A Surprising Density of Illusionable Natural Speech

Melody Y. Guan
Department of Computer Science
Stanford University
mguan@stanford.edu

Gregory Valiant
Department of Computer Science
Stanford University
gvaliant@stanford.edu

Abstract

Recent work on adversarial examples has demonstrated a brittleness of many state-of-the-art machine learning systems: most natural inputs can be carefully perturbed to fool the systems. In this work, we investigate one analog for humans, asking *what fraction of natural instances of speech can be turned into "illusions" which either alter humans' perception or result in different people having significantly different perceptions?* First we consider the McGurk effect—the phenomenon by which the perception of what we hear can be influenced by what we see [1]. We estimate that over 15% of words occurring in natural speech have some susceptibility to the McGurk effect. Specifically, for these words, by adding a carefully chosen video clip to the audio channel, the error rate increases by at least 30% over the baseline audio-only listening error rate. We further demonstrate that these McGurk effects can translate to miscomprehension at the sentence level. We then propose attack models based on illusionability and language modeling. Finally we discuss future extensions to our work on the lack of human robustness, such as understanding the frequency of other sensory illusions. As an example, the Yanny or Laurel [2] auditory illusion is not an isolated occurrence; we demonstrate this by generating several very different new instances of this phenomenon. We believe that the surprising density of illusionable instances warrants further investigation, both from the perspective of cognitive science, as well as from the security standpoint.

1 Introduction

A growing body of work on *adversarial examples* has identified that for nearly every machine-learning (ML) system that operates on high-dimensional data, for nearly *every* natural input—including points in the training set—there exists a small perturbation of the point that will be misclassified by the system. It is not necessarily surprising that there exist natural input points that can be adversarially perturbed. What is more surprising, is the density of such points. Even for state-of-the-art computer vision systems that achieve over 95% accuracy on the ImageNet dataset [3], for over 99% of the images there exist perturbations such that: 1) the vision system incorrectly classifies the perturbed image; 2) the confidence of the system in the incorrect response is high; and 3) the perturbations are nearly imperceptible to the human eye [4].

Work on adversarial examples has sparked intense interest for several reasons. First, as has been pointed out by a line of papers from the security community [5, 6, 7, 4, 8, 9, 10], the susceptibility of ML systems to such perturbations may pose an existential threat to deploying ML systems in certain critical settings. More broadly, the apparent brittleness of ML systems to adversarial examples has prompted a re-examination of the conceptual question of what it means to learn. One can question whether current ML systems are truly learning or if they are assemblages of tricks that are effective yet brittle and easily fooled [11]. Implicit in this line of reasoning is the assumption that instances of ”real" learning, such as human cognition, yield extremely robust systems. Indeed, at least in the setting of computer vision, human perception is regarded as the gold-standard for robustness to adversarial examples.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

	Audio Only	Illusory Video	Relative Increase in Error
Words Incorrectly Identified	10.0%	24.8%	+148%
Words Correctly Identified with Low Confidence	2.1%	5.1%	+30%

Table 1: Test results for word-level McGurk illusions among the 147 words predicted to be illusionable. Shown are average error rates for watching the illusory video vs watching the original video, as well as the percentage of words that are correctly identified but sound ambiguous to the listener.

Voice	Sentence	Original Video	Audio Only	Illusory Video
Female	Fay mocked their proof.	0.2	0.6	1.0
Female	That viking mapped Britain.	0.3	0.5	0.6
Male	That's a van theft tool.	0.1	0.4	0.7
Male	Tristian's pie is tart.	0.1	0.0	0.3

Table 2: Test results for sentence-level McGurk illusions, denoting error rates for watching the illusory video, watching the original video, and listening to the audio only.

Evidently, humans can be fooled by a variety of *illusions*, whether they be optical, auditory, or other; and there is a long line of research from the cognitive science and psychology communities investigating these [12]. In general, however, these illusions are viewed as isolated examples that do not arise frequently, and which are far from the instances encountered in everyday life.

In this work, we explore the density of certain classes of illusion, and attempt to understand how susceptible humans' perceptive systems could be to carefully planned "adversarial attacks". As a starting place, we focus on the McGurk effect, which is the well-studied phenomenon by which the perception of what we hear can be influenced by what we see [1]. The most commonly given illustration is that the phoneme "baa", can be perceived as "vaa" or "gaa" if the audio of "baa" is accompanied by a video of someone mouthing "vaa". This effect persists even when the subject is aware of the setup, though the strength of the effect varies significantly across people and varies with factors such as age, gender, languages, and disorders [13, 14, 15, 16, 17, 18, 19, 20, 21]. The McGurk effect has also been observed in multimodal deep neural networks [22].

We find that 1) a significant fraction of words that occur in everyday speech *can* be turned into McGurk-style illusions, and that 2) such illusions persist when embedded within the context of natural sentences. To the best of our knowledge, prior to our work, there has been little systematic investigation of the extent to which the McGurk effect, or other types of illusions, can be made dense in the set of instances encountered in everyday life. The closest work is [23], where the authors demonstrate that some adversarial examples for computer vision systems also fool humans—at least when humans were given less than a tenth of second to view the image. Some of these examples, however, seem less satisfying, as the perturbation essentially acts as an interpolation between the original image and the "incorrect" class. In general, researchers have not probed the robustness of human perception with the same tools, intent, or perspective, with which the security community is currently interrogating the robustness of ML systems.

The density of illusionable instances for humans presents similar types of security risks as adversarial examples present for ML systems. Given the density of McGurk-style illusionable phenomena, malicious attacks are easy to conceive. For example, consider the hypothetical setting where a malicious agent hacks into the video feed at an airport; this agent could then broadcast a carefully crafted (silent) video synchronized with the announcements spoken over the public announcement system, with the result that those people who happen to be glancing at the video screens either cannot understand the announcement, or end up perceiving the wrong message. More pernicious attacks are easy to imagine, especially in a future where much of people's time is spent in front of personalized screens and possibly wearing virtual or augmented reality headsets.

2 Experiments

We began by conducting a preliminary study to determine which phoneme sounds can be paired with video dubs of other phonemes to effect a perceived phoneme that is different from the actual sound. There are 20 vowel phonemes and 24 consonant phonemes in American English although /ʤ/ and /ʒ/ are redundant for our purposes. We created McGurk videos for all vowel pairs preceded with the consonant /n/ as well as for all consonant pairs followed by the vowel /a/, for both a male and a female speaker. Based on labels provided by 10 individuals (3 female, 7 male, ages 22-34) we found that although vowels were not easily confused, there are a number of illusionable consonants (Appendix A, Table 3). We note that that the illusionable phoneme pairs and strength of the effects depend both on the speaker and listener identities. We received Institutional Review Board approval for all experiments in our study (Protocol 46430).

Given this table of illusionable phonemes, the goal was to understand whether these could be leveraged within words or sentences; and if so, the fraction of natural speech that is susceptible. To this end, we sampled 200 unique words (Appendix B, Table 4) from the 10,000 most common words in the Project Gutenburg novels in proportion to their frequency in the corpus. The 10k words have a collective prevalence of 80.6%. We estimated that 147 of the 200 words (73.5%) might be susceptible to the McGurk effect because they contained illusionable phonemes, based on our preliminary study. For these 147 words, we paired audio clips spoken by the female speaker with adversarial video dubs of the speaker saying the words with appropriately switched out phonemes. We tested these videos on 20 naive test subjects (8 female, 12 male, age range 19-36) who did not participate in the preliminary study. Each subject watched half of the words and listened without video to the other half of the words, and wrote down what they heard. They were allowed a total of three plays of each video and were also asked to indicate whether a clip sounded ambiguous.

We found that watching the illusory videos led to an error rate of 24.8%, a relative 148% increase from the baseline of listening to the audio alone (Table 1). Beyond this, even when the responses were correct, the illusory videos made people less confident about their answers, with an additional 5.1% of words being heard correctly but ambiguous-sounding, compared to 2.1% for audio only. For 17.0% of the 200 words, the illusory videos increased the error rates by more than 30% above the audio-only baseline.

We further provided a proof-of-concept that these word-level effects can translate into misunderstood sentences. We created four sentences, two spoken by each of the male and female speakers. We tested watching the videos, listening to audio alone, and watching the original videos on 30 naive test subjects (15 female, 15 male, age range 19-36). Examples of mistakes made by listeners of the illusory videos are listed in Appendix C, Table 5. As expected, the illusory videos generally have the highest error rates and the original videos generally have the lowest error rates, with the audio-only samples falling in between (Table 2).

3 Attack Models

We now illustrate how the surprising density of this phenomenon could be systematically leveraged to produce attacks on human perceivers. We would need to model both the likelihood of words being illusionable and the likelihood of the words occurring in context. The latter is the task of language modeling, which under the n-gram model [24, 25] approximates the probability of a length-m sequence $(v_1,...v_m)$ as:

$$P(v_1,...v_m) = \Pi_{i=1}^{m} P(v_i|v_{i-(n-1)},...v_{i-1}),$$

where v can be phonemes, words, etc. Each token v is assumed to only depend on the previous n tokens.

We might approach the problem of generating McGurk-style illusions as follows. First let function A be the probability that a sequence will be illusionable under McGurk video pairing. We can tokenize words w into their phonemes x_i: $w = <x_0,x_1,...x_k>$. Let q be the likelihood that a given token is illusionable. Then

$$q(w) = \Pi_{i=0}^{k} q(x_i),$$

where $\{q(x_i)\}$ where determined from our experiments and are assumed to be independent of surrounding phonemes. Combining with a language model, we can write:

$$A(w_1,...w_m) = \Pi_{i=1}^{m} q(w_i)^a P(w_i|w_{i-(n-1)},...w_{i-1}),$$

for some parameter $a > 0$ controlling the tradeoff between obfuscation and generating a likely sentence. If we care about the sentence making sense, as we would if we do not want the viewer to be aware that they are being attacked, then we can use a smaller a, and if we want to prioritize confusing or misdirecting the viewer, then we can use a larger a. If we wish to generate a sequence of length m for which we can pair a video that would cause confusion and misunderstanding, we simply need to maximize the above equation.

To train the n-grams' probability distributions, one common method is Kneser-Ney smoothing [26]. Researchers also often use neural networks for modeling the contextual dependencies of words or tokens, e.g. feed forward neural networks [27] and more recently, recurrent neural networks such as LSTMs [28, 29, 30]. The neural networks are trained on textual datasets to predict a probability distribution over the vocabulary V given some fixed-size window of previous words (the context), using stochastic gradient descent with backpropagation [31, 27] and word embeddings [32, 33, 34].

We may also wish to attack perception of words within a given, naturally-occurring stream of audio, such as by hijacking a radio transmission. Here we would like the perturbed word to fit in with both preceding and succeeding adjacent words. One way we can do this is to train a skip-gram model [35], making the network learn the context given a word w_t by maximizing:

$$\sum_{-k \leq j \leq k} \mathrm{log} P(w_{t+j}|w_t).$$

Here, the context of a word comes both from the future and past, so its occurence probability is

$$P(w_t|w_{t-k},...,w_{t+k}).$$

If we wanted to change the meaning at time t, we can change the perception of the actual word w_t by dubbing with a video of the word given by:

$$argmax_{w \in V} q(w)^a P(w|w_{t-k},...,w_{t+k}).$$

Alternatively, to look both before and after in context, one can use a model trained for Cloze ("fill in the blanks") tests [36, 37] or take the product of a forward language model and a backward language model. One benefit these two approaches have over the skip-gram model is that they incorporate positional information of all the words in the context.

4 Future Directions

This work is an initial step towards exploring the density of illusionable phenomenon for humans, and the potential security implications. There are many natural directions for future work, both in the vein of further understanding the security risks posed by McGurk-style illusions, as well as the more broad questions of understanding the weaknesses of the human perception system and how those weaknesses could be exploited by malicious agents, and whether we can make ML systems more robust by better understanding when and why certain human perception systems nonrobust.

One next step in understanding McGurk-style illusions would be to actually implement a system which takes an audio input, and outputs a video dub resulting in significant misunderstanding. Such a system would need to combine a high-quality speech-to-video-synthesis system, with a fleshed-out language model and McGurk prediction model (as described at a high level in the previous section). There is also the question of how to guard against McGurk "attacks". For example, how one can rephrase a passage of text in such a way that the meaning is unchanged, but the rephrased text is significantly more robust to McGurk style manipulations. The central question in this direction is what fraction of natural language can be made robust (without significantly changing the semantics).

More broadly, as the tools for probing the weaknesses of ML systems develop further, it seems like a natural time to reexamine the supposed robustness of human perception. Our work suggests unexpected findings. To provide one example, an audio clip of the word "Laurel" recently gained widespread attention (with coverage by notable news outlets such as *The New York Times* and *Time*). Roughly 50% of listeners hear "Laurel" and the remaining 50% hear "Yanny" or some similar-sounding words, with high confidence on both sides. One of the reasons the public was intrigued is because examples of such phenomena are viewed as extremely rare, isolated instances that we do not expect to naturally occur in speech. In a preliminary attempt to investigate the density of such audio-only phenomena, we identified five more examples of the Yanny or Laurel phenomenon, which are all quite different from the initial example, and from each other.

These examples were generated by examining 5000 words, and selecting the 50 whose spectrograms contain a balance of high and low frequency components that most closely matched those for the word Laurel. Each audio file corresponded to the Google Cloud Text-to-Speech API synthesis of a word, after low frequencies were damped and the audio was slowed 1.3-1.9x. After listening to these top 50 candidates, we evaluated the most promising five on a set of 15 individuals (3 female, 12 male, age range 22-33). We found multiple distributional modes of perceptions for all five audio clips (Appendix D, Table 6). For example, a clip of "worlds" with the high frequencies damped and slowed down 1.5x was perceived by five listeners as "worlds", four as "yikes/yites" and six as "nights/lights". Meanwhile, a clip of "prologue" similarly damped and slowed down 1.9x was perceived by five listeners to sound similar to "prologue", six as "kayak/kayank/kayan", three as "turnip/tienap/tarzan", and one as unintelligible. While these experiments do not demonstrate a density of such examples—and it is unlikely that illusory audio tracks in this style can be created for the majority of words—they illustrate that even the surprising Yanny or Laurel phenomenon is not an isolated occurrence. It remains to be seen how dense such phenomenon can be, given the right sort of subtle audio manipulation.

In conclusion, we hope our work inspires future investigations into the discovery, generation, and density of multimodal and unimodal auditory and visual illusions for humans. Our work also raises the possibility that human vulnerabilities may inherently limit the robustness of machine learning systems, and that some vulnerability to adversarial examples might be inherent to complex learning systems.

References

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

[2] Wikipedia. Yanny or laurel — Wikipedia, the free encyclopedia, 2018. [Online; accessed 22-May-2018].

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[4] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. *arXiv preprint arXiv:1709.04114*, 2017.

[5] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[6] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

[7] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519, 2017.

[8] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2017.

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

[10] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 262–275. Springer, 2017.

[11] Laasya Samhita and Hans J Gross. The "clever hans phenomenon" revisited. *Communicative & integrative biology*, 6(6):e27122, 2013.

[12] James M Hillis, Marc O Ernst, Martin S Banks, and Michael S Landy. Combining sensory information: mandatory fusion within, but not between, senses. *Science*, 298(5598):1627–1630, 2002.

[13] Julia R Irwin, DH Whalen, and Carol A Fowler. A sex difference in visual influence on heard speech. *Perception & Psychophysics*, 68(4):582–592, 2006.

[14] Kaoru Sekiyama. Cultural and linguistic factors in audiovisual speech processing: The mcgurk effect in chinese subjects. *Perception & Psychophysics*, 59(1):73–80, 1997.

[15] Mireille Bastien-Toniazzo, Aurélie Stroumza, and Christian Cavé. Audio-visual perception and integration in developmental dyslexia: An exploratory study using the mcgurk effect. *Current psychology letters. Behaviour, brain & cognition*, 25(3, 2009), 2010.

[16] Linda W Norrix, Elena Plante, Rebecca Vance, and Carol A Boliek. Auditory-visual integration for speech by children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, 50(6):1639–1651, 2007.

[17] Elizabeth A Mongillo, Julia R Irwin, DH Whalen, Cheryl Klaiman, Alice S Carter, and Robert T Schultz. Audiovisual processing in children with and without autism spectrum disorders. *Journal of autism and developmental disorders*, 38(7):1349–1358, 2008.

[18] Kaoru Sekiyama, Takahiro Soshi, and Shinichi Sakamoto. Enhanced audiovisual integration with aging in speech perception: a heightened mcgurk effect in older adults. *Frontiers in Psychology*, 5:323, 2014.

[19] Linda W Norrix, Elena Plante, and Rebecca Vance. Auditory–visual speech integration by adults with and without language-learning disabilities. *Journal of Communication Disorders*, 39(1):22–36, 2006.

[20] Kathleen M Youse, Kathleen M Cienkowski, and Carl A Coelho. Auditory-visual speech perception in an adult with aphasia. *Brain injury*, 18(8):825–834, 2004.

[21] Xavier Delbeuck, Fabienne Collette, and Martial Van der Linden. Is alzheimer's disease a disconnection syndrome?: Evidence from a crossmodal audio-visual illusory experiment. *Neuropsychologia*, 45(14):3315–3323, 2007.

[22] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[23] Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both human and computer vision. *arXiv preprint arXiv:1802.08195*, 2018.

[24] William B Cavnar and John M Trenkle. N-gram-based text categorization. *Ann arbor mi*, 48113(2):161–175, 1994.

[25] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[26] Yee Whye Teh. A bayesian interpretation of interpolated kneser-ney. 2006.

[27] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[28] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.

[29] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[30] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *Association for the Advancement of Artificial Intelligence*, 2016.

[31] Holger Schwenk and Jean-Luc Gauvain. Training neural network language models on very large corpora. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 201–208. Association for Computational Linguistics, 2005.

[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[34] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[35] David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. A closer look at skip-gram modelling. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006.

[36] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[37] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433, 1953.

Speaker	Phoneme	Lip Movement	Perceived Sound	Effect Strength
Female	/b/	/w/	/v/, /f/, /p/	strong
	/ð/	/b/	/b/	strong
	/f/	/z/	/θ/, /t/, /b/	strong
	/m/	/ð/	/nθ/, /n/, /ml/	strong
	/p/	/t/	/t/, /k/	strong
	/v/	/b/	/b/	strong
	/d/	/v/	/v/, /t/	weak
	/l/	/v/	/v/	weak
	/θ/	/v/	/d/, /k/, /t/, /f/, /θ/	weak
	/w/	/l/	/l/	weak
Male	/ð/	/v/	/v/,/θ/	strong
	/m/	/j/	/n/, /l/	strong
	/p/	/t/	/t/	strong
	/t/	/p/	/k/	strong
	/θ/	/d/	/d/, /ð/, /v/	strong
	/v/	/ð/	/ð/	strong
	/b/	/f/	/v/, /f/	weak
	/b/	/n/	/v/, /ð/	weak
	/d/	/p/	/b/, /g/, /p/	weak
	/l/	/p/	/m/, /b/	weak
	/r/	/b/	/b/	weak

Table 3: Illusionable phonemes and effects for the female and male speakers based on preliminary phoneme-pair testing. Where a number of lip movements were available to affect a phoneme, the most effective one is listed. The strength of the effect is based on the proportion of listeners affected.

Appendix

A Illusionable McGurk Phonemes

In Table 3 we show the results of our preliminary study identifying audio-phoneme-video-phoneme pairs that are illusionable.

B Sampled Words for McGurk Study

In Table 4 we list the 200 unique words we sampled from the 10k most common words in the Project Gutenburg corpus and mark the ones which we predicted to be illusionable and for which we made illusory videos.

C Sample Mistakes for Sentence-level McGurk Study

In Table 5 we list the sentences used for the sentence-level McGurk experiments alongside examples of mistakes made by listeners of the illusory videos.

D Results for Yanny or Laurel illusions

In Table 6 we list the results of the Yanny or Laurel illusion experiments.

Illusion Attempted	about	addressed	all	also	and	anyone	arms
	away	bad	be	been	before	behind	besides
	blind	bought	box	brothers	but	by	call
	called	calling	came	child	close	coming	could
	days	dead	did	die	direction	done	else
	end	even	everything	far	features	fell	few
	fighting	fly	for	formed	from	game	gathered
	gave	general	generally	god	good	half	hands
	happened	hath	have	him	himself	idea	information
	july	large	let	letter	life	like	list
	made	many	mass	may	me	meet	men
	months	more	Mrs	my	myself	never	nothing
	of	off	old	one	open	opinion	ordinary
	other	outside	passion	perhaps	please	plenty	point
	possessed	present	put	questions	roof	said	save
	seized	shall	sharp	ship	should	slow	some
	speech	still	successful	summer	terms	than	that
	the	their	them	themselves	there	they	things
	though	time	top	upon	used	very	waited
	was	water	we	went	what	when	which
	will	wisdom	with	working	world	would	wounded
No Attempt	a	act	air	an	any	are	as
	at	change	city	country	eyes	go	going
	hair	has	he	heart	her	higher	his
	house	i	in	into	is	it	its
	king	know	nature	new	no	not	now
	on	or	our	out	rest	saw	see
	seen	she	sorrow	strange	take	talking	to
	turn	who	writing	your			

Table 4: The 200 unique words sampled from the Project Gutenberg novel corpus in proportion to their naturally occurring frequency. The 147 of those for which an illusory video was created are listed on top. Ordering is otherwise alphabetical.

Sentence	**Sample Listener Perceptions**
Fay mocked their proof.	they locked beer proof faith lark bear clueth They knocked their proof
That viking mapped Britain.	That biking matt written bad viking, mat britain that my king not Britain bet fiking met written
That's a van theft tool.	that's a than theft tool That's a then left tool
Tristian's pie is tart.	christian's pie is tart tristen's tie is tart trishan's pie is tarped

Table 5: Sample verbatim mistakes for sentence-level McGurk illusions. Variations on the name "Tristian" starting with the sound /trɪʃ/ were all considered correct.

Word	**Slowdown**	**Perceived Sound**	**N**
worlds	1.5x	worlds	5
		yikes/yites	4
		nights/lights	6
bologna	1.7x	bologna	2
		alarming	2
		alarmy/ayarmy/ignore me	3
		uoomi/ayomi/wyoming	3
		anomi/anolli/amomi	3
		–	2
growing	1.3x	growing	8
		pearling	3
		curling	3
		crowing	1
potent	1.7x	potent/poatin/poden	4
		pogie/bowie/po-ee	3
		pone/paam/paan	3
		power/poder/pair	3
		tana	1
		–	2
prologue	1.9x	prologue	2
		prelude/pro-why/pinelog	3
		kayak/kayank/kayan	6
		turnip/tienap/tarzan	3
		–	1

Table 6: Test results for Yanny or Laurel illusions. Each line displays a cluster of similar-sounding reported words, with counts. Emdashes indicate unintelligible.