1   **Training listeners to detect auditory-visual temporal coherence enhances their ability to exploit visual**

2   **information for auditory scene analysis**

3   Huriye Atilgan[1, 2] and Jennifer K. Bizley[1]

4   [1] The Ear Institute, University College London, UK

5   [2] Current address: Department of Psychiatry, Yale University School of Medicine, New Haven,

6   Connecticut, USA.

7

8   3 figures, 1 table

9   1835 words

10 **Abstract**

11 Listeners engaged in an auditory selective attention task are better able to report brief deviants in a

12 target auditory stream when a task-irrelevant visual stimulus is coherently modulated with the target

13 stream, than when it is coherent with the distractor stream (Maddox et al., 2015). Here, we

14 demonstrate that learning to better discriminate auditory-visual temporal coherence, but not simple

15 exposure to temporally coherent AV stimuli, enhances the ability of listeners to exploit visual

16 information in this task. After 5 short training sessions listeners were able to benefit from auditory-

17 visual temporal coherence both when the visual stimulus is temporally coherent with the target or with

18 the distractor stream, relative to an independently modulated condition. These findings indicate that

19 training to discriminate cross-modal temporal coherence fundamentally changes how listeners exploit

20 visual information for auditory scene analysis.

21

22

23

24

25

1

Introduction

3   While some auditory-visual (AV) correspondences, such as temporal and spatial relations (Spence and

4   Deroy, 2012), seem to be innate or established very early in life, others, such as those that rely on

5   semantic relations, are learned through experience (Navarra et al., 2010). In our previous study (Maddox

6   et al., 2015), human participants performed an auditory selective attention task in which they were

7   required to report brief frequency or timbre deviants in a target auditory stream, while ignoring those

8   occurring in a simultaneous distractor. Listeners were better able to perform this task when changes in

9   the size of a task-irrelevant visual stimulus were temporally coherent with intensity changes in the

10 target auditory stream. Since the visual stimulus conveyed no information about whether or when

11 auditory deviants occurred, we concluded that the only way in which listeners could benefit from AV

12 temporal coherence was if this helped them to better segregate the competing auditory streams. We

13 have since proposed that enhancement in a stimulus dimension (here, pitch or timbre) orthogonal to a

14 cross-modal binding feature (here, temporally coherent changes in auditory intensity and visual size) is

15 strong evidence for cross-modal binding (Bizley et al., 2016) and demonstrated that the integration of

16 visual information into early auditory cortex provides a potential mechanism for these effects (Atilgan et

17 al., 2018).

18 In the present study, we hypothesized that 1) that training participants to better detect AV temporal

19 coherence might facilitate an improved ability to use visual information in the selective attention task,

20 and 2) that the ability of an observer to detect auditory-visual temporal coherence might determine

21 their ability to utilize such information to assist with auditory scene analysis.

Results

23 We recruited participants and randomly assigned them to one of three groups, each of which performed

24 a pre-test and post-test which comprised of the timbre variant of the selective attention task in Maddox

25 et al., (2015) and a measurement of their ability to detect AV temporal coherence. In between the pre-

26 and post-test, one group trained on an AV temporal coherence task (AVTC group, n=12), one group were

27 trained on an AM rate discrimination task with temporally coherent AV stimuli (AVAM group, n=12), and

28 a third group simply performed the pre-test and post-test (control, n=12; Fig. 1A).

1   We first determined whether the five brief (< 40 minutes per session) training sessions experienced by

2   participants in the training groups were sufficient to improve performance in the trained task. Both

3   groups of participants showed improved performance between session 1 and 5 (Fig. 1B/C, pairwise t-

4   test on S1 and S5 AVTC thresholds, t_{22} = 2.961, p=0.007; AVAM thresholds: t_{22} = 4.529, p<0.001).

5   Pairwise comparison between the AV temporal coherence thresholds measured in the pre- and post-

6   test for the three experimental groups, revealed that coherence thresholds were significantly decreased
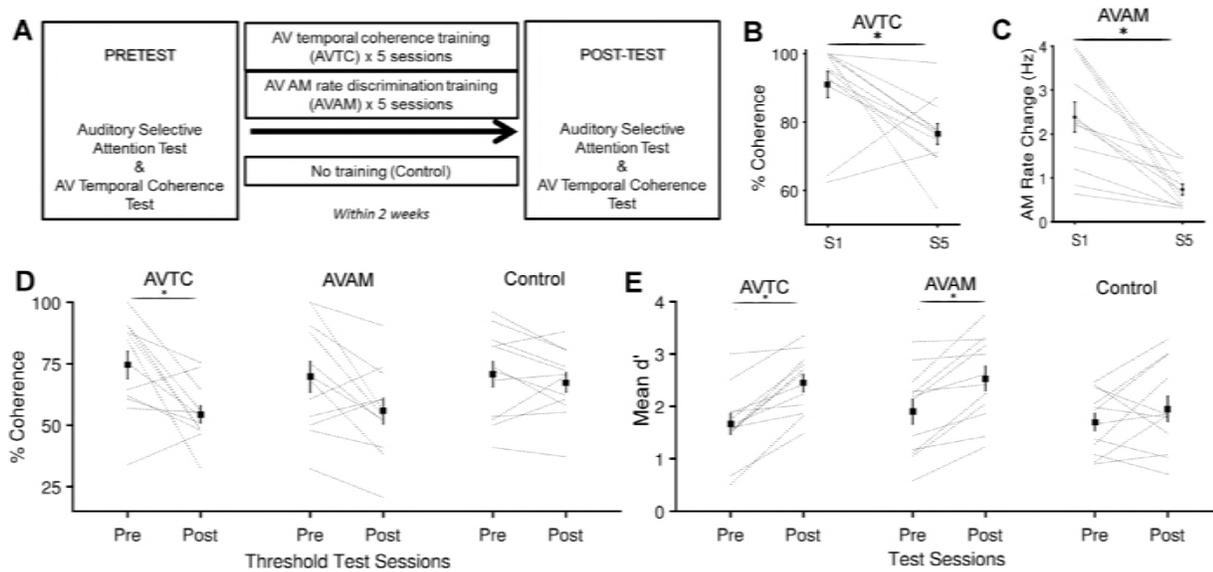
Figure 1: A Experimental design. **B** Training in AVTC task was effective at driving an improvement in AV temporal coherence discrimination between session (S1) and session 5 (S5). Black line show the mean ± SEM across participants, gray lines are individual subjects. **C** Training in the AVAM task was effective at driving an improvement in AM rate discrimination between S1 and S5. **D** AV temporal coherence thresholds values of pretest and post-test for the three groups. **E** Mean d' across AV coherence condition of pretest and post-test for the three groups. *indicates significant differences (Pairwise t-tests p < 0.05)

7   only in the AVTC group (Fig. 1d, t_{22} = 3.081, p = 0.005) and not in the AVAM group (t_{22}=1.69, p=0.104) or

8   in the control group (t_{22} =0.234, p = 0.817). While the change in threshold for the AVTC group was

9   correlated with the change in performance between session 1 and 5 (r = 0.632, p = 0.027), there was no

10   correlation between the changes in AVAM threshold and AV temporal coherence threshold (r = 0.392, p

11   = 0.207).

12   Having confirmed that training was effective, we turned to performance in the selective attention task.

13   While neither training paradigm exposed participants to the deviants they were required to detect in

1   the selective attention task, both groups were exposed to otherwise similar AV stimuli. We predicted

2   that perceptual learning might drive improvements in performance in both trained groups but

3   hypothesized that only the AVTC group would show a change in their ability to exploit visual stimuli. To

4   detect overall changes in performance, we calculated the across condition sensitivity (d'), and directly

5   compared the pre- and post-test data across the three experimental groups. A 2 x 3 way mixed ANOVA

6   with factors of training (pre- and post-test) and experimental group (AVTC, AVAM and control) revealed

7   a significant effect of training ($F_{(1,71)}=10.66$, p=0.002) but the not experimental group ($F_{(2,71)}=1.75$,

8   p=0.181) and no interaction between experimental group and training ($F_{(2, 71)} = 0.86$, p = 0.427).

9   Pairwise comparisons between pre- and post-test data revealed that d' values were significantly

10   different only for the trained groups (Fig. 1D; AVTC group: t_{22} = 3.065, p=0.006, AVAM group: $t2_2$ = 1.920,

11   p = 0.034; control group: t_{22} = 0.854, p = 0.402). Thus both experimental groups, but not the control

12   group, improved their ability to detect timbre deviants in a target auditory stream.
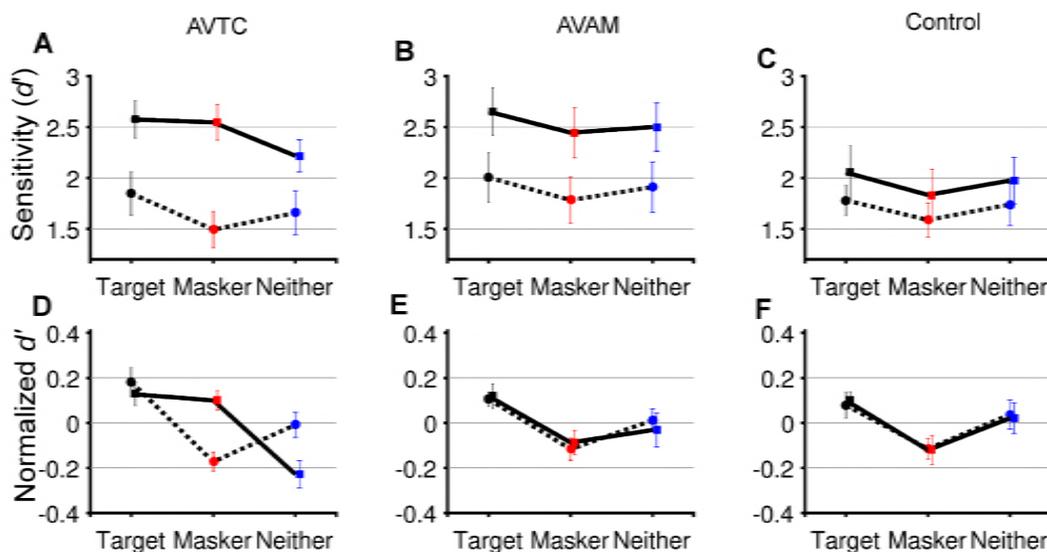
Figure 2: **A-C** pre-test (dashed line) and post-test (solid line) performance in the ASA task according to visual condition. A: AVTC group, B: AVAM group, C: control group. **D-F** Normalized mean ± SEM performance (calculated as within condition d' normalized to across condition performance for pre- and post-test separately).

13   To further understand the impact of training on the ability of participants to benefit from temporal

14   coherence between auditory and visual streams, we next consider each group in turn. We calculated d',

15   bias, hit rates, false alarm rates and visual hit rates for each AV condition and conducted two-way

16   repeated measures ANOVA on these values with factors of AV coherence (Target, Masker and Neither)

1 and training (Pre- and Post-test; Table 1) for each experiment group. In the AVTC group, for d', there

2 was a significant effect of training (F (1, 71) = 9.39, p = 0.006), AV coherence (F (2, 71) = 9.13, p<0.001)

3 and a significant interaction (Fig. 2A, D; F (2, 71) = 7.26, p = 0.002). Post-hoc comparison (p<0.05) across

4 AV coherence condition in the pre-test revealed that participants performed better when the visual

5 stimulus was coherent with the target auditory stream versus the masker auditory stream (Target >

6 Masker; p = 0.0031, Bonferroni corrected α = 0.017). Similar results were obtained for hit rates (see

7 Table 1). In contrast, post-hoc comparisons of the post-test d' scores revealed that, after training,

8 performance was better when the visual stimulus was coherent with either the target or the masker

9 stream than the independent condition (Target> Neither, p = 0.0046; Masker> Neither, p = 0.0055,

10 Bonferroni-corrected α = 0.017), indicating that participants were using visual information in a

11 qualitatively different way after training.

12

		AV Coherence		Training		Interaction between AV Coherence and training	
		F	p	F	p	F	p
AVTC Group	*d'*	9.125	**<.001**	9.393	**0.006**	7.258	**0.002**
	Hit Rates	6.660	**0.002**	6.118	**0.021**	3.210	**0.004**
	False Alarm	2.481	0.095	3.838	0.062	2.755	0.074
	Visual hit rates	0.027	0.972	3.300	0.083	0.083	0.920
AVAM Group	*d'*	5.307	**0.009**	3.688	**0.037**	0.171	0.844
	Hit Rates	3.536	**0.037**	3.676	**0.038**	0.386	0.682
	False Alarm	2.377	0.105	1.629	0.215	0.366	0.695
	Visual hit rates	0.002	0.998	0.640	0.432	0.456	0.637
Control	*d'*	4.600	**0.015**	0.730	0.402	0.039	0.961
	Hit Rates	3.279	**0.023**	0.394	0.537	0.044	0.956
	False Alarm	3.602	0.035	0.074	0.788	0.205	0.816
	Visual hit rates	0.293	0.748	0.628	0.436	0.859	0.430

13

14 **Table 1**: The results of two-way repeated measures ANOVA for each variables (p < 0.05 in bold) for d',
15 Hit rates, false alarm and visual hit rates for three experimental groups.

16

17 Like the participants in the AVTC group, participants in AVAM group were exposed to the target vowel

18 sounds used in the ASA task but were not actively discriminating temporal coherence and were only

19 exposed to temporally coherent stimuli. Training improved their performance in the ASA task (Fig. 2B,

1   E). Both training (F(1,71)=5.31, p = 0.009) and AV coherence condition (F(2,71) = 3.69, p = 0.044)

2   influenced d', but – importantly – there was no interaction (F(2,71)=0.17, p=0.844). Post-hoc

3   comparison (p<0.05) across AV coherence conditions in the pre- and post-test revealed that subjects

4   performed better when the visual stimulus was coherent with the target auditory stream vs the masker

5   auditory stream (Target> Masker, Pre-test: p = 0.0049; Post-test: p= 0.0063). Therefore, this suggests an

6   overall improvement in performance after AVAM group, but no change in the way in which subjects

7   were able to exploit visual cues.

8   Performance in the control group did not differ between pre- and post-test (Fig 2C,G): there was no

9   effect of test on d' scores (F (1,71) = 0.730, p = 0.402), but a significant effect of AV coherence (F (2,71) =

10 4.600, p = 0.015), with no interaction (F(1,71) = 0.039, p = 0.961). Participants performed better when

11 the visual stimulus was coherent with the target auditory stream versus the distractor auditory stream

12 (Target > Masker, Pretest: p = 0.0032; Post-test: p= 0.013).
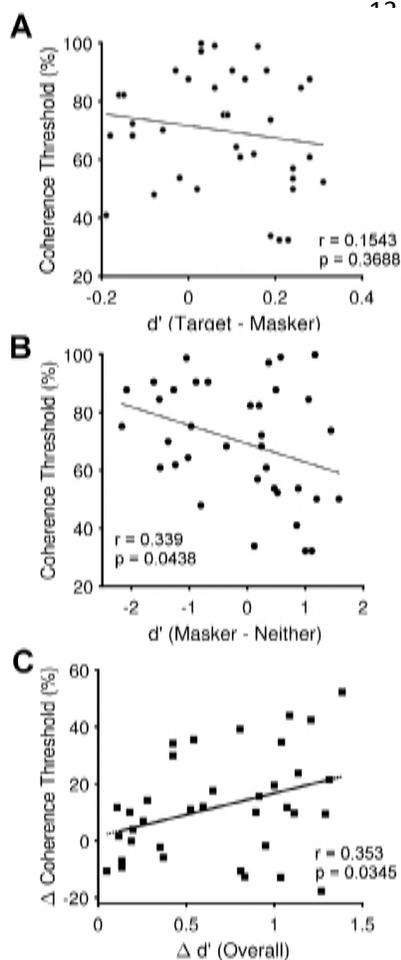
Figure 3

A Scatter plot showing the Target-Masker d' difference for the 36 naïve listeners that completed the pre-test (positive values indicate superior target performance) versus AV coherence threshold (low values indicate better thresholds) for the pre-test data. There was no statistical relationship.

B Scatter plot showing the Masker-Neither d' difference for the 36 naïve listeners that completed the pre-test (negative values indicate masker performance is impaired relative to the independent condition) versus AV coherence threshold for the pre-test data. Participants who showed a benefit for the masker-coherent condition had lower AV coherence thresholds.

C scatter plot showing a positive correlation between the change in overall performance between the pre-test and post-test (positive values indicate improvement) versus post- and pre-test change in AV coherence threshold (positive values indicate improvement).

1

2

3

4   We explored individual differences in performance to test the hypothesis that the ability of naïve

5   listeners to detect AV temporal coherence would predict their ability to benefit from AV temporal

6   coherence in the ASA task: We correlated each listener's AV temporal coherence threshold with the

7   difference between the d' score in the Target coherent and Masker coherent visual condition (Fig 3A).

8   Contrary to our hypothesis, there was no relationship between these values ($r = 0.1543$, $p = 0.3688$), nor

9   was there any relationship between overall performance (across condition d') and AV temporal

10 coherence thresholds ($r = 0.2888$, $p = 0.0882$). Having observed that the AVTC group improved their

11 ability to utilize masker-visual stimulus temporal coherence, we considered whether temporal

12 coherence thresholds might be correlated with the magnitude of the *impairment* that the masker-

13 coherent condition had over the independent condition. The Masker-coherent – Neither comparison

14 was weakly negatively correlated with AV temporal coherence thresholds (Fig.3B; $r = 0.339$, $p = 0.0438$)

15 suggesting a trend where participants with better AV temporal coherence thresholds were more able to

16 exploit the temporal coherence between masker and visual stimulus to yield a performance benefit

17 relative to the independent condition. This finding mirrors the effect of training whereby improving AV

18 coherence thresholds led to an improvement in the masker-coherent condition.

19 Finally, participants' change in performance between pre- and post-test was correlated with their

20 change in ability to detect temporal coherence between auditory and visual stimuli. Participants with a

21 larger change in their AV coherence threshold showed larger improvements in overall performance (Fig.

22 3C; $r =0.353$, $p=0.0347$).

23 **Discussion**

24 Here we demonstrate that five short training sessions can improve both a listener's ability to detect AV

25 temporal correspondence and their ability to exploit cross-modal temporal coherence to segregate a

26 sound mixture. This effect is principally driven by an improvement in the ability of listeners to exploit

27 temporal coherence in the masker-coherent condition. We have demonstrated that the enhancement of

28 one sound in a mixture by a temporally coherent visual stimulus is a stimulus driven, attention

29 independent, bottom-up effect supported by the early integration of auditory and visual information in

1 auditory cortex (Atilgan et al., 2018). In keeping with our behavioral data from naïve listeners, such an

2 enhancement seems likely to facilitate selective attention when the temporally coherent stream is a

3 target, and impair it when that sound is a distractor. Nonetheless, if AV temporal coherence allows the

4 representation of each of two competing sounds to be more distinct within sensory cortex then

5 temporal coherence between target *or* masker should offer an advantage over an independently

6 modulated visual stimulus. That this advantage is weakly present in some naïve listeners (Fig.3B) but

7 appears strongly after training (Fig.1D) implies that it potentially arises from an interaction between the

8 stimulus driven effects we observe in the absence of attention and a top-down process. Substantiating

9 such speculation requires further behavioral and neurophysiological investigation. Previous studies have

10 illustrated that visual cues can assist speech processing in noise (Grant et al., 1998; Schwartz et al., 2004;

11 Helfer and Freyman, 2005). While speech reading abilities are strongly predictive of audiovisual benefit

12 for speech reception thresholds (MacLeod and Summerfield, 1987), lip reading can influence auditory

13 streaming (Devergie et al., 2011), supporting they idea that lip reading benefits in noise potentially

14 comprise of both bottom-up sensory effects that facilitate auditory scene analysis (Atilgan et al., 2018)

15 in addition to conveying phonetic information. An important question in interpreting the significance of

16 our findings is whether the benefits in the auditory selective attention task transfer to other more real-

17 world tasks such as utilizing speech reading in noisy listening conditions.

18 **Materials and methods**

19 **Subjects**

20 42 adults (age range 18–34 years; mean age 28 years; 11 males) with normal hearing and normal or

21 corrected-to-normal vision, participated in the study. Six participants were excluded after the pretest

22 due to poor performance (mean d'<0.8, n=4), or low visual hit rates (<70%, n=2). The remaining 36

23 participants were randomly allocated to three groups. Participants were paid for their participation and

24 gave written informed consent to the study approved by the Ethics Committee of the University College

25 London (ref: 5139).

26 **Stimuli and testing procedure**

27 For the **auditory selective attention** task, the stimuli were generated and presented as described in the

28 *timbre* variant of the previous study (Maddox et al., 2015). That is, on each trial they heard two diotically

29 presented artificial vowels with distinct pitches and timbres whose amplitudes were independently

30 modulated with a low pass (<7 Hz) envelope. Unlike Maddox et al., we did not assess individual timbre

1   discrimination thresholds but instead used a fixed level of difficulty determined using the individual

2   thresholds measured in Maddox et al. For [i] deviants in [a] stimuli this corresponded to a shift of 42 Hz

3   in F1 frequency and 143 Hz for F2, and for [e] deviants in [u] stream there was a shift of 75 Hz for F1,

4   196 Hz for F2. Both vowels could take either F0 value (175 Hz or 190 Hz, counterbalanced) and were

5   equally likely to be target or masker. For AVTC and AVAM training and temporal coherence testing we

6   used the same auditory (either [u] or [e], F0 = 175 or 190 Hz, counterbalanced) and visual (a radius-

7   modulated gray disc) stimuli. Each pre- and post-test session lasted 90 minutes in total and was

8   separated by a maximum of 2 weeks. Participants in the control group did no training sessions but

9   performed the pre- and post-test within 2 weeks (mean ± SD = 5 ± 3 days).

10   For the **AV coherence threshold** test, two artificial vowel sounds (duration 5 seconds, with identical

11   pitch and timbre within a trial, randomly drawn across trials, amplitude modulated < 7 Hz) were

12   consecutively presented, each accompanied by a visual stimulus. In one interval, the radius modulation

13   of the visual stimulus was independent of the envelope of the sound, while in the other interval, the

14   auditory and visual stimulus maintained some degree of temporal coherence. The method of constant

15   limits was used to determine the threshold with subjects performing 20 trials at each coherence level.

16   AV stimuli were generated from 100% coherent in 10% steps to 10% coherent by multiplying the

17   temporally coherent envelope with an independent envelope. Participants were required to select the

18   interval (by pressing 1 or 2 on a button box) in which the temporally coherent pair was presented.

19   Feedback was provided on every trial.

20   The stimuli and procedure in the **AV coherence training** were identical to those used in the threshold

21   test, but with an adaptive three-down one-up rule to determine the coherence level of the stimulus in

22   the next trial, as the goal was require that participants work near to threshold during the training

23   session. In the first training session, the stimuli in the first trial were 100% coherent, and 100%

24   independent. For the first 6 reversals coherency was decreased in 10% steps followed by 5% steps for

25   the following six reversals and by 2.5% steps for the remainder. The procedure was terminated at 18

26   reversals unless a maximum of 150 trials was reached first. For the 2nd-5th training session the first

27   "coherent" stimulus was generated with the average coherence level of the last ten reversals in the

28   previous session. Each training session lasted less than 40 minutes. Feedback was provided on every

29   trial.

30   **Stimuli for AM rate discrimination task:** Two temporally coherent AV stimuli (duration 5 s) were

31   sequentially presented. In one interval the envelope was always generated with a 7 Hz cut off rate,

1   whereas the other was generated with a higher rate (maximum AM cut off rate = 11 Hz). In the sessions

2   of AM rate training, an adaptive three-down one-up rule was used to determine the AM rate of the

3   stimulus in the next trial. In the first session, the first stimulus was generated at the maximum AM rate

4   and differed in AM rate by 1Hz for the first six reversals and 0.5 Hz for the next six reversals and 0.25Hz

5   for the rest of the trials. The procedure was terminated at 18 reversals unless a maximum of 150 trials

6   was reached first. In each consequent session the first stimulus was generated with the average

7   coherence level of the last ten reversals in the previous session. Participants pressed "1" or "2" on the

8   press box to indicate the interval of the faster AV pair. Feedback was provided on each trial.

9   **Statistical Analysis**

10   We used a two-way repeated measures analysis of variance (ANOVA) to test for differences in the d', hit

11   rates, false alarm, and visual hit rates across AV coherence and pretest versus post-test. Statistical tests

12   were performed using the SPSS statistical software (version 20.0, IBM Corp., Armonk, NY, USA) and

13   Matlab (2011b, MathWorks, USA)

14   **Acknowledgments**

19   **References**

20   Atilgan H, Town SM, Wood KC, Jones GP, Maddox RK, Lee AKC, Bizley JK (2018) Integration of Visual
21   Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding.
22   Neuron 97:640-655 e644.
23   Bizley JK, Maddox RK, Lee AK (2016) Defining Auditory-Visual Objects: Behavioral Tests and Physiological
24   Mechanisms. Trends Neurosci 39:74-85.
25   Devergie A, Grimault N, Gaudrain E, Healy EW, Berthommier F (2011) The effect of lip-reading on
26   primary stream segregation. J Acoust Soc Am 130:283-291.
27   Grant KW, Walden BE, Seitz PF (1998) Auditory-visual speech recognition by hearing-impaired subjects:
28   consonant recognition, sentence recognition, and auditory-visual integration. J Acoust Soc Am
29   103:2677-2690.
30   Helfer KS, Freyman RL (2005) The role of visual speech cues in reducing energetic and informational
31   masking. J Acoust Soc Am 117:842-849.
32   MacLeod A, Summerfield Q (1987) Quantifying the contribution of vision to speech perception in noise.
33   British journal of audiology 21:131-141.

1   Maddox RK, Atilgan H, Bizley JK, Lee AK (2015) Auditory selective attention is enhanced by a task-
2   irrelevant temporally coherent visual stimulus in human listeners. eLife 4:e04995.
3   Navarra J, Alsius A, Velasco I, Soto-Faraco S, Spence C (2010) Perception of audiovisual speech synchrony
4   for native and non-native language. Brain Res 1323:84-93.
5   Schwartz JL, Berthommier F, Savariaux C (2004) Seeing to hear better: evidence for early audio-visual
6   interactions in speech identification. Cognition 93:B69-78.
7   Spence C, Deroy O (2012) Crossmodal correspondences: Innate or learned? i-Perception 3:316-318.

8