Modeling Multistable Auditory Perception: Bayesian Insights from the Viral Yanny-Laurel Phenomenon

Clara Chen
clara_chen@college.harvard.edu, clarache@mit.edu
MIT 9.66

Introduction

In 2018, a viral audio clip took the internet by storm, after a group of high school students studying for a vocabulary test were split over hearing either "Yanny" or "Laurel" in the same recording for "laurel" on vocabulary.com. The Yanny-Laurel clip was deemed the audio version of #TheDress (Figure 1), a viral image in 2015 of a dress that appeared to some as black and blue and to others as white and gold. In both cases, users on internet were equally divided, and each side was entirely confident in what they perceived, completely unable to understand how the other side could see or hear something different. This differs from other types of illusions, often visual illusions, where after looking at the stimulus long enough, the alternative image can be perceived as well. For example, in the famous illusion Rubin's face/vase, people are typically able to see both the vase and face after viewing the image for long enough (Figure 2). While viral phenomenons such as Yanny/Laurel and #TheDress are fun and unserious, studying their multi-stability helps us to better understand how humans respond to uncertainty in sensory stimulus.

Figure 1: The Dress

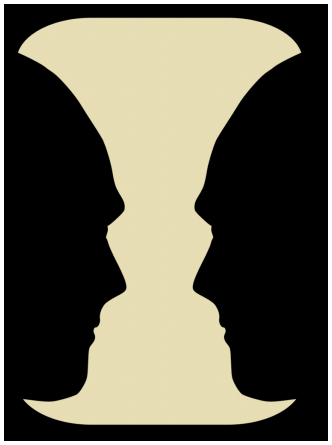


Figure 2: Rubin's vase/face

The most common explanation for the ambiguity of the Yanny-Laurel audio file is whether your ear focuses more on high or low frequencies (Matsakis, 2018). Those who hear Laurel are likely hearing more of the low frequencies in the audio, while those who hear Yanny are likely hearing the higher ones. Other factors that relate to frequency perception also play a role. Older people are typically more attuned to lower frequencies and may hear "Laural" more easily. Different devices ranging from phones and laptops to high quality speakers may emphasize certain frequencies over others, and furthermore, different websites may also compress audio files differently, leading to even more discrepancies.

The New York Times released a tool where you can listen to morphed versions of the original recording with either the high or low frequencies filtered to varying degrees, allowing you to hear the recording on a continuum from Laurel to Yanny (Katz, Corum, & Huang, 2018). In experimenting with this tool, one will find that what you hear depends not only on the absolute level of frequency modification but also the relative frequency modification compared to what was just heard previously. For instance, when the frequency is gradually morphed from "Laurel" to "Yanny," the perception is more likely to stay with "Laurel" for high frequencies than if the frequency had been changed suddenly. In other words, the interpretation of what one hears depends on the initial starting point.

This study will explore multi-stability in audio perception

by modeling the likelihood of people hearing Yanny or Laurel in morphed versions of the audio clip as well as the effect of audio priming. Taking inspiration from studies on multistable vision perception, we will apply Bayesian methods to model human perceptions of the Yanny-Laurel audio controversy from a probabilistic perspective.

Related Works

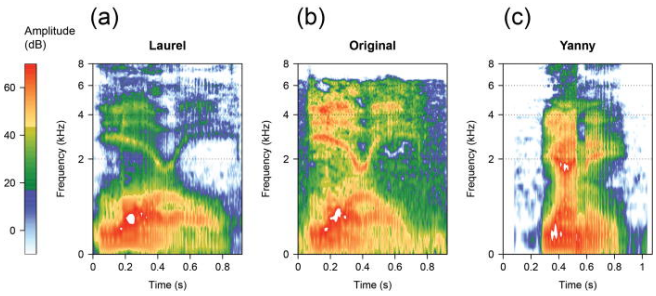


Figure 3: Spectrograms of (a) the recording from vocabulary.com, (b) the viral audio clip, and (c) a simulated "Yanny" audio from Bosker (2018).

Several studies and media articles have compared and morphed recordings of Laurel and Yanny to qualitatively demonstrate the role of frequency in people's perception of the original viral recording (Katz et al., 2018; Bosker, 2018; Pressnitzer, Graves, Chambers, de Gardelle, & Egré, 2018). Using spectrograms, one can see that recordings of Laurel have more lower frequencies present while recordings of Yanny have more higher frequencies present (Figure 3). In the original viral recording, both low and high frequencies are present, which elucidates why it was possible for people to hear both Yanny and Laurel in that audio file. Through audio processing, such studies have been able to create audio clips with different frequency compositions that gradually range from "Laurel" to "Yanny" (Figure 4). Low-pass filters are used on the audio file to create clips that sound more "Laurel-like" while high-pass filters are used to create clips that sound more "Yanny-like." Bosker (2018) also used low and high pass filters to create an ambiguous Harry-Megan audio file, similar to the Yanny-Laurel phenomenon, demonstrating how ambiguous audio files can be retroactively made by editing and combining frequencies appropriately.

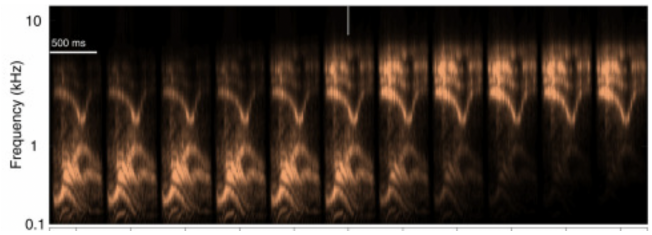


Figure 4: Spectrograms of morphed audio files from Pressnitzer et al. (2018).

In addition to the role of frequency in audio perception, academic studies of the Laural and Yanny controversy have uncovered other interesting aspects of human auditory perception. Pressnitzer et al. (2018) surveyed 289 online participants on what they heard in a range of morphed audio files and their confidence level about the word they heard in each clip. Although a large proportion of participants heard the same word regardless of the audio morphing, most participants were more likely to hear Laurel for the low-pass files and Yanny for the high-pass files as expected. However, regardless of whether participants' perceptions flipped based on audio morphing, they were all extremely confident in what they heard for each recording.

Bosker (2018) also conducted an online survey (N=532) asking people to listen to the Yanny-Laurel recording. In addition to morphing the audio files, they also randomly assigned participants to a priming condition. Before listening to the morphed Laurel-Yanny file, they would listen to an unrelated recording that was either put through a low-pass filter or a high-pass filter. Similar to Pressnitzer et al. (2018), Bosker (2018) found that a large portion of participants were very stable in their auditory perception, hearing either Yanny or Laurel over 90% of the time regardless of audio morphing. Overall, however, those who were primed with the low-pass filter were statistically significantly more likely to hear "Yanny" for every level of audio morphing compared to those who were primed with the low-pass filter (Figure 5). This suggests that perception of ambiguous audio highly depends on previously heard audio in addition to the objective adjustments made to the current audio. In particular, the authors highlight the impact of spectral contrast effects. Literature on spectral contrast suggests that the ear attunes itself to novelty. After hearing a lower frequency recording, for example, the ear is primed to hear higher frequencies. This explains why after hearing a low-pass filtered recording, which emphasizes lower frequencies, people were more likely to hear "Yanny" from the higher frequencies of the recording.

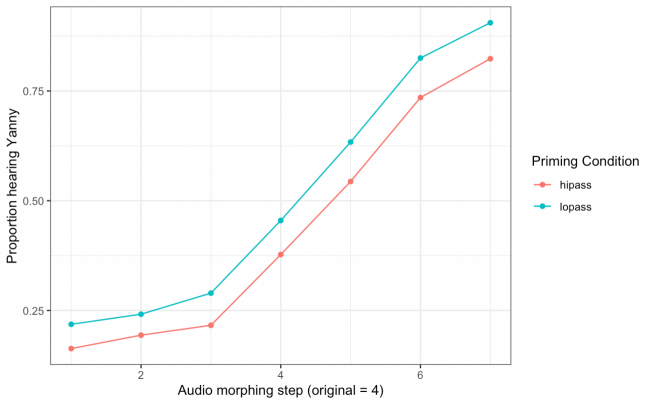


Figure 5: Rate of hearing Yanny based on audio morphing and priming. Figure created using data from Bosker (2018).

The two aforementioned studies both use standard logistic regression to model their survey results. However, studies on multi-stable vision perception suggest that probabilistic methods or a Bayesian framework may be more appropriate for modeling human sensory perception in uncertain settings. Many studies on multistability and color constancy, including those on #TheDress, use probabilistic models such as graphical models and Bayesian nets to model vision perception (Gershman, Vul, & Tenenbaum, 2012; Chandra, Li, Tenenbaum, & Ragan-Kelley, 2022; Brainard et al., 2006, Lafer-Sousa, Hermann, & Conway, 2015, Lafer-Sousa & Conway, 2017). Compared to multi-stable visual perception, multi-stable audio perception is not as well-studied, especially when it comes to specific "illusion" such as Yanny-Laurel or #TheDress, but there is some research suggesting that Bayesian frameworks may work for auditory scenes as well (Elhilali, 2013).

Gershman et al. (2012) and Chandra et al. (2022) also suggest parallels between MCMC sampling and human perception. Given that perceptions of the audio file seem to depend not only on absolute frequency but also audio priming, the MCMC sampling structure may apply to Yanny and Laurel as well since the interpretation of what you hear depends on the initial starting point. Taking inspiration from studies on inference in visual perception, this study aims to expend previous studies on the Yanny-Laurel controversy.

Methods

All code files and data files can be found at `https://github.com/cchen2125/yanny-laurel-analysis`

Bosker Data Analysis

The data collected from the survey by Bosker (2018) is publicly available at `https://osf.io/63wdh/`. This data was used to create two probabilistic models, a graphical Bayes net model and a hierarchical logistic regression model.

Table 1 lists the relevant variables from the Bosker (2018) data. In Bosker (2018), participants were first primed with either a low-pass filtered recording of a voice reading a phone number or a high-pass filtered recording of a phone number, represented by the `cond` variable. Then, they would hear a morphed Yanny-Laurel audio clip. The original recording was defined as step 4. Lower steps 1-3 sound more Laurel-like while higher steps 5-7 sound more Yanny-like. Participants were presented with each combination of condition and step multiple times.

Table 1: Bosker (2018) Data Description

Variable	Description
participantID	Participant uique identifier
cond	Low pass or high pass priming
step	Audio-modification level
yannyresp	Indicator of hearing Yanny

A new variable `bias` was derived from the existing data. A participant's bias was defined as the proportion of times they heard "Yanny" or "Laurel" when listening to the original audio (step=4). This variable was later used in the Bayes net to capture the inherent bias that an individual might have toward hearing Laurel over Yanny or vice versa. As seen in Figure 6, the probabilities of hearing Yanny for Laurel-biased individuals and Yanny-biased individuals are quite different. An individual's bias seems to play a stronger role on the probability of hearing "Yanny" than the audio priming, making this an important variable to consider.

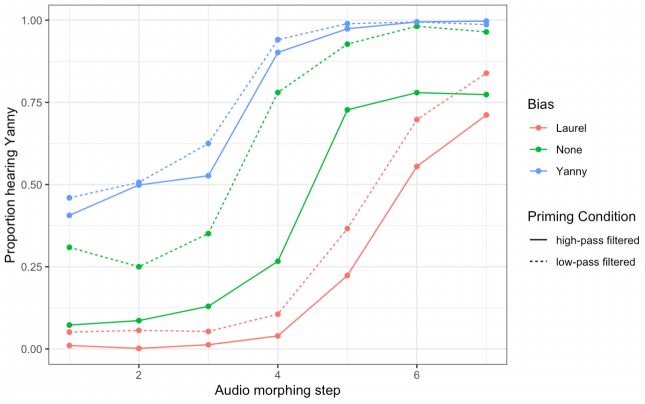


Figure 6: Rate of hearing Yanny based on individual bias toward hearing Yanny or Laurel

A Bayes net model was created using WebPPL to capture the relationship between the audio priming condition, absolute step modification of the audio, the participant's bias, and the participant's probability of responding Yanny. The model assumes that each participant is either Yanny-biased or Laurel-biased. The priming condition and the step modification could vary by audio, and a person's probability of hearing "Yanny" depends on all both audio conditions and their bias. Probability priors of the Bayes net were set using the proportions in the dataset. The model structure is visualized in Figure 7, and probabilities of hearing Yanny are given in Table 2.

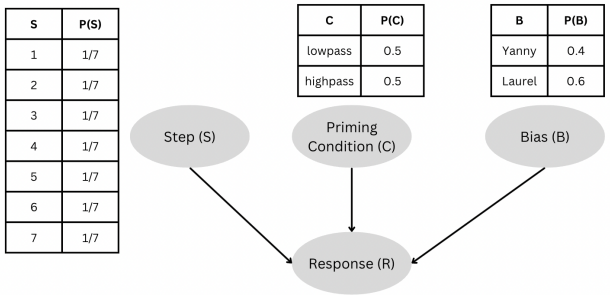


Figure 7: Bayes net structure for modeling the Bosker data

Table 2: Probabilities of hearing Yanny in Bosker data

P(R=Yanny|S, C, B)	S	C	B
0.45974026	1	lowpass	Yanny
0.50656168	2	lowpass	Yanny
0.625	3	lowpass	Yanny
0.94072165	4	lowpass	Yanny
0.98927614	5	lowpass	Yanny
0.99452055	6	lowpass	Yanny
0.98648649	7	lowpass	Yanny
0.40625	1	highpass	Yanny
0.49869452	2	highpass	Yanny
0.52673797	3	highpass	Yanny
0.90180879	4	highpass	Yanny
0.97382199	5	highpass	Yanny
0.99465241	6	highpass	Yanny
0.99717514	7	highpass	Yanny
0.051236749	1	lowpass	Laurel
0.056363636	2	lowpass	Laurel
0.053191489	3	lowpass	Laurel
0.105802048	4	lowpass	Laurel
0.366024518	5	lowpass	Laurel
0.697345133	6	lowpass	Laurel
0.838709677	7	lowpass	Laurel
0.010600707	1	highpass	Laurel
0.001754386	2	highpass	Laurel
0.012844037	3	highpass	Laurel
0.039586919	4	highpass	Laurel
0.223443223	5	highpass	Laurel
0.554964539	6	highpass	Laurel
0.711610487	7	highpass	Laurel

The `pymc` package in Python was used to fit the hierarchical logistic regression model, which included the step and priming condition as predictors of the probability of responding Yanny. The intercept and step coefficient could vary by participant to model individual differences. The model had the following structure and priors:

$$\beta_{0j} \sim \mathcal{N}(\mu_0, 2^2), \;\; \mu_0 \sim \mathcal{N}(0, 100^2)$$

$$\beta_{step,j} \sim \mathcal{N}(0, 10^2)$$

$$\beta_{cond} \sim \mathcal{N}(0, 10^2)$$

$$R_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}(p_{ij}) = \beta_{0j} + \beta_{step,j} \times \text{step} + \beta_{cond} \times \text{priming condition}$$

New Survey Design and Analysis

In order to assess the effect of audio priming using morphed versions of the Yanny-Laurel itself on the perception of the original recording, a new survey was conducted. The survey was conducted online via Qualtrics and distributed to Harvard student mailing lists. The survey asks 3 questions, which are included fully in the appendix. The first question asks the participant whether they hear Yanny or Laurel in the original recording (step 4 from Bosker, 2018), which is used to determine whether they are Yanny-biased or Laurel-biased. Then, they are asked to listen to the morphed recordings in order from step 1 to 7 and to report on which recording their perception changed, if at all. In the third question, they repeat this task but listening to the recordings from step 7 to 1.

A logistic regression was used to model the probability of hearing Yanny on step 4 using the bias, listening direction, and their interaction term as predictors. The logistic regression followed the following model:

$$Y_i \sim \text{Bernoulli}(p_i)$$

$$\begin{aligned}\text{logit}(p_i) = \;& \beta_0 + \beta_1 \times \text{bias} + \beta_2 \times \text{listening direction} \\ & + \beta_3 \times \text{bias} \times \text{listening direction}\end{aligned}$$

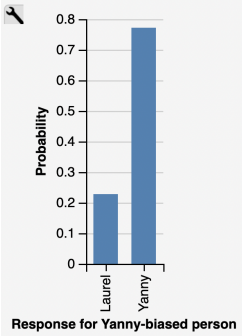
Results

Bosker Data Results

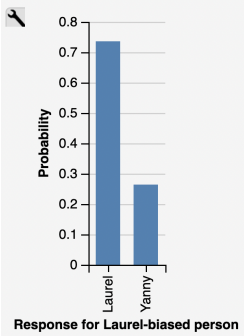
Bayes Net Figures 8a and 8b show the marginal distribution of responses for a Yanny-biased and a Laurel-biased person. As expected, a Yanny-biased person is much more likely to hear Yanny regardless of the audio conditions than a Laurel-biased person. The benefits of the Bayes net model over the standard logistic regression are that it can be used to predict the latent attributes of the audio file given a person's response. Figures 8c and 8d show the posterior distributions of the step and priming condition given that "Yanny" was heard. As expected, a Yanny response makes it more likely that the audio file had a higher step and that the priming condition was lowpass, reflecting the patterns in the data.

The Bayes net could also be used to model the posterior distribution of responses for a particular individual. To test this functionality, we selected a Laurel-biased participant from the data and provided four of their responses, shown in Table 3. Given this information, their posterior probability of being Laurel-biased was very high at 0.999. The model was also used to find the posterior distribution of what this person would respond to two new audio conditions. As seen in the Table 4, the model's predictions were similar to their actual response proportions, demonstrating that the Bayes net seems to reflect actual human stochastic behavior well.

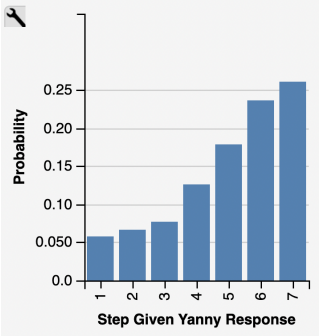
The same analysis was repeated on a participant who was not identified as being Yanny or Laurel biased because their responses to the original audio were evenly split between "Yanny" and "Laurel." The four audio files used as data input and the model's posterior distributions are shown in Table 5. As expected, this person was not as clearly defined as being Yanny or Laurel biased. Their posterior probability of being Yanny-biased was 0.3. The posterior probabilities given by the model for the new audio clips were not as closely matched to the actual participant responses (Table 6). Therefore, it seems that the Bayes net is not able to accurately predict posterior probabilities as easily for people with a weaker bias.



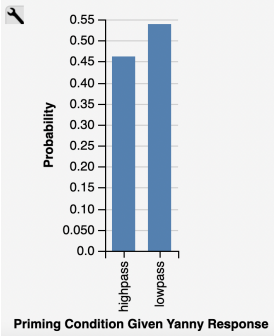
(a) Yanny-biased person response distributions



(b) Laurel-biased person response distributions



(c) Yanny-biased person response distributions



(d) Laurel-biased person response distributions

Figure 8: Overall posterior distributions from Bayes net model of Bosker (2018) data

Table 3: Data inputs for a Laurel-biased participant

Response	Audio Step	Priming Condition
Laurel	4	lowpass
Laurel	6	highpass
Laurel	3	highpass
Yanny	7	highpass

Table 4: Laurel-biased participant's response to new audio

Step	Priming	Predicted Yanny Probability	Actual Yanny Probability
6	lowpass	0.697	0.8
5	highpass	0.224	0.2

Table 5: Data inputs for a non-biased participant

Response	Audio Step	Priming Condition
Yanny	7	highpass
Yanny	6	lowpass
Laurel	2	lowpass
Laurel	1	highpass

Table 6: Non-biased participant's response to new audio

Step	Priming	Predicted Yanny Probability	Actual Yanny Probability
4	highpass	0.294	0.6
1	lowpass	0.172	0.0

Hierarchical Logistic Regression Figure 9 shows the distributions of μ_0 and β_{cond} from MCMC sampling. μ_0 seems to be centered around -9 and β_{cond} around 1.15. Figure 10 shows the 95% HDI from MCMC sampling for the intercepts and step coefficients of 20 randomly selected Yanny-biased participants and Laurel-biased participants. There is considerable variation between participants, demonstrating individual variability in what people hear. In general, Yanny-biased participants have larger intercepts and step coefficients than Laurel-biased participants, so the model was able to capture systematic differences between Yanny and Laurel-biased participants even though the bias was not a variable explicitly incorporated into the logistic regression model.

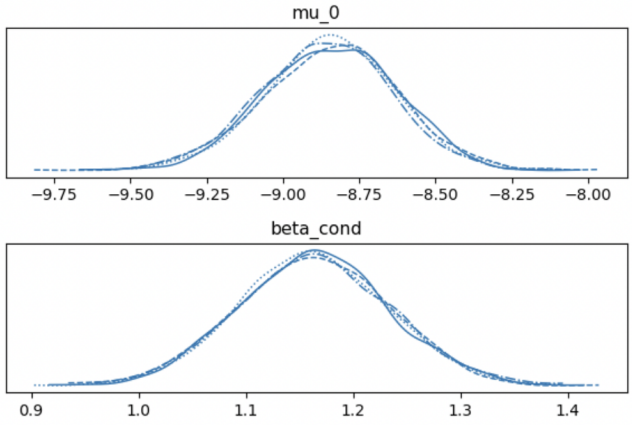


Figure 9: Fixed effect distributions

Survey Results

Exploratory Results The survey was completed by 57 participants. The participants were very evenly split in their responses to the original recording with 29 Laurel responses and 28 Yanny responses. Figure 11 shows the distribution of when people's perception changed when the audio files progressed from being most Laurel-like to most Yanny-like, and Figure 12 shows the distribution when the audio files progressed from being most Yanny-like to most Laurel-like. Overall, it seems people were very likely to stick to hearing

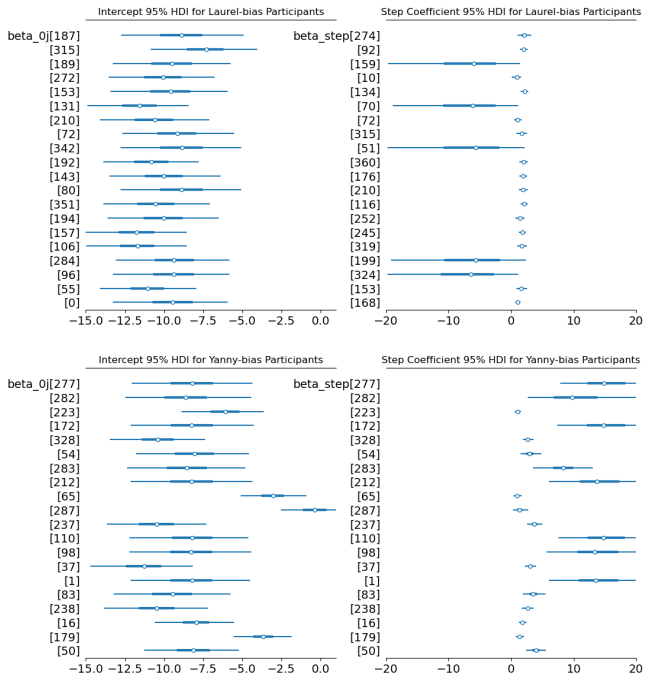


Figure 10: Varying effect distributions for a selection of participants

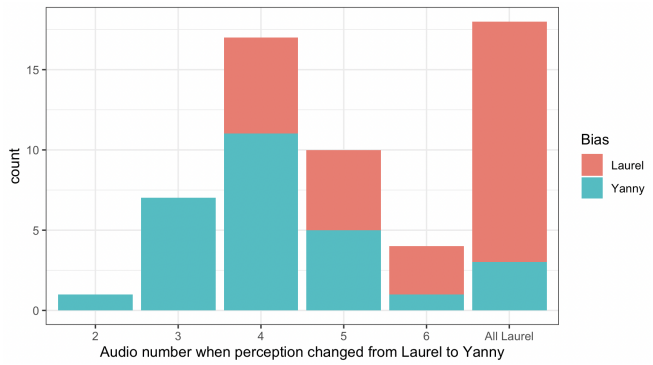


Figure 11: Reported audio perception change when participants heard recordings from most Laurel-like to most Yanny-like (Question 2 responses)

Figure 12: Reported audio perception change when participants heard recordings from most Yanny-like to most Laurel-like (Question 3 responses)

the word in the starting audio. For both listening directions, many people heard Laurel for all audio files, which may be because the first audio in the second question was the most Laurel-like. Laurel-biased participants' perceptions were particularly sticky toward Laurel when the frequencies gradually modulated. Their perception stayed as Laurel longer in the Laurel-starting condition and changed to Laurel faster in the Yanny-starting condition. On the other hand, Yanny-biased participants had much more variation in when their perceptions changed. Overall, as hypothesized, people were more likely to change their perception on audio files 5-7 than on files 1-3, indicating that when the frequency composition of the file gradually changes, the perception stays with the first audio heard for longer.

Table 7: Logistic regression coefficients

Coefficient	Estimate	p-value
Intercept	-1.3437	0.003376
Yanny Bias	2.0909	0.000627
Listening Direction (reference=Laurel start)	1.6920	0.004362
Interaction	-2.1516	0.008178

Logistic Regression Coefficient estimates for the logistic regression predicting whether a participant hears Yanny for the original recording (step 4) are shown in Table 7. All coefficients had significant coefficients. If a participant was Yanny-biased, they were more likely to hear Yanny on the original recording when it was presented in the gradual continiums later in the survey. When the listening direction was switched from starting Laurel-like to starting Yanny-like, participants were much more likely to hear Yanny in the original recording.

Discussion

This study provides insights into the factors influencing multi-stable auditory perception by analyzing individual differences in responses to the Yanny-Laurel phenomenon. By using probabilistic approaches, we demonstrate that auditory perception is not only influenced by external stimuli such as audio priming and frequency modulation but also by intrinsic factors like individual biases toward perceiving "Yanny" or "Laurel."

The results highlight the significant role of individual auditory bias in determining perception. As seen in the hierarchical logistic regression model on the Bosker (2018) data, there

were major differences in coefficients estimates between participants. In the analysis on the survey data, intrinsic bias toward Yanny or Laurel also had a significant on perceptual stability.

Audio priming also played a critical role, with participants exposed to low-pass filtered priming being more likely to hear "Yanny" in ambiguous recordings, consistent with the spectral contrast effect. However, analysis on the survey data also reveals an anchoring effect, where participants tended to stick with their initial perception longer when frequencies were gradually modulated.

There were several limitations in the survey design and results. First, the sample size for the survey (N=57) was relatively small, limiting the generalizability of the results. Second, demographic variables such as age and the device used to take the survey were not collected, though these factors are known to influence frequency perception. Given that the survey was only explicitly sent to Harvard undergraduates, age is likely to be a confounding factor that may bias the newly collected data. Third, the survey design may have introduced order effects, as participants always began with hearing the most Laurel-like audio first, potentially biasing responses. Randomization of stimuli order in audio morphing could be improved in future research designs.

On the modeling side, many assumptions were made in the structure of the Bayes net and logistic regression models including assumptions about the possible outcomes and the relationships between variables. Other model structures could be tested to determine which is most suitable. Future research could address these limitations to further deepen understanding of human auditory perception in uncertain settings.

Appendix

Survey Questions

1. Listen to the following audio: [Step 4 audio from Bosker (2018), or the original viral recording]

   Which is most similar to what you heard?

   - Response options were "Laurel" and "Yanny"

2. Listen to the following audio files in order:

   Audio 1: [Step 1 audio from Bosker (2018)]

   Audio 2: [Step 2 audio from Bosker (2018)]

   Audio 3: [Step 3 audio from Bosker (2018)]

   Audio 4: [Step 4 audio from Bosker (2018)]

   Audio 5: [Step 5 audio from Bosker (2018)]

   Audio 6: [Step 6 audio from Bosker (2018)]

   Audio 7: [Step 7 audio from Bosker (2018)]

   On which numbered audio did you hear something different from the earlier audio?

   (ex. If you heard Laurel on audio 1 and 2, then Yanny on audio 3, you would select 3)

   - Response options were "I heard Laurel for all of them", "I heard Yanny for all of them", and 2 through 7

3. Listen to the following audio files in order:

   Audio 1: [Step 7 audio from Bosker (2018)]

   Audio 2: [Step 6 audio from Bosker (2018)]

   Audio 3: [Step 5 audio from Bosker (2018)]

   Audio 4: [Step 4 audio from Bosker (2018)]

   Audio 5: [Step 3 audio from Bosker (2018)]

   Audio 6: [Step 2 audio from Bosker (2018)]

   Audio 7: [Step 1 audio from Bosker (2018)]

   On which numbered audio did you hear something different from the earlier audio?

   (ex. If you heard Laurel on audio 1 and 2, then Yanny on audio 3, you would select 3)

   - Same options as question 2

References

Bosker, H. R. (2018). Putting laurel and yanny in context. *The Journal of the Acoustical Society of America*, *144*, EL503–EL508.

Brainard, D. H., Longère, P., Delahunt, P. B., Freeman, W. T., Kraft, J. M., & Xiao, B. (2006). Bayesian model of human color constancy. *Journal of Vision*, *6*. doi: https://doi.org/10.1167/6.11.10

Chandra, K., Li, T.-M., Tenenbaum, J. B., & Ragan-Kelley, J. (2022). Designing perceptual puzzles by differentiating probabilistic programs. *SIGGRAPH '22 Conference Proceedings*. doi: https://doi.org/10.1145/3528233.3530715

Elhilali, M. (2013). Bayesian inference in auditory scenes. *Conf Proc IEEE Eng Med Biol Soc*. doi: 10.1109/EMBC.2013.6610120

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*, 1–24.

Katz, J., Corum, J., & Huang, J. (2018). We made a tool so you can hear both yanny and laurel. *New York Times*.

Lafer-Sousa, R., & Conway, B. R. (2017). thedress: Categorical perception of an ambiguous color image. *Journal of Vision*, *17*. doi: https://doi.org/10.1167/17.12.25

Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in color perception uncovered by 'the dress' photograph. *Current Biology*, *25*, R545-R546. doi: https://doi.org/10.1016/j.cub.2015.04.053

Matsakis, L. (2018). The true history of 'yanny' and 'laurel'. *WIRED*.

Pressnitzer, D., Graves, J., Chambers, C., de Gardelle, V., & Egré, P. (2018). Auditory perception: Laurel and yanny together at last. *Current Biology*, *28*(13), R739-R741. doi: https://doi.org/10.1016/j.cub.2018.06.002