Circuits, Systems, and Signal Processing (2019) 38:3399–3405
https://doi.org/10.1007/s00034-019-01161-7

# Guest Editorial: Algorithms and Architectures for Machine Learning Based Speech Processing

**Tokunbo Ogunfunmi**[1] · **Ravi P. Ramachandran**[2] · **Roberto Togneri**[3] · **Brett Smolenski**[4] · **Visar Berisha**[5]

Published online: 26 June 2019

The Special Section in this issue of the Circuits, Systems and Signal Processing journal focuses on "Algorithms and Architectures for Machine Learning based Speech Processing."

It is well known that the applications of machine learning in all areas have been growing rapidly. In the field of speech processing, the growth is tremendous and this Special Section brings the readers a collection of papers that focus on many of these applications.

The underlying theme of the Special Section is on the theory, architectures, and applications that use machine learning techniques to achieve high performance and robustness to unknown or adverse conditions.

In general, the current trend that is emerging in applying machine learning approaches to many types of speech processing systems is referred to as deep learning. Deep learning employs deep neural networks (DNNs), which are neural networks with more than one hidden layer, with recently developed initialization and training strategies using massive amounts of diverse data as examples. These techniques have recently shown considerable promise in applications ranging from speech enhancement for hearing aids to speech, speaker, and language recognition.

Although several layer neural networks are not new, developing techniques to initialize and train them so that they converge to useful classifiers is a relatively recent breakthrough. Unlike other classifier architectures, DNNs are capable of exploiting vast amounts of development data resulting in robust classifiers that can generalize to non-pertinent variations in their targeted data. They also tend not to over-fit their

✉ Tokunbo Ogunfunmi
togunfunmi@scu.edu

[1] Santa Clara University, Santa Clara, USA

[2] Rowan University, Glassboro, USA

[3] University of Western Australia, Perth, Australia

[4] State University of New York Polytechnic Institute, New York, USA

[5] Arizona State University, Tempe, USA

large number of parameters to any irrelevant intricacies in their development data. Since it appears that this trend will continue, there will exist opportunities for circuits and systems researchers to develop hardware architectures to efficiently implement systems that use DNNs.

The Special Section has several themes:

1. Survey papers
2. Speech recognition papers
3. Phoneme recognition papers
4. Speaker verification papers
5. Speech quality and intelligibility papers
6. Compressive sensing applications in speech paper

The first theme is covered by the first two papers and is a collection of survey papers.

The first paper authored by Tokunbo Ogunfunmi et al. is a primer on the broad subject of deep learning architectures and applications in speech processing. The authors introduce the subject and provide a review of state-of-the-art and popular discriminative deep learning neural networks (DNN), convolutional neural networks (CNN) and recurrent neural networks (RNN) deep learning techniques, the basic framework and algorithms, hardware implementations, applications in speech, and the overall benefits of deep learning for speech processing. It is a useful guide to the newcomer to the field and discusses the issues and applications areas further discussed in subsequent papers in the Special Issue.

The second paper authored by Xianjun Xia et al. provides a comprehensive survey for the neural network-based deep learning approaches on acoustic event detection. Different deep learning-based acoustic event detection approaches are investigated with an emphasis on both strongly labeled and weakly labeled acoustic event detection systems. This paper also discusses how deep learning methods benefit the acoustic event detection task and the potential issues that need to be addressed for prospective real-world scenarios.

TThe third and fourth papers cover the general theme of speech recognition.

The third paper authored by Douglas O'Shaughnessy is titled "Recognition and Processing of Speech Signals using Neural Networks." In the past decade deep learning, which generally refers to the application of artificial neural networks with multiple layers between the input and output layers, has revolutionized the processing of speech signals. Large-scale automatic speech recognition was one of the first and most convincing successful cases of deep learning, which is now being applied in several other disparate fields. In the paper, the author takes us on a tour of this revolution from its modest beginnings to the current state of the art. The paper provides an overview of recent approaches to deep learning as applied to a range of speech processing tasks, primarily for automatic speech recognition (ASR), but also text-to-speech and speaker, language, and emotion recognition. The author focuses on efficient methods, addressing issues of accuracy, computation, storage, and delay. The discussion puts the tasks in the broader context of pattern recognition for processing speech, comparing with other signals. It also compares machine learning with other recent methods of speech analysis, e.g., hidden Markov models (HMM). The paper emphasizes a thorough understanding of the choices made in analyzing and interpreting speech signals.

The fourth paper authored by Pradeep Rangan et al. investigates and interprets the hidden layers representation of a deep network trained on the phoneme recognition task. The variants of recurrent neural networks (RNN) such as long short-term memory (LSTM) and gated recurrent unit (GRU) are successful in sequence modeling such as automatic speech recognition (ASR). However, the decoded sequence is prone to have false substitutions, insertions, and deletions. In this paper, the authors investigate the outcome of the hidden layers in LSTM trained on the TIMIT database. They found interestingly that the first hidden layer was capturing information related to some broad manners of articulation. The successive hidden layers try to cluster among the broad manners of articulation. The authors define an additional gate called the manner of articulation gate that is high if the broad manners of articulation of the $\mathbf{t}$'th frame are the same as that of the $(\mathbf{t}+1)$'th frame. The manner of articulation detection is embedded at the output of the activation gate of the LSTM at the first hidden layer. By doing so, the sonorants (vowels, semi-vowels, nasals) being substituted as obstruents (fricatives, stops, affricates) are minimized at the output layer. The proposed method decreased the phone error rates when evaluated on the core test set of the TIMIT database.

The fifth and sixth papers are on phoneme recognition.

The fifth paper authored by Abdolreza Sabzi Shahrebabaki et al. is a comparative study of deep learning techniques on frame-level speech data classification. This paper deals with the classical robustness issue of performance degradation under mismatched training and testing conditions. The mismatched condition is the speaking rate. The phoneme classification accuracy diminishes when a slower or faster speaking rate is used to test the system. The paper shows that various CNN architectures achieve a higher accuracy in classifying vowels and consonants over their DNN counterpart. Classification is performed at the frame level. The best features are the filter bank energies and the MFCC cepstrum.

The sixth paper authored by Xiyue Wang et al. covers the topic of automatic hypernasality detection in cleft palate speech using a convolutional neural network (CNN). Hypernasal speech refers to the perception of excessive nasal resonance in speech, caused by air escaping through the nasal cavity as a person is speaking. It negatively impacts intelligibility, and it is a common symptom in individuals with cleft palate. Cleft palate is the most common birth defect all over the world. The paper provides a thorough overview of the existing literature on hypernasality detection and presents an end-to-end algorithm based on convolutional neural networks to detect hypernasality from the speech of individuals with cleft palate and avoids the procedure of feature extraction. Various CNN architectures are explored. The results are presented on heterogeneous corpora and demonstrate that the approach proposed by the authors is better able to handle nuisance parameters in speech when compared to other state-of-the-art methods.

To cover the fourth theme of speaker verification, we have the seventh paper authored by Azharuddin Laskar et al. and titled "Integrating DNN-HMM technique with Hierarchical multi-Layer Acoustic Model (HiLAM) for text dependent speaker verification." The HiLAM uses Gaussian mixture model (GMM) and hidden Markov model (HMM) to incorporate the temporal information of the pass phrase, and has been found to outperform the subspace techniques of i-vector/PLDA (probabilistic linear discriminant analysis) and joint factor analysis (JFA) which have been the most commonly used techniques in the field of text-dependent speaker verification. These techniques, however, do not model the temporal structure of the pass phrase, which is an important cue in the case of text-dependent speaker verification. The authors propose integrating deep neural network (DNN)-HMM technique with the HiLAM method where the state alignment information from the HiLAM is used to discriminatively train a DNN to further improve the system performance. This allows the neural network to learn the actual context of the pass phrase, which is not the case with the DNN trained for automatic speech recognition (ASR). Besides, the network also models the speaker idiosyncrasies with specific text units. The use of the DNN posteriors to replace the GMM likelihood probabilities of HiLAM has led to significant improvement in performance over the baseline HiLAM system.

The eighth and ninth papers cover the theme of quality and intelligibility of speech.

Improving the speech intelligibility remains a challenging problem in digital hearing aids. The eighth paper authored by S. Shoba and R. Rajavel deals with improving speech intelligibility in monaural segregation system by fusing voiced and unvoiced speech segments using the genetic algorithm. The voiced speech segments are obtained using perceptual speech cues such as auto-correlation, cross-channel correlation, and pitch. Similarly, the unvoiced speech segments are obtained using another perceptual speech cue onset/offset after subtracting the voiced segments. The speech onset- and offset-based segregation process actually produce segments for both voiced and unvoiced. The unvoiced speech segments are obtained by subtracting the voiced speech segments from the segments obtained using speech onset and offset. The unvoiced segments obtained using onset and offset may contain interference. This paper proposes a scheme to remove those interferences from the unvoiced speech segments and effectively fuse the segments of voiced and unvoiced speech using a genetic algorithm. The performance of the proposed algorithm is evaluated using various intelligibility measures. The experimental results show that the proposed algorithm significantly improves the speech intelligibility compared with other existing systems.

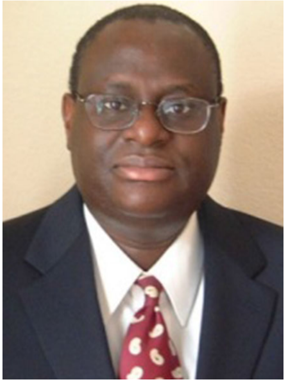
The ninth paper authored by Asutosh Kar et al. discusses ways to improve sound quality for hearing aids in the presence of multiple inputs. This paper is a biomedical application aimed at improving the performance of hearing aids if the wireless input and the input from the acoustic environment are not highly correlated. This relatively low correlation causes much confusion to the user. The feed-forward path and the shaping filter for the wireless signal are optimized to enhance the quality of the overall output signal.

To cover the final theme of compressive sensing, we have the tenth paper titled "Recurrent Neural Network Based Dictionary Learning for Compressive Speech Sensing." The authors Yunyun Ji et al. propose a novel dictionary learning technique for compressive sensing of speech signals based on the recurrent neural network. First, they utilize the recurrent neural network for estimating the linear prediction coding (LPC) coefficients for voiced and unvoiced speech, respectively. Then, the extracted LPC coefficient vectors are clustered through an improved Linde–Buzo–Gray (LBG) algorithm to generate codebooks for voiced and unvoiced speech, respectively. A dictionary is then constructed for each type of speech by concatenating a union of structured matrices derived from the column vectors in the corresponding codebook. Next, a decision module is designed to determine the appropriate dictionary for the recovery algorithm in the compressive sensing system. Finally, based on the sequential linear prediction model and the proposed dictionary, a sequential recovery algorithm is proposed to further improve the quality of the reconstructed speech. Experimental results show that when compared to the selected state-of-the-art approaches, the proposed method can achieve superior performance in terms of several objective measures including segmental signal-to-noise ratio, perceptual evaluation of speech quality, and short-time objective intelligibility under both noise-free and noise-aware conditions.

We thank all the authors, the reviewers, the CSSP journal administrative staff, and the CSSP Editor-in-Chief for all their contributions to making this Special Section possible.

We hope you enjoy reading the articles!!!

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Tokunbo Ogunfunmi** received the B.S. (with first class honors) degree from the University of Ife, Ile-Ife, Nigeria, and the M.S. and Ph.D. degrees from Stanford University, Stanford, California, all in Electrical Engineering. He is currently a Professor of Electrical Engineering and Director of the Signal Processing Research Laboratory at Santa Clara University (SCU), Santa Clara, California. From 2010 to 2014, he served as the Associate Dean for Research and Faculty Development for the SCU School of Engineering. At SCU, he teaches a variety of courses in circuits, systems, signal processing and related areas. His current research interests include machine learning, deep learning, speech and multimedia (audio, video) compression, digital and adaptive signal processing and applications, and nonlinear signal processing. He has published over 170 refereed journal and conference papers in these areas. Dr. Ogunfunmi currently serves as an Associate Editor of the IEEE *Transactions on Signal Processing* and the *Circuits, Systems and Signal Processing* journal. He also serves as Lead Guest Editor for the *Journal of Signal Processing Systems (JSPS)* (2019 Special Issue on SiPS 2018). He has been involved with many IEEE conference committees as a member of the organizing and technical committees. He served as the General Chair of the 2018 IEEE Workshop on Signal Processing Systems (SiPS 2018) and as the Technical Program Co-Chair of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS 2019). Dr. Ogunfunmi served the IEEE as a Distinguished Lecturer from 2013 to 2014 for the Circuits and Systems Society.

**Ravi P. Ramachandran (SM'08)** received the B.E. degree (with great distinction) from Concordia University in 1984, the M. Eng. degree from McGill University in 1986 and the Ph.D. degree from McGill University in 1990. From October 1990 to December 1992, he worked at the Speech Research Department at AT&T Bell Laboratories. From January 1993 to August 1997, he was a Research Assistant Professor at Rutgers University. He was also a Senior Speech Scientist at T-Netix from July 1996 to August 1997. Since September 1997, he is with the Department of Electrical and Computer Engineering at Rowan University where he has been a Professor since September 2006. He has served as a consultant to T-Netix, Avenir Inc., Motorola, and FocalCool. From September 2002 to September 2005, he was an Associate Editor for the IEEE Transactions on Speech and Audio Processing and was on the Speech Technical Committee for the IEEE Signal Processing Society. From September 2000 to December 2015, he was on the Editorial Board of the IEEE Circuits and Systems Magazine. Since May 2002, he has been on the Digital Signal Processing Technical Committee for the IEEE Circuits and Systems Society. Since May 2012, he has been on the Education and Outreach Technical Committee for the IEEE Circuits and Systems Society. His research interests are in digital signal processing, speech processing, biometrics, pattern recognition, machine learning, and filter design.

**Roberto Togneri** received the B.E. degree in 1985, and the Ph.D. degree in 1989 both from the University of Western Australia. He joined the School of Electrical, Electronic and Computer Engineering at The University of Western Australia in 1988, where he is now currently an Associate Professor. Professor Togneri leads the Signal Processing and Recognition Lab, and his research activities in signal processing and pattern recognition include feature extraction and enhancement of audio signals, statistical and neural network models for speech and speaker recognition, audiovisual recognition and biometrics, and related aspects of language modeling and understanding. He has published over 150 refereed journal and conference papers in the areas of signal processing and recognition, is the chief investigator on three Australian Research Council Discovery Project research grants since 2009, and is currently the Area Editor for IEEE Signal Processing Magazine Columns and Forums. He is a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE), a Member of the Australian Speech Science and Technology Association (ASSTA), and a Member of the International Speech Communication Association (ISCA).

**Brett Smolenski** has over 25 years of experience in research, development, and design of signal processing and electronic systems. He is currently a senior researcher at North Point Defense and lecturer at the State University of New York Polytechnic Institute. He received a Ph.D. in Electrical and Computer Engineering from Temple University, a Bachelor's Degree in Electrical Engineering from Pennsylvania State University, and a Master's Degree in Mathematics from West Chester University. His current research interests include low-noise and mixed-signal circuit design, array and nonlinear signal processing, and unsupervised learning. He was awarded a patent for his work in enabling speaker identification systems to be robust to overlapping speakers and has published over 35 scientific articles. He was inducted into the Eta Kappa Nu honor society and is currently a member of the IEEE and the Acoustical Society of America.

**Visar Berisha** received the Ph.D. degree from Arizona State University, Tempe, USA, in 2007. From 2007 to 2009, he was a member of the technical staff at the Massachusetts Institute of Technology—Lincoln Laboratory, Cambridge, USA. Following his appointment at Lincoln Labs, he was Principal Engineer at Raytheon Co. He is now an Associate Professor at Arizona State University in the School of Electrical Computer and Energy Engineering with a joint appointment in the Department of Speech and Hearing Sciences. His research interests include statistical signal processing, machine learning, and computational models of speech and audio perception.