

4006. Learning to integrate auditory and visual information in speech perception

Joseph D.W. Stephens & Lori L. Holt

Psychology Dept., Carnegie Mellon University, and Center for the Neural Basis of Cognition



1 Introduction

Audiovisual speech

Visual information from speakers' faces greatly affects speech perception.

Visual information can improve intelligibility (Sumbly & Pollack, 1954)

Visual information can change identification of speech segments entirely (McGurk & MacDonald, 1976)

2 Novel visual cues

Can speakers learn new A/V associations?

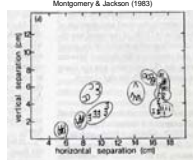
Stimuli

We used novel visual stimuli to convey phonetic information.

A "speech robot" contained moving parts synchronized with acoustic VCV tokens.

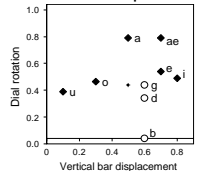
/aba/

Natural lip apertures (vowels)



The robot's parts moved to positions defined by phonetic categories. The combinations of cues were derived from natural mouth positions, but bore no resemblance to them (see figures). Consonants /b/, /d/, /g/ varied only in the position of the rotating component.

Robot's movement parameters



Examples of stimuli may be viewed at: <http://www.andrew.cmu.edu/~jds2/robot.html>

Example: the "McGurk Effect"

Auditory Visual Perceived
/ba/ + /ga/ → "da"

Is A/V integration learned?

"NO" – Integration happens because auditory and visual information specify same event (Fowler, 1986)

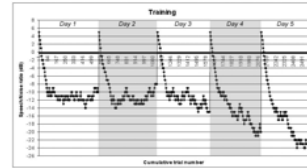
Prediction: Only articulatory cues can be integrated with auditory speech

"YES" – Integration = finding the best match across auditory & visual categories (Massaro, 1998)

Prediction: Other visual patterns may be integrated with auditory speech

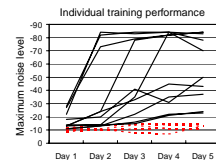
Training

Fifteen participants were trained to identify consonants in an adaptive bimodal task, across five sessions.



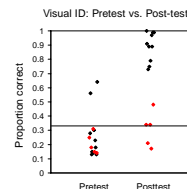
Participants identified the consonant spoken by the robot on each trial. Correct responses resulted in increased acoustic noise.

The training task forced participants to use visual cues from the robot.



Across multiple days, most participants learned to identify the consonants at increasingly higher noise levels.

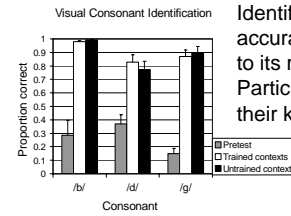
After training, most participants were able to identify consonants based on visual cues alone.



3 Results: Post-training

Visual cue learning

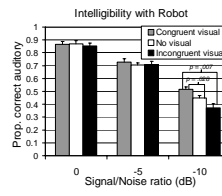
Participants' responses to novel visual cues were consistent with structure of stimulus space



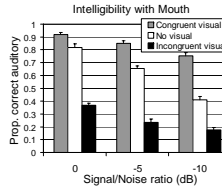
Identification of /b/ was more accurate than /d/ and /g/ due to its relative distinctiveness. Participants also generalized their knowledge of visual consonants to untrained vowel contexts.

Intelligibility

After training, visual cues from the robot improved intelligibility of consonants in noise



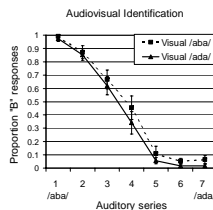
At high noise levels, congruent visual cues from the robot improved identification, and incongruent cues impaired identification, relative to a static picture of the robot.



The effects of viewing the robot on intelligibility were similar to the effects of lipreading, though smaller.

Audiovisual categorization

Visual cues from the robot influenced the identification of ambiguous consonants

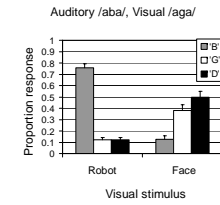


Along an auditory series ranging from /aba/ to /ada/, participants reported hearing /b/ more often when the accompanying robot video showed visual cues for /aba/ rather than /ada/.

($F(3,24) = 6.45, p = .002$; some data not shown)

McGurk effect?

Integration of visual cues from the robot did not produce a McGurk effect



When auditory /aba/ was paired with robot video of /aga/, participants rarely reported hearing /d/. When they viewed video of a speaker's lips, /d/ was most often reported (data from intelligibility post-tests).

4 Conclusions

1. Non-gestural visual speech cues can be learned and integrated with auditory speech

Two-thirds of participants successfully learned artificial visual cues for consonants

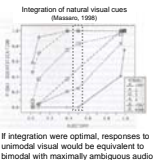
Newly-learned visual cues affected speech perception similarly to natural visual cues

2. Effects of newly-learned visual cues on speech perception are small

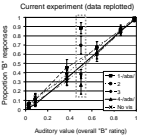
Intelligibility influenced only at high noise levels

Visual influence insufficient for McGurk effect

Sub-optimal use of visual cues with ambiguous auditory information (see figures)



If integration were optimal, responses to unimodal visual would be equivalent to bimodal with maximally ambiguous audio.



Future questions

Can larger effects or optimal integration be developed through more training?

Can effects be manipulated by changes in the stimulus space?

References

Fowler, C. A. (1986). *J. Phonetics*, 14, 3-28.
 Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
 McGurk, H., & MacDonald, J. (1976). *Nature*, 264, 746-748.
 Montgomery, A., & Jackson, P. (1983). *J. Acoust. Soc. Am.*, 73(6), 2134-2144.
 Sumbly, W. H., & Pollack, I. (1954). *J. Acoust. Soc. Am.*, 26, 212-215.