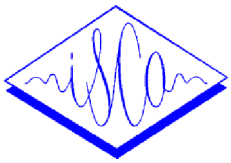

Auditory-Visual Speech Processing
2005 (AVSP'05)
British Columbia, Canada
July 24-27, 2005

USING GRAPHICS TO STUDY THE PERCEPTION OF SPEECH-IN-NOISE, AND VICE VERSA.

Harold Hill[1] and Eric Vaikiotis-Bateson[2],

1. Department of Vision Dynamics, Human Information Science Labs., ATRi, Kyoto, Japan.
2. Department of Linguistics, University of British Columbia, Canada.

ABSTRACT

This work aims to use the speech in noise task to assess talking head animations, and talking head animations to investigate the perception of speech-in-noise. The theoretical aim is to determine what visual information is important for speech, while the practical aim is to develop an effective talking head animation system adaptable to robots.

The first experiment used the "cuboid", a deliberately abstract face. Head, jaw and mouth movement were presented separately and in combination. Results showed an advantage of mouth movement independent of the other factors. This shows that even an abstract structure can carry useful facial speech information and that mouth movement is an essential component.

Two other experiments reported used ATR's in-house animation system [1] to look at the relative contribution of face and head movement. The first experiment replicated a combined head and face movement advantage [2]. A 2 Head Movement (present/absent) x 2 Face Movement (present/absent) experiment showed a main effect of face movement, but no effect of head movement or any interaction.

We conclude that abstract faces can carry useful visual speech information and that, while mouth and face movement are primary, head and jaw movement do not interfere with and can help.

1. INTRODUCTION

The combination of faces and voices is so central to human communication that a controllable, synthetic talking head would be an invaluable tool for many applications including research. By asking basic questions about visual speech perception we are seeking to understand what is needed for such a system.

Previous work using animations has shown speech-in-noise advantages both when driven by visemes [3,] and motion capture data [2]. While far from realistic, these animations are intended to look like faces. Little work has been done with deliberately abstract representations. One exception is an abstract animation of skull plus lips driven by parameters derived from video produced significant speech-in-noise advantages, although these were less than those for a whole face model driven by the same parameters [5]. Informal comparison of "Polar Express" and "The Incredibles" suggests that deliberately unrealistic animations may often be more effective [4].

In this paper we test perception of facial speech using both a deliberately abstract structure and a more realistic head model.

2. GENERAL METHODS

All animations were based on a set of Northern Digital Optotrak recordings of movement and audio from a male native speaker reading a set of Japanese sentences as described previously [2]. Stimuli were played back in babble noise using a computer controlled dvd, mixing console, monitor and speakers. Performance was calculated in terms of the number of kana correct (each kana corresponding to a mora in Japanese). Signal to noise ratio was adjusted for each subject using a different set of sentences spoken by the same speaker to give approximately 40% performance for audio only.

3. CUBOID EXPERIMENT

The aim of this experiment was to determine whether an abstract shape can convey useful speech-in-noise information and, if so, which components of the movement are used. Motion capture data imported into Maya was used to drive rigid movement of the head, articulated movement of the

jaw and non-rigid movement of the mouth through eight vertices. The experiment was run as a 2 Head movement (present/absent) x 2 Jaw Movement (present/absent) x 2 Mouth Movement (present/absent) design.

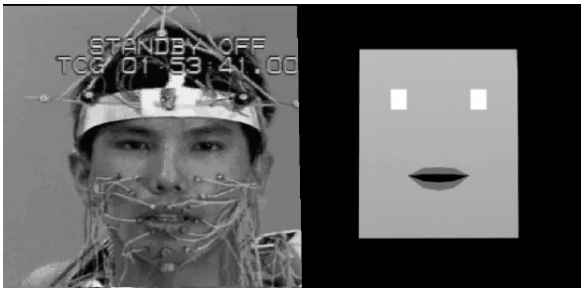

Figure 1. The marker setup used for motion capture (left) and the "cuboid" (right) animation used as stimuli. The cuboid was produced in collaboration with the Visualization Design Institute, Sheridan College.

3.1 Results and discussion

Results showed a main effect of mouth movement, $F(1,15) = 17.1$, $p<<.05$. (With mouth 58%, without 52%, both SEM's 2%). This shows that an abstract structure can convey useful facial speech when it includes a simple, 2D, model of mouth movement. Jaw or head movement alone or combined are not sufficient to improve speech in noise performance, but they do not interfere with recovery of mouth movement.

4. ATR ANIMATION SYSTEM

Two experiments assessed the relative contribution of head and face movement using a face model derived from the speaker. The ATR animation drives the face movement in different way than the cuboid, using marker data to estimate position in the space of possible face postures rather than by direct movement of vertices [1].

The first experiment included three movement conditions: no movement, face movement alone and face plus head movement. ANOVA showed a marginally significant effect of type of movement, $F(2,30) = 3.1$, $p<0.1$, with paired t-tests giving a significant difference between the combined head and face and auditory alone conditions, $t(14) = 2.3$, $p<0.05$. Performance is summarized in Table 1.

The second experiment used a 2 Head Movement (present/absent) x 2 Face Movement (present/absent) design to include a test of whether head movement alone is useful for this task. Results given in Table 1 showed a significant main effect of face movement $F(1,18) = 82.4$, $p<<.05$, but no effect of head movement or any interaction. Head movement alone does not appear to be sufficient (although see [6]). We are current running a control to determine why face movement alone was as good as head + face in this experiment.

Exp.	A only	F only	H only	H + F
ATR 1	39 (4)	42 (4)	N/A	49 (4)
ATR 2	44 (3)	58 (4)	46 (4)	60 (3)

Table 1. Results using ATR animation system. % correct (sem) for Audio, Face and Head movement.

5. GENERAL DISCUSSION

These experiments showed that even deliberately abstract animations give significant speech in noise advantages and suggests that some useful facial speech is perceived independently of facial structure. Non-rigid mouth and face movement appear critical for facial speech but, in contrast to automatic tracking, the human visual system can recover these components regardless of head and jaw movement. Rigid head movement can, as has been reported before [2], even facilitate performance.

6. REFERENCES

[1]. 1. Kuratate, T., H.C. Yehia, and E. Vatikiotis-Bateson. *Kinematics-based synthesis of realistic talking faces.* in *International conference on audio-visual speech processing*. 1998. Terrigal-Sydney, Australia: Causal Productions.

[2]. 2. Munhall, K., J.A. Jones, D.E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, *Visual prosody and speech intelligibility.* Psychological Science, 2004. **15**(2): p. 133-137.

[3]. 3. Massaro, D., *Perceiving talking faces: from speech perception to a general principal*. Bradford books series in cognitive psychology. 1997, Cambrudge, MA: MIT Press.

[4]. 4. Katzman, A. *Mimesis and praxis in the art of traditional facial animation*. in *ATR symposium on the cross-modal processing of faces and voices*. 2005. Kyoto, Japan.

[5]. 5. Benoit, C., B. Guiard-Marigny, B. Le Goff, and A. Adjoudani, *Which components of the face do humans and machines best speechread*, in *Speechreading my humans and machines*, D.G. Stork and M.E. Hennecke, Editors. 1996, Springer-Verlag: Berlin Heidelberg. p. 316-328.

[6]. 6. Kim, J. and C. Davis. *Perceiving speech related head and upper-face movements*. in *ATR Symposium on the cross-modal processing of faces and voices*. 2005. Kyoto, Japan.