



Investigating the Role of Familiar Face and Voice Cues in Speech Processing in Noise

Jeesun Kim¹, Sonya Karisma¹, Vincent Aubanel^{1,2}, Chris Davis¹

¹The MARCS Institute, Western Sydney University, Australia

²Gipsa-Lab, Grenoble-Alpes University, France

j.kim@westernsydney.edu.au, 17401135@student.westernsydney.edu.au,
vincent.aubanel@gipsa-lab.fr, chris.davis@westernsydney.edu.au

Abstract

The speech of a familiar talker is better recognized in noise than an unfamiliar one, suggesting that listeners access talker-specific models to assist with degraded input. This study investigated whether a talker model could be accessed by presenting the face of a talker. In the experiment, participants were trained in recognizing three talkers' faces and voices to ceiling-level. Participants were then given a speech in noise recognition task consisting of four talker conditions: familiar face then familiar voice; unfamiliar face then familiar voice, familiar face then unfamiliar voice; and unfamiliar face then unfamiliar voice. A talker familiarity effect was found, i.e., speech perception was more accurate in the familiar face and familiar voice condition than all other ones. A familiar voice did not produce a talker familiarity effect when paired with an unfamiliar face. The familiar face and unfamiliar voice condition had the poorest performance, indicating that pairing a familiar face and unfamiliar voice had a disruptive effect. The results suggest that listeners develop a talker model that includes details of both the voice and the face; and that accessing this model can in some circumstances be wholly determined by face cues.

Index Terms: talker familiarity effects, speech perception in noise, paralinguistic cues

1. Introduction

Speech recognition in noise is better for a familiar talker compared with an unfamiliar one. For example, Nygaard, Sommers and Pisoni [1] 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 was given a 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.

The talker familiarity effect shows that the perceiver has learned something about a talker's voice (a talker model) that facilitates subsequent speech recognition. Several (not mutually exclusive) proposals have been put forward concerning what sort of information is learned about a talker. On the one hand, it has been proposed that what is learned is information about an individual's speech production, e.g., the perceiver learns information that specifies how a specific talker's vocal tract anatomy differs from that of other talkers, see Joos, [2]). Consistent with this idea is the finding that the

talker familiarity effect occurs with visual speech [3]; [4] and that having seen a talker silently speaking can subsequently enhance the intelligibility of that talker's voice in noise [5].

On the other hand, it has been proposed that the perceiver extracts phonetically detailed information from the instances of speech that she/he has been exposed to and learns to use this exemplar-based knowledge to facilitate the perceptual operations that analyze and encode the talker's voice [1]. Consistent with this proposal is the finding that the talker familiarity effect is sensitive to the specific perceptual task required of the perceiver. That is, exposure to a person uttering a word has been shown to facilitate recognition of a word in noise but not a sentence, whereas exposure to a person speaking a sentence does not facilitate the recognition of an isolated word but does facilitate sentence recognition in noise [6].

The talker familiarity effect indicates that specific knowledge (a talker model) can interact with speech processing. Here our research question concerns how such knowledge can be accessed. One straightforward proposal is that auditory speech cues (e.g., vocal characteristics) can lead to activation of the relevant talker model (Johnson [7]). Somewhat more controversially, Johnson has been suggested that non-linguistic cues can play a role in activating a talker model or that these can modify how the activated knowledge is considered. For example, Johnson suggested that prior expectations, visual cues, and other factors that affect the perceived identity of the talker have a role to play in whether a talker model will be activated or not (see also [8]). Given this, the current study followed up Johnson's suggestion by testing effects on a SPIN task of a familiar talker's face on speech processing. As a talker's face is a clear-cut cue to their identity, the presentation of a face should be an effective way of accessing a talker model.

On each trial in the SPIN experiment we first presented a target face then a spoken sentence. By doing so, we were able to test the recognition of a familiar and unfamiliar talkers' speech in the presence of a familiar talker's face or an unfamiliar one as baseline. The accuracy of speech recognition was measured to provide an index of the familiarity effect. Several predictions were made. First, a talker familiarity effect was expected when processing a familiar talker's speech, when a picture of the familiar talker's face was presented. Based on our previous study using animated character faces, we expected that a familiar voice would not produce a familiarity effect when an unfamiliar face was presented [9]. Finally, when the unfamiliar talker's speech was presented with a familiar face, it was expected the familiar talker's face would activate the familiar talker's speech information, i.e.,

prompt a frame of reference tuned specifically to a talker's speech different to that of the presented unfamiliar speech. If so, this may have a detrimental effect on processing speech by an unfamiliar talker.

2. Method

2.1. Participants

Fifty-five undergraduate students ($M_{\text{age}} = 23.05$, $SD = 7.34$, including 12 males) from the Western Sydney University participated in the experiment for course credit. All the participants were fluent Australian English speakers (50 native speakers and five who had learnt English from an early age). All the participants reported normal hearing and normal or corrected-to-normal vision.

2.2. Stimuli

The study used video recordings of 5 male native Australian English talkers in their twenties, each uttering two hundred sentences (5×200). Each sentence contained five key words and was selected from IEEE Harvard sentence list (1969) [10]. This list consists of phonetically balanced sentences that are semantically unpredictable (e.g., "The latch on the back gate needs a nail").

The recordings captured each talker's face and shoulder. The videos were recorded using a Sony HXR-NX30P video camera recorder with audio recorded separately using an AT4033a audio-technica microphone. In the recording, each sentence to utter was presented one at a time on a monitor which was set up in front of the talker. The talker was instructed to look at the sentence and then to say aloud the sentence clearly and in a neutral emotion while looking directly at the camera.

2.3. Research Design

The experiment consisted of two parts: talker familiarization and the SPIN task. The talker familiarization and the SPIN task were run using Psych Toolbox [11].

2.3.1. Familiarization Session

In the familiarization session, three talkers were presented via 140 trials in total. This included 135 main trials (using 45 sentences) and 5 practice trials (using 3 sentences). The sentences were presented in pseudo-random order to reduce the influence of any systematic practice or order effects.

In each familiarization trial, three talkers' faces were presented side by side on a monitor, along with a spoken sentence uttered by one of the talkers. Two of the talkers were static (i.e., static pictures) while the other talker was uttering a spoken sentence. The participants' task was to match the voice to one of the three faces by positioning the cursor over the image and clicking the mouse key. Note that the talker with the moving face in a trial was not necessarily the talker whose voice was presented. That is, in 1/3rd of the trials, the speech was uttered by the talker with the moving face while in the other 2/3rds, the speech heard was by one of talkers with the static faces (1/3 each). This encourage participants to use the face to match the spoken utterance.

That is, the purpose of including a dynamic face was twofold: First, to provide more realistic and ecologically valid talker-related information (in addition to what the face looks like and how their speech is like). Second, to make the

familiarization task more challenging and thus more interesting and/or engaging. That is, it is likely that a moving face would draw attention, as such it would be easily associated with the concurrently presented speech. When this was not always the case, the task was more difficult and required participants to ignore the movement and concentrate instead on who uttered the heard speech.

Over the course of the trials, all the three talkers were equally presented as dynamic faces so that participants got familiarized to the same extent with how each talker moved when speaking. The positioning of the static pictures and the dynamic face was counterbalanced. The talkers' positioning and the position of the moving face was pseudo-randomized to reduce any possible order effects.

In each trial, a click on a talker's face (as a response indicating who the talker for the spoken sentence was) was required in order to be able to move on to the next trial. Upon clicking, feedback was provided by the presentation of a green frame surrounding the talker's face for a correct response; and a red frame surrounding the talker's face which was incorrectly selected, along with a green frame surrounding the correct face. A click on the correct face was required to start the next trial.

A criterion of 85% correct response in the second half of the familiarization session was set for inclusion of a participant's data in the results. Only the second half of the familiarization session was used to establish this criterion since this was more indicative of whether the participants had familiarized themselves with the talker, i.e., whether they had learned which voice belonged to which talker.

2.3.2. The SPIN task

In order to test the effect of a familiar face on speech recognition, we first presented a target face then a spoken sentence in each trial of the SPIN task. In the task, three familiar and two unfamiliar talkers' faces/voices were combined in four different ways: familiar face/voice (F_rF_v), familiar voice paired with unfamiliar face (U_rF_v), unfamiliar voice paired with familiar face (F_rU_v) and unfamiliar face and voice (U_rU_v).

It was important to ensure that participants paid attention to the face presented in each trial so that any association between the talker's face and voice could influence speech perception in noise. To make sure that the presentation of a static face (along with the spoken sentence) would draw attention, each trial included a visual task at the beginning which involved recognizing the target talker's face.

In this visual task, five static pictures were presented on the screen in a row. These pictures consisted of four talkers with one of the faces appearing twice. The visual task presented the same four faces throughout the experiment. To the participants, two were familiar faces and two unfamiliar. The participants' task was to indicate which face was duplicated by clicking on either of the duplicated faces. When a correct response was made, a green frame surrounded the duplicated faces and other three faces disappeared. Following this, one of the duplicated static pictures was presented in isolation for 500ms and then the auditory speech in noise was presented for the SPIN task, for which participants were asked to type what they had heard.

For the SPIN task stimuli, each talker's speech was mixed with noise created by computing the long-term spectrum of all

recorded tokens of that talker (tailored speech shaped noise (SSN)) so that for each of the talkers the speech and noise had similar spectral properties. Tailored noise was used to control for possible intelligibility differences across talkers that might occur if the same noise was used for all the talkers. The speech sound and noise were combined at -3 dB SNR. The SNR level was determined through pilot testing to avoid ceiling and floor effects. The speech and noise stimuli had the same duration and onset.

Overall, there were 100 experimental trials and four practice trials. The presentation order of talkers and sentences was pseudo-randomized to minimize any order effects.

Since talkers can vary in terms of face/voice distinctiveness and speech intelligibility in noise, this variation could add noise to any talker familiarity effect produced. In order to control such talker effects, five versions of the experiment were created. Across these five versions, each of the five talkers used in the current study was allocated to one of the four different conditions.

In scoring participant's typed responses in the SPIN task, only key words were counted, and credit was given only 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.

2.4. Procedure

All participants were individually tested in a sound attenuated booth. The stimuli presentation and typed response data collection were controlled using Psych Tool Box [11]). The audio was presented through a Sennheiser HD-555 headset at a comfortable level, which was held constant for every participant.

The participants were instructed that there would be two sessions consisting of a familiarization session and a SPIN task. The participants were then assigned to one of the five versions of the experiment in a consecutive manner where participant one completed version one, participant two completed version two and so forth.

In the familiarization session, participants were informed that they would be presented with three different talkers' faces and voices over the entire session, and their task was to get to know which face and voice went together by listening to what the talker was like in terms of voice and how they spoke, etc. The participants were told that in each trial, they would hear one spoken sentence and see three faces and that they were required to click on the face that corresponded with the voice. The participants were informed that the faces consisted of two still pictures and one moving one and that the moving one did not always match the voice. They were also informed that they would get feedback on their response such that a green frame would surround the selected talker that matched the voice and a red frame indicated an incorrect response. When an incorrect response was made, participants were instructed to click on the correct face and then to move on to the next trial. The participants were given five practice trials followed by 135 trials. The participants were given two breaks, one in the middle of the exposure session trials and one at the end of the exposure session.

After the completion of the familiarization session, participants were informed that in the second session they would be presented with speech in noise. They were instructed

to type in any words that they heard and informed that at the beginning of each trial, there would be still pictures of five faces, two of which were duplicated; and they had to click on either of the duplicated faces in order to hear speech in noise. The participants were given four practice trials followed by 100 experimental trials.

After the completion of the SPIN task, participants were asked about their recognition of familiar voices and whether they noticed that the face matched/mismatched the subsequently presented voice in noise. This question was asked to find out whether any talker familiarity effect might have been modified by the participants' conscious awareness of the match/mismatch.

3. Results

All the participants achieved levels of more than 85% accuracy in the familiarization training task indicating that they were familiar with each of the talkers and able to match the face and voice. All the participants' data in the SPIN task were included for the examination of differences across the found conditions of talker familiarity: F_rF_v, U_rF_v, F_rU_v and U_rU_v.

The talker familiarity effect was measured based on the speech intelligibility scores that were calculated as the percentage of correct key words identified. The mean, standard error and 95% confidence intervals for the four conditions are reported in Table 1. A mixed repeated measures ANOVA was conducted with the four conditions as a within-participant factor and the 5 versions as a between-participant factor.

Table 1: SPIN Task Performance (mean percent key words correct) for the Four Talker Conditions.

Talker Conditions	Mean	SE	95% Confidence Interval	
			Lower bound	Upper bound
F _r F _v	48.4	1.72	45.0	51.9
U _r F _v	42.0	1.55	38.9	45.1
F _r U _v	38.7	1.47	35.8	41.7
U _r U _v	41.2	1.32	38.6	43.9

The analysis showed that there was a significant difference across the four conditions, $F(3,150) = 38.48$, $p < .001$, $\eta^2 = .44$. In order to determine which talker familiarity conditions were different from which, planned pairwise comparisons were conducted with the Sidak correction for multiple comparisons. The results were clear-cut, showing that the performance in the F_rF_v condition was significantly higher than the other three conditions (each $p < .001$); (2) performance in the U_rF_v condition was significantly higher than in the F_rU_v condition ($p < .01$) but not different from U_rU_v condition ($p = .98$); (3) performance in the F_rU_v condition was significantly lower than the U_rU_v condition ($p < .05$).

Nearly all of the participants (i.e., 51 out of the 55) reported that they had been aware that the SPIN task included familiar voices and that sometimes the voice did not always match the face.

4. Discussions

The research question posed in the current study was whether a talker's face could act as a cue to trigger a talker-specific model, and thus influence subsequent speech processing. This was tested by examining whether the presentation of a familiar talker's face (contrasted with an unfamiliar face as baseline), affected the recognition of subsequently presented familiar and unfamiliar talkers' speech in noise.

If the presentation of a face had no effect on speech recognition, it should have been the case that speech recognition performance would be better for a familiar compared to an unfamiliar voice regardless of which face was shown. This was not what occurred. The results showed a robust facilitatory effect of a familiar talker face for the familiar talker's speech and also a significant interference effect (worse speech recognition) when it was presented ahead of the unfamiliar talker's speech. These results suggest that a familiar talker's face was able to activate the relevant talker model that then acted as a guide for subsequent speech processing. In the case where the face matched the subsequently presented speech, speech processing was facilitated (F_{iF_v}); whereas when the face did not match the upcoming speech this caused interference (F_{iU_v}). It is possible that this interference effect was due to the listener paying attention to spectral/temporal regions in the speech and noise combination that worked against normal segmentation strategies.

Interestingly, there was no talker familiarity effect when a familiar voice was presented after an unfamiliar talker's face (U_{iF_v}). Previous studies that have demonstrated a voice familiarity effect did not present a face cues and so the listener would have been able to pay his/her full attention to the speech and noise mixture in order to activate a familiar talker model (if available) based on any auditory cues picked up. However, when an unfamiliar face is presented first, listeners may simply give up on trying to activate a familiar talker model, i.e., they ignore available cues. In this regard, the result is similar to that of Hay and Drager [8], who showed that the presentation of a stuffed toy (e.g., a kangaroo or a kiwi) likely to evoke the concept of Australia or New Zealand, affected their New Zealand listener's matching performance on New Zealand vowels – shifting it towards raised and fronted Australian-like tokens when the kangaroo was shown; here the visual cue also appeared to affect a listener's ability to drawn on the available acoustic cues.

In conclusion, the current results suggest that face cues can play a role in activation of the relevant talker specific models and interfere as well as facilitate speech processing.

5. Acknowledgements

The experiment was run as the second author's Honours thesis project. We thank all the participants and acknowledge the support from Australian Research Council (DP150104600 & DP 130104447).

6. References

- [1] Nygaard, L. C. Sommers M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5, 42-6.
- [2] Joos, M. A. (1948). Acoustic Phonetics, *Language*, 24, Suppl. 2, 1-136.

- [3] Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability?. *Quarterly Journal of Experimental Psychology*, 61(7), 961-967.
- [4] Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, 57(8), 1124-1133.
- [5] Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-read me now, hear me better later. *Psychological Science*, 18(5), 392-396.
- [6] Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & psychophysics*, 60(3), 355-376.
- [7] Johnson, K. (2005). Speaker normalization in speech perception. *The handbook of speech perception*, 363-389.
- [8] Hay, J. & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, 48, 865-892.
- [9] Kim, J., & Davis, C. (2011). Testing audio-visual familiarity effects on speech perception in noise. *Proceedings of the International Congress of Phonetic Sciences*, Vol. 4, 1062-1065.
- [10] Institute of Electrical and Electronic Engineers (1969). IEEE Recommended Practice for Speech Quality Measures (IEEE, New York).
- [11] Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, 36(14), ECVF Abstract Supplement.