Word Identification Under Multimodal Uncertainty

Abdellah Fourtassi
afourtas@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Identifying the visual referent of a spoken word – that a particular insect is referred to by the word "bee" – 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 optimal baseline. Model predictions are tested in two experiments where word recognition is made under two kinds of uncertainty: category ambiguity and distorting noise. In both cases, the ideal observer model explains much of the variance in human judgments. But when one modality had noise added to it, human perceivers systematically preferred the unperturbed modality to a greater extent than the ideal observer model did.

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

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 honeybee buzzes around the food. Determining reference involves processing the auditory information and linking it with 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).

This multimodal integration task takes place in a noisy world. 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 (e.g., 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.

Thus, establishing reference requires reasoning under a great deal of uncertainty in both modalities. The goal of this work is to characterize such reasoning. Imagine, for example, that someone is uncertain whether they heard "pea" or "bee", does it make them rely more on the visual modality (e.g., the object being pointed at)? Vice versa, if they are not sure if they saw a bee or a fly, does that make them rely more on the auditory modality (i.e., the label)? More importantly, when input in both modalities is uncertain to varying degrees, do they weight each modality according to its relative reliability, or do they over-rely on a particular modality?

In this paper, we propose a probabilistic framework where such reasoning can be expressed precisely. We characterize uncertainty in each modality with a probability distribution, and we predict ideal responses by combining these probabilities in a optimal way. Our work can be seen as an extension to previous Bayesian models of phoneme identification (e.g., Feldman, Griffiths, & Morgan, 2009), where, instead of a unimodal input, we model a bimodal one. A few studies have explored some aspects of audio-visual processing in a probabilistic framework (e.g., Bejjanki, Clayards, Knill, & Aslin, 2011). In these studies, the researchers focused on the specific case of phoneme recognition from speech and lip movement, however, where information is tightly correlated across modalities.

In the present work, we study, rather, the case of arbitrary mapping between sounds and visual objects. We test participants on their ability to process audio-visual stimuli and use them to recognize the underlying word. More precisely we study the case where both the word's form and the word's referent are ambiguous, and we examine the extent to which humans conform to, or deviate from the predictions of the ideal observer model. Moreover, some previous studies on audio-visual processing documented cases of modality preference, when people rely predominantly on the visual modality (e.g., Colavita, 1974) or the auditory modality (e.g., Sloutsky & Napolitano, 2003). Thus, we will explore if participants in our task show evidence of a modality preference.

The paper is organized as follows. First, we introduce our audio-visual recognition task. We next present the ideal observer model. Then we present two behavioral experiments where we test word recognition under audio-visual uncertainty. In Experiment 1, audio-visual tokens are ambiguous with respect to their category membership. In Experiment 2, we intervene by adding noise to one modality. In both experiments participants show qualitative patterns of optimal behavior. Moreover, while participants show no modality preference in Experiment 1, in Experiment 2 they over-rely on visual input when the auditory modality is noisy.

The Audio-Visual Word Recognition Task

We introduce a new task that tests audio-visual word recognition. 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/).

[1] In 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.

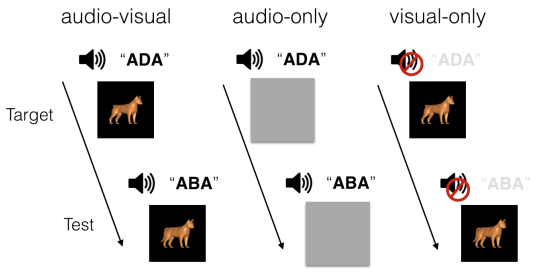

Figure 1: Overview of the task

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 research to probe audio-visual encoding. However, unlike this previous line of research, here participants are not asked whether the two audio-visual presentations are identical. Instead, the task is category-based: They 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 word, in the object, 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 were hidden (audio-only) or where sounds were muted (visual-only). These unimodal trials provide us with participants' categorization functions for the auditory and visual categories and are used as inputs to the ideal observer model, described below.

Ideal Observer Model

The basis of our ideal observer model is that individual categorization functions from each modality should be combined optimally. In each modality, we have two categories: /ada/ ($A = 1$) and /aba/ ($A = 2$) in the auditory dimension, and *cat* ($V = 1$) and *dog* ($V = 2$) in the visual dimension. We assume, for the sake of simplicity, that the probability of membership in each category is normally distributed:

$$p(a|A) \sim N(\mu_A, \sigma_A^2)$$
$$p(v|V) \sim N(\mu_V, \sigma_V^2)$$

In the bimodal condition, participants see word tokens with audio-visual input, and have to make a categorization decision. We define word tokens as vectors in the audio-visual space, $\mathbf{w} = (a, v)$. A word category W is defined as the joint distribution of auditory and visual categories. It can be characterized with a bivariate normal distribution:

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

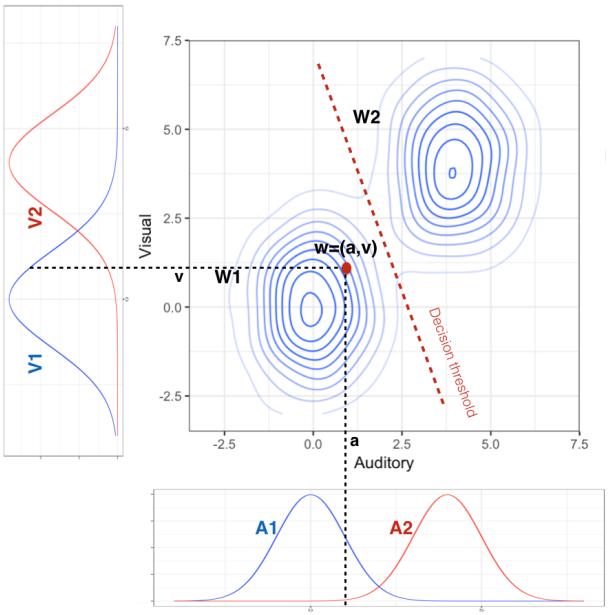

Figure 2: Illustration of our model using simulated data. A word category is defined as the joint bivariate distribution of an auditory category (horizontal, bottom panel) and a visual semantic category (vertical, left panel). Upon the presentation of a word token w, participants guess whether it is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

We have two word categories: dog-/aba/ (W_1) and cat-/ada/ (W_2). 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 $\mathbf{w} = (a, v)$ is the posterior probability of this category:

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

We make the assumption that, given a particular word category, the auditory and visual tokens are independent:

$$p(\mathbf{w}|W) = p(a, v|W) = p(a|W)p(v|W) \quad (2)$$

Under this assumption, the posterior probability reduces to:

$$p(W_2|\mathbf{w}) = \frac{1}{1 + (1+\varepsilon)\exp(\beta_0 + \beta_a a + \beta_v v)} \quad (3)$$

with $\beta_a = \frac{\mu_{A1} - \mu_{A2}}{\sigma_A^2}$, $\beta_v = \frac{\mu_{V1} - \mu_{V2}}{\sigma_V^2}$, $\beta_0 = \frac{\mu_{A2}^2 - \mu_{A1}^2}{2\sigma_A^2} + \frac{\mu_{V2}^2 - \mu_{V1}^2}{2\sigma_V^2}$ and $1 + \varepsilon = \frac{p(W_1)}{p(W_2)}$ is the proportion of the prior probabilities. If the identity of word categories is randomized, and if W_1 is the target, then ε measures a response bias to "same" if $\varepsilon > 0$, and a response bias to "different" if $\varepsilon < 0$.

In sum, the posterior 3 provides the ideal observer's predictions for how probabilities that characterize uncertainty in

each modality can be combined to make categorical decision about bimodal input.

Experiment 1

In Experiment 1, we test the predictions of the model in the case where uncertainty is due to similar auditory categories, and similar visual categories. Crucially, the similarity is such that the distributions overlap. To simulate such uncertainty in a controlled fashion, we use a continuum along the second formant (F2) linking the words /aba/ and /ada/, and we use a morph that links a dog prototype and a cat prototype.

Methods

Participants We recruited a planned sample of 100 participants, recruited 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. Data were excluded if participants completed the task more than once (2 participants). Moreover, as specified in the preregistration (`https://osf.io/h7mzp/`), participants were excluded if they had less than 50% accurate responses on the unambiguous training trials (6), and if they reported having experienced a technical problem of any sort during the online experiment (14). The final sample consisted of 76 participants.

Stimuli For auditory stimuli, we used the continuum introduced in Vroomen, van Linden, Keetels, de 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 end-points (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 corresponding sound. 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 (∼800ms). 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/. For each participant, we randomized the sound-object mapping as well as the identity of the target.

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.

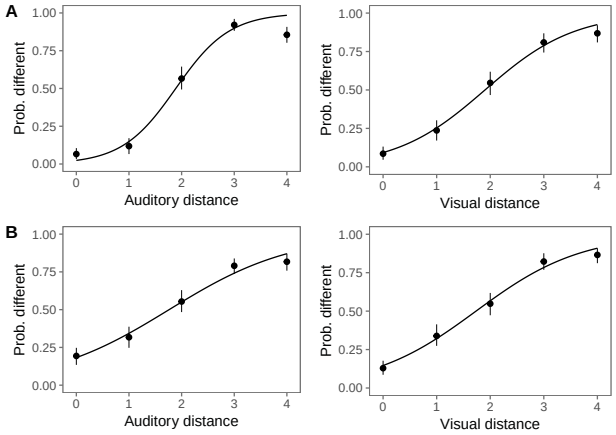

Figure 3: Average human responses in the auditory-only condition (left), and visual-only condition (right). A) represents data from Experiment 1, and B) data from Experiment 2. Error bars are 95% confidence intervals. Solid lines represent unimodal logistic fits.

Results

Unimodal conditions this is the case where the pictures were hidden, or where the sounds were muted. Average categorization judgments and fits are shown in Figure (3, A). The categorization function of the auditory condition was steeper than that of the visual condition. The fit was done using the Nonlinear Least Squares (NLS) R package, as follows. For an ideal recognizer, the probability of choosing category 2 (that is, to answer "different") when presented with an audio instance a, is the posterior probability of this category $p(A_2|a)$. If we assume that both categories have equal variances, the posterior probability reduces to:

$$p(A_2|a) = \frac{1}{1+(1+\varepsilon_A)\exp(\beta_{a0}+\beta_a a)} \tag{4}$$

with $\beta_a = \frac{\mu_{A_1}-\mu_{A_2}}{\sigma_A^2}$ and $\beta_{a0} = \frac{\mu_{A_2}^2-\mu_{A_1}^2}{2\sigma_A^2}$. ε_A is the response bias in the auditory-only trials.

For this model (as well all other models), we fixed the values of the means to be the end-points of the corresponding continuum: $\mu_{A1} = 0$ and $\mu_{A2} = 4$ (and similarly $\mu_{V1} = 0$, and $\mu_{V2} = 4$). To determine the values of the bias and the variance, we fit a model for each modality, collapsed across participants. For the auditory modality, we obtained $\varepsilon_A = -0.20$ and

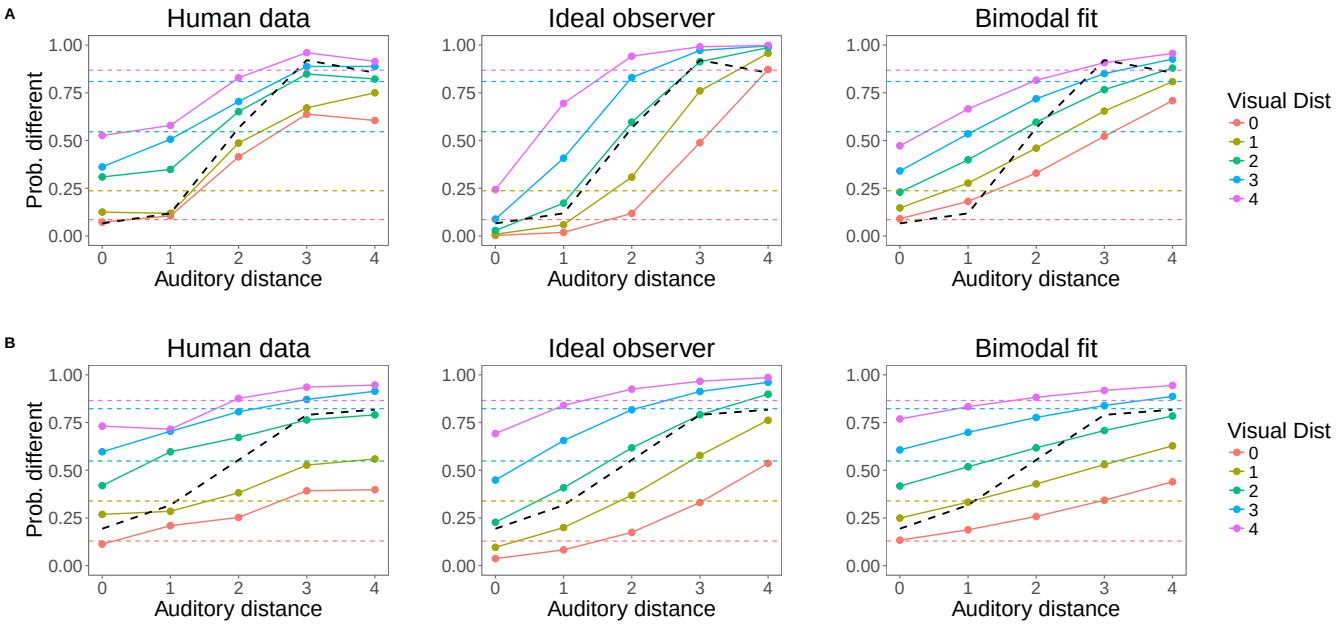

Figure 4: Proportion of "different" judgments as a function of auditory distance. Solid lines represent average human responses (left), predictions of the ideal observer (middle), and the bimodal fit (right). Dashed lines represent average human responses in the unimodal conditions. Colors represent values in the visual continuum. A) represents data from Experiment 1, and B) data from Experiment 2.

$\sigma_A^2 = 2.04$. For the visual modality, we obtained $\varepsilon_V = -0.11$ and $\sigma_V^2 = 3.34$.

Bimodal condition We fit a model to human responses in the bimodal condition, collapsed across participants, finding $\varepsilon = -0.32$, $\sigma_{Ab}^2 = 5.00$ and $\sigma_{Vb}^2 = 7.27$. The fit explained 94% of total variance.

Ideal observer model We derived the predictions of the ideal observer model by using the values of the variances derived from the unimodal conditions, and the response bias derived from the bimodal condition, and by substituting these values into the expression of the posterior in Eq. 3. Figure (4, A) shows participants' responses in the bimodal condition (left), as well as the prediction of the ideal observer (middle), and the bimodal fit models (right).

Response bias We found negative values in all response bias terms, which suggests a general bias toward answering "different." This bias is probably due to the categorical nature of our same-different task: when two items are ambiguous but perceptually different, this could cause a slight preference for "different" over "same".

Modality preferences We next analyzed whether there was a preference for one or the other modality when making decisions in the bimodal condition, beyond that explained by the variance in categories implied by the unimodal responses. This preference would manifest as a deviation from the decision threshold predicted by the ideal observer model. The decision threshold is defined as the set of values in the audio-visual space along which the posterior (Eq. 3) is equal to 0.5. The decision threshold takes the following form:

$$v = -\frac{\sigma_V^2}{\sigma_A^2} a + v_0 \tag{5}$$

If the slope derived from the bimodal fit is greater than the slope of the ideal observer, this finding would suggest a general preference for the auditory modality (similarly, a smaller slope would suggest 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). Figure 5 (top left) shows the decision threshold in the audio-visual space with a constant intercept; the fit to human data (solid black line) was very close to the ideal observer threshold (red line). Non-parametric resampling of the data showed no evidence of a deviation from the slope of the ideal observer (5, bottom left).

Discussion

Qualitatively, participants' judgments were similar to the predictions of the ideal observer model (remember that the latter was obtained by optimally combining fits to the unimodal data). Consider, for example, the contrast between the auditory-only case (dashed black line in Figure 4, top left) and the bimodal case (solid colored lines). Higher certainty in the visual modality generally influenced responses, with greater visual distance leading to more "different" ratings and less visual distance leading to more "same" ratings. Similar ob-

servations can be made about the contrast between the visual-only case and the bimodal case.

Overall, we found that the ideal observer model explained much of the variance in judgments ($r^2 = 0.89$). But although we see a qualitative resemblance between human data and the model, there were quantitative differences. For example, model predictions were more influenced by the visual modality at the auditory midpoint (the point of highest uncertainty) than human judgements, and were more compressed at the endpoints (the points of lowest uncertainty).

Formally, there was an increase in the value of the variance associated with each modality. Whereas the ideal observer model predicted the weights to be proportional to $1/\sigma_A^2$ and $1/\sigma_V^2$, for the auditory and the visual modalities, respectively (see expression 3), the fit to human data suggested that the real weights were proportional to $1/\sigma_{Ab}^2$ and $1/\sigma_{Vb}^2$. Our analysis of modality preference showed that the relative values of these variances were almost the same (Figure 5, left). Thus, 1) the bimodal presentation introduced a certain level of randomness in the participants' responses, and 2) this increased randomness did not affect the relative weighting of both modalities, i.e., participants were weighting modalities according to their relative reliability. The latter explains the qualitative resemblance between the predictions of the ideal observer and human data, and the former explains the quantitative discrepancy.

In sum, we found that participants followed the ideal observer model in that they weighted modalities according to their reliabilities. In real life, however, tokens can undergo distortions due to noisy factors in the environment. In Experiment 2, we explore this additional level of uncertainty.

Experiment 2

Imagine that the speaker generates a target production t from an auditory category $t|A \sim N(\mu_A, \sigma_A^2)$. In Experiment 1, we assumed that the observer could directly retrieve this production token. But if the observer is in a noisy environment, they do not hear exactly this produced target, but the target perturbed by noise, which we assume, following Feldman et al. (2009), that it is normally distributed: $a|t \sim N(t, \sigma_N^2)$. When we integrate over t, we get:

$$a|A \sim N(\mu_A, \sigma_A^2 + \sigma_N^2) \tag{6}$$

In this experiment, we explored the effect of this added noise[2] on performance in our task. We tested a case where one modality was ambiguous and noisy (auditory), and where the other modality was ambiguous but not noisy (visual). We were interested to know if participants would treat this new source of uncertainty as predicted by the ideal observer model, and whether noise in one modality would lead to some systematic preference for the non-noisy modality.

[2]Note that we are considering environmental noise, which is different from the noise inherent to perception.

Methods

Participants A planned sample of 100 participants was recruited online through Amazon Mechanical Turk. We used the same exclusion criteria as in the previous experiment; the final sample consisted of 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 audio editor Audacity (2.1.2). The procedure was exactly the same as in the previous experiment, except that test stimuli were presented with the new noisy auditory stimuli.

Results

Unimodal conditions We fit a model for each modality, collapsed across participants. For the auditory modality, our parameter estimates were $\varepsilon_A = -0.18$ and $\sigma_A^2 + \sigma_N^2 = 4.70$. For the visual modality, we found $\varepsilon_V = -0.24$ and $\sigma_V^2 = 3.93$. Figure 3 (bottom) shows responses in the unimodal conditions as well as the unimodal fits. In contrast to Experiment 1, auditory responses were flatter (showing more uncertainty).

Bimodal condition We fit a model to human responses in the bimodal condition, collapsed across participants. We estimated $\varepsilon = -0.38$, $\sigma_{Vb}^2 = 5.21$, and $\sigma_{Ab}^2 + \sigma_{Nb}^2 = 9.84$. The fit explained 97% of total variance.

ideal observer model We generated the predictions of the ideal observer model by using the values of the variances derived from the unimodal conditions, and the response bias derived from the bimodal condition, and by substituting these values into the expression of the posterior in Eq. 3. Results are shown in Figure 4 (bottom).

Modality preferences Participants' decision threshold suggested a preference for the visual modality (the non-noisy modality). Indeed non-parametric resampling of the data showed a decrease in the value of the slope (5, right).

Discussion

We found, similar to Experiment 1, that participants generally showed qualitative patterns similar to the ideal observer model ($r^2 = .91$). But we also found a similar discrepancy at the quantitative level. The ideal observer model predicted the modality weights to be proportional to $1/(\sigma_A^2 + \sigma_N^2)$ and $1/\sigma_V^2$, for the auditory and the visual modalities, respectively. The fit to human data suggested that the empirical weights were proportional to $1/\sigma_{Ab}^2$ and $1/\sigma_{Vb}^2$. An interesting difference with Experiment 1, however, was that participants had a clear preference for the non-noisy modality, as the values of the relative variances were different (Figure 5, right). This preference affected the relative weighting, where, contrary to Experiment 1, the visual modality had greater weight than what could be expected from its relative reliability alone.

It is important to understand that this preference was not the mere consequence of the added noise increasing the variance of the auditory modality, since this increase was already

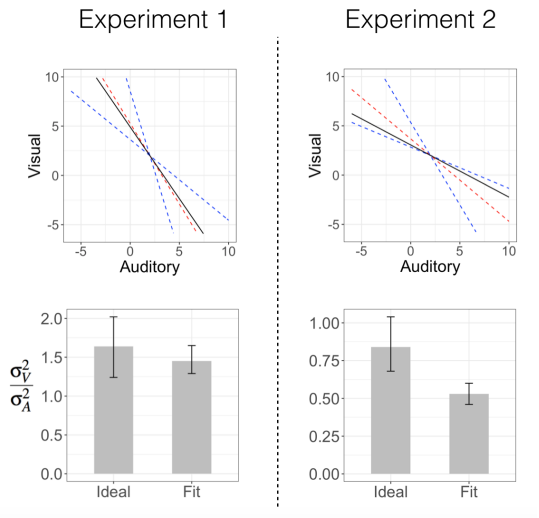

Figure 5: Top: decision thresholds in the audio-visual space. Red dotted line is the prediction of the ideal observer. Blue dotted lines are cases where modality preference is twice as strong as the ideal observer. Solid line is the threshold derived from human data. Bottom: comparison of the threshold slope between the ideal observer and the fit to human data. Error bars are 95% confidence intervals computed via non-parametric bootstrap.

accounted for in the ideal observer model. The preference was, rather, a form of over-reliance on the visual modality.

General Discussion

Understanding language requires both the ability to integrate multimodal input, and the ability to deal with uncertainty. In this work, we explored a case where both abilities were at play. We studied the case of identifying a word when both its form (auditory) and its referent (visual) were ambiguous with respect to their category membership (Experiment 1), and when the form was perturbed with additional noise (Experiment 2). We introduced a model that instantiated an ideal observer, predicting how information from each modality could be combined in an optimal way. In both experiments, participants showed the qualitative patterns of the ideal observer.

There were, however, quantitative differences. Audio-visual presentation increased the level of randomness in the participants' responses. One possible explanation is that this phenomenon was caused by the arbitrary nature of the form-meaning mapping. Previous studies suggest that while redundant multimodal information improves performance (e.g., determining the frequency of a bouncing ball from visual and auditory cues), arbitrary mappings generally tends to hinder performance (for review, see Robinson & Sloutsky, 2010).

Interestingly, however, in Experiment 1 this increase in randomness occurred at a similar rate for both the auditory and the visual modality, and thus, it did not affect their relative weighting. The latter was primarily determined by informational reliability. Only when we intervened by adding noise to one modality in Experiment 2, did participants show a systematic preference for the non-noisy modality. One possible explanation for this preference could be that people do not combine cross-modal uncertainties of a similar kind (e.g., ambiguity in both modalities) in the same way they would combine uncertainties of different kinds (e.g., ambiguity in one modality and noise in the other). For instance, it could be that the latter, but not the former, cause the over-reliance on a particular modality.

Overall, in both Experiments, the majority of the variance could be explained by an ideal observer that combined multimodal information optimally. In the light of this main result, we can revisit some previous findings in the literature. For instance, Sloutsky and Napolitano (2003) reported a dominance for the auditory modality in children. This dominance disappears or reverses in adults. Could this difference be driven by changes across development in the level of perceptual noise affecting the intrinsic relative reliability of modalities (by analogy to Experiment 2)? More work is needed to carefully examine this (and other) speculations, and more generally, to determine the extent to which the optimal combination account helps us better understand the mechanisms of word processing and learning.

Acknowledgements

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

References

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLoS ONE*, *6*.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, *16*.

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.

Freedman, D., Riesenhuber, M., Poggio, T., , & Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*.

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*.

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

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

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