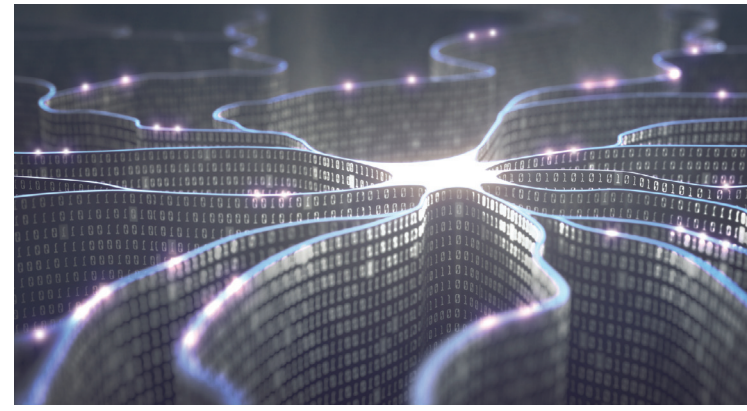
Can a Machine Learn to Solve our Speech-in-Noise Problem?

By Eric W. Healy, PhD

Anyone who has worked with individuals with hearing loss has heard something like this: *"They work just fine when it's my wife and me in our kitchen, but if we go anywhere noisy, they just don't work well at all."* "They" in this case are typically hearing aids, although the same applies to cochlear implants.



iStock/ktsimage

WHAT'S THE PROBLEM?

This statement highlights the most common complaint of hearing-impaired (HI) individuals–poor speech recognition when background noise is present. Accordingly, effective noise reduction can be considered one of our most important goals. Despite its importance, noise reduction currently implemented into modern devices is notoriously ineffective (after all, if it were effective, the complaints would stop). In fact, the literature shows that noise reduction often produces a subjective preference, but all too often, no actual increase in intelligibility.

WHAT'S A GOOD SOLUTION?

A goal that has been long sought involves a single-microphone technique to improve speech intelligibility in noise. Single-microphone approaches have distinct advantages over microphone-array techniques like beamforming, making them preferable. But historically, they just haven't worked. Single-microphone techniques have been able to remove noise from speech, but in doing so, they produced distortions. So one starts with speech that is not intelligible because it's noisy, and ends up with speech that is not intelligible because it's distorted. Phillip Loizou summarized this work in "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions" (*IEEE Trans Audio Speech Lang Proc.* 2011;19[1]:47).

WHAT'S OUR APPROACH?

Recent years have brought about advances, including a solution pursued by our group at The Ohio State University that involves a machine-learning algorithm to improve intelligibility of noisy speech for HI listeners. This work, done in collaboration with DeLiang Wang, PhD, has two main components: The first involves time-frequency (T-F) masking, and the second involves machine learning.

In T-F masking, the speech-plus-noise signal is divided by time and frequency into small units–think checkerboard squares in the spectrogram view. In its simplest form, the units dominated by speech are retained, and the units dominated by noise are simply discarded. By dominated, I mean having a favorable/unfavorable signal-to-noise ratio (SNR). When only the speech-dominated T-F units are presented to listeners, the speech can typically be understood perfectly.

In time-frequency masking, the speech-plus-noise signal is divided by time and frequency into small units—think checkerboard squares in the spectrogram view.

The second component in our approach involves machine learning. This task of classifying T-F units into two piles is well-suited for machine learning. In the classic example, a machine is shown images of apples and oranges. During a training phase in which it learns, the machine is also told the answer–what each one is (apple or orange). This makes it "supervised learning." After being shown many examples, the machine enters an operation phase. It's shown new images of apples and oranges, ones it hasn't seen before, and is not told the answers. But it can effectively classify them.

Dr. Healy is a Joan N. Huber Fellow and a professor at The Ohio State University.

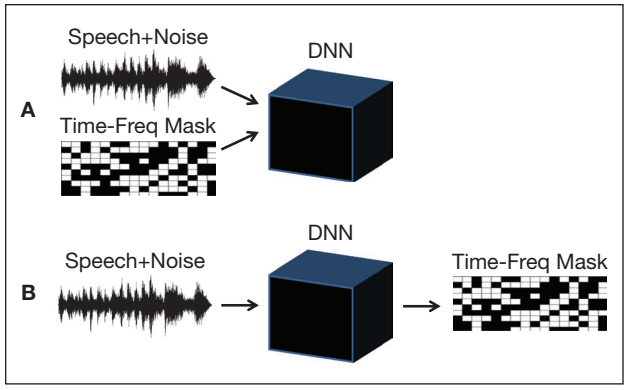


Figure 1. Illustration of the machine-learning noise-reduction algorithm. During a training phase (A), a deep neural network (DNN) is fed a speech-plus-noise mixture, along with the ideal time-frequency (T-F) mask for that mixture. This is repeated hundreds or thousands of times with different mixtures. Once trained this way, the DNN is simply fed the speech-plus-noise mixture and is able to estimate the T-F mask for that mixture (B). The application of the T-F mask to the mixture is a simple matter that results in substantial intelligibility increases.

In our use of machine learning, the machine (a Deep Neural Network or DNN) is provided with a sentence mixed with noise. During the training phase, it is also given the answer, which is whether each T-F unit is dominated by speech or noise. Once trained this way, with many examples of speech in noise, the DNN can classify on its own. It learns what speech-dominant units look like and what noise-dominant units look like. Once the DNN labels the speech-dominant units for us, we can simply present those, and only those, to listeners (Fig. 1).

CI processing lends itself well to T-F unit processing because it already performs a somewhat similar decomposition of the signal.

Does it work? Keep in mind that increases in intelligibility have historically been almost entirely elusive. Figure 2 shows some of our data from typical hearing aid users. Vast improvements are clearly obtainable using our approach. In fact, many HI listeners improved intelligibility from very low scores to scores of roughly 80 percent. Furthermore, in some conditions, the HI listeners with access to our algorithm actually outperformed young normal-hearing listeners (without the algorithm, of course) on speech intelligibility in noise. This suggests that an older (~70's) HI listener using our algorithm could potentially understand speech as well or better than their young (~20's) normal-hearing conversation partner in a noisy setting.

Can it work for cochlear implants too? Most of our effort has been focused on the largest population in need–individuals

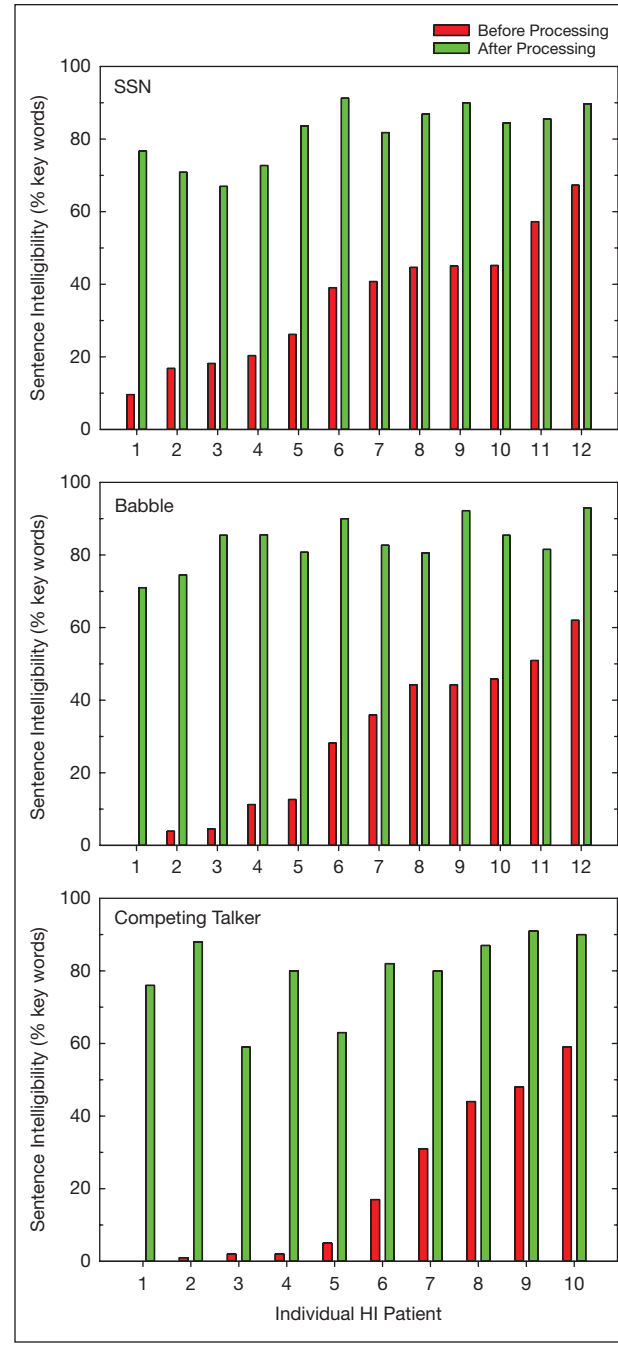


Figure 2. Sentence intelligibility scores for individual hearing-impaired listeners (typical hearing aid users) before and after processing by our algorithm. Shown here are scores in three types of background noises (speech-shaped noise, multi-talker babble, and a single competing talker). Apparent is the wide range of relatively low scores prior to processing (red) and the relatively uniform high scores following algorithm processing (green; *J. Acoust Soc Am.* 2013:134[4]:3029**;** *J. Acoust Soc Am.* 2017; 1141[6]:4230**).**

with sensorineural hearing loss and who wear hearing aids. But as previously noted, the speech-in-noise problem is similar for people with cochlear implants (CIs). If anything, they

are even more hindered by noise than individuals who use hearing aids. Reasons our algorithm may also be well-suited for CI users include the following:

1. Since the speech-in-noise problem is even worse for these folks, they typically need a highly favorable SNR to understand speech. So even if an algorithm could only operate effectively at relatively high SNRs, it could produce improvements for CI users.
2. CI processing lends itself well to T-F unit processing because it already performs a somewhat similar decomposition of the signal.
3. CI processors are typically more powerful than hearing aid processors, potentially opening additional processing options.

But can it work in the real world? Two main considerations for implementing a machine-learning algorithm involve generalizing to conditions not encountered during training and operating in real time. Our series of papers have focused on this first aspect, and strides have been made. We have demonstrated generalization to untrained sentences, SNRs, segments of noise, and entirely novel noise types. Regarding real-time operation, the algorithm is essentially "feed-forward." So although we have focused on effectiveness and haven't implemented real-time operation yet, there is no fundamental barrier.

Will it ever go behind the ear? The short answer is no, but as I'll describe, that's the wrong question to ask. First, remember that most of the heavy lifting, the "computational load," is encountered during training and prior to a person actually using it. We can use our most powerful computers to train the DNN, and we can train it for as long as we want—for a day, a few days, or a month—it makes little difference. What does matter is how efficiently it runs once it's trained—after all, that's what it will do once a person puts it on and goes out into the world. Fortunately, the computational load associated with operation is far smaller than that associated with training.

But even with this advantage—that the computational load is largely shifted to an earlier training stage—it's important to understand just how limited even the most powerful ear-worn devices are in terms of battery and processing power. Getting a trained DNN to operate on those platforms could be a challenge. We suggest thinking about the problem differently. Current smartphones possess massive battery and processing power that rival personal computers. They also possess the ability to transmit bi-directionally and wirelessly. We suggest a solution involving a smartphone-like device that can be placed in a pocket or handbag and can transmit wirelessly to earpieces. This solution provides convenient packaging, massive power, and earpieces potentially even smaller than those of current ear-worn devices, because they will simply need to include powered microphones and output transducers (speakers). Perhaps a machine can learn to solve the problem that has beset us for so long.