The time-course of audiovisual interactions affecting perception of distorted speech

Ediz Sohoglu, Jonathan Peelle, Robert Carlyon and Matthew Davis

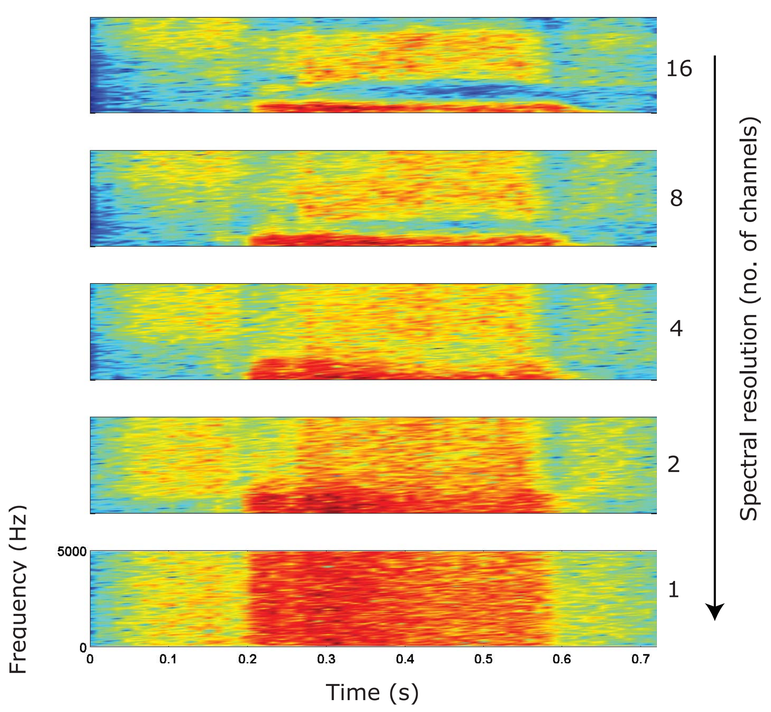
MRC Cognition and Brain Sciences Unit, Cambridge, UK

Introduction

- Cortical integration across sensory modalities is fundamental to successfully identifying objects or events in our environment. One example is intelligibility of distorted speech, which is enhanced if listeners are provided with clear written feedback on speech content (Frost et al., 1988).
- One factor that is known to influence the magnitude of perceptual facilitation is the order of clear and distorted presentations: enhancement is most pronounced when clear feedback is supplied prior to distorted speech (Hannemann et al., 2007; Jacoby et al., 1988).
- To further investigate these timing characteristics, we developed an intelligibility estimation procedure to measure the perceived clarity of speech presented with varying levels of spectral distortion. Speech clarity was manipulated by presenting written words before or after spoken presentations (Experiment 1). Timing was then manipulated in a finer-grained manner by presenting written words with onset asynchronies (SOAs) ranging from -1600 to +1600 ms relative to speech onsets. (Experiment 2).

Stimuli

- Stimuli were sampled from a set of 396 monosyllabic words and presented in spoken and written format.
- Written words were composed of lowercase characters and were presented for 200 ms.
- Spoken words had a mean duration of ~600 ms. They were distorted using a noise-vocoding procedure, which reduces the number of spectral channels in the speech signal (Shannon et al., 1995). Five levels of spectral distortion were used (1, 2, 4, 8 and 16 channels).



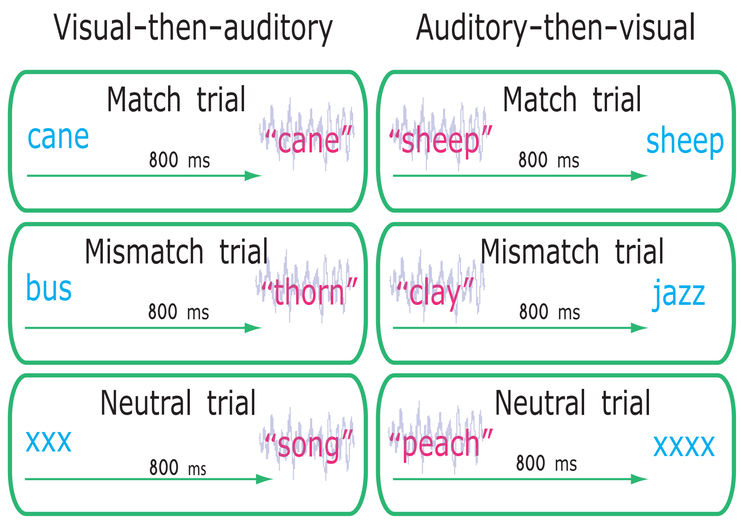
Methods

Task: To rate the clarity of each spoken word on a scale from 1 to 8 (where 1 is hard to understand and 8 is easy to understand).

Participants: 12 (Exp 1) and 14 (Exp 2) participants were tested. All were native speakers of English, aged between 18 and 40 years and had no history of hearing impairment.

Experiment 1

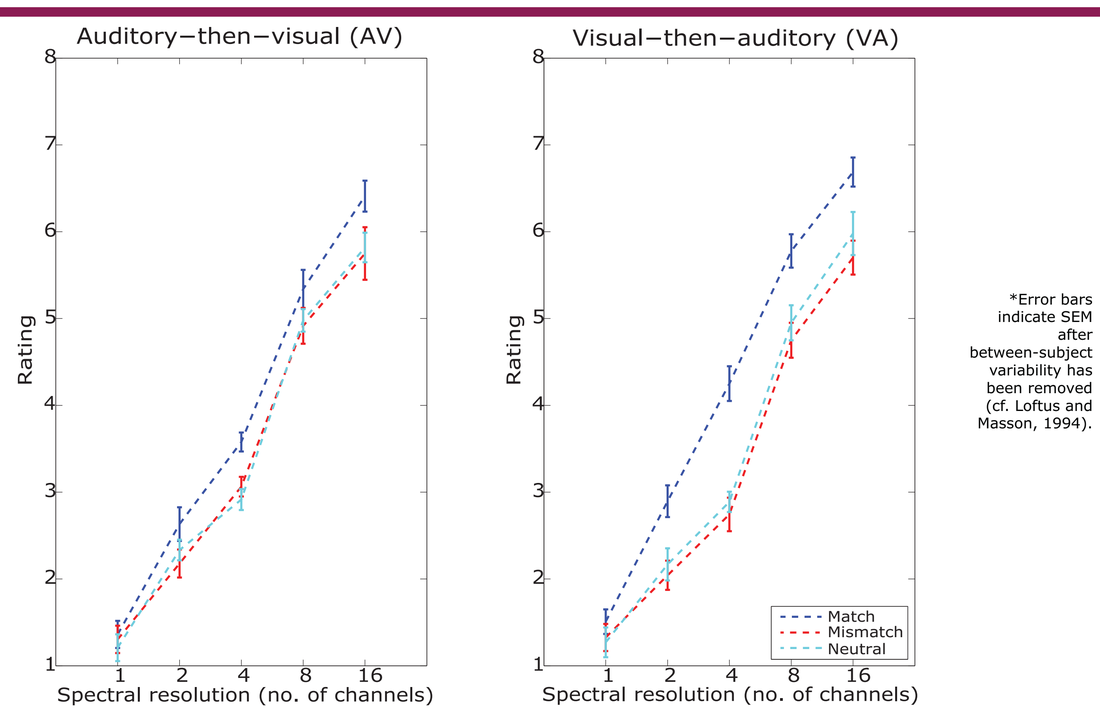
- Written words were presented 800 ms before (visual-then-auditory VA) or after (auditory-then-visual AV) speech onsets. Written words were the same (match) or different (mismatch) to the spoken words. An additional control condition was presented (neutral) in which the visual input consisted of a series of 'x' characters.
- A fully crossed 2 X 3 X 5 design was used with 12 trials per condition combination (360 trials in total). The factors were presentation order (VA and AV), written word congruency (Match, Mismatch and Neutral) and speech spectral resolution (1, 2, 4, 8 and 16 channels).



Experiment 2

- A 2 X 2 X 3 X 5 design was used with 12 trials per condition combination (792 trials in total). The factors were presentation order (VA and AV), written word congruency (match and mismatch), speech spectral resolution (2, 4 and 8 channels) and SOA (100, 200, 400, 800 and 1600 ms). An additional SOA condition (0 ms) was presented but not included in the reported ANOVA.

Results: Experiment 1

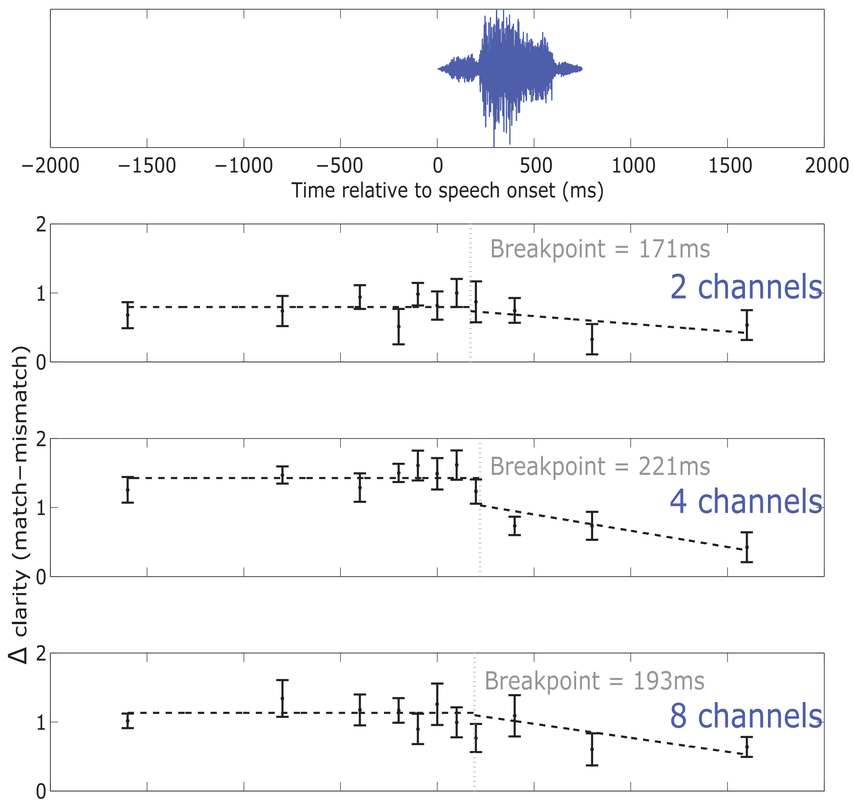


*Error bars indicate SEM after between-subject variability has been removed (cf. Loftus and Masson, 1994).

- Speech clarity improved with increasing spectral resolution, $F(4, 8) = 144.57$, $p < 0.01$ (Greenhouse-Geisser corrected).
- Clarity was affected by written word congruency, $F(2, 10) = 37.65$, $p < 0.01$ (Greenhouse-Geisser corrected). Post-hoc tests revealed that ratings for match trials were greater than mismatch, $t(11) = 6.41$, $p < 0.01$ and neutral trials, $t(11) = 6.24$, $p < 0.01$ (Bonferroni corrected for multiple comparisons). The difference between mismatch and neutral conditions was nonsignificant.
- Speech clarity was higher when written words were presented before spoken words, $F(1, 11) = 5.14$, $p < 0.05$. This effect was most pronounced for matching trials, as revealed by an interaction between presentation order and congruency, $F(2, 10) = 12.14$, $p < 0.01$.

Results: Experiment 2

- The perceptual enhancement for match relative to mismatch trials (Δ clarity) was higher when written words were presented before spoken words, $F(1, 13) = 14.01$, $p < 0.01$, and varied with SOA, $F(4, 10) = 4.9$, $p < 0.01$.
- There was a significant interaction between presentation order and SOA, $F(4, 10) = 4.00$, $p < 0.01$, indicating that whereas Δ clarity declined over positive SOAs, there was no difference over negative SOAs.



- To determine the SOA at which Δ clarity declined, the data were modelled by an invariant function (mean Δ clarity) over SOAs where perceptual enhancement was constant and by a monotonically decreasing function (least-squares fit) for where it declined. The transition between these two functions (**the breakpoint**) was systematically varied from 0 to 400 ms and the total root mean square error (RMSE) calculated for each model obtained. The breakpoint that gave the smallest RMSE was then taken to be the SOA at which Δ clarity declined. Breakpoint estimates were subsequently averaged across distortion level and participants (shown in figure above). The mean breakpoint was found to be **~200 ms** with nonsignificant differences between distortion levels.

Discussion

- The crossmodal perceptual enhancement effects observed may arise from top-down retuning of low-level acoustic representations (**the feedback hypothesis**). Alternatively, they may arise from changes at a late-cognitive stage where information from the auditory and visual inputs are combined to form a perceptual decision about speech clarity (**the feedforward hypothesis**).
- The time-course of audiovisual interactions can be accommodated by either feedback or feedforward model. Top-down retuning would be expected to be less effective past 200 ms when auditory-echoic representations are estimated to decay (Massaro, 1970). Similarly, neural correlates of perceptual decisions have been observed as early as 150 ms (cf. Kaiser et al., 2007) past which the arrival of visual inputs could have less of an influence.
- An MEG study is soon to commence that will aim to distinguish between feedforward or feedback accounts.

References

Frost et al. (1988) Can speech-perception be influenced by simultaneous presentation of print? Journal of Memory and Language 27:741-755.
Hannemann et al. (2007) Top-down knowledge supports the retrieval of lexical information from degraded speech. Brain Research 1153:134-143.
Kaiser et al. (2007) Dynamics of oscillatory activity during auditory decision making. Cereb Cortex 17:2258-2267.
Loftus et al. (1994) Using confidence-intervals in within-subject designs. Psychonomic Bulletin & Review 1:476-490.
Massaro et al. (1970) Preperceptual auditory images. Journal of Experimental Psychology 85:411-417.