Beyond Laurel/Yanny: An Autoencoder-Enabled Search for Polyperceivable Audio

Kartik Chandra
Stanford University
kach@cs.stanford.edu

Chuma Kabaghe
Stanford University
chuma@alumni.stanford.edu

Gregory Valiant
Stanford University
gvaliant@cs.stanford.edu

Abstract

The famous "laurel/yanny" phenomenon references an audio clip that elicits dramatically different responses from different listeners. For the original clip, roughly half the population hears the word "laurel," while the other half hears "yanny." How common are such "polyperceivable" audio clips? In this paper we apply ML techniques to study the prevalence of polyperceivability in spoken language. We devise a metric that correlates with polyperceivability of audio clips, use it to efficiently find new "laurel/yanny"-type examples, and validate these results with human experiments. Our results suggest that polyperceivable examples are surprisingly prevalent, existing for $>2\%$ of English words.[1]

1 Introduction

How robust is human sensory perception, and to what extent do perceptions differ between individuals? In May 2018, an audio clip of a man speaking the word "laurel" received widespread attention because a significant proportion of listeners confidently reported hearing *not* the word "laurel," but rather the quite different sound "yanny" (Salam and Victor, 2018). At first glance, this suggests that the decision boundaries for speech perception vary considerably among individuals. The reality is more surprising: almost everyone has a decision boundary between the sounds "laurel" and "yanny," without a significant "dead zone" separating these classes. The audio clip in question lies close to this decision boundary, so that if the clip is slightly perturbed (e.g. by damping certain frequencies or slowing down the playback rate), individuals switch from confidently perceiving "laurel" to confidently perceiving "yanny," with the exact point of switching varying slightly from person to person.

How common is this phenomenon? Specifically, what fraction of spoken language is "polyperceivable" in the sense of evoking a multimodal response in a population of listeners? In this work, we provide initial results suggesting a significant density of spoken words that, like the original "laurel/yanny" clip, lie close to unexpected decision boundaries between seemingly unrelated pairs of words or sounds, such that individual listeners can switch between perceptual modes via a slight perturbation.

The clips we consider consist of audio signals synthesized by the Amazon Polly speech synthesis system *with a slightly perturbed playback rate* (i.e. a slight slowing-down of the clip). Though the resulting audio signals are not "natural" stimuli, in the sense that they are very different from the result of asking a human to speak slower (see Section 5), we find that they are easy to compute and reliably yield compelling polyperceivable instances. We encourage future work to investigate the power of more sophisticated perturbations, as well as to consider natural, ecologically-plausible perturbations.

To find our polyperceivable instances, we (1) devise a metric that correlates with polyperceivability, (2) use this metric to efficiently sample candidate audio clips, and (3) evaluate these candidates on human subjects via Amazon Mechanical Turk. We present several compelling new examples of the "laurel/yanny" effect, and we encourage readers to listen to the examples included in the supplementary materials (also available online at https://theory.stanford.edu/~valiant/polyperceivable/index.html). Finally, we estimate that polyperceivable clips can be made for $>2\%$ of English words.

[1] This research was conducted under Stanford IRB Protocol 46430.

2 Method

To investigate polyperceivability in everyday auditory input, we searched for audio clips of single spoken words that exhibit the desired effect. Our method consisted of two phases: (1) sample a large number of audio clips that are likely to be polyperceivable, and (2) collect human perception data on those clips using Amazon Mechanical Turk to identify perceptual modes and confirm polyperceivability.

2.1 Sampling clips

To sample clips that were likely candidates, we trained a simple autoencoder for audio clips of single words synthesized using the Amazon Polly speech synthesis system. Treating the autoencoder's low-dimensional latent space as a proxy for *perceptual* space, we searched for clips that travel through more of the space as the playback rate is slowed from $1.0\times$ to $0.6\times$. Intuitively, a longer path through encoder space should correspond to a more dramatic change in perception as the clip is slowed down (Section 3 presents some data supporting this).

Concretely, we computed a score S proportional to the length of the curve swept by the encoder E in latent space as the clip is slowed down, normalized by the straight-line distance traveled: that is, we define $S(c) = \frac{\int_{r=1.0\times}^{0.6\times} ||dE(c,r)/dr||dr}{||E(c,0.6\times)-E(c,1.0\times)||}$. Then, with probability proportional to $e^{0.2 \cdot S}$, we importance-sampled 200 clips from the set of audio clips of the top 10,000 English words, each spoken by all 16 voices offered by Amazon Polly (spanning American, British, Indian, Australian, and Welsh accents, and male and female voices). The distributions of S in the population and our sample is shown in Figure 2.

Autoencoder details Our autoencoder operates on one-second audio clips sampled at 22,050 Hz, which are converted to spectrograms with a window size of 256 and then flattened to vectors in $\mathbb{R}^{90,000}$. The encoder is a linear map to $\mathbb{R}^{512}$ with ReLU activations, and the decoder is a linear map back to $\mathbb{R}^{90,000}$ space with pointwise squaring. We used an Adam optimizer with lr=0.01, training on a corpus of 16,000 clips (randomly resampled to between 0.6x and 1.0x the original speed) for 70 epochs with a batch size of 16 ($\approx$ 8 hours on an AWS c5.4xlarge EC2 instance).

2.2 Mechanical Turk experiments

Each Mechanical Turk worker was randomly assigned 25 clips from our importance-sampled set of 200. Each clip was slowed to either 0.9x, 0.75x, or 0.6x the original rate. Workers responded with a perceived word and a confidence score for each clip. We collected responses from 574 workers, all of whom self-identified as US-based native English speakers. This yielded 14,370 responses ($\approx$ 72 responses per clip).

Next, we manually reviewed these responses and selected the most promising clips for a second round with only 11 of the 200 clips. Note that because these selections were made by manual review (i.e. listening to clips ourselves), there is a chance we passed over some polyperceivable clips — this means that our computations in Section 3 are only a conservative lower bound. For this round, we also included clips of the 5 words identified by Guan and Valiant (2019), 12 potentially-polyperceivable words we had found in earlier experiments, and "laurel" as controls. We collected an additional 3,950 responses among these 29 clips ($\approx$ 136 responses per clip) to validate that they were indeed polyperceivable.

Finally, we took the words associated with these 29 clips and produced a new set of clips using each of the 16 voices, for a total of 464 clips. We collected 4,125 responses for this last set ($\approx$ 3 responses for each word/voice/rate combination).

3 Results

Are the words we found polyperceivable? To identify cases where words had multiple perceptual "modes," we looked for clusters in the distribution of responses for each of the 29 candidate words. Concretely, we treated responses as "bags of phonemes" and then applied K-means. Though this rough heuristic discards information about the order of phonemes within a word, it works sufficiently well for clustering, especially since most of our words have very few syllables (more sophisticated models of phonetic similarity exist, but they would not change our results).

We found that the largest cluster typically contained the original word and rhymes, whereas other clusters represented significantly different perceptual modes. Some examples of clusters and their relative frequency are available in Table 1, and the relative cluster sizes as a function of playback rate are shown in Figure 1. As the rate is perturbed,

Perceived sound	Playback rate		
	0.90×	0.75×	0.60×
laurel/lauren/moral/floral	0.86	0.64	0.19
manly/alley/marry/merry/mary	0.0	0.03	0.35
thrilling	0.63	0.47	0.33
flowing/throwing	0.34	0.50	0.58
settle	0.65	0.25	0.33
civil	0.32	0.64	0.48
claimed/claim/climbed	0.58	0.34	0.11
framed/flam(m)ed/friend/ find	0.33	0.52	0.43
leg	0.50	0.31	0.10
lake	0.46	0.34	0.14
growing/rowing	0.50	0.47	0.26
brewing/booing/boeing	0.19	0.23	0.26
third	0.40	0.10	0.10
food/foot	0.18	0.29	0.13
idly/ideally	0.38	0.30	0.03
natalie	0.25	0.27	0.09
fiend	0.22	0.34	0.32
themed	0.11	0.17	0.24
bologna/baloney/bellany	0.26	0.00	0.00
(good)morning	0.03	0.28	0.77
thumb	0.66	0.74	0.79
fem(me)/firm	0.06	0.10	0.12
frank/flank	0.72	0.96	0.43
strength	0.08	0.00	0.15
round	0.53	0.38	0.65
world	0.03	0.00	0.14

Table 1: Some polyperceivable words (bold) and their alternate perceptual modes (below). Each row gives representative elements from the mode, and the proportion of workers whose response fell in that mode.

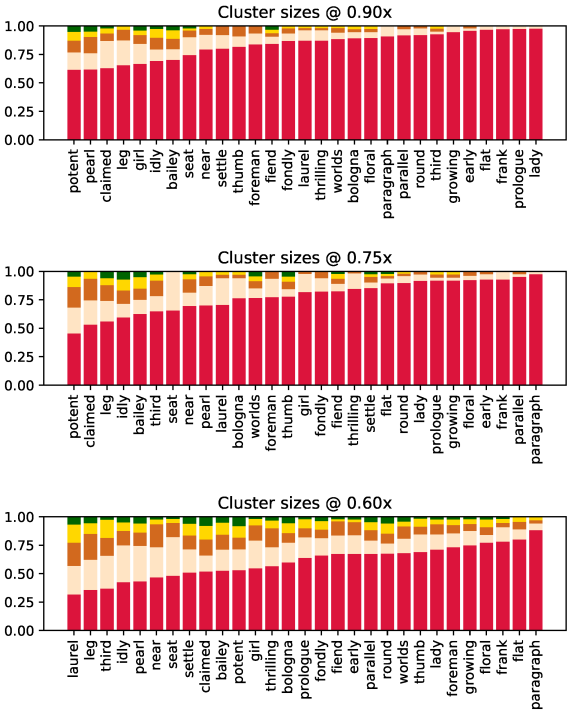

Figure 1: Relative cluster sizes across different playback rates. When the rate is slightly perturbed, the prevalence of alternate modes increases.

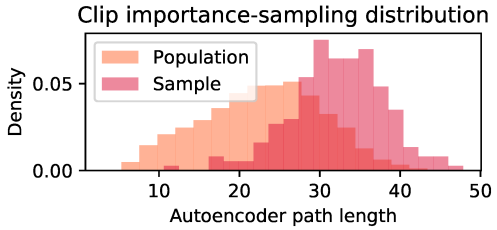

Figure 2: Distribution of path lengths (the S metric) in the population (top 10k English words, all 16 voices) and our sample of 200.

the prevalance of alternate modes among our clips increases.

How prevalent are polyperceivable words? Of our initial sample of 200 words, 11 ultimately yielded compelling demonstrations. To compute the prevalence of polyperceivable words in the population of the top 10k words, we have to account for the importance sampling weights we used when sampling in Section 2.1. After scaling each word's contribution by the inverse of the probability of including that word in our nonuniform sample of 200, we conclude that polyperceivable clips exist for at least 2% of the population: that is, of the 16 voices under consideration, at least one yields a polyperceivable clip for >2% of the top 10k English words.

We emphasize that this is a conservative lower bound, because it assumes that there were no other polyperceivable words in the 200 words we sampled, besides the 11 that we selected for the second round. We did not conduct an exhaustive search among those 200 words, instead focusing our Mechanical Turk resources on only the most promising candidates.

Is S a good metric? We consider the metric S to be successful because it allowed us to efficiently find several new polyperceivable instances. If the 200 words were sampled uniformly instead of being importance-sampled based on S, we would only have found 4 polyperceivable words in expectation (2% of 200). Thus, importance sampling increased our procedure's recall by almost 3×.

For a more quantitative understanding, we analyzed the relationship between "autoencoder path length" S and "perceptual path length" T. Our measure T of "perceptual path length" for a clip is *change in average distance between source word and response* as we slow the clip down from 0.75× to 0.6×. As with clustering above, distance

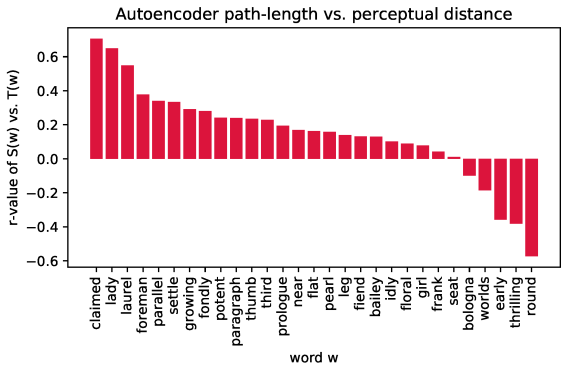

Figure 3: Correlation between S and T across the $n = 16$ voices for each of our 29 words. Nearly all words correlate positively, though with varying strengths (note that "laurel" correlates quite strongly).

is measured in bag-of-phonemes space. For each word, we computed the correlation between S and T among the 16 voices (both S and T vary significantly across voices). For all but 5 of our 29 words these metrics correlated positively, though with varying strength (Figure 3). This suggests that S indeed correlates with polyperceivability.

4 Discussion: Why study quirks of human perception in an ACL paper?

Perceptual instability in human sensory systems offers insight into ML systems. The question of what fraction of natural inputs lie close to decision boundaries for trained ML systems has received enormous attention. The surprising punchline that has emerged over the past decade is that *most* natural examples (including points in the training set) actually lie extremely close to unexpected decision boundaries. For most of these points, a tiny but carefully-crafted perturbation can lead the ML system to change the label. Such perturbations are analogous to the slight perturbation in playback speed for the polyperceivable clips we consider. In the ML literature, these perturbations, referred to as "adversarial examples" seem pervasive across complex ML systems (Szegedy et al., 2013; Goodfellow et al., 2014; Nguyen et al., 2015; Moosavi-Dezfooli et al., 2016; Madry et al., 2017; Raghunathan et al., 2018; Athalye et al., 2017).

While the initial work on adversarial examples focused on computer vision, more recent work shows the presence of such examples across other settings, including reinforcement learning (Huang et al., 2017), reading comprehension (Jia and Liang, 2017), and speech recognition (Carlini and Wagner, 2018; Qin et al., 2019). Studying perceptual illusions would provide a much-needed reference when evaluating ML systems in these domains. For vision tasks, for example, human vision provides the only evidence that current ML models are far from optimal in terms of robustness to adversarial examples. However, while humans are certainly not *as* susceptible to adversarial examples as ML systems, we lack quantified bounds on human robustness. More broadly, understanding which systems (both biological and ML) have decision boundaries that lie surprisingly close to many natural inputs may inform our sense of what settings are amenable to adversarially robust models, and what settings inherently lead to vulnerable classifiers.

Perceptual instability in ML systems offers insight into human sensory systems. Recent research on adversarial robustness of ML models has provided a trove of new tools and perspectives for probing classifiers and exploring the geometry of decision boundaries. These tools cannot directly be applied to study the decision boundaries of biological classifiers (e.g. we cannot reasonably do "gradient descent" on human subjects). However, using standard data-driven deep learning techniques to *model* human perceptual systems can allow us to apply these techniques by proxy.

An example can be found in the study of "transferability." Adversarial examples crafted to fool a specific model often also fool other models, even those trained on disjoint training sets (Papernot et al., 2016a; Tramèr et al., 2017; Liu et al., 2016). This prompts the question of whether adversarial examples crafted for an ML model might also transfer to *humans*. Recent surprising work by Elsayed et al. (2018) explores this question for vision. Humans were shown adversarial examples trained for an image classifier for $\approx$ 70ms, and asked to choose between the correct label and the classifier's (incorrect) predicted label. Humans selected the incorrect label more frequently when shown adversarial examples than when shown unperturbed images. Similarly, Hong et al. (2014) trained a low-dimensional representation of "perceptual space," and used the decision boundaries of the model to find images that confused human subjects.

5 Related work

An enormous body of work from cognitive sciences communities explores the quirks of human/animal sensory systems (Fahle et al., 2002). These works

often have the explicit goal of exploring isolated "illusions" that provide insights into our perceptual systems (Davis and Johnsrude, 2007; Fritz et al., 2005). However, there are few efforts to quantify the extent to which "typical" instances are polyperceivable or lie close to decision boundaries.

Miller (1981) studies the effect of speaking rate on how listeners perceive phonemes. The perceptual shifts studied therein are between phonetically adjacent perceptions (e.g. "pip" vs. "peep") rather than dramatically different perceptions (e.g. "laurel" vs. "yanny"). The "perturbation" of increasing human *speaking* rate is much more complex than simply linearly scaling the *playback* rate of an audio clip. Speaking-rate induced shifts also seem to hold more universally across voices, as opposed to the polyperceivable instances we examine.

6 Future work

Priming effects It is possible to use additional stimuli to alter perceptions of the "laurel/yanny" audio clip. For example, Bosker (2018) demonstrates the ability to control a listener's perception by "priming" them with a carefully crafted recording before the polyperceivable clip is played. Similarly, Guan and Valiant (2019) investigated the "McGurk effect" (McGurk and MacDonald, 1976), where what one "sees" affects what one "hears." The work estimated the fraction of spoken words that, when accompanied by a carefully designed video of a human speaker, would be perceived as significantly different words by listeners. Such phenomena raise questions about how our autoencoder-based method can be extended to search for "priming-sensitive" polyperceivability.

Security implications Just as adversarial examples for DNNs have security implications (Papernot et al., 2016b; Carlini and Wagner, 2017; Liu et al., 2016), so too might adversarial examples for sensory systems. For example, if a video clip of a politician happens to be polyperceivable, an adversary could lightly edit it with potentially significant ramifications. A thorough treatment of such security implications is left to future work.

7 Conclusion

In this paper, we leveraged ML techniques to study polyperceivability in humans. By modeling perceptual space as the latent space of an autoencoder, we were able to discover dozens of new polyperceivable instances, which were validated with Mechanical Turk experiments. Our results indicate that polyperceivability is surprisingly prevalent in spoken language. More broadly, we suggest that the study of perceptual illusions can offer insight into machine learning systems, and vice-versa.

Acknowledgements

We would like to thank Melody Guan for early discussions on this project, and the anonymous reviewers for their thoughtful suggestions. This research was supported by a seed grant from Stanford's HAI Institute, NSF award AF-1813049 and ONR Young Investigator Award N00014-18-1-2295.

References

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2017. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*.

Hans Rutger Bosker. 2018. Putting laurel and yanny in context. *The Journal of the Acoustical Society of America*, 144(EL503).

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE.

Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE.

Matthew H Davis and Ingrid S Johnsrude. 2007. Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing research*, 229(1-2):132–147.

Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. 2018. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems*, pages 3910–3920.

Manfred Fahle, Tomaso Poggio, Tomaso A Poggio, et al. 2002. *Perceptual learning*. MIT Press.

Jonathan B Fritz, Mounya Elhilali, and Shihab A Shamma. 2005. Differential dynamic plasticity of a1 receptive fields during multiple spectral tasks. *Journal of Neuroscience*, 25(33):7623–7635.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Melody Y. Guan and Gregory Valiant. 2019. A surprising density of illusionable natural speech. In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society (CogSci)*. cognitivesciencesociety.org.

Ha Hong, Ethan Solomon, Dan Yamins, and James J DiCarlo. 2014. Large-scale characterization of a universal and compact visual perceptual space. *space (P-space)*, 10(20):30.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Joanne L Miller. 1981. Effects of speaking rate on segmental distinctions. *Perspectives on the study of speech*, pages 39–74.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016a. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016b. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE.

Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International Conference on Machine Learning (ICML)*.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. 2018. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*.

Maya Salam and Daniel Victor. 2018. Yanny or laurel? how a sound clip divided america. *The New York Times*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*.