recurrent neural network for speech enhancement*,
arXiv preprint;, . arXiv:2104.05267

- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Silovsky, J.
The Kaldi speech recognition toolkit
(2011) *IEEE 2011 workshop on automatic speech recognition and understanding*,
IEEE Signal Processing Society

- Pascual, S., Bonafonte, A., Serrà, J.
(2017) *SEGAN: speech enhancement generative adversarial network*,
Interspeech

- Baby, D., Verhulst, S.
Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty
(2019) *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 106-110.
IEEE

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

- Lv, S., Fu, Y., Xing, M., Sun, J., Xie, L., Huang, J., Wang, Y., Yu, T.
S-dccrn: Super wide band dccrn with learnable complex feature for speech enhancement
(2022) *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP), pp. 7767-7771.
IEEE

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

- Chen, J., Wang, Z., Tuo, D., Wu, Z., Kang, S., Meng, H.
Fullsubnet+: channel attention fullsubnet with complex spectrograms for speech enhancement
(2022) *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7857-7861.
IEEE

- Passos, L.A., Papa, J.P., Hussain, A., Adeel, A.
Canonical cortical graph neural networks and its application for speech enhancement in audio-visual hearing aids
(2023) *Neurocomputing*, 527, pp. 196-203.

- Hussain, T., Wang, W.-C., Gogate, M., Dashtipour, K., Tsao, Y., Lu, X., Ahsan, A., Hussain, A.
A novel temporal attentive-pooling based convolutional recurrent architecture for acoustic signal enhancement
(2022) *IEEE transactions on artificial intelligence*, 3 (5), pp. 833-842.

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

Publisher: Springer

ISSN: 18669956
Language of Original Document: English
Abbreviated Source Title: Cognitive Comput.
2-s2.0-85191772206
Document Type: Article
Publication Stage: Final
Source: Scopus