PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Audiovisual Speech Perception: Time for a Paradigm Shift

Nancy TYE-MURRAY[1]

[1] Washington University School of Medicine, USA

ABSTRACT

Most adults cannot lipread very well, and when viewing only the visual speech signal, recognize very little of the content. Even so, the visual speech greatly enhances a degraded auditory speech signal, as when listening in the presence of background noise or with hearing loss. Although previous researchers have suggested that this super-additive effect is due to a distinct audiovisual integration ability, recent findings indicated that it can be accounted solely by unimodal performance.
Keywords: Audiovisual, integration, lipreading

INTRODUCTION

Although researchers often focus on the auditory aspect of speech perception, in daily conversation speech perception is typically an audiovisual perceptual event because most conversations and other forms of spoken communication occur face-to-face. Moreover, whether individuals have normal or impaired hearing, there are times when they are highly reliant on lipreading for speech understanding, as when talking in a noisy restaurant or when talking with wind noise. Interestingly, although most people are relatively poor lipreaders---they can recognize less than 20% of words in sentence context—the benefit they receive when the visual signal is added to the degraded auditory signal is super-additive, meaning the number of words they can recognize is greater than the sum of how many words they can recognize in an auditory-only and vision-only condition. In this paper, we'll review some of the findings about audiovisual speech perception and then consider the super-additive effect.

REVIEW

In a study performed at Washington University in St. Louis, older adults recognized less than 10% of the test words in sentences when present in a vision-only condition and younger adults recognized only about 15% (Sommers, Tye-Murray, & Spehar, 2005, Figure 1). Performance in the auditory-only condition was "controlled" by presenting background babble, so on average, participants recognized 40% of the words. When the visual speech signal was added to this degraded auditory signal, performance jumped to over 65% words correct.

[1] nmurray@wustl.edu

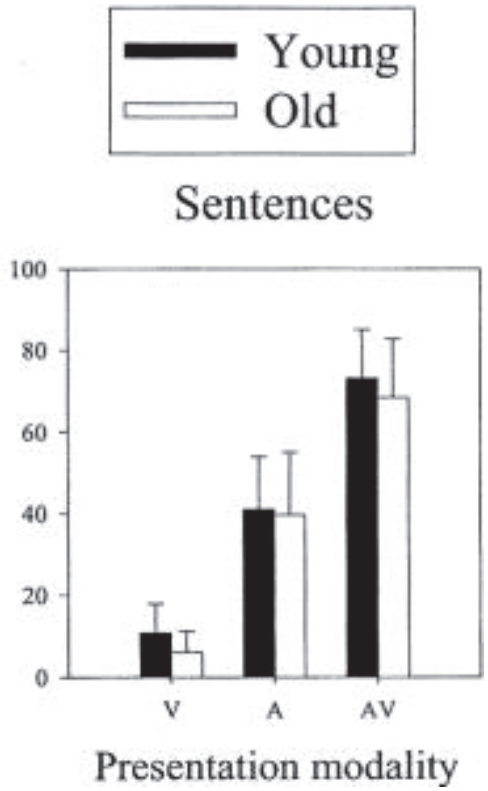

Figure 1. Words correct in sentence context, as presented in three different conditions (V: vision-only; A: audition-only; AV: audiovisual) for younger adults (ages 18-30 yrs) and older adults (age 65 yrs and older) (Adapted from Sommers et al., 2005).

This early experiment captures three robust findings that we have since found in other studies (e.g., Tye-Murray, Sommers, & Spehar, 2007a; Tye-Murray, Sommers, & Spehar, 2007b; Tye-Murray, Spehar, Myerson, Hale, and Sommers, 2016). First, vision-only speech recognition is very difficult and most people are not very good at it. Second, age causes vision-only speech recognition to decline. Older adults are significantly poorer lipreaders than are young adults. And thirdly, performance in an audiovisual condition is super-additive. This super-additive effect is sometimes referred to as the "audiovisual speech advantage." As Figure 1 indicates, younger and older adults tend to have similar degrees of the audiovisual speech advantage. That is, although younger adults have overall better audiovisual speech recognition than older adults, they have about the same amount of super-additive effect.

Not only does the audiovisual speech advantage appear unaffected by aging, but it also appears to be unaffected by learning. For example, we compared the performance of adults who have normal hearing and adults who have acquired hearing loss, meaning that they began to lose their hearing in adulthood. We reasoned that acquired hearing loss might afford increased practice with lipreading and force increased reliance on the visual signal for speech perception. We found that older adults who had hearing loss performed the same on tests of word and sentence recognition than older adults with normal hearing when tested in vision-only and audiovisual test conditions. Figure 2 shows performance for the sentence task. As in the Sommers et al. (2005) paper, the auditory-only condition was controlled by presenting background babble so participants would be equated in this condition.

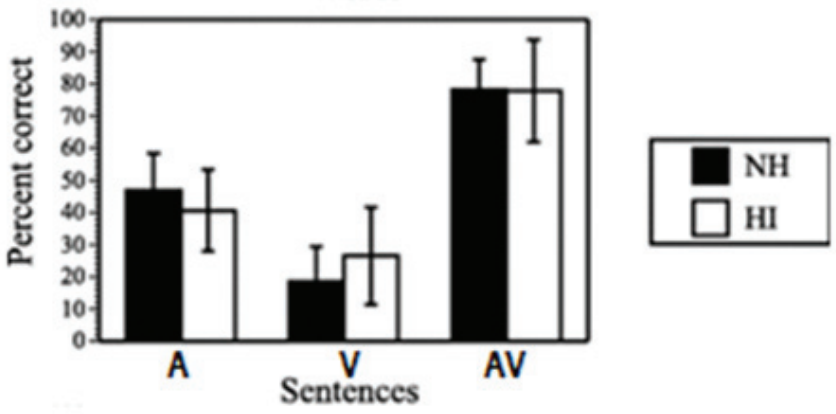

Figure 2. Words correct in sentence context, as presented in three different conditions (V, A, AV). (Adapted from Tye-Murray et al., 2007a).

From these two studies, we can conclude that the audiovisual speech advantage appears to be unaffected by either aging or the presence of hearing loss (and implicitly, by the opportunity for perceptual learning). These findings led our research team to further question the basis for the super-additive effect of auditory-only and vision-only speech perception on audiovisual speech performance.

Some researchers have suggested that an individual's ability to integrate auditory and visual speech perception may account for the audiovisual speech advantage, and that some people are simply better at integrating the auditory and visual speech signals than are others. In this view, it is postulated that audiovisual integration occurs as a distinct stage of the speech perception process (e.g., Grant, Walden, & Seitz, 1998; Massaro, 1998; Ouni, Cohen, Ishak, & Massaro, 2007). For example, Grant et al. (1998) proposed that audiovisual speech perception entails three sequential stages. First, there is a stage for perceiving the auditory and visual speech cues, then there is a stage for integrating the two kinds of speech cues, and finally, there is a stage for accessing the mental lexicon. Some people have greater ability at this second stage, integration, than do others, and this accounts for the variability in audiovisual speech perception, at least in part.

However, because neither age nor practice appears to affect one's ability to integrate, we sought to determine whether we could account for the super-additive effect without postulating an integrative ability and a distinct integration stage in the speech perception process. To this end, we tested a group of participants who ranged in age from 20 years to 93 years using a matrix sentence test. A matrix sentence test is a test that has a closed set of words embedded in a sentence context, "The ____ and the ____ saw the ____ and the ____." Use of a matrix test format precludes floor effects in a vision-only condition.

We equated participants' performance in an auditory-only condition, as in our previous experiments, by adding background babble until they reached about a 30% words correct performance. In a similar vein, we equated their vision-only performance by blurring the visual speech signal to the point where each participant could recognize about 30% of the words. As shown in Figure 3, when unimodal performance was equated, participants achieved similar audiovisual word recognition scores and similar audiovisual speech advantage. This finding suggests that auditory-only and vision-only speech recognition abilities alone can account for an individual's audiovisual speech recognition in noisy situations and that neither age nor an assumed integrative ability further enhances predictions.

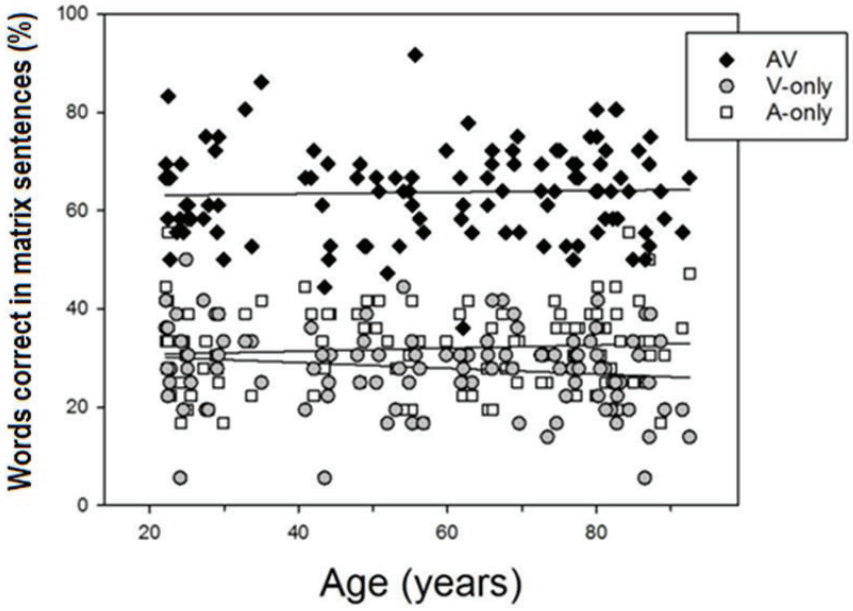

Figure 3. Percent word correct for participants who range in age, where performance was equated in the two unimodal conditions, V-only and A-only (Adapted from Tye-Murray et al., 2016).

To further bolster this conclusion, we performed a regression analysis. We found that after vision-only and auditory-only speech recognition scores were entered into the regression model, adding age as a predictor did not significantly increase the model's ability to account for variance in audiovisual performance. Based on these analyses, we concluded that the audiovisual speech advantage may stem from the complementary and reinforcing nature of the auditory and visual signals, and not to some integrative ability, and that age may have little direct influence apart from its effects on unimodal perceptual abilities.

Recently, we analyzed the performance on a word test completed by this same group of participants. We performed a nonlinear regression analysis using percent word correct scores that were obtained in an auditory-only, vision-only, and audiovisual condition (Myerson et al., in review). We considered participants' accuracy in the audiovisual condition as a function of the probability of their getting words correct in that condition as estimated based on the participants' accuracy in the auditory-only and vision-only conditions. Unimodal performance predicted 88% of the variance. Taken together with the Tye-Murray et al. (2016) results, these two studies, which differed in testing conditions and speech material, demonstrate that listening and lipreading abilities are sufficient to predict an individual's audiovisual speech recognition in noisy situations and that neither age nor an assumed integrative ability further enhance predictions.

CONCLUSIONS

The studies reviewed in this paper suggest that even though lipreading is difficult, the visual speech signal is important for face-to-face communication. The visual signal greatly enhances one's ability to perceive speech in difficult listening conditions, such as in the presence of background babble. In addition, aging diminishes one's ability to take advantage of visual speech information and simply having hearing loss does not make you become a better lipreader.

Finally the decline in older adults' audiovisual speech perception can be attributed almost entirely to declines in unimodal speech performance and not to some audiovisual integrative ability.

ACKNOWLEDGEMENTS

Support was provided by grant AG018029 from the National Institutes of Health.

REFERENCES

1. Grant KW, Walden BE, Seitz PF. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. J Acoust Soc Am. 1998;103(5):2677-2690.
2. Massaro D. Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, MA: MIT Press;1998.
3. Ouni S, Cohen MM, Ishak H, Massaro DW. Visual contribution to speech perception: measuring the intelligibility of animated talking heads. EURASIP Journal on Audio, Speech, and Music Processing. 2006;2007(1):047891.
4. Sommers M, Tye-Murray N, Spehar B. Audiovisual integration and aging. Ear and Hearing. 2005;26(3):263–275.
5. Tye-Murray N, Sommers MS, Spehar B. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. Ear and Hearing. 2007a;28(5):656-668.
6. Tye-Murray N, Sommers MS, Spehar B. The effects of age and gender on lipreading abilities. Journal of the American Academy of Audiology. 2007b;18(10):883-92.
7. Tye-Murray N, Spehar B, Myerson J, Hale S, Sommers M. Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. Psychology and Aging. 2016;31(4):380.