

B-293. Experience-Driven Audiovisual Integration 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; can change identification of speech segments entirely (Sumbly & Pollack, 1954; McGurk & MacDonald, 1976)

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

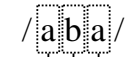
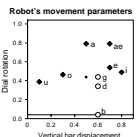
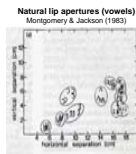
2 Novel visual cues

Stimuli

We used novel visual stimuli to convey phonetic information.

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

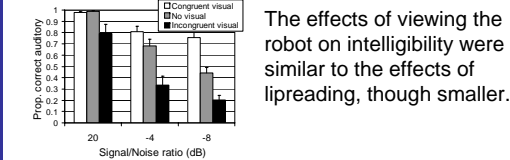
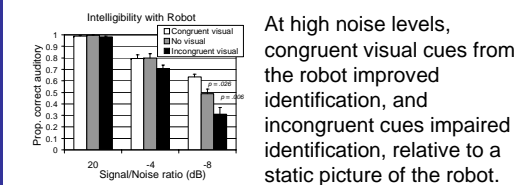
The robot's parts moved to positions defined by phonetic categories. The combinations of cues were derived from natural mouth positions, but bore no resemblance to them (see figures). Consonants /b/, /d/, /g/ varied only in the position of the rotating component.



3 Results: Post-training

Intelligibility

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

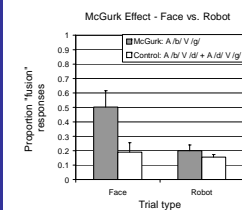


At high noise levels, congruent visual cues from the robot improved identification, and incongruent cues impaired identification, relative to a static picture of the robot.

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

McGurk effect?

Integration of visual cues from the robot hinted at a McGurk effect



When auditory /aba/ was paired with robot video of /aga/, participants reported hearing /d/, but not significantly more than in a strict control condition (data from intelligibility post-tests).

Audiovisual categorization

The influence of visual information developed over the course of training

Participants identified stimuli that consisted of an auditory series ranging from /aba/ to /ada/, paired with robot video ranging from /aba/ to /ada/.

An effect of visual information appeared after the third phase of training (Test 2) and developed into an interaction between auditory and visual information after the fifth phase of training (Test 4). The interaction is characteristic of natural AV integration (Massaro, 1998).

4 Conclusions

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

Participants successfully learned artificial visual cues for consonants

Over time, use of novel visual cues became more speech-like

2. Effects are small, but improved relative to a previous experiment we conducted

Intelligibility influenced only at high noise levels

McGurk effect not significant

Sub-optimal use of visual cues

Future research

Account for the development of AV integration using computational models

Study effects of manipulating stimulus space

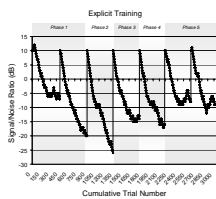
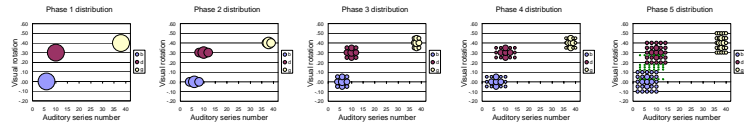
Evaluate integration in other tasks

References

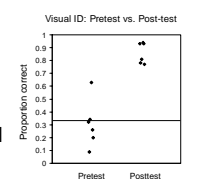
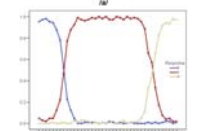
Fowler, C. A. (1986). *J. Phonetics*, 14, 3-28.
 Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
 McGurk, H., & MacDonald, J. (1976). *Nature*, 264, 746-748.
 Montgomery, A., & Jackson, P. (1983). *J. Acoust. Soc. Am.*, 73(6),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

Training Six participants were trained to identify consonants across multiple sessions.

In an **explicit task**, participants identified the consonant spoken by the robot on each trial. Correct responses resulted in increased acoustic noise. In an **implicit task**, participants watched AV stimuli while monitoring for "robot malfunctions." The tasks alternated daily.



The distributions of AV combinations broadened gradually throughout training. Auditory stimuli were taken from a set of 40-member series created from natural tokens. AV integration was tested after all training phases except the first.



The overall length of training depended on performance in the explicit task.

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

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