A range of new technologies have the potential to help people, whether traditionally considered hearing impaired or not. These technologies include more sophisticated personal sound amplification products, as well as real-time speech enhancement and speech recognition. They can improve user's communication abilities, but these new approaches require new ways to describe their success and allow engineers to optimize their properties. Speech recognition systems are often optimized using the word-error rate, but when the results are presented in real time, user interface issues become a lot more important than conventional measures of auditory performance. For example, there is a tradeoff between minimizing recognition time (latency) by quickly displaying results versus disturbing the user's cognitive flow by rewriting the results on the screen when the recognizer later needs to change its decisions. This article describes current, new, and future directions for helping billions of people with their hearing. These new technologies bring auditory assistance to new users, especially to those in areas of the world without access to professional medical expertise. In the short term, audio enhancement technologies in inexpensive mobile forms, devices that are quickly becoming necessary to navigate all aspects of our lives, can bring better audio signals to many people. Alternatively, current speech recognition technology may obviate the need for audio amplification or enhancement at all and could be useful for listeners with normal hearing or with hearing loss. With new and dramatically better technology based on deep neural networks, speech enhancement improves the signal to noise ratio, and audio classifiers can recognize sounds in the user's environment. Both use deep neural networks to improve a user's experiences. Longer term, auditory attention decoding is expected to allow our devices to understand where a user is directing their attention and thus allow our devices to respond better to their needs. In all these cases, the technologies turn the hearing assistance problem on its head, and thus require new ways to measure their performance.

**Key words:** Audiology, Auditory attention decoding, EEG, HRTF, Listening effort, MEG, PSAP, Speech enhancement, Speech recognition.

**Abbreviations:** AAD, Auditory attention decoding; DNN, Deep neural network; EEG, Electroencephalography; ML, Machine learning; NBU, Next billion user; NR, Noise reduction; PCA, Principal component analysis; PSAP, Personal sound-amplification products; RL, Reinforcement learning; SNR, Signal to noise ratio; UI, User interface

(Ear & Hearing 2020;41:131S–139S)

**INTRODUCTION**

Mobile phones with powerful computer and always-on connections to the global Internet are common in many parts of the world, providing useful services to their users, but these technologies have not reached everybody. If the current technology is widely available to billions of users, then the next frontier is to provide the same technology to the next billion users (NBUs).

Arora (2019). These users typically have fewer financial resources, as well as less support for technical or medical issues. A device the user already owns because it satisfies a number of other needs is an excellent platform for users who need help with their hearing.

At the same time, accessibility is becoming more important for all of us. Everyone may experience limited perceptual abilities, whether it is due to pathology (trauma, genetics, or aging), or ambient acoustic circumstances. We all wish we could communicate better in noisy environments. Thus, we see better communication technologies as important not only to the NBUs, but also to the first billions and the last billion users, too. In this article, we emphasize the NBU.

The widespread availability of devices that recognize the speech and other audio around us, whether we carry them on our heads or in our pockets, will change the nature of users' lives and how we characterize their devices' performance. In hearing science, ecological validity refers to the degree to which research findings reflect real-life hearing-related function, activity, or participation (Keidser et al. 2020). The goal of this article is to describe new technology to help people communicate better and how this might change the type of metrics we use to determine scientific or commercial success. These new technologies allow a different set of solutions, more than just hearing aids, and thus drive the need for new metrics. In addition, since these devices are often fully connected to the Internet, their real-time performance metrics have high ecological validity—throughout their day a user must decide whether a tool is helping them—and these data can guide research and development.

The latest breakthroughs in machine learning (ML) are key to some of these new solutions. State-of-the-art mobile phones have computing resources that dwarf the supercomputers of 35 years ago, along with high-quality transducers (microphones, cameras, etc) and easy connectivity with peripherals (e.g., Bluetooth devices). These resources are used to improve the user experience, entertain the user, connect them to their friends, and answer their questions. Increases in computer resources, both on mobiles and in the cloud, allow mobile devices to recognize objects in images, as well as speech and other sounds. It is quite common now for modern ML and artificial intelligence systems to demonstrate superhuman abilities (Bishop 2017), often performing image and sound recognition with error rates better than humans. While all these capabilities do not yet run on the phones in our pockets, it is clearly the direction in which the technology is moving.

Finally, personal electronics are becoming ubiquitous. A walk through any city shows many people displaying their auditory accessories as a fashion statement, and these devices have more processing power than conventional hearing aids, which due to their latency requirements and battery technology are limited. These new mobile technologies can distribute the computation on the body (ears, pockets, or purses), and the machines on the Internet and thus are freed from the need for batteries that are...
Upon which many manufacturers build their mobile devices. This software provides a standard set of tools for multiband filtering, compression, limiting, and equalization. These signal-processing tools are available for all developers to use in their own applications, as well as being used to build our Sound Amplifier application. This application enhances the sounds captured by a microphone to provide a more comfortable and natural listening experience and plays it to the user over headphones. Likewise, National Institutes of Health’s “Open-Source Audio-Processing Tools for Hearing Research” project (NIH 2014) funded efforts to create state-of-the-art hardware and software to further accelerate these efforts.

A multiband compression system, such as the dynamics processing effects, requires dozens, if not hundreds of parameters. Setting such a large number of parameters is far from easy, especially for users without signal processing or audiology backgrounds. The Sound Amplifier application, an accessibility application for Android, manages this complexity by connecting scores of parameters to a small number of user controls.

Sound Amplifier’s settings are determined by two user interface (UI) sliders connected to equalization and compression parameters derived from hundreds of audiograms (of people with and without hearing loss). The design of these sliders was based on the set of 60 “typical” audiograms published by Bisgaard et al. (2010), which had been derived by a k-means algorithm from a database of more than 28,000 audiograms. Given the audiograms summarized by these 60 clusters, we can perform principal component analysis (PCA) to find the principal components that account for most of the variance. PCA components are calculated so that the top k components create the best (lowest error) approximation of the original data. Figure 1B shows the amount of error as we vary the number of PCA components used to approximate the 60 audiogram cluster centroids. The error goes down as we increase the number of components, but we are most interested in the error when using two (or maybe 3) components because this is more manageable for a user, see Keidser et al. (2007) or Dreschler et al. (2008). Figure 1A also shows three typical audiograms and the resulting two-component PCA approximation. This approximation will not help with all audiograms, such as “cookie bite” or U-shaped audiograms. More importantly, the underlying Android software can implement any desired filter shape, and developers can implement any desired fitting procedure, all of which can run on any Android phone.

In a smartphone application, we ask the user to control the amplifier using two parameters, implemented as sliders in a mobile application, easily sweeping over possible filters to find the most useful settings. One slider is labeled “boost,” representing the overall slope of the hearing loss as determined by the first PCA component. The other is labeled “fine tuning” and is connected to the second PCA component. These two sliders represent a large fraction of potential audiograms, and for each corresponding audiogram, we design compression and equalization settings, implementing wide-range dynamic compression, to compress the incoming audio to fit a prototypical user’s range of hearing. The settings are translated into filter settings and the user can hear the difference instantaneously. This translation can be an approximation because the user controls the effect, based on what they hear, not based on a premeasured audiogram.

Users can find their own settings (Boothroyd & Mackersie 2017), but expecting users to set the optimum parameters every time their environment or listening conditions change is
Hearing aids have evolved to be highly optimized, power-efficient, and nearly invisible devices that fit snugly on or inside the ear. They tend to be more concealable, less updateable, and offer much longer battery compared to smartphone-powered hearing apps. Smartphone apps, when paired with consumer headphones, offer comparatively unlimited processing power and flexibility to update and control the hearing experience in real time. That stated, challenges such as trying to reduce the level of audio latency to transparency remain as long as smartphones rely on off-the-shelf protocols such as Bluetooth, which was not originally designed to be either low power or low latency. As technology evolves, smartphone-based hearing solutions are likely to improve, but may never reach full experience parity with purpose-built devices.

Objective measures of auditory function, such as hearing thresholds, allow an audiologist to fit a hearing aid. Without professional assistance, or after the user returns home, a hearing aid user must find their own best settings, or fine-tune the aid to best help in a new environment. However, several studies have shown that listeners who have hearing loss have difficulties consistently selecting a preferred setting for different listening situations (e.g., Keidser et al. 2008), in particular when the listening situation becomes more complex and includes speech (Walravens et al. 2020). It remains to be seen how the NBU will find their best settings and which objective measures can help designers optimize the UI of our devices. Reinforcement learning (RL), described at the end of this article, is one option for long-term customization of future devices.

At the other end of the scale, especially for those with profound hearing issues, closed captioning and real-time transcripts provide users with words that can be read from a screen. Since 1993, television signals in the United States must be closed-captioned to provide greater access to these important signals (Wikipedia 2019). YouTube (and other Internet video services) often provide a similar capability, using automatic speech recognition so that they can operate at the scale of the Internet. While captions are critical to making information accessible to people with hearing loss, they are also incredibly useful for anyone who needs to consume audio when there is loud background noise, such as in a restaurant or bar, or simply when one would like to consume spoken audio silently without disturbing anyone nearby.

On a smartphone, online transcription applications such as Live Transcribe use cloud-based speech recognition to recognize the speech picked up by the microphone and display the recognition results as text. Live Transcribe is a mobile accessibility app designed for individuals who are deaf or hard of hearing, and is usable by anyone. Using Google’s automatic speech recognition technology, Live Transcribe converts speech and ambient sound to text on the screen. The application provides real-time results, generally within 1 second of the words being spoken, rewriting the results if the recognizer’s language model changes its decision. Figure 2 shows several aspects of the interface. Live Transcribe supports more than 80 languages and additionally recognizes sound categories such as music, crowd noise, applause, and coughs. A sound level indicator, in the upper right of the screen, is useful so users know what kind of signal is being heard by the microphones.

Live Transcribe is built using models trained on many sources, including YouTube videos, so the recognizer is capable of high-accuracy recognition in many different environments. Real-time voice transcription, whether from microphones or other content like movies on the phone, can serve as an attractive alternative or adjunct to hearing aids, especially for the NBU, because mobile technology is increasingly ubiquitous and is an effective adjunct for so many other parts of our lives.

Fig. 1. The process of fitting a low-dimensional model to audiogram data. A, Sample fits for three different hearing losses. Solid lines are three users’ audiograms, while the dashed lines are the best two-parameter fits. B, The root-mean-square (RMS) errors for the 60 prototypical audiograms as a function of the number of principal component analysis (PCA) components. The black dashed line shows the RMS error overall the audiograms.
A common way to evaluate the efficacy of speech recognition technology is with the metric word-error rate. Google’s cloud-based automatic speech recognizer has achieved a word error rate of 4.9% as of 2017 (VentureBeat 2017) and is rapidly being driven down by the entire speech-recognition community as larger training sets and novel neural networks are applied to the problem. The word error rate is objective when there is a transcript, but less accurate when there are disfluencies or conversation restarts (Bortfeld et al. 2001). Most importantly, the word error rate is not necessarily a meaningful reflection of how useful the recognizer is to a user. Function words are recognized more accurately, partly because they are usually stressed and are often longer, so getting these words right is easier than minimizing the total word error rate.

Real-time speech recognizers pose an interesting UI question: how to deal with latency and rewrites? Modern recognizers decode speech with large probabilistic models. They look at the probabilities of the words given the acoustic input using both the immediate past and the future. A classic ambiguous sentence is the list of phonemes that sounds like “It’s hard to wreck a nice beach.” But if the next sentence starts with the words “However neural networks,” then the original sentence should probably be rewritten as “It’s hard to recognize speech.” Modern recognizers are willing to predict the first, perhaps 200 ms after each word, and then change their mind when they see more evidence in the future. Users want to see the recognition results as soon as possible, so they can formulate their reply. We call the delay between the end of the word and when it first appears on the screen the recognition latency. In a real-time system that aims to reduce latency, rewrites are necessary when the recognizer updates its word probabilities and then changes its final decision. It is difficult to make these changes on the screen and not add to the user’s cognitive load. Keeping the surrounding text fixed is an important design principle.

NEW TECHNOLOGIES

Many new auditory aids for the NBU are made possible by the surge of data and ML techniques using deep neural networks (DNNs). These networks are universal function approximations, finding the words, for example, that best explain the audio data. Ever deeper networks with hundreds of millions of parameters consume large amounts of data, tens of thousands of hours of training data, as well as use new techniques to train the DNNs that solve audio understanding in new ways. In this section, we describe systems for speech recognition, sound-event recognition, and speech enhancement. Again, we illustrate these new technologies with those cases with which we are most familiar, but these ideas are well known in the ML literature.

Modern speech-recognition technology is based on time-dependent translation models. A DNN predicts the next symbol, perhaps a piece of a word, from the input audio. This is implemented with a form of nonlinear regression, and the network converts input symbols, audio in a high-dimensional spectrogram space, into a probability of 16,000 different word parts. The Listen-Attend-Spell system described by Chiu et al. (2017) does this with a network of 100 million parameters and achieves a word-error rate of 5.8%, when trained with approximately 12,500 hr of English sentences. These state-of-the-art systems are in daily use, on phones and in homes, serving millions of queries a day.

A related technology is sound-event recognition or recognizing nonspeech audio. Similar deep networks but without the language model look at a window of sound and pick labels that explain the sound. AudioSet is a large-scale database of audio events that can be used for training a sound-classification system (Hershey et al. 2017). This dataset consists of more than 2 million 10-second sound clips that are manually annotated with 632 class labels. This size allows training a DNN with upwards of 30 million parameters to converge.

A sound classifier can provide additional context to a person who is hard-of-hearing and is concentrating on a written, real-time transcript. It is difficult to make a general statement about performance due to the large number of classes, the wide variability in class probabilities, and the task-dependent nature of any application. Systems based on AudioSet are optimized based on classification rates, where all sound classes are equally
important. Yet, it is not clear what the right error measure should
be since not all sound classes are equally important to users.

Finally, speech enhancement or noise reduction (NR) has the
opportunity to provide the biggest assistance to those who have
challenges hearing in noisy conditions. Speech-enhancement
systems commonly generate a time-frequency mask by which
the system can select the portions of an audio signal that con-
tains a speech signal (Wilson et al. 2018). These systems con-
vert the incoming audio into a spectral representation, such as
a spectrogram. Speech enhancement systems are trained from
noisy speech, often by adding noise to a clean waveform. From
the spectral representation, a DNN learns a model of what
speech looks like and predicts a mask that will produce the best
approximation to the original speech signal. This is a straight-
forward optimization problem.

A common metric to optimize a speech enhancement net-
work is signal to noise ratio (SNR), as it is just the ratio of the
powers of the desired clean signal divided by the power of the
noise left after enhancement. However, our ears do not hear
based on simple measures of power, since some noises are
masked by others. As described by Bosi and Goldberg (2002):

The inadequacy of simple objective measures was made dramat-
ically clear in the late eighties when J. Johnston and K. Bran-
denburg, then researchers at Bell Labs, presented the so-called
“13 dB Miracle.” In that example, two processed signals with
a measured SNR of 13 dB were presented to the audience. In
one processed signal, the original signal was injected with white
noise, while in the other, the noise injection was perceptually
shaped. In the case of injected white noise, the distortion was
quite annoying background hiss. In contrast, the distortion in the
perceptually shaped noise case varied between being just barely
noticeable to being inaudible (i.e., the distortion was partially or
completely masked by the signal components).

Thus, even though both signals had the same SNR, the percep-
tion of the noise could not be more different. Perceptual met-
rics, such as auditory quality (Kates & Arehart 2014), speech
intelligibility (Ma et al. 2009), listening effort (Pichora-Fuller et
al. 2016), or mean opinion score (Streijl et al. 2016) are perhaps
more appropriate for evaluating speech enhancement.

The power of these enhancement methods is that they build a
detailed model of what speech sounds like. That is what allows
them to decide that this bit of the spectral-temporal map is
noise and not part of the harmonic stack of a vowel. As shown
in Figure 3, Wilson et al.’s (2018) DNN enhancer looks at the
noisy speech spectrogram to decide how much of the spectral-
temporal energy is going to give the best enhanced speech
signal. Networks described in modern speech enhancement pa-
ters (i.e., Wilson’s) learn a model of speech so it can remove
nonspeech sounds. More recent work (Ephrat et al. 2018) learns
a model of sounds and facial images, so when pointed to a face
can separate out the speech corresponding to the desired face
movements and ignore other speakers. This is a form of speech
enhancement tied to lipreading.

One might expect that these systems work best when they
have access to both the speech before and after the current
frame. However, Wilson et al. (2018) got good results by lim-
iting the lookahead—the enhancement performance did not
improve when frames in the future are included in the optimiza-
tion. Surprisingly, the performance only drops a few dB when
the system is asked to predict the mask for unseen speech up
to 50 ms in the future. In other words, this system is predicting
the upcoming speech and noise even though it has not seen it
yet. While hearing aids endeavor to introduce less than 10 ms of delay, it is useful to have a system that can look ahead even a few tens of milliseconds, though it is not clear how general Wilson et al.’s findings are. Figure 4 shows this speech enhancement system’s performance.

Humans and machines both have the power to impute missing speech. Phonemic restoration is a perceptual phenomenon, where under certain conditions, sounds actually missing from a speech signal can be restored by the brain and may appear to be heard. In an example from Richard Warren (1970), a cough is heard in the middle of a sentence. Before prompting, listeners cannot recall the position of the cough within the sentence. This is an example of auditory streaming, where different streams are heard as different objects, and their relative timing is unimportant (Bregman 1990). In actual fact, the waveform of the middle of the word has been set to zero, and then a cough is added to mask the silence. Listeners hear the word normally, because we have strong expectations of how the English word sounds in a sentence. The SNR is quite low at this point, yet the cough is not heard as speech and is easy to ignore.

Wilson’s speech enhancement network is not able to completely restore the missing section. It largely zeros out the cough, leaving the original gap. But the network’s speech model does fill in a portion of the signal, as shown in Figure 5, where the original missing section is shown as a straight line, and the “enhanced” version shows some small speech-like sounds that have leaked into the quiet section. This is noteworthy because the speech-enhancement network is largely subtractive, yet like the human perceptual system, it uses some of the energy in the cough to fill in, or explain, the missing speech. To be fair, it is not clear whether the speech enhancement has failed to remove all the cough, or has learned a form of packet concealment, where noise is a better approximation to the speech signal than silence (Perkins et al. 1998).

There are many ways to characterize the performance of a speech enhancement network, often based on variants of the SNR (Le Roux et al. 2019). But as the 13 dB Miracle shows, this is not realistic. More importantly, Moore and Skidmore (2019) note: “ Speakers and listeners continuously balance the effectiveness of communication against the effort required to communicate effectively,” and the same could be said of speech enhancement. What matters is that the message is received with minimal human effort; however, that might be defined.

**FUTURE DIRECTIONS**

Future work toward helping the NBU with their auditory needs will focus on a user’s brain and specifically their attention and the effort they put into understanding an auditory signal.

Recently, a number of laboratories (see below) have reported success at determining to which sound a subject is attending. On the visual side, eye tracking gives a good clue about a user’s information-seeking intent (Hakkani-Tür et al. 2014). No such direct signal exists for the auditory system.

Instead, auditory attention decoding (AAD) works by finding the brain’s response to attended (or unattended) sounds and determining if the sound and brain waves correspond. This has been demonstrated for electrocorticography (Mesgarani & Chang 2012), magnetoencephalography (Ding & Simon 2012), and electroencephalography (EEG; O’Sullivan et al. 2015). Both electrocorticography and magnetoencephalography are expensive, so the recent EEG results are promising for mobile devices. This process is diagrammed in Figure 6. In this example, a subject is hearing two different audio sources but attending to just one. The two arrows illustrate how the signals might be processed by the subject’s brain. The blue (left) arrow shows the signal that is being attended and passed all the way “up” into the brain. It is processed and acted upon by the subject. The red (right) arrow shows the unattended signal, which perhaps gets only as far as semantic processing (Brodbeck et al. 2018). The goal of the decoder is to estimate the audio from the measured EEG signals (a backwards model), and then choose the original audio with the higher correlation to the estimated signal. This is the attended signal, which is fully processed by the user’s brain.

Work by O’Sullivan et al. (2015) showed that a model based on linear regression can take EEG signals measured at the scalp and connect the brain signals with an audio signal. The output of this decoding step is a correlation, and when applied to the two signals, the larger correlation indicates the sound that best fits the model and is thus attended (Choi et al. 2013).

In O’Sullivan et al.’s (2014) experiment, the model was performing well above chance, and for 7 subjects, the decoder was able to determine exactly to which audio story the user was attending. For many reasons, but certainly the structural issues related to the folding of the brain into sulci, the ability to decode auditory attention from some subjects is better than for...
Auditory attention decoding finds the acoustic signal, via one of many methods, that best explains the measured brain signals. Further shows that active NR enhances our ability to decode the active foreground signal. Furthermore, most interesting for those interested in hearing impairment, our ability to decode attention is not adversely affected by hearing loss (Fuglsang et al. 2020).

One of the surprising results is that our ability to decode the speech signal from brain waves correlates with the subject’s performance on a comprehension task. When this analysis was applied to O’Sullivan’s data, there was a significant and non-zero correlation for portions of the signal between 125 and 225 ms. Most interestingly, not only could this system determine to which signal the user was attending, it had a measure of how well they were understanding the desired signal, at least as measured by a comprehension test. This is a measure of correlation, not causation, and bears further study.

Accuracy and latency characterize AAD system performance. One common scenario envisions the AAD decision driving an amplifier that increases the volume of the attended speaker (or cuts the gain to the unattended speaker). Nobody has done a test yet to see how this kind of changing gain, which is a function of AAD performance, improves a user’s ability to hear what they want to hear.

More abstractly, our community needs to move toward objectives that quantify a device’s usability in the broadest sense. New tools allow people to communicate in ways that are not limited to amplifying audio. We not only want to find the most “usable” device, but we want to use new metrics to further adapt the device’s behavior in the user’s best interests. Thus we want to use, for example, a measure of (listening) effort as an optimization criterion in an adaptive system. The ideas are discussed in the remainder of this section.

A colleague of ours states “After a day of listening through my hearing aids, I’m exhausted.” It’s not the hearing aid that is the problem, but the active effort that is needed to understand the world around her. Research into a relationship between hearing and fatigue is still new, but there is some evidence to suggest that effortful listening is fatiguing, especially in people with hearing loss (Hornsby et al. 2016). In this context, the use of objective measures such as SNR or speech intelligibility to characterize performance is also problematic. Sarampalis et al. (2009) notes “Some listeners claim a subjective improvement from NR, yet it has not been shown to improve speech intelligibility, often even making it worse.”

The disconnect between objective measures such as SNR and working memory measures such as sentence-final word identification (Ng et al. 2013), and personal preferences leads to interest in measures like listening effort (Pichora-Fuller et al. 2016). Listening effort is objectively defined using a dual-task experiment (Gagné et al. 2017). Dual-task experiments trade off performance on two different tasks, which are both limited by some sort of cognitive limit. If the user is putting more effort into one task, necessarily the other task gets less effort and there is a reduction in performance. In one dual-task experiment (Sarampalis et al. 2009), listeners were tasked with both identifying words in noisy speech and speech that has been enhanced and then performing an independent visual reaction time experiment. The ability of a subject to recall words in the past was not significantly affected by the use of a speech enhancement (NR) system. Yet, on the second visual task, the NR system did allow the subject to more quickly respond to a visual stimulus. Other approaches such as pupillometry or subjective rating scales can also be used (Zhao et al. 2019).

In the longer term, the best systems will adapt to our personal usage. The first billion users are already at the point where much time is spent with personal electronics. We do not usually share our phones or earbuds; therefore, these devices have the potential to learn from our electronically mediated activities and their results. One such example is a 3D sound reproduction system that learns from the user’s behavior. In a computer game, sound and visuals combine to make a holistic environment. One could test to see if the user’s eyes correctly saccade to the source of a new salient sound. Thus, when a sudden sound is made in the upper right of the user’s field of view, the user’s eyes should at least glance in the direction of the sound. If the eyes go to the wrong place, we can use that as an error signal and adjust the head-related transfer function. Any difference between the desired location of the sound and the user’s reaction is an error signal that can be used to improve the device’s simulated head-related transfer function.

New systems based on RL use feedback from the user to better learn each user’s preferred settings. RL is a branch of ML that optimizes the long-term behavior of a system. This is based on self reporting, and the device uses the feedback to learn. The feedback tells the system that it did not make the right
prediction, and when it sees that condition again, it should do something different. RL is the basis of the superhuman results from chess or the game Go but can also be used in more personal settings, like setting the temperature of a house (Mozer 2005). The same thing could be done in hearing aids. For example, when my aid senses my partner is speaking, it should turn the volume WAY up because I will get in trouble if I ignore them.

These error signals are a form of self-report (Keidser et al. 2020). The signal is not perfect since the user might not feel it is worth the effort to correct the system (even though it just means pressing one button; Shamma & Slaney 2012) or because it knows that the effort is pointless. But this form of self report is important information to our devices and will become better as our devices become more intelligent.

Finding the right metric for either scientific or commercial use is hard. As applied researchers, we half-jokingly say that the only metric that matters is this one, and pull out our wallet. But as engineers we need measurable metrics by which to optimize these technologies. Is it better for users to use extra processing power for better neural network speech enhancement, or for better beamforming? How does one compare the efficacy of speech recognition versus attention decoding? How do we make it easier for people to adjust their tools to fit the acoustic environment, not to mention their level of energy to devote to understanding it? Systems have been designed that allow a user to specify to whom at the table they want to listen (Mobin & Olshuasen 2019) but how should we evaluate their performance? There is no single answer.

A constant stream of real-life data from sensors on a device can be used to refine algorithms with high ecological validity while maintaining the privacy and anonymity of users. Hearing aids that predict the user’s needs get feedback from the user, who either accepts the prediction or tells the device to do something different. This is an important error signal, which the developers can use to further tune the way their algorithms work in the “real world.” We will eventually learn under which circumstances a technology is used and perhaps more importantly how well it works. Privacy is always an issue with user data, especially outside a clinic. But new tools such as Federated Learning (McMahan et al. 2017) suggest ways to learn from private data, without the user’s data ever leaving their device. This can be done by updating the on-device model, that is, the network weights, based on private data and then consolidating the new models across the network of users.

CONCLUSIONS

From our point of view, the new technologies enabled by DNNs and the widespread adoption of these tools will dramatically impact the way we measure user satisfaction. Easily available hearing assistance, as well as technologies such as speech enhancement and speech recognition that remove the need for amplification, will increase the number of people we can help. New technologies such as attention decoding may make it easier for all of us to communicate.

We are no longer limited to understanding hearing capacity by the audiogram but can contemplate more holistic measures of communication ability. An assessment by users in their native environments is inherently an assessment of high ecological validity. But as hearing professionals and developers, we need matching metrics for these new technologies that can guide our development efforts.

The NBUs of our devices will choose the tools that work best for them, whether it is better amplification, speech enhancement or recognition, or more advanced tools that are keyed to the user’s needs as measured by external signals such as EEG.

ACKNOWLEDGMENTS

We are grateful for ideas and feedback from many colleagues, including Gitte Keidser, Graham Naylor, Thomas Lunner, Volker Hohmann, and Brent Edwards.

The article was largely written by M. S., with significant contributions, both in words and results, from R. L., R. G., B. K., C. G., K. W., D. K., S. S., and V. C.

This writing of this point of view was funded by Google Inc.

The authors have no conflicts of interest to disclose.

Address for correspondence: Malcolm Slaney, Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA. E-mail: malcolm@ieee.org

Received May 26, 2020; accepted August 14, 2020.

REFERENCES

American Academy of Audiology. (2018). The audiologist’s guide to hearing aids, PSAPs, hearables and OTC devices. https://www.audiology.org/sites/default/files/publications/resources/20180130_AuD_Guide_OTC.pdf
Apple. (2020). iPhone Accessibility. Retrieved from https://www.apple.com/accessibility/iphone/ hearing/. Retrieved July 6, 2020.
Arora, P. (2019). The Next Billion Users: Digital Life Beyond the West. Harvard University Press.
Bisgaard, N., Vlaming, M. S., Dahlquist, M. (2010). Standard audiograms for the IEC 60118-15 measurement procedure. Trends Amplif., 11, 134–140.
Bishop, T. (2017). Microsoft claims new speech recognition record, achieving a super-human 5.1% error rate. Geekwire. https://www.geekwire.com/2017/microsoft-claims-new-speech-recognition-record-achieving-super-human-5-1-error-rate.
Boothroyd, A., & Mackersie, C. (2017). A “Goldilocks” approach to hearing-aid self-fitting: User interactions. Am J Audiol, 26(3S), 430–435.
Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. Lang Speech, 44(Pt 2), 123–147.
Bosi, M., & Goldberg, R. E. (2002). Introduction to Digital Audio Coding and Standards (pp. 237–238). Kluwer Academic Publishers.
Bregman, A. S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound. Bradford Books, MIT Press.
Brodbeck, C., Hong, L. E., Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. Curr Biol, 28, 3976–3983.e5.
Chiu, C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gorina, E., Jaitly, N., Li, B., Chorowski, J., Bacchiani, M. (2017). State-of-the-art speech recognition with sequence-to-sequence models. International Conference on Acoustics, Speech and Signal Processing. https://doi.org/10.1109/ICASSP.2018.8462105
Choi, L., Rajaram, S., Varghese, L., Shinn-Cunningham, B. G., (2013). Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. Front Hum Neurosci., 7, 115.
de Cheveigné, A., Wong, D. D. E., Di Liberto, G. M., Hjortkjær, J., Slaney, M., Lalor, E. (2018). Decoding the auditory brain with canonical component analysis. Neuroimage, 172, 206–216.
Ciccarelli, G., Nolan, M., Perricone, J., Calamina, P. T., Haro, S., O’Sullivan, J., Mesgarani, N., Quatieri, T. F., Small, C. J. (2019). Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods. Sci Rep, 9, 11538.
Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A, 109, 11854–11859.
