

1 Word-Referent Identification Under Multimodal Uncertainty

2 Abdellah Fourtassi¹ & Michael C. Frank¹

3 ¹ Department of Psychology, Stanford University

4 Author Note

5 Abdellah Fourtassi

6 Department of Psychology

7 Stanford University

8 50 Serra Mall

9 Jordan Hall, Building 420

10 Stanford, CA 94301

11 Correspondence concerning this article should be addressed to Abdellah Fourtassi,

12 Postal address. E-mail: afourtas@stanford.edu

Abstract

Identifying a spoken word in a referential context requires both the ability to process and integrate multimodal input and the ability to reason under uncertainty. How do these tasks interact with one another? We introduce a task that allows us to examine how adults identify words under joint uncertainty in the auditory and visual modalities. We propose an ideal observer model of the task which provides an account of how auditory and visual cues are combined optimally. Model predictions are tested in three experiments where word recognition is made under two kinds of uncertainty: category ambiguity and/or distorting noise. In all cases, the optimal model explains much of the variance in human mean judgments. In particular, when the signal is not distorted with noise, participants weight the auditory and visual cues optimally, that is, according to the relative reliability of each modality. But when one modality has noise added to it, human perceivers systematically prefer the unperturbed modality to a greater extent than the optimal model does. The study provides a formal framework which helps us understand precisely how word form and word meaning interact in word recognition under uncertainty. Moreover it offers a first step towards a model that accounts for form-meaning synergy in early word learning.

Keywords: Language understanding; audio-visual processing; word learning; speech perception; computational modeling.

Word count: X

Word-Referent Identification Under Multimodal Uncertainty

Language uses symbols expressed in one modality (e.g., the auditory modality, in the case of speech) to communicate about the world, which we perceive through many different sensory modalities. Consider hearing someone yell “bee!” at a picnic, as a honey bee buzzes around the food. Identifying a word involves processing the auditory information as well as other perceptual signals (e.g., the visual image of the bee, the sound of its wings, the sensation of the bee flying by your arm). A word is successfully identified when information from these modalities provide convergent evidence. However, word identification takes place in a noisy world, and the cues received through each modality may not provide a definitive answer. On the auditory side, individual acoustic word tokens are almost always ambiguous with respect to the particular sequence of phonemes they represent, which is due to the inherent variability of how a phonetic category is realized acoustically (Hillenbrand, Getty, Clark, & Wheeler, 1995). And some tokens may be distorted additionally by mispronunciation or ambient noise. Perhaps the speaker was yelling “pea” and not “bee”. Similarly, a sensory impression may not be enough to make a definitive identification of a visual category.¹ Perhaps the insect was a beetle or a fly instead.

How does the mind deal with uncertainty to identify the speaker’s intended word? To address this question, many researchers in recent years have adopted the rational approach to cognition (Anderson, 1990), which corresponds to Marr’s computational level of analysis. This approach has had a tremendous impact not only on speech related research (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, Griffiths, & Morgan, 2009; Kleinschmidt & Jaeger, 2015; Norris & McQueen, 2008), but also on many other disciplines in the cognitive sciences (for reviews, see Chater & Manning, 2006; Knill & Pouget, 2004; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Within this framework, researchers typically consider an “ideal observer” model which provides a characterization of the task/goal and shows what the

¹In the general case, language can of course be visual as well as auditory, and object identification can be done through many modalities. For simplicity, we focus on audio-visual matching here.

optimal performance should be under this characterization. When there is uncertainty in the input, the ideal observer performs an optimal probabilistic inference. For example, in order to recognize an ambiguous linguistic input, the model uses all available probabilistic knowledge in order to maximize the accuracy of this recognition. Thus, when the task is well specified, the ideal observer model can be seen as a theoretical upper limit on performance. It is often used, not so much as a realistic model of human performance, as much as a baseline against which human performance can be compared (Geisler, 2003). When there is a deviation from the ideal, this can reveal extra constraints on human cognition, such as limitations on working memory or attentional resources.

In this line of research, Clayards et al. (2008) simulated auditory uncertainty by manipulating the probability distribution of a cue (VOT) that differentiated similar words (e.g., “beach” and “peach”). They found that humans were sensitive to these probabilistic cues and their judgments closely reflected the optimal predictions. In another work, Feldman et al. (2009) studied the perceptual magnet effect, which is related to how categories influence the way the stimuli are perceived. They showed that this effect can be explained as the consequence of optimally solving the problem of perception under uncertainty. Both studies explored optimal performance under uncertainty from one modality, i.e., the auditory modality. There is however extensive evidence that information from the visual modality, in particular the speaker’s facial features, influences the perception of auditory information (see Campbell, 2008 for a review). Bejjanki, Clayards, Knill, and Aslin (2011) offered a mathematical characterization of how cues from speech and lip movements can be optimally combined and showed that human performance during audio-visual phonemic labeling was consistent (at least at the qualitative level) with the behavior of the ideal observer.

The speaker’s facial features are not the only visual information that can be used to disambiguate the auditory input, however. For instance, in a referential context, information about the identity of the semantic referent can be integrated with linguistic information to resolve lexical and syntactic ambiguities in speech (e.g., Eberhard, Spivey-Knowlton, Sedivy,

84 & Tanenhaus, 1995; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus,
85 Spivey-Knowlton, Eberhard, & Sedivy, 1995). To our knowledge, however, no study offered
86 an ideal observer analysis of word identification in such context, that is, when the mind has
87 to combine cues from the sound and the referent. Imagine, for example, that someone is
88 uncertain whether they heard “pea” or “bee”, does this uncertainty make them rely more on
89 the referent (e.g., the object being pointed at)? Vice versa, if they are not sure if they saw a
90 bee or a fly, does it make them rely more on the sound? More importantly, when input in
91 both modalities is uncertain to varying degrees, do they weight each modality according to
92 its relative reliability (which is the optimal strategy), or do they over-rely on a particular
93 modality (which is a sub-optimal strategy)?

94 The visual information provided by the referent is arguably different from the visual
95 information provided by the speaker’s facial features (such as lip movements). There seems
96 to be at least two fundamental differences between these two cases, and both can influence
97 the cue combination strategy:

98 First, in the case of audio-visual speech, modalities are understood to differ only in
99 terms of their reliability. In a referential context, however, modalities also differ in terms of
100 their roles in the referential process: the auditory input represents the symbol (i.e., the word)
101 whereas the visual input represents the referent. It has been suggested that because of its
102 referential property, speech is a privileged signal for humans, starting in infancy (see
103 Vouloumanos & Waxman, 2014 for a review).² Thus, in a referential context, it is possible
104 that we do not treat the auditory and visual modalities as equivalent sources of information.
105 Instead, there could be a bias for the auditory modality beyond what is expected from
106 informational reliability alone.

²There is a debate over whether speech is privileged for infant and adults for the same reasons. Whereas Waxman and colleagues suggest that speech is privileged for both infants and adults because of its ability to refer, Sloutsky and colleagues suggest that speech might not have a referential status from the start. Rather, according to Sloutsky and colleagues, speech seems to be preferred by infants only because of a low level auditory "overshadowing".

Second, in the case of audio-visual speech, the associations are expected to be perceptually correlated. The expectation for this correlation is such that when there is a mismatch between auditory and visual input, people still integrate them into a unified (but illusory) percept (McGurk & MacDonald, 1976). In the case of referential language, however, the multimodal association is by nature *arbitrary* (Greenberg, 1957; Saussure, 1916). For instance, there is no logical/perceptual connection between the sound “bee” and the corresponding insect. Moreover, variation in the way the sound “bee” is pronounced is generally not expected to correlate perceptually with variation in the shape (or any other visual property) in the category of bees. Thus, one could imagine that cue combination in the case of arbitrary audio-visual associations (word-referent) is less automatic, more effortful, and therefore less conducive to optimal integration than it is in the case of perceptually correlated associations (as in audio-visual speech perception).

In the current study, we investigate how people combine cues from the auditory and the visual modality to recognize words in a referential context. In particular, we study how this combination is performed under various degrees of uncertainty in both the auditory and the visual modality. We perform a rational analysis of the task. First we propose an ideal observer model that performs the combination in an optimal fashion. Second we compare the predictions of the optimal model to human responses. Humans can deviate from the ideal for several reasons. For instance, as mentioned above, a sub-optimality can be induced by the suggested privileged status of speech or by the arbitrariness of the referential association. In order to study possible patterns of sub-optimality, we compare the optimal model (which provides a normative benchmark) to a descriptive model which is identical to the optimal model, except that the parameters in the former are fitted to human responses.

We tested the ideal observer model’s predictions in three behavioral experiments where we varied the source of uncertainty. In Experiment 1, audio-visual tokens were ambiguous with respect to their category membership only. In Experiment 2, we intervened by adding background noise to the auditory modality, and in Experiment 3, we intervened by adding

background noise to the visual modality. In all experiments, participants were quantitatively near-optimal, though overall response precision was slightly lower than expected. Moreover, in Experiment 1 where neither of the modalities was perturbed with background noise, participants weighed auditory and visual cues according to the relative reliability predicted by the optimal model. In other words, we found no evidence for a modality bias towards either the auditory or the visual modality. However, in Experiment 2 and 3, participants over-relied on one modality when the other modality was perturbed with additional noise.

Models and predictions

In this section we describe the multimodal cue combination model. First, we briefly introduce the experimental paradigm, and second we explain how behavior in this paradigm can be characterized optimally with an ideal observer model.

The Audio-Visual Word Recognition Task

We introduce a new task that tests word recognition in a referential context. We use two visual categories (cat and dog) and two auditory categories (/b/ and /d/ embedded in the minimal pair /aba/-/ada/). For each participant, an arbitrary pairing is set between the auditory and the visual categories, leading to two audio-visual word categories (e.g., dog-/aba/, cat-/ada/). In each trial, participants are presented with an audio-visual target (the prototype of the target category), immediately followed by an audio-visual test stimulus (Figure 1). The test stimulus may differ from the target in both the auditory and the visual components. After these two presentations, participants press “same” or “different”.

This task is similar to the task introduced by Sloutsky and Napolitano (2003), and used in subsequent studies in both children and adults (see Robinson & Sloutsky, 2010 for a review). However, in this original task, participants are asked whether or not the two audio-visual presentations are *identical*. Such task would allow us to probe audio-visual encoding, but not necessarily language understanding. The latter requires—in addition to the perceptual encoding—a categorical judgement, i.e., determining whether or not two

similar tokens are members of the same phonological/semantic category. Thus, our task is category-based: Participants are asked to press “same” if they think the second item (the test) belonged to the same category as the first (target) (e.g., dog-/aba/), even if there is a slight difference in the sound, in the referent, or in both. They are instructed to press “different” only if they think that the second stimulus was an instance of the other word category (cat-/ada/). The task also includes trials where pictures are hidden (audio-only) or where sounds are muted (visual-only). These unimodal trials provide us with the participants’ evaluation of the probabilistic information present in the auditory and visual categories. This unimodal distributions are used as inputs to the optimal combination model, described below.

Optimal Model

The ideal observer model combines probabilistic information from the auditory and visual modalities, and makes a categorical decision. How can this task be characterized in a formal way, and what would be the optimal behavior? If the stimuli were continuous (e.g., size, location), cues derived from each modality would be characterized with their sensory noise variance only, and (assuming conditional independence) their combined information would be characterized with a simple linear combination rule where the combined estimate is an average of the individual estimates, each weighted according to its relative reliability (Alais & Burr, 2004; Ernst & Banks, 2002). However, in our case decision is made over categorical variables (phonemes and concepts), the optimal model should take into account, not only the noise variability around an individual perceptual estimate, but also its categorical variability (i.e., the probability with which the estimate belongs to a given category).

One way to account for the categorical variability is to model the decision step which maps sensory input onto task-dependent categories (Bankieris, Bejjanki, & Aslin, 2017; Bejjanki et al., 2011). We represent the decision step as follows (see Figure 2 for illustration).

If we assume that both the visual and the auditory cues vary along a single dimension, an audio-visual signal (a, v) can be represented as a point in this two-dimensional space, and a word category can be characterized with a bivariate probability distribution over audio-visual tokens.³ Each category is peaked around the most typical audio-visual token, and the probability of membership in the category (the likelihood probability) decreases gradually as the auditory token deviates from the typical sound instance (depending on the variance of the auditory category) and/or as the visual token deviates from the typical visual instance (depending on the variance of the visual category).

A crucial feature of the ideal observer model is that it makes use of the entire probability distribution of a given category in order to make optimal decisions (e.g., Clayards et al., 2008). This means that the probability of picking a word (the posterior probability) should mirror closely the probability of the combined information (the likelihood probability). For instance, the model becomes less ambiguous and more certain as the tokens get closer to the typical instance of the category. In what follows, we formalize this sense of optimality using a Bayesian framework.

An auditory category A and a visual category V are defined as distributions in the auditory and the visual dimensions, respectively. A word category W is defined as a bivariate distribution in the audio-visual. More specifically, we define it as the joint distribution of the auditory and the visual categories. For simplicity, we assume both auditory and visual categories to be normally distributed. That is, if $w = (a, v)$ is an audio-visual token, then:

$$p(a|A) \sim N(\mu_A, \sigma_A^2)$$

where μ_A and σ_A^2 are respectively the mean and the variance of the auditory category,

$$p(v|V) \sim N(\mu_V, \sigma_V^2)$$

³For the sake of clarity, we assume here that the sensory noise around this point is small/negligible compared to the categorical variability. But see below how the noise can be integrated when its effect is relatively large

207 where μ_V and σ_V^2 are the mean and the variance of the visual category, and

$$p(w|W) \sim N(M_W, \Sigma_W)$$

208 where $M_W = (\mu_A, \mu_V)$ and Σ_W are mean and the covariance matrix of the word category.

209 We assume that auditory and visual variables are independent (i.e., uncorrelated), so the

210 covariance matrix is simply:

$$\Sigma_W = \begin{bmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_V^2 \end{bmatrix}$$

211 This assumption simply says that, given a word-object mapping, e.g., $W=(\text{“cat”}-\text{CAT})$,
 212 variation in the way “cat” is pronounced does not correlate with changes in any visual
 213 property of the object CAT, which is a valid assumption.⁴

214 Now we turn to the crucial question of modeling how optimal decision should proceed
 215 based on the probabilistic information in the auditory and the visual modalities. In our task,
 216 we have two word categories: dog-/aba/ (W_1) and cat-/ada/ (W_2).⁵ Participants can be
 217 understood as choosing one of these two word categories (Figure 2). For an ideal observer,
 218 the probability of choosing category 2 when presented with an audio-visual instance
 219 $w = (a, v)$ is the posterior probability of this category:

$$p(W_2|w) = \frac{p(w|W_2)p(W_2)}{p(w|W_2)p(W_2) + p(w|W_1)p(W_1)}$$

220 Using our assumption that the cues are uncorrelated, we have

$$p(w|W) = p(a, v|W) = p(a|A)p(v|V)$$

221 Under this assumption, the posterior probability reduces to:⁶

⁴Note that this assumptions is more adequate in the case of arbitrary associations such as ours, and less so in the case of redundant association such as audio-visual speech. In the latter, variation in the pronunciation is expected to correlate, at least to some extent, with lip movements.

⁵This mapping is randomized in the experiments.

⁶See the appendix for the details of the derivation

$$p(W_2|w) = \frac{1}{1 + (1 + \epsilon) \exp(\beta_0 + \beta_a a + \beta_v v)} \quad (1)$$

where

$$1 + \epsilon = \frac{p(W_1)}{p(W_2)}$$

$$\beta_0 = \frac{\mu_{A2}^2 - \mu_{A1}^2}{2\sigma_A^2} + \frac{\mu_{V2}^2 - \mu_{V1}^2}{2\sigma_V^2}$$

$$\beta_a = \frac{\mu_{A1} - \mu_{A2}}{\sigma_A^2}$$

$$\beta_v = \frac{\mu_{V1} - \mu_{V2}}{\sigma_V^2}$$

The parameter ϵ represents the differential between the categories' prior probabilities. However, since the identity of word categories is randomized across participants, ϵ measures, rather, a response bias to “same” if $\epsilon > 0$, and a response bias to “different” if $\epsilon < 0$. We expect a general bias towards answering “different” because of the categorical nature of our same-different task: When two items are ambiguous but perceptually different, participants might have a slight preference for “different” over “same”.

The values of the means are fixed: They correspond to the value of the most typical tokens in our stimuli. Thus, observations from each modality (a and v) are weighted according to their reliability in the posterior:

$$\beta_a \propto \frac{1}{\sigma_A^2}$$

$$\beta_v \propto \frac{1}{\sigma_V^2}$$

Note that the weights account for categorical variability only. For instance, if the speaker generates a target production a_t from an auditory category $p(a_t|A) \sim N(\mu_A, \sigma_A^2)$, the ideal model assumes that it has direct access to this production token (i.e., $a = a_t$), and that all uncertainty is about the category membership of this token. However, we might also

want to account for noise in the brain and/or in the environment. For example, the observer might not have access to the exact produced target, but only to the target perturbed by noise. If we assume this noise to be normally distributed, that is, $p(a|a_t) \sim N(a_t, \sigma_N^2)$, then integrating over a_t lead to the following simple expression:

$$p(a|A) \sim N(\mu_A, \sigma_A^2 + \sigma_N^2)$$

Using an identical reasoning for the case of the visual modality, we end up with the following multimodal weighting scheme in the optimal combination model:

$$\beta_a \propto \frac{1}{\sigma_A^2 + \sigma_{N_A}^2}$$

$$\beta_v \propto \frac{1}{\sigma_V^2 + \sigma_{N_V}^2}$$

To summarize, the posterior provides the optimal model’s predictions for how probabilities that characterize uncertainty in each modality can be combined to make categorical decision about the bimodal input. The model will be constructed based on data from the unimodal trials, but the predictions of the optimal model will be compared to human responses in the bimodal trials.

Auditory and Visual models

Besides the optimal model, we also test the predictions of two baselines models: a visual model, which assumes that participants rely only on visual information, and an auditory model, which assumes that participants rely only on auditory information. If participants rely on both the auditory and the visual modalities, the optimal model would explain more variance in human responses than the visual model alone or the auditory model alone.

Descriptive model

The visual, auditory, and optimal models are *normative* models. Their predictions are made about human data in the bimodal condition, but their crucial parameters (i.e.,

variances associated with the visual and auditory modalities) are derived from data in the unimodal conditions. In addition to these normative models, we consider a *descriptive* model. The parameters of this model are fit to human data in the bimodal condition. If the referential task induces sub-optimality (due, for instance, to the arbitrary nature of the sound-object association), then we predict that the descriptive model should explain more variance than the optimal model does.

A systematic comparison of the optimal and the descriptive models would allow us, not only to quantify how much people deviate from optimality, but also to understand precisely how they deviate from this optimality. Let σ_A^2 and σ_V^2 be the values of the variances used in the optimal model (derived from the unimodal conditions), and σ_{Ab}^2 and σ_{Vb}^2 be the values observed through the descriptive model in the bimodal condition. Deviation from optimality is measured in two ways. First, we measure the change in the values of the variances specific to each modality, that is, how σ_A^2 compares to σ_{Ab}^2 , and how σ_V^2 compares to σ_{Vb}^2 . Second, we measure changes in the proportion of the visual and auditory variances, i.e., how $\frac{\sigma_A^2}{\sigma_V^2}$ compares to $\frac{\sigma_{Ab}^2}{\sigma_{Vb}^2}$. The first measure allows us to test if precision changes for each modality when we move from the unimodal to the bimodal conditions. The second allows us to test whether or not the weighting scheme follows the prediction of the optimal model. The reason we used the proportion of the variances as a measure of cross-modal weighing is because it corresponds to the slope⁷ of the decision threshold in the audio-visual space (Figure 2). The decision threshold is defined as the set of values in this audio-visual space along which the posterior is equal to 0.5. Formally speaking, the decision threshold takes the following form:

$$v = -\frac{\sigma_V^2}{\sigma_A^2}a + v_0$$

If the absolute value of the slope derived from the descriptive model is greater than that of the optimal model, the corresponding shift in the decision threshold indicates that participants have a preference for the auditory modality in the bimodal case. Similarly, a

⁷Or more precisely the absolute value of the slope

smaller absolute value of the slope would lead to a preference for the visual modality. The limit cases are when there is exclusive reliance on the auditory cue (a vertical line), and where there is exclusive reliance on the visual (a horizontal line).

It follows that there are three possible scenarios describing how humans can deviate from optimality. These scenarios are illustrated in Figure 3 and are as follows:

- 1) Both variances may increase, but their proportion remains the same. That is, $\sigma_{Ab}^2 \geq \sigma_A^2$ and $\sigma_{Vb}^2 \geq \sigma_V^2$, but $\frac{\sigma_{Ab}^2}{\sigma_{Vb}^2} \approx \frac{\sigma_A^2}{\sigma_V^2}$. In this case, sub-optimality would be due to increased randomness in human responses in the bimodal condition. However, this randomness would not affect the relative weighting of both modalities, i.e., participants would still weigh modalities according to the relative reliability predicted by the optimal model.
- 2) The auditory variance increases at a higher rate. That is, $\sigma_{Ab}^2 \gg \sigma_A^2$ and $\sigma_{Vb}^2 \geq \sigma_V^2$, leading to $\frac{\sigma_{Ab}^2}{\sigma_{Vb}^2} > \frac{\sigma_A^2}{\sigma_V^2}$. In this case, sub-optimality would consist not only in participants being more random in the bimodal condition, but also in having a systematic preference for the visual modality, even after accounting for informational reliability.
- 3) The visual variance increases at a higher rate. That is, $\sigma_{Vb}^2 \gg \sigma_V^2$, and $\sigma_{Ab}^2 \geq \sigma_A^2$, leading to $\frac{\sigma_{Ab}^2}{\sigma_{Vb}^2} < \frac{\sigma_A^2}{\sigma_V^2}$. This case is the reverse of case 2, i.e., in addition to increased randomness in the bimodal condition, there is a systematic preference for the auditory modality, even after accounting for informational reliability.

We tested the predictions of the optimal model in three experiments. In Experiment 1, we studied the case where bimodal uncertainty was due to ambiguity in terms of category membership, without any additional background noise. In Experiment 2 and 3 we added background noise on top of ambiguity in category membership.

Experiment 1

In this Experiment, we start with testing the predictions in the case where uncertainty is due to ambiguity in terms of category membership only, that is, according to the following

weighting scheme:

$$\beta_a \propto \frac{1}{\sigma_A^2}$$

$$\beta_v \propto \frac{1}{\sigma_V^2}$$

We do not introduce any additional noise in the background, and we assume that sensory noise is negligible compared to categorical variability.

Methods

Participants: We recruited a planned sample of 100 participants from Amazon Mechanical Turk. Only participants with US IP addresses and a task approval rate above 85% were allowed to participate. They were paid at an hourly rate of \$6/hour. Participants were excluded if they reported having experienced a technical problem of any sort during the online experiment (N=14), or if they had less than 50% accurate responses on the unambiguous training trials (N=6). The final sample consisted of (N = 80) participants.⁸

Stimuli: For auditory stimuli, we used the continuum introduced in Vroomen, Linden, Keetels, Gelder, and Bertelson (2004), a 9-point /aba/-/ada/ speech continuum created by varying the frequency of the second (F2) formant in equal steps. We selected 5 equally spaced points from the original continuum by keeping the endpoints (prototypes) 1 and 9, as well as points 3, 5, and 7 along the continuum. For visual stimuli, we used a morph continuum introduced in Freedman, Riesenhuber, Poggio, and Miller (2001). From the original 14 points, we selected 5 points as follows: we kept the item that seemed most ambiguous (point 8), the 2 preceding points (i.e., 7 and 6) and the 2 following points (i.e., 9 and 10). The 6 and 10 points along the morph were quite distinguishable, and we took them to be our prototypes.

Design and Procedure. We told participants that an alien was naming two objects: a dog, called “aba” in the alien language, and a cat, called “ada”. In each trial, we presented the first object (the target) on the left side of the screen simultaneously with the

⁸The sample size and exclusion criteria were specified in the pre-registration at <https://osf.io/h7mzp/>.

corresponding sound. For each participant, the target was always the same (e.g., dog-/aba/). The second sound-object pair (the test) followed on the other side of the screen after 500ms and varied in its category membership. For both the target and the test, visual stimuli were present for the duration of the sound clip (~ 800 ms). We instructed participants to press “S” for same if they thought the alien was naming another dog-/aba/, and “D” for different if they thought the alien was naming a cat-/ada/. We randomized the sound-object mapping (e.g., dog-/aba/, cat-/ada/) as well as the identity of the target (dog or cat) across participants.

The first part of the experiment trained participants using only the prototype pictures and the prototype sounds (12 trials, 4 each from the bimodal, audio-only, and visual-only conditions). After completing training, we instructed participants on the structure of the task and encouraged them to base their answers on both the sounds and the pictures (in the bimodal condition). There were a total of 25 possible combinations in the bimodal condition, and 5 in each of the unimodal conditions. Each participant saw each possible trial twice, for a total of 70 trials/participant. Trials were blocked by condition and blocks were presented in random order.

Model fitting details.

Unimodal condition. Remember that data in this conditions allow us to derive the variances of both the auditory and the visual categories, and that these variances are used to make predictions about bimodal data (in the visual and auditory baselines as well as in the optimal model). These individual variances were derived as follows (we explain the derivation for the auditory-only case, but the same applies for the visual-only case). We use the same Bayesian reasoning as we did in the derivation of the bimodal model: When presented with an audio instance a , the probability of choosing the sound category 2 (that is, to answer “different”) is the posterior probability of this category $p(A_2|a)$. If we assume that both sound categories have equal variances, the posterior probability reduces to:

$$p(A_2|a) = \frac{1}{1 + (1 + \epsilon_A) \exp(\beta_{a0} + \beta_a a)}$$

with $\beta_a = \frac{\mu_{A1} - \mu_{A2}}{\sigma_A^2}$ and $\beta_{a0} = \frac{\mu_{A2}^2 - \mu_{A1}^2}{2\sigma_A^2}$. ϵ_A is the response bias in the auditory-only condition. For this model (as well as all other models in this study), we fixed the values of the means to be the end-points of the corresponding continuum, since these points are the most typical instances in our stimuli. Thus, we have $\mu_{A1} = 0$ and $\mu_{A2} = 4$ (and similarly $\mu_{V1} = 0$, and $\mu_{V2} = 4$). This leaves us with two free parameters: the bias ϵ_A and the variance σ_A^2 . To determine the values of these parameters, we fit the unimodal posterior to human data in the unimodal case.

Bimodal condition. In this condition, only the descriptive model is fit to the data, using the expression of the posterior (Equation 1). Since the values of the means are fixed, we have 3 free parameters: the variances for the visual and the auditory modalities, respectively, and ϵ , the response bias. The visual and auditory baselines as well as the optimal model are not fit to bimodal data, but their predictions are tested against these bimodal data. All these normative models use the variances derived from the unimodal data and the bias term derived from the bimodal data.

Although the paradigm is within-subjects, we did not have enough statistical power to fit a different model for each individual participant. Instead, models are constructed with data collapsed across all participants. The fit was done with a nonlinear least squares regression using the NLS package in R (Bates & Watts, 1988). We computed the values of the parameters, as well as their 95% confidence intervals, through non-parametric bootstrap (using 10000 iterations).

Results and analysis

Unimodal conditions. Average categorization judgments and best fits are shown in Figure 4. The categorization function of the auditory condition was slightly steeper than that of the visual condition, meaning that participants perceived the sound tokens slightly

more categorically and with higher certainty than they did with the visual tokens. For the auditory modality, we obtained the following values:⁹ $\epsilon_A = -0.20$ [0.02, -0.38] and $\sigma_A^2 = 2.04$ [1.66, 2.53]. For the visual modality, we obtained $\epsilon_V = -0.12$ [0.06, -0.28] and $\sigma_V^2 = 3.33$ [2.83, 3.92].

Bimodal condition.

Normative models. Figure 5 compares the predictions of the normative models against human responses. The visual, auditory and optimal model explained, respectively, 30%, 67%, and 89% of total variance in mean responses.

Descriptive model. In the descriptive model, all parameters are fit to human responses in the bimodal condition. We found $\epsilon = -0.34$ [-0.28, -0.39], $\sigma_{Ab}^2 = 4.96$ [4.58, 5.40] and $\sigma_{Vb}^2 = 7.06$ [6.40, 7.84]. Note that the variance of both the auditory and visual modalities increased compared to the unimodal conditions.

The descriptive model explained 95% of total variance. However, since the descriptive model was fit to the same data, there is a risk that this high correlation is due to overfitting. To examine this possibility, we cross-validated the model using half the responses to predict the other half (averaging across 1000 random partitions). The predictive power of the model remained very high ($r^2=0.92$).

Cue combination and Modality preference. We next analyzed if cue combination was performed in an optimal way, or if there was a systematic preference for one modality when making decisions in the bimodal condition. As explained above, modality preference can be characterized formally as a deviation from the decision threshold predicted by the optimal model. Figure 7 (top) shows both the decision threshold derived from the descriptive model (in black) and the decision threshold predicted by the optimal model (in red). The deviation from optimality is compared to two hypothetical cases of modality preference (dotted lines). We found that the descriptive and optimal decision thresholds were almost identical. Indeed, non-parametric resampling of the data showed no evidence of

⁹all CIs in the paper are 95% confidence intervals

a deviation from the optimal prediction (Figure 7, bottom).

Discussion

Overall, we found that the optimal model explained much of the variance in the mean judgments, and largely more than what can be explained with the auditory or the visual models alone. Moreover, the high value of the coefficient of determination in the optimal model ($r^2=0.89$) suggests that the population was near-optimal. However, we see in Figure 5 that the mean responses deviated systematically from the optimal prediction in that they were slightly pulled toward chance (i.e., the probability 0.5). This is due to the increase in the value of the variance associated with each modality. Note however that, despite this increase in randomness, our analysis of modality preference showed that the relative values of these variances were not different (Figure 7), meaning that there was no evidence for a modality preference. Thus, 1) There was a simultaneous increase in the values of the auditory and visual variances in the bimodal condition compared to the unimodal condition, meaning that the bimodal input lead to an increase in response randomness, and 2) this increased randomness did not affect the relative weighting of both modalities, i.e., the population was weighting modalities according to the relative reliability predicted by the optimal model. This situation corresponds to the first case of sub-optimally described in Figure 3.

As we noted earlier, the model addresses the question of optimality at the population level. However, it is important to know how individual responses are distributed. In fact, one could think of an extreme case where optimality at the population level would be misleading. Imagine, for instance, that in the bimodal condition half the participants relied exclusively on the visual modality, whereas the other half relied exclusively on the auditory modality. This case could still lead to an aggregate behaviour which appears optimal, but this optimality would be spurious.

To examine this possibility, we consider the distribution of individual cross-modal weighting in the bimodal condition (i.e., $\frac{\sigma_{Vb}^2}{\sigma_{Ab}^2}$). Using a factor of 10 as a cut-off, we found

that 5 participants relied almost exclusively on the visual modality, and 12 relied almost exclusively on the auditory modality. The percentage of both cases was relatively small compared to the total number of participants (21.25%). An additional number of participants ($N=7$) relied on both modalities, but provided noisy responses which lead to negative variances (probably due to mistaking “same” for “different” or vice versa). When these outliers were removed, the distribution had a rather unimodal shape (Figure 6). This finding indicates that the population’s near optimality is not spurious, but based mostly on genuine cue combination at the individual level.

As a second analysis, we asked whether the observed variance in the individual distribution was due to mere sampling errors or whether it corresponded to a real between-subject variability. We simulated individual responses from the posterior distribution whose parameters were fit to the population as a whole (i.e., the descriptive posterior). The resulting distribution is shown in Figure 6. For ease of comparison, the simulated distribution was superimposed to the real distribution. We found that the real distribution was broader than the simulated distribution, indicating real between-subject variation beyond sampling errors. This means that the participants varied in terms of how they weighed modalities: Compared to the predictions of the global descriptive model, some participants relied more on the auditory modality, and some relied more on the visual modality.

In Experiment 1, we tested word recognition when there was multimodal uncertainty in terms of category membership only. In real life, however, tokens can undergo distortions due to noisy factors in the environment (e.g., car noise in the background, blurry vision in a foggy weather,...). In Experiment 2 and 3, we explore this additional level of uncertainty.

Experiment 2

In this Experiment, we explored the effect of added noise on performance. We tested a case where the background noise was added to the auditory modality. We were interested to

know if participants would treat this new source of uncertainty as predicted by the optimal model, that is, according to the following weighting scheme

$$\beta_a \propto \frac{1}{\sigma_A^2 + \sigma_{N_A}^2}$$

$$\beta_v \propto \frac{1}{\sigma_V^2}$$

The alternative hypothesis is that noise in one modality leads to a systematic preference for the non-noisy modality.

Methods

Participants. A sample of 100 participants was recruited online through Amazon Mechanical Turk. We used the same exclusion criteria as in Experiment 1. 7 participants were excluded because they had less than 50% accurate responses on the unambiguous training trials. The final sample consisted of ($N = 93$) participants.

Stimuli and Procedure. We used the same visual stimuli as in Experiment 1. We also used the same auditory stimuli, but we convolved each item with Brown noise of amplitude 1 using the free sound editor Audacity (2.1.2). The average signal-to-noise ratio was - 4.4 dB. The procedure was exactly the same as in the previous experiment, except that the test stimuli (but not the target) were presented with the new noisy auditory stimuli.

Results and analysis

Unimodal conditions. We fit a model for each modality. For the auditory modality, our parameter estimates were $\epsilon_A = -0.18$ [-0.05, -0.30] and $\sigma_A^2 + \sigma_N^2 = 4.70$ [4.03, 5.55]. For the visual modality, we found $\epsilon_V = -0.24$ [-0.10, -0.36] and $\sigma_V^2 = 3.93$ [3.43, 4.55]. Figure 4 shows responses in the unimodal conditions as well as the corresponding best fits. The visual data is a replication of the visual data in Experiment 1. As for the auditory data, in contrast to Experiment 1, responses were flatter, showing more uncertainty.

Bimodal condition.

Normative models. Figure 5 compares the predictions of the visual, auditory and optimal models to human responses. These normative models explained, respectively, 77%, 21%, and 91% of total variance in mean judgements. Note that, in contrast to Experiment 1, the visual model explains more variance than the auditory model.

Descriptive model. We estimated $\epsilon = -0.38$ $[-0.33, -0.42]$, $\sigma_{Ab}^2 + \sigma_{Nb}^2 = 9.84$ $[8.75, 11.27]$, and $\sigma_{Vb}^2 = 5.21$ $[4.84, 5.64]$. The fit explained 0.97% of total variance. Cross-validation using half the responses to predict the other half yielded $r^2 = 0.96$.

Modality preferences. Figure 7 (top) shows that the participants' decision threshold deviated from optimality, and that this deviation was biased towards the visual modality (the non-noisy modality). Indeed non-parametric resampling of the data showed a decrease in the value of the slope in the descriptive model compared to the optimal model (Figure 7, bottom).

Discussion

We found, similar to Experiment 1, that the population was generally near optimal ($r^2 = 0.91$), and that the optimal model explained more variance than the auditory or the visual models alone. We also found a similar discrepancy from the optimal model as precision dropped for both the auditory and the visual modalities. As for the weighting scheme used by participants, contrary to Experiment 1 where modalities were weighted according to their relative reliability, we found in this experiment that the visual modality had a greater weight than what is expected from its relative reliability. This situation corresponds to the second case of sub-optimally described in Figure 3.

We were also interested in whether noise in the auditory modality lead more participants to rely exclusively on the visual modality at the individual level. Using the same cut-off as in Experiment 1 (a factor of 10), the percentage of participants who relied exclusively on either modalities was 34.41%, which is much higher than the percentage obtained in Experiment 1 (21.25%). Moreover, the subset of participants relying exclusively

on the visual modality (compared to those who relied exclusively on the auditory modality) increased from 29.41% in Experiment 1 to 68.75% in Experiment 2, indicating that noise in the auditory modality prompted more participants to rely exclusively and disproportionately on the visual modality.

In Experiment 2, we tested the case of added background noise to the auditory modality. In Experiment 3, we test the case of added noise to the visual modality.

Experiment 3

In this Experiment, we tested the case where the background noise was added to the visual modality. Similar to Experiment 2, we were interested to know if participants would treat this new source of uncertainty as predicted by the optimal model, that is, according to the following weighting scheme

$$\beta_a \propto \frac{1}{\sigma_A^2}$$

$$\beta_v \propto \frac{1}{\sigma_V^2 + \sigma_{N_V}^2}$$

The alternative hypothesis is that, just like noise in the auditory modality lead to a preference for the visual input in Experiment 2, noise in the visual modality would lead to a preference for the auditory input.

Methods

Participants. A planned sample of 100 participants was recruited online through Amazon Mechanical Turk. We used the same exclusion criteria as in both previous experiments. 2 participants were excluded because they reported having a technical problem, and 10 participants were excluded because they had less than 50% accurate responses on the unambiguous training trials. The final sample consisted of ($N = 88$) participants.

Stimuli and Procedure. We used the same auditory stimuli as in Experiment 1. We also used the same visual stimuli, but we blurred the tokens using the free image editor

GIMP (2.8.20). We used a Gaussian blur with a radius¹⁰ of 10 pixels. The experimental procedure was exactly the same as in the previous Experiments.

Results and analysis

Unimodal conditions. For the auditory modality, our parameter estimates were $\epsilon_A = -0.24$ $[-0.04, -0.42]$ and $\sigma_A^2 = 1.94$ $[1.61, 2.33]$. For the visual modality, we found $\epsilon_V = 0.11$ $[0.27, -0.03]$ and $\sigma_V^2 + \sigma_N^2 = 13.00$ $[9.92, 18.94]$. Figure 4 shows responses in the unimodal conditions as well as the corresponding fits. The auditory data is a replication of the auditory data in Experiment 1. As for the visual data, we see that, in contrast to Experiment 1 and 2, responses were flatter, showing much more uncertainty.

Bimodal condition.

Normative models. Figure 5 compares the predictions of the visual, auditory and optimal models to human responses. These normative models explained, respectively, 1%, 98%, and 97% of total variance in mean judgement.

Descriptive model. We estimated $\epsilon = -0.35$ $[-0.29, -0.40]$, $\sigma_{Ab}^2 = 3.00$ $[2.75, 3.25]$, and $\sigma_{Vb}^2 + \sigma_{Nb}^2 = 39.42$ $[25.06, 98.96]$. The fit explained 97% of total variance. Cross-validation using half the responses to predict the other half yielded $r^2 = 0.96$.

Modality preferences. Participants' decision threshold suggested a preference for the auditory modality (the non-noisy modality). Indeed non-parametric resampling of the data showed an increase in the value of the slope in the descriptive model compared to the optimal model (Figure 7).

Discussion

We found that the optimal model accounted for almost all the variance ($r^2 = 0.97$). However, whereas in previous experiments, the optimal model explained more variance than the auditory or the visual models, here the auditory model explained at least as much variance ($r^2 = 0.98$). Thus, though participants were still sensitive to variation in the noisy

¹⁰A features of the filter that affects the intensity of the blur

visual data in the unimodal condition, they tended to ignore this information in the bimodal condition, and relied almost exclusively on the non-noisy auditory modality. The reason why we saw this (floor) effect when we added noise to the visual modality (Experiment 3), and not when we added noise to the auditory modality (Experiment 2), is the fact that our visual stimuli were originally perceived less categorically and with less certainty than the auditory stimuli. This made it more likely for the visual categorization function to become flat and uninformative after a few drops in precision due to noise on the one hand, and to the additional randomness induced by the bimodal presentation on the other hand.

The general finding corresponds to the third case of sub-optimality described in Figure 3. Indeed, precision dropped for both modalities in the bimodal condition compared to the unimodal condition. But the drop was much greater for the visual modality, resulting in a much lower weight assigned to it than what is expected from its optimal relative reliability. Therefore, just like participants over-relied on the visual modality when the auditory modality was noisy (Experiment 2), they also over-relied on the auditory modality when the visual modality was noisy (Experiment 3).

The percentage of participants who relied exclusively on either the visual modality or the auditory modality was 38.64%, which is closer to the percentage of Experiment 2, except that now almost all of them relied on the auditory modality (94.12%).

General Discussion

When identifying a spoken word under uncertainty, one often needs to combine and make the most of the available multimodal cues. While most previous studies focused on the case of audio-visual speech where people combine cues from speech and facial features, this work explored the case of identifying an ambiguous word when the available multimodal cues are speech and the visual referent. More specifically, we conducted an ideal observer analysis of the task whereby a model provided predictions about how information from each modality should be combined in an optimal fashion. The predictions of the model were tested in a

series of three experiments where instances of both the form and the referent were ambiguous with respect to their category membership only (Experiment 1), when instances of the form were perturbed with additional background noise (Experiment 2), and when instances of the referent were perturbed with additional visual noise (Experiment 3).

In all Experiments, we found many patterns of optimal behaviour. Quantitatively speaking, the optimal model accounted, respectively, for 89%, 91%, and 97% of the variance in mean responses. When compared to the predictions of the visual or the auditory models, participants generally relied on both modalities to make their decisions in the bimodal condition. Indeed, in Experiment 1 and 2, the optimal model accounted for more variance in mean responses than the auditory or the visual models did. In Experiment 3, participants appeared to rely on one modality, but this was likely a floor effect, due to the fact that noise made the visual input barely perceivable. In Experiment 1, which did not involve background noise, participants not only relied on both modalities, but generally weighted these modalities according to the prediction of the optimal model, that is, according to their relative reliability. At the individual level, however, we found evidence of a between-subject variation: Some participants relied slightly more on the visual modality, whereas others relied slightly more on the auditory modality.

We documented two major cases of sub-optimality. First, in all Experiments, the variance associated with each modality increased in the bimodal condition compared to the unimodal conditions. This means that participants responded slightly more randomly in the bimodal condition than they did in the unimodal conditions. This finding contrasts with research on multisensory integration where associations tend to lead to a higher precision (e.g., Ernst & Banks, 2002). Nevertheless, there is a crucial difference between these two situations. Research on multisensory integration (of which audio-visual speech is arguably an instance) deals with redundant multimodal cues, and these cues are integrated into a unified percept. In contrast, the word-referent association is usually arbitrary and, in particular, the cues are not expected to be correlated perceptually. Therefore the mind cannot form a

unified percept, rather, it must encode information separately from both modalities and retain this encoding through the decision making process. Retaining two separate cues at the same time instead of forming one unified percept (as in multisensory integration of redundant cues), or instead of retaining only one cue (as in the unimodal case), is likely to place extra-demand on cognitive resources, which, in turn, can cause general performance to drop. Indeed, there is evidence that cognitive load has a detrimental effect on word recognition, which can be due to a reduction in perceptual acuity (Mattys & Wiget, 2011). Similar to our finding, previous experimental studies reported that in both children and adults, identifying new ambiguous form-referent mappings is usually associated with what appears to be a decrease in speech perception acuity (Pajak, Creel, & Levy, 2016; Stager & Werker, 1997). Hofer and Levy (2017) provided a probabilistic model of this phenomenon, which they similarly characterized as an increase in the noise variance of the auditory modality. Our finding replicates this previously documented fact, and additionally suggests that the reduction in perceptual acuity occurs simultaneously in both the auditory and the visual modalities.

The second case of sub-optimality is related to how participants weighed the cues from the visual and the auditory modalities in a noisy context. In contrast to Experiment 1 where the combination was indistinguishable from the optimal prediction, results of Experiment 2 and 3 which both involved background noise in one modality, showed that participants had a systematic preference for the other (non-noisy) modality. From previous empirical studies, we know that when the speech signal is degraded, people tend to compensate by relying more on other sources of information such as the accompanying visual cues (i.e., lip movements) or the semantic/syntactic context (see Mattys, Davis, Bradlow, & Scott, 2012 for a review). However, and generally speaking, these studies do not differentiate between an optimal compensatory strategy (i.e., relying more on the alternative source while using all information still available in the distorted signal), and a sub-optimal strategy (i.e., relying more on the alternative source while ignoring at least some of the information still available

in the distorted signal). The formal approach followed in this paper allowed us to tease apart these two possibilities, and the analysis supports the sub-optimal compensatory strategy: The preference for the non-noisy modality is above and beyond what can be explained by the relative reliability alone, meaning that the participant tend to ignore at least part of the information still available in the noisy modality.

This second case of sub-optimal behavior is possibly related to the fact that language understanding under degraded listening conditions is cognitively more taxing than language understanding under normal conditions (Ronnberg, Rudner, Lunner, & Zekveld, 2010). This fact can lead to a preference/bias towards alternative sources of information when these sources are available. One could explain this phenomenon in terms of the metacognitive experience about the fluency with which information is processed. The perceived perceptual fluency (e.g., the ease with which a stimulus' physical identity can be identified) can affect a wide variety of human judgements (see Schwarz, 2004 for a review). In particular, variables that improve fluency tends to increase liking/preference (Reber, Winkielman, & Schwarz, 1998). In our case, the subjective experience of lower fluency in the noisy modality might cause people to underestimate information that can be extracted from this modality, especially when presented simultaneously with a higher fluency alternative.

In this study, we tested the extent to which people would be optimal in combining uncertain auditory and visual cues to recognize words. This required that we use a case of double ambiguity, that is, a case where both the word forms (“ada”-“aba”) and the referents (cat-dog) were similar and, thus, confusable. One could wonder about the extent to which such case occurs in real languages? Cross-linguistic corpus analyses suggest that the lexical encoding tends towards double ambiguity in many languages (Dautriche, Mahowald, Gibson, & Piantadosi, 2017; Monaghan, Shillcock, Christiansen, & Kirby, 2014; Tamariz, 2008). For instance, Dautriche et al. (2017) analysed 100 languages and found that word that are similar phonologically tend to be similar semantically as well, beyond what could be explained by chance. Crucially, they found this trend in a set of monomorphemic lemmas in

four languages, suggesting that the effect is not driven by morphological regularities. These studies suggest that the case of double uncertainty, though perhaps not pervasive, could be a real issue in language as it increase the probability of confusability. Moreover, it is important to keep in mind that ambiguity in both the form and the referent can be induced by an external noisy context even when these forms and referents are not confusable in normal situations.

Though we only tested adults in this paper, the problem of word recognition under uncertainty, as well as the need to make the most of ambiguous cues, is a particularly pressing issue for children. In fact, whereas adults are mostly faced with uncertainty in the *input*, children have to deal with the additional uncertainty that results from their early unrefined *representations* of both phonological and semantic categories. For example, upon hearing a noisy instance of “bee”, adults may have to decide whether the speaker intended to say “pea” or “bee”, but children can additionally be uncertain whether “bee” is a different word from “pee” (as opposed to, say, a valid within category variation), especially if these similar sounding words are newly learned (Creel, 2012; Merriman & Schuster, 1991; Stager & Werker, 1997; Swingley, 2016; White & Morgan, 2008). Though new word form representation can be shown to be differentiated under some circumstances (e.g., Yoshida, Fennell, Swingley, & Werker, 2009), this differentiation is still not mature enough and is probably noisier than the adult-like representation and/or encoded with lower confidence (see Swingley, 2007).

At the semantic level, early representations have, similarly, an intrinsically fragile and uncertain status. For example, upon seeing a bee in a foggy weather, adults may be uncertain if they saw a bee or a fly. But on top of this perceptual uncertainty, children may not be certain if the semantic category being named is that of bees and only bees, or if it includes other small flying insects like flies and beetles. In fact, though children can be fast at learning a first approximation of a given word’s referent (Carey & Bartlett, 1978), the refinement of this early approximation into a mature semantic category is a slow and gradual

process (see also Bion, Borovsky, & Fernald, 2013; Carey, 2010; Fernald, Perfors, & Marchman, 2006; McMurray, Horst, & Samuelson, 2012). Among other things, children have to enrich this early representation with new features, and revise its extension in the light of new referential exposures.

Thus, uncertainty in the representation associated with one modality (e.g., a bee and a fly) can be mitigated through the possibly more differentiated representations associated with the other modality (e.g., the sound “bee” is acoustically different from the sound “fly”). That being said, a multi-modal cue combination strategy might help children not only recognize an individual word instance, but also refine the underlying phonological and semantic representations in the process. Previous research in early word learning has—whether implicitly or explicitly—largely treated the process of learning form and of learning meaning as independent. However, the developmental data reviewed above shows that children do not wait to have completed the acquisition of form to start learning meanings, and that both form and meaning representations develop, rather, in a parallel fashion. A few studies pointed to the possibility of an interaction between sound and meaning in early acquisition. For instance, Waxman and Markow (1995) showed that labeling various objects with the same name helps infants form the broad semantic category (but see Sloutsky & Napolitano, 2003). Vice versa, Yeung and Werker (2009) showed that pairing similar sounds with different objects help infants pay attention to subtle phonological contrasts. The present study proposes a first step towards a formal framework where accounts of sound-meaning synergy in development can be unified and further explored. For example, one could imagine that, initially, visual and auditory categories have relatively large noise variances, and that development consists in reducing the values of the variances through a mutually constraining process as further multimodal data accumulates.

In conclusion, this work studied the mechanism of word identification under uncertainty in both the word form and the word meaning. To our knowledge, this is the first study that performs an ideal observer analysis of this task. We found people to be near

optimal in their cue combination: They weighted each modality according to its relative reliability. However, they also showed patterns of sub-optimality especially when the stimuli were perturbed with additional background noise. Future work will extend the present formal framework in order to account for how children use multimodal cues to refine their early phonological and semantic representations.

Appendix: derivation of the posterior (equation 1)

For an ideal observer, the probability of choosing category 2 when presented with an audio-visual instance $w = (a, v)$ is the posterior probability of this category:

$$p(W_2|w) = \frac{p(w|W_2)p(W_2)}{p(w|W_2)p(W_2) + p(w|W_1)p(W_1)}$$

Which reduces to:

$$p(W_2|w) = \frac{1}{1 + \frac{p(w|W_1)p(W_1)}{p(w|W_2)p(W_2)}}$$

In order to further simplify the quantity $\frac{p(w|W_1)}{p(w|W_2)}$, we use our assumption that the cues are uncorrelated,

$$p(w|W) = p(a, v|W) = p(a|A)p(v|V)$$

Using the log transformation, we get:

$$\ln\left(\frac{p(w|W_1)}{p(w|W_2)}\right) = \ln\left(\frac{p(a|W_1)}{p(a|W_2)}\right) + \ln\left(\frac{p(v|W_1)}{p(v|W_2)}\right)$$

Under the assumption that the categories are normally distributed and that, within each modality, the categories have equal variances, we get (after simplification):

$$\ln\left(\frac{p(a|W_1)}{p(a|W_2)}\right) = \frac{\mu_{A1} - \mu_{A2}}{\sigma_A^2} \times a + \frac{\mu_{A2}^2 - \mu_{A1}^2}{2\sigma_A^2}$$

and similarly:

$$\ln\left(\frac{p(v|W_1)}{p(v|W_2)}\right) = \frac{\mu_{V1} - \mu_{V2}}{\sigma_V^2} \times v + \frac{\mu_{V2}^2 - \mu_{V1}^2}{2\sigma_V^2}$$

When putting all these terms together, we the get this final expression for the posterior:

$$p(W_2|w) = \frac{1}{1 + (1 + \epsilon) \exp(\beta_0 + \beta_a a + \beta_v v)}$$

where

$$1 + \epsilon = \frac{p(W_1)}{p(W_2)}$$

$$\beta_0 = \frac{\mu_{A2}^2 - \mu_{A1}^2}{2\sigma_A^2} + \frac{\mu_{V2}^2 - \mu_{V1}^2}{2\sigma_V^2}$$

$$\beta_a = \frac{\mu_{A1} - \mu_{A2}}{\sigma_A^2}$$

$$\beta_v = \frac{\mu_{V1} - \mu_{V2}}{\sigma_V^2}$$

References

- Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3), 257–262.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Bankieris, K. R., Bejjanki, V., & Aslin, R. N. (2017). Sensory cue-combination in the context of newly learned categories. *Scientific Reports*, 7(1), 10890.
- Bates, D., & Watts, D. (1988). *Nonlinear regression analysis and its applications*. Wiley.
- Bejjanki, V., Clayards, M., Knill, D., & Aslin, R. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLoS ONE*, 6.
- Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, 126(1), 39–53.
- Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases.

Philosophical Transactions of the Royal Society of London B: Biological Sciences,
363(1493), 1001–1010.

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development*, 6(3), 184–205.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. In *Proceedings of the Stanford Child Language Conference* (Vol. 15, pp. 17–29).

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10, 335–344.

Clayards, M., Tanenhaus, M., Aslin, R., & Jacobs, R. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108.

Creel, S. (2012). Phonological similarity and mutual exclusivity: On-line recognition of atypical pronunciations in 3–5-year-olds. *Developmental Science*, 15(5), 697–713.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, 41(8), 2149–2169.

Eberhard, K., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24(6), 409–436.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870), 429–433.

Feldman, N., Griffiths, T., & Morgan, J. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4), 752–782.

Fernald, A., Perfors, A., & Marchman, V. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, 42(1), 98–116.

Freedman, D., Riesenhuber, M., Poggio, T., & Miller, E. and. (2001). Categorical

representation of visual stimuli in the primate prefrontal cortex. *Science*, 291.

Geisler, W. S. (2003). Ideal observer analysis. In *The visual neurosciences* (pp. 825–837).
Cambridge, MA: MIT Press.

Greenberg, J. (1957). *Essays in linguistics*. Chicago: University of Chicago Press.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of
american english vowels. *Journal of the Acoustical Society of America*, 97.

Hofer, M., & Levy, R. (2017). Modeling Sources of Uncertainty in Spoken Word Learning. In
Proceedings of the 39th Annual Meeting of the Cognitive Science Society.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the
familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 148.

Knill, D., & Pouget, A. (2004). The bayesian brain: The role of uncertainty in neural coding
and computation. *Trends in Neurosciences*, 27(12), 712–719.

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal
of Memory and Language*, 65(2), 145–160.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in
adverse conditions: A review. *Language and Cognitive Processes*, 27(7-8), 953–978.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the
interaction of online referent selection and slow associative learning. *Psychological
Review*, 119.

Merriman, W., & Schuster, J. (1991). Young children’s disambiguation of object name
reference. *Child Development*, 62(6), 1288–1301.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is
language? *Philosophical Transactions of the Royal Society of London B: Biological
Sciences*, 369(1651).

Norris, D., & McQueen, J. M. (2008). Shortlist B: A bayesian model of continuous speech

recognition. *Psychological Review*, 115(2), 357–395.

Pajak, B., Creel, S., & Levy, R. (2016). Difficulty in learning similar-sounding words: A developmental stage or a general property of learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9).

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, 9(1), 45–48.

Robinson, C. W., & Sloutsky, V. M. (2010). Development of cross-modal processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1.

Ronnberg, J., Rudner, M., Lunner, T., & Zekveld, A. (2010). When cognition kicks in: Working memory and speech understanding in noise. *Noise and Health*, 12(49), 263–269.

Saussure, F. (1916). *Course in general linguistics*. New York: McGraw-Hill.

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, 14(4), 332–348.

Sloutsky, V. M., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development*, 74.

Spivey, M. J., Tanenhaus, M., Eberhard, K., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45(4), 447–481.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388(6640).

Swingle, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology*, 43(2), 454–464.

Swingle, D. (2016). Two-year-olds interpret novel phonological neighbors as familiar words. *Developmental Psychology*, 52(7), 1011–1023.

Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence

831 representations of the mental lexicon. *The Mental Lexicon*, 3(2).

832 Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of
833 visual and linguistic information in spoken language comprehension. *Science*,
834 268(5217), 1632–1634.

835 Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind:
836 Statistics, structure, and abstraction. *Science*, 331(11 March 2011), 1279–1285.

837 Vouloumanos, A., & Waxman, S. (2014). Listen up! Speech is for thinking during infancy.
838 *Trends in Cognitive Sciences*, 18(12), 642–646.

839 Vroomen, J., Linden, S. van, Keetels, M., Gelder, B. de, & Bertelson, P. (2004). Selective
840 adaptation and recalibration of auditory speech by lipread information: Dissipation.
841 *Speech Communication*, 44.

842 Waxman, S., & Markow, D. (1995). Words as invitations to form categories: Evidence from
843 12-to 13-month-old infants. *Cognitive Psychology*, 29(3), 257–302.

844 White, K., & Morgan, J. (2008). Sub-segmental detail in early lexical representations.
845 *Journal of Memory and Language*, 59.

846 Yeung, H., & Werker, J. (2009). Learning words’ sounds before learning how words sound:
847 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*,
848 113, 234–243.

849 Yoshida, K., Fennell, C., Swingley, D., & Werker, J. (2009). 14-month-olds learn
850 similar-sounding words. *Developmental Science*, 12.

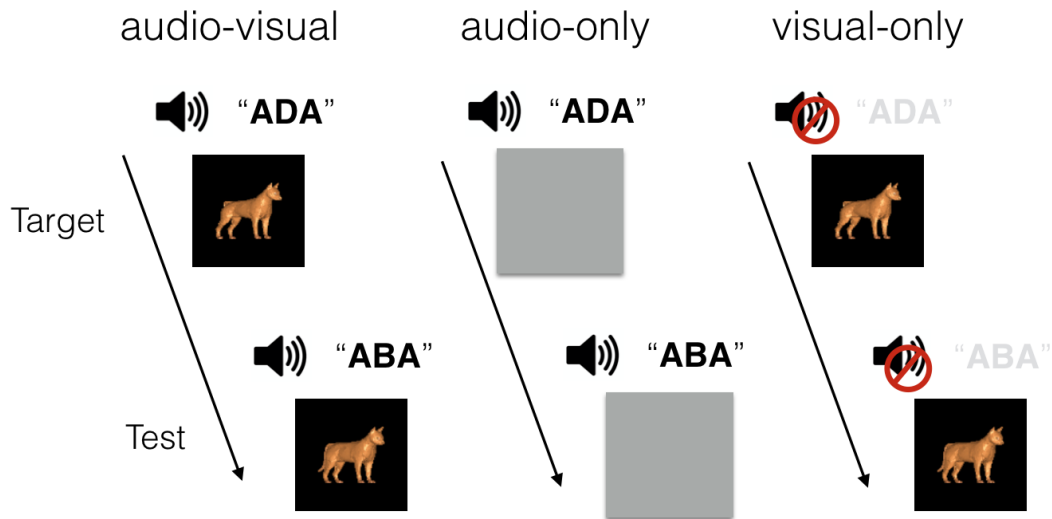


Figure 1. Overview of the task. In the audio-visual condition, participants are first presented with an audio-visual target (the prototype of the target category), immediately followed by an audio-visual test. The test may differ from the target in both the auditory and the visual components. After these two presentations, participants press ‘same’ (i.e., the same category as the target) or ‘different’ (not the same category). The auditory-only and visual-only conditions are similar to the audio-visual condition, except that only the sounds are heard, or only the pictures are shown, respectively.

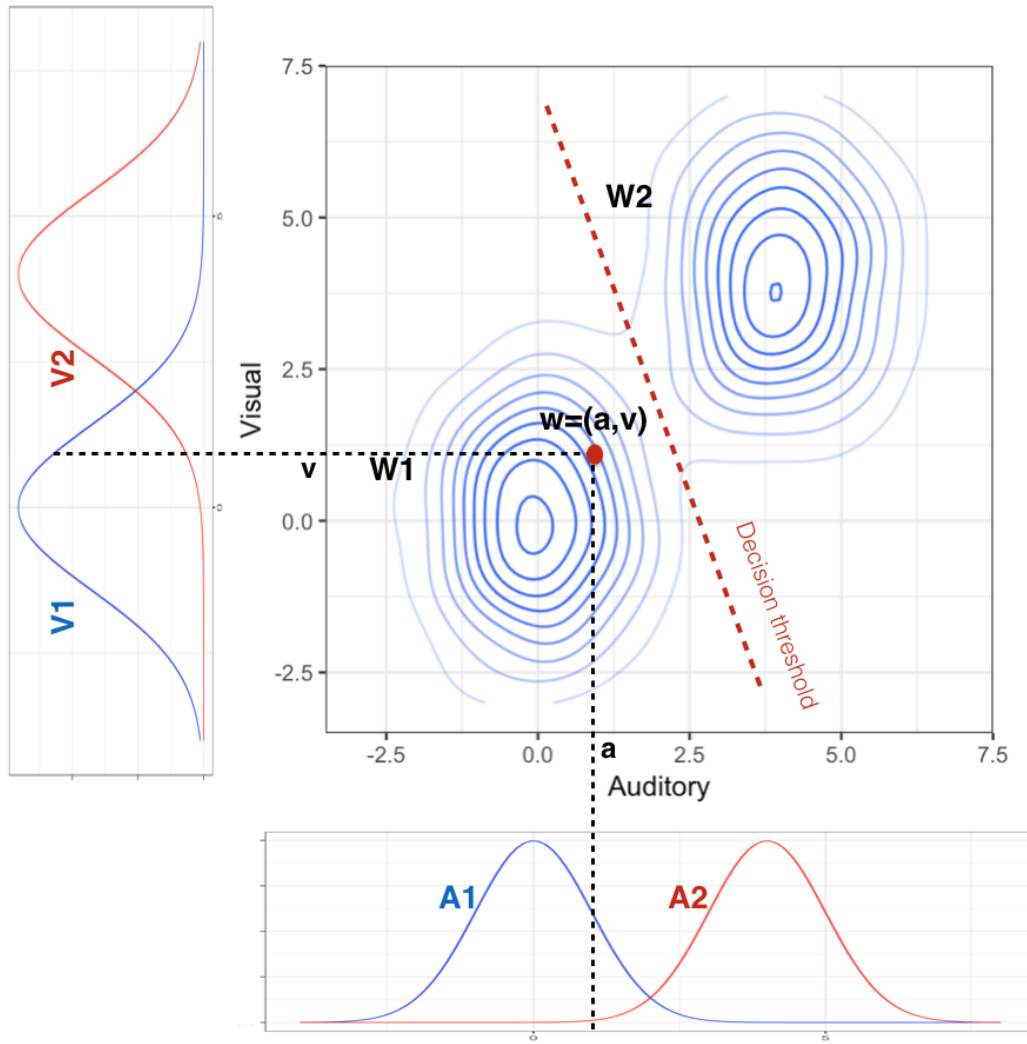


Figure 2. Illustration of the model using simulated data. A word category is defined as the joint bivariate distribution of an auditory category (horizontal, bottom panel) and a visual semantic category (vertical, left panel). Upon the presentation of a word token w , participants guess whether it is sampled from the word type W_1 or from the word type W_2 . Decision threshold is where the guessing probability is 0.5.

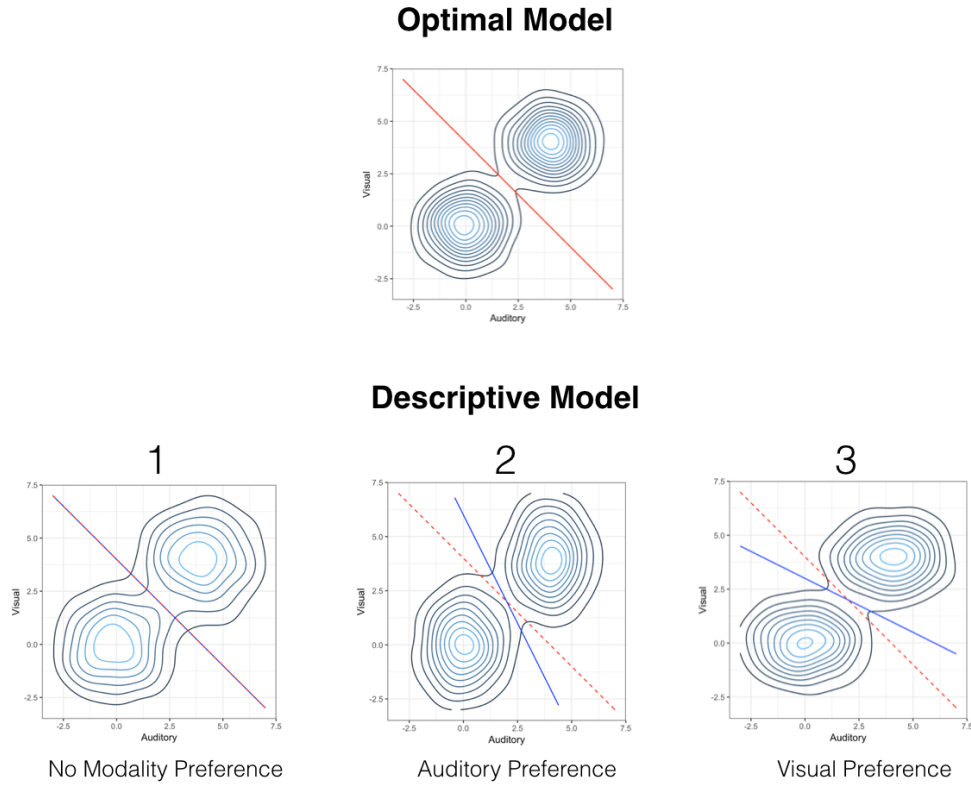


Figure 3. Illustration using simulated data showing the example of a prediction made by the optimal model (top), and the three possible ways human participants can deviate from this prediction (bottom). These cases are the following: 1) The variance increases equally for both modalities, but the weighting scheme (characterized by the decision threshold) is optimal, 2) The auditory variance increases at a higher rate, leading to a preference for the auditory modality, and 3) The visual variance increases at a higher rate, leading to a preference for the visual modality

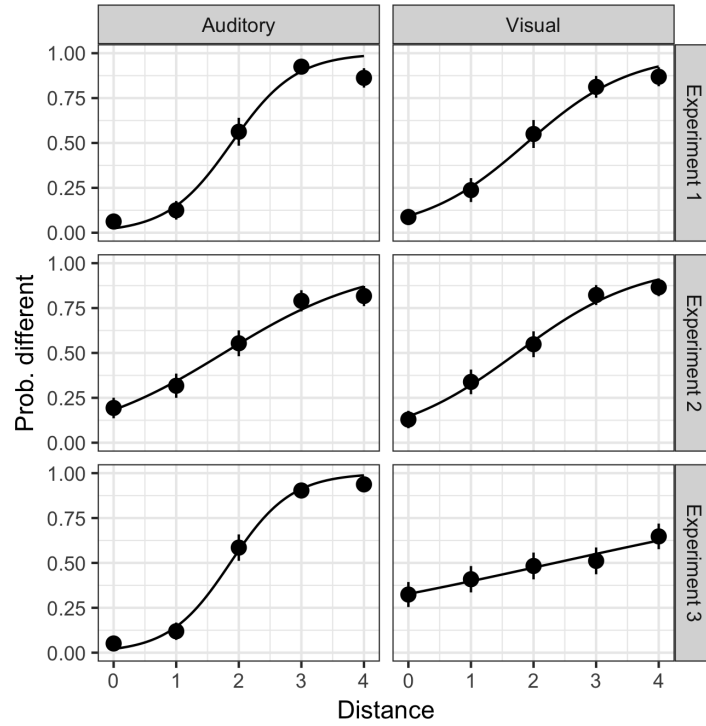


Figure 4. Human responses in the unimodal condition. Points represent the proportion of ‘different’ to ‘same’ responses in the auditory-only condition (left), and visual-only condition (right). Error bars are 95% confidence intervals. Solid lines represent logistic fits.

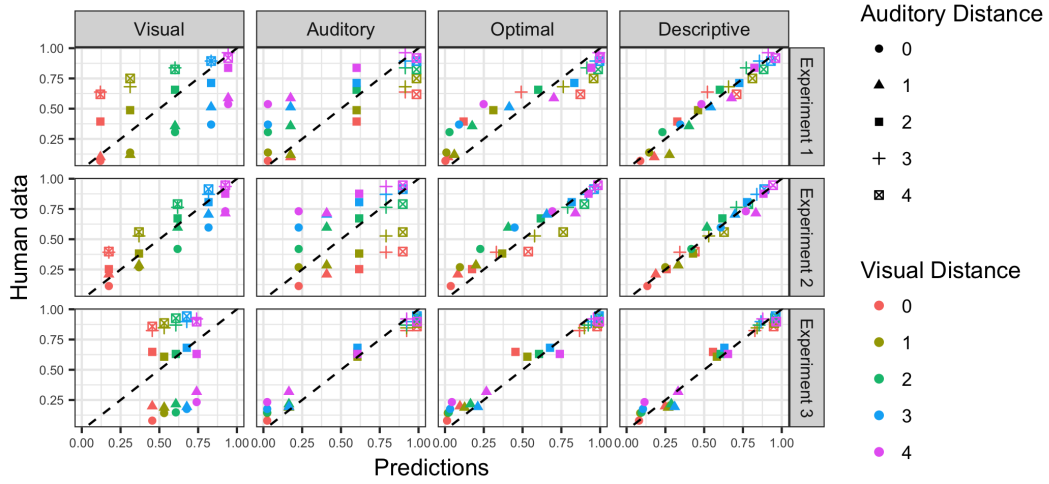


Figure 5. Human responses vs. models' predictions in the bimodal condition. Shape represents auditory distance from the target, and color represents visual distance from the target.

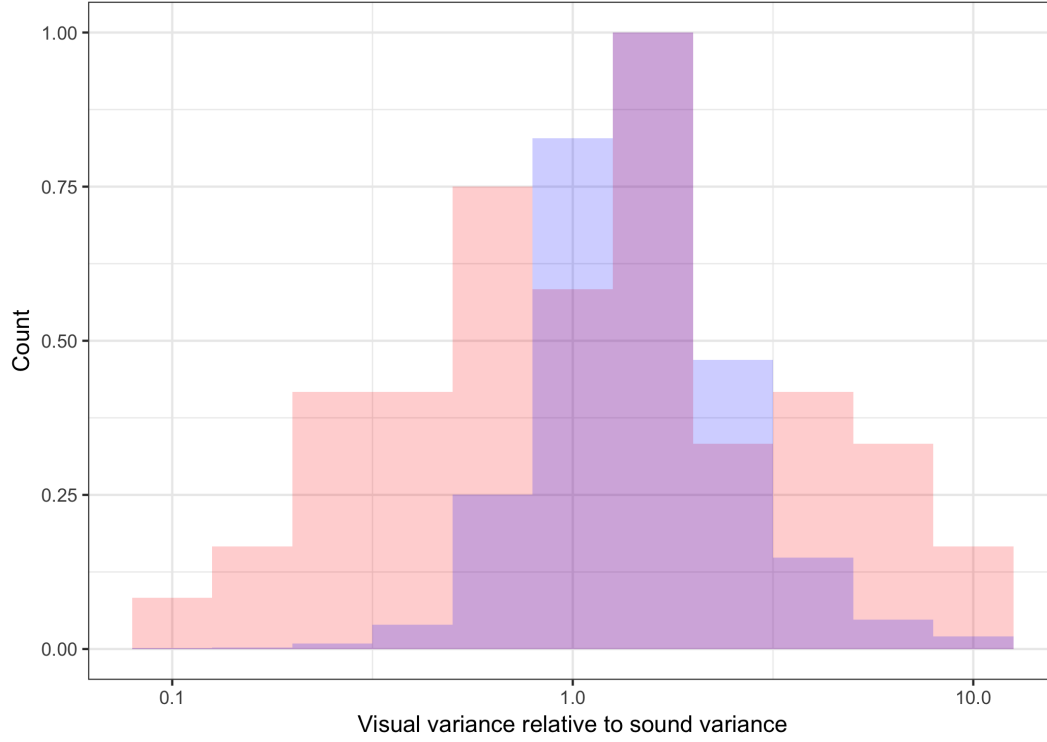


Figure 6. Histograms of the values of the visual variance relative to the auditory variance in Experiment 1. Light color represents the values derived from each individual participant, and dark color represents simulated data sampled from the descriptive model.

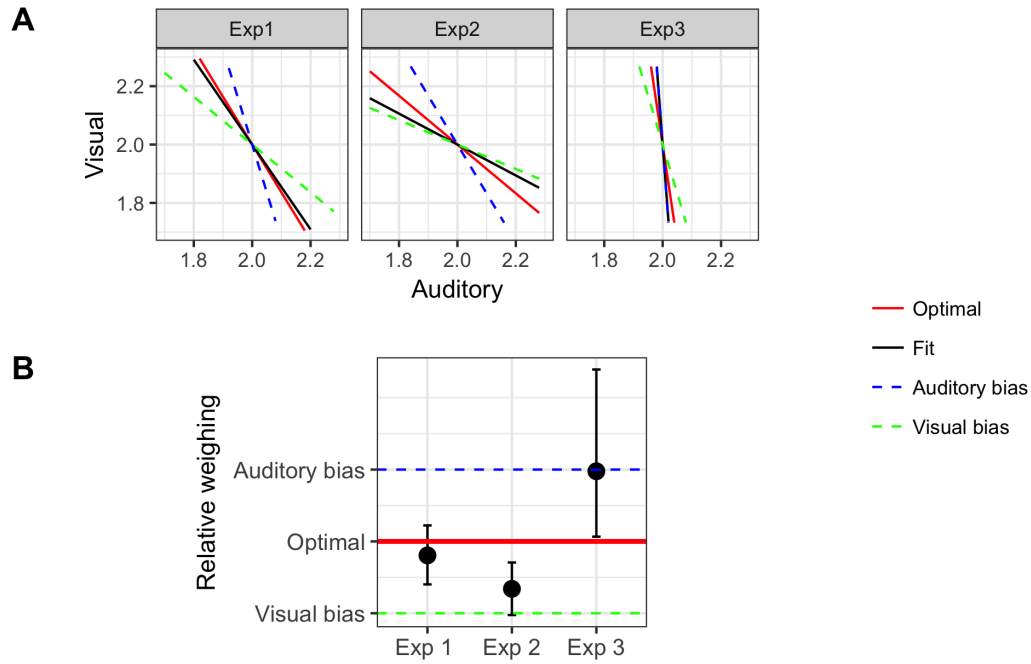


Figure 7. Modality preference is characterized as a deviation from the optimal decision threshold. A) The decision thresholds of both the optimal and the descriptive models (solid red and black lines respectively). Deviation from optimality is compared to two hypothetical cases of modality preference. In these cases, deviation from optimality is due to over-lying on the visual or the auditory input (green and blue dotted lines, respectively) by a factor of 2. B) The value of the decision threshold's slope derived from the descriptive model relative to that of the optimal model. Error bars represent 95% confidence intervals over the distribution obtained through non-parametric resampling.