Towards Efficient Recurrent Architectures: A Deep LSTM Neural Network Applied to Speech Enhancement and Recognition

By Wang, J (Wang, Jing) ; Saleem, N (Saleem, Nasir) ; Gunawan, TS (Gunawan, Teddy Surya)

View Web of Science ResearcherID and ORCID (provided by Clarivate)

Source COGNITIVE COMPUTATION

Volume: 16 Issue: 3 Page: 1221-1236

DOI: 10.1007/s12559-024-10288-y

Published MAY 2024

Early Access APR 2024

Indexed 2024-05-12

Document Type Article

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. Ou…

	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.
Keywords	**Author Keywords:** Deep learning; Speech enhancement; Speech recognition; Skip connections; LSTM; Acoustic features; Attention process **Keywords Plus:** NOISE
Addresses	[1] Yunnan Univ, Sch Mat Sci & Engn, Kunming City, Yunnan Province, Peoples R China [2] Gomal Univ, Fac Engn & Technol, Dept Elect Engn, Dera Ismail Khan 29050, Pakistan [3] Int Islamic Univ Malaysia IIUM, Dept Elect & Comp Engn, Kuala Lumpur, Malaysia
Categories/ Classification	Research Areas: Computer Science; Neurosciences & Neurology
Web of Science Categories	Computer Science, Artificial Intelligence; Neurosciences
Language	English
Accession Number	WOS:001210732600001

ISSN	1866-9956
eISSN	1866-9964
IDS Number	SY1S2

– See fewer data fields

Citation Network

In Web of Science Core Collection

0 Citations

57 Cited References

How does this document's citation performance compare to peers?

← Open comparison metrics panel

New

Data is from InCites Benchmarking & Analytics

Use in Web of Science

1	1
Last 180 Days	Since 2013

This record is from:

Web of Science Core Collection

- Science Citation Index Expanded (SCI-EXPANDED)

Suggest a correction

If you would like to improve the quality of the data in this record, please Suggest a correction