Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Transfer of Talker-Familiarity Effects

Chris Davis and Jeesun Kim

MARCS Auditory Laboratories, University of Western Sydney, Milperra, Australia

PACS: 43.71.Gv, 43.71.Es, 43.71.Sy

ABSTRACT

Familiarity with a talker facilitates perception for both heard speech (where speech from a familiar talker is better identified in noise) and for visual speech (where familiarity with a talker's face assists visual speech recognition). Recently, it has even been shown that the talker familiarity effect can be produced cross modally, i.e., experience in speech-reading a talker facilitates performance on a SPeech-In-Noise (SPIN) task. The current study examined within and across modal speaker familiarity effects with short-term familiarity training and test of transfer to SPIN performance from auditory only (AO), visual only (VO) and Auditory-visual (AV) exposure. The results showed that there was transfer from AO and VO talker familiarization, but not from AV speech. The results are discussed in terms of how the familiarity effect might be sensitive to the degree of bottom-up attention initially paid to a talker's speech.

I. INTRODUCTION

Speech perception is plagued with uncertainties. Unlike print, the speech signal is evanescent, is often commingled with other signals and there is no speech equivalent of a standard font. Given this indefinite state, factors that act to reduce the speech signal's uncertainty generally facilitate its perception. For example, spoken word recognition in noise is more accurate with a single talker compared to when the talker is unpredictable from trial to trial, (i.e., mixed talkers, Creelman, 1957). Even without noise, word recognition times are faster in single-talker lists (Summerfield & Haggard, 1973). Furthermore, talker-list effects have also been found with visual speech. For instance, speech reading sentences is easier from a single speaker list than a multiple speaker list (Yakel, Rosenblum, & Fortier, 2000).

Talker familiarity, a related factor that reduces signal uncertainty, also facilitates speech processing, especially when processing occurs under difficult circumstances (in noise) or with difficult stimuli, e.g., low-frequency words with many high-frequency neighbours (Nygaard & Pisoni, 1998; Bradlow & Pisoni, 1999). For example, Nygaard, Sommers, and Pisoni (1994) had listeners learn to identify the voices of ten talkers (five male and five female) from single word utterances. Listeners were familiarized with each talker's voice for each of nine days of training and learned to associate a name with each. Following training, one group of listeners were given SPIN task with the trained talkers and another group a SPIN task with talkers they had not been trained on. It was found that more words were correctly identified in noise when the voices were familiar (although, there were large individual differences in the application of this effect).

Yonan and Sommers (2000) used familiar and unfamiliar talkers in a SPIN task (that used both single words and sentences) to investigate an explanation of the familiarity advantage put forward by Nusbaum and Morin (1992). This explanation proposed that when a voice is familiar a listener can use stored information to compute perceptual normalization; a benefit accrues because this is more efficient than a computation conducted from scratch (Logan, 1998). To test this proposal, both young and elderly participants were examined to see if the latter would benefit from talker familiarity. It was also tested whether explicit training was required for a familiarity benefit to show.

Yonan and Sommers found that older listeners, who had impaired ability to explicitly identify talkers' voices, showed a familiarity benefit that was similar in size as to that shown by the younger listeners. Furthermore, it was found that incidental exposure to the voices (in a semantic judgment task) produced the same familiarity benefit as explicitly directing participants to attend to the voices. Given these two results, it was argued that the talker familiarity effect is mediated by implicit memory.

As with the talker-list effects mentioned above, it has also been demonstrated that the facilitatory effect of talker familiarity extends to visual speech. That is, speech reading gets better as participants became increasingly familiar with the same speaker (Lander & Davies, 2008). What makes the link between auditory and visual speech even clearer is the recent finding that familiarity with a talker in one modality (i.e., having seen the talker speaking) can enhance the perception of speech of that talker in the other modality, i.e., hearing the talker speak in noise (Rosenblum, Miller & Sanchez, 2007). Rosenblum and colleagues argued that this cross-modal familiarity effect was due to amodal, talker-specific articulatory-style information that acts to facilitate the perception of speech in both modalities.

The current study followed up the Rosenblum et al one, so it is useful to consider the method used in this study in more detail. In Rosenblum et al study, participants first took part in a speech reading task. The spoken stimuli consisted of silent videos of the lower half (bottom of the chin to upper cheeks) of two female talkers. Speech reading performance was scored by a key word method in which three key open class words were scored in each of 100 BKB sentences (Bench & Bamford, 1979). Each sentence was presented twice and the task was to say what the words were (participants were not told about the SPIN task that followed). For the SPIN task, 150 new sentences were presented in white noise (each presented twice at +5, 0 and -5 SNR, as used in Yonan & Sommers, 2000). The same key word scoring (maximum 3 words/ sentence) method was used to evaluate SPIN performance. In the task, 30 participants were presented with same talker who they had seen in the speech reading task and 30 with a different talker. The results showed that participants who had the same talker in the speech reading and SPIN tasks performed approximately 5% better in the latter than did participants who had different talkers in the two tasks.

The current study presents a modified version of the talker familiarity paradigm in which participants are given relatively limited exposed to a talker and then immediately given a SPIN test. It is interesting to determine whether short-term exposure rather than extensive multi-day training will produce effects. Furthermore, the current study tested the effect on SPIN performance of exposure to auditory, visual and auditory-visual speech information. That is, participants were familiarized with four talkers in three conditions: auditory only, visual, or auditory-visual. After exposure to each talker, participants were required to identify the talker's speech in noise. The result will determine if there is a general talker familiarity effect, i.e., whether the overall performance in speech recognition in noise is better than the control condition (that consisted of reading written words) and whether the cross-modal talker familiarity effect is different from that produced by auditory only and auditory-visual exposure.

Other modifications that were introduced in the current study include: 1. The use of full face (& neck) videos since it has been shown that visual speech information from the whole head can affect speech perception and specifically speech perception in noise (Cvejic, Kim & Davis, 2010; Davis & Kim, 2006). 2. The use of multitalker babble speech as noise source (rather than white noise). Babble speech tends to be a more commonly experienced noise source and it has been shown that both auditory and visual speech perception can be affected by the type of noise used (Davis, Kim, Grauwinkel & Mixdorff, 2006). 3. The use of the same exposure (training) sentence for all the talkers as this would make talker differences readily apparent. 4. The use of a scoring procedure that used all the presented words rather than a set of three key words.

II. METHOD

A. Participants

Thirty-two undergraduate university students from the University of Western Sydney participated in the experiment. All participants were native speakers of English, 18 years of age or over and had self-reported normal or corrected-to-normal vision and none reported a history of hearing loss.

B. Materials

The materials consisted of two different sets: An exposure (training) set and a SPIN task set. The exposure set consisted of the same 10 sentences spoken by four different native talkers of Australian-English processed to be presented in an Auditory-Only version; a Visual-Only version; an Auditory-Visual version and a text version (control). The SPIN task sentences comprised of four sets of 10 different sentences per set spoken by the four talkers.

In all then, 50 sentences were selected from the IEEE sentence list (IEEE, 1969) and recorded as auditory and visual speech stimuli. Four male native speakers of Australian English (in their early twenties) were recruited as talkers. Audiovisual recordings were made using a digital video camera (25 fps) and an externally connected lapel microphone (44.1 kHz, 16-bit stereo). Each talker was seated in a well-lit IAC booth and instructed to say aloud all the 50 sentences in neutral emotion. The talkers were video recorded against a uniform grey background, facing the camera and the recording showed the head and shoulders. Overall, the experimental items consisted of speech stimuli in auditory-only (AO), visual-only (VO) and auditory-visual (AV) conditions.

For the SPIN task, the digitized auditory sentences were equated for peak root mean square amplitude (using Praat, Boersma & Weenink, 2010) at 69 dB and then combined with different samples of babble speech (consisting of three female talkers and one male, obtained from Auditec, St. Louis, MO) at 70dB. Thus the average SNR was -5dB. The onset of noise and speech stimuli had the same duration.

Sixteen versions of experiment were constructed so that the sentences from each of the talkers could be presented in each of the SPIN tasks that followed the exposure session without any sentences being repeated within a version. Figure 1 presents a cartoon of the design and shows four of the 16 versions. Each participant was run on one of the versions (the order of the trials in a version is shown horizontally). So for example, for version 1, a participant would be presented with 'Talker 1" auditory-only, then the SPIN test; "Talker 2" visual-only, then the SPIN test; "Talker 3" auditory-visual, then the SPIN test and finally a text control and then the SPIN test on "Talker 4". Each version of the experiment consists of 4 conditions and in each condition there were 10 exposure (training) sentences and an associated SPIN task. The four conditions differed in that speech presented in the training session was Auditory-Only, Visual-Only, Auditory-Visual or text (Control). In the associated SPIN task session participants had to identify speech in noise (10 sentences) that were produced by the same talker as in the exposure session. Each participant was allocated to one of the 16 versions (across which the presentation of the talkers was rotated).

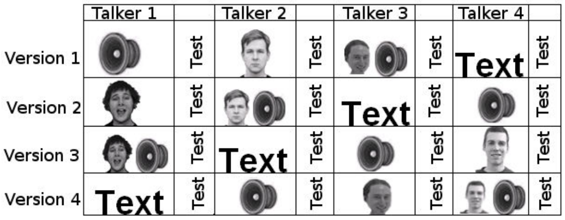

Figure 1. An example of the design of four out of the 16 version versions of the experiment. In the other 12 conditions, the talkers depicted above were rotated across each presentation order.

C. Procedure

Participants were first informed about what they would be required to do in the experiment. Each participant was told about the exposure sessions in which he/she was asked to type out precisely what the talker had said (in the Auditory only, Visual-Only and Auditory-Visual conditions) or type out what words they had read in each of the briefly presented sentences (each sentence was presented for 1.5 seconds).

Each of the sentences in the exposure session was presented twice and then the participant had to write down what he/she had heard/seen. Participants were then told that after each exposure session she/he would hear a person speaking in noise and that this time each sentence would be only presented once. They were told that the in-noise sentences were different from those presented in the exposure session. Participants were told that the task was to type out as many of the words that she/he had heard. Before the experiment proper began, participants were given 10 written practice trials (presented for 1.5 seconds each) to give them an idea of the type of sentences that she/he would see/hear in the experiment.

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 (Forster & Forster, 2003) on a ViewSonic G810 21 inch monitor.

Sentences within each condition were presented in a block and the presentation of sentences within each block was randomized. The order of the presentation conditions depended upon the particular version a participant did. After each stimulus presentation in the SPIN task, participants typed their responses. In scoring these data, all words were scored with credit 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.

III. RESULTS AND DISCUSSION

Mean percent correct scores are shown in Figure 2. As can be seen, SPIN performance was approximately 4% better when this test was preceded by same talker exposure in the Auditory-Only and Visual only conditions compared to the no speech text presentation control condition. The difference between the Auditory-Visual exposure condition and the text control was smaller than the other two exposure conditions.

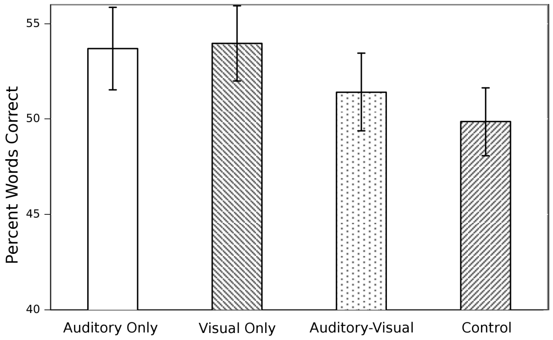

Figure 2. Mean percent correct word recognition scores (and Standard Error) in the SPIN task for each of the four different exposure conditions

The overall familiarity effect was tested using repeated measures ANOVA on the percent correct items scores. As it turned out, the difference between the four exposure conditions was marginal, F(3,108) = 2.58, p = 0.057, η2 = 0.07. The basis of this weak effect can be determined by pair-wise tests of each of the experimental conditions against the control.

SPIN performance was significantly better than control in the Auditory-Only condition, F(1,36) = 5.11, p < 0.05, η2 = 0.12. Likewise, performance was also better in the Visual-Only condition, F(1,36) = 5.56, p < 0.05, η2 = 0.13. Performance in the AV condition did not differ from that in the control condition, F < 1; it seems likely that scores in this condition contributed to the marginal overall difference between the presentation conditions.

In general, the results confirm those that have previously been reported. That is, there was an advantage in the SPIN task for talkers whose voice was familiar due to recent pre-exposure in the Auditory-Only condition or pre-exposed in the Visual-Only condition. The size of this familiarity advantage was similar for the two different types of exposure, lending support to the idea that the familiarity effect is based upon amodal information about speech production (Rosenblum et al, 2007).

This amodal view fits nicely with the idea that the familiarity effect is based on representations of the *person* who is talking (see Johnson, 2005). A straightforward proposal that fleshes out this talker-as-central view is that talker normalization is based upon perceptions of properties of the talker's vocal tract. In this regard, the proposal meshes with that of Nusbaum and Morin (1994), mentioned above, that the advantage of a familiar voice accrues because a listener can use stored information to compute perceptual normalization. What is important, as Johnson (2005) points out, are the perceiver's expectations about the talker that rather than the veridical vocal tract parameters themselves.

There is of course an obvious problem with the above proposal. If it were the perception of the talker that was paramount in producing the familiarity advantage, then the Auditory-Visual presentation condition should have been the one to produce the most robust effect, since it is in auditory-visual presentation that auditory and visual speech are tied together by a dynamic talking person. As it turned out, however, the Auditory-Visual exposure condition produced the least robust effect; one that was not significantly different from the no speech exposure text control.

Our proposal for why this was the case (in this particular experiment) is that the familiarity effect was modulated by the degree to which perceivers paid attention to details of the talker's way of speaking in the different exposure conditions. Consider the two conditions that showed robust effects, the Auditory and Visual only ones. The task in the exposure trials was to write down what was heard or seen; a task that would have drawn attention to details of how the speech was articulated and been quite challenging. However, because the Auditory-Visual condition provides a much more robust speech signal, perceivers may not have needed to invest as much attention in the task. We suggest that it was this variation in stimulus-driven bottom-up attention that might have modulated the robustness of the familiarity effect. It should be noted that the type of attentional effect we are proposing is not something that would be necessarily altered by explicit instruction and so we do not see that this proposal conflicts with the demonstration by Yonan and Sommers (2000) that the familiarity benefit was the same following intentional or incidental voice learning.

ACKNOWLEDGEMENTS

The authors wish to thank Erin Cvejic and Michael Fitzpatrick for assistance with the auditory and video recording and editing and for running participants. We also thank the four talkers for their time and patience. The authors acknowledge support from the Australian Research Council (DP0666857 & TS0669874).

REFERENCES

J. Bench and J. Bamford, "*Speech-hearing tests and the spoken language of hearing impaired children*" (London: Academic Press, 1979).

P. Boersma and D. Weenink. "Praat: doing phonetics by computer" [Computer program]. Version 5.1.32, retrieved 30 April 2010 from http://www.praat.org/

A.R. Bradlow and D.B. Pisoni, "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors" *J. Acoust. Soc. Am*. **106**, 2074–85 (1999).

C.D. Creelman, "The case of the unknown talker" *J. Acoust. Soc. Am*. **29**, 655 (1957).

E. Cvejic, J. Kim and C. Davis, "Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion" *Speech Commun*. **52**, 555–564 (2010).

C. Davis, J. Kim, K. Grauwinkel and H. Mixdorff, "Lombard speech: Auditory (A), visual (V) and AV effects" in *Proceedings of Speech Prosody 2006* eds. R. Hoffmann & H. Mixdorff (TUD Press, 2006) pp 361–364.

IEEE Subcommittee on Subjective Measurements, "IEEE recommended practices for speech quality measurements" *IEEE Trans. on Audio & Electroacoustics*, **17**, 227–246 (1969).

K.I. Forster and J.C. Forster, "DMDX: A Windows display program with millisecond accuracy" *Behav. Res. Methods* **35**, 116–124 (2003).

K. Johnson, "Speaker normalization in speech perception"in *The Handbook of Speech Perception. Oxford* eds. D. Pisoni & R.E.Remez (Oxford: Blackwell, 2005) pp 363–389.

K. Lander and R. Davies, "Does face familiarity influence speechreadability?" *Q. J. Exp. Psychol.* **61**, 961–967 (2008).

G.W. Logan, "Toward an instance theory of automatization" Psychol. Rev. **95**, 492–527 (1988).

H.C. Nusbaum and T.M. Morin, "Paying attention to differences among talkers" in *Speech Perception, Speech Production, and Linguistic Structure* eds. Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Tokyo: OHM, 1992) pp. 113–34.

L.C. Nygaard and D.B. Pisoni, "Talker-specific learning in speech perception" *Percept. Psychophys.* **60**, 355–76. (1998).

L.C. Nygaard, M.S. Sommers and D.B. Pisoni, "Speech perception as a talker-contingent process" *Psychol. Sci.* **5**, 42–6 (1994).

L.D. Rosenblum and R.M. Miller and K. Sanchez, "Lip-Read Me Now, Hear Me Better Later: Cross-Modal Transfer of Talker-Familiarity Effects" *Psychol. Sci.* **18**, 392–396 (2007).

Q. Summerfield and M.P. Haggard, "Vocal tract normalization as demonstrated by reaction times" *Report of Speech Research in Progress*, **2**, 1–12, The Queen's University of Belfast, Ireland (1973).

D.A. Yakel and L.D.Rosenblum and Fortier. M.A. "Effects of talker variability on speechreading" *Percept. Psychophys.* **62**, 1405–12 (2000).

C.A. Yonan and M.S. Sommers, "The effects of talker familiarity on spoken word identification in younger and older listeners" *Psychol. Aging* **15**, 88–99 (2000).