Immediate ambiguity resolution in speech perception based on prior acoustic experience

Laura Gwilliams & Pascal Wallisch
Department of Psychology
New York University

Abstract

Speech perception relies on the rapid resolution of uncertainty. Here we explore whether auditory experiences contribute to this process of ambiguity resolution. ~8000 participants were surveyed online for their (i) subjective percept of a speech stimulus with ambiguous formant allocation; (ii) demographic profile and auditory experiences. Both linguistic and non-linguistic auditory experiences significantly predict speech perception. Listeners were more likely to perceive the ambiguous stimulus in accordance with their own name, and were biased towards lower formant allocation as a function of being exposed to lower auditory frequencies in their environment. Overall, our results show that the subjective interpretation of an ambiguous stimulus in the auditory domain is determined by prior acoustic exposure, suggesting the operation of an exposure-dependent mechanism impacting sensitivity that resolves ambiguity in speech perception.

Significance statement

Organisms must contend with fundamental uncertainty due to missing information in sensory inputs, resulting in perceptual ambiguity. There are several known ways in which perceptual ambiguity of speech is resolved, most importantly by the immediate linguistic context. Here, we show that prior acoustic experience - be it speech specific or more general - determines how listeners resolve a segment of ambiguous speech, the LaurelYanny stimulus. This is important both because it suggests that the brain uses the principle described in this paper to resolve ambiguous speech in general and that divergent auditory inputs will yield immediate disagreement about the conscious percept, given ambiguous stimulus situations.

Keywords: *speech, ambiguity resolution, formant, auditory, perception, disagreement, #LaurelYanny*

Introduction

In the competitive struggle for survival, organisms face inherent, vast and irreducible uncertainty, yet need to rapidly produce a suitable motor action. Much of cognition is dedicated to deal with ambiguity stemming from missing information. For instance, the veridical motion of moving objects - confounded by the aperture problem - is readily recovered by processing in the visual system (Shimojo et al., 1989; Movshon et al., 1992; Pack et al., 2001). In the color domain, ambiguities due to variable lighting conditions are resolved by discounting the illuminant, a mechanism known as color constancy (Ebner, 2007).

Human verbal communication is no exception; it is also beset with inherent ambiguities. The uncertainty in speech perception derives from the many-to-many mapping between the acoustic signals and the linguistic units ("phonemes") such as /b/ or /p/ they refer to (Liberman, 1967). This is owed to inherent variability in the production of the acoustic signals - both within and between speakers - so that listeners have to infer which discrete meaning to assign to any given spectro-temporal pattern. This also poses challenges for speech recognition by artificial intelligence systems, like in the presence of accents (O'Shaughnessy, 2008).

For human listeners, handling ambiguity in the sensory input is typically not too difficult, because context sufficiently disambiguates possible interpretations. For instance, the word or sentence that the phoneme is embedded in (Gwilliams et al., 2018), the characteristics of a particular speaker (Cai et al., 2017) and the situation (Mattys et al., 2012) usually provide a context that is informative enough to resolve stimulus-inherent uncertainties.

One potential factor that could *also* serve to disambiguate speech lies in domain-general experience with auditory signals, e.g. exposure to low frequency sounds. Up until now, this has not been tested, owing to the fact that it is hard to estimate one's domain-general lifetime exposure to high or low frequency sounds.

In May 2018, a speech-relevant stimulus was discovered that lends itself to such exploration. Originally recorded for vocabulary.com, it has two radically divergent interpretations: Some listeners hear it as "Laurel", whereas other listeners hear it as "Yanny". Importantly, this stimulus is a person's name, so it is not easily disambiguated by the context of a sentence. The listener's percept can be adjusted by applying different filters to the signal: boosting the power in the lower frequencies shifts the percept towards Laurel and boosting higher frequencies shifts it towards Yanny (Pressnitzer et al., 2018). However, in its natural form, the stimulus is perceptually stable within listeners. This is in contrast to other popular illusions like Rubin's vase or the Necker cube (Peterson et al., 1992), which are bistable, i.e. the percept switches over time within the same observer. The stability of the Laurel/Yanny stimulus could hint at the fact that the interpretation is determined by global and therefore stable prior exposure.

These perceptual characteristics are akin to a similar stimulus in the visual domain, #theDress. In this stimulus - a photograph of a dress - the color of the dress is unambiguously perceived as black and blue by some and as white and gold by others. Which of these percepts is experienced has been linked to prior differential exposure to daylight (Wallisch, 2017). Viewers make different assumptions about the illuminant, depending on their experience with daylight, thus leading to the differential perception of dress color.

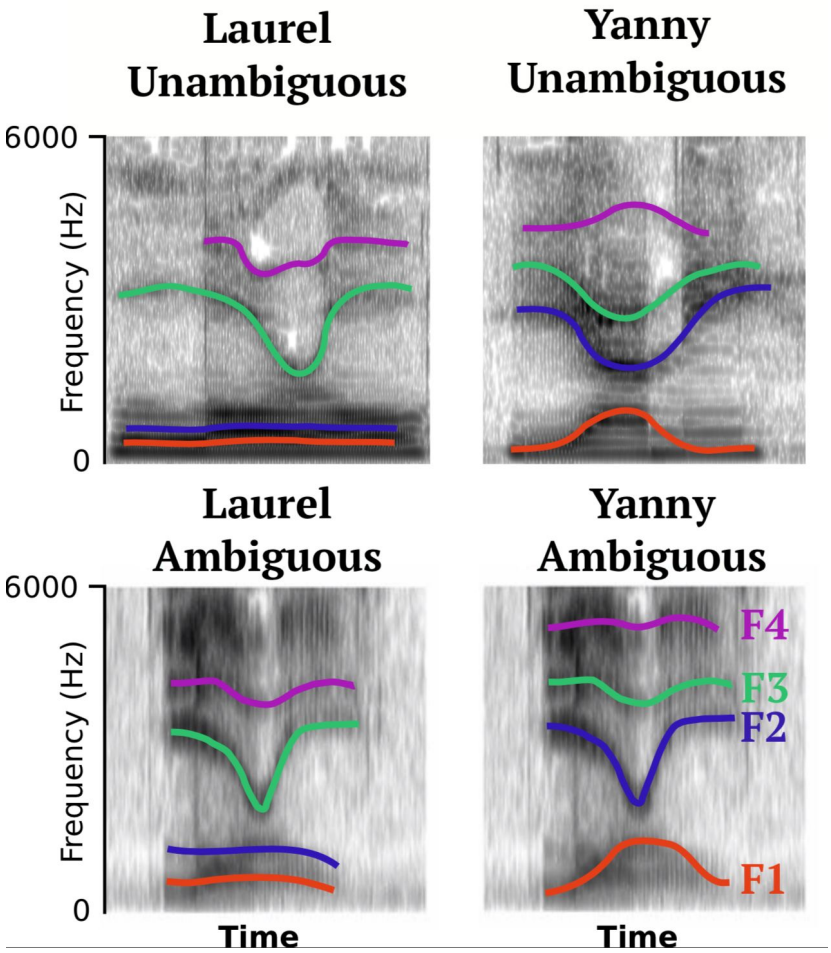

Figure 1. Top: Spectrogram of a female speaker saying the words “Laurel” (left) and “Yanny” (right). We superimposed the first four formants (orange = F1, blue = F2, green = F3, purple = F4). Bottom: Spectrogram of the ambiguous LaurelYanny stimulus with the perceived formant allocation when listeners assume a “Laurel” interpretation (left) and a “Yanny” interpretation (right). Note that the critical difference between the two percepts is the mapping from frequency information to the allocation of formants.

The source of ambiguity in the LaurelYanny stimulus is the allocation of formant frequencies that determine which vowels and vowel transitions are present in speech (Figure 1). When the power of the distribution is altered, as shown by Pressnitzer et al. (2018), the percept is biased relative to formant resolution: Boosting low frequencies biases the percept towards “Laurel” because more power in these frequencies allows to resolve the F1 and F2 - which are very close together. Conversely, low power in these frequencies leads listeners to interpret them as a single F1, interpreting the next frequency band as the F2.

The goal of this study is to test whether prior experience with linguistic and non-linguistic auditory input modulates the perception of an ambiguous stimulus. Specifically we test the hypothesis whether more prior exposure to sounds with stronger power in the low frequencies will predispose listeners to hear the LaurelYanny stimulus as “Laurel”.

Method

Survey

In order to evaluate the role of prior experience on speech perception, we conducted an online survey. The survey was hosted on Survey Monkey (SurveyMonkey.inc, San Mateo, California, USA, www.surveymonkey.com). The survey included a total of 34 questions, and took less than 10 minutes to complete.

This survey included a range of questions that can be grouped into the following categories:

Stimuli
We asked participants to report on their percepts in response to the original LaurelYanny sound clip as well as a low-pass filtered and a high-pass filtered version of the same stimulus. Filtering was done in Matlab, using a Butterworth filter with a 4250 Hz cutoff. Prior work has shown that this filtering technique, at the group level, shifts perception towards one percept or another (Pressnitzer et al., 2018). This measure also allowed us to assess how consistent the individual was in perceiving “Laurel” or “Yanny”.

Physical hardware
Depending on whether participants were listening to the stimulus through high or low quality headphones or speakers will serve as a hardware filter on the auditory stimulus. Thus, we wanted to be able to account for this in our data.

Biological factors
We predicted that depending on the age and health of participants, the fidelity of the signal which reaches the auditory cortex from the early auditory system (the inner ear in particular) would be of a variable quality. Participants were therefore asked their age at time of survey completion and whether they have a hearing disorder.

Prior experience
Our primary theoretical interest is prior exposure to auditory input. Thus, we asked participants: (i) the similarity of their name to the auditory targets; (ii) whether they grew up in a rural, suburban or urban environment, (iii) whether they play a musical instrument and if so, what kind; (iv) the relative pitch of their voice, (v) what kind of music they listen to.

Questions and personality, personal beliefs and demographics
Some of the demographics might relate to frequency content of one's prior auditory experience (e.g. pitch of voice is related to biological sex). Moreover, we did not think that political beliefs or personality would be related to how one perceives the LaurelYanny stimulus, allowing us to use these questions to estimate whether our dataset is prone to false positives.

Participants

3987 participants were recruited from Amazon Mechanical Turk in summer 2018. Participants had an approval rate greater than or equal to 95%, and at least 1000 approved previous tasks completed. They were compensated for their time. Data from 644 of these participants was not included in further analysis, because they indicated in self-report that the data they provided are not reliable. We included data from 83.9% (n=3343) of participants from this data source in our analysis.

3920 participants were recruited through media and social media. No compensation was provided to these participants. Data from 68 of these participants was not included in further analysis, because they indicated in self-report that the data they provided are not reliable. We included data from 98.3% of participants (n=3852) from this data source in our analysis.

Overall, we retained 91% of the data for further analysis. Of the remaining 7915 participants, 3821 identified as female; 3317 as male; 57 as non-binary; median age=36 (SD=13.5). Participants took part from all continents except Antarctica (Africa=30, Asia=678, Oceania=79, Europe=367, North America=5949, South America=92).

All participants provided informed consent and all procedures were approved by the New York University Institutional Review Board (UCAIHS).

Data Analysis

All data were analyzed with MATLAB (Mathworks, Natick). Due to the categorical nature of the data, we used Chi-squared tests to assess the significance of any given comparison. We estimated 95% confidence intervals by bootstrapping. Specifically, we draw 1e5 times from the empirical distribution to generate virtual samples. We also performed a regression analysis.

Results

Are percepts of the LaurelYanny stimulus modulated by filtering?
First we tested whether physically filtering the LaurelYanny stimulus modulates the perceptual reports of listeners, as suggested by Pressnitzer et al. (2018). We expect, based on the reasoning laid out in figure 1 that low-pass filtering of the stimulus will change the balance of

power distribution in the LaurelYanny stimulus, impoverishing the contribution of higher frequencies, shifting the formant allocation downwards, which should shift percepts of listeners towards Laurel. Conversely, high-pass filtering the stimulus should bias the percept towards Yanny, based on the same rationale, mutatis mutandis, with perceptions of the original, unfiltered stimulus somewhere in between. We present the results of this analysis in figure 2.

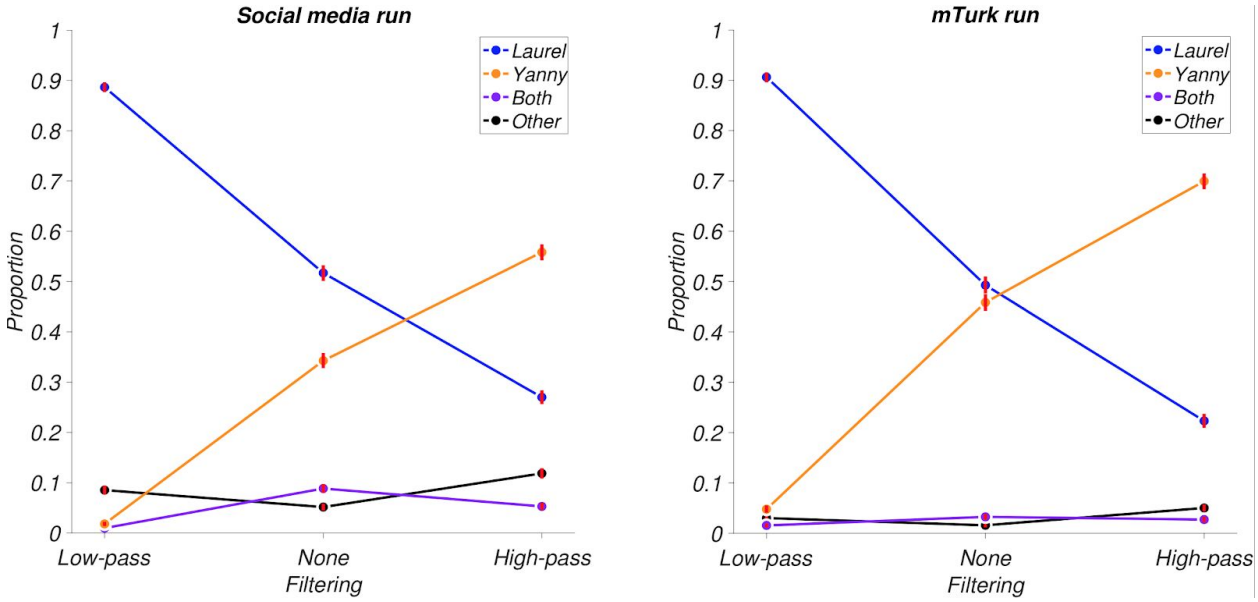

Figure 2. Effect of signal filtering on reported percept across data collection sessions. Left: Data from social media collection run (n=3852); Right: Data from Amazon Mechanical Turk (n=3343). The x-axis of both plots corresponds to the filter applied to the auditory signal of the ambiguous LaurelYanny stimulus. Low-pass filter corresponds to X-YHz; High-pass filter X-YHz; No filter was the original LaurelYanny stimulus. Y-axis corresponds to the proportion of "Laurel" responses. Red error bars represent 95% bootstrapped confidence intervals.

As you can see, reported percepts of listeners are strongly modulated by physical filtering of the stimulus in both runs, the social media run (χ^2 = 492.90, df = 2, p = 3.33e-109, ϕ = 0.253) as well as the mTurk run (χ^2 = 529.41, df = 2, p = 3.79e-117, ϕ = 0.281), replicating Pressnitzer et al. (2018). The overall pattern of results is consistent with our predictions based on the rationale laid out above. In the low-pass filtered version of the stimulus, almost 90% observers report to hear the LaurelYanny stimulus as "Laurel". This proportion drops to around 30% for the high-pass filtered stimulus. The reports of the Yanny percept follow the opposite pattern, but reports seems to be somewhat biased towards reporting Laurel overall, in the social media run. The listeners in the mTurk run broadly replicate all of the patterns observed in the social media run, with one notable difference: Reports of "switching" (both) percepts or "other" percepts almost entirely resolved towards "Yanny", rendering the overall Laurel/Yanny proportion in this run more symmetric than in the social media run. However, because the Laurel proportion is stable between the two runs, and we report on the proportion of the modal percept, we feel justified to pool the data from the two runs for the following analyses.

Do physical factors of stimulus delivery modulate the percept?

We established, replicating Pressnitzer et al. (2018), that direct signal filtering of the stimulus strongly modulates the percept of the LaurelYanny stimulus. Next we tested whether the filtering characteristics of the physical stimulus delivery system, i.e. whether the stimulus is delivered via headphones or speakers produces analogous effects on perception. This is plausible because speakers tend to emphasize bass, whereas headphones exhibit higher fidelity across the entire frequency range (Møller et al., 1995; Guttenburg, 2016).

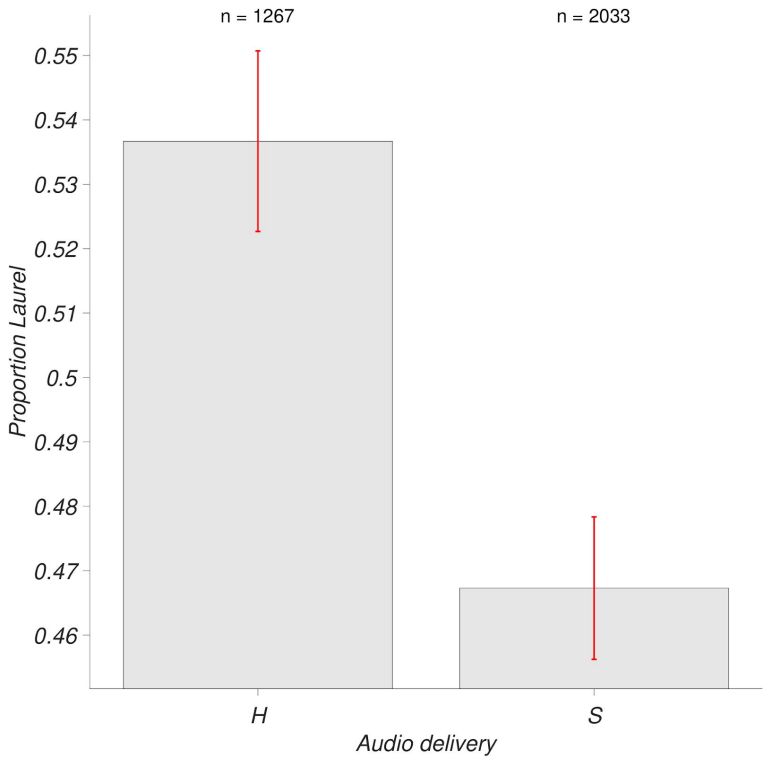

Figure 3. Effect of physical signal presentation. Trials are grouped into whether participants were listening on headphones (H) or speakers (S). Note that the reduced number of participants in the analysis is due to not adding this question to the survey until after data collection had started.

There is a relationship between audio delivery system and how listeners experience the LaurelYanny stimulus, it goes in the direction we predicted theoretically and is statistically significant, $\chi^2 = 15.04$, df = 1, p = 1.05e-4, $\phi = 0.067$.

Does biological filtering modulate the percept?

We showed that filtering the physical stimulus with signal processing as well as filtering via the physical delivery system modulates the percept of listeners responding to the LaurelYanny stimulus, see figure 2 and 3. Next we tested whether “biological” filtering can produce similar effects. This is plausible because the basilar membrane of the inner ear effectively represents the fourier transform of the frequency components of the auditory stimulus (von Békésy & Weber, 1960; Mills et al., 2006). However, this is contingent on the inputs to the basilar membrane, specifically the state of hair cells. It is well known that age drastically affects these

hair cells, in a frequency dependent fashion, with severe decay of hair cells processing higher frequencies as a function of age (Liberman, 2017). Thus, we expect that - using chronological age as a proxy for the state of hair cells in the inner ear - younger listeners hear the stimulus with higher fidelity in the higher frequency ranges, whereas older listeners will be exposed to a stimulus that is effectively overrepresenting the lower frequency ranges. Based on the same rationale as laid out previously, we therefore predict that younger listeners will be biased towards hearing Yanny and older listeners will be more likely to hear Laurel.

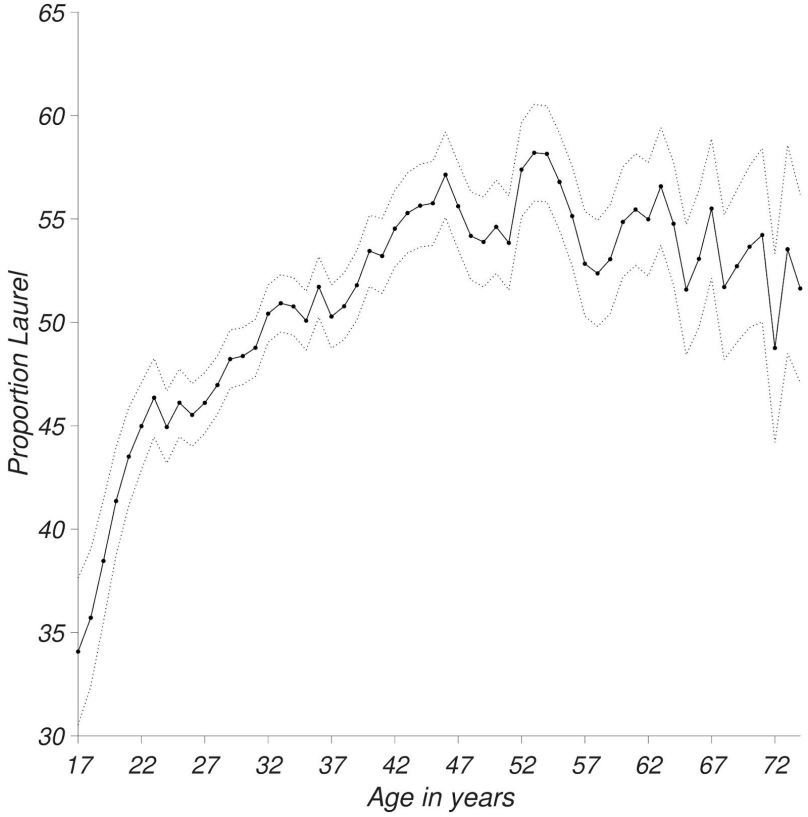

Figure 4. Effect of age on the likelihood of reported a “Laurel” percept. Solid line corresponds to the average number of “Laurel” reports for each chronological age. Hashed lines correspond to 95% bootstrapped confidence intervals.

There are several notable things about figure 4. As predicted, the probability of reporting Laurel increases as a function of age. In addition, most of this rise happens in the age range from 17 to 37, in line with what one would expect from biology - most of the high frequency hearing loss that affects the hair cells in the inner ear happens during that period (Liberman, 2017). Finally, note that the error bands on the higher end of the age range are rather large, which is due to the fact that we had only relatively few observers in our sample that fall into this age range. Overall, this pattern of results suggests a rapid rise of Laurel percept as a function of age, asymptoting in the late 40s, consistent with what one would expect from the biology of the inner ear. The Pearson correlation between age and proportion hearing Laurel in the age range from 17 to 46 is 0.949, p = 1.63e-15, asymptoting after that.

Does specific linguistic experience modulate the percept?

We showed that filtering the stimulus with signal processing, physically or biologically affects the percept of listeners in a way that is consistent with the notion that observers allocate formant frequencies to portions of the frequency spectrum that is relatively overrepresented in the stimulus. In line with this rationale, we reason that specific linguistic experience (i.e. hearing one's own name) would effectively have the same effect, overrepresenting the discernment of a specific phonetic pattern. It is known that one's own name has profound effects on domains of cognitive processing, e.g. in attention (Treisman, 1969), due to repeated exposure to the same phonetic pattern. Thus, we predict that people with a name close to Laurel, e.g. "Laura" or "Lauren" are more likely to hear "Laurel", whereas people with a name closer to Yanny, e.g. "Yan" or "Jenny" are more likely to hear "Yanny".

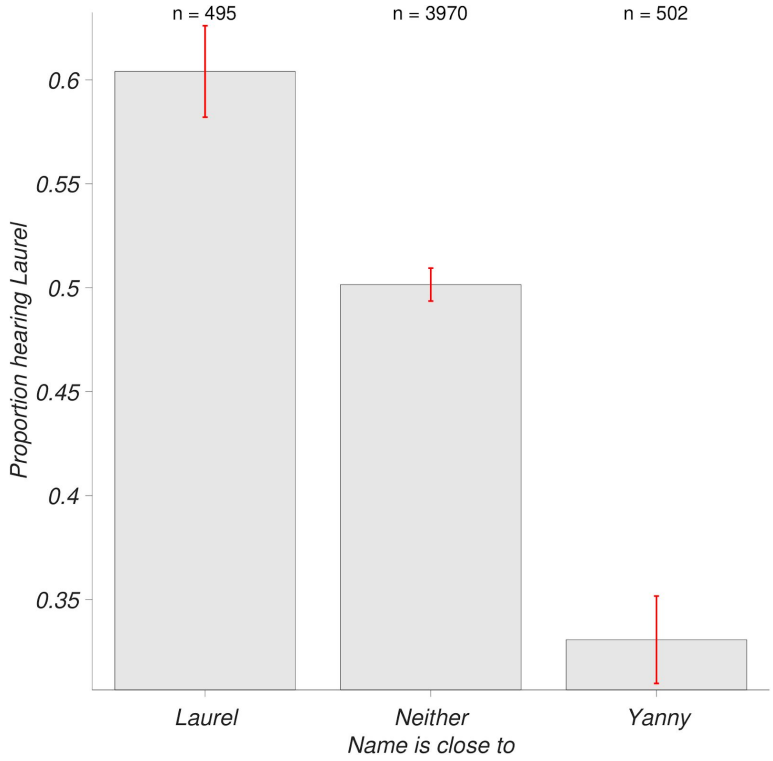

Figure 5. Effect of the listener's name on the proportion of reported "Laurel" percepts. The x-axis corresponds to whether the participant reported having a name that is similar to "Laurel", similar to "Yanny" or similar to neither of them. The number of participants that identify with each of these three categories is shown along the top of the figure. The y-axis represents the proportion of "Laurel" responses as a function of the three name categories. Error bars represent the 95% bootstrapped confidence intervals. In sum, participants were more likely to hear the percept that was closer to their own name.

We observe a pattern of results predicted by the rationale above. One's own name - and its closeness to one of the interpretations of the stimulus - strongly modulates the percept, on par with changing the physical stimulus with signal processing, this effect is statistically significant ($\chi^2 = 78.44$, df = 2, p = 9.27e-18, $\phi = 0.126$).

Does general acoustic experience modulate the percept?

All of the effects described here suggest a mechanism based on the allocation of formant frequencies to the ambiguous frequency content in the stimulus as a function of what is over-represented in stimulus, usually by filtering. Now, we expand this notion to a wider range of acoustic experiences, starting with the pitch of one's own voice. The reason we believe that the pitch of one's own voice might bias percepts is that people vary quite widely in terms of pitch of voice, and the sound of this voice will be a prominent feature in the auditory experience of most people. It is of course challenging to measure - particularly online - the precise pitch of someone's voice, as well as how much someone uses this voice. Thus, for the purposes of this study, we use biological sex as a proxy for pitch of voice - it is known that on average, male voices are pitched 80 Hz lower than female voices (Takefuta et al., 1971; Szakay & Torgersen, 2015). Therefore, we predict that - everything else being equal, biological males will be more likely to hear Laurel and biological females will be more likely to hear Yanny.

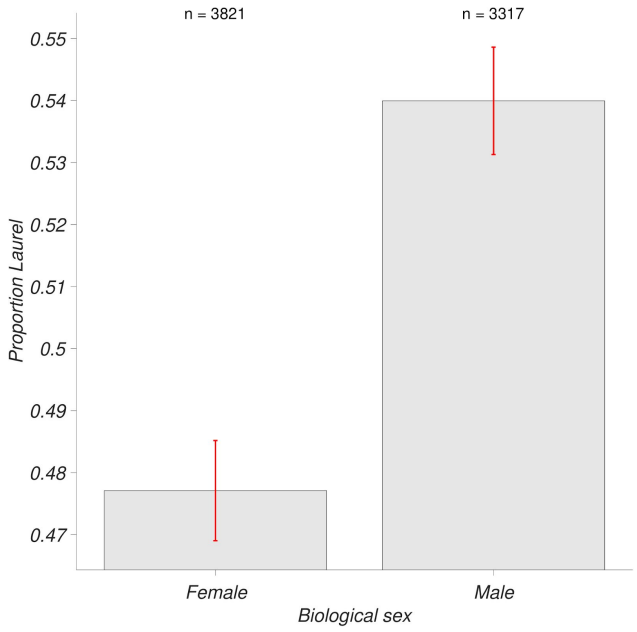

Figure 6. Effect of biological sex on reported "Laurel" percepts. Self-identifying males are more likely to perceive a "Laurel" percept; self-identifying females are more likely to perceive a "Yanny" percept.

The effect of biological sex on modulation of the percept is not large, but it is consistent with our theoretical predictions and statistically significant: $\chi^2 = 28.06$, df = 1, p = 1.18e-7, $\phi = 0.063$.

Just like the use of speakers might reflect a general propensity to be exposed to auditory environments with more base, we wondered whether this is true for content as well. An obvious choice is music, see figure 7.

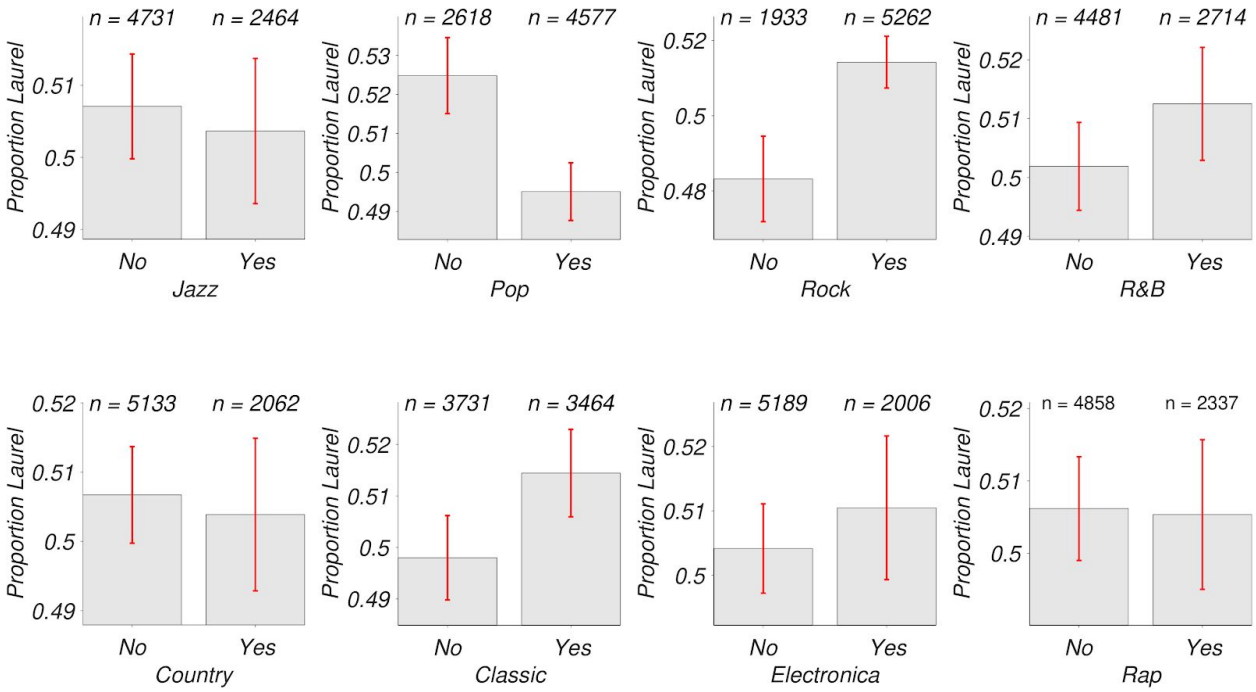

Figure 7. Effect of music genre on proportion of "Laurel" percepts. Participants were asked to select all genres, from the eight choices, that they enjoyed listening to. In all, participants who liked listening to pop music were more likely to hear "Yanny", and those who liked rock music were more likely to hear "Laurel". As we discuss, this is most likely due to the typical frequency range of these musical genres.

Two of these effects are statistically significant, namely the effect of pop ($\chi^2 = 5.89$, df = 1, p = 0.015, $\phi = 0.029$) and rock ($\chi^2 = 5.46$, df = 1, p = 0.019, $\phi = 0.028$) music on the perception of the LaurelYanny stimulus, and in different directions.

To ascertain that these effects are not due to obvious confounds, i.e. it could be possible that the use of audio-equipment is a cohort effect. If younger people are more likely to use headphones, they would be more likely to hear Yanny because of either effect. If headphone use is correlated with biological sex, we would see the propensity to hear Yanny as a function of biological sex, even if biological sex had no effect by itself. To control for this, we ran partial correlations on all of the effects, and none could be attributed to headphone use.

To extend this logic to other factors, we performed two regressions - one logistic regression against "hearing Laurel" - in the original stimulus as the dependent variable and one linear regression against "strength of percept" where we add up the number of times people report hearing "Laurel" (out of the 3 possibilities) as a dependent variable, see figure 8.

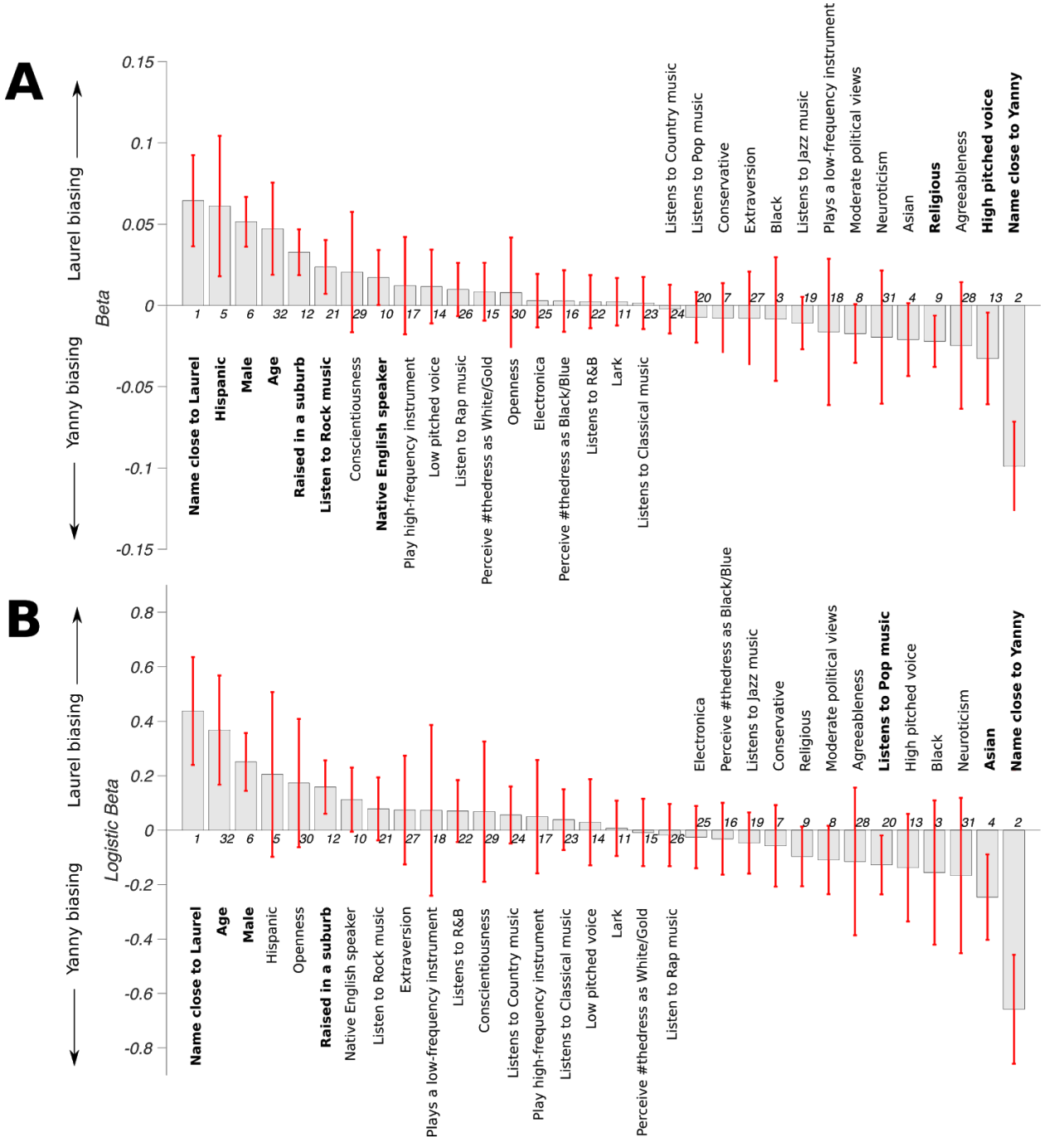

Figure 8. Results of the multiple regression analysis from linear regression predicting the "strength" of the Laurel percept (A) and logistic regression predicting whether listeners report hearing Laurel in the original unfiltered stimulus (B). All regressors were included in the model simultaneously, and the resulting beta coefficients and 95% confidence intervals are plotted in order of relative magnitude. Positive beta coefficients indicate that the regressor was predictive of a "Laurel" percept; negative beta coefficients are predictive of a "Yanny" percept. Bold font for the regressor name indicates that it significantly predicts the dependent variable at $p < .05$.

--- INSERT TABLE 1 about here ---

Table 1: Regression effects for both regression analyses

This regression analysis confirms the effects we showed previously in all cases, except for pop music, which might have emerged due to an age confound - younger people who are more likely to hear Yanny are more fond of pop music. This is not the case for the other effects - controlled by other factors, they are still significant, in the direction we expected, based on our theoretical rationale. In addition, this analysis revealed several effects we did not expect: Identifying as hispanic, being raised in a suburb, being a native English speaker and being religious. It is possible that some of these relationships are spurious - given the large number of comparisons - however, we think this is an opportunity for a future, targeted study of these effects. We could offer post-hoc rationalizations of all of these effects that are in line with our theory, but this would be speculation. For instance, it is known that speakers are very sensitive to the phonotactics of their native language. Given that the initial phoneme sequence in Laurel (i.e. /lɔːr/) is more frequent than in Yanny (i.e. /jan/), it follows that native English speakers would have perceptual preference for the more likely phoneme sequence (i.e. Laurel),(Jusczyk et al., 1994). Similarly, it is possible that suburban soundscapes emphasize frequency ranges that are different from Urban ones, but this has not been explored systematically (Mydlarz et al., 2017). The two cultural factors might also be worth exploring. It is not implausible that someone who is religious would be more likely to attend church on a regular basis, and hymns are notoriously high pitched, which would bias those who attend such services towards hearing “Yanny”. However, we want to emphasize that such rationalizations are purely speculative - as we did not expect them, and best left for future, targeted study.

Also note that this analysis does not include audio equipment as a factor, as we only asked about this in the mTurk sample. Concerned about statistical power (we do not have any information about audio equipment in half of our sample), we decided not to include this factor in the regression analysis. However, a separate analysis - a partial correlation between audio equipment used and the factors laid out above (name, age, gender) suggests that these effects are not simply due to differential use of audio equipment. It is conceivable that younger males are more likely to use headphones, but that does not seem to be the case empirically.

Note that the effect sizes outside of filtering the stimulus and the name effect are small. However, this just means that these smaller effects would likely not be noticeable for individuals, from everyday life. They are still of theoretical interest, mirroring the situation in #theDress research (Wallisch, 2017), where it is important to use adequately powered samples of thousands of observers to reliably detect subtle effects, as any given observer will be affected by many effects simultaneously, pushing and pulling the perceptual outcome in different directions. Finally, we would also like to note that the effects that were significant in both regressions are likely most reliable, and the ones that did not become significant in both are likely absent. Those effects that were significant in one, but not another analysis, i.e. rock music in the linear regression and pop music in the logistic regression are probably more subtle and it is perhaps worth exploring in future research which of these are statistically reliable. It is quite

possible that these reflect subtle cultural factors, for instance Asian languages such as Mandarin or Korean are known to have a higher fundamental frequency (Lee & Sidtis, 2017), which could for the observation that Asian self-identity results in a shift towards a Yanny percept in the logistic regression.

Discussion

In this study, we show that listeners resolve an ambiguous speech stimulus differently as a function of a variety of factors including filtering of the auditory input by signal processing, by audio hardware and by the biological state of the auditory system. In addition, we show that acoustic experience - both specifically with speech and more generally with auditory frequency distributions - modulates the percept in the direction consistent with our theoretical expectations. Finally, we establish that these effects are fairly specific - incidental personal characteristics like political beliefs or how someone perceives #theDress do not predict how listeners hear the LaurelYanny speech stimulus, although it is somewhat surprising that none of the personality measures predicts the subjective perceptual experience of observers here.

We believe that there is a parsimonious explanation that accounts for all of these observations - and in addition for phenomenological qualities, i.e. that the percept emerges instantaneously and is hard to override consciously.

First, we would like to recount why the LaurelYanny stimulus is ambiguous in the first place. Under everyday conditions, the auditory system of listeners needs to resolve the formant frequencies in the spectro-temporal auditory input in order to correctly identify the vowels in the speech (Lindblom and Studdert-Kennedy, 1967). The reason this stimulus is fundamentally ambiguous results from the fact that it is presented without context - which often helps to disambiguate formants - as well as the fact that the middle region of the spectro-temporal power distribution was corrupted upon recording of the speech snippet, likely by the braking of a passing truck (Hughes, 2018).

Second, it is the case that organisms are more readily able to resolve minute differences in stimuli if they allocate more resources to processing it. This is true in a variety of domains, including somatosensation (Duncan & Boynton, 2007), vision (Li et al., 2003) and audition (Recanzone et al., 1993) and perhaps due to the stimulus statistics of the respective domain (Simoncelli & Olshausen, 2001). Cognitive systems might simply have more evidence to discriminate stimuli in regions of the stimulus space that are represented well, yielding higher resolution. Prior research (Pressnitzer et al. 2018) showed that how much power - in the high vs. low frequency range - is available in the stimulus determines how finely listeners can resolve each band, and consequently are able to allocate formants to these frequency bands (see figure 1).

Our results dovetail nicely with this theoretical account - anything that is likely to overrepresent a given processing range, be it from physical or biological filtering, or a changing of the neural substrate from experience, shifts perception in that direction. It also explains why the percept is effectively instantaneous and hard to override consciously, as by this account, this shift is baked into the processing itself, not added at a later stage, nor infused “top-down”, akin to processing of a similarly ambiguous stimuli, like #theDress (Wallisch, 2017).

There are several limitations of the current study that put this theoretical account in question. Most importantly, all of our measures are self-reported proxies of auditory experience and stimulus filtering. The advantage of relying on such measures is statistical power - we could probe a large space of possibilities online, with many listeners. The drawback of doing so lies in the inherent coarseness of this approach. For instance, letting participants judge whether their name is “close” to Laurel or Yanny will yield inherently coarse categories. It would be better to use the real names of participants and parametrically determine how close their name is. This would likely yield a more precise link between someone’s name and their perception of the LaurelYanny stimulus. For many reasons - including privacy concerns - we did not do this here, but doing so provides a ready opportunity for future research to improve on our efforts, which could also be of theoretical interest. For instance, it would be interesting to see how sharp are the perceptual transitions in such a psychometric space.

Similarly - on the stimulus delivery end - we relied on a binary description of the auditory hardware used by research participants (speakers versus headphones). We did this in the interest of robustness. It is not realistic to expect members of the general public to know the details of their audio hardware. Nevertheless - and this is another opportunity to improve on our work - it would be interesting to see how the EQ characteristics of audio hardware maps onto percepts. We predict that audio hardware that emphasizes bass will shift percepts towards Laurel whereas those that provide high fidelity across the audio range shift them towards Yanny. On a related note, we do not know whether use of speakers vs. headphones impacted the percept at the point of delivery, or whether using speakers in this moment is indicative of being more likely to use speakers in general, confounding the factors of filtered stimulus vs. experience in our measure. If someone were to make such measurements in a lab environment, they could assign audio equipment at random to tease these factors apart.

There are also more technical limitations to this study. For instance, we did not anticipate some of the factors that turned out to matter, most importantly the impact of one’s own name. We only realized this once reviewing the verbal accounts of early study participants and added this question midway through the study, complicating statistical analyses of our data. Fortunately, this effect seems to be strong enough that we nevertheless had sufficient power to detect it. Given our results, we do believe that there is a predisposing effect of any acoustic experience. However, the details do matter - this exposure will have to be persistent, strong and specific enough to induce changes in auditory cortical substrates to have perceptual consequences. We were somewhat surprised that the impact of listening to particular types of music was not stronger, but this does make sense: Any given song covers a wide frequency range, the broad

genre labels we used are quite heterogeneous and most people have broad musical palates. Judged in the context of the effect of voice pitch, expecting stronger results from musical exposure might therefore not be realistic. Also - again - this issue is plagued by our reliance on self-reported proxies. If we knew how much and which specific music any given participant actually listens to, we predict that this relationship would emerge much more strongly, given sufficient power. This is perhaps an opportunity for a major music streaming service to test this theoretical possibility empirically. Finally, the unexpected effects revealed by the regression analysis (particularly the soundscape of the suburb, see figure 8) are an obvious jumping off point for future research.

Although we recognize that our measures can be refined, we think that our basic theoretical account is sound and also quite important. We live in highly polarized times (van Bavel & Pereira, 2018), with rising disagreement and potential for conflict. Here, we add to a body of evidence suggesting that experience determines the interpretation of ambiguous stimuli, a principle which has been termed “SURFPAD” (Wallisch & Karlovich, 2019). Importantly, the impact of experience on perception is unconscious, involuntary and immediate, yielding unambiguously polarized percepts to observers. This research extends the SURFPAD principle to the auditory domain, suggesting generality. Finally, whereas this stimulus - much like #theDress - arose on social media and caused a internet frenzy, we believe that the effects we describe in the current study reveal mechanisms fundamental to understanding not just speech perception, but auditory perception more generally.

Acknowledgments
This research was supported by the Abu Dhabi Institute Grant G1001 and the Dingwall Foundation Dissertation Fellowship. We would also like to thank the many people who volunteered their time to contribute their data to this study.

Author contributions
PW conceived of the project, LG and PW designed the survey and stimuli, LG and PW recorded the data, PW analyzed the data, PW and LG wrote and edited the manuscript.

References

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, *98*, 73-101.

Duncan, R. O., & Boynton, G. M. (2007). Tactile hyperacuity thresholds correlate with finger maps in primary somatosensory cortex (S1). Cerebral Cortex, 17(12), 2878-2891.

Ebner, M. (2007). *Color constancy* (Vol. 7). John Wiley & Sons.

Hughes, V. (2018, May 27). Where Did The Yanny/Laurel Recording Really Come From? *BuzzFeedNews.* Retrieved from: https://www.buzzfeednews.com/article/virginiahughes/yanny-laurel-audio-conspiracy-theory

Guttenburg, S. (2016, April 16). What's more accurate: Speakers or headphones? *CNET*. Retrieved from: https://www.cnet.com/news/world-ufo-day-see-how-your-state-rates-when-it-comes-to-sightings/

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*(35), 7585-7599.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630.

Lee, B., & Sidtis, D. V. L. (2017). The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing*, *20*(3), 174-185.

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique effect: a neural basis in the visual cortex. Journal of neurophysiology, 90(1), 204-217.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431

Liberman, M. C. (2017). Noise-induced and age-related hearing loss: new perspectives and potential therapies. *F1000Research*, *6*.

Lindblom, B. E., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical society of America*, *42*(4), 830-843.

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.

Mills, J. H., Schmiedt, R. A., Schulte, B. A., & Dubno, J. R. (2006, November). Age-related hearing loss: A loss of voltage, not hair cells. In *Seminars in Hearing* (Vol. 27, No. 04, pp. 228-236). Copyright© 2006 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA.

Movshon, J., Adelson, E. H., Gizzi, M. S., & Newsome, W. T. (1992). The analysis of moving visual patterns. In Frontiers in cognitive neuroscience. MIT Press.

Møller, H., Jensen, C. B., Hammershøi, D., & Sørensen, M. F. (1995). Design criteria for headphones. *Journal of the Audio Engineering Society*, *43*(4), 218-232.

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical society of America*, *68*(5), 1294-1300.

Mydlarz, C., Salamon, J., & Bello, J. P. (2017). The implementation of low-cost urban acoustic monitoring devices. *Applied Acoustics*, *117*, 207-218.

O'Shaughnessy, D. (2008). Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, *41*(10), 2965-2979.

Pack, C. C., & Born, R. T. (2001). Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature*, *409*(6823), 1040-1042.

Peterson, M. A., Kihlstrom, J. F., Rose, P. M., & Glisky, M. L. (1992). Mental images can be ambiguous: Reconstruals and reference-frame reversals. *Memory & Cognition*, *20*(2), 107-123.

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.

Recanzone, G. H., Schreiner, C. E., & Merzenich, M. M. (1993). Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. Journal of Neuroscience, 13(1), 87-103.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. Annual review of neuroscience, 24(1), 1193-1216.

Shimojo, S., Silverman, G. H., & Nakayama, K. (1989). Occlusion and the solution to the aperture problem for motion. *Vision research*, *29*(5), 619-626.

Szakay, A., & Torgersen, E. (2015). An acoustic analysis of voice quality in London English: The effect of gender, ethnicity and f0. In *ICPhS*.

Takefuta, Y., Jancosek, E. G., & Brunt, M., "A statistical analysis of melody curves in the intonation of American English", Proceedings of the 7th International Congress of Phonetic Sciences - Montreal (1971), 1035-1039, 1972.

Traunmüller, H., & Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. Unpublished manuscript.

Treisman, A. M. (1969). Strategies and models of selective attention. Psychological review, 76(3), 282.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. Trends in cognitive sciences, 22(3), 213-224.

Von Békésy, G., & Wever, E. G. (1960). Experiments in hearing (Vol. 8). New York: McGraw-Hill.

Wallisch, P. (2017). Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain:“The dress”. *Journal of Vision*, *17*(4), 5-5.

Wilson, J. P. (1968). High-quality electrostatic headphones. *Wireless World*, *74*, 440-443.