Audiovisual Speech Enhancement: Effects of Development and Hearing Loss, in partnership with American Auditory Society

Presenter: Kaylah Lalonde, PhD

- [Christy] It's my pleasure to welcome Dr. Kaylah Lalonde. This presentation is in partnership with the American Auditory Society. Annually we host two courses with AAS. And this year Dr. Lalonde is gonna be presenting Audiovisual Speech Enhancement: Effects of Development and Hearing Loss. If you'd like to learn more about Dr. Lalonde, you can read about her bio on the course registration page, but at this time I'll hand the mic over to you, Dr. Lalonde.

- [Kaylah] Thank you. Good morning. These are my disclosures. Here are the learning outcomes for today's presentation. I'd like to say thank you to both ContinuED and the American Auditory Society for inviting me for this presentation today. Okay. The research in this talk is funded by the National Institutes of Health and was conducted in collaboration with a large team of individuals who I'd like to quickly acknowledge here. Okay, in this presentation, I'll share some research findings focused on audiovisual speech perception. Much speech communication occurs in face-to-face settings where we have access to both acoustic speech and to the visible movements of the face and mouth that coincide with acoustic speech. Audiovisual speech enhancement refers to the way these visual cues from a talker's face help to compensate for environmental noise, hearing loss, and other forms of degradation.

This graph is one of many examples of audiovisual speech enhancement in the literature. Here, accuracy of word recognition is plotted on the Y axis as a function of signal-to-noise ratio on the X axis. The solid line represents performance in auditory only conditions with no visual cues. The dashed line represents performance in audiovisual conditions with visual cues. One way to quantify audiovisual enhancement in this graph is to compare accuracy between auditory only and audiovisual conditions at each signal to noise ratio. As shown by the arrows across signal to noise ratios. Adults recognize words more accurately in audiovisual conditions than in auditory only

conditions. Another way to quantify AV enhancement is what amount of noise can you tolerate and still reach a certain level of accuracy.

For example, what signal to noise ratio corresponds to 50% speech recognition thresholds? And you can see here that value is about -8 dB for auditory only conditions and about -11 dB for audiovisual conditions. So that's a three decibel benefit of the visual speech, but speech doesn't only help in noise. Actually previous studies show audiovisual benefit in quiet listening conditions when the task is particularly difficult or the speech input is very complex, it seems to help with listening effort for example. There are multiple mechanisms that underlie audiovisual speech enhancement. First visual speech tells you when to listen. Second, it provides cues about the auditory amplitude envelope of speech. Third, it helps you to access audiovisual speech representations. And fourth, it helps you to segregate and attend to a target talker.

We'll talk about each of these turn. So first, visual speech tells you when to listen. Seeing this woman's lips pursed in this way tells you that a burst of acoustic energy is about to happen. One way we can see the benefits of this type of temporal cue is in measures of speech detection and noise. And the example I'll show you, you'll see this woman's face appear on the left and then on the right you'll see her mouth the word ba in both intervals. Your job is to determine whether you heard the woman say ba during the first interval when she's shown on the left or the second interval when she's shown on the right.

- Ba.

- [Kaylah] Okay? We do the same thing with just pictures.

- [Subject] Ba.

- [Kaylah] So even though you see the same thing in both intervals, her mouth movements help you to pick out exactly when in the noise to listen for that ba, And that actually allows you to detect the speech at a 2 dB lower signal-to-noise ratio than when you only see the pictures. Visual speech also helps you to follow the auditory amplitude envelope of speech. I'll explain what that means using this figure. Here in gray is the waveform of a sentence. The auditory amplitude envelope is the slow variation in overall acoustic energy that's shown in blue. And you can see that the auditory amplitude envelope follows the general outline of the shape of the waveform. It turns out that the area of the mouth opening over time is correlated with that auditory amplitude envelope with greater mouth opening area corresponding to louder sound.

And smaller area or closed mouth corresponding to quieter moments in speech. That's shown here by the fact that the peaks and valleys are relatively aligned between the blue and red auditory and visual signals. Recent studies have shown that this envelope information can improve speech perception on its own. Yuan and colleagues made videos, where they had a circle on the screen which varied in size over time to match the envelope of the auditory speech. Here's an example of what that might look like.

- [Subject] Dad came to say hello.

- [Kaylah] This type of circle video actually improved recognition of sentences in multi talker babble by over listening alone by 7%. So seeing just this visual cue to that amplitude envelope is beneficial on its own. A third wave visual speech helps is that you can use knowledge about what mouth gestures correspond to a particular speech sound to access speech representations in long-term memory. And this is the mechanism that I'm going to focus on mostly today. So we have this long-term knowledge that certain speech mouth shapes go along with certain sounds. And we can use that information during speech perception. Here's an example of that in action.

I'm going to play a silent video and you try to determine which of these four words she said.

- Ready. Ready.

- [Kaylah] Okay.

- Ready.

- [Kaylah] You probably guessed that she said mash as opposed to any of these other words, and that's because you know, maybe without really thinking about it, that the closure of her mouth goes with P sounds, B sounds and M sounds and so forth. In fact, just as there are phonemes and auditory speech, there are categories called visemes in visual speech. When we're lip reading there are patterns to the sounds that we confuse for one another. So you might've guessed that she said mash or match, bash, batch, patch, and that's 'cause all of these sounds contain the same visemes in the initial position and in the final position. Several researchers have looked at those sort of patterns of errors from lip reading by making consonant confusion matrices.

Here's an example of how that works. We present words or syllables and plot what consonant we presented against what consonant was in the response. The diagonal here represents correct responses and the gray represents errors. Several researchers have looked at those patterns. Here's an example of data showing the same thing. So again, we have the target in this case syllable on the X axis and the response syllable on the Y axis. More, let's see, sorry, the color represents how many responses were given to that particular target. So the dark blue indicates that you never made that particular response to that target. For example, you never responded Z when the

person said P, the red colors or the redder colors, the warmer colors indicate that there were more responses there.

So for example, when V was presented, the person always said it was V. So once again, the correct responses are the boxes along the diagonal and all of the off diagonal responses are the consonant errors. What I'd like you to see from this graph is that even in the errors, responses kind of cluster together. As we noticed in the lip reading example, P, B and M are confused for one another. Those are the ones on the bottom left, but they aren't confused for other consonants and they make up one viseme category. So highly confusable phonemes belong to the same viseme category. Visual information cannot reliably distinguish phonemes in the same viseme category. Visually distinct phonemes belong to different viseme categories.

They're not often confused with one another in lip reading testing because the information from the visual signal reliably distinguishes them. The way we can see this visual knowledge at play in audiovisual speech perception is by looking at the types of errors that people's confuse in auditory only conditions and in auditory visual conditions. The idea is that if you're using this visual phonetic knowledge when you have access to visual speech and get the consonant wrong, it should be substituted for another consonant that looks like the one that was presented. It should be substituted for a consonant inside these gold boxes, one in the same viseme category. And so we're gonna look at those types of errors in adults.

So here I'm showing you data from a study where we presented words to adults in a speech spectrum noise. And you can see that in auditory only conditions in white, and then the audiovisual conditions in dark gray, you can see their performance was better on the audiovisual conditions, their accuracy. Than in the auditory only conditions. But then what we were really interested in, as I said, are the kind of errors that they made. And so what we did was to look at only the errors and say what proportion of

consonant confusions in each modality are within the viseme category are visually confusing. And you can see that even though there were more errors in the auditory only condition, the audiovisual confusions were more likely to be within viseme category.

So that tells us that they were using that visual knowledge whenever they were listening to the audiovisual speech to kind of narrow down when they couldn't quite make out a sound, they can narrow down to the ones that had the same mouth shape. All right, finally, a fourth way that visual speech helps is that in complex environments like this one where the background sounds are very similar to speech, as in when there are other people talking, visual speech can help us to pick out the parts of the incoming acoustic signal that go with the target talker and attend to those parts. It helps us to better ignore competing talkers. Because visual speech helps in this way.

Audiovisual benefit is actually larger when background sounds consist of other people talking. For example, a two talker competing speech masker, which are harder to segregate than when it consists of steady state noise, which is easier to segregate. All right, so those are the four mechanisms that I wanted to tell you about, but what I wanna talk about now is, how does audiovisual speech enhancement develop in children with typical hearing? First, let's talk a little about infants. So infants actually show pretty early audiovisual speech perception abilities. For example, infants are sensitive to the correspondence between audiovisual, auditory and visual speech cues. If you play a sound such as oo to babies, while presenting them both of these pictures, they will actually look more at the face that matches the ooh sound, the face on the right than the face on the left.

That doesn't match the oo sound. And this shows us again that they are sensitive to the fact that that kind of rounded gesture goes with the oo sound. Infants also use temporal cues to improve detection and discrimination of speech and noise. So

remember we saw earlier that this kind of, preparatory gesture in the woman on the left tells us that a burst of acoustic energy is about to happen, that we were able to benefit from that visual speech even though all it told us is when speech might happen. It turns out that infants can actually do the same. They benefit from that same sort of temporal queuing. It helps them to detect speech better and to discriminate between different speech sounds better.

Okay, in terms of children, both preschool and school-aged children also benefit from visual cues. There have been several cross-sectional studies examining the development of audiovisual speech enhancement in this age range. These studies typically show an increase in enhancement over age, but they differ with respect to the youngest age that significant enhancement is observed. And the age at which enhancement becomes adult-like. However, there are also some studies that suggest a U-shaped trajectory with a decrease in benefit in the early school years. And some studies actually show no change in benefit over age or a decrease in benefit just during those early school years. We recently conducted a study to try to get a grasp on what's happening developmentally across school age.

And to do so, we completed a study in which participants completed a battery of auditory only and audiovisual speech recognition tests. This battery included three measures of audiovisual speech enhancement. All of the tasks required children to sit in a sound booth and listen to speech in a steady state speech spectrum noise. Their task was to repeat whatever the talker said. The participants were 63 children and 18 adults with normal hearing and normal or corrected to normal vision. I'll describe each test a bit in the following slides. For the first measure, we obtained auditory only and audiovisual 50% sentence recognition in noise thresholds using high probability SPIN sentences and a one down one up adaptive method. All sentences had four target words that were with an early age of acquisition.

Here are the results from the first measure. Many of the subsequent figures will be very similar. On the left panel, 50% thresholds are plotted as a function of child age with auditory only thresholds in red. And audiovisual thresholds in purple. Age is shown on a log scale because we expect greater rate of change in younger participants than in older participants. In the right panel box plots show the distribution of adults thresholds for the same measures using the same color scale. The diamonds showed the average for each condition. To analyze these data, we used a mixed effects linear model. When looking at just the children, we checked for an interaction of modality and age which would signify age-related change in AV enhancement.

When comparing the adult and child groups, we included an interaction of age group and modality to determine whether there were group differences between adults and children in audiovisual benefit. In general, children's thresholds decreased with age and they had a two decibel audiovisual enhancement. But AV enhancement did not vary across the age range tested. Adults benefited significantly more than children with about 4 dB audiovisual enhancement. So there was no increase in benefit across school age, but there was greater benefit in the adults than in the children. On the measures that follow, we use the auditory only thresholds from this task shown in red as an individualized test signal to noise ratio for the other tests. So you could tell that younger children were tested at lower signal to noise ratios than the older children and the adults as well.

For our second measure, we assessed accuracy of auditory only and audiovisual sentence recognition in noise. Once again, there was a significant enhancement in children and in adults. However, there was no increase in benefit across school age. And in this case there was no difference in benefit between adults and children either. So what did predict individual differences if it wasn't age? Here I've plotted audiovisual enhancement, the difference between visual and auditory only accuracy on the vertical axis as a function of auditory only proportion correct. You can see that the audiovisual

enhancement is related to auditory only accuracy. So for this task rather than age audiovisual enhancement in school-aged children is predicted by auditory access by how much information you're getting just from the auditory only signal, even though the children were not performing at ceiling.

So it's not just that you're benefiting less because there was no room for growth. For the third task, we measured auditory only and audiovisual syllable recognition and noise. The stimuli, were C, V syllables, consonant vowel syllables with the vowel E and the carrier say. For example, say me, say key, we used all word initial English consonants excluding the voiceless fricative fi. For this last measure, there were significant audiovisual enhancement. I should mention also that we did test every child's articulation abilities before conducting this test. And we had flashcards so that if a child made a certain articulation error, we knew that in advance from our articulation test and we had pictures we could show them to say, I couldn't quite tell did you say that was key or did you see say that was T using pictures of a key and the letter T or a cup of tea.

And so the results you're seeing here should not be a reflection of differences in articulation ability across age. So for this last measure, there was a significant audiovisual enhancement. They did on average 7% better when they had the visual speech in children. There was no significant effective age, and that's because we use those individualized signal to noise ratios in the auditory condition. There's no change over age, but there is a significant audiovisual, excuse me, age by modality interaction showing us that there was an increase in benefit across school age for this particular task. There was also a significant audiovisual benefit in adults they showed a 23% benefit compared to the child's average of 7%. And so there was a group by modality interaction there.

Where greater benefit was observed in adults than in children. Okay, for this last measure, we actually also tested lip reading using our visual only speech recognition using the same stimuli. And so here I've added those results in blue, you can see that visual only accuracy was lower than auditory only accuracy. And that visual only accuracy also increased with age. All right, I'll share one more set of cross-sectional data with you. These are from a separate study completed by Elizabeth Gijbel who's a PhD student at the University of Washington. This online study included 161 English speaking children between four and 15 years of age. They're presented auditory only audio, audiovisual, and visual only words in noise and the auditory only and audiovisual conditions included signal to noise ratio of negative five, negative eight and negative 11 decibels.

Children were instructed to choose one of four pictures to indicate which word the talker said. Here audiovisual scores in green, auditory only scores in blue and visual only scores in red are plotted as a function of age. Overall accuracy significantly increased in all modalities with increasing age, but there was no modality by age interaction. So audiovisual enhancement did not increase once again across school age. And so with these two pretty large groups of children across four different measures, we now have kind of the summary of what is it that's happening developmentally in audiovisual speech enhancement between five and 16 years of age. Now to summarize so far only one out of four measures indicated an increase in audiovisual speech enhancement across school age.

This was surprising given that the past literature had shown age-related increases in audiovisual enhancement when comparing across age groups. The only task that showed an increase in audiovisual enhancement across school age was the syllable recognition task. There are a number of potentials reasons for this, but what I think is happening is that the syllable task was very explicitly designed to test children's use of visual phonetic knowledge. Their knowledge of what visual cues coincide with different

speech sounds, the use of those visemes. And so we found that age-related changes in audiovisual speech enhancement across school age are inconsistent across measures. But we're gonna look a little bit more closely at that syllable recognition task that was related to visual phonetic knowledge.

All right, so I showed you this continent confusion error analysis earlier when we were talking about what it means to use visual phonetic knowledge for audiovisual speech perception. We use this method to quantify adult and children's visual phonetic knowledge. So we took all of their responses on that syllable task in the visual only condition, and we plotted them out in this way. So remember the diagonal represents correct responses, but were really interested in the errors around that diagonal and how they group together. Okay, so in a previous study with adults and three and four year olds, we used this methodology and we compared as I showed you, between how many of your errors are within a viseme category.

So in the colored boxes here, those that P, B, me category or that wi, ri category or ki, gi, li, all going together. And we compared between errors that were within viseme category in the colored boxes here. And errors that were outside viseme categories or between viseme categories shown in gray here. And I showed you earlier that adults showed a much greater proportion of consonant confusions within viseme category for audiovisual conditions than auditory only conditions. We did the same thing with four-year-olds and three-year-olds. And we found that four-year-olds also showed a greater proportion of constant confusions that were within viseme category showing that they too were using visual phonetic knowledge during audiovisual speech perception.

The same was not true for three year olds. You can see that the difference between the auditory only condition in white, and the audiovisual condition in dark gray is much greater for the adults on the left than for the children in the middle and on the right.

And so now that we have similar data from children between five and 15 years of age, we can kind of see what's happening in development between these adults and these four year olds. So to look at this, we broke up our data from those 63 children and 18 adults into five age categories. So there are four roughly equal size groups of children and then a fifth adult group. And then we wanted to quantify and characterize children's visual phonetic knowledge.

So we examined the consonant confusions in each of these five groups. This is a figure I showed you earlier, and these are actually the data from the adults in this particular study. So this is the constant confusion matrices from combining all of the trials from the adults. You can see that the target syllables, as I said, are on the X axis and the responses are on the Y axis. We use something called agglomerative hierarchical clustering analysis to define those viseme categories present in the adult data. And in the matrix, the consonants are organized according to these viseme categories. So the gold boxes show the viseme categories present in the adult data. And on the left you can see exactly what those categories are.

So from left to right from bottom to top Excuse me, in the table on the left, you can see that the categories are organized primarily by place of articulation. So we have bilabial consonants on the bottom left. We have wi and ri together, followed by fi and ve labiodentals. The interdental V is on its own. Li categorized on its own as well. Post-alveolar sounds, shi, chi and gi grouped together. Alveolar sounds, si, zi, ti and di grouped together. And then the velar and glottal sounds, ni, ji, gi, ki, hi, all grouped together. And so like I said, this is all kind of by place of articulation essentially. So what we did is plot the same results for children to see how well these viseme categories fit the child data.

And here they are. The gold boxes once again show the viseme categories in the adult data. The percentage, excuse me, we plotted the results for the children in a similar

manner. You can see the adult clusters apply quite well to the child data. So we have older children on the right all the way down to our youngest children on the left. And you can see that although the clusters apply quite well to the child data as we go from older children on the right down to younger children on the left, the specificity decreases. There are more responses that are outside of those viseme categories, especially for consonants at the top of the graphs, which have more dorsal places of articulation are produced farther back in the oral tract.

What we did to really characterize in a simpler manner, what's going on here is to calculate the proportion of visual only responses within each of our viseme categories in each of the five groups. So here they are, you can see, so we're looking at age group on the X-axis and the proportion of responses within the viseme categories. You can see some viseme categories. Especially the ones produced at the front of the mouth with the lips are pretty well developed by five years of age in our youngest group. And they plateau by nine to 12 years of age. That includes the P, B, me, category, the wi, ri category and the shi, chi, gi category. The fi, vi category in yellow plateaus a little bit sooner than the others.

The V category shows a really drastic early change between five to seven and seven to nine and a half years. And then these more dorsal consonant categories, si, zi, ti, di and ni, ji, gi, ki, hi, show more change over age. And so overall what we see here is that there is some really early visual phonetic knowledge in place by five years of age. So children down to five years of age can do some lip reading. They understand something about what gestures go with which speech sounds, but there is change over the school years that we see where there's an increasing specificity for those more dorsal sounds. So that's just the lip reading part, right? That's only the visual only.

What we wanted to do next is say how are you applying that information to audiovisual speech perception? And so in this graph we use those viseme categories again to calculate the proportion of errors that were within viseme category for each of the three modalities. Just as we saw for the adults, the four-year-olds, and the three-year-olds. In other words we're asking when you made an error, how likely were you to substitute a phoneme that is visually similar. We see some strong age-related change in visual phonetic knowledge in the visual only condition on the left where the proportion of errors within viseme category on the Y axis increases pretty drastically over age on the X axis for the visual only on the left, we also see that occurring on the right for audiovisual stimuli.

So the proportion of errors within advising categories also increases over age in the audiovisual condition. By comparison, there's no change over age in the proportion of errors within viseme category in the auditory only condition, which makes sense 'cause you don't really have the visual information to narrow that down. We can also tell that the visual information is being used by comparing between the auditory only and the audiovisual condition. And so we see that real big difference between the proportion of errors within viseme category for audiovisual conditions in purple compared to auditory only conditions in red. So to summarize what I've shown you about typical development. There are some early developing signs of visual phonetic knowledge. Children as young as five years of age are showing us that they have some understanding of what speech sounds correspond to which visible gestures.

And there is also age related change in the use of visual phonetic knowledge across the school years, which could explain why we see these changes in audiovisual benefit on some measures, those that rely really heavily on visual phonetic knowledge or really isolate the use of visual phonetic knowledge. But we don't see those age-related changes on other measures that maybe don't rely as much on visual phonetic knowledge or don't isolate that particular mechanism quite as much. So the last thing I

wanna talk about today is the effective pediatric hearing loss on audiovisual speech enhancement. In particular, we're gonna talk about children with bilateral mild to moderately severe sensory and oral hearing loss. Primarily children who are using hearing aids, although we've seen a couple children with mild losses who are unaided.

So this work is actually motivated by a finding that visual speech enhancement is greater in six to 12 year old children who are hard of hearing than six to 12 year old children with typical hearing. And so here is an example of that from a recent study. We tested sentence recognition in noise in six to 12 year old children in these two age groups. The graph, excuse me, shows auditory only accuracy in white and audiovisual accuracy in gray. You can see that in children with typical hearing there was a 13% audiovisual benefit, whereas in children who are hard of hearing, there's a 28% audiovisual benefit. So audiovisual benefit was twice as large in children who are hard of hearing as children who had typical hearing, even though we used some methodologies that equated auditory access between, excuse me, equated auditory accuracy between the two conditions.

And so this is really promising obviously for children who are hard of hearing, who are going to struggle a lot more recognizing speech and noise. And our current research is really about examining why this is the case. So is it that children who are hard of hearing, does their experience with reduced or inconsistent acoustic access during the early critical period of brain plasticity fundamentally alter audiovisual speech processing development? Or is there a more stimulus driven explanation for what we're seeing here? Excuse me. To answer that question, we're gonna look at the super simple model of audiovisual speech perception. There are three main components, acoustic phonetic access, essentially, which speech sounds are audible, visual phonetic access, what information are you getting from lip reading?

And then the combination of auditory and visual cues. So each of these could impact how you perform on audiovisual speech perception. Now if we're talking about audibility, this first component, there are obviously measurable differences in acoustic phonetic speech access that depend on a child's configuration of hearing loss. If we're talking about lip reading, we're asking are children who are hard of hearing simply better lip readers? Or is it that children who are hard of hearing, weight auditory and visual speech cues differently than children with typical hearing. Are they better at shifting between relying on auditory cues and relying on visual cues? So first, looking at lip reading. The literature on lip reading in children with hearing loss is really mixed.

Whereas some studies suggest enhanced lip reading in children with hearing loss. Others have found no difference. There's a lot of variability in lip reading among children with hearing loss. And some have suggested that the group differences that we see in some of the studies are actually caused by a couple of extremely skilled lip readers in the group of children with hearing loss rather than an overall greater ability to lip read among children who have hearing loss. What is clear is that children who are hard of hearing show similar consonant confusions as children who have typical hearing, excuse me. So both children with typical hearing and children who are hard of hearing. Are better at distinguishing a consonants place of articulation than it's voicing or manner information when they're using only visual cues.

The study I wanna share with you next really focuses more on this second component. The acoustic phonetic speech access or audibility. So this figure, which I recreated from Grant Walden & Seitz 1998, demonstrates how differences in acoustic cues you have access to could influence audiovisual benefit. These are data from two imaginary subjects. The first on the left has perfect access to acoustic place information. The second on the right has perfect access to acoustic voicing and manner information. They have the same access to visual speech. So their their auditory accuracy in blue is similar and their visual accuracy in red or orange is the same. However, because the

cues that are provided by the visual signal are redundant with the acoustic place cues in the subject on the left, there isn't very much audiovisual benefit.

The predicted audiovisual scores are rather low and pretty similar to the auditory and visual scores. However, in the imaginary subject on the right, the acoustic voicing and manner cues are complimentary to the visual place cues. Resulting in a much greater audiovisual score. So we think this could be what's happening in children who are hard of hearing, who have decreased access to high frequency audibility particularly. So decreased high frequency audibility could lead to a decrease in access to acoustic place cues, which decreases redundancy between the auditory and visual speech, thereby increasing the potential for audiovisual benefit, which might be why they have greater audiovisual benefit or audiovisual enhancement. The alternate explanation is that early experience with reduced acoustic access either increases their lip reading ability or increases their reliance on visual speech, which in turn also increases audiovisual benefit.

As a first step towards assessing which of these might be the case in assessing the role of frequency specific audibility in explaining differences between children with typical hearing and children who are hard of hearing. We conducted a study that examines the extent to which acoustic frequency content influences auditory and audiovisual word recognition and sentence recognition in children with normal or typical hearing. So today I'm gonna show you data from 30 children. With typical hearing who are between seven and 12.9 years of age, they all had normal or corrected to normal vision. Children completed hearing, vision, and articulation screenings, followed by the speech perception testing. Their job was to repeat words and sentences presented via computer. And the articulation screening results were once again used to guide our testers in seeking clarifications about potentially ambiguous responses.

The stimuli in this case were phonetically balanced lists of consonant vowel consonant words within the productive vocabulary of children by seven years of age. They were presented in auditory only, and audiovisual conditions with acoustic speech, either low-pass filtered or high-pass filtered with a two kilohertz cutoff frequency. Here's an example of the low-pass stimulus

- [Subject] Ready, shell.

- [Kaylah] And a high-pass stimulus.

- [Subject] Ready, shell.

- [Kaylah] She said ready shell in both of those, you might've noticed that for the low-pass filtered speech, it was really hard to tell that s from the sh. And so that's kind of one of the impacts of low-pass filtering there. We also included a visual only condition. Participants heard the word ready and then they had to lip read the target, which was presented silently

- Ready.

- [Kaylah] So we hypothesize that for, given we're looking at the stimulus driven explanation, we thought that there would be less redundancy between auditory and visual speech queue access in the low-pass filtered condition than the high-pass filtered condition. Due to high access to place cues in visual speech and decreased access to place cues in the low-pass filtered condition. Our second hypothesis was that greater audiovisual benefit would occur when acoustic speech was low-pass filtered than when it was high-pass filtered. So what this would look like is an interaction of filter condition and modality where the slope for the low-pass filter condition is steeper than the slope for the high-pass condition. If this turned out to be

the case, it would suggest to us that maybe children who are hard of hearing are showing greater audiovisual benefit because of differences in redundancy between accessible auditorium and visual cues rather than some fundamental difference in the development of audiovisual speech processing.

So looking at the data to test the first hypothesis. Here I've plotted accuracy of continent feature recognition with the X-axis showing place of articulation on the left and voicing and manner of articulation on the right. The blue circle shows the auditory only low-pass filtered condition. The red square shows the auditory only high-pass filtered speech. And the teal diamond shows results for the visual only lip reading. Lip reading of place information is much more accurate than lip reading of voice and manner information. Additionally, the effect of the low-pass filtering on auditory place cues are greater than the effects of high-pass filtering on auditory place queues. This supports our hypothesis that there's less redundancy between visual speech queue access and low-pass filtered acoustic speech than visual speech queue access and high-pass filtered acoustic speech.

Here I'm showing you some pilot data from children with typical hearing in green. And children who are hard of hearing in blue, listening to auditory only speech and noise. The children who are hard of hearing were tested at zero dB signal to noise ratio. And the children with typical hearing were tested at individually set signal to noise ratios, which were lower than zero dB signal to noise ratio. The X axis and Y axis are the same as in the graph on the left. And what I want you to see is that where is the children with typical hearing have approximately equal accuracy for place and voice manner. The children who are hard of hearing in blue have greater accuracy for voicing in manner than for place.

So this resembles what we see in blue on the left for children listening with typical hearing, listening to low-pass filtered speech, suggesting that the low-pass filtered

speech is more aligned with this group of children who are hard of hearing in that this is a pretty good sort of way of maybe looking at what might be happening using children who have typical hearing. And then of course you can see that in both children with typical hearing and children who are hard of hearing on the right, that visual speech is providing greater access to place cues than voice manner cues. So the low-pass filtered condition, what I want you to take away here, the low-pass filtered condition has decreased access to place cues which has decreased the redundancy between the auditory and visual speech.

And so our question next is, does that result in greater audiovisual benefit? All right, so to test this hypothesis, we look at whether there's an interaction of modality and filter condition for auditory only and audiovisual conditions. So here we're looking at whole word recognition accuracy. The X axis shows auditory only on the left audiovisual in the middle and visual scores on the right. The low-pass filtered condition and high-pass filtered conditions are once again shown in blue circles and red squares. And the visual only is shown in teal for reference. There is a significant audiovisual benefit. There's also interaction of modality and filter, which shows that audiovisual benefit at the word level is greater for low-pass filtered speech than for high-pass filtered speech.

And so this follows our hypothesis that we would see greater audiovisual benefit for low-pass filtered speech due to that decrease in auditory visual redundancy. Here, consonant accuracy is shown on the left, and vowel accuracy is shown on the right. So we took the same words and we broke them up into consonants and vowels. What you might notice here is that there is audiovisual benefit for both consonant and for vowels. There is an effective filter in both conditions. So low-pass filtering has a greater impact on consonants and high-pass filtering has a greater impact on vowels. There's also a, excuse me, a modality by filter interaction where for consonants we see greater audiovisual benefit for the low-pass filtered condition than the high-pass filtered condition.

However, AV scores are a little bit close to ceiling, so we can't totally rule out that ceiling effects are the cause of that interaction. The last thing I wanna show you is we're gonna break up the consonants now into place and voice and manner cues. So here we're showing place on the left and voicing and manner on the right. We once again see audiovisual benefit for both place and voicing and manner cues. There is an effect of filter in both conditions, but it particularly in the place condition because voicing and manner such high accuracy. So for place cues, once again, we see that greater effect of the low-pass filtering. And we also do in fact see that interaction of modality and filter.

So particularly for place of articulation, we're seeing greater audiovisual benefit for that low-pass filtered speech than the high-pass filtered speech. One last bit of data forgot this was in here. We also asked children to do the same task using sentences. So the same conditions, auditory only, and audiovisual with low-pass and high-pass filtered sentences. Here are examples,

- [Subject] This string can stay in my pocket.

- [Kaylah] That was low-pass.

- [Subject] The string can stay in my pocket.

- [Kaylah] and that's high-pass. The talker said the string can stay in my pocket. So here for sentences, we once again see a significant audiovisual benefit. It's about 9% overall, we see the effective filter, but actually in this case, high-pass filtering has a greater impact on performance than low-pass filtering. And there's no significant difference in audiovisual benefit across the two filter conditions. All right, so to summarize, we found less redundancy between auditory and visual speech queue

access in low-pass filtered conditions than in high-pass filtered conditions. We also found greater audiovisual benefit when acoustic speech is low-pass filtered than when it's high-pass filtered when examining perception of words phon names, consonants and consonant features, especially place of articulation.

This preliminarily suggests that some differences between children with typical hearing and children who are hard of hearing are due to differences in redundancy. However, there were situations where we couldn't fully rule out that this was caused by ceiling effects. We did not find greater audiovisual benefit when acoustic speech was low-pass filtered than when it was high-pass filtered when we were examining perception of vowels and perception of sentences. And this preliminarily suggests that there are some differences between children with typical hearing and children who are hard of hearing that are not simply due to differences in redundancy because we still see that increased audiovisual benefit in children who are hard of hearing at the level of sentences and not just for words and for consonants.

So overall, what I hope you take away from this last set of data is that frequency specific aided audibility may impact the degree to which children who are hard of hearing benefit from visual speech. Our data suggests that visual speech may be particularly helpful for children who have decreased high frequency audibility. Now our ongoing studies are going to test that further by focusing specifically on looking at audiovisual benefit and Q waiting and frequency waiting in children who are hard of hearing, who have varying audiograms. So for a very quick summary, what I showed you today is that visual speech cues can help to compensate for environmental noise and for hearing loss. They help in quiet conditions.

And when the listening task is difficult or the speech is complex as well. Visual speech cues also help because they tell you when to listen. They provide cues about the auditory amplitude envelope of speech. They help you to access audiovisual speech

representations, and they can help you to segregate and attend to a target talker. Regarding development. Infants, children and adults all show benefit from visual speech cues. The cues that we use for visual speech seem to differ across development potentially. We saw that school-aged children who are hard of hearing benefit more from visual speech in noisy conditions than children with typical hearing. This might be due to enhanced lip reading skills in a subset of children.

It might also occur because decreased high frequency audibility decreases the redundancy between auditory and visual speech increasing the potential for audiovisual benefit. It is also possible that early experience with reduced acoustic access causes children who are hard of hearing to rely more on those visual speech cues or makes them better at choosing when to rely on auditory versus visual cues. So I wanna thank you all for listening and I'm happy to take any questions you may have.

- [Christy] Thank you so much Dr. Lalonde. We're gonna go ahead and open up the floor for any questions or comments that you might have.

- [Kaylah] All right, if there are no comments, I wanted to again, just take a moment to acknowledge all my collaborators and the funding from the National Institutes of Health that makes this work possible. We're really looking forward to the results from the children who are hard of hearing in the coming months. We hope to be done collecting those data by the end of the summer. What we're doing actually is, so instead of doing things like filtering, we're testing kids both with typical hearing and who are hard of hearing in noise and kinda looking at those patterns of errors again, at whether they're making errors in place of articulation, voicing or manner across auditory and audiovisual conditions.

But we're also going to do something called frequency importance weighting testing. Where we what we've found so far and what others have found in adults is that when

you are listening to auditory speech, so it's well established from other researchers that listening to auditory speech, what we see is that around the two kilohertz region is super important for speech perception. That's kind of where the weighting comes from for speech intelligibility index and things. But when you add audiovisual, when you add visual speech to the mix, the importance of the low frequencies becomes higher and the importance of the high frequencies becomes relatively lower. And so what we wanna see is that if children know how to change their weighting of the frequencies, depending on whether that visual information is there or not, and whether the weighting that children who are hard of hearing use in auditory only conditions might be more aligned with audiovisual conditions.

So they're just kind of naturally queued up for a sort of audiovisual listening paradigm. So that's whats to look forward to in the future.

- [Christy] Thank you again to Dr. Kaylah Lalonde and also to the American Auditory Society. We hope you enjoy today's presentation, and feel free to tune into the other recorded courses in partnership with AAS. Until next time, thank you everyone.