Voice-to-Voice Cloning for the Laryngectomee Community: Not Quite There, But Close

By J.H. Snider, February 4, 2024

After a biopsy on July 12, 2023, my doctors informed me that my survival depended on removing my cancerous vocal cords—and thus my natural voice—in an operation I had on Sept. 5, 2023. My voice was an essential part of my identity, so this was a traumatic next step in my battle with cancer. I addressed it by embarking on a race to record my voice so I could later take advantage of the emerging technology of AI-based voice cloning, including both text-to-speech and real-time voice-to-voice cloning.

By the summer of 2023, the idea of cloning one's voice was hardly science fiction. Indeed, it was front page news in every major newspaper in the United States. Fake voice clones of politicians and celebrities were widely discussed as a threat to the integrity of social media, the mainstream media, our democracy, and even pornography. Hollywood and other actors from throughout the United States were on strike--one of the largest and longest strikes of the past few decades--seeking to protect their rights to digital replicas of themselves so that they could profit from them. If during 2024 the quality of the new TV shows you watch is much worse than in recent years, you can probably blame it on the actors' strike during 2023 to protect their image and voice cloning rights. At least a dozen companies, some quite sophisticated, such as llElevenLabs, PlayHT, and Respeecher, were marketing voice cloning services to the non-disability community. Reviews rating such companies proliferated over the Internet.

By the time I was told that my life depended on removing my vocal cords, I had already been on a four-year battle with laryngeal cancer; in my case, centered on my glottis. This included three laser surgeries to remove the cancer. After that failed, I was put on a full course of radiation during 2022. And after that failed, my voice box was finally placed on the chopping block, to be removed in an operation called a supracricoid laryngectomy. Most patients who get a laryngectomy get a total laryngectomy, but my doctors at Penn Medicine said I could get by with a partial laryngectomy. This meant I would still be allowed to breathe through my mouth afterward, even if my voice box had to go.

In terms of the effects on my voice, I would still have a voice after the laryngectomy, but it would be a much hoarser and softer one. Indeed, I'd sound more like what an adult bear with laryngitis might sound like if it could speak like a human. Some fellow members of the laryngectomee community said I shouldn't care so much about preserving my old voice, as I'd grow into my new one and come to think of it as me. But the reality was that my new voice not only didn't sound even like a human being but also only my family and friends were likely to have the patience to hear me like they used to. Talking to other people would become a huge hassle--assuming they were even willing to stick around to try to decipher what I was saying. So regardless of whether I cloned my own voice or another human voice, I could greatly improve my ability to communicate from a clone of any human-sounding voice. And if I were to use a clone anyway, why not one of my pre-laryngectomy voice?

Voice Banking vs. Cloning

Cloning one's voice involves two separate tasks: First, recording (called "banking") one's voice. Second, using that banking to create a clone. Thus, my immediate task was to bank my current voice for future cloning. This proved to be far more difficult than I initially expected.

One reason for that difficulty is that the voice cloning industry for text-to-speech products to the disability community is geared to selling voice cloning products, not voice banking services. This is reflected in the fact that voice cloning services typically provide some type of free voice banking service as a way to entice potential customers to purchase their voice cloning product. Customers, including speach language pathologists (SLPs) who occasionally recommend such products to their patients, seem to assume that if the voice clone is good, then the voice banking is also good. And it is on this assumption that I believe they fall off a deep cliff.

The problem I confronted is that it's in each vendor's interest to lock you in to their voice cloning product so you cannot easily transfer your voice recording to a competitor. Most customers seemed to be different from me by focusing on the end product of the voice cloning service rather than the portability of the voice recordings used to make it. A dozen or so companies competed in the text-to-speech voice cloning market for the disability community, (the MND Association and PreserveYourVoice had useful but incomplete lists), the voice cloning technology was fast-changing, and I concluded that I wasn't at all clear which vendors would be the survivors. I didn't want to spend a lot of time investing in a voice recording with one vendor only to discover years later that it wasn't the best service for my needs.

For my long-term horizon, I didn't care much about the short-term voice cloning output; I was most concerned about whether I'd have perpetual access to my input, my banked voice. Here I found most of the vendors highly vague in their policies. And even if I could later get access to my voice, I didn't know how valuable that would be if the vendor had already gone out of business, changed its voice banking terms, or lacked a viable customer service option to access the voice banking.

Apple, which was launching a free voice cloning service as part of its operating system, allowed users to download their voice recording. But that amounted to more than 100 separate files, one for each sentence recorded, and I wasn't sure how useful that would really be as I didn't know a single vendor who would accept Apple's voice recordings as the basis of their own voice cloning service. Using another vendor's recordings just wasn't how the business seemed to work.

Another problem is that I considered pretty much all the current text-to-speech voice clones unusable for my type of demanding use, where I wanted not only to communicate but communicate as effectively as I had done before my operation. Thus, the only thing of real value to me in the short-term was preserving a good and useful recording of my voice.

The "useful" criterion for voice banking proved highly confusing to me. All the text-to-speech vendors I tried had the same basic voice banking model: I would need to read lots of sentences, one-by-one, and then they would use that input to create a voice clone. However, within that basic framework there was lots of variation. For example, the amount of time required to train a voice clone varied from 20 minutes for Apple to fifteen hours for ModelTalker. Some vendors

gave users a choice of text to read (e.g., fun text like reading *Alice in Wonderland*) while others provided tortured sentences that would be sure to include, say, all the possible phonetic combinations of the English language. I had no idea which type of voice banking would be future proof.

Ultimately, I decided to unbundle my voice banking and voice cloning needs, even though the text-to-speech services I had tried treated them as a bundle by having customers follow their company-specific and sentence-by-sentence voice recording prompts. For high-end voice cloning outside the disability market, which is often for voice-to-voice rather than text-to-speech cloning, that's largely how the voice cloning business already operated. By high-end, I mean the type of voice cloning services that Hollywood produces (e.g., for actors) and Fortune 500 companies (e.g., for automated customer service representatives) They were often using conversation-style banked voices to clone the voices of their talent. The resulting clones were already incredibly impressive, by which I meant realistic sounding. The catch for me was that none of it was geared to real-time cloning. That is, none could turn my new hoarse voice *instantaneously* into my former natural voice; all these services required an intermediate production step that took time to transform one voice into another, which wasn't an option for my primary use of my voice, which was to converse with people in real-time.

On the other hand, given that I was more geared to future rather than present voice cloning technology, it seemed to me that if I got a good banking of my voice, it didn't matter if the technology that allowed me to converse in real-time didn't yet exist. My banked voice would be available when it did.

The Challenge of Voice Banking

The next problem I ran into was conflicting advice—or a lack of any advice within the laryngectomee community—on how to do such a voice recording. Press reports (e.g., see segment beginning at 5:50) and vendors (see Google's advice to professionals) indicated that proper voice banking was a highly sophisticated endeavor. Early in my research, an academic specializing in voice technology told me that I should use four mics for an excellent recording—a near, mid-range, far-afield, and throat mic. I wasted a lot of time trying to follow that microphone advice and then eventually gave up when I learned that that is not how professional voice recording studios worked.

In the end, I decided that a single high-quality mic—the type of specialized mic professional recording studios used—was all that was needed for my purposes.

I rented my first professional studio, Future Fields, for close to four hours. It was in Burlington, Vermont, where I was then vacationing. The first ninety minutes was devoted to reading sentences one-by-one for SpeakUnique's text-to-speech voice cloning service. I'd read a sentence and then decide if I did a good job. If I decided the answer was no, I'd click a button to reread it. If the answer was yes, I'd click a button to move on to the next sentence. The recording process was actually pretty fun because I chose *Alice in Wonderland* as my text—one of a dozen or so the company offered. What wasn't so fun were the bugs in the software interface in transitioning from sentence to sentence. But after a half hour or so of usage, including learning a

sign language to communicate with the studio's sound producer, I figured out how to get around those bugs.

Next up, I read chapters from a variety of fiction and non-fiction books. But that didn't work well because I'm not a natural actor and my voice was too monotone. The producer decided that I should bring a friend into the study so I could engage in more natural-sounding conversation. But since my friend's voice couldn't be recorded, his side of the conversation was based on written questions. This resulted in a more natural-sounding version of my voice. But the producer decided I was still using a stilted voice that wasn't ideal for training an AI to reproduce my voice as a clone to be used in a conversational setting. Alas, I had no more time to fix these problems, as I was leaving Vermont the next day.

The next week I reserved a studio in my hometown of Severna Park, Maryland. I had decided that one hour of natural conversational speech would probably be adequate for a high-quality future voice clone. This time my conversational partner was my wife, and we crowded into a small recording room. But yet again it was decided that my conversational style didn't adequately correspond to the conversational style of real human beings. The problem wasn't the technology; it was me.

Fortunately, I had two children who were professional musicians and had also done a lot of acting. They turned my home's walk-in closet into a recording studio (the walk-in closet was like a professional recording studio because it has lots of clothes to absorb sound and was secluded so had no outside noise interference.) They eventually figured how to get me to speak in a natural, conversational style. But the price was their voices occasionally interjecting in the middle of the conversation.

The Challenge of Using Voice Clones

So finally, I had what I thought was a future-proof voice banking of my natural voice. Alas, there was little I could do with it because I hadn't found a usable real-time voice-to-voice cloning service and the text-to-speech vendors I was looking at all had their own sentence-by-sentence voice models to build their voice clones.

I figured that for the foreseeable future I'd have to rely on text-to-speech voice cloning. This included my immediate need to communicate with staff during my hospital stay to recuperate from my laryngectomy. For weeks I was wildly excited about Apple's promised new text-to-speech voice cloning service built into the next generation of its IOS operating system, IOS17, to be released in mid-September 2023. I loved the concept behind Apple's vision: a free text-to-speech voice cloning system that was part of its operating system and that Apple app vendors could build into their products. Given Apple's reputation for user friendly, high-quality software, I was confident it would shake up the market and be the benchmark for evaluating all other text-to-speech voice cloning products. By mid-August, I had signed up as an IOS17 beta user and began using it. Alas, I was deeply disappointed in the product Apple delivered. The output was no worse than the other text-to-speech voice cloning software products on the market, even though the number of sentences required to train Apple's app was less. At the level of the individual word, the cloning was good. But at the level of sentence, the prosody was poor; it didn't at all sound like how a real person would speak. What most annoyed me was the small

text box Apple provided to enter the text to be converted to speech. I'm the type of person that speaks and writes in long sentences and in paragraph length units, and Apple's small text box was just too limiting for me. I was astounded that such a simple problem with a potentially easy fix would prevent me from being reasonably happy with the Apple product. Another consideration was that I was a Windows user and couldn't bring myself to abandon Windows to use such a flawed Apple product.

Eventually, I settled on the UK's SpeakUnique as my text-to-speech product. It was endorsed by The Swallows, a UK nonprofit neck cancer support group that provides resources for UK laryngectomees. Also, if my U.S. speech language pathologist (SLP) filled out a grant form provided by The Swallows, The Swallows would give me a grant to use SpeachUnique for free. This was a service The Swallows provided to all laryngectomees, not just me.

I hated all the clutter on the SpeakUnique interface. All I really wanted was a simple text box where I could type in the text I wanted converted into a clone of my voice. But I recognized that those interface options were useful for people with different disabilities than myself, so I just focused on the core part of the interface useful to me.

Lower-Tech Solutions

Alas, within a day of losing my vocal cords and trying to communicate with hospital staff, I ended up using a traditional boogie board for my communications with the staff. This was a big letdown for me, as I had spent so much time training and choosing a text-to-speech program. But I found the boogie board far more efficient and effective in my hospital setting than any of my text-to-speech programs.

One facet of hospital life as a laryngectomee patient was dealing with huge numbers of different hospital staff who transition every ten hours and sometimes more. All told, I had about ten different caretakers come into my room every day. I got an extreme taste of hospital staff turnover the first day after my surgery, when I transitioned from intensive to regular care and a completely new set of staff.

Some of the staff knew I couldn't speak; others did not. When I used my text-to-speech app, the initial reaction was often like deer in the headlight. One reason is that the cloned speech didn't sound natural because the prosody was so off. In particular, when I had a question, it didn't sound like a question; it sounded like a statement. The result was that I usually had to replay the sentence at least once and sometimes more times. That meant I couldn't go on to the next sentence until I was assured the first was already understood. I found that when I wrote my text on a boogie board and handed it to the staff, they only had to read it once to understood what I wanted. I also could write longer passages that were more easily understood when people could read what I was trying to communicate. Just "launching" the boogie board also proved faster than launching either my iPad or boogie board, both of which had a locked screen and an app I'd have to locate before beginning the text-to-speech. With the boogie board, I was instantly able to begin writing.

I chose a boogie board over a pad of paper because it left no record, including of my comments to my family, which sometimes included critiques of the hospital staff. For example, I would

report to my family my experience with a nurse who didn't know I couldn't speak. I buzzed her repeatedly during the wee of the night when I had trouble breathing and wanted my throat suctioned. Each time she popped into my dark room, looked at my head without hearing anything, and assumed that I had made a mistake by hitting the buzzer, which often happens with patients. After this happened several times, she got angry with me and stopped coming. My eventual solution was to stop buzzing her at multi-minute intervals, which I thought was polite, and instead harass her with incessant buzzes, which communicated to her that I wasn't buzzing by mistake.

After I left the hospital and began communicating with my wife during car rides or with vendors over the telephone, the text-to-speech applications proved superior to the boogie board, which was unusable in such settings. But I still wasn't very happy. Communicating with my wife, it was fine. But the prosody was still awful, the communication rate awkwardly slow, and sometimes strangers, such as telephone reps for the many vendors and prospective vendors I deal with, gave up. Vendors that I've communicated with in recent months include ad placement, airline, cruise operator, dentist, driveway, e-commerce, election office, fence, financial, grass cutter, handyman, hotel, HVAC, newspaper, package delivery, pest control, plumbing, pool maintenance, primary care physician, restaurant, roof, security, taxi, telecommunications, tourist destination, and voice cloning.

Since the vendor often didn't initially get what I was saying, I was unable to type ahead to my next statement. I had to wait until I knew whether they understood what I had said and then found myself often replaying what I had already typed rather than engaging in a natural conversation. Ultimately, I decided to first try vendor automated chat services before calling via telephone. Alas, I got a lesson in how awful these chat services tended to be. Many started with chat robots who didn't respect my time and were rarely helpful given the questions I asked. I longed to return to talking to live reps.

After leaving the hospital, I also began exploring other non-AI, much lower-tech voice enhancement technology. I wanted to find out if I could amplify my voice with smart sound technology just as sound engineers routinely enhance sound through dozens of different controls. My son, who is a musician, found dozens of inexpensive apps that could enhance music and other sound recordings without recourse to voice cloning technology. For example, I hoped to deemphasize my voice's low, hoarse frequencies and emphasize its higher, more human-sounding frequencies. But for inexplicable reasons, I didn't find one of these apps that worked in real-time. They worked fine on pre-recorded music, but even when I tried the various apps in real-time, I got unpleasant feedback. (But please note that at the January 9 to 12, 2024 Consumer Electronics Show, the largest electronics trade show in the world, there seemed to be dozens of AI-based sound enhancement products either on display or announced, so the information above, even if it was correct for last year, may already be obsolete.)

I then decided to go even lower tech. Tens of thousands of teachers at all grade levels use voice amplifiers to amplify their voice in the classroom. These teachers have perfectly fine voices but find it taxing on their voices when they must speak all day long and sometimes at elevated volumes in large classrooms. So they use a lapel mic or something similar attached to a voice amplifier hanging from their neck or belt. This was a very simple and inexpensive voice amplification technology, but my wife didn't like it because it amplified my grating, hoarse

voice. I may try using it again in settings such as in a noisy restaurant with friends or family. But otherwise, I doubt I'll have much use for it.

Maryland's Relay Service

I submitted via the web a jargon-filled Maryland Relay service profile to get a free, live person to translate what I was saying in a voice that would be more intelligible than my own.

This service was enabled by the American Disability Act, which includes a provision requiring the FCC to provide a shared set of foundational rules and funding for state provided "Telecommunications Relay Service" (TRS). For example, the FCC requires that TRS be provided 24/7 and offer e911 service. The FCC's base funding for TRS can be supplemented by additional state fees on telecommunications services, which 33 (66%) of states charge, including Maryland, and with fees varying widely among those states. Overall annual funding combining state and federal expenditures appears to be in the multi-billion-dollar range. But only a small fraction of that is likely to be for speech-to-speech services. Those who have difficulty seeing or hearing, not those with difficulty speaking per se, are likely to be the largest TRS users. About ten different companies compete to provide TRS services at the state level. The U.S. General Accounting Office oversees the FCC's administration of TRS services. The National Council on Disability, an independent federal agency reporting to Congress and the President, observed in an October 31, 2019 report that TRS has "problems with choice, competition, and quality." Some agencies that provide TRS services at the federal and state level have advisory committees including representatives from the disability community.

The branding for Maryland's relay service is Maryland-, not Federal- or provider-, based, including a picture of the Governor on the brochure touting the service as a free Maryland service for the disabled. The only initial hint I got that the day-to-day speech-to-speech service was actually provided by a company rather than the State of Maryland was the instant receipt I received from Hamilton Relay after I submitted my profile (but when I looked harder, I found this information). From my limited experience with Maryland Relay operators, I have nothing but praise for the work done by Hamilton's phone operators.

A month went by after I submitted my profile, I assumed was an application, without my receiving any follow-up notification from Maryland Relay. I then decided to see if I could use the service by calling 711. To my surprise, Maryland Relay treated me as an already qualified user.

The high quality of the subsequent service also surprised me, as I didn't expect that from a free-to-user government-provided service. If the purpose of the service was to provide excellent speech-to-speech repetition of my voice to help my listeners understand me, Maryland Relay excelled. Nevertheless, I have some reservations, which may not apply to most other users.

I am ambivalent about using Maryland Relay because most of my vendor relationships involve first providing a password and other personal identification information to access my account, and I'm not sure I can trust that information with an anonymous Maryland Relay service rep. I recently learned that Maryland Relay promises confidentiality to users like me, but even if I had

seen that fine print confidentiality promise in some obscure location, I doubt I'd trust it for certain highly sensitive transactions, such as communicating with a financial institution.

When I used the service the first time, I felt guilty wasting a half hour of the Relay operator's time because it took me that long just to get through to a live telephone representative from the vendor (T-Mobile) I wanted to talk with. The Relay service won't let a disabled person call a vendor directly; instead, the Relay service has to call the vendor. So, if, as in my case, the vendor took half an hour to pick up the phone, the Relay operator had to wait during that interval. The Relay operator then must provide the access information, including a PIN in the case of T-Mobile. While waiting, I thought that if the operator is a volunteer (my initial assumption), it is pretty inconsiderate of me to take up so much of her time. Later, I learned that operators were paid, but I would still feel guilty for wasting money by having a highly professional person, presumably paid reasonable compensation by the minute, wait before I could use her services. The fact that taxpayers were paying it rather than me would not eliminate my guilt.

Lastly, after taking up so much of the Relay operator's time, I wouldn't know ahead of the call if the Relay operator's services would really be needed. If the telephone rep. has an excellent understanding of English (often not the case with today's overseas phone reps) and is a patient listener willing to learn how to listen to my strange voice, then the Relay operator's service won't be needed. I feel I occupy an awkward gray zone between needing and not needing this type of highly professional free service. I simply couldn't imagine myself using the Relay service multiple times a day; it would damage my self-esteem too much, even though for most users, such as those who use an electrolarynx and have a much worse communication problem than I do, I'd expect it to boost their self-esteem by empowering them to live a more normal life.

In any case, I still had my heart set on a voice cloning solution, which would be less labor-intensive and more efficient for all concerned.

The Holy Grail: Real-Time Voice-To-Voice Cloning

From the very beginning of my quest to clone my original voice, my heart had been set on finding a real-time voice-to-voice, rather than text-to-speech, voice cloning app. Real-time voice-to-voice was clearly the Holy Grail—in theory, vastly superior to text-to-speech—if I wanted to communicate like I had before my laryngectomy. I immediately found one company that was working on such technology and expected to have a usable product by last November 2023 (it's now January 2024 and no such product has appeared). I found another company that was working on an advanced text-to-speech app and was considering moving into real-time voice-to-voice as an enhancement to its asynchronous voice-to-voice product. But it's now four months later and multiple deadlines for just their text-to-speech app have come and gone with no delivered app.

During the past few months, including at the recent CES2024, a slew of claimed real-time voice-to-voice cloning apps seem to have been launched. I claim to be no expert on them. But the few I tried seem to define real-time as recording a voice and then having it repeated a few seconds later in any voice one wants, including potentially one's own voice. For me, however, that is still not what I'd consider real-time, which must mean instantaneous in the same way a conversational voice is instantaneous.

In early January, my friend and fellow laryngectomee Steve Cooper, who is a font of information about companies providing voice enhancement services, pointed me to a Dutch company, WHISPP, exhibiting at CES2024 that claimed to provide a real-time voice-to-voice voice cloning service for not only telephone calls but the disability community. Eureka, I thought! The company's vision of "smart amplification" or "voice-to-voice AI" was exactly what I was looking for--except that the promised product was for real-time voice-to-voice cloning for telephone, not in-person, communications.

Alas, my experience so far with this company is that it has employed the common Silicon Valley fake-it-until-you-make-it strategy, which seems to be endemic to the current voice cloning marketplace. I could get WHISPP's real-time voice-to-voice cloning service to work like the other so-called real-time voice-to-voice cloning services described above. But it only worked with two normal voices from my family, not my hoarse voice. Perhaps that was because the product was trained on Dutch, not English, and the company's founders seemed to be focused on helping stutterers, not those missing vocal cords. Also, the telephone conversation version wasn't yet available. To get a clone of my voice, I had to email a five-minute version of my voice to the company, which I did in early January. I then sent a follow-up reminder, where I also offered to be a beta tester. But three weeks later, I haven't heard back from the company. Still, I'd recommend that people watch this company to see if it eventually delivers on its compelling and publicly announced vision.

I remain confident that real-time voice-to-voice cloning will eventually be a usable product--and within the coming decade. I'm hoping, for example, that the new generation of AI PCs being launched in 2024 by Intel and AMD, with support from Microsoft's next-generation Windows software, Windows 12, launching later this year, will have the necessary processing power to instantly translate my hoarse voice (or any other type of voice, such as from an electrolarynx) into a cloned voice. Traditional PCs only had CPUs (computer processing units) and GPUs (graphical processing units). The new generation of computers adds NPUs (neural processing units) to allow for fast local AI processing. I imagine it's only a matter of time before both vendors and consumers realize the potential of this technology for localized real-time voice cloning. Already at CES2024, the AI PC chip makers were touting instantaneous voice language translation (e.g., for Google Gemini's "automatic speech translation"), which seems to me a much more difficult technological challenge than same-language voice cloning. However, the devil for these products is often in the detailed implementation.

What I think is clearly doable right now is what I'd call *approximate real-time voice-to-voice cloning*. The idea here is that the user could say a sentence or some other set of words, and this recording would be immediately cloned in a live or remote setting. The cloning technology but not the interface for this type of voice-to-voice cloning is already available (e.g., see voice.ai). For all I know, some company may also have already created a workable interface for it, by which I mean a fast and easy way to quickly jump from sentence to sentence. My take for the moment is that the dozens of markets these voice cloning companies cater do don't seem to think this an important use case.

Approximate real-time voice-to-voice cloning is hardly a perfect voice cloning solution, but I think it could be a highly usable interim solution. For example, it is essentially the type of voice-

to-voice service Maryland Relay provides, except that it would be automated, of uniformly consistent quality, and available to anyone regardless of state. That is, it would be much more scalable—a favorite term used in Silicon Valley to describe a desirable product quality—than Maryland Relay. (But recall that Maryland Relay provides many other services than just voice-to-voice for the disability community, so my point here is quite limited.) In my case, I could use this type of cloning to quickly train listeners to understand my voice; as soon as these training wheels weren't necessary for my listener, I could revert to speaking with just my hoarse voice, which I think could often be done after fewer than a half dozen sentences.

Recommendations

I have to live life as it is, not how I want it to be. Thus, I'm relying on multiple, lower-tech voice enhancement technologies, including boogie boards, text-to-speech voice cloning, voice amplifiers, Maryland Relay/TRS, asynchronous voice-to-voice cloning, and my own voice, depending on the particular context. That's also the messy advice I must recommend to others for the time being: use multiple techniques to enhance your communication power.

I am hopeful that once voice-to-voice AI finally takes off, this messy reality will be greatly simplified, just as general-purpose PC technology replaced dozens of previously separate products such as the typewriter, calculator, camera, and telephone. The best possible scenario might be if computer and smartphone vendors include AI-based real-time voice-to-voice cloning in operating systems such as IOS, Android, and Windows. That way the voice cloning service would be free to use by end users who have one of those operating systems, and the service could be readily incorporated in apps built on one of those operating systems. Perhaps we could even get rid of the clutter of dozens of different and confusing disability applications in current operating systems and have something like Microsoft's Co-Pilot (based on ChatGPT) and Gemini (built on Google's Bard), two of the highest profile general AI's currently on the market, be the universal and simple interface for real-time voice-to-voice voice cloning, just like it's expected to be for instantaneous voice translation. But it's hard for me to imagine that specialized hardware, such at Atos Medical's voice hardware for laryngectomees, won't also play an important role.

Meanwhile, too, I'd be cautious about any vendor claiming to offer any type of AI-based sound improvement product, not just voice cloning services. The term "AI" is not well defined and in recent months appears to have become a must-use marketing term for companies selling sound enhancement products, even if those products haven't much changed from their pre-AI incarnations. It's gotten so out-of-hand that the term might even come into disrepute, such as 5G did when mobile phone companies underdelivered on its claimed advantages. Perhaps the best way to think of AI is in aspirational terms: the holy grail of communication for the laryngectomee community.

In terms of public policy and consumer education for the laryngectomee community, I think the priority should be focused on enhancing the input of voice banking rather than the output of voice cloning. The European Union's law that social media users should have access to their own data is an interesting precedent for voice banking, as I think voice cloning services should have a standardized and enforceable policy to provide their customers with access to their voice recordings. One reason for this European directive is to allow social media users to transfer their

data from one social media provider to another so as to reduce the monopoly power of those providers.

In the more free-market U.S. context, a more realistic recommendation might be for SLPs and nonprofit organizations that focus on the needs of the laryngectomee community to shift their focus from the output to the input of text-to-speech voice cloning apps. Specifically, they should not endorse any text-to-speech software that doesn't give customers an interoperable version of their voice recording so it can be used with other voice cloning vendors. They should also provide guidance as to what constitutes and how to make a high-quality voice recording. The current practice of reaching out to laryngectomees with voice enhancement advice only after their laryngectomy obviously is not ideal, as voice cloning technology cannot replicate a pre-laryngectomee voice if it does not have a reasonably high-quality recording of the pre-laryngectomee voice. Moreover, the recording must be made as early as possible because voices have often heavily deteriorated prior to the laryngectomy operation. My guess is that solving the technological problems of real-time voice-to-voice cloning will be relatively easy compared to this needed change in institutionalized patient care. On the other hand, it might be viewed as empowering to allow laryngectomees to create any voice they want for themselves, thus turning what has traditionally been viewed as a defect into a sort of superpower. Most laryngectomees appear to make peace with their new voice and that may be even more so when they'll have so much control over what that new voice sounds like. I for one want my old voice back, but the future probably belongs to those seeking voice superpowers, not just replication of the voice of one's genes.

The laryngectomee community represents only a tiny fraction of the voice banking and cloning market. One only has to look at how voice cloning products are marketed to see that. But since much of the non-disability market has an unsavory reputation in both the public policy community and the general public due to the problem of deep fakes, the laryngectomee community may have surprising corporate and political leverage. This may help explain, for example, why Apple launched its first publicized voice cloning product with a free voice product for the disability community rather than a more profitable market that might have sparked more controversy. No one can claim that its voice cloning product is harming the lives of politicians, celebrities, or women, which appears to still be the dominant public perception of voice cloning. Consider this recent passage from a Bloomberg article:

> The recurring story of new technology is unintended consequences. Take AI-powered image generators. Their creators have claimed they are enhancing human imagination and making everyone an artist, but they often fail to mention how much they're helping to create illicit deepfake pornography too. Lots of it. Over the weekend, X had to shut down searches for "Taylor Swift" because the site formerly known as Twitter had been flooded with so many faked porn images of the singer that it couldn't weed them all out. One image alone was viewed more than 45 million times before being taken down. Swift's scandal points to a broader problem: Around 96% of deepfakes on the web are pornographic. But it could also be the final tipping point before some genuine solutions are introduced….
>
> The technology is being used in more scams and bank frauds, it's making Google search results worse and it's duping voters with fake robocalls from President Joe Biden…. Lawmakers are up in arms…. When Microsoft Chief Executive Officer Satya

> Nadella was recently asked about the Swift deepfakes and jumped straight into platitudes about "guardrails" — rather than make any specific policy recommendations — that may have been because his firm is at the heart of today's booming generative AI business.

Here are some recent headlines similar in tone:

- A fake recording of a candidate saying he'd rigged the election went viral; Experts say it's only the beginning, CNN, February 1, 2024.
- The deepfake era of US politics is upon us, CNN, January 24, 2024.
- Even President Biden's Voice Is Now Too Easy to Fake; Beware: AI and a host of free tools have made voice cloning — and misleading political robocalls — something almost anyone can do, Bloomberg, January 24, 2024.
- Deepfake Audio of Biden Alarms Experts in Lead-Up to US Elections: While many have warned of deepfake videos and images in the lead-up to this year's elections, experts say fake audio worries them the most, Bloomberg, January 22, 2024.

When hearings are held in Congress about the dangers of voice cloning for our democratic, family, and other values, I wouldn't be surprised if companies like Apple, Google, and Microsoft trot out their voice cloning products for the disability community to show the extraordinary good that voice cloning can do. China has already passed extensive regulations to prevent the use of voice clones as deep fakes. Companies in the voice cloning business may come to consider disability applications as a sort of pro bono service that helps give them and their employees a good reputation. This type of potential power, which is still more in the realm of speculation than fact, is something the laryngectomee community—and the disability community more generally—should consider as it strives to foster voice cloning technology that helps its members.

Voice cloning technology for laryngectomees is both rapidly improving and sold by companies prone to hype. The good news is that these companies' visions are often good, and the emergence of powerful AI processors in everyday consumer computers may help them realize it. The bad news is that if you aren't willing to suffer the pangs of being an early adopter, it's still likely to be a wait-and-see moment. Meanwhile, laryngectomee consumers and trade associations have a responsibility to keep the companies honest by pushing for policies, such as interoperable voice banking, that companies will otherwise have incentives to resist.

#

J.H. Snider is a public policy analyst. An earlier version of these remarks was posted in **Lary's Speakeasy Laryngectomy throat cancer**, August 16, 2023.

Acknowledgement

Special thanks to Steven Cooper for suggesting that I write this article and then finding it a home in the laryngectomee community. Thanks to Larry Alexander for his inspired choice of graphics for this article.