Voice & Swallow Mechanics Lab

Speech and Technology

AUTHORS: VICTORIA MCKENNA, PH.D., CCC-SLP AND BRITTANY FLETCHER M.S., CCC-SLP

	TIME (MM:SS)	TRANSCRIPT PAGE 1/5
1	00:00 — 00:36	***Dr. Victoria McKenna, Ph.D., CCC-SLP:*** Hello and welcome to Culturally Inclusive Education for the Speech Sciences. Today, we will be discussing speech and technology. My name is Dr. Victoria McKenna, and I am an assistant professor in the Department of Communication Sciences and Disorders and I am the director of the Voice and Swallow Mechanics Lab at the University of Cincinnati. I am here with Brittany Fletcher, a certified speech-language pathologist and doctoral student currently enrolled at UC.
2	00:36 — 00:52	This module series is funded by the Advancing Academic-Research Careers Award from the American Speech-Language-Hearing Association, also known as ASHA, and the College of Allied Health Sciences at the University of Cincinnati.
3	00:52 — 01:21	This module series assumes prerequisite knowledge across two areas. First, that you have a basic understanding of speech sounds and their corresponding frequencies. And second, that you have a broad understanding of the phonological and linguistic differences in various English dialects. That is, dialects of Southern American English and African American English, as compared to Standard American English.
4	01:21 — 02:10	Here is a brief outline of today's presentation. First, I will begin by discussing the foundations of speech technology including different examples commonly used in American culture, such as speech-to-text capabilities. I will discuss the hardware and software required to transform speech into text, and finally discuss the benefits and drawbacks of speech-to-text technology. Next, Brittany Fletcher will delve into the biases associated with automatic speech recognition processes. She will discuss how artificial intelligence, or AI, is trained based on narrow examples of American speech. She will discuss the implications for speakers of African American English and identify considerations for speech-language pathologists.
5	02:10 — 02:18	To begin, I will discuss speech technology currently used in American culture.
6	02:18 — 03:59	There are several different speech technologies used today, with only some included on this list. Speech technology can include text-to-speech, in which a person can use orthographic or pictorial input to create speech output. One primary example of this technology is augmentative and alternative communication, or AAC, used by some individuals who have difficulty with oral speech and/or language. But speech technology also encompasses platforms that use oral speech as a primary means of communication. For example, telehealth and telecommunication have become exceedingly important to work, social, and daily activities during the COVID-19 pandemic. These platforms must ensure accurate and efficient transmission of acoustic signals in order to effectively communicate from remote locations. Our last example here is speech-to-text technology. This technology uses automatic speech recognition to take an acoustic representation of speech and translate that into a digital form that can be interpreted by a device. Some examples that come to mind are voice assistants, such as Siri or Cortana, as well as educational and business softwares that use speech-to-text to improve learning, efficiency, and/or work productivity. It is expected that the global speech and voice recognition market will be worth approximately twenty-two billion dollars by the end of 2022, due to its multiple uses and the recent advances in detection accuracy.
7	03:59 — 04:12	Regardless of age, gender, race, or ethnicity, speech-related technology has become part of everyday life for many people across the U.S.

Voice & Swallow Mechanics Lab

Speech and Technology

AUTHORS: VICTORIA MCKENNA, PH.D., CCC-SLP AND BRITTANY FLETCHER M.S., CCC-SLP

	TIME (MM:SS)	TRANSCRIPT PAGE 2/5
8	04:12 — 05:06	***Dr. Victoria McKenna, Ph.D., CCC-SLP (cont.):*** Today we are going to be focusing in on automatic speech recognition, also known as ASR. ASR allows for hands-free access to technology that is unrestricted by someone's physical abilities. ASR may be a tool someone can use who's unable to type on a keyboard. Likewise, it can also be a tool for someone who has difficulty writing, also known as agraphia. A combination of text-to-speech and speech-to-text can be used by someone with reading impairments due to vision problems or other disability. Speech recognition has been shown to be life changing for some, in that it can assist with activities of daily living, such as turning lights on and off at the request of your voice. Likewise, it can also help with social connectivity, as many social platforms often communicate via text.
9	05:06 — 07:04	In order to acquire an acoustic speech signal, there needs to be a microphone. But not all microphones have the same capabilities or are made specifically for speech acquisition. First, there can be variation in the directionality of a microphone, meaning that there can be variation in where the microphone is trying to listen for a speech signal. A unidirectional microphone only picks up sound from a specific direction. Head-mounted microphones that you see for singers and public speakers are usually unidirectional and designed to only pick up sound from the direction of the speaker's lips. Conversely, omnidirectional microphones are designed to pick up sounds in any direction. Interestingly, the microphones on most phones are omnidirectional. Likewise, an Amazon Echo box, which you can place in your home to access Alexa, would also have an omnidirectional microphone so that it can pick up your voice regardless of your location. You're probably thinking, "Well then, aren't omnidirectional microphones better?" Well, omnidirectional microphones may be helpful when you're speaking into a device where you or the device might change location. For example, when you're holding your phone to your ear versus when you're reading on it or speaking on speakerphone. However, omnidirectional microphones can unintentionally also pick up a lot of background noise. When there is high levels of background noise in the recording, it can reduce your signal-to-noise ratio, a measurement that tells us how strong the signal we actually want is to the background noise that we do not want. When signal-to-noise ratios are low, meaning that the noise in the signal is high, then ASR algorithms can have more difficulty distinguishing speech sounds.
10	07:04 — 08:46	Besides microphone capabilities, the sampling rate that speech is acquired can also impact the accuracy of ASR. The standard sampling rate for high quality recordings is 44.1 kHz, which is the same as 44,100 Hz. That means in order to make a high-quality recording, more than 40,000 samples have to be captured during a single second. This process is often referred to as speech digitization and can be modified to sample at higher or lower rates. The benefits of having a higher rate is a clearer signal quality that truly reflects the original sound source. But a drawback is the high amount of processing power needed to acquire that information plus the size of the recording files are larger and take up more space on your recording equipment. Therefore, there's a trade-off between the quality of the recording and the processing and storage capabilities of your device. This leads to the question, "What is the minimum sampling rate for speech?" To answer this question, we need to use our knowledge of speech frequencies relevant to each phonemic sound and understand how eliminating some frequencies might impact speech intelligibility. For example, we know that the first four formants convey important information that allow us to distinguish between different voiced phonemes. Most of the information can be found at frequencies less than 5 kHz, or 5000 Hz. However, voiceless phonemes require mid-to-high range frequencies to help us distinguish between sounds. Keep that in mind as we move to the next slide.

Voice & Swallow Mechanics Lab

Speech and Technology

AUTHORS: VICTORIA MCKENNA, PH.D., CCC-SLP AND BRITTANY FLETCHER M.S., CCC-SLP

TIME (MM:SS)	TRANSCRIPT PAGE 3/5
11 08:46 — 10:47	***Dr. Victoria McKenna, Ph.D., CCC-SLP:*** Dr. Victoria McKenna Ph.D., CCC-SLP: Here we have the same speech recording sampled at several different sampling rates, from as low as 4 kilohertz all the way up to 44.1 kHz, which is our high-quality recording rate. What's important to remember when listening to these samples is that the sampling rate must be double the frequency range you wish to examine. This is due to Nyquist theory, in which a sampling rate x2 (times two) of your target must be used in order to prevent aliasing of the signal. For example, if we were interested in examining frequencies up to 5,000 Hz, we would then need to sample at 10,000 Hz. You can also do the math backwards, in which you know that if you sampled at 20,000 Hz, you would only have frequency information up to 10,000 Hz. Knowing that, let's listen to our first two samples: the rainbow sentence at 4 kHz and 8 kHz. Keep in mind that the frequency range you're listening to is half of that, or only 2 kHz and 4 kHz." ***Recording 1, 4 kHz:*** *"A rainbow is a division of white light into many beautiful colors."* ***Recording 2, 8 kHz:*** *"A rainbow is a division of white light into many beautiful colors."* ***Dr. Victoria McKenna Ph.D., CCC-SLP:*** The speech is more clear at the higher sampling rate of 8 kHz, which would then capture frequency information up to 4 kHz. Interestingly, this is the same frequency information transmitted in the original landline telephones, as they were only developed to transmit information below 4 kHz. Now, let's listen to the other examples of higher sampling rates and consider their quality and the clarity of their speech sounds. ***Recording 3, 12 kHz:*** *"A rainbow is a division of white light into many beautiful colors."* ***Recording 4, 16 kHz:*** *"A rainbow is a division of white light into many beautiful colors."* ***Recording 5, 55.1 kHz***: *"A rainbow is a division of white light into many beautiful colors."*
12 10:47 — 11:50	So, how does speech recognition work? Well, we just established the hardware and software specifications needed to acquire and digitize the speech signal. Next, your device needs to have the processing capabilities to extract and categorize speech features. This is performed using artificial intelligence, or AI. AI is a special kind of technology that is able to learn and adapt over time based on the input it is given. This means that AI can learn from its mistakes to improve accuracy over time. However, initial AI rules are pre-established ahead of time. For the case of speech recognition, AI is fed speech samples from something called the speech corpus, or a collection of speech from a wide range of speakers. This allows the AI system to learn initial features of speech and informs categorization for future decisions. Therefore, AI systems are heavily dependent on the speech corpus they are developed from.
13 11:50 — 12:20	Subsequently, the accuracy of automatic speech recognition software is highly dependent on the speech corpus and its AI programming. Inaccuracies arise when a user has speech patterns that deviate from the training corpus of a particular AI system. This has resulted in issues across many users in the U.S. Next, Brittany Fletcher will discuss biases in AI systems and how that information is relevant to speech-language pathologists.
14 12:20 — 12:42	***Brittany Fletcher M.S., CCC-SLP:*** Hello. My name is Brittany Fletcher. I am a clinician and a PhD student at the University of Cincinnati, studying speech sound disorders in minority populations. And today, I'm here to talk to you about the biases of artificial intelligence, or AI, and its impacts and relevancy to us as clinicians.

Voice & Swallow Mechanics Lab

Speech and Technology

AUTHORS: VICTORIA MCKENNA, PH.D., CCC-SLP AND BRITTANY FLETCHER M.S., CCC-SLP

TIME (MM:SS)	TRANSCRIPT PAGE 4/5
15 12:42 — 13:57	***Brittany Fletcher M.S., CCC-SLP:*** When we think about artificial intelligence, we usually assume that it has nothing to do with us as speech therapists. We can categorize the world of technology into the AAC division with high tech devices, and we tend to not consider any other realms technology may fall into within our practices. So first, let's consider some of the common uses of artificial intelligence, which is automated speech recognition, or ASR software. This has an impact on our everyday lives, at home, in the community, and in the clinic. And the ASR recognition is utilized by big names, such as Alexa, Google, Cortana, and Siri. It is on almost everybody's phone and laptop today and allows for us to speak to our phones and easily ask questions and make commands. I have even begun to create goals for speech sound disorders, or SSD, for older kids and teenagers to practice and increase intelligibility by using their own Googles and Siris.
16 13:57 — 16:19	With AI and this great power of technology, comes great responsibilities that we need to address. Years ago, I listened to a TED Talk by Cathy O'Neil from 2017, titled "The Era of Blind Faith in Big Data Must End." In this speech, she spoke about how the bias of artificial intelligence is one of the reasons that the blind faith in artificial intelligence must end. Based on that talk, I learned the bias comes from the fact that even though artificial intelligence is programmed to learn for itself, the knowledge that is given to it is given through the data researchers collect and input. At the end of the day, these researchers are human with their own inherent biases cultivated through how they were raised through their communities, through their beliefs, even through their specific areas of specialization. This impacts how the data is collected and interpreted, and eventually is fed into the artificial intelligence system. So, if the foundational knowledge and data given to the system is biased, no matter how much the artificial intelligence learns, it and the information it will provide will inherently be biased too. So, within the speech science realm, one way that this affects us is thinking about, "When using these common day tools of Alexa and Google, who all can really utilize these to its fullest potential?" Think about when you use Google. If you are a native English speaker that mostly speaks Standard American English, you don't have to think too much about how you speak and articulate to Google. You may just have to increase your loudness, if you are far away from the speaker. However, this is not the case for people from culturally and linguistically diverse communities. What the software is able to understand from your speech production depends upon if it has the information to understand it. Different languages and dialects have different patterns. So, if all the system has learned is the patterns of Standard American English, it will be hard for the system to understand culturally and linguistically diverse people that have different patterns.
17 16:19 — 17:16	Therefore, we need to consider the populations that are affected by biased AI data. We need to think about the people from culturally and linguistically diverse communities and groups. People who may speak Southern American English or African American English, or people who have a dialect due to English being their second language. Most speech data analyzed and given to AI is a speech from Standard American English speakers. The software then becomes a little more difficult to use and it requires for people in these communities to either code-switch or over articulate to fit the patterns and forms that the system will recognize as Standard American English. So, a tool provided to everyone now seems like it was not made for everyone. It requires more effort for them to use these items that are offered for everybody's use.
18 17:16 — 18:02	Now we are going to take some time to see what the research has found about AI bias in minority communities. Let's specifically look at a population that is affected by the automatic speech recognition, or ASR, which is a specific type of artificial intelligence that is used for Alexa and Google and Siri. A great article that was published wrote about the impact of ASR errors on African Americans. In this study, African American participants were interviewed to understand how they use ASR and how that impacts them, their everyday life, their identity, and their beliefs around the system.

Voice & Swallow Mechanics Lab

Speech and Technology

AUTHORS: VICTORIA MCKENNA, PH.D., CCC-SLP AND BRITTANY FLETCHER M.S., CCC-SLP

	TIME (MM:SS)	TRANSCRIPT PAGE 5/5
19	18:02 — 18:59	***Brittany Fletcher M.S., CCC-SLP:*** Here are the results explaining the lack of satisfaction of ASR technology among African American participants. 27% of the participants said that dictating, sending, or reading a message contributed most to their lack of satisfaction. Some other contributions included: playing music; dictating, sending or reading an email; or that the technology was not understanding the words, phrases, or names, especially the ethnic ones that were used. So, this means that a good chunk of the frustration of the participants came from the basic tasks of dictating and sending or reading messages, which is what most people use ASR for on their phones. Therefore, accessibility and use of this tool is reduced in the minority population.
20	18:59 — 19:54	Here we can see some results on why African American participants aren't satisfied with the ASR use. Most of them said it's because the ASR gave incorrect results for the speech that they provided. Secondly, ASR didn't understand their commands, or they just ended up having to do the task manually. Third, at 27%, they said the transcribed the messages were incorrect. And lastly, 18% said they still had to proofread or edit after use of the ASR tool. So, when they're trying to use ASR for one of its purposes, including dictating, reading, sending messages, the system is not working for them., and that can be very discouraging."
21	19:54 — 21:07	Let's look at the top three attribution of errors among the participants when using ASR. Meaning, why did the African American participants believe the technology was not working for them or that the technology was creating errors. 30% said the technology wasn't designed to pick up accents and slang. 20% said that technology didn't understand me, or my speech patterns. And then 10% said technology isn't familiar with the words, phrases, or ethnic names that are used. Here we see a glimpse of how people from diverse communities can feel undervalued or separated from the larger community. We have this common tool used for people all over the country and all over the world; however, for these African American participants, you can see that they think, "This tool cannot understand me. I know that I am different, and I am a minority in the community, and yet this tool cannot understand what I'm saying so that it can be used in a way that the majority of speakers and other speakers can use it and what it was intended for."
22	21:07 — 21:54	Here with these results, we can get a understanding of who African American participants believe the technology is made for. We see 36% say they think it's made for white people, mostly, for people without an accent. 21% say people who don't use slang. And then the 14% said people with an American accent or people who use correct Standard English. So once again, you're getting this underlying confirmation of the African American participants saying, "I know I'm a minority. I know I speak differently than the majority of people in this country, and this tool that is supposedly for everybody is not meant for me and for my community to use."
23	21:54 — 23:46	The last piece of results that we're looking at for this lecture are the frequency of speech modification of the African American participants using ASR. So, how often did the African American participants have to modify, change their speech, make it fit into a certain pattern for the ASR system to understand them. And we see that most people said "yes most of the time," meaning most of the time when they're ready to speak to Google, that they have to code-switch. They have to change their pattern. They have to make a conscious decision for their speech to sound more like the majority of what they think Americans sound like, for the system to understand them. They have to articulate in a manner that they know is not the way they normally speak, but it is the way that they know this system will understand them. And they have to alter the essence of who they are to use this technology that everybody has access to and that they want to use along with everyone else. This is a huge issue, and it's something that has been researched more recently in the coming 10 years. But it is up to researchers to be able to improve the systems and that just doesn't mean looking at the technology and improving the technology itself. That means improving the understanding and acknowledging biases within the researchers that are making these products and how that affects the product's ability to provide its services to all people, from all communities. We need more representation within our data so that the data is more diverse to input into the systems, so the system can recognize more diversity.

Voice & Swallow Mechanics Lab

Speech and Technology

AUTHORS: VICTORIA MCKENNA, PH.D., CCC-SLP AND BRITTANY FLETCHER M.S., CCC-SLP

TIME (MM:SS)	TRANSCRIPT PAGE 5/5	
24	23:46-25:32	***Brittany Fletcher M.S., CCC-SLP:*** I will end today's lecture by starting a conversation on: Now that we know this information about bias in AI, what is our role as clinicians? Whether it's working with the adult or pediatric population, I have observed more functional communication goals around the independent use of ASR systems at home and in the community. For example, a goal of having the client use Siri to ask for their favorite music to be played or to send an email or text to a friend or family member. As speech therapists, we need to be aware of our clients' attitudes and emotions around the use of these systems. Not only with our clients who may have motor speech and speech sound disorders, but also with our clients from diverse communities, backgrounds, and who speak various dialects and languages, who we now see can have possible ASR errors when using the system as well. For linguistically diverse population, there may be underlying frustrations with use of ASR tools at home. So, if you try to implement it into your therapy session and a client is already frustrated with ASR because it does not understand them when they have used it previously, that may further feed into their communication frustrations surrounding their impairment and cause them to shut down or impede the progress in the session. AI is a great tool and allows for a new level of accessibility and creativity within our sessions. We, as clinicians, just need to be aware of the tool's capabilities to understand people who speak various dialects and languages other than Standard American English and consider the positive, or negative, relationship people in this community may have when using ASR
25	25:32 — 25:38	***Dr. Victoria McKenna Ph.D., CCC-SLP:*** Please find a list of references and resources below.
26	25:38 — 25:48	***Brittany Fletcher M.S., CCC-SLP:*** Thank you for listening to today's presentation. Should you have any questions, do not hesitate to contact me or Brittany Fletcher.