TESTING AUDIO-VISUAL FAMILIARITY EFFECTS ON SPEECH PERCEPTION IN NOISE

Jeesun Kim & Chris Davis

MARCS Auditory Laboratories, University of Western Sydney, Australia
j.kim@uws.edu.au; chris.davis@uws.edu.au

ABSTRACT

The current study investigated the effect of familiarity with the voice and face of a talker on a subsequent speech intelligibility task. Participants were first familiarized with four animated characters that they learned to identify by voice, by face and by both face and voice. Ceiling level performance was reached. Following this training, participants were given a speech in noise identification task. Three types of talker condition were tested: familiar voice with familiar face, familiar voice with unfamiliar face, and unfamiliar face and voice. The results showed speech perception was more accurate in the familiar voice and face condition compared to the unfamiliar face and voice condition (a talker familiarity effect). Performance in the familiar voice with unfamiliar face condition did not differ from the unfamiliar face and voice baseline. In part, these results support the proposal that talker familiarity effects arise as a product of exemplar-based processing in speech recognition.

Keywords: talker familiarity, speech perception in noise, visual cues

1. INTRODUCTION

Spoken word recognition is affected by the familiarity of the talker [12]. One approach to explaining familiarity effects has been in terms of exemplar-based model of speech perception [6]. Although there are a variety of such models, e.g. [4, 8, 14] the basic idea is that speech is represented in phonetically detailed exemplars that the speaker/perceiver has experienced.

These stored perceptual details are integral to later speech perception, with talker appropriate exemplars activated and inappropriate exemplars deactivated. These activated exemplars form the representations over which inputs are matched for the purposes of identification, classification and so on. This is basically how the effect of talker familiarity is explained, experience with a talker enables a representation of his/her talking style, habits of articulation, etc, to be generated and act as a frame of reference from which to perceive speech.

Johnson [6] has proposed that all kinds of talker cues can be involved in refining the set of activated exemplars from which the perceptual/linguistic response emerge. In addition to the familiarity of a specific voice, Johnson has suggested that non-linguistic cues can play a role in how the activated set of exemplars is tuned, e.g., prior expectations, visual cues, and other factors that affect the perceived identity of the talker, see also [5].

The idea that visual cues can influence auditory speech perception is consistent with findings that speech intelligibility can be affected by perceptions of the language competence of the talker based on visual cues. For example, the performance on a speech intelligibility test dropped when USA college-age listeners associated a voice with Asian-looking face [13].

In the current study we examined the effect that auditory and visual talker familiarity had on a speech perception in noise (SPIN) task. We manipulated both auditory and visual familiarity of talkers by pairing voices with animated characters. Animated characters were chosen to enable greater control in the swapping of faces and voices and because it is clear that perceivers can form a strong association between a voice and a character (e.g., Homer Simpson). Further, the faces of such characters have been demonstrated to affect the perception of talker voice characteristics [9]. More importantly, the aim of this manipulation was not to tap the intrinsic link between real faces and voices [10] but to employ specific (but non-linguistic) visual stimuli that could be readily associated with particular voices. In this regard, the manipulation was similar to that used by [5] who used stuffed toys (e.g., kangaroos or kiwis) to evoke the concept of regions associated with different vowels (Australia and New Zealand).

To present familiar and unfamiliar faces and voices we adapted a recently developed version of a talker familiarization procedure that has demonstrated familiarity effects from short-term exposure [2]. In this adapted procedure we had participants learn to pair a character's face with a particular voice and later we tested to see whether this pairing would influence speech intelligibility.

Talker familiarity effects appear to show specificity in how learning from one episode is transferred to another. So, for instance, exposure to a person speaking a word facilitates recognition of a word in noise but not a sentence, whereas exposure to a person speaking a sentence does not facilitate isolated word recognition but does facilitates recognition of a sentence in noise [11].

Given this finding of transfer specificity, and the idea that visual aspects of a talker become part of a set of activated exemplars that influence speech recognition, it should be expected that the largest facilitation from talker familiarity would occur when both the talker's voice and face (or representation of this) were presented at exposure and at test.

2. EXPERIMENT

2.1. Method

2.1.1. Participants

Thirty-four participants took part in the experiment for course credit at the University of Western Sydney. All were native speakers of Australian-English and reported normal hearing and normal or corrected-to-normal vision.

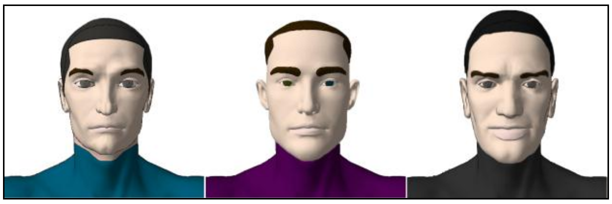
2.1.2. Stimuli

Fifty sentences were selected from the IEEE sentence list (IEEE, 1969) and recorded as auditory speech stimuli. Six male native speakers of Australian English (in their early twenties) were recruited as talkers. The recordings were made using a lapel microphone (44.1 kHz, 16-bit stereo). Each talker was seated in an IAC booth and instructed to say aloud all the 50 sentences in a neutral emotion.

The visual stimuli to be combined with the auditory stimuli consisted of 8 animated 3D talking heads. These heads were created by using the Daz3D® (*http://www.daz3d.com/i/3d-models*) models (based on 3D scans of human models, see Figure 1). The rigid and non-rigid motion of each head was animated from auditory speech input using DAZ3D® Mimic pro. This program generates the lip, jaw and face/head motion of a character from speech sounds and transliterated text. Given, as mentioned above, that the aim was to use non-linguistic voice/visual associations, we rendered the characters so that their mouth was obscured by a depiction of a microphone. This was also done to curtail ceiling effects in the SPIN task as visual speech information (even generated from an automatic process) might boost speech intelligibility to near ceiling levels.

Altogether 138 video clips were constructed. In the exposure task participants were presented 8 sentences spoken by four talkers (n = 32) and three versions of the exposure task were constructed (n = 32 x 3) so that a sentence by each talker could be presented in each of the three test conditions (familiar voice and face, familiar voice and unfamiliar face and unfamiliar voice and face). For the SPIN task, 6 talkers each said 7 new sentences without any sentence being repeated (N = 42).

Figure 1: Examples of three of the animated talking heads used in the study.



The SPIN stimuli were constructed by first equating the digitized auditory sentences for peak RMS amplitude using Praat [1] at 69 dB and then combining these with different samples of babble speech (consisting of three female talkers and one male, Auditec, St. Louis, MO) at -5dB. The noise and speech stimuli had the same duration.

Participants were allocated to one of the versions at random but care was exercised to keep the total number of participants across versions as similar as possible.

2.1.3. Procedure

Participants were tested individually in a sound attenuated IAC booth. Auditory stimuli were presented through Sennheiser HD580 headphones. The video clips were played back using the DMDX software [3] on a ViewSonic G810 21 inch monitor.

The experiment consisted of an exposure (training) phase and a SPIN task. The materials for the exposure phase were made up of the 8 sentences spoken by four different talkers. These sentences were initially introduced in association with four animated characters and were then repeatedly presented by voice only, by face only and by face and voice for identification in the training session.

Participants were informed that in the exposure session they would have to learn the names, voices and faces of four characters. Initially, a talking video of each of the four characters was presented (one at a time) along with a name. Following this, participants were presented with only the voices of the same characters and the associated name. Next 32 auditory-only sentences (8 sentences for each character) were presented one at a time and after each the participant was presented with the four names and asked to move the mouse and click on the correct one. This was repeated until the participant's performed better than 80% correct. Following this, participants were presented the name of a character and then the four faces and they were asked to click on the matching face (this time it was necessary to reach 90% correct). Participants then were shown the talking videos and had to match the correct name. Next, participants were presented the voices and had to match these to the correct faces (again it was necessary to reach 90% correct). Finally, participants received a mixture of the previous trials (voice/name, face/name, video/name and voice/face) and had to reach the 90% correct criterion. The exposure session lasted about 1/2 hour.

After the exposure session had been completed, participants were informed that they would see and hear animated characters speaking in noise and that their task was to type out as many of the words that they had heard at the end of each trial.

The SPIN task sentences were comprised of six sets that each consisted of 7 different sentences spoken by the six animated characters. There were three different familiarity conditions: familiar voice with matched face (Fam V/F), familiar voice with unfamiliar face (Fam V & Unfam F), and unfamiliar voice and face (UnFam V/F). Two animated characters were used in each condition.

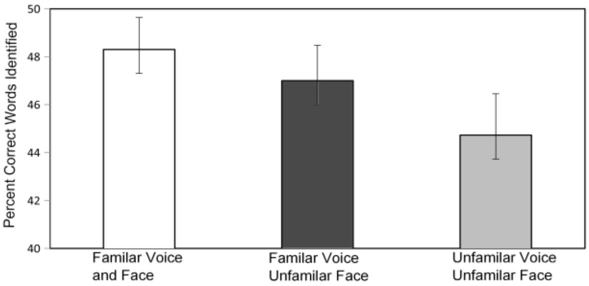
The presentation of sentences/talkers was randomized. After each stimulus presentation in the SPIN task, participants typed their responses. In scoring these data, all words were scored and credit was only given if the typed word exactly matched the spoken word (except where the response was an obvious typo). The percentage correct word identification was calculated as the measure of speech recognition for each condition.

2.2. Results & discussion

Participants all achieved levels of more than 90% accuracy in the training task (showing that they were familiar with each of the talkers).

Figure 2 presents the summary of the percent correct scores (and SE) for the SPIN task. A repeated measure ANOVA was conducted to compare the scores of the three presentation conditions (experimental version was treated as a non-repeated factor). Overall, the difference between the three conditions was not significant, although it was borderline, F(2,62) = 2.82, p = 0.07.

Figure 2: Mean percent words correctly indentified (SE) for familiar voice/face, familiar voice & unfamiliar face, and unfamiliar voice/face conditions.



A Planned comparison showed a significant difference between the Fam V/F and Unfam V/F conditions (4%), F(1,31) = 5.45, $p < 0.05$. The size of this effect is similar to that reported for familiarity effects in [2]. There was no difference between the Fam V/F and the Fam V & Unfam F conditions, F < 1, nor between Fam V & Unfam F and the Unfam V/F conditions, F(1,31) = 2.02, $p > 0.05$.

There was a significant talker familiarity effect shown in SPIN performance when the talker's face as well as voice was familiar. This effect was not shown when the talker's face was not familiar even though participants had been familiarized with the voice. There was, however, no difference between the intelligibility of the familiar voice and familiar face pairing and the familiar voice, unfamiliar face one. Unfortunately, the intermediate status of the latter condition with respect to the two others makes the interpretation of the results less clear.

3. GENERAL DISCUSSION

A reliable talker familiarity effect was demonstrated for the familiar voice and face condition compared to the unfamiliar voice and face one. The effect was small (4%) but it shows that the short-term paradigm offers a tractable research tool for the ready investigation of familiarity effects without intensive training. It should be noted though, that the size of the effect might be enhanced with additional training.

Having shown a significant familiarity effect when both the voice and face were familiar, the chief interest was on what happened to the familiar voice/unfamiliar face condition. For this condition, it was suggested that perceivers might take into account a range of talker properties when processing speech. That is, just as the image of a female or male talker can shift vowel identification boundaries [7] so too might the perceived identity of a talker influence how a speech exemplar is encoded and latter retrieved. If this was the case, then it might be expected that a reduced voice familiarity effect would be found in this condition.

What was found for the familiar voice/ unfamiliar face condition was that the mean percent speech identification score did not differ from that of the unexposed control. On the face of it, this appears to support the idea that the voice familiarity effect was reduced due to it being paired with an unfamiliar talking face. A result that is consistent with the idea that where both the face and voice are familiar, the properties of recent stored exemplars can assist in the recovery of that voice in noise; but where the voice is familiar but the face not, access to stored exemplars is less effective. Of course, the intelligibility of the familiar voice in the unfamiliar face condition also did not differ from the familiar voice and face condition. In this regard, it would seem that a more powerful design (possibly involving more extensive training) may be required in order for a clearer outcome to be produced.

4. ACKNOWLEDGEMENTS

The authors wish to thank Tim Paris for his assistance in conducting the experiments and acknowledge support from the Australian Research Council (DP0666857 & TS0669874).

5. REFERENCES

[1] Boersma, P., Weenink, D. 2010. Praat: Doing phonetics by computer [Computer program], Version 5.1.32. *http://www.praat.org/*. Retrieved by 2010/30/04.

[2] Davis, C., Kim, J. 2010. Transfer of talker-familiarity effects. *Proceedings of 20th International Congress on Acoustics* Sydney, Australia.

[3] Forster, K.I., Forster, J.C. 2003. DMDX: A windows display program with millisecond accuracy. *Behavioral Research Methods: Instruments & Computers* 35, 116-124.

[4] Goldinger, S.D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105, 251-279.

[5] Hay, J., Drager, K. 2010. Stuffed toys and speech perception. *Linguistics* 48, 865-892.

[6] Johnson, K. 2005. Speaker normalization in speech perception. In Pisoni, D., Remez, R.E. (eds.), *The Handbook of Speech Perception*. Oxford: Blackwell, 363-389.

[7] Johnson, K., Strand, E.A., D'Imperio, M. 1999. Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27, 359-384.

[8] Jusczyk, P.W. 1993. From general to language-specific capacities. The WRAPSA model of how speech-perception develops. *Journal of Phonetics* 21, 3-28.

[9] Kamachi, M., Hill, H., Lander, K., Vatikiotis-Bateson, E. 2003. Putting the face to the voice: matching identity across modality. *Curr Biol* 13, 1709-1714.

[10] Kim, J., Davis, C. 2010. Knowing what to look for: Voice affects face race judgments. *Visual Cognition* 18, 1017-1033.

[11] Nygaard, L.C., Pisoni, D.B. 1998. Talker-specific learning in speech perception. *Perception & Psychophysics* 60, 355-376.

[12] Nygaard, L.C., Sommers, M.S., Pisoni, D.B. 1994. Speech perception as a talker-contingent process. *Psychol. Sci.* 5, 42-46.

[13] Rubin, D.L. 1992. Non-language factors affecting undergraduates' judgments of non-native English-speaking teaching assistants. *Research in Higher Education* 33, 4.

[14] Skousen, R. 1989. *Analogical Modeling of Language.* Dodrecht, the Netherlands: Kluwer Academic Publishers.