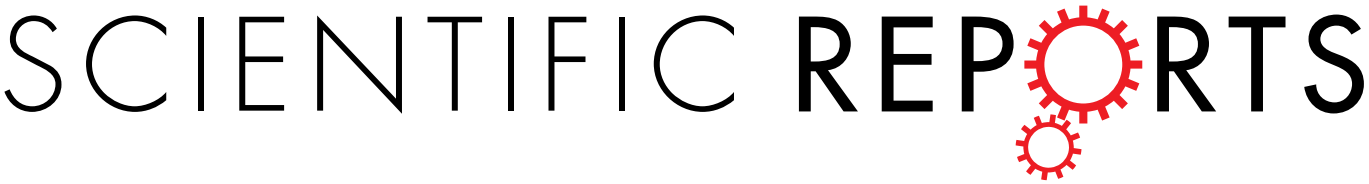


OPEN

Cross-modal noise compensation in audiovisual words

Martijn Baart[1,2], Blair C. Armstrong[1,3], Clara D. Martin[1,4], Ram Frost[1,5,6] & Manuel Carreiras[1,4,7]

Received: 03 August 2016
Accepted: 06 January 2017
Published: 07 February 2017

Perceiving linguistic input is vital for human functioning, but the process is complicated by the fact that the incoming signal is often degraded. However, humans can compensate for unimodal noise by relying on simultaneous sensory input from another modality. Here, we investigated noise-compensation for spoken and printed words in two experiments. In the first behavioral experiment, we observed that accuracy was modulated by reaction time, bias and sensitivity, but noise compensation could nevertheless be explained via accuracy differences when controlling for RT, bias and sensitivity. In the second experiment, we also measured Event Related Potentials (ERPs) and observed robust electrophysiological correlates of noise compensation starting at around 350 ms after stimulus onset, indicating that noise compensation is most prominent at lexical/semantic processing levels.

Information about external events in the world enters the system through our senses, but the input is often degraded. The system can overcome unisensory noise in the signal and stabilize the percept by integrating information from multiple senses (e.g., refs 1 and 2). This is also true for complex signals such as human speech, wherein noise in the auditory speech signal can be compensated for by simultaneously presented visual speech (i.e., the articulating mouth of a speaker, e.g., refs 3–5) or printed text (e.g., ref. 6). Intriguingly, this latter type of compensation occurs despite the fact that print-speech mappings emerged relatively recently in evolution, and the system has not been biologically tailored for this type of audiovisual (henceforth, AV) correspondence.

Noise compensation in speech and in print can be assessed with a "matching task" in which a written and a spoken word are presented simultaneously, with one or both inputs masked by noise, and participants have to indicate whether the same word (congruent trials) or different words (incongruent trials) were presented across the two modalities. The underlying rationale is that asking participants to detect AV correspondence requires one or more interaction(s) between the unisensory inputs, during which cross-modal compensation for noise can be instantiated. This task has been used as an alternative to lexical decision or naming tasks in which one cannot be certain whether participants' responses are driven by the auditory/phonological or visual/orthographic aspects of the signal[6]. Likewise, when AV noise compensation is assessed by comparing noisy unimodal trials to AV trials where there is one noisy signal and one clear signal, one cannot rule out the possibility that the AV gain relative to unimodal trials is driven by the clear unimodal signal, rather than cross-modal noise compensation per se (except when AV stimuli generate percepts that are different from either A or V in isolation, see ref. 7).

Although the matching task has produced evidence for interactions between speech and print (see e.g., refs 6 and 8), the interpretation of correct "match" responses on AV congruent trials (in which one of the signals is masked by noise) is complicated by several factors. Firstly, it is well known that participants may trade-off response speed (i.e., reaction times, henceforth RT) for accuracy (e.g., ref. 9). Therefore, high/low proportions of correct "match" responses may be the result of (i) a general tendency to respond more quickly or more slowly, (ii) successful noise compensation, or (iii) a combination of both (assuming that noise compensation takes time, the degree of compensation could vary with RT). Secondly, the proportion of incorrect "match" responses (when responding "match" to AV incongruent trials) needs to be considered, as participants may adopt a bias to respond "match" or "mismatch" on all trials, irrespective of AV congruency. Thirdly, accuracy is related to the actual noise levels as participants will become more sensitive to the match/mismatch as the signal-to-noise ratio increases. In two experiments, we isolated noise compensation effects in words when controlling for RT, bias and sensitivity. This rigorous assessment approach allowed us to conclude that noise compensation can be explained by accuracy (Experiment 1), and mainly occurs on a lexical/semantic level of processing (Experiment 2).

[1]BCBL. Basque Center on Cognition, Brain and Language, Donostia - San Sebastián, Spain. [2]Department of Cognitive Neuropsychology, Tilburg University, Tilburg, The Netherlands. [3]Department of Psychology & Centre for French & Linguistics at Scarborough, University of Toronto, Toronto, Canada. [4]IKERBASQUE Basque Foundation for Science, Bilbao, Spain. [5]Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, Israel. [6]Haskins Laboratories, New Haven, CT, USA. [7]University of the Basque Country. UPV/EHU, Bilbao, Spain. Correspondence and requests for materials should be addressed to M.B. (email: m.baart@bcbl.eu)

To assess the effects of noise compensation more formally, we quantified this construct using the following logic: First, we assumed that performance on AV stimuli in the clear (henceforth No_{noise}) reflects a base-line ability to detect correspondence of print and speech under optimal conditions. From this baseline, we subtracted the accuracy differences under unimodal noise conditions (i.e., A_{noise} and $V_{noise,}$ for noise in the auditory and visual modality), and summed them. The resulting value was then subtracted from the accuracy detriment for bimodal noise (AV_{noise}) relative to No_{noise} (i.e., noise compensation, or $Acc_{comp} = (No_{noise} - AV_{noise}) - [(No_{noise} - A_{noise}) + (No_{noise} - V_{noise})]$). This compound noise compensation score increases when the $No_{noise} - A_{noise}$ and $No_{noise} - V_{noise}$ differences become smaller (and the $No_{noise} - AV_{noise}$ difference becomes larger). In Experiment 1 participants were asked to respond as quickly as possible and we assessed the relationship between these accuracy differences relative to No_{noise} and noise compensation, when taking into account the interplay between accuracy, RT, bias, and sensitivity.

Experiment 2 was a non-speeded variant of the first task in which we probed the time-course of noise compensation via Event-Related Potentials (ERPs). This allowed us to disentangle the noise compensation process from processes related to bias and sensitivity in post-lexical response processes. Although it is debated whether orthography and phonology interact at all (e.g., ref. 10), there is also evidence for interactions occurring at a (sub) lexical level (e.g., refs 6, 11 and 12). As highlighted above, we assume that noise compensation has to be instantiated during at least one such interaction, and the time-course of the neural correlates that underlie noise compensation would reveal at which level(s) of processing noise compensation occurs.

To preview the main results, noise compensation increased when the difference between the unimodal noise conditions and the No_{noise} condition decreased, which persisted when RT, bias and sensitivity were controlled for. Furthermore, in a 350–390 ms window after stimulus onset, the ERP difference waves of the unimodal noise conditions relative to No_{noise} were not modulated by bias or sensitivity, but could explain noise compensation in accuracy. This result indicated that that compensation for noise in speech and print predominantly occurs at a lexical/semantic level of processing.

Results

Behavioral. First, we explored how accuracy, bias, sensitivity and accuracy-based noise compensation were modulated by RT, using the noise compensation score described in the introduction (i.e., Acc_{comp}: $[No_{noise} - AV_{noise}] - ([No_{noise} - A_{noise}] + [No_{noise} - V_{noise}])$. Based on signal detection theory, we quantified sensitivity through d-prime (d′) and used response criterion *c* as a measure of response bias (e.g., refs 13 and 14). As illustrated in Fig. 1, the linear trends between RT and accuracy, bias, sensitivity and Acc_{comp} values in the speeded task (Experiment 1) always carried over to the non-speeded task (Experiment 2). This is reflected by the fact that averages in Experiment 2 increased/decreased relative to Experiment 1 (although differences between were not always significant), according to the positive/negative linear relationships related to RT.

Because of the way we defined the compound Acc_{comp} score, it should increase when accuracy differences between A_{noise}/V_{noise} and No_{noise} become smaller, and conversely, when unimodal noise detriments are high, Acc_{comp} is low. Before assessing this however, we first determined how accuracy related to bias and sensitivity. To do so, we correlated accuracy with bias and sensitivity for each noise condition separately, which was preferred over multiple linear regression because of multicollinearity between the predictors. In both experiments, bias and sensitivity were highly and significantly correlated with accuracy in most conditions: accuracy increased with a decrease in bias, and an increase in sensitivity. This finding also held for the partial correlations in which we controlled for RT (in Experiment 1 only, see Table 1).

Taken together, it is clear that effects of RT, bias and sensitivity cannot be ignored when explaining Acc_{comp} through accuracy. Therefore, the correlations between Acc_{comp} and the A_{noise}, V_{noise} and AV_{noise} accuracy differences relative to No_{noise} were also computed when controlling for bias and sensitivity in all conditions (the last column in Table 1). These partial correlations showed that in both experiments, noise compensation increased when the unimodal noise accuracy detriments relative to No_{noise} decreased. This was not the case for the $No_{noise} - AV_{noise}$ accuracy difference, as it was not significantly correlated with Acc_{comp} in Experiment 1, and was negatively correlated with Acc_{comp} in Experiment 2 (i.e., the smaller the difference between AV_{noise} accuracy and No_{noise} accuracy, the larger Acc_{comp}). The partial correlations in Experiment 2 were also clearly stronger than the uncontrolled correlations, which suggests that in the non-speeded task, processes related to bias and sensitivity somehow interfered with noise compensation. It is therefore likely that processing of noise compensation, bias and sensitivity overlap in time (at least partially). This was assessed empirically in the ERP data.

ERPs. To determine the neural correlates of Acc_{comp}, we constructed the ERP_{comp} difference wave (using the same formula described previously, see Supplementary Information for an expanded rationale) and correlated this waveform with the behavioral Acc_{comp} measure (see Fig. 2a). Based on this exploratory correlation analysis, we selected five 40-ms time-windows in which ERP_{comp} correlated with Acc_{comp} in at least three electrodes that formed a spatial cluster. As illustrated in Fig. 2a, these criteria included all correlations that coincided in time and topography, and were purposely chosen to include all potentially relevant spatial-temporal clusters.

Next, we correlated ERP_{comp} with bias and sensitivity in all conditions. Figure 2b plots these relationships and shows that there was temporal-spatial overlap between these correlations and those between ERP_{comp} and Acc_{comp} (see for example the 500–540 ms window).

Next, we computed the ERP_{comp} averages (over both time and electrodes) for each of the temporal spatial clusters, and correlated the averaged ERP_{comp} with Acc_{comp}, bias, and sensitivity. The results of these analyses are summarized in Fig. 3. Because the temporal-spatial clusters were selected based on the $ERP_{comp} \times ACC_{comp}$ correlations, it is not surprising that these correlations were significant in all clusters. More critically, the analyses confirmed the visual pattern in Fig. 2b: V_{noise} bias and AV_{noise} sensitivity were correlated with ERP_{comp} in the 500–540 ms window, and V_{noise} bias was also correlated with ERP_{comp} in the 590–630 ms window. To determine

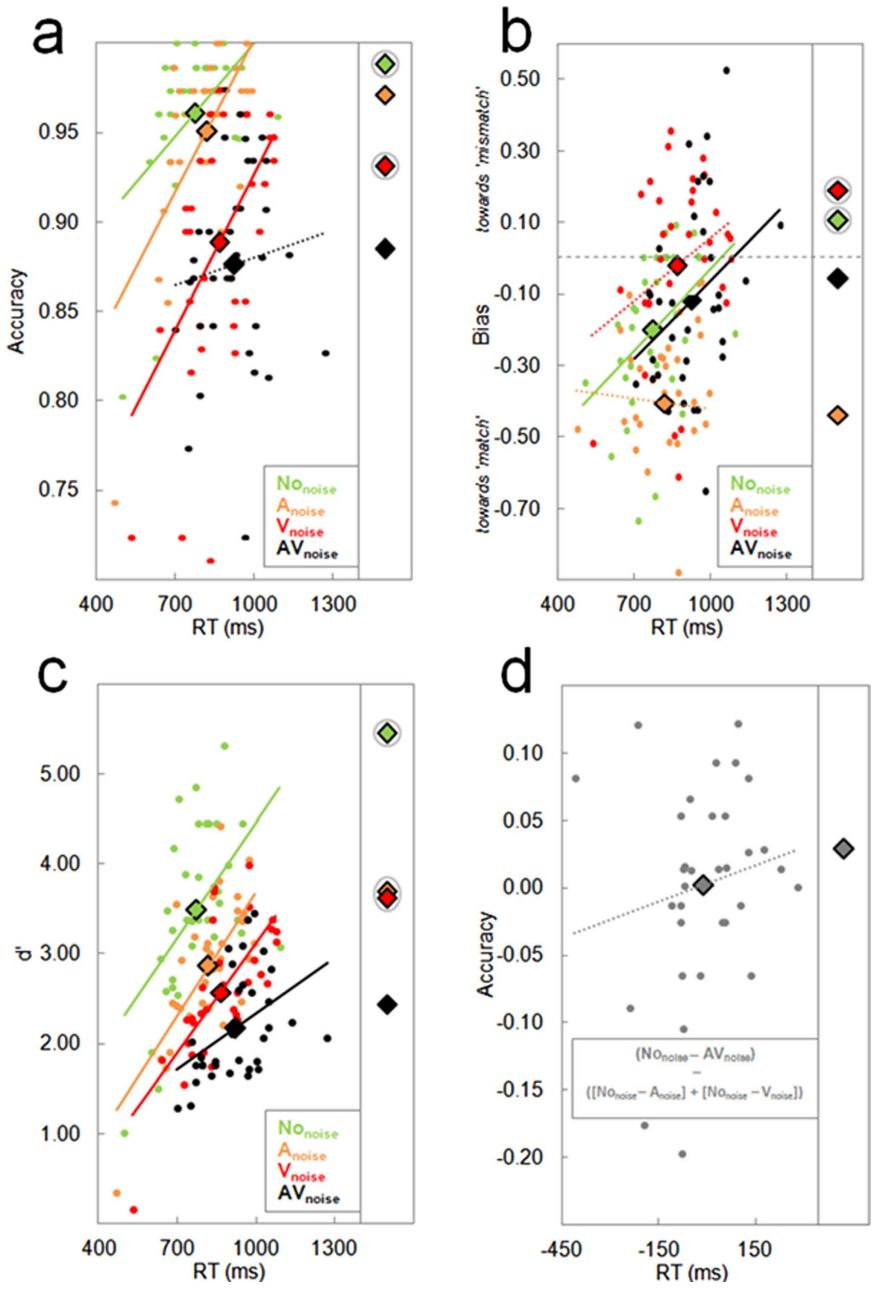


Figure 1. Scatter plots, linear trends and group averages between RT and accuracy (panel a), bias (panel b), sensitivity (panel c) and Acc_{comp} (panel d). The individual data (dots) and linear trends are data from Experiment 1, with group averages represented by diamonds. The narrow plots on the right of each panel represent the averages in Experiment 2 (where RT was not informative given the non-speeded nature of the task). All linear trends between RT and accuracy were positive, except for A_{noise} bias. Dotted lines indicate that linear trends were not significant, and grey circles indicate significant differences between group averages in Experiments 1 and 2.

which of the components in the ERP_{comp} waveform drove the correlation with Acc_{comp}, and thereby understand the neural basis of the behavioral effects, we correlated the amplitude difference-waves (relative to No_{noise}) with Acc_{comp}. For the temporal-spatial clusters in which ERP_{comp} was also correlated with bias and/or sensitivity, these correlations involved the residual ERP difference waves after regressing out the bias/sensitivity effect(s).

As illustrated in Fig. 3, the larger the difference between the No_{noise} and AV_{noise} ERPs in the 260–300 ms window, the larger Acc_{comp} was. In the 350–390 ms window, correlations in the same direction became significant for both unimodal noise ERP differences. These correlations were also observed in the 500–540 ms window, and the $No_{noise} - A_{noise}$ ERP difference was also correlated with Acc_{comp} in the 590–630 ms window.

				r	p	r_{p1}	p	r_{p2}	p
Exp. 1	No_{noise} acc	×	No_{noise} c	−0.021	>0.250	−0.283	0.110	—	—
	A_{noise} acc	×	A_{noise} c	−0.364	0.034	−0.446	0.009	—	—
	V_{noise} acc	×	V_{noise} c	−0.292	0.094	−0.559	0.001	—	—
	AV_{noise} acc	×	AV_{noise} c	−0.464	0.006	−0.537	0.001	—	—
	No_{noise} acc	×	No_{noise} d′	0.860	<0.001	0.814	<0.001	—	—
	A_{noise} acc	×	A_{noise} d′	0.904	<0.001	0.821	<0.001	—	—
	V_{noise} acc	×	V_{noise} d′	0.811	<0.001	0.734	<0.001	—	—
	AV_{noise} acc	×	AV_{noise} d′	0.567	<0.001	0.577	<0.001	—	—
	Acc_{comp}	×	$No_{noise} - A_{noise}$ acc	−0.487	0.003	−0.512	0.004	−0.895	<0.001
	Acc_{comp}	×	$No_{noise} - V_{noise}$ acc	−0.545	0.001	−0.552	0.002	−0.686	<0.001
	Acc_{comp}	×	$No_{noise} - AV_{noise}$ acc	0.369	0.032	0.250	0.183	0.309	0.162
Exp. 2	No_{noise} acc	×	No_{noise} c	−0.771	<0.001	—	—	—	—
	A_{noise} acc	×	A_{noise} c	−0.828	<0.001	—	—	—	—
	V_{noise} acc	×	V_{noise} c	−0.559	0.002	—	—	—	—
	AV_{noise} acc	×	AV_{noise} c	−0.514	0.005	—	—	—	—
	No_{noise} acc	×	No_{noise} d′	0.635	<0.001	—	—	—	—
	A_{noise} acc	×	A_{noise} d′	0.672	<0.001	—	—	—	—
	V_{noise} acc	×	V_{noise} d′	0.569	0.002	—	—	—	—
	AV_{noise} acc	×	AV_{noise} d′	0.723	<0.001	—	—	—	—
	Acc_{comp}	×	$No_{noise} - A_{noise}$ acc	−0.425	0.024	—	—	−0.820	<0.001
	Acc_{comp}	×	$No_{noise} - V_{noise}$ acc	−0.415	0.028	—	—	−0.765	<0.001
	Acc_{comp}	×	$No_{noise} - AV_{noise}$ acc	−0.062	>0.250	—	—	−0.457	0.043

Table 1. Correlations between accuracy and bias (criterion c), accuracy and sensitivity (d′), and noise compensation (Acc_{comp}) and accuracy differences of the noise conditions relative to No_{noise}. Correlation values r_{p1} are partial correlations controlled for RT (correlations involving Acc_{comp} were controlled for RT in all noise conditions). Correlation values r_{p2} are partial correlations that controlled for RT, bias and sensitivity in all conditions. Columns labeled 'p' represent the corresponding p-values.

Discussion

Across two experiments, we quantified noise compensation in simultaneously presented speech and print by subtracting the summed unimodal noise detriments relative to No_{noise} from the bimodal noise detriment relative to No_{noise}. Because RT, bias, and sensitivity modulated accuracy, we assessed noise compensation after controlling for those variables, thereby providing a more rigorous treatment of noise compensation effects than in previous work (e.g., ref. 6).

The ERP data from Experiment 2 allowed us to disentangle processes related to noise compensation from processes related to bias and sensitivity at the neural level, and we observed that at 350–390 ms after stimulus onset, the ERP differences for the unimodal noise conditions relative to No_{noise} were correlated with Acc_{comp}: the larger the ERP difference was, the larger Acc_{comp} was (as measured about 1500 ms later). The same correlations were observed in a 500–540 ms window, but during this time, the ERPs were also modulated by V_{noise} bias and AV_{noise} sensitivity. However, after regressing out the direct effect of bias and sensitivity on the ERP difference waves, the correlations with Acc_{comp} remained significant. As explained in the Supplementary Material, the timing of these effects aligns with an N400 effect that is usually associated with lexical/semantic processing, suggesting that noise compensation in audiovisual word processing occurs predominantly on a lexical/semantic level.

This is particularly interesting when considering that other research has demonstrated that integration of print and speech can occur on sub-lexical levels[15,16]. For example, effects of AV incongruence between simultaneously printed and spoken single letters takes place already at around 200 ms (i.e., when assessed in electrophysiological oddball paradigms, see refs 17 and 18. In the current study, we did observe positive correlations between the ERP_{comp} score and the Acc_{comp} score starting at around 200 ms, but these were not as robust as the later negative correlations. Moreover, the earliest time-window in which Acc_{comp} was correlated with an individual ERP difference wave was 260–300 ms, but since it involved the $No_{noise} - AV_{noise}$ ERP difference (i.e., the larger the ERP difference, the larger noise compensation), and not the $No_{noise} - A_{noise}$ or $No_{noise} - V_{noise}$ differences, it is not directly interpretable as noise compensation. What, then, is the source of the discrepancy between our findings and previous work? One possibility is that by using full words, our task engages a different, potentially more natural/automatic set of audiovisual integration processes than single letter tasks, and this interaction occurs at a higher level of representation. A second possibility is that noise compensation is (partially) constituted on a sub-lexical level (see also refs 6 and 11). However, the effects of such compensation would then be fed forward (and potentially magnified) in higher order representations on lexical/semantic levels that are most relevant for the response system (although it is currently not exactly clear how and why response systems engage different levels of representation in different tasks, see ref. 19 for discussion).

A second set of highly relevant observations were that the correlations between Acc_{comp} and $No_{noise} - A_{noise}$ accuracy were stronger than for $No_{noise} - V_{noise}$ accuracy (see Table 1), and the correlations between the $No_{noise} - A_{noise}$ ERP difference and Acc_{comp} were observed in more time-windows than the correlations between

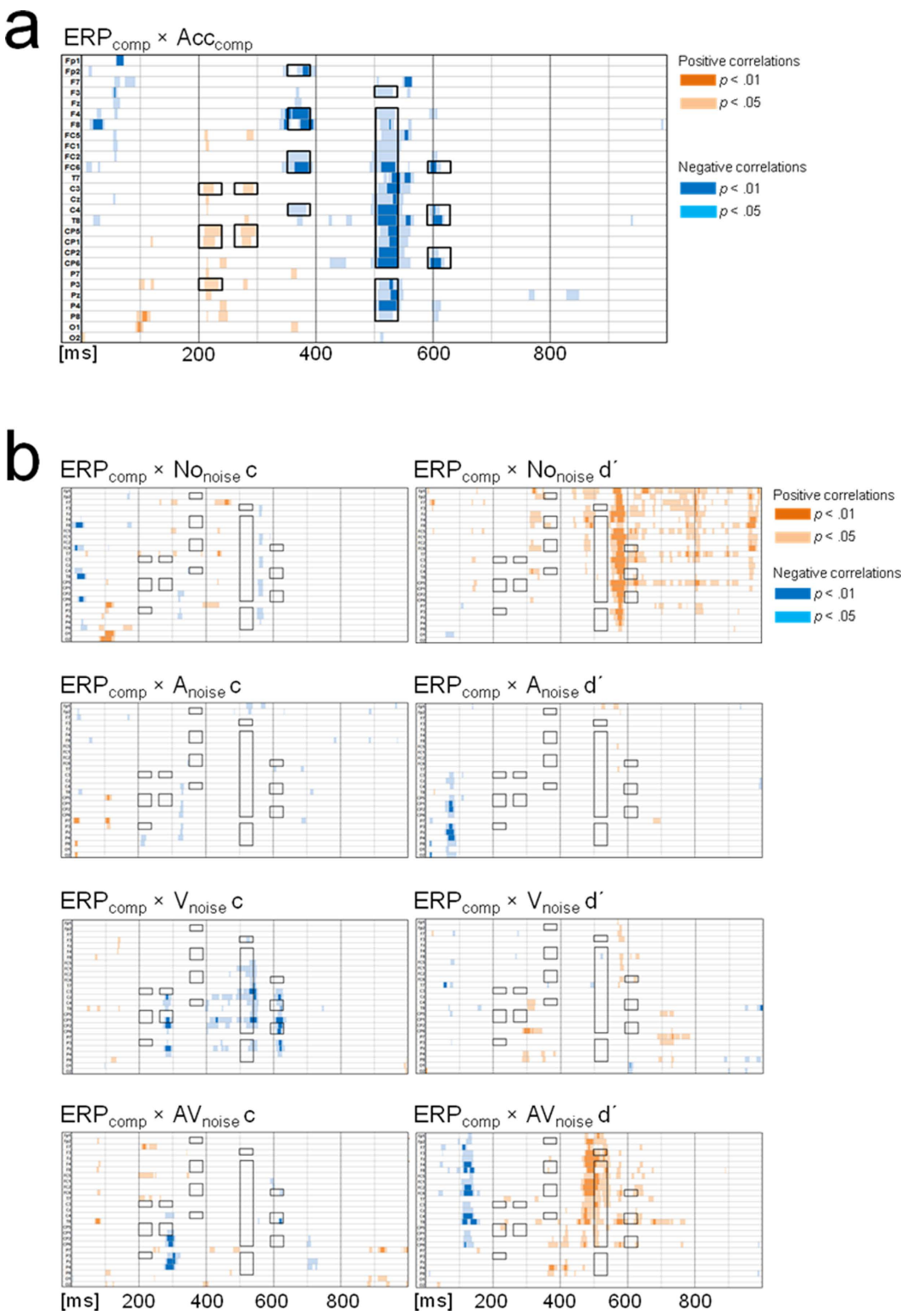
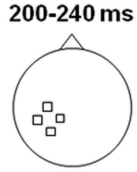


Figure 2. Correlations between ERP_{comp} and Acc_{comp} (panel a), between ERP_{comp} and bias (criterion c; panel b, left plots) and between ERP_{comp} and sensitivity (d′; panel b, right plots). The black outlines represent the clusters of electrodes in five 40-ms time-windows where ERP_{comp} and Acc_{comp} correlated in at least three electrodes (i.e., 200–240 ms, 260–300 ms, 350–390 ms, 500–540 ms, 590–630 ms).

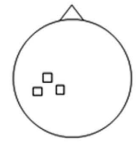
the $No_{noise} - V_{noise}$ ERP difference and Acc_{comp}. This suggests that print-driven noise compensation is more prominent than speech-driven compensation, and is consistent with similar asymmetric effects reported by Borowsky, Owen and Fonos[11]. In two related tasks, they presented participants with printed syllables in combination



200-240 ms

ERP_{comp} × Acc_{comp}, bias and sensitivity	r	p
ERP_{comp} × Acc_{comp}	.436	.021
ERP_{comp} × No_{noise} c	.103	>.250
ERP_{comp} × A_{noise} c	-.138	>.250
ERP_{comp} × V_{noise} c	-.057	>.250
ERP_{comp} × AV_{noise} c	.250	.200
ERP_{comp} × No_{noise} d′	.167	>.250
ERP_{comp} × A_{noise} d′	.146	>.250
ERP_{comp} × V_{noise} d′	.035	>.250
ERP_{comp} × AV_{noise} d′	.311	.108

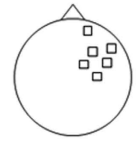
ERP differences × Acc_{comp}	r	p
$No_{noise} - A_{noise}$	-.232	.235
$No_{noise} - V_{noise}$	-.266	.172
$No_{noise} - AV_{noise}$	-.037	>.250



260-300 ms

ERP_{comp} × Acc_{comp}, bias and sensitivity	r	p
ERP_{comp} × Acc_{comp}	.423	.025
ERP_{comp} × No_{noise} c	.292	.132
ERP_{comp} × A_{noise} c	.110	>.250
ERP_{comp} × V_{noise} c	-.252	.196
ERP_{comp} × AV_{noise} c	-.009	>.250
ERP_{comp} × No_{noise} d′	.046	>.250
ERP_{comp} × A_{noise} d′	-.083	>.250
ERP_{comp} × V_{noise} d′	.064	>.250
ERP_{comp} × AV_{noise} d′	.134	>.250

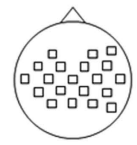
ERP differences × Acc_{comp}	r	p
$No_{noise} - A_{noise}$	.296	.126
$No_{noise} - V_{noise}$	.194	>.250
$No_{noise} - AV_{noise}$	.453	.016



350-390 ms

ERP_{comp} × Acc_{comp}, bias and sensitivity	r	p
ERP_{comp} × Acc_{comp}	-.503	.006
ERP_{comp} × No_{noise} c	-.199	>.250
ERP_{comp} × A_{noise} c	-.285	.141
ERP_{comp} × V_{noise} c	-.010	>.250
ERP_{comp} × AV_{noise} c	.067	>.250
ERP_{comp} × No_{noise} d′	.336	.080
ERP_{comp} × A_{noise} d′	.062	>.250
ERP_{comp} × V_{noise} d′	-.056	>.250
ERP_{comp} × AV_{noise} d′	.276	.155

ERP differences × Acc_{comp}	r	p
$No_{noise} - A_{noise}$	.384	.044
$No_{noise} - V_{noise}$	.505	.006
$No_{noise} - AV_{noise}$	.355	.063

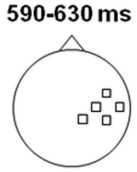


500-540 ms

ERP_{comp} × Acc_{comp}, bias and sensitivity	r	p
ERP_{comp} × Acc_{comp}	-.539	.003
ERP_{comp} × No_{noise} c	-.264	.175
ERP_{comp} × A_{noise} c	.008	>.250
ERP_{comp} × V_{noise} c	-.448	>.017
ERP_{comp} × AV_{noise} c	-.084	>.250
ERP_{comp} × No_{noise} d′	.157	>.250
ERP_{comp} × A_{noise} d′	-.061	>.250
ERP_{comp} × V_{noise} d′	-.263	.176
ERP_{comp} × AV_{noise} d′	.401	.035

ERP differences × Acc_{comp} (r_2 = correlation on residual ERP difference waves after regressing out the direct effect of V_{noise} bias and AV_{noise} d′)

	r	p	r_2	p
$No_{noise} - A_{noise}$	.438	.020	.444	.018
$No_{noise} - V_{noise}$	.544	.003	.564	.002
$No_{noise} - AV_{noise}$	.270	.165	.294	.129



590-630 ms

ERP_{comp} × Acc_{comp}, bias and sensitivity	r	p
ERP_{comp} × Acc_{comp}	-.458	.014
ERP_{comp} × No_{noise} c	-.222	>.250
ERP_{comp} × A_{noise} c	.042	>.250
ERP_{comp} × V_{noise} c	-.412	>.029
ERP_{comp} × AV_{noise} c	-.351	.067
ERP_{comp} × No_{noise} d′	.325	.092
ERP_{comp} × A_{noise} d′	-.127	>.250
ERP_{comp} × V_{noise} d′	.142	>.250
ERP_{comp} × AV_{noise} d′	.343	.074

ERP differences × Acc_{comp} (r_2 = correlation on residual ERP difference waves after regressing out the direct effect of V_{noise} bias)

	r	p	r_2	p
$No_{noise} - A_{noise}$	.417	.027	.437	.020
$No_{noise} - V_{noise}$	.339	.078	.342	.074
$No_{noise} - AV_{noise}$	.216	>.250	.213	>.250

Figure 3. Correlations between ERP_{comp} averaged across electrodes (electrode sites correspond to Fig. 2 and are indicated on the scalp) and Acc_{comp}, bias (criterion c) and sensitivity (d′) in 40 ms windows of averaged activity (i.e., 200–240 ms, 260–300 ms, 350–390 ms, 500–540 ms, 590–630 ms). The right panels show the correlations between the ERP difference waves (relative to Nonoise) and Acc_{comp}, in which the direct effects of V_{noise} bias (500–540 ms, and 590–630 ms) and AV_{noise} sensitivity (500–540 ms) were regressed out.

with noise-masked auditory syllables (e.g., /ta/), and asked participants to indicate what they had heard via a two-alternative-forced-choice response probe (e.g., *heard /ta/* or *heard /da/*); or, they presented the auditory syllable in the clear, and the printed one in noise, and asked participants what they had seen. Critically, the syllables could be incongruent as well, and the authors observed that the facilitating effect of congruent print on noise-masked speech was larger than the cost induced by incongruent print. In contrast, the facilitation/cost of congruent/incongruent speech on noise-masked print was symmetrical. In a follow-up study with words rather than syllables[12], the same pattern of results was observed and the authors therefore argued in favor of facilitation-dominant connections from orthography to phonology on both sublexical and lexical levels of processing, that outweigh connections from phonology to orthography.

A similar asymmetry is observed when the visual signal consists of an articulating mouth rather than print. A clear example is provided by past work in which auditory and visual speech are purposely selected to reflect the most ambiguous stimulus on the boundary of a phonetic or lip-read contrast (i.e. an ambiguous speech sound

or lip-read video in between /aba/ and /ada/ that is perceived as either /aba/ or /ada/ in ~50% of trials). Repeated exposure to these ambiguous speech segments in combination with clear unambiguous input in the other modality has shown that the effect of unambiguous visual speech on the perception of ambiguous sounds (e.g., ref. 20) is larger than the effect of unambiguous auditory speech on ambiguous visual speech[21]. Clearly, visual speech and printed text do not engage identical neurocomputational circuits as auditory and visual speech. However, it is argued that mechanisms underlying AV speech integration for stimuli in which the visual input consists of print or visual speech are similar to some extent (e.g., ref. 22). A general asymmetry in cross-modal effects could therefore indicate that the system is biased towards compensating for auditory noise via visual speech (e.g., ref. 4) or print (e.g., ref. 6), and not vice versa due to the properties of domain-general noise compensation mechanisms.

Despite the fact that interactions between speech and print may occur on sub-lexical (e.g., refs 15 and 16) and lexical levels (e.g., ref. 12), our data suggest that noise compensation occurs mainly on a lexical/semantic level in the context of words. Auditory lexical processing is often observed before 350 ms [23–28], whereas we observed robust effects of noise compensation at 350–390 ms, and at 500–540 ms. Furthermore, the combined set of findings suggest a general primacy of visual stimuli in AV integration. This provides an intriguing direction for future research: Why would visual stimuli take precedence over auditory stimulation when constraint satisfaction is performed using cross-modal information? This is especially relevant if one considers the differences in acquisition trajectories of the different sources of information. That is, auditory and visual-speech information are both available early during development (e.g., two month old infants can already detect the correspondence between auditory and visual speech, ref. 29). In contrast, linguistic information conveyed by print becomes available only relatively late, with reading typically bootstrapping off the spoken language system.

The current paradigm, analytical method and findings can be used to conduct rigorous, direct comparisons of the time-courses that underlie cross-modal noise compensation for visual speech and print, and other analogous cross-modal integration processes (such as integration of speech and visual speech). This is particularly crucial because auditory and visual speech are tightly linked (e.g., the temporal characteristics are highly correlated) and biologically grounded through evolution, whereas this is not the case for the relationship between auditory speech and print. The present work therefore offers an innovative avenue for contrasting flexible domain-general vs. neurobiologically optimized domain-specific processing.

Methods

Experiment 1. *Participants.* Spanish adults (34 in total, 33 right-handed, 23 females, mean age = 21 years, SD = 2.3 years) participated in the experiment for payment. All participants had normal or corrected-to-normal vision and reported no language, hearing, or motor impairments. All participants used Spanish as their primary language and gave their informed consent prior to testing. The experiment was conducted in accordance with the Declaration of Helsinki and approved by the BCBL ethics committee.

Stimuli. The candidate stimulus population was determined using the EsPal subtitle database[30] and was constrained using three criteria: word frequency (frequency ranged between 1 and 20 per million), word length (the number of phonemes ranged in between 3 and 10), and number of syllables (which was restricted to 2 or 3). Homophones (e.g., 'vaca' [cow], 'baca' [roof box]) were excluded, and 2659 words fulfilled the inclusion criteria. With the Stochastic Optimization of Stimuli (SOS) algorithm and software[31], 400 pairs of words were identified (i.e., 800 words) in which the differences on the aforementioned criteria were minimal. For each pair, one word formed the basis for an audiovisually congruent stimulus, whereas the other was manually matched with a different word with similar psycholinguistic properties to form an incongruent stimulus. All stimuli were inspected by two native speakers who removed inflections, compound words, foreign words, and dialect-inappropriate words, which resulted in a final set of 304 test stimuli. The AV incongruent items were included for task purposes and not analyzed in detail (see also, ref. 32), but they were used to compute sensitivity and bias. A female native speaker of Spanish was recorded in a sound-proof booth while reading two differently ordered lists that included all items (i.e., each item was recorded twice). After cutting all items at on/offset, a native speaker of Spanish selected the recording of each stimulus that sounded most natural for use in the experiment.

Procedure. The experiment was run in a sound attenuated and dimly-lit booth. Participants were seated ~80 cm from a 48 cm (19-in) CRT monitor (100 Hz vertical refresh, 1024 px × 768 px resolution) on which the visual words were displayed in Arial font (font height was 5% of the display height, or ~38 px). Auditory words were delivered at a comfortable listening level through headphones.

In total, there were 608 trials (presented with PsychoPy, see ref. 33), with 152 trials per noise level (No_{noise}, A_{noise}, V_{noise} and AV_{noise}). For each of the four noise levels, 76 trials were AV congruent, and 76 were incongruent. Visual noise was created by superimposing a rectangular field of 950 randomly positioned white dots (3 pixels in diameter) over an area slightly larger than the longest word in the stimulus set, and auditory noise was created by replacing 85% of the auditory waveform with signal-correlated noise.

Trials were distributed in random order across 17 blocks, with the first and last block containing 19 trials, and the 15 middle blocks containing 38 trials each. Blocks were separated by self-paced breaks. As shown in Fig. 4, each trial began with a fixation cross (750 ms), that was followed by a black screen (50 ms), after which the stimulus was delivered. Participants were asked to indicate whether speech and print matched or not by pressing the left or right CTRL key (as fast as possible, and within a 2500 ms window after stimulus onset) with the left or right index finger ("match" responses were indicated with the dominant hand). Once participants responded, the next trial began automatically after 250 ms. After each block, participants received feedback on the monitor regarding their accuracy and RT (based on correct responses) to keep them motivated. The experiment was preceded by a short practice session (10 trials), during which participants were acquainted with the task and procedure.

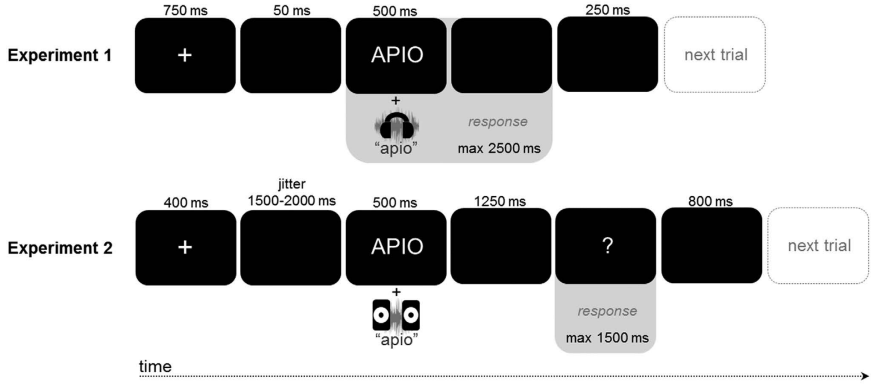


Figure 4. Trial Overview. When participants responded outside the response windows (indicated by the gray areas), they received a 2000 ms message on the screen stating that they should try to respond faster (Experiments 1 and 2) or later (Experiment 2), and such trials were excluded from analyses. A response within the window immediately prompted a 250 ms (Experiment 1) or 800 ms (Experiment 2) black screen.

Experiment 2. *Participants.* 35 new right-handed Spanish adults with the same profile as those in Experiment 1 participated in return for payment. As described below, six participants were excluded because of substantial artifacts in the EEG signal. One participant was excluded because responses fell outside of the response time window (see Fig. 4) in 19% of trials. In the final set of 28 participants, there were 19 females, and mean age was 23 (SD = 2.3 years). All participants used Spanish as their primary language and gave their informed consent prior to testing. The experiment was conducted in accordance with the Declaration of Helsinki and approved by the BCBL ethics committee.

Stimuli and procedure. The stimuli were the same as in Experiment 1, but procedural details were tailored toward a non-speeded ERP paradigm: auditory words were delivered via two regular computer speakers (at ~65 dBA) placed on both sides of the monitor, and trial timing differed from experiment 1 (see Fig. 4).

EEG recording and analyses. The EEG was recorded at a 500 Hz sampling rate using a 32-channel BrainAmp system (Brain Products GmbH) and 28 Ag/AgCl electrodes that were placed in an EasyCap recording cap (electrode locations were Fp1, Fp2, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, O2, and FCz [ground]). Four electrodes (2 on the orbital ridge above and below the right eye and 2 next to the lateral canthi of both eyes) recorded the vertical- and horizontal Electro-oculogram (EOG). Two additional electrodes were placed on the mastoid bones, of which the left was used to reference the signal on-line. Electrode impedance was adjusted to $<5\,k\Omega$ (scalp electrodes) and $<10\,k\Omega$ (EOG electrodes). Using Brain Vision Analyzer 2.0, the signal was re-referenced off-line to an average of the two mastoid electrodes and high-pass filtered (Butterworth zero phase filter at 24 dB/octave) at 0.5 Hz. Next, segments in the continuous EEG that contained course artifacts such as EMG bursts or glitches (defined as segments $+/-100$ ms around amplitude changes $>60\,\mu V/ms$) were identified, and the signal was decomposed into independent components (i.e., ICA, e.g., ref. 34). The ICA decomposition (restricted infomax) was based on the entire data-set (not including the previously identified artifacts) and components that captured blinks or horizontal eye-movements (identified through visual inspection based on components' energy and topography) were removed (the mean number of removed components was 3.32). Next, the data were low-pass filtered at 35 Hz (Butterworth zero phase filter at 24 dB/octave) and an additional 50 Hz notch filter was applied to remove residual electrical interference. The data were segmented into 1200 ms epochs (including 200 ms before, and 1000 ms after stimulus onset), and epochs that contained voltage steps $>50\,\mu V/ms$, had a voltage difference $>100\,\mu V/1000$ ms, minima/maxima $<-100/>100\,\mu V$, and/or activity $<0.5\,\mu V$ were rejected. Six participants with a substantial artifact rate (mean = 36%) were excluded from analyses. For the remaining participants, averaged artifact rate was less than 15%. Epochs were base line corrected using the 200 ms of data before stimulus onset, averaged per condition, and the resulting ERPs were exported for statistical analyses.

References

1. Ernst, M. O. & Bülthoff, H. H. Merging the senses into a robust percept. *Trends Cogn. Sci.* **8(4),** 162–169 (2004).
2. Spence, C. Crossmodal correspondences: A tutorial review. *Atten. Percept. Psycho.* **73(4),** 971–995 (2011).
3. Repp, B. H., Frost, R. & Zsiga, E. Lexical mediation between sight and sound in speechreading. *Q. J. Exp. Psychol.* **45(1),** 1–20 (1992).
4. Sumby, W. H. & Pollack, I. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* **26,** 212–215 (1954).
5. Alsius, A., Wayne, R. V., Paré, M. & Munhall, K. G. High visual resolution matters in audiovisual speech perception, but only for some. *Atten. Percept. Psycho.* **78,** 1–16 (2016).
6. Frost, R. & Katz, L. Orthographic depth and the interaction of visual and auditory processing in word recognition. *Mem. Cognition* **17(3),** 302–310 (1989).
7. McGurk, H. & MacDonald, J. Hearing lips and seeing voices. *Nature* **264,** 746–748 (1976).
8. Frost, R., Repp, B. H. & Katz, L. Can speech perception be influenced by simultaneous presentation of print? *J. Mem. Lang.* **27(6),** 741–755 (1988).
9. Garrett, H. E. 1922 A study of the relation of accuracy to speed. *Arch. Psychol.* **56,** l–104 (1922).

10. Fowler, C. A. & Dekle, D. J. Listening with eye and hand: cross-modal contributions to speech perception. *J. Exp. Psychol. Human.* **17(3),** 816–828 (1991).
11. Borowsky, R., Owen, W. J. & Fonos, N. Reading speech and hearing print: Constraining models of visual word recognition by exploring connections with speech perception. *Can J. Exp. Psychol.* **53(4),** 294–305 (1999).
12. Owen, W. J. & Borowsky, R. Examining the interactivity of lexical orthographic and phonological processing. *Can. J. Exp. Psychol.* **57(4),** 290–303 (2003).
13. MacMillan, N. A. & Creelman, C. D. Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychol. Bull.* **107,** 401–413 (1990).
14. Stanislaw, H. & Todorov, N. Calculation of signal detection theory measures. *Behav. Res. Meth. C.* **31,** 137–149 (1999).
15. Blau, V., van Atteveldt, N. M., Formisano, E., Goebel, R. & Blomert, L. Task-irrelevant visual letters interact with the processing of speech sounds in heteromodal and unimodal cortex. *Eur. J. Neurosci.* **28(3),** 500–509 (2008).
16. van Atteveldt, N. M., Formisano, E., Goebel, R. & Blomert, L. Integration of letters and speech sounds in the human brain. *Neuron* **43(2),** 271–282 (2004).
17. Froyen, D. J., Bonte, M. L., van Atteveldt, N. M. & Blomert, L. The long road to automation: neurocognitive development of letter-speech sound processing. *J. Cognitive Neurosci.* **21(3),** 567–580 (2009).
18. Froyen, D. J., van Atteveldt, N. M., Bonte, M. & Blomert, L. Cross-modal enhancement of the MMN to speech-sounds indicates early and automatic integration of letters and speech-sounds. *Neurosci. Lett.* **430(1),** 23–28 (2008).
19. Armstrong, B. C. & Plaut, D. C. Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Lang. Cogn. Neurosci.* **31(7),** 1–27 (2016).
20. Bertelson, P., Vroomen, J. & De Gelder, B. Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychol. Sci.* **14(6),** 592–597 (2003).
21. Baart, M. & Vroomen, J. Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neurosci. Lett.* **471(2),** 100–103 (2010).
22. Keetels, M., Schakel, L., Bonte, M. & Vroomen, J. Phonetic recalibration of speech by text. *Atten. Percept. Psycho.* **78(3),** 938–945 (2016).
23. MacGregor, L. J., Pulvermüller, F., van Casteren, M. & Shtyrov, Y. Ultra-rapid access to words in the brain. *Nat. Commun.* **3,** 711 (2012).
24. van den Brink, D., Brown, C. M. & Hagoort, P. Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *J. Cognit. Neurosci.* **13,** 967–985 (2001).
25. Baart, M. & Samuel, A. G. Early processing of auditory lexical predictions revealed by ERPs. *Neurosci. Lett.* **585,** 98–102 (2015).
26. Baart, M. & Samuel, A. G. Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. *J. Mem. Lang.* **85,** 42–59 (2015).
27. Pettigrew, C. M., Murdoch, B. E., Ponton, C. W., Finnigan, S., Alku, P., Kei, J., Sockalingam, R. & Chenery, H. J. Automatic auditory processing of English words as indexed by the mismatch negativity, using a multiple deviant paradigm. *Ear. Hear.* **25,** 284–301 (2004).
28. van Linden, S., Stekelenburg, J. J., Tuomainen, J. & Vroomen, J. Lexical effects on auditory speech perception: an electrophysiological study. *Neurosci. Lett.* **420,** 49–52 (2007).
29. Patterson, M. L. & Werker, J. F. Two-month-old infants match phonetic information in lips and voice. *Developmental Sci.* **6(2),** 191–196 (2003).
30. Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A. & Carreiras, M. EsPal: One-stop shopping for Spanish word properties. *Behav. Res. Methods* **45(4),** 1246–1258 (2013).
31. Armstrong, B. C., Watson, C. E. & Plaut, D. C. SOS! An algorithm and software for the stochastic optimization of stimuli. *Behav. Res. Methods* **44(3),** 675–705 (2012).
32. Frost, R., Feldman, L. B. & Katz, L. Phonological ambiguity and lexical ambiguity: Effects on visual and auditory word recognition. *J. Exp. Psychol. Learn.* **16(4),** 569–580 (1990).
33. Peirce, J. W. PsychoPy - Psychophysics software in Python. *J. Neurosci. Meth.* **162,** 8–13 (2007).
34. Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J. *et al.* Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* **37,** 163–178 (2000).

Acknowledgements

This work was supported by the Severo Ochoa program grant SEV-2015-049 awarded to the BCBL, and by grant 613465-AThEME from the FP7/2007-2013 Cooperation grant agreement. MB was supported by grants FPDI-2013-15661 and PSI2014-51874-P from the Spanish Ministry of Economy and Competitiveness (MINECO) and VENI grant 275-89-027 from the Netherlands Organization for Scientific Research (NWO). BCA was supported by the Marie Curie International Incoming Fellowship (IIF) PIIF-GA-2013-689 627784. CDM was supported by MINECO Grant PSI2014-54500 and grant PI_2015_1_25 from the Basque Government. RF was supported by grant ERC-ADG-692502 from the European Research Council. MC was supported by MINECO grant PSI2015-67353-R, and grant ERC-2011-ADG-295362 from the European Research Council. The authors would like to thank Arthur Samuel for sharing his software for adding noise to an auditory signal.

Author Contributions

B.C.A., C.D.M., R.F. and M.C. designed Experiment 1. M.B., B.C.A. and C.D.M. designed Experiment 2. B.C.A. programmed the experiments. M.B. and B.C.A. analyzed all data. M.B. wrote the initial draft, and the final manuscript was reviewed, adapted and approved by all authors.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Baart, M. *et al.* Cross-modal noise compensation in audiovisual words. *Sci. Rep.* **7**, 42055; doi: 10.1038/srep42055 (2017).