Introduction to the Special Issue on End-to-End Speech and Language Processing

SPEECH and Language processing (SLP) is essentially a series of sequence-to-sequence learning problems. Conventional SLP systems map input to output sequences through module-based architectures where each module is independently trained. These have a number of limitations including local optima, assumptions about intermediate models and features, and complex expert knowledge driven steps. It can be difficult for non-experts to use and develop new applications. Integrated End-to-End (E2E) systems aim to simplify the solution to these problems through a single network architecture to map an input sequence directly to the desired output sequence without the need for intermediate module representations. E2E models rely on flexible and powerful machine learning models such as recurrent neural networks. The emergence of models for end-to-end speech processing has lowered the barriers to entry into serious speech research. This special issue showcases the power of novel machine learning methods in end-to-end speech and language processing.

Following an open call for papers, we received a total of 41 submissions for this special issue, spanning a range of topics, including automatic speech recognition (ASR) and multimodal emotion recognition, information retrieval systems such as spoken keyword search, and text processing applications such as machine translation. After an extensive and competitive review process, we selected 11 papers for final publication.

Several papers of this special issue are devoted to ASR. There are two major types of E2E architectures for ASR; attention-based methods that use an attention mechanism to perform alignment between acoustic frames and recognized symbols, and connectionist temporal classification (CTC) that uses Markov assumptions to efficiently solve sequential problems by dynamic programming. S. Watanabe *et al.* propose a hybrid CTC/attention end-to-end ASR, which effectively utilizes the advantages of both architectures for training and decoding. The hybrid approach allows the use of the forward–backward algorithm in CTC to enforce monotonic alignments and speed up the process of estimating the desired alignment, instead of solely depending on data-driven attention methods to estimate the desired alignments in long sequences.

While E2E approaches considerably simplify the build process for ASR systems, several challenges remain, including computational complexity, dealing with large sequences, and real-time performance. H. Tang *et al.* provide a combined review of neural segmental models from several prominent research groups. These models can be viewed as consisting of a neural network-based acoustic encoder and a finite-state transducer decoder. This study analyzes several weight functions that map the input acoustic features to the label sequence and loss functions that combine segmental and frame-level information for E2E training. The study draws connections between the various types of models, setting the stage for future research in ASR to benefit from its findings.

P. Doetsch *et al.* propose a novel, fully stochastic, inverted HMM architecture as an alternative to the CTC/attention-based E2E systems. This integrated discriminative model can be trained end-to-end from scratch by inversely aligning each element of an HMM state sequence to a segment-wise encoding of several consecutive input frames. In the inverted HMM, the HMM concept is generalized to drop the conditional independence assumption on the frame or label level while keeping the alignment as part of the underlying stochastic model.

While all these approaches on E2E frameworks focus on the ASR problem in a single-channel setup without speech enhancement, T. Ochiai *et al.* propose a unified architecture to incorporate a multi-channel setup with speech enhancement within an E2E framework. The proposed approach is a single, fully-differentiable, E2E neural network with a neural beamformer that can be optimized with a combined CTC and attention-based ASR objective. A similar integrated approach to jointly learn the front-end signal processing and back-end acoustic modeling of an ASR system is proposed by B. Wu *et al.* A joint speech dereverberation for signal enhancement and a deep neural network based acoustic modeling architecture is presented.

An end-to-end audio-visual emotion recognition system is proposed by P. Tzirakis *et al.* Instead of using hand-crafted features, a convolutional neural network (CNN) is used to extract emotional features from speech and a deep residual network is used for vision. A long short-term memory (LSTM) neural network is then used to combine the automatically extracted audio and visual feature streams.

This Special Issue presents several articles on keyword search (KWS), which investigate the relationship between learning and different components of KWS systems. T. Fuchs and J. Keshet introduce a calibrated loss function that maximizes performance at a specific decision threshold, rather than simply trying to maximize the margin between positive and negative examples. They apply it to both classical phonetic queries and query-by-example (QbyE). B. Gündoğdu *et al.* jointly learn a trainable distortion measure for sub-sequence dynamic time warping and a model to map phone sequences with durations to pseudo-posteriorgrams, so that techniques from QbyE search can be used to find OOV

Digital Object Identifier 10.1109/JSTSP.2017.2767938

queries in a system where queries are presented in textual format. In this case, both a loss function and a query representation are learned.

H. Chen *et al.* focus on the problem of keyword search in a low-resource setting where no transcribed speech is available in the target language. They propose using a multi-task learning (MTL) approach to train a bottleneck neural network feature extractor. The MTL network is trained to map features from the target language to phoneme-like units induced using a Dirichlet process Gaussian mixture model, and to map features from an auxiliary, high-resource language for which transcripts are available to HMM states.

The final KWS paper, by K. Audhkhasi *et al.* proposes an ASR-free KWS system built from three neural models that are trained jointly: an RNN acoustic encoder that maps variable-length feature sequences to a fixed-length embedding, a CNN-RNN encoder that maps the text query to a fixed-length embedding, and a feedforward network that predicts the presence or absence of the query in the audio.

E2E neural machine translation has become state-of-the-art for language pairs with large amounts of parallel data. The paper by C. España-Bonet *et al.* focuses on multilingual translation over a set of languages (i.e., Arabic, English, French, German, and Spanish). They investigate the interlingual nature of the context vectors generated by their multilingual neural machine translation system and study their power in the assessment of monolingual and cross-language similarity. Four research questions are addressed: whether embeddings learned for a source text also depends on the target language, how distinguishable representations of semantically similar and semantically-distant sentence pairs are, how close representations of sentence pairs within and across languages are, and how representations evolve throughout the training.

Finally, we would like to thank all the authors and reviewers whose contributions have made this special issue possible. We would like to thank Prof. Shri Narayanan, Editor-in-Chief, for his support, and encouragement. Warmest thanks also to Allison Fleisher from the IEEE publication office for keeping the issue on track.

BHUVANA RAMABHADRAN, *Lead Guest Editor*
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598 USA

NANCY F. CHEN, *Guest Editor*
Institute for Infocomm Research
Agency for Science, Technology and Research
Singapore 138632

MARY P. HARPER, *Guest Editor*
Army Research Laboratory
Adelphi, MD 20783 USA

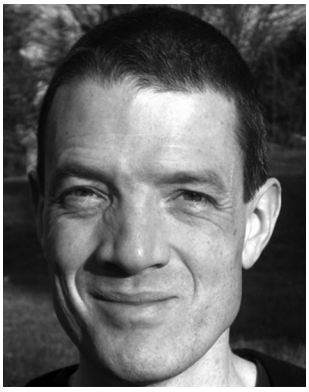
BRIAN KINGSBURY, *Guest Editor*
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598 USA

KATE KNILL, *Guest Editor*
University of Cambridge
Cambridge CB2 1PZ, U.K.

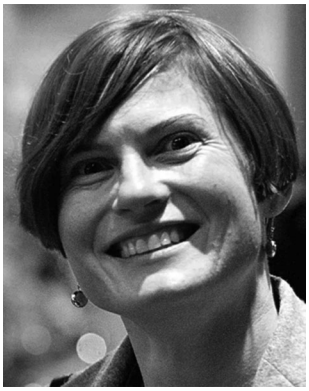
Bhuvana Ramabhadran (F'17) is a Distinguished Research Staff Member and Manager in IBM Research AI, Yorktown Heights, NY, USA. She leads a team of researchers in the Speech Technologies Group and coordinates activities across IBM's world-wide laboratories in the areas of speech recognition, synthesis, and spoken term detection. She was the elected Chair of the IEEE SLTC (2014–2016), Area Chair for ICASSP (2011–2017) and Interspeech (2012–2016), was on the editorial board of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2011–2015), and is currently an ISCA board member. She is a Fellow of ISCA.

Nancy F. Chen (SM'15) received the Ph.D. degree from Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, and Harvard, Cambridge, MA, USA, in 2011, based on her research at MIT Lincoln Laboratory in multilingual speech processing. She is currently leading initiatives in deep learning and human language technology at the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. She is an elected member of the IEEE SLTC (2016–2018) and has received multiple awards, including the Best Paper at APSIPA ASC (2016), the Microsoft-sponsored IEEE Spoken Language Processing Grant (2011), and the NIH Ruth L. Kirschstein National Research Award (2004–2008).

Mary P. Harper (SM'02) received the Ph.D. degree in computer science from Brown University, Providence, RI, USA. She is Deputy Chief Scientist at Army Research Laboratory, Adelphi, MD, USA. She has also served as a Program Manager for the Babel Program at IARPA. She has served as a Guest Member of the IEEE SLTC (2007–2010). She has served on the editorial board of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2005–2009) and on Flanagan Speech and Audio Processing Award Committee (2014–2016). She is a Fellow of ISCA.



Brian Kingsbury (SM'09) received the B.S. degree in electrical engineering from Michigan State University, Lansing, MI, USA, and the Ph.D. degree in computer science from the University of California, Berkeley, CA, USA. He is a Principal Research Staff Member in IBM Research AI, Yorktown Heights, NY, USA. He has served on the IEEE SLTC (2009–2011). He was an ICASSP Speech Area Chair (2010–2012), an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2012–2016), and a Program Chair for the International Conference on Representation Learning (2014–2016).



Kate Knill (M'01) received the Ph.D. degree in digital signal processing from Imperial College London, London, U.K. She is a Senior Research Associate in the Department of Engineering, Cambridge University, Cambridge, U.K. She has more than 25 years' experience in speech and language processing in industry and academia, including the establishing and leading of Speech Technology Group, Toshiba Cambridge Research Laboratory, Cambridge, U.K. (2002–2012). She was a member of the IEEE SLTC (2009–2012), is an ISCA Board member (2013–2021), and is currently the Secretary of ISCA.