How Optimal is Word Recognition Under Multimodal Uncertainty?

Abdellah Fourtassi¹ & Michael C. Frank¹

¹ Department of Psychology, Stanford University

Author Note

- Abdellah Fourtassi
- 6 Department of Psychology
- 7 Stanford University
- s 50 Serra Mall
- Jordan Hall, Building 420
- Stanford, CA 94301
- 11 Correspondence concerning this article should be addressed to Abdellah Fourtassi,
- Postal address. E-mail: afourtas@stanford.edu

Abstract

Identifying a spoken word in a referential context requires both the ability to integrate 14 multimodal input and the ability to reason under uncertainty. How do these tasks interact 15 with one another? We study how adults identify novel words under joint uncertainty in the auditory and visual modalities, and we propose an ideal observer model of how cues in these modalities are combined optimally. Model predictions are tested in four experiments where recognition is made under various sources of uncertainty. We found that participants use 19 both auditory and visual cues to recognize novel words. When the signal is not distorted 20 with environmental noise, participants weight the auditory and visual cues optimally, that is, 21 according to the relative reliability of each modality. In contrast, when one modality has 22 noise added to it, human perceivers systematically prefer the unperturbed modality to a 23 greater extent than the optimal model does. This work extends the literature on perceptual 24 cue combination to the case of word recognition in a referential context. In addition, this 25 context offers a link to the study of multimodal information in word meaning learning. 26

27 Keywords: Language understanding; audio-visual processing; word learning; speech 28 perception; computational modeling. How Optimal is Word Recognition Under Multimodal Uncertainty?

30

31

49

29

Introduction

Language uses symbols expressed in one modality — 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 provides 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). Moreover, 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. Perhaps the insect was a beetle or a fly
instead. How does the listener deal with such multimodal uncertainty to recognize the
speaker's intended word?

As a simplified case study of early word learning, the task of matching sounds to corresponding visual objects has been studied extensively in the developmental literature.

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

For example, many studies focus on how children might succeed in this type of task despite referential ambiguity (Medina, Snedeker, Trueswell, & Gleitman, 2011; Pinker, 1989; L. B. Smith & Yu, 2008; Suanda, Mugwanya, & Namy, 2014; Vlach & Johnson, 2013; Vouloumanos, 2008; Yurovsky & Frank, 2015). However, even when they have learned the exact meaning of a word, observers (both children and adults) often still find it challenging to recognize which word the speaker has uttered, especially under noise (Mattys, Davis, Bradlow, & Scott, 2012; Peelle, 2018). The purpose of the current study is thus to explore word recognition by adults under multimodal uncertainty, focusing on the special case where people have access to multimodal cues from the auditory speech and the visual referent. In the General Discussion, we return to the question of how these findings relate to questions about word learning.

One rigorous way to approach this question is through conducting an *ideal observer* 62 analysis. This research strategy provides a characterization of the task/goal and shows what 63 the optimal performance should be under this characterization.² 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. The ideal observer model can be seen as a theoretical upper limit on performance. It is not so much a realistic model of human performance, as much as a baseline against which human performance can be compared (Geisler, 2003; Rahnev & Denison, 2018). When there is a deviation from the ideal, it can reveal extra constraints on human cognition, such as limitations on the working memory or on attentional resources. This approach has had a tremendous impact not only 72 on speech-related research (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, Griffiths, 73 & 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 &

²It is, thus, a general instance of the rational approach to cognition (Anderson, 1990), instantiating Marr's computational level of analysis (Marr, 1982).

Pouget, 2004; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

Some prior ideal observer studies are closely related to the question we are addressing in the current work. For instance, Clayards et al. (2008) simulated auditory uncertainty by manipulating the probability distribution of a cue (Voice Onset Time) 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. Moreover, Feldman et al. (2009) studied the perceptual magnet effect, a phenomenon that involves reduced discriminability near prototypical sounds in the native language (Kuhl, 1991), showing that this effect can be explained as the consequence of optimally solving the problem of perception under uncertainty (see also Kronrod, Coppess, & Feldman, 2016).

Besides the acoustic cues explored in Clayards et al. (2008) and Feldman et al. (2009), 86 there is extensive evidence that information from the visual modality, such as the speaker's 87 facial features, also influences speech understanding (see Campbell, 2008 for a review). Bejjanki, Clayards, Knill, and Aslin (2011) offered a mathematical characterization of how probabilistic cues from speech and lip movements can be optimally combined. They showed that human performance during audio-visual phonemic labeling was consistent (at least at the qualitative level) with the predictions of an ideal observer. This previous research did not, however, study speech understanding when visual information was obtained through the referential context rather than through observation of the speaker's face. Although some experimental findings show that information about the identity of a referent can be integrated with linguistic information to resolve lexical and syntactic ambiguities in speech (e.g., Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), to our knowledge no study has offered an ideal observer analysis of this task (as we do here).

Combining information between words and visual referents might seem similar to audio-visual speech integration (e.g., Bejjanki et al., 2011), but there are at least two

fundamental differences between these two cases, and both can influence the way the auditory and visual cues are combined.

First, in the case of audio-visual speech, both modalities offer information about the 104 same underlying speech category. They differ only in terms of their informational reliability. 105 In a referential context, however, the auditory and visual modalities play different roles in 106 the referential process — the auditory input represents the symbol whereas the visual input 107 represents the meaning (and these differences are in addition to possible differences in 108 informational reliability). Speech is claimed to have a privileged status compared to other 109 sensory stimuli (Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012; Vouloumanos 110 & Waxman, 2014; Waxman & Gelman, 2009; Waxman & Markow, 1995), and this privilege 111 is suggested to be specifically related to the ability to refer (Waxman & Gelman, 2009). 112 Thus, in a referential context, it is possible that listeners do not treat the auditory and 113 visual modalities as equivalent sources of information. Instead, there could be a (potentially 114 sub-optimal) bias for the auditory modality beyond what is expected from informational 115 reliability alone. 116

Second, in the case of audio-visual speech, the auditory and visual stimuli are expected to be perceptually correlated. The expectation for this correlation is strong enough that when there is a mismatch between the auditory and visual input, they are still integrated into a unified (but illusory) percept (e.g., the McGurk Effect; 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 or perceptual connection between the sound "bee" and the corresponding insect. Moreover, variation in

³There is, however, a debate as to whether speech is privileged for children and adults for similar reasons. Whereas some researchers suggest that speech is privileged for both children and adults because of its ability to refer (e.g., Waxman & Gelman, 2009), others suggest that speech might *not* have a referential status from the start. Rather, speech might be preferred by children only because of a low level auditory "overshadowing" (e.g., Sloutsky & Napolitano, 2003). We do not take a stance on this issue here.

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. In sum,
cue combination in the case of arbitrary audio-visual associations (word-referent) is likely to
be 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).

129 The current study

147

We investigate how cues from the auditory and the visual modality are combined to 130 recognize novel words in a referential context. In particular, we study how this combination 131 is performed under various degrees of uncertainty in both the auditory and the visual 132 modality. Imagine, for example, that someone is uncertain whether they heard "pea" or 133 "bee." Does this uncertainty make them rely more on the referent (e.g., the object being 134 pointed at)? Or, if they are not sure if they saw a bee or a fly, does this uncertainty make 135 them rely more on the sound? More importantly, when input in both modalities is uncertain 136 to varying degrees, do they weight each modality according to its relative reliability (the optimal strategy), or do they over-rely on a particular modality? 138

We begin by proposing an ideal observer model that performs the combination in an 139 optimal fashion. We then compare the predictions of the optimal model to human responses. 140 Humans can deviate from the ideal for several reasons. For instance, as mentioned above, a 141 sub-optimality can be induced by the privileged status of a particular modality or by the 142 arbitrariness of the referential association. In order to study possible patterns of 143 sub-optimality, we compare the optimal normative model to a descriptive model (which is fit 144 to actual responses). Comparing parameter estimates between these two formulations allows 145 us to quantify the degree of deviation from optimality. 146

We tested the ideal observer model's predictions in four behavioral experiments where

we varied the source of uncertainty. In Experiment 1, audio-visual tokens were ambiguous
with respect to their category membership (in addition to sensory noise). In Experiment 2,
we intervened by adding environmental noise that degraded information from the auditory
modality, and in Experiment 3 we intervened by adding environmental noise that degraded
information from the visual modality. Finally, in Experiment 4, we replicated Experiment 1
with a higher power design, allowing us to test cue combination at the individual level.

Paradigm and Models

In this section, we first briefly introduce the multimodal combination task. Then we explain how behavior in this paradigm can be characterized optimally with an ideal observer model.

58 The Audio-Visual Word Recognition Task

154

170

We introduce an experimental paradigm adapted from a task used by Sloutsky and 159 Napolitano (2003). The original was used with both children and adults to probe 160 audio-visual encoding (see Robinson & Sloutsky, 2010 for review). Here we use a slightly different version to test word recognition in a referential context. We use two visual 162 categories (cat and dog) and two auditory categories (/b/ and /d/ embedded in the minimal 163 pair /aba/-/ada/). For each participant, an arbitrary pairing is set between the auditory and 164 the visual categories, leading to two audio-visual word categories (e.g., dog-/aba/, 165 cat-/ada/). In each trial, participants are presented with an audio-visual target (the 166 prototype of the target category), immediately followed by an audio-visual test stimulus 167 (Figure 1). The test stimulus may differ from the target in both the auditory and the visual 168 components. After these two presentations, participants press "same" or "different." 169

In the testing phase of the original task (Sloutsky & Napolitano, 2003), participants

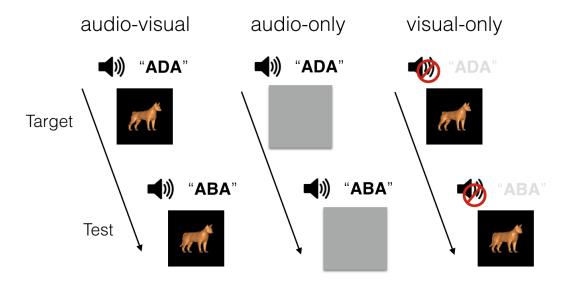


Figure 1. Overview of the task. In the audio-visual condition, participants are first presented with an audio-visual target (the prototype of the target category), immediately followed by an audio-visual test. The test may differ from the target in both the auditory and the visual components. After these 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.

were asked whether or not the two audio-visual presentations are *identical*. In the current 171 study, we are interested, rather, in the categorization, i.e., determining whether or not two 172 similar tokens are members of the same phonological/semantic category. Therefore, testing 173 in our task is category-based: Participants are asked to press "same" if they think the second 174 item (the test) belongs 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 176 press "different" only if they think that the second stimulus was an instance of the other 177 category (cat-/ada/). The task also includes trials where pictures are hidden (audio-only) or 178 where sounds are muted (visual-only). These unimodal trials provide us with the 179 participants' evaluation of the probabilistic information present in the auditory and visual 180

categories. As we shall see, these unimodal distributions are used as inputs to the optimal cue combination model.

3 Optimal Model

We construct an ideal observer model that combines probabilistic information from the 184 auditory and visual modalities. In contrast to the model used in most research on multisensory integration (e.g., Ernst & Banks, 2002), which typically studies continuous 186 stimuli (e.g., size, location), the probabilistic information in our case cannot be characterized with sensory noise only. Since our task involves responses over categorical variables 188 (phonemes and concepts), the optimal model should take into account not only the noise 189 variability around an individual perceptual estimate but also its categorical variability, i.e., the uncertainty related to whether this perceptual estimate belongs to a given category (see 191 also Bankieris, Bejjanki, & Aslin, 2017; Bejjanki et al., 2011). In what follows, we describe a 192 probabilistic model that accounts for both types of variability. First, we describe the model 193 in the simplified case of categorical variability only. Second, we augment this simplified 194 model to account for sensory and environmental noise. 195

Categorical variability. We assume that both the auditory categories (i.e., /aba/ and /ada/) and the visual categories (cat and dog) are distributed along a single acoustic and semantic dimension, respectively (Figure 2). Moreover, we assume that all categories are normally distributed. Formally speaking, if A denotes an auditory category (/ada/ or /aba/), then the probability that a point a along the acoustic dimension belongs to the category A is

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

where μ_A and σ_A^2 are respectively the mean and the variance of the auditory category. Similarly, the probability that a point v along the visual dimension belongs to the category V is

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

where μ_V and σ_V^2 are the mean and the variance of the visual category. An audio-visual signal w=(a,v) can be represented as a point in the audio-visual space. These audio-visual tokens define bivariate distributions in the bi-dimensional space. We call these bivariate distributions $Word\ categories$, noted W, and are distributed as follows:

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

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

The main assumption of the model is that the auditory and visual variables are independent

(i.e., uncorrelated), so the covariance matrix is simply:

$$\Sigma_W = \left[egin{array}{cc} \sigma_A^2 & 0 \\ 0 & \sigma_V^2 \end{array}
ight]$$

This assumption says that, given a word-object mapping, e.g., W = (``cat"-CAT'), variation in the way "cat" is pronounced does not correlate with changes in any visual property of the object CAT, which is a valid assumption in the context of our task.⁴

Now we turn to the crucial question of modeling how the optimal decision should proceed given the probabilistic (categorical) information in the auditory and the visual modalities, as characterized above. We have two word categories: dog-/aba/ (W_1) and cat-/ada/ (W_2) . When making decisions, participants can be understood as choosing one of these two word categories (Figure 2). For an ideal observer, the probability of choosing category 2 when presented with an audio-visual instance w = (a, v) is the posterior

⁴Note that this assumption 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.

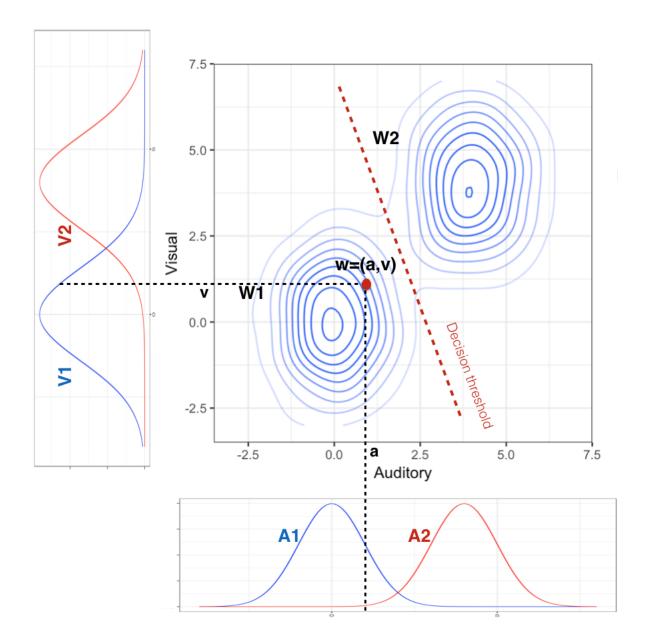


Figure 2. Illustration of the model using simulated data. A word category is defined as the joint bivariate distribution of an auditory category (horizontal, bottom panel) and a visual semantic category (vertical, left panel). Upon the presentation of a word token w, participants guess whether it is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

220 probability of this category:

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

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

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

Under this assumption, the posterior probability reduces to the following formula (see Appendix 1 for the details of the derivation):

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

where

225

226

$$1 + b = \frac{p(W_1)}{p(W_2)}$$
$$\beta_0 = \frac{\mu_{A2}^2 - \mu_{A1}^2}{2\sigma_A^2} + \frac{\mu_{V2}^2 - \mu_{V1}^2}{2\sigma_V^2}$$

$$\beta_a = \frac{\mu_{A1} - \mu_{A2}}{\sigma_A^2}$$
$$\beta_v = \frac{\mu_{V1} - \mu_{V2}}{\sigma_V^2}.$$

The parameter b represents the differential between the categories' prior probabilities. 227 However, since the identity of word categories is randomized across participants, b measures, 228 rather, a response bias to "same" if b > 0, and a response bias to "different" if b < 0. We 229 expect a general bias towards answering "different" because of the categorical nature of our 230 same-different task: When two items are ambiguous but perceptually different, participants 231 might have a slight preference for "different" over "same." As for the means, their values are 232 fixed, and they correspond to the most typical tokens in our stimuli. Finally, observations from each modality (a and v) are weighted in Equation 1 according to their reliability (that 234 is, according to the *inverse* of their variance): 235

$$\beta_a \propto \frac{1}{\sigma_A^2}$$
 $\beta_v \propto \frac{1}{\sigma_V^2}$.

236

Sensory variability. So far, we have only accounted for categorical variability, i.e., 237 $\sigma_A^2 = \sigma_{A_C}^2$. For instance, if the speaker generates a target production a_t from an auditory 238 category $p(a_t|A) \sim N(\mu_A, \sigma_{A_C}^2)$, the ideal model assumes that it has direct access to this 239 production token (i.e., $a = a_t$), and that all uncertainty is about the category membership of 240 this token. However, we might also want to account for internal noise in the brain and/or 241 external noise in the environment. For example, the observer might not have access to the 242 exact produced target, but only to the target perturbed by noise. If we assume this noise to 243 be normally distributed, that is, $p(a|a_t) \sim N(a_t, \sigma_{A_N}^2)$, then integrating over a_t leads to this new expression of the probability distribution: 245

$$p(a|A) \sim N(\mu_A, \sigma_{A_C}^2 + \sigma_{A_N}^2)$$

Similarly, in the case of sensory noise in the visual modality, we get:

250

$$p(a|V) \sim N(\mu_V, \sigma_V^2 + \sigma_{V_N}^2)$$

Finally, using exactly the same derivation as above, we end up with the following multimodal weighting scheme in the optimal combination model (Equation 1) which takes into account both categorical and sensory variability:

$$\beta_a \propto \frac{1}{\sigma_{A_C}^2 + \sigma_{A_N}^2}$$

$$\beta_v \propto \frac{1}{\sigma_{V_C}^2 + \sigma_{V_N}^2}.$$

Optimal cue combination. Equation 1 provides the optimal model's predictions
for how probabilities that characterize uncertainty in the auditory and the visual modalities
can be combined to make categorical decisions. Parameter estimates of the probability
distributions in each modality are derived by fitting unimodal posteriors to the participants'
responses in the unimodal conditions, i.e., the condition where only the sounds are heard or

only the pictures are seen (Figure 1).⁶ Using these derived parameters, the optimal model makes predictions about responses in the bimodal (i.e., audio-visual) condition where participants both hear the sounds and see the pictures.

Auditory and Visual baselines. The predictions of the optimal model will be 259 compared to two baselines. The first baseline is a visual model which assumes that 260 participants rely only on visual information, and an auditory model, which assumes that 261 participants rely only on auditory information. More precisely, these baseline models assume 262 that the participants' responses in the bimodal condition will not be different from their 263 response in either the visual-only or the auditory-only condition. However, if the participants 264 rely on both the auditory and the visual modalities to make decisions in the bimodal 265 condition, the optimal model would explain more variance in human responses than the 266 visual or the auditory model does. 267

Descriptive model and analysis of (sub-)optimality

The optimal model (as well as the auditory and visual baselines) are *normative* models. 260 Their predictions are made about human data in the bimodal condition, but their crucial 270 parameters (i.e., variances associated with the visual and auditory modalities) are derived 271 from data in the unimodal conditions. In addition to these normative models, we consider a 272 descriptive model. It is formally identical to the normative optimal model (Equation 1), 273 except that the parameters are fit to actual responses in the bimodal condition. If the 274 referential task induces sub-optimality (due, for instance, to the arbitrary nature of the sound-object association), then the descriptive model should explain more variance than the 276 optimal model does.

how much people deviate from optimality but also to understand precisely the nature of this 279 deviation. Let σ_A^2 and σ_V^2 be the values of the variances used in the optimal model (derived 280 from the unimodal conditions), and σ_{Ab}^2 and σ_{Vb}^2 be the values observed through the 281 descriptive model in the bimodal condition. Deviation from optimality is measured in two 282 ways. First, we measure the change in the values of the variance specific to each modality, 283 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 284 in the proportion of the visual and auditory variances, i.e., we examine how $\frac{\sigma_A^2}{\sigma_V^2}$ compares to 285 $\frac{\sigma_{Ab}^2}{\sigma_{Vb}^2}$. The first measure allows us to test if response precision changes for each modality when 286 we move from the unimodal to the bimodal conditions. The second allows us to test the 287 extent to which the weighting scheme follows the prediction of the optimal model. The 288 reason we used the proportion of the variances as a measure of cross-modal weighting is that 289 this proportion 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 291 along which the posterior is equal to 0.5. Formally speaking, the decision threshold has 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
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).

There are three possible ways human responses can deviate from optimality. These

300

⁷Or more precisely the absolute value of the slope.

scenarios are illustrated in Figure 3, and are as follows: 301

304

311

312

313

314

315

316

317

318

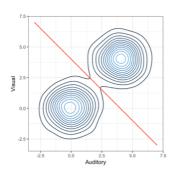
320

- 1) Both variances may increase, but their proportion remains the same. That is, $\sigma_{Ab}^2 \geqslant \sigma_A^2$ 302 and $\sigma_{Vb}^2 \geqslant \sigma_V^2$, 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 303 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 305 weigh modalities according to the relative reliability predicted by the optimal model. 306
- 2) The auditory variance increases at a higher rate. That is, $\sigma_{Ab}^2 \gg \sigma_A^2$ and $\sigma_{Vb}^2 \geqslant \sigma_V^2$, 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 308 being more random in the bimodal condition, but also in having a systematic 309 preference for the visual modality, even after accounting for informational reliability. 310
 - 3) The visual variance increases at a higher rate. That is, $\sigma_{Vb}^2 \gg \sigma_V^2$, and $\sigma_{Ab}^2 \geqslant \sigma_A^2$, 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 compared these models to human responses in four experiments. In Experiment 1, we studied the case where bimodal uncertainty was due to categorical variability and sensory noise. In Experiment 2 and 3 we added environmental noise to the auditory and the visual modalities, respectively. Finally, in Experiment 4, we replicated Experiment 1 with a higher power design in order to test cue combination at the individual level.

Experiment 1

In this experiment, we test the predictions of the model in the case where uncertainty 321 is due to categorical variability (i.e., ambiguity in terms of category membership) and 322 inherent sensory noise. We do not add any external noise to the background. Thus, we test 323





Descriptive Model

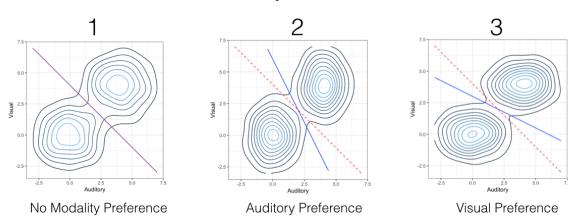


Figure 3. Illustration using simulated data showing the example of a prediction made by the optimal model (top), and the three possible ways human participants can deviate from this prediction (bottom). These cases are the following: 1) The variance increases equally for both modalities, but the weighting 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.

the following (normative) cue weighting scheme.⁸

⁸In all our analyses, we study the combined effect of both categorical and sensory uncertainties in the process of cue combination. We do not examine the specific role of each source of uncertainty separately. For example, in the auditory modality, the quantity $\sigma_{AC}^2 + \sigma_{AN}^2$ will be treated as a single variable. This variable

$$\beta_a \propto \frac{1}{\sigma_A^2} = \frac{1}{\sigma_{A_C}^2 + \sigma_{A_N}^2}$$

$$\beta_v \propto \frac{1}{\sigma_V^2} = \frac{1}{\sigma_{V_C}^2 + \sigma_{V_N}^2}.$$

Methods

325

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. 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. All participants provided informed consent
before taking the experiment. 9

For auditory stimuli, we used the continuum introduced in Vroomen, 334 Linden, Keetels, Gelder, and Bertelson (2004), a 9-point /aba/-/ada/ speech continuum 335 created by varying the frequency of the second (F2) formant in equal steps. We selected 5 336 equally spaced points from the original continuum by keeping the endpoints (prototypes) 1 337 and 9, as well as points 3, 5, and 7 along the continuum. For visual stimuli, we used a 338 cat/dog morph continuum introduced in Freedman, Riesenhuber, Poggio, and Miller (2001). 339 From the original 14 points, we selected 5 points as follows: we kept the item that seemed 340 most ambiguous (point 8), the 2 preceding points (i.e., 7 and 6) and the 2 following points 341 (i.e., 9 and 10). The 6 and 10 points along the morph were quite distinguishable, and we took them to be our prototypes.

will be measured in the unimodal condition and tested in the bimodal condition.

⁹The sample size and exclusion criteria were specified in the pre-registration at https://osf.io/h7mzp/.

Design and Procedure. We told participants that an alien was naming two 344 objects: a dog, called "aba" in the alien language, and a cat, called "ada". In each trial, we 345 presented the first object (the target) on the left side of the screen simultaneously with the 346 corresponding sound. For each participant, the target was always the same (e.g., dog-/aba/). 347 The second sound-object pair (the test) followed on the other side of the screen after 500ms 348 and varied in its category membership. For both the target and the test, visual stimuli were 349 present for the duration of the sound clip ($\sim 800 \text{ms}$). We instructed participants to press "S" 350 for same if they thought the alien was naming another dog-/aba/, and "D" for different if 351 they thought the alien was naming a cat-/ada/. We randomized the sound-object mapping 352 (e.g., dog-/aba/, cat-/ada/) as well as the identity of the target (dog or cat) across 353 participants. 354

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. The experiment lasted around 15 minutes.¹⁰

Model fitting details.

363

364

$Unimodal\ conditions.$

Remember that data in these conditions allows us to derive the variances of both the auditory and the visual categories and that these variances are used to make predictions

The experiment can be accessed and played from the github repository: https://github.com/afourtassi/

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 to 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:

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

with $\beta_a = \frac{\mu_{A_1} - \mu_{A_2}}{\sigma_A^2}$ and $\beta_{a0} = \frac{\mu_{A_2}^2 - \mu_{A_1}^2}{2\sigma_A^2}$. b_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 b_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.

$Bimodal\ condition.$

381

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 b, the
response bias. The visual and auditory baselines, as well as the optimal model, are not fit to
the 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 fit to bimodal data.

Although the paradigm is within-subjects, we did not have enough statistical power to
fit a different model for each individual participant (but see Experiment 4). Instead, models
were 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 using non-parametric bootstrap (with 10000
iterations).

Table 1
Estimates of the models' parameters.

	Auditory		Visual		Bimodal		
Experiment	b_A	$\mathrm{Var}_{\mathrm{A}}$	b_{V}	$\mathrm{Var}_{\mathrm{V}}$	b_b	Var_{Ab}	Var_{Vb}
Experiment1	-0.20	2.04	-0.12	3.33	-0.34	4.96	7.06
Experiment2	-0.18	4.70	-0.24	3.93	-0.38	9.84	5.21
Experiment3	-0.24	1.94	0.11	13.00	-0.35	3.00	39.42
Experiment4	-0.40	1.92	-0.22	3.24	-0.42	4.17	7.28

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. The
unimodal models' estimates in each modality are shown in Table 1.

Bimodal condition. Figure 5 compares the predictions of the normative and
descriptive models against human responses. Remember that the normative models use the
parameters estimated from the unimodal conditions (where people see input from only one
modality) to predict behavior in the bimodal condition (where people see input from both

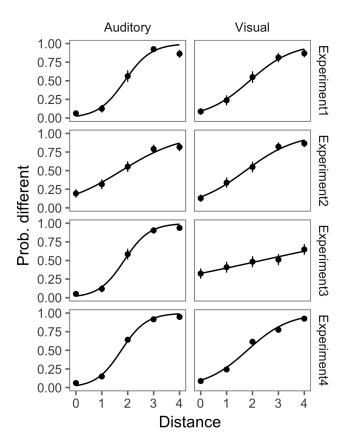


Figure 4. Human responses in the unimodal conditions across the four experiments. Points represent the proportion of 'different' to 'same' responses in the auditory-only condition (left), and visual-only condition (right). Error bars are 95% confidence intervals. Solid lines represent best unimodal posterior fits.

modalities). The descriptive model has a similar structure than the optimal model but is
directly fit to human responses in the bimodal condition in order to allow us to assess
deviation from optimality.

We found, by comparing the correlation values, that the optimal model explained more variance than the visual and auditory models did. However, the optimal model was not perfect: It explained less variance than the descriptive model did, which indicates a deviation from optimality. To investigate this deviation, we compare the parameter values of the optimal model to the values obtained in the descriptive model (Table 1). We note an

increase in *both* the auditory and visual variances. This increase in noise is compatible with the fact that human responses appear to be pulled towards chance (i.e., the value 0.5) when compared to the optimal model (see Figure 5). Below we investigate if this deviation from optimality can be related to the cue combination strategy.

Cue combination. We analyzed if the cue combination was performed in an 417 optimal way, or if there was a systematic preference for one modality when making decisions 418 in the bimodal condition. As explained in Figure 3, modality preference can be characterized 419 formally as a deviation from the decision threshold predicted by the optimal model. The 420 results in Figure 6 (top) show both the decision threshold derived from the descriptive model 421 (in black) and the decision threshold predicted by the optimal model (in red). We found that 422 the descriptive and optimal decision thresholds were almost identical. Indeed, 423 non-parametric resampling of the data showed no evidence of a deviation from the optimal 424 prediction (Figure 6, bottom). 425

426 Discussion

This experiment studied the way participants combine multimodal information to 427 recognize novel words. We found that the optimal model explained more variance than the 428 auditory or the visual models did, indicating that participants take into account both the 429 auditory and visual cues when making a decision. That said, Figure 5 shows that the 430 participants deviated slightly — but systematically— from the optimal prediction in that they were slightly pulled toward chance (i.e., the probability 0.5). This fact was captured by the increase in the value of the variance associated with each modality (as can be noted from Table 1). Note, however, that despite this increase in response randomness, our analysis of 434 modality preference showed that the relative values of these variances were not different 435 (Figure 6), meaning that there was no evidence for a modality preference.

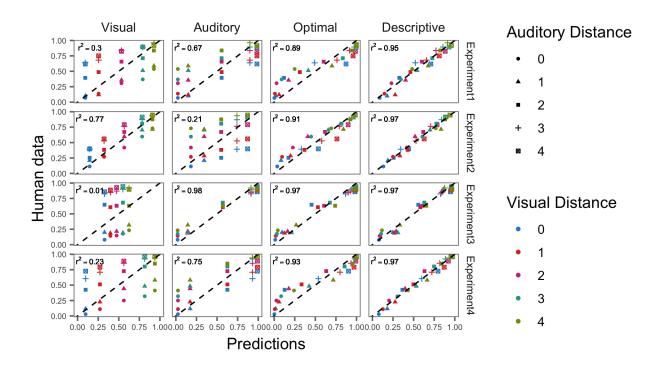


Figure 5. Human responses vs. Models' predictions in the bimodal condition across the four experiments. Each point represents data from a particular audio-visual matching (corresponding to an instance from the set of 5x5 possible matchings in the audio-visual space). Shape represents auditory distance from the target, and color represents visual distance from the target. Thus, each point is characterized by both shape and color.

To sum up, 1) the participants used both the auditory and visual information, 2) they responded slightly more randomly that what was predicted, but 3) this increased randomness was general and did not influence the cue combination strategy, i.e., the participants still weighted modalities according to their relative reliability as predicted by the optimal model. This situation corresponds to the first case of sub-optimality described in Figure 3.

437

439

In Experiment 1, we tested word recognition when there was multimodal uncertainty in terms of category membership and perceptual noise. In real life, however, both sound and

visual tokens can undergo distortions due to noisy factors in the environment (e.g., car noise in the background, blurry vision in 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 environmental noise σ_E^2 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:

$$\beta_a \propto \frac{1}{\sigma_A^2} = \frac{1}{\sigma_{A_C}^2 + \sigma_{A_N}^2 + \sigma_{A_E}^2}$$
$$\beta_v \propto \frac{1}{\sigma_V^2} = \frac{1}{\sigma_{V_C}^2 + \sigma_{V_N}^2}.$$

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

$_{ ext{456}}$ Methods

447

Participants. A sample of N=100 participants was recruited online through

Amazon Mechanical Turk. We used the same exclusion criteria as in Experiment 1. N=7participants 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

467

The analysis is similar to the analysis we did in Experiment 1.

Unimodal condition. We fit a model for each modality. Figure 4 shows human responses together with their best fits. The visual data is a replication of the visual data in Experiment 1. The auditory data, in contrast, were flatter, showing more uncertainty.

Bimodal condition. We used the values derived from the unimodal condition to
construct the visual, auditory and optimal models. In addition, we fit a descriptive model
which allowed us to assess real human performance in this condition. Figure 5 shows that,
similar to Experiment 1, the optimal model explained more variance than the auditory and
visual models did (note, however, that the visual model explained more variance than the
auditory model did). Also similar to Experiment 1, the values of the variances increased in
the bimodal condition (Table 1).

Cue combination. Here we investigated whether the observed increase in the
auditory and visual variances affected the relative weighting of the corresponding modalities.
Figure 6 (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 6, bottom).

484 Discussion

497

501

Experiment 2 tested audio-visual combination in the case where the auditory input was
noisy. We found, similar to Experiment 1, that the optimal model explained more variance
than the auditory or the visual models did. In other words, despite additional noise,
participants still used information from the noisy modality to recognize words. We also
found a similar discrepancy between the descriptive and optimal models as response
randomness increased along both the auditory and the visual modalities. As for the relative
weighting, and contrary to Experiment 1 where modalities were weighted optimally, we
found in this experiment that the visual modality had a greater weight than what was
expected from its relative reliability. This situation corresponds to the second case of
sub-optimality described in Figure 3.

Whereas 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

Similar to Experiment 2, we were interested to know if participants would treat additional uncertainty as predicted by the optimal model, that is, according to the following weighting scheme:

$$\beta_a \propto \frac{1}{\sigma_A^2} = \frac{1}{\sigma_{A_C}^2 + \sigma_{A_N}^2}$$
$$\beta_v \propto \frac{1}{\sigma_V^2} = \frac{1}{\sigma_{V_C}^2 + \sigma_{V_N}^2 + \sigma_{V_E}^2}.$$

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

preference for the visual input in Experiment 2.

505 Methods

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

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 radius¹¹ of 10 pixels. The experimental

procedure was exactly the same as in the previous Experiments.

516 Results

Unimodal conditions. 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 found that, in contrast to Experiment 1 and 2,
responses were flatter, showing much more uncertainty.

Bimodal condition. Figure 5 shows that almost all the variance was captured by
the auditory model alone, the addition of visual information in the optimal model did not
improve the prediction of human responses. Similar to Experiments 1 and 2, the values of
the variances increased in the bimodal condition (Table 1).

 $^{^{11}\}mathrm{A}$ feature that modulates the intensity of the blur.

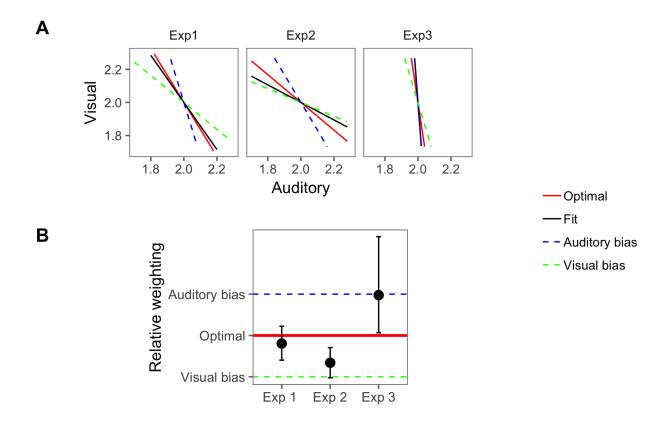


Figure 6. Modality preference is characterized as a deviation from the optimal decision threshold. A) The decision thresholds of both the optimal and the descriptive models (solid red and black lines, respectively). Deviation from optimality is compared to two hypothetical cases of modality preference. In these cases, deviation from optimality is due to over-lying on the visual or the auditory input by a factor of 2 (green and blue dotted lines, respectively). B) An alternative way to represent the same data. Each point represents the value of the decision threshold's slope derived from the descriptive model relative to that of the optimal model (log-scaled). The lines represent the optimal case as well as the two hypothetical cases of modality preference. Error bars represent 95% confidence intervals over the distribution obtained through non-parametric resampling.

Cue Combination. Figure 6 indicates that the decision threshold was biased towards the auditory modality (the non-noisy modality). Indeed non-parametric resampling

525

526

of the data showed an increase in the value of the slope in the descriptive model compared to
the optimal model (Figure 6).

Discussion

545

Experiment 3 tested audi-visual combination in the case where the visual input was 530 noisy. Whereas in previous experiments the optimal model explained more variance than the 531 auditory or the visual models did, here the auditory model alone explained almost all the 532 variance. In other words, though participants were sensitive to variation in the noisy visual 533 input when presented in isolation (as shown in Figure 4), they tended to ignore this 534 information when the visual input was presented simultaneously with the auditory input 535 (i.e., in the bimodal condition). Instead, they relied almost exclusively on the non-noisy 536 auditory modality.¹² 537

This 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 the optimal model. 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).

So far we have studied the problem of cue combination at the population level — the — 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 (see Experiment 1 in Figure 4). This fact 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.

552

554

563

models were fit to the data aggregated across all participants. However, it is important to 546 investigate individual variability, especially for cases when we reported optimal cue 547 combination (i.e., Experiment 1). In fact, optimality at the population level can be spurious 548 if it is obtained only on average while individuals have sub-optimal strategies (e.g., 549 over-relying on the visual or the auditory modalities). In Experiment 4, below, we examine 550 how the average cue combination relates to individual strategies.¹³ 551

Experiment 4

As we noted earlier, we did not have enough statistical power in Experiment 1 to fit a 553 different model for each participant. Thus, here we used a higher power design, allowing us to collect the number of data points necessary to model cue combination at the individual level. 555

Participants. We recruited a planned sample of N=50 participants from Amazon 556 Mechanical Turk. Only participants with US IP addresses and a task approval rate above 557 99% were allowed to participate. Participants were excluded if they reported having 558 experienced a technical problem of any sort during the online experiment (N=0), or if they 559 had less than 75\% accurate responses on the unambiguous training trials (N=7). The final 560 sample consisted of N=43 participants. All participants provided informed consent before 561 taking the experiment.¹⁴ 562

We used the same stimuli as in Experiment 1.

 $^{^{13}}$ Note that while the claim of (near-)optimality (which we made in Experiment 1) should be verified at the individual level — since it is possible to have a spurious optimal pattern on average while individuals show suboptimal behavior, it is not necessary to verify sub-optimality (which we found at the population level in Experiments 2 and 3) for every individual — since a sub-optimal average pattern cannot emerge from a predominantly optimal behavior at the individual level.

¹⁴The sample size, exclusion criteria and the main analyses were pre-registered at https://osf.io/h7mzp/.

Design and Procedure. The design and procedure were similar to Experiment 1.

There were, however, two differences: 1) We increased the number of responses elicited per subject from 70 to 300, and 2) we randomized the order of the three blocks (i.e., visual-only, auditory-only, and audio-visual) within subject: Each participant saw the 3 blocks exactly 6 times, covering all possible ordering combinations. Unlike the between-subject randomization that we used in Experiment 1, this choice allowed us to avoid a possible confound linked to the order of exposure.

71 Results

Unimodal and Bimodal conditions. In order to replicate the analysis of
Experiment 1, we started by fitting population-level models to the aggregated data. Indeed,
we found that the results — as shown in Figure 4, Figure 5, and Table 1 — mirror closely
the patterns obtained in Experiment 1.

Cue combination. We analyzed the cue combination strategies at the individual level. For each participant, we computed the optimal weighting, $\frac{\sigma_A^2}{\sigma_V^2}$, relative to the observed (i.e., descriptive) weighting, $\frac{\sigma_{Ab}^2}{\sigma_{Vb}^2}$. We show the resulting distribution in Figure 7. We note first that the distribution has a rather unimodal shape, centered around the optimal cue combination strategy. This finding rules out the hypothesis that optimality at the population level is a spurious finding, i.e., only obtained by aggregating over various sub-optimal strategies.

In addition, we asked whether the observed variance in the individual distribution was
due to the randomness inherent to the process of sampling from a probabilistic model or
whether it corresponded to a real between-subject variability induced by different cue
weighting strategies. We simulated responses through sampling from the population-level
models and we computed the resulting distribution of cue weighting for each simulated

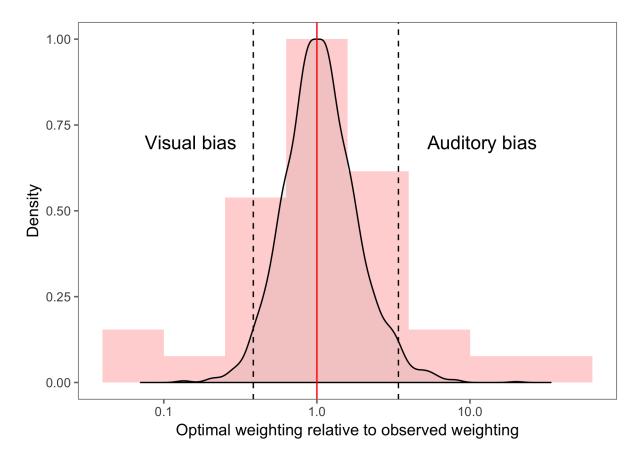


Figure 7. The histogram shows the distribution of the participants' predicted (i.e., optimal) cue weighting relative the observed (i.e., descriptive) weighting (see Figure 3 for the details). The density plot shows the distribution of simulated data sampled from the population-level probabilistic model. The dashed lines represent 95% confidence interval on this simulated distribution. Optimal behavior is exactly obtained when the value of the relative cue weighting is 1 (red solid line). Participants whose values are outside the confidence interval of the simulated distribution can be understood to be over-relying on the visual modality (left side) or on the auditory modality (right side) beyond sampling-related variation (i.e., with p < 0.05).

individual leading to the density function in Figure 7. We can observe that most empirical values fall within the 95% confidence interval of the simulated density, showing that this part of the variance can be due to mere sampling variation. However, a few participants had

values outside this interval, indicating that they systematically over-relied on the visual modality (N = 5, p < 0.05) or the auditory modality (N = 6, p < 0.05).

3 Discussion

602

This experiment was an extension to Experiment 1. Through collecting denser data per subject, we were able to analyze the cue combination optimality, not only at the population level but also at the individual level. The population-level analysis replicated the results of Experiment 1. The individual-level analysis showed that the distribution of cue combination scores had a unimodally-shaped distribution centered around the optimal combination, thus reflecting genuine cue combination at the individual level. That said, the variance of this distribution indicates that a few participants tended to over-rely on the auditory modality and others tended to over-rely on the visual modality beyond sampling-related randomness.

General Discussion

In the current paper, we explored word recognition under uncertainty about both 603 words and their referents. We conducted an ideal observer analysis of this task whereby a 604 model provided predictions about how information from each modality should be combined 605 in an optimal fashion. The predictions of the model were tested in a series of four experiments where instances of both the form and the meaning were ambiguous with respect to their category membership only (Experiment 1 and 4), when instances of the form were perturbed with additional background noise (Experiment 2), and when instances of the 609 referent were perturbed with additional visual noise (Experiment 3). We discuss the findings 610 of these studies first with respect to our ideal observer model and inferences about optimality 611 and second with respect to their implications for word identification more generally. 612

Patterns of optimality and sub-optimality

In all of our experiments, and when compared to the predictions of the visual or the 614 auditory models, participants generally relied on both modalities to make their decisions in 615 the bimodal condition. Indeed, in Experiment 1 and 2, the optimal model accounted for 616 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 perceptible. Further, in Experiment 1 and 4, 619 which did not involve background noise, participants not only relied on both modalities but 620 generally weighted these modalities according to the predictions of the optimal model, that 621 is, according to their relative reliability. At the individual level, Experiment 4 showed that 622 most participants were near-optimal. Only a few subjects over-relied on the auditory or 623 visual modalities beyond sampling errors. 624

Despite this overall near-optimal behavior, 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: Participants
responded slightly more randomly. This increase in randomness could be due to limitations
on cognitive resources: Processing two separate — and perceptually uncorrelated — cues
instead of one cue (as in the unimodal case) is likely to place extra demands on working
memory, causing general performance to drop (see Mattys & Wiget, 2011).

Previous research has found similar cases of sub-optimal behavior. For instance, studies
that have explored the identification of ambiguous, newly learned pairs of word-referent
associations have reported what appears to be a decrease in speech perception acuity in both
children (Stager & Werker, 1997) and adults (Pajak, Creel, & Levy, 2016). In agreement
with the findings of our study, Hofer and Levy (2017) characterized the apparent reduction in
perceptual acuity as an increase in the noise variance of the auditory modality. Our findings,

besides providing more evidence to this documented fact, suggest that the reduction in perceptual acuity may occur simultaneously in both the auditory *and* the visual modalities.

The second case of sub-optimality is related to how participants weighted the cues 640 from the visual and the auditory modalities in a noisy context. In contrast to Experiment 1 641 and 4 where the cue combination was indistinguishable from the optimal predictions, results 642 of Experiment 2 and 3 suggested that participants had a systematic preference for the 643 non-noisy modality. This finding is reminiscent of the fact that humans tend to compensate 644 for a degraded speech signal by relying more on other sources of information such as the 645 accompanying visual cues, the semantic/syntactic context, or the top-down expectations. 646 This kind of compensation has been observed with adults (Mattys et al., 2012; Tanenhaus et 647 al., 1995), and recent evidence suggests that it starts in childhood (K. MacDonald, Marchman, Fernald, & Frank, 2018; Yurovsky, Case, & Frank, 2017).

Generally speaking, previous experimental studies have not differentiated 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), however. The formal approach followed in this paper allowed us to tease apart these two possibilities, and our 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 participants tend to ignore at least part of the information still available in the noisy modality.

This second case of sub-optimal behavior may be related to the fact that language understanding under degraded conditions is cognitively more taxing than language understanding under normal conditions (Mattys et al., 2012; Peelle, 2018; Rönnberg, Rudner, Lunner, Zekveld, & others, 2010). Perhaps these demands lead to sub-optimal behavior (i.e., over-reliance on the less noisy cue) as participants seek to minimize cognitive effort. One

could also explain this phenomenon in terms of the metacognitive experience about the 664 fluency with which information is processed. The perceived perceptual fluency (e.g., the ease 665 with which a stimulus' physical identity can be identified) can affect a wide variety of human 666 judgments (see Schwarz, 2004 for a review). In particular, variables that improve fluency 667 tend to increase liking/preference (Reber, Winkielman, & Schwarz, 1998). In our case, the 668 subjective experience of lower fluency in the noisy modality might cause people to 669 underestimate information that can be extracted from this modality, especially when 670 presented simultaneously with a higher fluency alternative. 671

Word recognition in the wild

688

An important question to ask is how the combination mechanism — as revealed in our 673 controlled study — scales up to real-life situations. Note that in order to test audio-visual 674 cue combination under uncertainty, we had to use a case of double ambiguity, that is, a case 675 where both the word forms ("ada"-"aba") and the referents (cat-dog) were similar and, thus, 676 confusable. To what extent does such a case occur in real languages? Cross-linguistic corpus 677 analyses suggest that lexical encoding tends, surprisingly, towards double ambiguity in many 678 languages (Dautriche, Mahowald, Gibson, & Piantadosi, 2017; Monaghan, Shillcock, Christiansen, & Kirby, 2014; Tamariz, 2008). For instance, Dautriche et al. (2017) analyzed 100 languages and found that words that are similar phonologically tend to be similar semantically as well. These studies suggest that the case of double uncertainty, though 682 perhaps not pervasive, could be a real issue in language as it increases the probability of 683 confusability for many words. That said, the inferences discussed here might play a more 684 significant role in naturalistic language comprehension when ambiguity in both the form 685 and/or the referent is induced by an external noisy context — e.g., a very noisy party or a far 686 away referent — even when these forms and referents are not confusable in normal situations. 687

Though we only studied adult performance in this paper, the problem of word

recognition under uncertainty is likely more pressing for children. In fact, young children 689 have greater difficulties learning the meanings of novel similar-sounding words (e.g., "bin" vs. 690 "din"), even when these words are uttered very clearly (Creel, 2012; Merriman & Schuster, 691 1991; Stager & Werker, 1997; Swingley, 2016; White & Morgan, 2008). Such similar-sounding 692 words can be shown to be differentiated by infants in simplified experimental settings (e.g., 693 Yoshida, Fennell, Swingley, & Werker, 2009). Nevertheless, Swingley (2007) suggested that 694 the ability to make this differentiation is likely not mature in early childhood; children's 695 representations are almost certainly noisier than the adults' representations and may also be 696 encoded with lower confidence. It is true that the children's lexicon is not as dense in terms 697 of the phonological neighborhood as the adults' lexicon is, but it is still fairly dense (Coady 698 & Aslin, 2003). Thus, children even more than adults might benefit from additional 699 disambiguating cues during new word-referent encoding and recognition.

A multi-modal cue combination strategy might help children not only recognize words, 701 but also refine their underlying phonological and semantic representations in the process. 702 Previous research in early word learning has — whether implicitly or explicitly — largely 703 treated the process of refining the word form and of refining the word meaning as following a 704 linear timeline. However, developmental data reveal that children do not wait to have 705 completed the acquisition of word forms before they start learning their meanings (Bergelson 706 & Swingley, 2012; Tincoff & Jusczyk, 1999). Rather, both form and meaning representations 707 develop in a parallel fashion. A few studies have already suggested the possibility of an 708 interaction between sound and meaning in early acquisition. For instance, Waxman and 709 Markow (1995) showed that labeling various objects with the same name helps infants form the underlying semantic category (but see Sloutsky & Napolitano, 2003). And in the opposite 711 direction, Yeung and Werker (2009) showed that pairing similar sounds with different objects can help infants enhance their sensitivity to subtle phonological contrasts in their native 713 language. The present study offers a stepping stone towards a formal framework where these 714 sorts of sound-meaning interactions in development can be unified and further explored. 715

Note, however, that though our framework allows in principle for the study of how 716 sound-meaning interaction influences the underlying categories, we did not explore this 717 question in the current task. In fact, we used familiar categories both at the semantic level 718 (cat and dog) and at the phonological level (the phonemes p/2 and p/2). This choice allowed 719 us to test the mechanisms of cue integration separately from the question of category 720 learning. Using novel categories would create a confound as participant would additionally 721 have to learn the boundaries of the novel categories, making it difficult for us to tease apart 722 how much of the suboptimal behavior is due to the difficulty of audio-visual integration and 723 how much it is due to the difficulty of estimating the boundaries of the novel categories. 724

Though our task deals primarily with the mechanism of word recognition, the 725 conclusions of the current study could potentially apply to some word learning situations as 726 well. While the typical task in the word learning literature involves mapping a new label to a 727 novel object (e.g., Landau, Smith, & Jones, 1988; Markman, 1991), several cases of word 728 learning in the wild involve familiar objects (as in our task). For example, adults learning a 729 second language do map new labels to objects that are already familiar to them. Besides, 730 children come to the task of word learning with fairly refined knowledge about semantic 731 categories, at least in the case of basic-level object categories (e.g., Quinn, Eimas, & 732 Rosenkrantz, 1993). That is, in many situations, they do not associate new labels with 733 objects that they see for the first time (or for that matter, labels that they hear for the first 734 time). Rather, they often map familiar sequences of phonemes onto concepts that they have 735 refined over the course of earlier exposures (e.g., Bloom, 2000). 736

Our task was an adaptation of a task introduced originally by Sloutsky and Napolitano (2003). This original work is part of a rich literature about modality-dominance in cross-modal processing (Colavita, 1974; Robinson & Sloutsky, 2010). Generally speaking, this literature suggests that, when presented with an audio-visual stimulus, adults pay more attention to the visual component whereas children may pay more attention to the auditory

component (Barnhart, Rivera, & Robinson, 2018; Hirst, Cragg, & Allen, 2018). Our work
shows that adults do not systematically prefer one modality over the other. Rather, they
modulate this preference as a function of noise (Experiments 2 and 3). That said, it is
important to keep in mind that while the literature on modality-dominance deals with
low-level perceptual encoding, our task targets specifically high-level linguistic categorization.
It is possible that the underlying mechanisms in both cases are related (or even similar).
However, as it stands, our work cannot speak to this possibility.

749 Limitations

One salient limitation of our current work is that we used a restricted and highly simplified stimulus set. For the auditory modality, we used speech categories that varied along a single acoustic dimension. While this dimension might be sufficient to recognize words in our specific case, in general the speech signal is far more complex, varying along several acoustic/phonetic dimensions. Additionally, these dimensions may be highly variable due to various kinds of speaker and context differences.

Concerning the visual dimension, simulating meaningful variability has been a 756 notoriously difficult problem. Following previous studies (Freedman et al., 2001; Havy & 757 Waxman, 2016; Sloutsky & Fisher, 2004), we used a visual continuum along a 758 one-dimensional morph. This simplification was motivated by the need to construct a 759 multimodal input where the auditory and visual components are parametrized in a symmetrical fashion, allowing us to compare graded effects of auditory and visual information on categorical judgment. Though such a visual variability is clearly artificial 762 (one does not encounters in real life an animal that is, e.g., 30 % dog and 70 % cat), we 763 assume that the induced uncertainty form these visual stimuli has a similar effect on word 764 recognition as the uncertainty induced by more naturalistic semantic variability. 765

It is an open question whether people use the same strategy in controlled laboratory conditions and more naturalistic settings where they have to deal with various levels of variability. An answer to this question is likely to involve a multifaceted research approach that goes beyond controlled experimentation. We believe that one fruitful approach is to test computational mechanisms with an input that more accurately represents the full extent of multimodal variability in the learning environment (Dupoux, 2018; Fourtassi, Schatz, Varadarajan, & Dupoux, 2014; Harwath, Torralba, & Glass, 2016; B. C. Roy, Frank, DeCamp, & Roy, 2015).

Finally, though we used the term "word recognition," our work is only indirectly related to the literature about how a rich lexicon is accessed (e.g., McClelland & Elman, 1986). We have used this term in a more specific way, describing access to a simplified lexicon made of two novel, ambiguous words. Such a simplified experimental context is not new and has been crucial to our understanding of early word learning and recognition (e.g., Stager & Werker, 1997).

Conclusions and future research directions

780

Our work used an ideal observer model to study word recognition under audio-visual uncertainty. This framework enabled us not only to test optimality but also to examine systematically how and by how much people deviate from optimality in their combination strategies. Thus, our work is part of a growing effort to go beyond optimality tests — which have limited explanatory power — and use models that also allow us to identify and explain various patterns of sub-optimality in human behavior (Rahnev & Denison, 2018).

While we focused on the case of arbitrary associations in novel word recognition, it is possible to extend the framework to other cases such as that of *iconicity*, i.e., when there is a resemblance between the sound of a word and its referent. Previous work has suggested that

iconicity, among other things, helps with learning (and generalizing the meaning of) new words (see Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015 for a review). Using the research strategy in this paper, we can, for example, test whether iconicity has such an advantage because it mitigates the sub-optimal patterns observed with more arbitrary pairings.

Finally, though the current framework only characterizes adult novel word recognition,
it provides a first step towards a model where developmental questions can also be
investigated. For instance, future work should explore whether children, like adults, use
probabilistic cues from both the auditory and the visual input to recognize ambiguous words,
the extent to which they combine these cues in an optimal fashion, and whether this cue
combination helps them learn words and refine their early phonological and semantic
representations.

All data and code for these analyses are available at https://github.com/afourtassi/WordRec

802

803

804

805

Acknowledgements

This work was supported by a post-doctoral grant from the Fyssen Foundation.

Disclosure statement

None of the authors have any financial interest or a conflict of interest regarding this work and this submission.

Appendix 1: 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:

808

811

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

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:

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

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 obtain this final expression for the posterior:

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

819 where

820

821

822

$$1 + b = \frac{p(W_1)}{p(W_2)}$$
$$\beta_0 = \frac{\mu_{A2}^2 - \mu_{A1}^2}{2\sigma_A^2} + \frac{\mu_{V2}^2 - \mu_{V1}^2}{2\sigma_V^2}$$

$$\beta_a = \frac{\mu_{A1} - \mu_{A2}}{\sigma_A^2}$$
$$\beta_v = \frac{\mu_{V1} - \mu_{V2}}{\sigma_V^2}.$$

References

- 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.
- Barnhart, W. R., Rivera, S., & Robinson, C. W. (2018). Different patterns of modality dominance across development. *Acta Psychologica*, 182.
- 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.
- Bergelson, E., & Swingley, D. (2012). At 6 to 9 months, human infants know the meanings

- of many common nouns. Proceedings of the National Academy of Sciences, 109(9).
- Bloom, P. (2000). How children learn the meanings of words. Cambridge, MA. MIT Press.
- ⁸³⁴ 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.
- 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.
- Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon.

 Journal of Child Language, 30(2).
- Colavita, F. B. (1974). Human sensory dominance. Attention, Perception, & Psychophysics, 16(2), 409–412.
- 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.
- Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015).

 Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*,

 19(10), 603–615.
- Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for

- reverse-engineering the infant language-learner. Cognition, 173, 43–59.
- 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.
- Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues.

 **Cognition, 143, 93–100.
- 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.
- Fourtassi, A., Schatz, T., Varadarajan, B., & Dupoux, E. (2014). Exploring the relative role of bottom-up and top-down information in phoneme learning. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (Vol. 2, pp. 1–6).
- 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.
- ⁸⁷³ Greenberg, J. (1957). Essays in linguistics. Chicago: University of Chicago Press.
- Harwath, D., Torralba, A., & Glass, J. (2016). Unsupervised learning of spoken language with visual context. In *Advances in neural information processing systems* (pp.

- 1858–1866).
- Havy, M., & Waxman, S. (2016). Naming influences 9-month-olds' identification of discrete categories along a perceptual continuum. *Cognition*, 156, 41–51.
- 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.
- Hirst, R. J., Cragg, L., & Allen, H. A. (2018). Vision dominates audition in adults but not children: A meta-analysis of the colavita effect. Neuroscience & Biobehavioral

 Reviews, 94.
- 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.
- Kronrod, Y., Coppess, E., & Feldman, N. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, 23(6).
- Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2), 93–107.
- Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning.

 Cognitive Development, 3.
- Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of

- concepts by verbal and nonverbal means. Journal of Experimental Psychology:

 General, 141(1), 170.
- MacDonald, K., Marchman, V., Fernald, A., & Frank, M. C. (2018). Adults and preschoolers
 seek visual information to support language comprehension in noisy environments. In

 Proceedings of the 40th Annual Conference of the Cognitive Science Society.
- Markman, E. M. (1991). The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In S. A. Gelman & J. P. Byrnes (Eds.),.
- 905 Marr, D. (1982). *Vision*. WH Freeman.
- 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.
- McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, 18(1).
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746–748.
- Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22), 9014.
- 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).
- Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, 39(2), 204.
- Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure.

 Cambridge, MA: MIT press.
- Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants.

 Perception, 22(4), 463–475.
- Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making.

 Behavioral and Brain Sciences, 1–107.
- 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. (2010). Development of cross-modal processing. Wiley

 Interdisciplinary Reviews: Cognitive Science, 1.
- Roy, B. C., Frank, M. C., DeCamp, M., P., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112.
- Rönnberg, J., Rudner, M., Lunner, T., Zekveld, A. A., & others. (2010). When cognition

- kicks in: Working memory and speech understanding in noise. *Noise and Health*, 12(49), 263.
- 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., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, 133(2), 166.
- Sloutsky, V., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development*, 74.
- Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568.
- 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).
- Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word
 learning in young children. Journal of Experimental Child Psychology, 126.
- Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds.

 Developmental Psychology, 43(2), 454–464.
- 962 Swingley, 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 representations of the mental lexicon. *The Mental Lexicon*, 3(2).
- Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268 (5217), 1632–1634.
- Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind:

 Statistics, structure, and abstraction. *Science*, 331 (6022), 1279–1285.
- Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10(2), 172–175.
- Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, 127.
- Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. Cognition, 107(2), 729–742.
- Vouloumanos, A., & Waxman, S. (2014). Listen up! Speech is for thinking during infancy.

 Trends in Cognitive Sciences, 18(12), 642–646.
- Vroomen, J., Linden, S. van, Keetels, M., Gelder, B. de, & Bertelson, P. (2004). Selective
 adaptation and recalibration of auditory speech by lipread information: Dissipation.

 Speech Communication, 44.
- Waxman, S., & Gelman, S. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, 13(6).
- Waxman, S., & Markow, D. (1995). Words as invitations to form categories: Evidence from

- 12-to 13-month-old infants. Cognitive Psychology, 29(3), 257–302.
- White, K., & Morgan, J. (2008). Sub-segmental detail in early lexical representations.
- Journal of Memory and Language, 59.
- Yeung, H., & Werker, J. (2009). Learning words' sounds before learning how words sound:
- 9-month-olds use distinct objects as cues to categorize speech information. Cognition,
- 990 113, 234–243.
- Yoshida, K., Fennell, C., Swingley, D., & Werker, J. (2009). 14-month-olds learn
- similar-sounding words. Developmental Science, 12.
- 993 Yurovsky, D., & Frank, M. C. (2015). An Integrative Account of Constraints on
- 994 Cross-Situational Learning. Cognition, 145.
- ⁹⁹⁵ Yurovsky, D., Case, S., & Frank, M. C. (2017). Preschoolers flexibly adapt to linguistic input
- in a noisy channel. Psychological Science, 28(1), 132–140.