Harnessing the Power of Artificial Intelligence in Otolaryngology & the Communication Sciences

BLAKE S. WILSON,[1,2,3,4,5] DEBARA L. TUCCI,[1,6] DAVID A. MOSES,[7,8] EDWARD F. CHANG,[7,8] NANCY M. YOUNG,[9,10,11] FAN-GANG ZENG,[12,13,14,15,16] NICHOLAS A. LESICA,[17] ANDRÉS M. BUR,[18] HANNAH KAVOOKJIAN,[18] CAROLINE MUSSATTO,[18] JOSEPH PENN,[18] SARA GOODWIN,[18] SHANNON KRAFT,[18] GUANGHUI WANG,[19] JONATHAN M. COHEN,[1,20] GEOFFREY S. GINSBURG,[4,21,22,23,24,25] GERALDINE DAWSON,[26,27,28] AND HOWARD W. FRANCIS[1]

[1]*Department of Head and Neck Surgery & Communication Sciences, Duke University School of Medicine, Durham, NC 27710, USA*

[2]*Duke Hearing Center, Duke University School of Medicine, Durham, NC 27710, USA*

[3]*Department of Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA*

[4]*Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA*

[5]*Department of Otolaryngology – Head & Neck Surgery, University of North Carolina, Chapel Hill, NC 27599, USA*

[6]*National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Bethesda, MD 20892, USA*

[7]*Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA 94143, USA*

[8]*UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94117, USA*

[9]*Division of Otolaryngology, Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, IL 60611, USA*

[10]*Department of Otolaryngology - Head and Neck Surgery, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA*

[11]*Knowles Hearing Center, Department of Communication, Northwestern University, Evanston, IL 60208, USA*

[12]*Center for Hearing Research, University of California, Irvine, Irvine, CA 92697, USA*

[13]*Department of Anatomy and Neurobiology, University of California, Irvine, Irvine, CA 92697, USA*

[14]*Department of Biomedical Engineering, University of California, Irvine, Irvine, CA 92697, USA*

[15]*Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92697, USA*

[16]*Department of Otolaryngology – Head and Neck Surgery, University of California, Irvine, Irvine, CA 92697, USA*

[17]*UCL Ear Institute, University College London, London WC1X 8EE, UK*

[18]*Department of Otolaryngology - Head and Neck Surgery, University of Kansas Medical Center, Kansas City, KS 66160, USA*

[19]*Department of Computer Science, Ryerson University, Toronto, ON M5B 2K3, Canada*

[20]*ENT Department, Kaplan Medical Center, Rehovot 7661041, Israel*

[21]*MEDx (Medicine & Engineering at Duke), Duke University, Durham, NC 27708, USA*

[22]*Center for Applied Genomics & Precision Medicine, Duke University School of Medicine, Durham, NC 27710, USA*

[23]*Department of Medicine, Duke University School of Medicine, Durham, NC 27710, USA*

[24]*Department of Pathology, Duke University School of Medicine, Durham, NC 27710, USA*

[25]*Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA*

[26]*Duke Institute for Brain Sciences, Duke University, Durham, NC 27710, USA*

[27]*Duke Center for Autism and Brain Development, an NIH Autism Center of Excellence, Duke University School of Medicine and the Duke Institute for Brain Sciences, Durham, NC 27705, USA*

[28]*Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, NC 27701, USA*

Word and number counts

Introduction: 829

Discussion: 1950

Number of tables: none

Number of figures: 12

Correspondence to: Blake Wilson · 2410 Wrightwood Ave. · Durham, NC 27705 · USA. email: blake.wilson@duke.edu

Email addresses for the other authors:

Debara L. Tucci: debara.tucci@duke.edu and debara.tucci@nih.gov

David A. Moses: David.Moses@ucsf.edu

Edward F. Chang: Edward.Chang@ucsf.edu

Nancy M. Young: NYoung@luriechildrens.org

Fan-Gang Zeng: fzeng@hs.uci.edu

Nicholas A. Lesica: n.lesica@ucl.ac.uk

Andrés M. Bur: abur@kumc.edu

Hannah Kavookjian: hkavook1@jhu.edu

Caroline Mussatto: Cmussatto2@kumc.edu

Joseph Penn: josephpenn601@gmail.com

Sara Goodwin: Sara.goodwin8@gmail.com

Shannon Kraft: skraft3@kumc.edu

Guanghui Wang: wangcs@ryerson.ca

Jonathan M. Cohen: yonicohen2@gmail.com and JONATHANCO@clalit.org.il

Geoffrey S. Ginsburg: geoffrey.ginsburg@duke.edu

Geraldine Dawson: geraldine.dawson@duke.edu

Howard W. Francis: howard.francis@duke.edu

ABSTRACT

Use of artificial intelligence (AI) is a burgeoning field in otolaryngology and the communication sciences. A virtual symposium on the topic was convened from Duke University on 26 October 2020 and was attended by more than 170 participants worldwide. This review presents summaries of all but one of the talks presented during the symposium; recordings of all the talks, along with the discussions for the talks, are available at https://www.youtube.com/watch?v=ktfewrXvEFg and https://www.youtube.com/watch?v=-gQ5qX2v3rg. Each of the summaries is about 2500 words in length and each summary includes two figures. This level of detail far exceeds the brief summaries presented in traditional reviews and thus provides a more-informed glimpse into the power and diversity of current AI applications in otolaryngology and the communication sciences and how to harness that power for future applications.

Keywords: otolaryngology, machine learning, artificial intelligence, deep learning, human communication, hearing, speech production, speech perception, auditory prostheses, auditory system, hearing aids, hearing loss, cochlear implants, neural prostheses, neuroprostheses, brain-computer interfaces, laryngeal pathology, thyroid pathology

INTRODUCTION

The Duke Institute for Brain Sciences (DIBS), the Duke Medical School's Department of Head and Neck Surgery & Communication Sciences (HNSCS), and the Duke MEDx Program (a program to foster collaborations between medicine and engineering) convened a virtual symposium on 26 October 2020 to highlight and suggest applications of artificial intelligence (AI) in otolaryngology & the communication sciences (Fig. 1). The symposium had approximately 360 registrants and more than 170 attendees. Howard Francis, Chair of the Department of HNSCS, and Geraldine Dawson, Director of the DIBS, introduced the symposium and the speakers included Edward F. Chang, Professor and Chair of Neurological Surgery at the University of California, San Francisco; Nancy M. Young, Lillian S. Wells Professor of Pediatric Otolaryngology at the Northwestern University Feinberg School of Medicine; Fan-Gang Zeng, Professor of Otolaryngology, Anatomy & Neurobiology, Biomedical Engineering, and Cognitive Sciences at the University of California, Irvine; Roger L. Miller, Director of Neuroprosthesis Research at the National Institute on Deafness and Other Communication Disorders (NIDCD); Andrés M. Bur, Assistant Professor in the Department of Otolaryngology – Head and Neck Surgery at the University of Kansas Medical Center and now an Associate Professor there; and Jonathan M. (Yoni) Cohen, Adjunct Associate in Duke's Department of HNSCS and now also a head and neck surgeon at the Kaplan Medical Center in Rehovot, Israel. Blake S. Wilson, Director of the Duke Hearing Center and Adjunct or Consulting Professor in Duke's Departments of HNSCS, Biomedical Engineering, and Electrical & Computer Engineering, served as the Chair for the symposium, and Debara L. Tucci, Director of the NIDCD and Adjunct Professor in Duke's Department of HNSCS, served as the Co-Chair for the symposium, in her capacity as an Adjunct Professor at Duke. Wilson and Tucci each offered some concluding remarks at the end of the symposium, and Wilson and Geoffrey S. Ginsburg, Professor of Medicine at Duke and Director of the MEDx Program, served as the moderators for the two sessions within the symposium. The symposium was highly interactive and clearly indicated the power of AI (and particularly the deep learning branch of AI) in current applications in otolaryngology & the communication sciences and additionally indicated the high potential and best approaches – along with cautions and lessons learned – for further applications in those fields. The symposium was beautifully organized by Tyler Lee, Katherine Neal, and Dr. Nicole Schramm-Sapyta of the DIBS, and video recordings of the symposium are available at https://www.youtube.com/watch?v=ktfewrXvEFg and https://www.youtube.com/watch?v=-gQ5qX2v3rg; the latter recording is a recording of Dr. Cohen's complete talk with slides, which could not be included in the first recording due to a technical problem. Additionally, brief bios of the speakers are presented at https://dibs.duke.edu/file/harnessing-ai-speaker-bios.

An image of the full program for the symposium is presented in Fig. 2. The purpose of this present Review is to convey the essence of the symposium and to provide detailed examples of recent and significant advances in the fields of the symposium. In contrast to a traditional review, the present Review indicates how the advances were achieved, including advice on how to avoid pitfalls, and thereby provides models for future advances.

The examples are from the talks or based on the talks, and each of the presented examples is about 2500 words in length and includes two figures. Sections for all talks except the one by Dr. Miller are included in this Review; the talk by Dr. Miller was on NIDCD support to advance machine learning in otolaryngology & the communication sciences and not on an application of AI *per se*. Additionally, Dr. Miller cannot be among the authors of a paper in which any author other than him is a recipient or potential recipient of a grant or other support from the NIDCD. However, Dr. Miller provided wonderful guidance in his splendid talk and the talk is included in the first recording mentioned previously.

The sections for the talks present the examples but not much if any basic information about AI and machine learning. That latter information can be found in excellent overviews (e.g., LeCun et al. 2015; Bur et al. 2019; Hogarty et al. 2019; Sarker 2021) and in online courses such as the outstanding (and comprehensive) ones offered by Duke (https://www.coursera.org/learn/machine-learning-duke?utm_campaign=website--courses&utm_medium=institutions&utm_source=duke), Google (https://www.youtube.com/watch?v=lYWt-aCnE2U), the Massachusetts Institute of Technology (https://openlearninglibrary.mit.edu/courses/course-v1:MITx+6.036+1T2019/about, https://www.youtube.com/watch?v=h0e2HAPTGF4, https://www.youtube.com/watch?v=O5xeyoRL95U, and https://deeplearning.mit.edu/), Stanford (https://www.youtube.com/playlist?list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN), Simplilearn (https://www.youtube.com/watch?v=9f-GarcDY58 and https://www.youtube.com/watch?v=8Pyy2d3SZuM), and 3Blue1Brown (https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi).

Additionally, brief but informative (and sometimes humorous) descriptions are given at https://www.youtube.com/watch?v=2ePf9rue1Ao ("What is AI – for people in a hurry!"), https://www.youtube.com/watch?v=6M5VXKLf4D4 ("Deep learning in 5 minutes"), https://www.youtube.com/watch?v=kWmX3pd1f10 ("What is artificial intelligence exactly?"), https://www.youtube.com/watch?v=ukzFI9rgwfU ("Machine learning basics"), https://www.youtube.com/watch?v=ad79nYk2keg ("Artificial intelligence in 5 minutes"), https://www.youtube.com/watch?v=bfmFfD2RIcg ("Neural networks in 5 minutes"), https://www.youtube.com/watch?v=Ec7Wu2JMvPw ("What's the deal with AI in healthcare?"), and https://www.youtube.com/watch?v=gzGs1ke8idA ("The language of AI and how to speak it"). (All of the links in this and the preceding paragraph were accessed on 14 December 2021.)

Following the sections for the talks, we conclude this Review with remarks about the demonstrated and anticipated power of AI in otolaryngology & the communication sciences.

TOWARD A DIRECT-SPEECH NEUROPROSTHESIS: DECODING SPEECH FROM SENSORIMOTOR CORTEX USING ARTIFICIAL INTELLIGENCE (Authors DAM and EFC)

Speech is a unique and defining feature of human interaction. Although speech typically feels effortless, the mechanisms underlying speech production involve rapid, precise, and highly coordinated movements of many individual muscles to control the lips, jaw, tongue, and larynx (Fowler et al. 1980; Browman and Goldstein 1992). It is through this complex system that the brain orchestrates the translation of intent to speak into audible speech output.

In the past decade, researchers have characterized various aspects of speech-related neural activation patterns in human sensorimotor cortex, a brain area that has been heavily implicated in speech processing (Bouchard et al. 2013; Chakrabarti et al. 2015; Lotte et al. 2015; Carey et al. 2017; Chartier et al. 2018; Conant et al. 2018; Dichter et al. 2018). By leveraging advances from the fields of automatic speech recognition, which focuses on translating acoustic speech waveforms into text, and artificial intelligence (AI), researchers also showed that spoken speech can be decoded (both as text and as synthesized speech waveforms) directly from cortical activity in this brain area (Fig. 3) (Mugler et al. 2014; Herff et al. 2015; Angrick et al. 2019; Anumanchipalli et al. 2019; Moses et al. 2019; Makin et al. 2020; Sun et al. 2020). Informed by these insights and techniques, we developed a prototype of a direct-speech neuroprosthesis that used AI to decode words and sentences from the brain activity of a person with severe paralysis in real time as the person attempted to speak (Moses et al. 2021). Overall, advances in speech neuroscience and AI have the potential to enable restoration of speech to individuals with severe paralysis who are unable to communicate naturally.

Speech-Motor Encoding in the Ventral Sensorimotor Cortex

Efforts to understand the role of the human sensorimotor cortex during speech production have been ongoing for over a century, with major contributions from Dr. Pierre Paul Broca in the mid-1800s and Dr. Harvey Cushing in the early 1900s (Broca 1861; Cushing 1909; Pendleton et al. 2012). Dr. Cushing proposed a somatotopic organization of this brain area, attributing control of the various articulators (lips, jaw, tongue, etc.) to adjacent regions of the motor cortex along a dorsal-ventral axis. This proposed organization was further solidified by the somatosensory homunculus model described by Dr. Wilder Penfield in the mid-1900s (Penfield and Boldrey 1937; Penfield and Rasmussen 1950), which was largely based on observed relationships between electrical stimulation of specific locations of the sensorimotor cortex of surgical patients and the induced sensations reported by the patients.

Recent work has supported and expanded these findings. Results from a study using electrocorticography (ECoG) to record cortical activity during a syllable articulation task showed that somatotopically arranged speech-articulator representations are present in the ventral sensorimotor cortex (vSMC) (Bouchard et al. 2013).

Additionally, these results demonstrated that spatially distributed cortical activation patterns recorded as participants articulated various syllables were functionally organized by phonetic features such as constriction location (which is also referred to as "place of articulation"). Organization based on the constriction location was also found using functional magnetic resonance imaging while participants produced speech phones (Carey et al. 2017). As described in a separate study using ECoG recordings as participants spoke entire sentences and paragraphs, this functional organization seems to persist not just during execution of speech production, but also during speech preparation and monitoring of acoustic feedback (Lotte et al. 2015).

These findings implicate the vSMC in speech-motor processing and describe the somatotopic organization of speech in this brain area, but it was still not clear how this area orchestrates physical articulator movements. Subsequent research showed that ECoG signals recorded from the vSMC were directly correlated with kinematic articulator movements, more so than with produced acoustics or categorical phonetic feature descriptions (Conant et al. 2018). Separately, cortical activity in the dorsal aspect of the vSMC has been correlated with vocal pitch control during speech and non-speech vocalizations (Dichter et al. 2018). During continuous production of natural sentences, cortical activity in the vSMC encodes complex articulator kinematics for a variety of vocal-tract shapes, including representations of coarticulation that enable the production of fluid speech (Chartier et al. 2018). In general, although neural activation patterns in the vSMC are correlated with categorical phonetic features of produced speech, these patterns most robustly encode rapid, coordinated, and complex kinematic articulatory movements to directly control the vocal tract and produce natural, continuous speech.

Synthesizing Produced Speech from Brain Activity

To fully understand how neural activation patterns eventually give rise to fluent speech production, it is important to not only understand the relationship between the brain activity and the articulatory kinematics but also the relationship between articulatory kinematics and the produced speech waveforms. Efforts to mechanically "synthesize" speech (generate acoustic waveforms of speech) started as early as circa 1770 with Wolfgang von Kempelen's "Speaking machine," a simplified mechanical model that emulated the vocal tract (von Kempelen 1791; Dudley and Tarnoczy 1950). Homer Dudley and colleagues later pioneered the "Voder" (which is short for "Voice Operating Demonstrator"), an electrical device controlled by keys and levers to synthesize speech sounds (Dudley et al. 1939; Gold et al. 2011). These physical models used mechanical or electrical components to mimic the various articulators (lips, tongue, etc.) and their movements and positions during speech production.

Eventually, software-based speech-synthesis approaches were developed, such as the formant-based synthesizer created by Dennis Klatt (Klatt and Klatt 1990). More recently, researchers and engineers have developed "text-to-speech" (TTS) approaches involving the synthesis of speech directly from text, with some approaches using AI methods to enable generation of naturalistic speech waveforms (Dutoit 1997; Ze et al. 2013; Oord et al. 2016; Wang et al. 2017; Shen et al. 2018; Ning et al. 2019). It has also been shown that articulatory

features can be incorporated to improve TTS performance (Ling et al. 2009). These articulatory features can be measured directly using a recording approach such as electromagnetic articulography (EMA) (Schönle et al. 1987) or inferred from acoustic signals using a technique called acoustic-to-articulatory inversion (AAI) (Richmond 2002; Ghosh and Narayanan 2011; Mitra et al. 2017; Chartier et al. 2018). To characterize the relationship between cortical activity in the vSMC and articulatory kinematics, researchers used an AAI approach with a deep recurrent neural network to infer EMA features from the acoustic signals (Chartier et al. 2018).

Together, these findings and approaches facilitated recent efforts to decode articulator movements and reconstruct speech waveforms directly from brain activity (Angrick et al. 2019; Anumanchipalli et al. 2019; Salari et al. 2019; Angrick et al. 2021). In one study, researchers used recurrent neural networks to translate cortical activity into articulatory kinematic representations and then translate those intermediate representations into speech acoustics, resulting in intelligible synthesized speech (Anumanchipalli et al. 2019). These researchers showed that using articulatory kinematic representations as intermediate features improved the quality of synthesized speech compared to a model that tried to directly map neural activity to acoustic output. Even with advanced AI models, this finding suggests that using physiologically motivated representations during modeling of neural activity can lead to increased performance when the amount of available data is limited (which is often the case in studies involving invasive brain recordings). The researchers also showed that it was possible to synthesize speech from brain activity while a participant mimed sentences (i.e., the participant made articulatory movements as if she or he was saying the sentences but without overtly vocalizing them). Although speech synthesized during miming was less intelligible than during overt attempts, this demonstration still illustrates that important features of speech (e.g., spectral and temporal features) can be decoded from cortical signals underlying speech-motor processing even if the features are not audibly uttered. In a separate study, researchers developed a prototype to synthesize imagined speech in real time from stereotactic depth electrodes implanted in an able speaker, although the synthesized speech was not intelligible (Angrick et al. 2021). Overall, the various findings reviewed in this subsection suggested that further development of approaches that use AI techniques to synthesize speech directly from brain activity could lead to direct-speech neuroprostheses capable of restoring spoken communication for persons with severe paralysis.

Decoding Text from Brain Activity

Although evidence suggests that speech-related neural activity in the sensorimotor cortex is more strongly correlated with continuous articulator kinematics than categorical speech features, there are some advantages to decoding textual representations of speech (such as phonemes or words) from brain activity. One such advantage is the ability to leverage advances in speech recognition (also referred to as automatic speech recognition, or ASR) technologies and methodologies during text decoding. A primary focus of speech-recognition research is to enable the translation of speech acoustics into text (Jelinek 1976; Gold et al. 2011). Many traditional (earlier)

ASR systems use hidden Markov models (HMMs) that combine probabilities from a phoneme "emission" model (which yields phoneme likelihoods given acoustic feature vectors) and a natural-language "transition" model (which yields phoneme- and word-sequence likelihoods) (Bourlard and Morgan 1994; Chen and Goodman 1999; Benzeghiba et al. 2007; Gold et al. 2011; Mohamed et al. 2012). More recently, advanced AI methods using deep learning have enabled "end-to-end" decoding approaches with multi-layer artificial neural networks (ANNs) that are capable of translating acoustic waveforms directly into text without requiring an HMM or a separate language model, achieving superior performance over previous methods (Graves et al. 2013; Collobert et al. 2016; Oord et al. 2016; Kim et al. 2017; Zhang et al. 2017; Wang et al. 2019). Both types of ASR systems can inform the design of "neural speech recognition" (NSR) pipelines that decode text from brain activity; many AI model architectures and language modeling techniques are effective at decoding text from features containing speech information whether those features are acoustic or neural.

Informed by earlier descriptions of discriminable phonetic-feature encoding in sensorimotor cortex activity (Bouchard et al. 2013; Mugler et al. 2014), researchers used HMM-based ASR pipelines to decode textual representations of spoken speech from brain activity (Herff and Schultz 2016), replacing acoustic-to-phoneme emission models with neural-to-phoneme emission models. Results from one study showed that ECoG signals recorded as participants said various phrases could be translated into text using Gaussian phone models and a natural-language model (Herff et al. 2015). Separately, in our previous work, we showed that linear phone models and HMMs parameterized by phonetic sequences could be used to decode words and phrases that participants heard and said from ECoG signals in real time (Moses et al. 2019).

Although these HMM-based NSR systems illustrate the feasibility of decoding text from brain activity, it was hypothesized that performance could be improved by leveraging modern AI techniques such as deep learning. This hypothesis was informed not only by the fact that ASR systems using deep learning have achieved state-of-the-art performance on acoustic recognition tasks (Wang et al. 2019), but also by the validation of deep learning in various speech neuroscience studies (Berezutskaya et al. 2017; Rezazadeh Sereshkeh et al. 2017; Angrick et al. 2019; Anumanchipalli et al. 2019; Livezey et al. 2019). Results from two recent studies confirmed this hypothesis and outperformed previous NSR approaches, with both studies using deep end-to-end networks with ECoG recordings to decode text as participants verbally produced sentences (Makin et al. 2020; Sun et al. 2020).

In the first of these two studies, researchers used a deep encoder-decoder ANN model to directly decode word sequences (Makin et al. 2020). This ANN first performs temporal convolution to downsample the filtered ECoG input signals. Temporal convolution is a technique validated in ASR applications for extracting robust features from time-series data (Zhang et al. 2017). Then, for each sentence spoken by a participant, the encoder recurrent neural network (RNN) receives all of the downsampled features for that sentence and outputs a single high-dimensional encoding of the sentence. RNN layers with long short-term memory (LSTM) units were used because of their ability to integrate information over time and model long-term dependencies, making them well suited for applications involving temporally dynamic processes such as speech (Hochreiter and Schmidhuber

1997; Gers et al. 2000). Finally, the decoder RNN tries to reconstruct the spoken word sequence from the high-dimensional encoding, resulting in mean word error rates as low as 3% for one participant (meaning that roughly 97% of the spoken words were successfully decoded) across 50 unique sentences and approximately 250 unique words. During training of the encoder RNN, the network is also required to predict acoustic features of the spoken sentences, and an auxiliary loss function is introduced as a form of "multi-task learning" to help the network determine effective encoding strategies (Caruana 1997; Szegedy et al. 2015; Kim et al. 2017). In addition, decoding performance could be improved by pretraining networks on data from one participant before being trained and tested on data from a different participant (Caruana 1997). This "transfer learning" demonstration adds to growing evidence that data from multiple participants can be aggregated to improve individual decoding performance (Peterson et al. 2021), which may enable improved initialization of decoding models for new users in future speech-neuroprosthetic applications.

The other of these two NSR studies involved an end-to-end network with deep neural feature-extraction and text-decoding ANN modules to decode spoken character sequences (Sun et al. 2020). With some architectural similarities to existing ASR approaches (Collobert et al. 2016), this NSR approach included language-modeling functionality and was used to decode character sequences, and the output was generalizable to vocabulary sizes of up to 1900 words while maintaining a word error rate of approximately 10%. This pipeline also used a multi-task learning approach by including a latent feature regularization module with auxiliary loss functions during model training to encourage the model to learn intermediate neural representations that were correlated with articulatory and acoustic speech features (Caruana 1997; Szegedy et al. 2015; Kim et al. 2017). Including this regularization procedure led to improved text decoding performance, further cementing multi-task learning as a useful approach for incorporating physiologically motivated information into a deep end-to-end model and supporting other findings of performance enhancement through data-driven neural-feature extraction techniques (Peterson et al. 2021; Schaworonkow and Voytek 2021).

Creating an AI-Driven Direct-Speech Neuroprosthesis

A primary motivation underlying many scientific and engineering efforts to understand how the brain encodes speech and to decode speech from brain activity is to eventually enable a clinically viable speech neuroprosthesis. Thousands of people suffer from anarthria, which is the loss of the ability to articulate speech, due to conditions such as stroke or amyotrophic lateral sclerosis that impair control of the vocal tract (Beukelman et al. 2007; Nip and Roth 2017). Those who also suffer from severe paralysis may be cognitively intact but unable to operate commercially available assistive communication devices, significantly limiting their ability to communicate with family, friends, and caregivers and drastically reducing their quality of life (Rousseau et al. 2015; Felgoise et al. 2016; Branco et al. 2021). Through remarkable advances in brain-computer interface (BCI) technology, researchers have demonstrated that paralyzed individuals can communicate by controlling a spelling interface using visual evoked potentials or imagined hand or arm movements (Wolpaw et

al. 2002; Sellers et al. 2014; Vansteensel et al. 2016; Pandarinath et al. 2017; Brumberg et al. 2018; Linse et al. 2018; Oxley et al. 2021; Willett et al. 2021). In a separate set of studies, researchers demonstrated that a paralyzed person could use a BCI to generate vowel sounds and phonemes (Guenther et al. 2009; Brumberg et al. 2011). However, none of these studies attempted to restore communication by directly translating brain activity into full words and sentences as the user attempted to say them. Because speech is typically the fastest and most natural form of communication (Chang and Anumanchipalli 2019), the development of BCI-based techniques to restore speech production could be transformative for those who are unable to speak naturally (Rabbani et al. 2019).

Guided by scientific findings and engineering advances from our research group, other speech neuroscience researchers, and the AI community, we started the BRAVO (BCI Restoration of Arm and Voice) clinical trial in collaboration with another lab at the University of California, San Francisco. Our goal in this trial is to evaluate the potential of custom hardware and software approaches to enable persons with severe paralysis to control a speech neuroprosthesis. In a pilot study with our first clinical trial participant, who has anarthria and paralysis due to a brainstem stroke, we used AI techniques to decode words and sentences (expressed as text, the lower path of processing in the rightmost part of Fig. 3) directly from cortical activity in real time as he attempted to say them (Moses et al. 2021).

In this study, we first recorded ECoG signals from the participant's sensorimotor cortex while he attempted to say individual, isolated words from a 50-word vocabulary. To enable user-paced engagement of the system and overcome difficulties associated with temporally aligning neural data to attempted speech without a clear, intelligible speech waveform (Martin et al. 2018), we trained a speech-detection model, which contained an RNN with LSTM units, to detect the participant's attempts to speak by analyzing the neural activity sample-by-sample (Kanas et al. 2014; Moses et al. 2019; Dash et al. 2020). Then to predict the likelihood of each word given a brief segment of neural activity associated with a detected speech attempt, we trained a word-classification model that contained temporal-convolution and gated-recurrent unit layers that are well suited for nonlinear classification with time-series data (Cho et al. 2014; Zhang et al. 2017). To reduce (1) modeling variance caused by random initializations of model parameters and (2) overfitting on the training data, we used a technique called "model ensembling," which involved training multiple models with identical architectures (but with different parameter initializations) on the same data and then averaging their predicted outputs (Sollich and Krogh 1996; Szegedy et al. 2015). We also used data augmentation, a technique in which training datasets are artificially enlarged through the use of label-preserving transformations of the original data, to increase the model's robustness to minor temporal variabilities in the ECoG signals (Krizhevsky et al. 2012). After training these models, the participant performed a real-time sentence-production task using combinations of the 50 isolated words. During this task, we used a Viterbi decoding approach to combine predictions from these two deep-learning models with information from a natural-language model to decode the sentences that the participant attempted to say in real time. This natural-language model described how likely certain sequences of words are to occur in English (Kneser and Ney 1995; Chen and Goodman 1999), and the Viterbi decoding

approach is commonly used in traditional ASR applications to decode the most likely text string from input features (Viterbi 1967; Gold et al. 2011). The median performance was favorable, with only a 26% word error rate between the target and decoded sentences and with a decoding rate of 15 words per minute (which was a comfortable rate for the participant), demonstrating that functional representations of speech were present in the sensorimotor cortex of a person suffering from anarthria for over a decade (Moses et al. 2021). Furthermore, the word error rate was below the 30% threshold at which text-decoding approaches can generally enable functional communication (Watanabe et al. 2017).

These findings represent a promising first step in direct-speech neuroprosthetic technology development. Overall, deep learning and other AI techniques have facilitated breakthrough findings and demonstrations in speech neuroscience and BCI research (Fig. 4). Further development and deployment of AI techniques for speech BCI applications can continue to produce profound impacts on the design and efficacy of future speech neuroprostheses (Chang and Anumanchipalli 2019). In the foreseeable future, persons who have lost the ability to speak could regain a voice to more intimately and efficiently express themselves in their daily lives.

CORTICAL PREDICTORS OF LANGUAGE: A TOP-DOWN EXPLORATION OF PEDIATRIC COCHLEAR IMPLANTATION USING MACHINE LEARNING (Author NMY)

Artificial intelligence has the potential to improve language outcomes and transform habilitation of children with hearing loss. Cochlear implantation, the first and still the only effective treatment of profound sensorineural hearing loss, has improved the lives of many children and adults. Its effectiveness in enabling speech perception significantly influenced neuroscience, in particular the concept that brain plasticity is lifelong (Merzenich 2011). The focus of our research is to use machine learning to predict the language outcomes of young-implanted children (i.e., children receiving their implants before 3.5 years of age), based upon brain structure and function. This line of research fulfills an unmet clinical need because although cochlear implantation enables many young children to develop spoken language, there is great variability in outcomes that is not understood. Our research group is exploring use of pre-surgical brain imaging to forecast outcomes on the individual level. The goal of this research is creation of a *predict-to-prescribe* approach to habilitation to improve the performance of implanted children at risk for poorer outcomes.

Fascination with prediction of spoken language development, and appreciation of the tremendous importance of effective parental engagement of children to build listening skills, began early in my career. In the 1990s, while establishing the first pediatric cochlear implant (CI) program in the city of Chicago, our implant team members repeatedly witnessed children making little progress in listening and spoken language, dramatically improve once the family began working with an experienced auditory verbal therapist. At that time, access to this type of therapy was very limited, as only a handful of trained therapists were available. In addition,

no early intervention program was in place in Illinois. This experience gave me tremendous respect for behavioral therapy and the importance of training parents to effectively engage their child in skill building throughout normal daily activities.

In clinical practice, the sparseness of the auditory signal provided by a CI and the essential role of brain plasticity may be difficult to appreciate given the remarkable listening and spoken language of many implant recipients. Over the years our pediatric implant team has observed many examples of surprising CI outcomes that illustrate near miraculous benefits of CI made possible primarily by the brain, rather than engineering. One example is a child born at 24 weeks gestational age, at risk for long term intellectual and developmental disabilities due to the extreme prematurity, with hearing loss. After implantation at age two years, he developed age-appropriate speech and language, learned to speak German well in addition to English as his first language, and became an excellent violinist (https://www.luriechildrens.org/en/specialties-conditions/cochlear-implant-program/cochlear-implant-videos/). His musical ability is remarkable because CIs, designed primarily for understanding of spoken language, provide much information about the temporal elements of music, but very little about melody and pitch (Gfeller et al. 2002). Yet somehow his brain enabled him to perform at a level many individuals with normal hearing do not achieve.

A special population of children who may have surprising results after cochlear implantation is the population with bilateral cochlear nerve deficiency (CND). These are children who have an absent or very small cochlear nerve, often below the resolution of magnetic resonance imaging (MRI). One of my first CND patients was implanted at age 3 in an ear with no eighth nerve visible on high resolution MRI, as well as a severe cochlear malformation precluding full electrode insertion. She was able to detect "s" after six weeks of implant use. She enrolled in an oral school and developed intelligible speech. Measurable open-set speech perception was present after 3 years. In our published series of 10 children with CND with a mean age of implant of 2.6 years, 70% had improved detection thresholds, one developed closed-set monosyllable word recognition after 12 months, and three developed open-set skills after 20 months (range 6 – 36) (Young et al. 2012). All but one of these children had a significantly longer time course for speech perception skills to emerge than children with normal anatomy. This additional time is apparently necessary for the brain to meaningfully use auditory stimulation impoverished by nerve hypoplasia.

Fifteen years ago, our team began a collaboration with Patrick C.M. Wong, PhD, a distinguished cognitive neuroscientist, linguist, and speech and language pathologist at Northwestern University with an interest in non-invasive studies of brain structure and function to predict language abilities. Dr. Wong had demonstrated that MRI brain scans of normal hearing adults can be used to predict second language learning ability and to improve language learning by lower performers by optimizing language learning training paradigms (Wong et al. 2007; Perrachione et al. 2011; Wong et al. 2017). Our collaboration to predict language outcomes of young-implanted children has continued since Dr. Wong became a Professor of Linguistics and the founding Director of the Brain and Mind Institute at the Chinese University of Hong Kong in 2013.

Improvement of language abilities is an important clinical need because significant language variability of young-implanted children in comparison to normal hearing children is a hallmark of pediatric CI outcomes (Niparko et al. 2010; Nittrouer and Caldwell-Tarr 2016). More than language is at stake because language and cognition bootstrap one another during early childhood development. Because spoken language requires not only the peripheral auditory system but the central nervous system to encode bottom-up input, it is not surprising that brain variability would be associated with language variability after implantation. In addition, there is evidence that auditory deprivation causes neuronal reorganization resulting in differences in perception and understanding of spoken language. Most previous studies of auditory deprivation have implicated the middle portion of the superior temporal region, including the primary auditory cortex, in producing deficits in language processing (Emmorey et al. 2003; Shibata 2007; Smith et al. 2011).

Several decades of cochlear implant research using traditional methods have not yielded accurate predictions of language outcomes. It is known that young age at implantation and residual hearing are two factors influencing the outcomes of children born with bilateral severe to profound sensorineural hearing loss. As predictors they fare well in conventional regression analyses when the sample size is large. However, their predictive ability is limited on an individual level. Accurate prediction on the individual level is a necessary first step for a *predict-to-prescribe* method to enhance language learning.

Speech perception with a CI relies upon the brain for bottom-up processing of the acoustic signal and top-down processing that uses knowledge and meaning gained from experience. The need to address variability and improve language outcomes of implanted children by taking neurocognitive risk factors into account has been proposed by Kral et al., who recommended development of a clinical test battery to create an individualized approach (Kral et al. 2016). The importance of brain function and the need for a top-down or cognitive neuroscience approach to designing cochlear implant systems has also been proposed (Wilson et al. 2011).

Given the normal variability of brain structure and function, we hypothesize that neural predictors from brain imaging may be well suited to forecast language outcomes. Prediction based upon neural brain-based factors has the potential to enable development of brain-specific training. If accurate prediction could be achieved at the individual level, customization of listening and language therapy for different brain types would be possible. We further hypothesize that custom brain-based therapy and optimal dosing of therapy could improve outcomes and be cost effective.

To develop predictive models on an individual level, machine learning is essential. Machine learning, a subfield of artificial intelligence, uses principles from computing, optimization, and statistics to create algorithms to accomplish tasks that can improve the algorithm's own performance as it gains experience. Many believe that machine learning will transform diagnostic medicine by making it possible to individualize treatment. A recent example of clinical applications of artificial intelligence is the 2018 FDA approval of IDx-DR, the first device using machine learning to detect early diabetic retinopathy in primary care settings. This technology analyses retinal images unsupervised by a human being with a sensitivity of 87% and a specificity of 91% in identifying pathology requiring referral to a specialist (Abràmoff et al. 2018).

Our research group has taken the first steps to apply machine learning to predict language outcomes of young-implanted children on an individual level. Our overall strategy is to use pre-surgical MRI scans to predict outcomes. T1 images provide information about brain volume, tissue density, and surface area. This approach is advantageous compared to functional MRI because the approach is objective, task free, and not affected by sedation or hearing status. Another advantage of T1 MRI scans is that they are commonly used in the pre-surgical evaluations of implant candidates. Therefore, prediction models based on T1 MRI are likely to be easily incorporated into clinical practice.

Two steps are involved in our research process. The first step is to determine how hearing impairment affects neuroanatomical development. To do so, the brain anatomy of implant candidates is compared to the brain anatomy of children with normal hearing. The comparison enables determination of brain regions affected and not affected by hearing impairment. This step also assists in hypothesis generation about brain circuits that may best predict outcomes, and whether regions affected or unaffected by hearing impairment would be most predictive. As described below, this first step enables predictive models to be built and the second step, in which comparisons are made between models relying upon brain regions affected or unaffected by hearing loss, enables hypothesis testing.

Our research process and the results of our T1 imaging pilot study were published in the Proceedings of the National Academy of Sciences (Feng et al. 2018). T1 scans of 37 implant candidates under 3.5 years of age were compared to T1 scans of normal hearing children from a National Institutes of Health (NIH) brain bank. The two groups were matched on age, sex, and socioeconomic status. Two different types of neuroanatomical analyses including evaluation of the density and patterns were conducted. These analyses were done for both grey and white matter. Across these types of analyses, the auditory cortex, especially primary areas, was most affected by hearing impairment. From this first step, we gained an understanding of areas affected and not affected by hearing impairment. This allowed building of models to evaluate whether affected or unaffected brain regions were most predictive.

The primary outcome measure was Speech Recognition Index in Quiet (SRI-Q) (Wang et al. 2008), using the Infant Toddler Meaningful Auditory Integrations (IT-MAIS) or Meaningful Auditory Integration (MAIS) scale at baseline (before surgery) and at six months after implant activation. Using a machine learning algorithm, implant candidates' neuroanatomical data was used to predict improvement over the six months at the individual level. This method permitted analysis of the sensitive and specificity of models in predicting each child's improvement.

One way of analyzing our data was use of a median split in which the subjects were divided into two groups with separate halves of the subjects classified as high or low improvers (Fig. 5). Models were built to predict a binary, high versus low, improvement classification using machine learning techniques and cross validation methods. Traditional characteristics including residual hearing and age at implant were able to predict improvement above chance. However, models built using brain regions affected or unaffected by hearing impairment were much more sensitive, specific, and accurate in prediction. Combining neural predictors with

traditional variables did not improve prediction. Importantly, the regions unaffected by hearing impairment were best at predicting outcomes. The unaffected brain regions that were the best predictors were mostly higher-level auditory and cognitive regions including the auditory association cortex and the dorsal auditory network engaged in speech perception and processing. These findings support the hypothesis that preservation of higher-level processing regions would be most predictive of language outcomes.

In ongoing work, we are also obtaining diffusion tensor imaging (DTI) as part of pre-surgical MRI imaging for some patients. DTI is widely used for research in cognitive neuroscience and neurobiology because it is noninvasive and sensitive to subtle pathology. It provides information about white matter fiber tracts and can be used to follow fiber tracts through the brain. Therefore, DTI can be used to evaluate brain connectivity in implant candidates.

Prediction models are created using a support vector machine learning approach with methods similar to those used for neuroanatomical prediction model building. Maps of brain connectivity are generated that demonstrate brain areas most predictive of performance at baseline (before surgery) performance and improvement in that performance at six months after surgery (Fig. 6). Our preliminary data indicate that brain areas predicting pre-implant performance do not overlap with areas predicting improvement. This finding is consistent with our theory that unaffected areas of the brain, the higher-level auditory and cognitive regions, are essential to improvement after CI, and not brain regions affected by auditory deprivation.

Support from NIH R21DC016069-01A1 and our collaborators from the University of Southern California, University of Michigan, and University of Miami have enabled us to construct neural predictive models to forecast outcomes in learning English by implanted children. However, the prediction time window is short and does not reflect the need for development of more complex aspects of language. We thus hope to expand our studies to five years as this would enable hypothesis testing of a new theory of "Neural Readiness for Spoken Language Development." Another aim of future work includes development of models for Spanish learning children with comparison to English learning. In addition, we plan to evaluate Parent-Implemented Communication Treatment (PICT), an intensive therapy with demonstrated effectiveness for hearing impaired children (Roberts 2019). Our goal there is to determine whether PICT will result in larger language gains for children predicted to have relatively poor language outcomes with the CI. Finally, an overarching goal of our research is to build models that will be useful across CI programs. For this reason, we will assess whether models work equally well across collaborating sites.

We hope our research will create new technology, advance theory, and improve the lives of children with CIs by providing clinicians with cost effective tools to improve language outcomes. For those children likely to have a very poor outcomes, prediction should not be used to deny cochlear implantation and other auditory devices, but to improve their lives. Children unlikely to develop language do benefit from hearing, a basic sense that is fundamental to quality of life and safety because it provides critical information about the environment and promotes caregiver and social engagement.

Accurate brain-based prediction maybe the first step toward personalized language learning for children with hearing loss. Prediction, when done well, may be a powerful tool to determine the optimal type and dose of therapy necessary to improve language outcomes. The general approach of using brain scans in conjunction with AI also should be applicable to developmental disorders other than hearing loss.

COULD ARTIFICIAL INTELLIGENCE SOLVE THE HEARING IN NOISE PROBLEM? (Authors FGZ, NAL, and BSW)

One in five people has some degree of hearing loss (Wilson and Tucci 2021). The most common complaint by people with hearing loss is difficulty in understanding speech in noise (Kochkin 2007; Vas et al. 2017). Hearing aids and cochlear implants can restore audibility and support speech understanding in quiet, but do not allow their users to achieve anywhere near normal performance in recognizing speech that is presented in competition with noise or other talkers (Zeng 2017; Lesica 2018). One reason for this disparity is that we still lack a good understanding of the neural and perceptual mechanisms underlying separation of signal from noise, especially when both signal and noise consist of speech, such as at a cocktail party (Cherry 1962). Most signal enhancement algorithms improve speech quality but fail to improve speech understanding (Loizou 2013). Thus, the "cocktail party" problem remains as one of the greatest challenges for both basic and translational researchers in otolaryngology.

Artificial intelligence (AI) has the potential to solve this long-standing hearing-in-noise problem for people with hearing loss (Slaney et al. 2020; Lesica et al. 2021; Wasmann et al. 2021). There are complementary approaches that can be taken, depending on the knowledge about the sound source, the acoustic environment (or "soundscape"), and the listener.

AI can enhance the signal from a known target sound source

AI can handle sound management in cases when the target sound source is already known (left column in Fig. 7). Many hearing devices already attempt to obtain a clean copy of a known source and deliver it to the listener at the highest possible signal-to-noise (or signal-to-competitor) ratio. In near-field situations such as conversing across a table in a noisy restaurant, a directional microphone turned toward a talker is effective in accentuating the talker's sound while attenuating surrounding noise or other talkers (Chung and Zeng 2009). In far-field situations such as listening to a teacher in a classroom, a preacher in a church, or a public announcement at an airport, frequency-modulation (FM) links, infra-red (IR) links, or hearing loops are effective in capturing the source and delivering it to a listener's hearing device(s) remotely (Boothroyd 2004; Mecklenburger and Groth 2016). Bluetooth-based devices, due to deep market penetration, will likely replace the FM, IR, and hearing loop solutions in the future (Einhorn 2017). True Wireless Stereo (TWS), a more advanced form of Bluetooth technology that has been made popular by Apple's AirPods, is capable of not only preserving binaural cues, but

also tracking head movements and accommodating for bilateral hearing loss for improved localization and recognition of a known sound source.

AI-based software can enhance the benefits of these existing approaches for hearing in noise. By analyzing soundscapes and tracking a listener's preference in different situations, AI can learn to adjust device parameters as needed. Such software is already included in some hearing devices and may well become standard in the future.

Additionally, AI could increase the signal-to-noise ratio at the listener's eardrum by incorporating deep learning algorithms that separate the target sound from the background noise (deep denoising), which effectively converts hearing-in-noise into hearing-in-quiet. Such deep denoising algorithms can restore normal speech-in-noise intelligibility (understanding) to hearing aid users (Wang 2017; Healy et al. 2019). But practical implementation of such denoising algorithms in ear-level devices may be years away due to the processing complexity and associated greater power consumption and additional processing delays produced by the algorithms. However, implementations could be practical now in smartphones that deliver their outputs to the user via earbuds and that have the additional benefit of being fashionable and thus possibly reducing or even eliminating the stigmas often associated with using hearing assistive devices. Alternatively, AI algorithms once trained may be simplified to the point that would allow implementations in ear-level devices, simplified algorithms that would have relatively low complexities (e.g., with less than seven AI "neural network" layers and 1024 or fewer processing nodes), low power consumption (< 10 mW), and low processing delays (< 10 ms), but with performances that approximate the performances of the original, starting algorithms. At present, the processing capabilities of smartphones vastly exceed the processing capabilities of even the most advanced hearing aids or other ear-level devices, and the processing capabilities of smartphones could be increased further, and by a large margin, with accelerator chips specifically designed to speed up AI computations in conjunction with the other resources already provided with standard smartphones.

When the target sound source is known, AI can also be used to provide the listener with alternative or complementary information streams. Automatic speech recognition and natural language processing can provide real-time speech-to-text conversion. Computer vision capabilities can add support for lipreading and signed languages, including gestures and body movements. Visual information can be displayed in smart glasses, in combination with improved auditory signals, to create an augmented or virtual reality environment to help solve the hearing-in-noise problem for people with hearing loss (Mehra et al. 2020). Further, real-time and accurate translation of sign languages to spoken languages and *vice versa* are possible with computer vision, AI processing of sound, visual, or sound plus vision inputs, and presentation of the outputs to smart glasses or earbuds, depending on the users and the direction of the translation.

AI can enhance the signal from even an unknown target sound source

AI can also help even if the target talker is unknown or changes over time due to the listener's shifted attention (middle column in Fig. 7). This application is based on two key technologies. One is source separation, which can separate several sound sources into separate streams without any prior knowledge about the sounds, also known as "blind" separation (Nachmani et al. 2020). The other key technology is cognitive control, which monitors the listener's brain (EEG) potentials to infer which sound stream is the desired target and delivers it as an isolated signal (O'Sullivan et al. 2017; Das et al. 2020). Both source separation and cognitive control rely on deep learning algorithms and, thus, face the same technical constraints as the deep denoising algorithm with respect to the development of real-world applications. Fortunately, such applications may be feasible for the reasons mentioned previously and because the brain signals can be adequately monitored from within the ear canal, which could be part of an overall ear-level device (Fiedler et al. 2017; Goverdovsky et al. 2017).

AI can enhance the neural coding of all sounds

It is important to note that the above-mentioned technologies focus on turning the difficult hearing-in-noise problem into a relatively easy hearing-in-quiet problem; none address impaired hearing *per se*. An alternative approach is to use AI as a tool to discover sound transformations that help provide people with hearing loss with normal or even enhanced auditory perception (right column in Fig. 7). This approach has the advantage of not making any assumptions about the nature of the target sound source or the listener's intent or attention, and has applicability beyond understanding speech in noise, e.g., to other sounds such as music.

A speech sound ("hello" in the top-left panel in Fig. 8) is processed differently by normal and damaged ears. This results in an abnormal pattern of neural activity that distorts the internal representation of the sound in the impaired system ("garble" in the top-right panel in Fig. 8). Lesica (2018) proposed a machine learning approach to design an "ideal" hearing aid, whose main function is to transform the spectral and temporal components of sound so that their internal neural representation in the impaired system is brought as close to normal as possible (bottom panel in Fig. 8). Note that the transformed output of the ideal hearing aid may not necessarily preserve the original sound's properties or even be intelligible to a normal-hearing person. But this is, of course, irrelevant if the transformed sounds are intelligible to the hearing aid user. Also note that this ideal hearing aid is fundamentally different from standard hearing aids, which restore audibility but do not improve – and may even impair – the neural coding of speech in noise (Armstrong et al. 2021). The key element to the development of such an ideal hearing aid is to use machine learning (AI) to derive the optimal sound transformation for each individual listener.

The same approach could be applied to cochlear implants. The traditional implant design attempts to extract and deliver the spectral and temporal information in speech that is presumed to be important for perception. An AI-based implant would have a different design goal, using machine learning to deliver electrical (or optical) stimuli that generate as closely as possible an approximation to the normal patterns of neural activity in the

auditory nerve, inferior colliculus, or auditory cortex. The use of AI to address hearing loss directly has great potential but is still in its infancy.

Integrating AI-based approaches to improve hearing in noise

AI approaches that serve to enhance the sound signal itself are complementary to those that attempt to address hearing loss directly by enhancing neural coding. Even if an ideal hearing aid could restore normal hearing, signal enhancement (e.g., deep denoising) would still be helpful in many situations, as hearing in noise is often difficult even for listeners with normal hearing. Inasmuch as the effects of background noise on the neural coding of speech in noise appear to be similar for normal-hearing and hearing-impaired listeners (at least for mild-to-moderate impairment) (Armstrong et al. 2021), the same signal enhancement approaches could benefit both groups of listeners.

Possibly too, a focus on neural coding could improve signal enhancement. Current speech denoisers, for example, are optimized in the acoustic domain to convert noisy speech into a new sound that is as close as possible to a clean version of the same speech. But, given that the capacity for denoising is limited, this approach is likely suboptimal in that it removes noise that would not be perceived. For example, the world is full of echoes that we do not perceive because they are filtered out of our internal representation through processing in the early auditory pathways. If the goal is to enhance speech perception, then denoising should be optimized in the neuro-perceptual domain to convert noisy speech into a new sound that elicits a close approximation to the neural activity elicited by perfectly clean speech. With this approach, the limited capacity of the denoiser will be automatically focused on those aspects of the speech and noise that are most important for perception.

Ultimately, the success of solving the hearing-in-noise problem depends on fruitful interactions between AI and hearing research. To date, modern AI has been mostly driven by vision research and applications to computer vision with little-to-no attention to hearing (Anon. 2021; Lesica et al. 2021). The resulting technologies excel at feature detection, e.g., identifying a known face from a crowd or measuring body temperature of hundreds or even thousands of people per second. But current AI technologies lack features that are critical for hearing, such as the capacity for feature binding that is essential for object formation and stream segregation (Hinton 2021).

Contributions from hearing research can help improve the functionality of AI in ways that are important for hearing in noise. After all, a music lover with normal hearing can differentiate and appreciate the parts – as well as the whole – of Beethoven's glorious Ninth Symphony, in which the many instruments in an exceptionally large orchestra, four solo voices, and a large choir are playing and singing simultaneously at times and in harmony throughout. At present, no AI system with two sensors (corresponding to the two ears) can achieve this analysis-synthesis performance. If hearing researchers can elucidate the mechanisms that enable the auditory system to perform as it does, and how these mechanisms are disrupted by hearing loss, the results will provide direction for the development of AI with similar abilities (Lesica et al. 2021).

AUTOMATED CLASSIFICATION OF LARYNGEAL LESIONS USING DEEP NEURAL NETWORKS: PROGRESS MADE AND CHALLENGES AHEAD (Authors AMB, HK, CM, JP, SG, SK, and GW)

Introduction

Flexible transnasal laryngoscopy is a diagnostic procedure commonly performed by otolaryngologists to evaluate the upper aerodigestive tract in patients with dysphonia, dysphagia, or dyspnea. It takes an average of only six attempts for a novice to become competent in the mechanics of performing flexible laryngoscopy (Laeeq et al. 2010). However, developing the expertise and clinical acumen to correctly interpret laryngoscopic findings requires years of training (Brook et al. 2015). Consequently, flexible laryngoscopy is typically only performed by highly trained specialists. When a lesion is present, subtle differences in its appearance enable otolaryngologists to differentiate benign phonatory lesions from potentially malignant ones. Failure to appreciate these differences can be potentially harmful to patients, as the correct management changes greatly based on the pathology. For example, if laryngeal cancers are not recognized and treated promptly, disease progression, respiratory compromise, and even death can ensue. Conversely, if benign nodules are misdiagnosed as malignancies, patients may undergo unnecessary, irreversible, and possibly harmful surgery for lesions that may respond to conservative, non-surgical management.

Machine learning is a subset of artificial intelligence that enables computers to learn from historical data, gather insights, and make predictions about new data using the information learned (Bur et al. 2019). Machine learning has been shown to have high degrees of accuracy and precision that exceed the abilities of standard statistical techniques and human judgement to make predictions about outcomes in medicine (Michie et al. 1994; Obermeyer and Emanuel 2016). An artificial neural network (ANN) is a type of machine learning algorithm that is loosely based on how biological nervous systems process information. ANNs consist of multiple layers of interconnected nodes. Each node performs a series of nonlinear calculations based on its inputs and signals other nodes connected to it. Each connection, much like the synapses in the human brain, transmits data from one node to the next. Complex ANNs that contain multiple hidden layers are known as *deep neural networks* and are well suited to solve the most complex problems in machine learning. A convolutional neural network (CNN), a class of deep neural networks, are commonly used for computer vision applications such as image classification and facial recognition (Lawrence et al. 1997; LeCun et al. 2015). The applications of deep learning in medicine are rapidly expanding. However, it remains to be seen how machine learning can render diagnoses in real time in order to affect patient care, and what practical limitations may exist.

A system capable of generating an automated diagnosis of laryngoscopy findings would have multiple applications in both education and clinical practice. First, an automated diagnostic tool could improve access to care for patients living in communities with limited access to otolaryngologic care. The tool would also be a

valuable in the education of medical students, residents, and speech language pathology students who are first learning to interpret laryngoscopic exams. With this in mind, we sought to develop a convolutional neural network for the detection and classification of structural laryngeal lesions during flexible transnasal laryngoscopy. The objectives of this summary of our initial work are to briefly introduce how to create such an algorithm and to describe other lessons learned throughout the process.

The Initial training of a Convolutional Neural Network – Start Small

To evaluate the feasibility of training a CNN to identify structural lesions of the larynx, our efforts began with a simple experiment. Tutorials that teach the general public how to train their own CNNs are readily available online. After completing several tutorials and introductory courses, an initial dataset was created. Forty images of the larynx, both diseased and healthy, were identified and collected using Google's Image Search. Due to the small dataset, a single, simple task was posed – train a CNN to differentiate between images of normal larynges and locally advanced laryngeal cancers. A 12-layer CNN was trained with 10 images of the larynx using R Studio (Boston, MA). These images included five normal larynges and five larynges with clearly visible and locally advanced cancers. With some minor modifications in code, it was possible to successfully train the CNN to correctly differentiate normal organs from diseased organs. The model was then validated using an independent test set of 4 laryngeal images and successfully classified two normal larynges and two laryngeal cancers (Fig. 9).

Next, a set of laryngeal images was collected from 180 patients who underwent digital laryngeal stroboscopy as part of their care in the Department of Otolaryngology – Head and Neck Surgery at the University of Kansas Medical Center. The images were randomly assigned to training and validation sets in an 80:20 distribution, respectively. The original training image set was expanded to 51,953 images using image augmentation (a process in which images are manipulated by randomly shifting, horizontally flipping, or zooming the original image). A MobileNet CNN pre-trained with 1.4 million images was used as a base model. That CNN was then retrained in python using the collected laryngeal images, which were each labeled as "normal," "benign lesion," or "suspicious for malignancy." Overall classification accuracy on 36 independent test images was 72.2% (Fig. 10). The CNN performed best for normal and suspicious images, identifying these classes correctly in 83.3% of the validation images. However, performance was notably worse for images of benign lesions with classification accuracy of 58.3%. This lower accuracy may have been due to heterogeneity in the different types of benign lesions (e.g., polyp or papilloma) included within the dataset.

Image quality affects classification accuracy

Developing deep learning algorithms to analyze endoscopic images presents unique challenges that don't apply to the analysis of pathologic or radiographic images. To ensure a model is reliable, the data used for training must be carefully selected and the accuracy of the labels (or diagnoses) must be carefully reviewed. Although it

may seem obvious, images used in both the training and testing datasets must be of sufficient quality to allow for the correct identification of any existing pathology. Previous algorithms to classify laryngeal disorders have been trained and validated using carefully curated images in which the structures of interest are clearly visible. This, however, is not always feasible during clinical encounters.

Many variables may transiently distort or obscure visualization of the structures of interest during real-time fiberoptic laryngoscopy. First, lighting and perspective can dramatically alter the appearance of structures or lesions of interest. Depending on the position of the tip of the endoscope, a lesion may be clearly visible and well-lit, or it may be largely obscured by shadows. Additionally, laryngoscopy examines mobile structures which, by definition, change their positions during the exam. If the patient swallows during flexible laryngoscopy, visualization of the larynx is temporarily obscured. Normal physiologic movement of the larynx may also obscure or alter the appearance of a lesion of interest.

To overcome this inherent variability, human clinicians are trained to render a diagnosis based on interrupted views, knowledge of the pertinent anatomy, and the gestalt of the complete examination. Automated systems must be able to recognize when a clear view of the larynx is present and disable classification when it is not. Further, for systems capable of localizing small lesions within a larger image, models must be "taught" to localize the lesion. This requires thousands of images in which the boundaries of the lesion have been manually identified, to train the network. All this together means that development of a highly curated dataset of laryngeal images for training automated systems is extremely labor intensive and time consuming.

Using a CNN for laryngeal lesion identification requires massive quantities of data

It is well known that training and validating CNNs for computer vision tasks requires extremely large datasets. In a study by Ren et al. (2020), the authors used 24,667 images from 9,231 patients to train and validate a model for the classification of five different types of laryngeal lesions. Despite the size of this dataset, the authors acknowledge that their model cannot be used to identify other types of lesions, including laryngeal papilloma. Papilloma is a unique pathology in that it is exceptionally variable in its distribution within the larynx, creating no set location for which the CNN can predictably focus to classify the lesions. Thus, even more images are needed to develop a model capable of identifying a wider variety of laryngeal lesions.

Lesion localization is also critical to building the trust of clinicians and patients

Machine learning algorithms are often likened to a black box. Although we can see the data that go in and the prediction that is generated, we cannot comprehend how an algorithm determines its prediction because the computations that occur internally cannot be meaningfully interpreted (Yu et al. 2018). This is especially true for computer vision algorithms. We have limited tools to peer into the inner workings of the model to understand how it makes its decisions.

The intended role of automated diagnostics in flexible laryngoscopy is to support, not replace, the judgement of the clinician. Prior studies have shown that patients prefer to receive care from a human care provider than an automated one. The patients are even willing to accept a greater risk of an inaccurate diagnosis or treatment complication in order to receive care from a human provider (Longoni et al. 2019). As such, acceptance of this technology by both clinicians and patients is essential to its successful integration into clinical practice.

Lesion localization is one way to integrate the clinical judgement of the provider with the trained CNN. Highlighting the area of interest helps "explain" the diagnosis rendered by the algorithm, showing the lesion in an image so that clinicians can easily confirm or reject the automated diagnosis. Lesion localization also has important ramifications in the technical aspects of reliable classification. As a laryngeal lesion becomes smaller, the number of pixels that contain information useful for lesion classification also decreases. Therefore, a classifier that uses the entire image for classification would be expected to perform worse on small lesions. An automated system that first localizes the lesion within the image can overcome this issue by focusing a classification network on the pixels that are most likely to contain information necessary for classification. To address the need to localize lesions within an image, our team has developed an approach using a location-aware, anchor-based reasoning neural network with support from the National Cancer Institute (1R03CA2532212-01).

Comparisons of performance between human clinicians and automated systems must be fair

Results from previous studies have suggested that automated systems may be more accurate in the classification of laryngeal lesions compared to human clinicians. However, this suggestion does not take into account the way these data were reviewed. Ren et al. (2020) asked clinicians to determine a diagnosis based on still laryngeal images. Although data must be fed into deep learning classifiers as still images, this format is unfamiliar to human clinicians, who depend on the global assessment of tissue movement, moisture, and visualization of lesions from multiple perspectives to determine the best clinical diagnosis. This additional information, which clinicians depend on, can only be discerned from video. It is therefore not surprising that the authors found that clinicians were only able to determine the correct diagnosis in 62% of the images. To better compare the performance of humans and automated systems, each must be presented with the data in its preferred format: clinicians need laryngoscopic videos and CNNs need still images extracted from laryngoscopic videos. To achieve this, performance between human and automated classifiers should be compared between unique patient exams and not based on a single still image, which gives the CNN a significant advantage.

The case for using artificial intelligence to leverage digitized medical data to support clinicians

We live in an era that is characterized by an unprecedented explosion in data. Each minute approximately 300 hours of video are uploaded to YouTube alone. At least a trillion new digital photos are taken every day and

more than 80% are now taken with smartphones (Perret 2017). A similar trend exists in the growth of data in medicine. Between 1950 and 2010, the doubling time of medical knowledge decreased from 50 years to just 3.5 years and was predicted to be just 73 days by 2020 (Densen 2011). To put this in concrete terms, the prediction suggests that an otolaryngologist who completes residency training in 2020 or beyond will experience at least four doublings in medical knowledge during her or his years of medical school and the residency training.

Undoubtedly, knowledge is expanding faster than our ability to assimilate and apply it effectively. This explosion in knowledge and data has driven the need to develop new technologies, including artificial intelligence, to support clinicians in caring for their patients using the best available evidence. With so much digitized medical data, large sets of clinical, pathological, and radiographic images are now readily available. Computer-assisted diagnosis using deep learning has come to the forefront of technological innovation as a way to aggregate and apply these data (Komura and Ishikawa 2018; Bur et al. 2019). Deep learning has been used to detect metastatic breast cancer in whole-slide images of sentinel lymph node biopsies and was found to reduce human pathologists' error rate by 85% (Wang et al. 2016). In radiology, Henschke et al. (1997) trained a neural network to differentiate malignant from benign lung nodules using CT images. The authors' algorithm correctly identified all 14 malignant nodules and 11 of 14 benign nodules, yielding a diagnostic accuracy of 89%.

Several studies have reported excellent performance using deep learning diagnostic support in esophagogastroduodenoscopy (EGD) and colonoscopy. Takiyama et al. (2018) trained a CNN to recognize the anatomical location in EGD images. Using 27,335 images obtained from 1,750 patients, the authors trained a neural network to identify the larynx, esophagus, and stomach with greater than 93% sensitivity and specificity. Deep learning has also shown efficacy in the endoscopic evaluation of helicobacter pylori (HP) infection. Shichijo et al. (2017) trained a CNN to detect the presence or absence of HP infection using 32,208 EGD images. A separate test dataset of 11,481 images from 397 patients was then independently evaluated by their algorithm and 23 endoscopists. The authors found that the CNN achieved higher accuracy than the trained endoscopists in detecting HP infection using endoscopic images (by 5.3%, 95% CI 0.3-10.2%). In 2017, Komeda et al. (2017) described a CNN-based diagnostic tool to classify colorectal polyps using 1,200 colonoscopy images. When applied to new polyp images, the tool classified 7/10 images correctly.

Ren et al. (2020) developed a deep learning-based diagnostic system, which was trained using 19,433 images from 7,521 patients to classify laryngoscopic images by diagnosis as normal, vocal nodules, leukoplakia, benign, or malignancy. The authors used transfer learning, whereby a previously trained model, ResNet-101, was retrained for the task of classifying the images. The model was then validated using 5,234 images from 1,710 patients. The classifier achieved an impressive overall accuracy of 94% on test images, significantly higher than human experts who classified the same lesions with an accuracy of only 62% (p <0.001). The authors included images taken under a range of angles, zoom, and magnification and with different sizes of opening of the vocal cords to ensure their model would be robust to these variations that occur in real world data. While promising, the authors' study highlights some of the obstacles discussed above to develop a model that can feasibly be used by clinicians for real-time diagnostic support.

Conclusion

Deep learning using CNNs can feasibly automate the detection and classification of laryngeal lesions using laryngoscopic images. Further development of this technology has applications in training novice clinicians to interpret laryngoscopic exam findings and may improve access to care for patients living in communities with limited otolaryngology resources. Although automatic identification of laryngeal lesions shows promise as a diagnostic tool, much work is still needed for the automatic identification to become clinically relevant. Deep learning algorithms must be highly accurate in classifying a wide range of structural laryngeal disorders, including papilloma, polyp, nodules, and malignancy. To train an algorithm that will reliably identify and classify laryngeal lesions, tens of thousands of uniquely labeled images are needed, often requiring manual labeling. Development of large sets of correctly labeled medical images is labor intensive and costly. Importantly, algorithms that localize lesions within an image or video are much more likely to be accepted and trusted by clinicians and patients.

DIGITAL PATHOLOGY: IDENTIFYING THYROID MALIGNANCY USING DEEP LEARNING (Author JMC)

Introduction

Thyroid nodules are very common and approximately 10% of the population will develop a nodule in their lifetime. Despite the frequency, most of these nodules are benign with only 10% of them harboring malignancy. The single most important step in diagnosing these nodules is fine needle aspiration (FNA) biopsy. It entails inserting a needle into the suspicious thyroid nodule under ultrasound guidance to extract cells which are then examined under a microscope by a cytopathologist.

The universally accepted system for thyroid cytopathology is The Bethesda System for Reporting Thyroid Cytopathology. The Bethesda System (TBS) recognizes five diagnostic categories and provides an estimate of malignancy risk within each category. (The TBS also includes a sixth, non-diagnostic category.) Ultimately, clinical decision making is based on synthesis of the medical history, physical exam, imaging, and the FNA biopsy. The clinical decision is straightforward for Bethesda II and VI (benign and malignant lesions, respectively); however, there are many cases where the decision making is not clear-cut. Specifically, Bethesda categories III-V are considered the indeterminate group and pose a dilemma for the clinician; despite having all available clinical information, it remains uncertain as to what treatment will ultimately be most beneficial for the patient. In these indeterminate cases, where clear actionable evidence-based guidelines are lacking, treatment options include follow-up, repeat FNA, molecular testing, and surgery. Often, these patients are sent for surgery; however, the pathological result for most of these operations is a benign lesion. More specifically, the final pathology for more than 70% of the Bethesda III and IV patients, which comprise an average of 35% of all

surgical cases, is benign. Thus, retrospectively, those surgeries were "unnecessary." Therefore, the thyroid nodule evaluation process, though well accepted, is at times very inefficient, and entails substantial uncertainty, repeated tests, follow ups, and a significant proportion of surgeries on patients with benign nodules. Consequently, this standard evaluation process has adverse impacts on the patient, the clinician, and the cost of patient care.

In recent years, machine and deep learning algorithms have become the state-of-the-art computational approach for data analysis. In their simplest form – termed supervised learning – these algorithms learn to perform data analysis tasks (classification, regression, etc.) from a set of examples; this process is known as algorithm training. Convolutional Neural Networks (CNNs), a variant of these deep learning techniques, are now in widespread use in computer vision and image analysis, where they consistently achieve substantial improvements over classical methods for various classifications and for regression and statistical inference tasks. In the context of medical imaging, use of deep learning is becoming more common as well. For example, for the classification of skin cancer, the detection of diabetic retinopathy, and in histopathology and cytopathology.

The use of deep learning for determination of thyroid cancer has also been explored. Recent work studied the use of deep learning for the detection of thyroid malignancy in ultrasound images. The work employed supervised learning for the characterization of thyroid cells in histopathological sections.

Nevertheless, machine learning techniques have not been studied for the prediction of thyroid nodule malignancies in FNA cytopathology slides, whose analysis has the greatest impact on clinical decision making. We are unaware of clinically applicable algorithms for cytopathology-based prediction of thyroid malignancy.

This project aims to establish a clinical tool for the accurate determination of a thyroid nodule malignancy via the development of a machine-learning algorithm for the analysis of digitized cytopathology slides. The purpose of the algorithm is to detect and classify malignant cytopathological features of thyroid nodules. Consequently, the algorithm could reduce uncertainties associated with thyroid nodule work-up and, in turn, improve the accuracy of clinical decisions, thereby reducing the number of unnecessary operations.

Methods

Our cohort included patients who had a thyroidectomy with a preceding FNA biopsy. Exclusion criteria included patients with a non-diagnostic FNA (Bethesda I category) and cases with equivocal pathology. Also, cases where we couldn't correlate the biopsied nodule with the surgical pathology were excluded. For the included cases, one alcohol-fixed, PAP-stained, smear from each biopsied nodule was selected and scanned at x40 magnification with a Leica AT-2 scanner. Slides were divided into the training set, consisting of 88% of the slides, and the separate test set, which consisted of about a hundred slides. The test set was used for evaluating the performance of the proposed approach.

The algorithm designed was based on a sequence of two machine learning algorithms (MLAs). The first was the screening MLA, designed to identify regions of interest (ROIs), i.e., the follicular groups within the

slide. These groups are the "informative" portion of the slide and what the pathologist bases her or his diagnosis on. The groups typically comprise only 1% of the slide, leaving the remaining 99% as a non-informative background. Once trained, the screening MLA was applied to identify the 100 ROIs with the highest prediction values of being informative. The second MLA (the classifier) was designed to predict the Bethesda score as well as the final pathology (benign or malignant) based on the 100 ROIs the screening MLA identified.

The performance of the combined algorithm including the two MLAs was compared to the findings from three experienced and trained cytopathologists who independently reviewed and provided a Bethesda score for each of the test set slides. Additionally, the performance of the combined algorithm was compared with the finding reported in the medical record for each patient by another experienced and trained cytopathologist.

Results

The data included whole slide images (WSIs) that were divided to a training set of 799 WSIs and a test set of 109 consecutive FNA biopsies. The area under the receiver operating characteristic (ROC) curve for the screening MLA separating between ROIs and non-ROIs was 0.985. A "heat map" of the screening MLA's prediction is shown in Fig. 11 and Fig. 12 shows the ROC curves comparing the predictions of the final pathology. The performance of the combined MLAs and the three reviewing pathologists (Experts 1-3) is reflected in the areas under the curve (AUC) for each case. The MLAs outperformed experts 1 and 2 and was comparable in performance to that of expert 1. Combining the MLA predictions with information in the electronic medical records (EMRs) for the patients increased the AUC to 0.962 (not shown in Fig. 12).

Discussion

The MLA performance was at least comparable to human performance and even surpassed it in two cases. All FNAs the MLA predicted as Bethesda II (benign) or Bethesda VI (malignant) were indeed benign and malignant, respectively, on final pathology. Similarly, the MLA did not classify a malignant nodule (per final pathology) as benign or classify a benign nodule (per final pathology) as malignant.

We then examined accuracy of the algorithm when combining it with human performance recognizing that merging the two might enhance performance. We formed a rule that directed the algorithm to use the EMR cytologic diagnosis if it was Bethesda II or VI (benign or malignant, respectively) and to give predictions only for the indeterminate categories – Bethesda III-V. This improved the AUC from 0.931 to 0.962 and with the specificity increasing from 90.5% to 92.9%. This result demonstrates the potential of using the MLAs as an adjunct tool to assist the clinician and reduce the number of indeterminate cases.

These results from the current study indicate the potential for future clinical use. For example, the first of the MLAs in the combined MLA could be used as a screening tool for identifying and highlighting the follicular groups which are ~1% of the slide, and thus valuable time could be saved for the pathologist. To demonstrate the utility of such a tool we designed and implemented a machine-learning-based software that summarizes WSIs by

generating an image gallery of automatically identified ROIs containing follicular cells. The software selects to top 100 ROIs which are the most informative and projects them on the pathologist's screen eliminating the need for the pathologist to review the whole slides manually. We had an experienced board-certified cytopathologist blindly review and assign a TBS category to each WSI in the test set using software without the ROIs being presented. Then, using the image gallery created by our software (identifying the top 100 ROIs), the same cytopathologist reviewed the test set 117 days later, demonstrating almost perfect concordance across the TBS categories. The MLA could also be applied to specific FNAs with an indeterminate diagnosis in combination with the EMR as demonstrated by the ROC curve. This combination would categorize a significant amount of the indeterminate group into either the benign or malignant subgroups. Additionally, adding data – such as sonographic and other clinical data – to our current cascade of MLAs could boost performance.

Our approach has several advantages. First, unlike many other MLAs using cytopathological images as their dataset, our approach requires little human effort. Other approaches necessitated manual annotations and hours of physician time. The use of WSIs obviated the need for manual acquisition and analysis and allowed the use of large amounts of cells and other data for analysis.

But our study also has several limitations worth mentioning. MLAs necessitate datasets with ground truth, which in our case was surgical pathology. This reliance might create a selection bias towards cases undergoing surgery (as opposed to surveillance) that may favor malignant or complex cases. However, our cohort was like other studies when we examined the rate of malignancy and the distribution of the different BTS categories. Currently, we are conducting a prospective study to compare the performance of the MLAs among all FNA cases both with and without surgery to evaluate this potential bias. The number of indeterminate cases was relatively small and, to evaluate fully the MLA's robustness, a larger test set would be necessary. An additional limitation is the fact we did not scan all available slides for each given case due to scanning and storage challenges; rather, we scanned a selected single slide for each case. This case was selected by a pathologist who selected the slide that was most informative – the slide with the most follicular groups. The selection, despite being done by a trained person, might have resulted in choosing a slide that was not representative or fully representative of the cytologic diagnosis.

Cytology slides (opposed to a histology slide) have "depth" and are essentially three dimensional. This 3D aspect necessitates scanning in multiple slices (z-stacks) to capture all the data. Our dataset included nine z-stacks per WSI, however the MLAs use only the middle z-plane. Using multiple z-stacks would require considerable computer time for the MLA calculations, so we conducted a preliminary experiment with a subset of slides comparing use of all slides vs. only the single selected slide and saw no difference in the accuracy of the MLAs. Thus, the faster approach does not have a downside, at least according to the results from the preliminary experiment.

Conclusions

To our knowledge, our cohort is the largest cohort of thyroid FNA biopsy WSIs reported in the literature. We used this cohort to train and test a cascade of MLAs to predict pathology based on the FNA biopsies. The results demonstrate that the MLAs can serve as a useful clinical tool for both the pathologists and the surgeons with many advantages compared to current clinical practice.

GENERAL DISCUSSION (Authors BSW, DTL, GSG, GD, and HWF)

A wide diversity of applications of artificial intelligence (AI) and machine learning in otolaryngology & the communication sciences is presented in this Review, including uses or potential uses of AI in (1) decoding recordings from 128 sites overlying the sensorimotor cortex to produce speech for paralyzed persons who have little or no control over their articulators but still have intact cognitive abilities that allow the persons to attempt to make speech sounds, which activates discrete areas in the cortex that move over the time course of the attempted speech; (2) predicting language outcomes from preoperative brain images – from T1 and diffusor tensor imaging (DTI) MRI scans – for young (less than 3.5 years of age) candidates for cochlear implants (CIs), predictions that are far more accurate than forecasts using prior methods and predictions that can enable a *predict-to-prescribe* approach to habilitation after the CI to improve the performance of individual children determined to be at risk for a relatively poor outcome; (3) improving the performance of hearing aids and CIs, particularly for attending to and understanding an individual talker in typical acoustic environments – environments with other talkers, noise, and reverberation (e.g., in cocktail parties, bustling cities, workplaces, and restaurants) – and including a new approach in which AI is used to infer sound inputs that could produce discharge patterns in the auditory nerve in persons with hearing loss that approximate more closely than previously the patterns that are produced in normal hearing; (4) automated detection and classification of laryngeal lesions from laryngoscopy images in a pilot study with a small dataset of 180 carefully selected images for clarity (with good lighting and good positioning of the tip of the endoscope), which produced an 83% accuracy for discriminating a normal larynx from a diseased larynx in the 36 test images within the dataset and a 72% accuracy for classification of the same images as showing a normal larynx, a benign lesion, or a suspicious lesion; and (5) automated diagnoses of thyroid samples in which the performance of a cascade of two deep learning algorithms – the first to identify regions of interest (ROIs) in smears of biopsied tissue on slides and the second to focus on the ROIs to identify possible pathologies according to the standard Bethesda System for reporting thyroid cytopathology – matched or exceeded the performances of three expert pathologists examining the test tissue specimens separately.

With respect to the first application, we note that, although text outputs were produced in the tests for the prototype speech production prosthesis (using the processing shown in the lower right portion of Fig. 3), audible speech outputs could be provided in an implementation of a prosthesis that uses the processing indicated in the

upper right portion of Fig. 3. The audible outputs might be better for supporting fluent conversations; an option for selecting the type of outputs could be provided for the user in a next-generation prosthesis. (More information about the processing to produce the audible outputs is presented in Anumanchipalli et al., 2019.)

With respect to the fourth application, Andrés Bur and his coauthors note challenges posed by the classification of laryngoscopy images with AI including the needs for (1) large datasets that are representative of the full range of pathologies, (2) high image quality, and (3) trust by physicians and patients. Of course, large (and well labeled) datasets and trust are needed for all clinically useful applications of AI. The trust issue arises because the internal workings of AI algorithms are not known and are sometimes regarded as a "black box." Validation data may show that an algorithm is producing highly accurate results, sometimes even exceeding the performances of expert clinicians, but a concern may remain for an aberrant finding from an AI algorithm. The concern may be addressed by regarding AI algorithms for diagnoses as assistants to clinicians, who may review borderline classifications by the algorithms. In this partnership, the algorithms also may detect pathologies or produce classifications that are not observed even by highly trained clinicians, and thus alert the clinicians to those possibilities, and the algorithms may save clinicians considerable time in initial diagnoses that can be handled by the algorithms. Bur et al. also note that AI diagnoses could be especially valuable for screening in rural areas, or in low- or middle-income countries (LMICs), where expert clinicians for a particular specialty such as diagnosis and treatment of laryngeal cancers may be scarce or not exist at all. In these areas and countries, AI diagnoses via telemedicine could provide a first line of referrals to tertiary care hospitals for cases in which the diagnoses indicate a cancer or suspicious lesion.

Bur et al. further remark that clinicians make their judgements based on moving images of the larynx, as observed through the laryngoscope. The moving images provide more information than the still images used to date in AI analyses of possible pathologies, and thus comparisons of performances by clinicians versus performances by present AI algorithms should use the moving images for the clinicians and the still images for the AI algorithms. Additionally, a corollary to this observation is that the performances of AI algorithms could possibly be improved by using moving images (videos) as inputs to the algorithms rather than the still images.

As noted previously for the fifth application, a cascade of neural networks was used, with the first of the networks trained to identify ROIs in the scans of the tissue samples. That made the second network, trained to identify possible pathologies, far more efficient and accurate compared with using the full scans as inputs rather than the selected regions, which typically comprise about 1% of the area in the full scans. Bur et al. also advocate this two-step procedure and will be using it along with large and fully representative datasets in further development of their AI system for analyzing laryngoscopy images.

Also with respect to the fifth application, Jonathan Cohen notes that combining clinical findings from the electronic medical records with the findings from AI diagnoses can improve accuracy beyond either the clinical or AI findings alone. Further, Cohen agrees with Bur et al. that (1) AI algorithms could be used as an assistant to clinicians, e.g., to identify ROIs in images or to direct attention of the clinician to further examination by her or

him for borderline cases, and (2) AI diagnoses in conjunction with telemedicine could be especially valuable for screening in geographic regions where specialty care is overburdened or completely unavailable.

The high performance of the AI algorithms in the fifth application also supports the suggestion offered by Bur et al., in connection with the fourth application, that large datasets are needed to adequately train the neural networks (a conclusion also reached in many other studies; see, e.g., the review by LeCun et al., 2015). The dataset in the fourth application included 180 items, whereas the dataset in the fifth application included 908 items. Within these datasets, 144 of the items were used for training for the fourth application and 799 of the items were used for training for the fifth application, and the remaining items in each of the datasets were used for testing the applications. Accuracy was good (encouraging) for the fourth application and outstanding for the fifth application, with accuracy for that latter application exceeding human performance for two out of three expert pathologists and matching the performance of the third expert pathologist. Of course, other elements also contribute to performance, such as the quality and representativeness of the data; the accuracy of labeling the data; the overall design of the network(s) including for example the number of network layers in deep learning networks; the cascading of networks (as used in the first and fifth applications); and the use of data augmentation or transfer learning or both (as used in the first and fourth applications). As emphasized in LeCun et al. (2015) and by others, much larger datasets than even the 908 items in the dataset for the fifth application are recommended and indeed are essential to high performance for many applications of AI in other fields such as face recognition, language translation, and automated "hands free" driving. In general, the necessary minimum number scales with the difficulty of the problem and the desired accuracy of the predictions.

Although not described in the presented applications, we note the high importance of addressing the "digital divide" in uses of AI and machine learning in medicine and in most other fields for that matter. Many people in the world do not have adequate if any access to the internet and cell (or smart) phone services; many people who have access do not have the digital literacy or funds to take advantage of it; many populations such as women, children, residents of LMICs, indigenous populations, and refugee populations are underrepresented if represented at all in many digital datasets relating to health; and balanced representation across ages, races, ethnicities, castes, income levels, and world regions (with their varying conditions) is rare in the datasets. These are elements of the digital divide, and the listed inequities can and most likely do introduce biases in the data used to train neural networks and thus may produce errors in predictions from the networks when applied to a population that was not included in the training. A further problem is that experts in AI and machine learning, and digital technologies in general, are scarce in many of the LMICs and in the indigenous and refugee populations. This dearth of expertise, where present, exacerbates the digital divide. Constructive approaches for bridging the divide of course include recognition of the problem and acting on it. That and other, more-specific approaches are presented in Jones et al. (2017), Mollura et al. (2020), Troncoso (2020), Chakravorti (2021), Lesica et al. (2021), Saeed and Masters (2021), Wasmann et al. (2021), and Williams et al. (2021), and at https://www.youtube.com/watch?v=fzokRz1pgb0 (a TEDx talk by Jim Sevier on "Bridging the digital divide") and https://www.youtube.com/watch?v=IdBemHBN7xQ (a TEDx talk by Prashant Shukla on "Transforming

rural India using AI and digital technology"). Indeed, AI and machine learning when applied properly and respectfully (collaboratively) can greatly empower healthcare in LMICs and for other populations such as indigenous populations or populations that are far from tertiary care hospitals (including the sparsely populated regions in high-income countries), e.g., with accurate diagnoses from images or other data obtained via telemedicine where internet or smartphone access exists, as mentioned by Andrés Bur and his coauthors, and by Jonathan Cohen, in their sections in this Review.

In all, the range of the five present applications encompasses many aspects of otolaryngology & the communication sciences. The successes, in some cases great successes, provide existence proofs of the power of AI in these fields. Additionally, the lessons learned in the applications and the suggestions for additional applications (e.g., the application suggested by Fan-Gang Zeng and his coauthors to use AI to produce a close approximation to the normal patterns of discharge in the auditory nerve for persons with hearing loss) also point the way forward. Some broad shoulders on which others can stand have been provided by the authors of the middle sections in this Review. The future looks bright, with a high likelihood of an even greater diversity of applications and ever more powerful applications. AI has the potential to transform otolaryngology & the communication sciences, as shown by the present results and just as AI has already transformed many other fields including – for one example among the many – ophthalmology & vision science (Abràmoff et al. 2018; De Fauw et al. 2018; Hogarty et al. 2019; Anon. 2021; Benet and Pellicer-Valero 2021; Cheung et al. 2021; Lesica et al. 2021; Nuzzi et al. 2021; Scheetz et al. 2021).

ACKNOWLEDGMENTS

STATEMENTS AND DECLARATIONS

Conflict of Interest: The authors declare that they have no conflict of interest.

REFERENCES

ABRÀMOFF MD, LAVIN PT, BIRCH M, SHAH N, FOLK JC (2018) Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med 1:39. https://doi.org/10.1038/s41746-018-0040-6

ANGRICK M, HERFF C, MUGLER E, TATE MC, SLUTZKY MW, KRUSIENSKI DJ, SCHULTZ T (2019) Speech synthesis from ECoG using densely connected 3D convolutional neural networks. J Neural Eng 16(3):036019. https://doi.org/10.1088/1741-2552/ab0c59

ANGRICK M, OTTENHOFF MC, DIENER L, IVUCIC D, IVUCIC G, GOULIS S, SAAL J, COLON AJ, WAGNER L, KRUSIENSKI DJ, KUBBEN PL, SCHULTZ T, HERFF C (2021) Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. Commun Biol 4(1):1055-1055. https://doi.org/10.1038/s42003-021-02578-0

ANON. (2021) Listen to this. Nat Mach Intell 3(2):101. https://doi.org/10.1038/s42256-021-00313-2

ANUMANCHIPALLI GK, CHARTIER J, CHANG EF (2019) Speech synthesis from neural decoding of spoken sentences. Nature 568(7753):493. https://doi.org/10.1038/s41586-019-1119-1

ARMSTRONG AG, LAM CC, SABESAN S, LESICA NA (2021) Compression and amplification algorithms in hearing aids impair the selectivity of neural responses to speech. Nat Biomed Eng. https://doi.org/10.1038/s41551-021-00707-y

BENET D, PELLICER-VALERO OJ (2021) Artificial intelligence: the unstoppable revolution in ophthalmology. Surv Ophthalmol. https://doi.org/10.1016/j.survophthal.2021.03.003

BENZEGHIBA M, DE MORI R, DEROO O, DUPONT S, ERBES T, JOUVET D, FISSORE L, LAFACE P, MERTINS A, RIS C, ROSE R, TYAGI V, WELLEKENS C (2007) Automatic speech recognition and speech variability: a review. Speech Commun 49(10-11):763-786. https://doi.org/10.1016/j.specom.2007.02.006

BEREZUTSKAYA J, FREUDENBURG ZV, RAMSEY NF, GÜÇLÜ U, VAN GERVEN MAJ (2017) Modeling brain responses to perceived speech with LSTM networks. In: Duivesteijn W, Pechenizkiy M, Fletcher GHL, Menkovski V, Postma EJ, Vanschoren J, Van Der Putten P (eds) Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning. Technische Universiteit Eindhoven, pp 149-153. https://pure.tue.nl/ws/portalfiles/portal/72619856/benelearn_2017.pdf

BEUKELMAN DR, FAGER S, BALL L, DIETZ A (2007) AAC for adults with acquired neurological conditions: a review. Augment Altern Comm 23(3):230-242. https://doi.org/10.1080/07434610701553668

BOOTHROYD A (2004) Hearing aid accessories for adults: the remote FM microphone. Ear Hear 25(1):22-33. https://doi.org/10.1097/01.aud.0000111260.46595.ec

BOUCHARD KE, MESGARANI N, JOHNSON K, CHANG EF (2013) Functional organization of human sensorimotor cortex for speech articulation. Nature 495(7441):327-332. https://doi.org/10.1038/nature11911

BOURLARD HA, MORGAN N (1994) Connectionist speech recognition. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4615-3210-1

BRANCO MP, PELS EGM, SARS RH, AARNOUTSE EJ, RAMSEY NF, VANSTEENSEL MJ, NIJBOER F (2021) Brain-computer interfaces for communication: preferences of individuals with locked-in syndrome. Neurorehabil Neural Repair 35(3):267-279. https://doi.org/10.1177/1545968321989331

BROCA P (1861) Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). Bulletin et Memoires de La Societe Anatomique de Paris 6:330–357

BROOK CD, PLATT MP, RUSSELL K, GRILLONE GA, ALIPHAS A, NOORDZIJ JP (2015) Time to competency, reliability of flexible transnasal laryngoscopy by training level. Otolaryngol Head Neck Surg 152(5):843-850. https://doi.org/10.1177/0194599815572792

BROWMAN CP, GOLDSTEIN L (1992) Articulatory phonology: an overview. Phonetica 49(3-4):155-180. https://doi.org/10.1159/000261913

BRUMBERG JS, PITT KM, MANTIE-KOZLOWSKI A, BURNISON JD (2018) Brain-computer interfaces for augmentative and alternative communication: a tutorial. Am J Speech Lang Pathol 27(1):1-12. https://doi.org/10.1044/2017_AJSLP-16-0244

BRUMBERG JS, WRIGHT EJ, ANDREASEN DS, GUENTHER FH, KENNEDY PR (2011) Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. Front Neurosci 5:65. https://doi.org/10.3389/fnins.2011.00065

BUR AM, SHEW M, NEW J (2019) Artificial intelligence for the otolaryngologist: a state of the art review. Otolaryngol Head Neck Surg 160(4):603-611. https://doi.org/10.1177/0194599819827507

CAREY D, KRISHNAN S, CALLAGHAN MF, SERENO MI, DICK F (2017) Functional and quantitative MRI mapping of somatomotor representations of human supralaryngeal vocal tract. Cereb Cortex 27(1):265-278. https://doi.org/10.1093/cercor/bhw393

CARUANA R (1997) Multitask learning. Mach Learn 28(1):41-75. https://doi.org/10.1023/a:1007379606734

CHAKRABARTI S, SANDBERG HM, BRUMBERG JS, KRUSIENSKI DJ (2015) Progress in speech decoding from the electrocorticogram. Biomed Eng Lett 5(1):10-21. https://doi.org/10.1007/s13534-015-0175-1

CHAKRAVORTI B (2021) How to close the digital divide in the United States. Harv Bus Rev. https://hbr.org/2021/07/how-to-close-the-digital-divide-in-the-u-s (HBR Technology and Analytics online)

CHANG EF, ANUMANCHIPALLI GK (2019) Toward a speech neuroprosthesis. JAMA. https://doi.org/10.1001/jama.2019.19813 (Epub ahead of print)

CHARTIER J, ANUMANCHIPALLI GK, JOHNSON K, CHANG EF (2018) Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. Neuron 98(5):1042-1054. https://doi.org/10.1016/j.neuron.2018.04.031

CHEN SF, GOODMAN J (1999) An empirical study of smoothing techniques for language modeling. Comput Speech Lang 13(4):359-393. https://doi.org/10.1006/csla.1999.0128

CHERRY EC (1962) The cocktail party problem. Discovery March:32-35

CHEUNG R, CHUN J, SHEIDOW T, MOTOLKO M, MALVANKAR-MEHTA MS (2021) Diagnostic accuracy of current machine learning classifiers for age-related macular degeneration: a systematic review and meta-analysis. Eye (Lond). https://doi.org/10.1038/s41433-021-01540-y

CHO K, VAN MERRIENBOER B, GULCEHRE C, BAHDANAU D, BOUGARES F, SCHWENK H, BENGIO Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Stroudsburg, PA, pp 1724–1734. https://doi.org/10.3115/v1/d14-1179

CHUNG K, ZENG FG (2009) Using hearing aid adaptive directional microphones to enhance cochlear implant performance. Hear Res 250(1-2):27-37. https://doi.org/10.1016/j.heares.2009.01.005

COLLOBERT R, PUHRSCH C, SYNNAEVE G (2016) Wav2Letter: an end-to-end ConvNet-based speech recognition system. arXiv:1609.03193. https://arxiv.org/abs/1609.03193 (preprint)

CONANT DF, BOUCHARD KE, LEONARD MK, CHANG EF (2018) Human sensorimotor cortex control of directly measured vocal tract movements during vowel production. J Neurosci 38(12):2955-2966. https://doi.org/10.1523/JNEUROSCI.2382-17.2018

CUSHING H (1909) A note upon the faradic stimulation of the postcentral gyrus in conscious patients. Brain 32(1):44-53. https://doi.org/10.1093/brain/32.1.44

DAS N, ZEGERS J, VAN HAMME H, FRANCART T, BERTRAND A (2020) Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding. J Neural Eng 17(4):046039. https://doi.org/10.1088/1741-2552/aba6f8

DASH D, FERRARI P, DUTTA S, WANG J (2020) NeuroVAD: real-time voice activity detection from non-invasive neuromagnetic signals. Sensors (Basel, Switzerland) 20(8):2248. https://doi.org/10.3390/s20082248

DE FAUW J, LEDSAM JR, ROMERA-PAREDES B, NIKOLOV S, TOMASEV N, BLACKWELL S, ASKHAM H, GLOROT X, O'DONOGHUE B, VISENTIN D, VAN DEN DRIESSCHE G, LAKSHMINARAYANAN B, MEYER C, MACKINDER F, BOUTON S, AYOUB K, CHOPRA R, KING D, KARTHIKESALINGAM A, HUGHES CO, RAINE R, HUGHES J, SIM DA, EGAN C, TUFAIL A, MONTGOMERY H, HASSABIS D, REES G, BACK T, KHAW PT, SULEYMAN M, CORNEBISE J, KEANE PA, RONNEBERGER O (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24(9):1342-1350. https://doi.org/10.1038/s41591-018-0107-6

DENSEN P (2011) Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 122:48-58. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116346/

DICHTER BK, BRESHEARS JD, LEONARD MK, CHANG EF (2018) The control of vocal pitch in human laryngeal motor cortex. Cell 174(1):21-31.e29. https://doi.org/10.1016/j.cell.2018.05.016

DOV D, KOVALSKY SZ, ASSAAD S, COHEN J, RANGE DE, PENDSE AA, HENAO R, CARIN L (2021) Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. Med Image Anal 67:101814. https://doi.org/10.1016/j.media.2020.101814

DUDLEY H, RIESZ RR, WATKINS SSA (1939) A synthetic speaker. J Franklin Inst 227(6):739-764. https://doi.org/10.1016/s0016-0032(39)90816-1

DUDLEY H, TARNOCZY TH (1950) The speaking machine of Wolfgang von Kempelen. J Acoust Soc Am 22(2):151-166. https://doi.org/10.1121/1.1906583

DUTOIT T (1997) An introduction to text-to-speech synthesis. Springer Netherlands, Dordrecht, Netherlands. https://doi.org/10.1007/978-94-011-5730-8

EINHORN R (2017) Hearing aid technology for the 21st century: a proposal for universal wireless connectivity and improved sound quality. IEEE Pulse 8(2):25-28. https://doi.org/10.1109/mpul.2016.2647018

EMMOREY K, ALLEN JS, BRUSS J, SCHENKER N, DAMASIO H (2003) A morphometric analysis of auditory brain regions in congenitally deaf adults. Proc Natl Acad Sci U S A 100(17):10049-10054. https://doi.org/10.1073/pnas.1730169100

FELGOISE SH, ZACCHEO V, DUFF J, SIMMONS Z (2016) Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis. Amyotroph Lateral Scler Frontotemporal Degener 17(3-4):179-183. https://doi.org/10.3109/21678421.2015.1125499

FENG G, INGVALSON EM, GRIECO-CALUB TM, ROBERTS MY, RYAN ME, BIRMINGHAM P, BURROWES D, YOUNG NM, WONG PCM (2018) Neural preservation underlies speech improvement from auditory deprivation in young cochlear implant recipients. Proc Natl Acad Sci U S A 115(5):E1022-E1031. https://doi.org/10.1073/pnas.1717603115

FIEDLER L, WÖSTMANN M, GRAVERSEN C, BRANDMEYER A, LUNNER T, OBLESER J (2017) Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. J Neural Eng 14(3):036020. https://doi.org/10.1088/1741-2552/aa66dd

FOWLER CA, RUBIN PE, REMEZ RE, TURVEY MT (1980) Implications for speech production of a general theory of action. In: Butterworth B (ed) Language production. Academic Press, New York, pp 373-420

GERS FA, SCHMIDHUBER J, CUMMINS F (2000) Learning to forget: continual prediction with LSTM. Neural Comput 12(10):2451-2471. https://doi.org/10.1162/089976600300015015

GFELLER K, TURNER C, MEHR M, WOODWORTH G, FEARN R, KNUTSON JF, WITT S, STORDAHL J (2002) Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults. Cochlear Implants Int 3(1):29-53. https://doi.org/10.1179/cim.2002.3.1.29

GHOSH PK, NARAYANAN SS (2011) A subject-independent acoustic-to-articulatory inversion. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Publications, pp 4624-4627. https://doi.org/10.1109/icassp.2011.5947385

GOLD B, MORGAN N, ELLIS D (2011) Speech and audio signal processing: processing and perception of speech and music. 2nd edn. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118142882

GOVERDOVSKY V, VON ROSENBERG W, NAKAMURA T, LOONEY D, SHARP DJ, PAPAVASSILIOU C, MORRELL MJ, MANDIC DP (2017) Hearables: multimodal physiological in-ear sensing. Sci Rep 7(1):6948. https://doi.org/10.1038/s41598-017-06925-2

GRAVES A, MOHAMED A-R, HINTON G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE Publications, pp 6645-6649. https://doi.org/10.1109/icassp.2013.6638947

GUENTHER FH, BRUMBERG JS, WRIGHT EJ, NIETO-CASTANON A, TOURVILLE JA, PANKO M, LAW R, SIEBERT SA, BARTELS JL, ANDREASEN DS, EHIRIM P, MAO H, KENNEDY PR (2009) A wireless brain-machine interface for real-time speech synthesis. PLoS One 4(12):e8218. https://doi.org/10.1371/journal.pone.0008218

HEALY EW, DELFARAH M, JOHNSON EM, WANG D (2019) A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation. J Acoust Soc Am 145(3):1378. https://doi.org/10.1121/1.5093547

HENSCHKE CI, YANKELEVITZ DF, MATEESCU I, BRETTLE DW, RAINEY TG, WEINGARD FS (1997) Neural networks for the analysis of small pulmonary nodules. Clin Imaging 21(6):390-399. https://doi.org/10.1016/S0899-7071(97)81731-7

HERFF C, HEGER D, DE PESTERS A, TELAAR D, BRUNNER P, SCHALK G, SCHULTZ T (2015) Brain-to-text: decoding spoken phrases from phone representations in the brain. Front Neurosci 9:217. https://doi.org/10.3389/fnins.2015.00217

HERFF C, SCHULTZ T (2016) Automatic speech recognition from neural signals: a focused review. Front Neurosci 10:429. https://doi.org/10.3389/fnins.2016.00429

HINTON G (2021) How to represent part-whole hierarchies in a neural network. arXiv 2102.12627

HOCHREITER S, SCHMIDHUBER J (1997) Long short-term memory. Neural Comput 9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

HOGARTY DT, MACKEY DA, HEWITT AW (2019) Current state and future prospects of artificial intelligence in ophthalmology: a review. Clin Exp Ophthalmol 47(1):128-139. https://doi.org/10.1111/ceo.13381

JELINEK F (1976) Continuous speech recognition by statistical methods. Proc IEEE 64(4):532-556. https://doi.org/10.1109/proc.1976.10159

JONES L, JACKLIN K, O'CONNELL ME (2017) Development and use of health-related technologies in indigenous communities: critical review. J Med Internet Res 19(7):e256. https://doi.org/10.2196/jmir.7520

KANAS VG, MPORAS I, BENZ HL, SGARBAS KN, BEZERIANOS A, CRONE NE (2014) Real-time voice activity detection for ECoG-based speech brain machine interfaces. In: 2014 19th International Conference on Digital Signal Processing. IEEE Publications, pp 862-865. https://doi.org/10.1109/icdsp.2014.6900790

KIM S, HORI T, WATANABE S (2017) Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Publications, pp 4835-4839. https://doi.org/10.1109/icassp.2017.7953075

KLATT DH, KLATT LC (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. J Acoust Soc Am 87(2):820-857. https://doi.org/10.1121/1.398894

KNESER R, NEY H (1995) Improved backing-off for M-gram language modeling. In: 1995 International Conference on Acoustics, Speech, and Signal Processing. IEEE Publications, pp 181-184. https://doi.org/10.1109/icassp.1995.479394

KOCHKIN S (2007) MarkeTrak VII: obstacles to adult non-user adoption of hearing aids. The Hearing Journal 60(4):24-51. https://doi.org/10.1097/01.hj.0000285745.08599.7f

KOMEDA Y, HANDA H, WATANABE T, NOMURA T, KITAHASHI M, SAKURAI T, OKAMOTO A, MINAMI T, KONO M, ARIZUMI T, TAKENAKA M, HAGIWARA S, MATSUI S, NISHIDA N, KASHIDA H, KUDO M (2017) Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. Oncology 93(suppl 1):30-34. https://doi.org/10.1159/000481227

KOMURA D, ISHIKAWA S (2018) Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J 16:34-42. https://doi.org/10.1016/j.csbj.2018.01.001

KRAL A, KRONENBERGER WG, PISONI DB, O'DONOGHUE GM (2016) Neurocognitive factors in sensory restoration of early deafness: a connectome model. Lancet Neurol 15(6):610-621. https://doi.org/10.1016/S1474-4422(16)00034-X

KRIZHEVSKY A, SUTSKEVER I, HINTON GE (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems, vol 25. Curran Associates, Inc., Red Hook, NY, pp 1097-1105. https://doi.org/10.1145/3065386

KUMAR R, INDRAYAN A (2011) Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr 48(4):277-287. https://doi.org/10.1007/s13312-011-0055-4

LAEEQ K, PANDIAN V, SKINNER M, MASOOD H, STEWART CM, WEATHERLY R, CUMMINGS CW, BHATTI NI (2010) Learning curve for competency in flexible laryngoscopy. Laryngoscope 120(10):1950-1953. https://doi.org/10.1002/lary.21063

LAWRENCE S, GILES CL, AH CHUNG T, BACK AD (1997) Face recognition: a convolutional neural-network approach. IEEE Trans Neural Netw 8(1):98-113. https://doi.org/10.1109/72.554195

LECUN Y, BENGIO Y, HINTON G (2015) Deep learning. Nature 521(7553):436-444. https://doi.org/10.1038/nature14539

LESICA NA (2018) Why do hearing aids fail to restore normal auditory perception? Trends Neurosci 41(4):174-185. https://doi.org/10.1016/j.tins.2018.01.008

LESICA NA, MEHTA N, MANJALY JG, DENG L, WILSON BS, ZENG F-G (2021) Harnessing the power of artificial intelligence to transform hearing healthcare and research. Nat Mach Intell 3(10):840-849. https://doi.org/10.1038/s42256-021-00394-z

LING Z-H, RICHMOND K, YAMAGISHI J, WANG R-H (2009) Integrating articulatory features into HMM-based parametric speech synthesis. IEEE Trans Audio Speech Lang Process 17(6):1171-1185. https://doi.org/10.1109/tasl.2009.2014796

LINSE K, AUST E, JOOS M, HERMANN A (2018) Communication matters -- pitfalls and promise of hightech communication devices in palliative care of severely physically disabled patients with amyotrophic lateral sclerosis. Front Neurol 9:1-18. https://doi.org/10.3389/fneur.2018.00603

LIVEZEY JA, BOUCHARD KE, CHANG EF (2019) Deep learning as a tool for neural data analysis: speech classification and cross-frequency coupling in human sensorimotor cortex. PLoS Comput Biol 15(9):e1007091. https://doi.org/10.1371/journal.pcbi.1007091

LOIZOU PC (2013) Speech enhancement: theory and practice. 2nd edn. CRC Press, Boca Raton, FL, USA. https://doi.org/10.1201/b14529

LONGONI C, BONEZZI A, MOREWEDGE CK (2019) Resistance to medical artificial intelligence. J Consum Res 46(4):629-650. https://doi.org/10.1093/jcr/ucz013

LOTTE F, BRUMBERG JS, BRUNNER P, GUNDUZ A, RITACCIO AL, GUAN C, SCHALK G (2015) Electrocorticographic representations of segmental features in continuous speech. Front Hum Neurosci 9:97. https://doi.org/10.3389/fnhum.2015.00097

MAKIN JG, MOSES DA, CHANG EF (2020) Machine translation of cortical activity to text with an encoder–decoder framework. Nat Neurosci 23(4):575-582. https://doi.org/10.1038/s41593-020-0608-8

MARTIN S, ITURRATE I, MILLÁN JDR, KNIGHT RT, PASLEY BN (2018) Decoding inner speech using electrocorticography: progress and challenges toward a speech prosthesis. Front Neurosci 12:422. https://doi.org/10.3389/fnins.2018.00422

MECKLENBURGER J, GROTH T (2016) Wireless technologies and hearing aid connectivity. In: Popelka GR, Moore BCJ, Fay RR, Popper AN (eds) Hearing aids, vol 56. Springer Handbook of Auditory Research (SHAR). Springer International Publishing, Switzerland, pp 131-149. https://doi.org/10.1007/978-3-319-33036-5_5

MEHRA R, BRIMIJOIN O, ROBINSON P, LUNNER T (2020) Potential of augmented reality platforms to improve individual hearing aids and to support more ecologically valid research. Ear Hear 41(Suppl 1):140S-146S. https://doi.org/10.1097/AUD.0000000000000961

MERZENICH MM (2011) Michael M. Merzenich. In: Squire LR (ed) The history of neuroscience in autobiography, vol 7. Oxford University Press, Oxford, UK, pp 440-476. https://doi.org/10.1093/acprof:oso/9780195396133.003.0010

MICHIE D, SPIEGELHALTER DJ, TAYLOR CC (1994) Machine learning, neural and statistical classification. Ellis Horwood, New York, NY, USA. https://www1.maths.leeds.ac.uk/~charles/statlog/whole.pdf

MITRA V, SIVARAMAN G, BARTELS C, NAM H, WANG W, ESPY-WILSON C, VERGYRI D, FRANCO H (2017) Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Publications, pp 5205-5209. https://doi.org/10.1109/icassp.2017.7953149

MOHAMED A-R, DAHL GE, HINTON G (2012) Acoustic modeling using deep belief networks. IEEE Trans Audio Speech Lang Process 20(1):14-22. https://doi.org/10.1109/tasl.2011.2109382

MOLLURA DJ, CULP MP, POLLACK E, BATTINO G, SCHEEL JR, MANGO VL, ELAHI A, SCHWEITZER A, DAKO F (2020) Artificial intelligence in low- and middle-income countries: innovating global health radiology. Radiology 297(3):513-520. https://doi.org/10.1148/radiol.2020201434

MOSES DA, LEONARD MK, MAKIN JG, CHANG EF (2019) Real-time decoding of question-and-answer speech dialogue using human cortical activity. Nat Commun 10(1):3096. https://doi.org/10.1038/s41467-019-10994-4

MOSES DA, METZGER SL, LIU JR, ANUMANCHIPALLI GK, MAKIN JG, SUN PF, CHARTIER J, DOUGHERTY ME, LIU PM, ABRAMS GM, TU-CHAN A, GANGULY K, CHANG EF (2021) Neuroprosthesis for decoding speech in a paralyzed person with anarthria. N Engl J Med 385(3):217-227. https://doi.org/10.1056/nejmoa2027540

MUGLER EM, PATTON JL, FLINT RD, WRIGHT ZA, SCHUELE SU, ROSENOW J, SHIH JJ, KRUSIENSKI DJ, SLUTZKY MW (2014) Direct classification of all American English phonemes using signals from functional speech motor cortex. J Neural Eng 11(3):035015. https://doi.org/10.1088/1741-2560/11/3/035015

NACHMANI E, ADI Y, WOLF L Voice separation with an unknown number of multiple speakers. In: Hal D, Iii, Aarti S (eds) Proceedings of the 37th International Conference on Machine Learning. Proc Mach Learn Res (PMLR) 119:7164-7175. https://proceedings.mlr.press/v119/nachmani20a.html

NING Y, HE S, WU Z, XING C, ZHANG L-J (2019) A review of deep learning based speech synthesis. Appl Sci 9(19):4050. https://doi.org/10.3390/app9194050

NIP I, ROTH CR (2017) Anarthria. In: Kreutzer J, Deluca J, Caplan B (eds) Encyclopedia of clinical neuropsychology. Springer International Publishing. https://doi.org/10.1007/978-3-319-56782-2_855-4

NIPARKO JK, TOBEY EA, THAL DJ, EISENBERG LS, WANG N-Y, QUITTNER AL, FINK NE, TEAM CDI (2010) Spoken language development in children following cochlear implantation. JAMA 303(15):1498-1506. https://doi.org/10.1001/jama.2010.451

NITTROUER S, CALDWELL-TARR A (2016) Language and literacy skills in children with cochlear implants: past and present findings. In: Young NM, Kirk KI (eds) Pediatric cochlear implantation. Springer, New York, NY, pp 177-197. https://doi.org/10.1007/978-1-4939-2788-3_11

NUZZI R, BOSCIA G, MAROLO P, RICARDI F (2021) The impact of artificial intelligence and deep learning in eye diseases: a review. Front Med (Lausanne) 8:710329. https://doi.org/10.3389/fmed.2021.710329

O'SULLIVAN J, CHEN Z, HERRERO J, MCKHANN GM, SHETH SA, MEHTA AD, MESGARANI N (2017) Neural decoding of attentional selection in multi-speaker environments without access to clean sources. J Neural Eng 14(5):056001. https://doi.org/10.1088/1741-2552/aa7ab4

OBERMEYER Z, EMANUEL EJ (2016) Predicting the future -- big data, machine learning, and clinical medicine. N Engl J Med 375(13):1216-1219. https://doi.org/10.1056/NEJMp1606181

OORD A, VAN DEN D, SANDER, ZEN H, SIMONYAN K, VINYALS O, GRAVES A, KALCHBRENNER N, SENIOR A, KAVUKCUOGLU K (2016) WaveNet: a generative model for raw audio. arXiv:1609.03499. http://arxiv.org/abs/1609.03499 (preprint)

Oxley TJ, Yoo PE, Rind GS, Ronayne SM, Lee CMS, Bird C, Hampshire V, Sharma RP, Morokoff A, Williams DL, MacIsaac C, Howard ME, Irving L, Vrljic I, Williams C, John SE, Weissenborn F, Dazenko M, Balabanski AH, Friedenberg D, Burkitt AN, Wong YT, Drummond KJ, Desmond P, Weber D, Denison T, Hochberg LR, Mathers S, O'Brien TJ, May CN, Mocco J, Grayden DB, Campbell BCV, Mitchell P, Opie NL (2021) Motor neuroprosthesis implanted with neurointerventional surgery improves capacity for activities of daily living tasks in severe paralysis: first in-human experience. J Neurointerv Surg 13(2):102-108. https://doi.org/10.1136/neurintsurg-2020-016862

Pandarinath C, Nuyujukian P, Blabe CH, Sorice BL, Saab J, Willett FR, Hochberg LR, Shenoy KV, Henderson JM (2017) High performance communication by people with paralysis using an intracortical brain-computer interface. eLife 6:e18554. https://doi.org/10.7554/elife.18554

Pendleton C, Zaidi HA, Chaichana KL, Raza SM, Carson BS, Cohen-Gadol AA, Quinones-Hinojosa A (2012) Harvey Cushing's contributions to motor mapping: 1902–1912. Cortex 48(1):7-14. https://doi.org/10.1016/j.cortex.2010.04.006

Penfield W, Boldrey E (1937) Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. Brain 60(4):389-443. https://doi.org/10.1093/brain/60.4.389

Penfield W, Rasmussen T (1950) The cerebral cortex of man: clinical study of localization of function. Macmillan, New York, NY

Perrachione TK, Lee J, Ha LYY, Wong PCM (2011) Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. J Acoust Soc Am 130(1):461-472. https://doi.org/10.1121/1.3593366

Perret E (2017) Here's how many photos will be taken in 2017. Tech Today. https://focus.mylio.com/tech-today/heres-how-many-digital-photos-will-be-taken-in-2017-repost-oct.

Peterson SM, Steine-Hanson Z, Davis N, Rao RPN, Brunton BW (2021) Generalized neural decoders for transfer learning across participants and recording modalities. J Neural Eng 18(2):026014. https://doi.org/10.1088/1741-2552/abda0b

Pratt LU, Mostow J, Kamm CA (1991) Direct transfer of learned information among neural networks. In: Ninth National Conference on Artificail Intelligence. AAAI Press, Menlo Park, CA, pp 584-589. https://www.aaai.org/Papers/AAAI/1991/AAAI91-091.pdf

Rabbani Q, Milsap G, Crone NE (2019) The potential for a speech brain-computer interface using chronic electrocorticography. Neurotherapeutics 16(1):144-165. https://doi.org/10.1007/s13311-018-00692-2

Ren J, Jing X, Wang J, Ren X, Xu Y, Yang Q, Ma L, Sun Y, Xu W, Yang N, Zou J, Zheng Y, Chen M, Gan W, Xiang T, An J, Liu R, Lv C, Lin K, Zheng X, Lou F, Rao Y, Yang H, Liu K, Liu G, Lu T, Zheng X, Zhao Y (2020) Automatic recognition of laryngoscopic images using a deep-learning technique. Laryngoscope 130(11). https://doi.org/10.1002/lary.28539

REZAZADEH SERESHKEH A, TROTT R, BRICOUT A, CHAU T (2017) EEG classification of covert speech using regularized neural networks. IEEE Trans Audio Speech Lang Process 25(12):2292-2300. https://doi.org/10.1109/taslp.2017.2758164

RICHMOND K (2002) Estimating articulatory parameters from the acoustic speech signal. Ph.D. dissertation, University of Edinburgh.

ROBERTS MY (2019) Parent-implemented communication treatment for infants and toddlers with hearing loss: a randomized pilot trial. J Speech Lang Hear Res 62(1):143-152. https://doi.org/10.1044/2018_JSLHR-L-18-0079

ROUSSEAU M-C, BAUMSTARCK K, ALESSANDRINI M, BLANDIN V, BILLETTE DE VILLEMEUR T, AUQUIER P (2015) Quality of life in patients with locked-in syndrome: Evolution over a 6-year period. Orphanet J Rare Dis 10:88. https://doi.org/10.1186/s13023-015-0304-z

SAEED SA, MASTERS RM (2021) Disparities in health care and the digital divide. Curr Psychiatry Rep 23(9):61. https://doi.org/10.1007/s11920-021-01274-4

SALARI E, FREUDENBURG ZV, BRANCO MP, AARNOUTSE EJ, VANSTEENSEL MJ, RAMSEY NF (2019) Classification of articulator movements and movement direction from sensorimotor cortex activity. Sci Rep 9(1):14165-14165. https://doi.org/10.1038/s41598-019-50834-5

SARKER IH (2021) Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2(3):160. https://doi.org/10.1007/s42979-021-00592-x

SCHAWORONKOW N, VOYTEK B (2021) Enhancing oscillations in intracranial electrophysiological recordings with data-driven spatial filters. bioRxiv. https://doi.org/10.1101/2021.02.26.433127

SCHEETZ J, ROTHSCHILD P, MCGUINNESS M, HADOUX X, SOYER HP, JANDA M, CONDON JJJ, OAKDEN-RAYNER L, PALMER LJ, KEEL S, VAN WIJNGAARDEN P (2021) A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. Sci Rep 11(1):5193. https://doi.org/10.1038/s41598-021-84698-5

SCHÖNLE PW, GRÄBE K, WENIG P, HÖHNE J, SCHRADER J, CONRAD B (1987) Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. Brain Lang 31(1):26-35. https://doi.org/10.1016/0093-934x(87)90058-7

SELLERS EW, RYAN DB, HAUSER CK (2014) Noninvasive brain-computer interface enables communication after brainstem stroke. Sci Transl Med 6(257):257re257. https://doi.org/10.1126/scitranslmed.3007801

SHEN J, PANG R, WEISS RJ, SCHUSTER M, JAITLY N, YANG Z, CHEN Z, ZHANG Y, WANG Y, SKERRV-RYAN R, SAUROUS RA, AGIOMVRGIANNAKIS Y, WU Y (2018) Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Publications, pp 4779-4783. https://doi.org/10.1109/icassp.2018.8461368

SHIBATA DK (2007) Differences in brain structure in deaf persons on MR imaging studied with voxel-based morphometry. Am J Neuroradiol 28(2):243-249. https://www.ncbi.nlm.nih.gov/pubmed/17296987

SHICHIJO S, NOMURA S, AOYAMA K, NISHIKAWA Y, MIURA M, SHINAGAWA T, TAKIYAMA H, TANIMOTO T, ISHIHARA S, MATSUO K, TADA T (2017) Application of convolutional neural networks in the diagnosis of *helicobacter pylori* infection based on endoscopic images. EBioMedicine 25:106-111. https://doi.org/10.1016/j.ebiom.2017.10.014

SLANEY M, LYON RF, GARCIA R, KEMLER B, GNEGY C, WILSON K, KANEVSKY D, SAVLA S, CERF VG (2020) Auditory measures for the next billion users. Ear Hear 41(Supplement 1):131S-139S. https://doi.org/10.1097/aud.0000000000000955

SMITH KM, MECOLI MD, ALTAYE M, KOMLOS M, MAITRA R, EATON KP, EGELHOFF JC, HOLLAND SK (2011) Morphometric differences in the Heschl's gyrus of hearing impaired and normal hearing infants. Cereb Cortex 21(5):991-998. https://doi.org/10.1093/cercor/bhq164

SOLLICH P, KROGH A (1996) Learning with ensembles: how overfitting can be useful. In: Touretzky DS, Mozer MC, Hasselmo ME (eds) Advances in neural information processing systems 8. MIT Press, Cambridge, MA, pp 190-196. http://papers.nips.cc/paper/1044-learning-with-ensembles-how-overfitting-can-be-useful.pdf

SUN P, ANUMANCHIPALLI GK, CHANG EF (2020) Brain2Char: a deep architecture for decoding text from brain recordings. J Neural Eng 17(6):066015. https://doi.org/10.1088/1741-2552/abc742

SZEGEDY C, WEI L, YANGQING J, SERMANET P, REED S, ANGUELOV D, ERHAN D, VANHOUCKE V, RABINOVICH A (2015) Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Publications, pp 1-9. https://doi.org/10.1109/cvpr.2015.7298594

TAKIYAMA H, OZAWA T, ISHIHARA S, FUJISHIRO M, SHICHIJO S, NOMURA S, MIURA M, TADA T (2018) Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. Sci Rep 8(1):7497. https://doi.org/10.1038/s41598-018-25842-6

TRONCOSO EL (2020) The greatest challenge to using AI/ML for primary health care: mindset or datasets? Front Artif Intell 3:53. https://doi.org/10.3389/frai.2020.00053

VANSTEENSEL MJ, PELS EGM, BLEICHNER MG, BRANCO MP, DENISON T, FREUDENBURG ZV, GOSSELAAR P, LEINDERS S, OTTENS TH, VAN DEN BOOM MA, VAN RIJEN PC, AARNOUTSE EJ, RAMSEY NF (2016) Fully implanted brain-computer interface in a locked-in patient with ALS. N Engl J Med 375(21):2060-2066. https://doi.org/10.1056/NEJMoa1608085

VAS V, AKEROYD MA, HALL DA (2017) A data-driven synthesis of research evidence for domains of hearing loss, as reported by adults with hearing loss and their communication partners. Trend Hear 21:2331216517734088. https://doi.org/10.1177/2331216517734088

VITERBI A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 13(2):260-269. https://doi.org/10.1109/tit.1967.1054010

VON KEMPELEN W (1791) Mechanismus der menschlichen sprache nebst beschreibung einer sprechenden maschine. J. B. Degan, Vienna, Austria. https://doi.org/10.6083/sx61dm64r

WANG D (2017) Deep learning reinvents the hearing aid: finally, wearers of hearing aids can pick out a voice in a crowded room. IEEE Spectr 54(3):32-37. https://doi.org/10.1109/MSPEC.2017.7864754

WANG D, KHOSLA A, GARGEYA R, IRSHAD H, BECK AH (2016) Deep learning for identifying metastatic breast cancer. arXiv:1606.05718. https://arxiv.org/abs/1606.05718

WANG D, WANG X, LV S (2019) An overview of end-to-end automatic speech recognition. Symmetry 11(8):1018. https://doi.org/10.3390/sym11081018

WANG N-Y, EISENBERG LS, JOHNSON KC, FINK NE, TOBEY EA, QUITTNER AL, NIPARKO JK, TEAM CDI (2008) Tracking development of speech recognition: longitudinal data from hierarchical assessments in the Childhood Development after Cochlear Implantation Study. Otol Neurotol 29(2):240-245. https://doi.org/10.1097/MAO.0b013e3181627a37

WANG Y, SKERRY-RYAN RJ, STANTON D, WU Y, WEISS RJ, JAITLY N, YANG Z, XIAO Y, CHEN Z, BENGIO S, LE Q, AGIOMYRGIANNAKIS Y, CLARK R, SAUROUS RA (2017) Tacotron: towards end-to-end speech synthesis. In: Proc. Interspeech 2017. International Speech Communication Association (ISCA), Grenoble, France, pp 4006-4010. https://doi.org/10.21437/interspeech.2017-1452

WASMANN J-WA, LANTING CP, HUINCK WJ, MYLANUS EAM, VAN DER LAAK JWM, GOVAERTS PJ, SWANEPOEL DW, MOORE DR, BARBOUR DL (2021) Computational audiology: new approaches to advance hearing health care in the digital age. Ear Hear 42(6):1499-1507. https://doi.org/10.1097/aud.0000000000001041

WATANABE S, DELCROIX M, METZE F, HERSHEY JR (eds) (2017) New era for robust speech recognition: exploiting deep learning. Springer-Verlag, Berlin. https://doi.org/10.1007/978-3-319-64680-0

WILLETT FR, AVANSINO DT, HOCHBERG LR, HENDERSON JM, SHENOY KV (2021) High-performance brain-to-text communication via handwriting. Nature 593(7858):249-254. https://doi.org/10.1038/s41586-021-03506-2

WILLIAMS D, HORNUNG H, NADIMPALLI A, PEERY A (2021) Deep learning and its application for healthcare delivery in low and middle income countries. Front Artif Intell 4:553987. https://doi.org/10.3389/frai.2021.553987

WILSON BS, DORMAN MF, WOLDORFF MG, TUCCI DL (2011) Cochlear implants: matching the prosthesis to the brain and facilitating desired plastic changes in brain function. Prog Brain Res 194:117-129. https://doi.org/10.1016/B978-0-444-53815-4.00012-1

WILSON BS, TUCCI DL (2021) Addressing the global burden of hearing loss. Lancet 397(10278):945-947. https://doi.org/10.1016/s0140-6736(21)00522-5

WOLPAW JR, BIRBAUMER N, MCFARLAND DJ, PFURTSCHELLER G, VAUGHAN TM (2002) Brain–computer interfaces for communication and control. Clin Neurophysiol 113(6):767-791. https://doi.org/10.1016/s1388-2457(02)00057-3

WONG PCM, PERRACHIONE TK, PARRISH TB (2007) Neural characteristics of successful and less successful speech and word learning in adults. Hum Brain Mapp 28(10):995-1006. https://doi.org/10.1002/hbm.20330

WONG PCM, VUONG LC, LIU K (2017) Personalized learning: from neurogenetics of behaviors to designing optimal language training. Neuropsychologia 98:192-200. https://doi.org/10.1016/j.neuropsychologia.2016.10.002

YOUNG NM, KIM FM, RYAN ME, TOURNIS E, YARAS S (2012) Pediatric cochlear implantation of children with eighth nerve deficiency. Int J Pediatr Otorhinolaryngol 76(10):1442-1448. https://doi.org/10.1016/j.ijporl.2012.06.019

YU MK, MA J, FISHER J, KREISBERG JF, RAPHAEL BJ, IDEKER T (2018) Visible machine learning for biomedicine. Cell 173(7):1562-1565. https://doi.org/10.1016/j.cell.2018.05.056

ZE H, SENIOR A, SCHUSTER M (2013) Statistical parametric speech synthesis using deep neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE Publications, pp 7962-7966. https://doi.org/10.1109/icassp.2013.6639215

ZENG F-G (2017) Challenges in improving cochlear implant performance and accessibility. IEEE Trans Biomed Eng 64(8):1662-1664. https://doi.org/10.1109/TBME.2017.2718939

ZHANG Y, CHAN W, JAITLY N (2017) Very deep convolutional networks for end-to-end speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Publications, pp 4845-4849. https://doi.org/10.1109/icassp.2017.7953077

FIGURE CAPTIONS

FIG. 1. Flyer for the symposium

FIG. 2. Program for the symposium

FIG. 3. Speech decoding approaches using brain activity and artificial intelligence. To decode speech from neural activity in the brain, two general types of approaches can be used: an approach to synthesize speech or an approach to decode text. In both approaches, neural signals are acquired from a human user and the relevant features are extracted from these signals. To synthesize speech, articulatory kinematic features can be inferred from the neural activity, and then these features can be used to synthesize speech. To decode text, sub-word (for example, phonetic) features can be inferred from the neural activity, and then language modeling techniques can be used to decode text from these features. For either approach, neural features can be mapped directly to the target output using end-to-end modeling. It is possible to convert decoded text into synthesized speech using text-to-speech synthesis and vice versa using speech recognition. Artificial intelligence techniques can be used to enable or improve the quality of each step in this schematic

FIG. 4. Artificial intelligence techniques for speech brain-computer interface (BCI) applications. Schematic depictions and brief descriptions are provided for artificial intelligence (AI) techniques to model speech-related neural activity. This is not an exhaustive compilation of all relevant AI techniques; multiple variants of a depicted technique may exist, and the techniques are depicted in an arbitrary order. The depicted techniques (and references for each technique) are: recurrent neural network modeling (Hochreiter and Schmidhuber 1997; Gers et al. 2000; Berezutskaya et al. 2017; Anumanchipalli et al. 2019; Makin et al. 2020; Sun et al. 2020; Moses et al. 2021), temporal convolution (Zhang et al. 2017; Makin et al. 2020; Moses et al. 2021), end-to-end network modeling (Graves et al. 2013; Collobert et al. 2016; Oord et al. 2016; Kim et al. 2017; Wang et al. 2017; Zhang et al. 2017; Wang et al. 2019; Makin et al. 2020; Sun et al. 2020), data augmentation (Krizhevsky et al. 2012; Moses et al. 2021), model ensembling (Sollich and Krogh 1996; Szegedy et al. 2015; Moses et al. 2021), multitask learning (Caruana 1997; Szegedy et al. 2015; Kim et al. 2017; Makin et al. 2020; Sun et al. 2020), and transfer learning (Pratt et al. 1991; Caruana 1997; Makin et al. 2020; Peterson et al. 2021)

FIG. 5. Accuracy of prediction by neural predictors and traditional characteristics. Panel **A** shows the median split of subjects' improvement on the Speech Recognition Index in Quiet (SRI-Q) test. Panel **B** presents a comparison of prediction accuracy (Acc), sensitivity (Sens), specificity (spec), and area under the curve (AUC), which combines sensitivity and specificity, for brain-based models versus the characteristics of age at implant and residual hearing. Panel **A** is from Feng et al. (2018) and is reproduced here with permission

FIG. 6. Brain connectivity predictive area mapping from diffusor tensor imaging (DTI) scans. Probability maps of combined fractional anisotropy, and radial and axial diffusivity, that demonstrate little overlap of brain areas predicting baseline and improvement at six months (blue = baseline; green = six-month improvement; red = overlap); coronal slices of the brain are shown

FIG. 7. Three artificial intelligence approaches to addressing the hearing-in-noise complaint by people with hearing loss; the label "Either" refers to "Yes" or "No"

FIG. 8. How artificial intelligence might restore normal or nearly normal hearing in hearing aid or cochlear implant users (Modified from Lesica, 2018, and presented here with permission)

FIG. 9. Loss and accuracy functions for training of a 12-layer convolutional neural network to classify laryngeal images as normal versus cancerous; acc = accuracy and epochs indicate iterations in the training of the network

FIG. 10. Normalized confusion matrix showing performance on a test set of 36 images of a convolutional neural network to classify laryngeal images as normal, benign, or suspicious for malignancy

FIG. 11. The top panel shows a whole-slide cytopathology scan in which the image was produced with a stained smear of cells and other tissue obtained from a fine needle aspiration (FNA) biopsy of a thyroid nodule and the middle panel shows a heat map of predictions of regions of interest (ROIs) identified by the first of the two machine learning algorithms (MLAs) developed in the present study. The bottom panels show magnified images corresponding to the red rectangles in the top and middle panels. The figure is from Dov et al., 2021, and is reproduced here with permission

FIG. 12. Receiver operating characteristic (ROC) curves for three expert pathologists (Experts 1-3), the expert pathologists whose reports were included in the electronic medical records for the patients (MR), and the cascade of the two machine learning algorithms (Proposed). The experts and the algorithms used the five diagnostic categories II-VI of The Bethesda System (TBS) as outcome measures (the first category is a non-diagnostic category). The areas under the ROC curves (auc) indicate performance. TPR = True Positive Rate (or sensitivity) and FPR = False Positive Rate (or 1.0 – specificity). The illustration is from Dov et al. (2021) and is reproduced here with permission. A description of ROC curves and their meanings is presented in Kumar and Indrayan (2011)