Modulated Signal Filtering:
Wouldn't it be Noise?

A Thesis
Presented to
The Division of Philosophy, Religion,
Psychology, & Linguistics
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Aidan M. Mokalla

May 2025

Approved for the Division
(Linguistics)

_____

Sameer ud Dowla Khan

Acknowledgements

I thank, in no particular order, you, Rube Goldberg, Sameer ud Dowla Khan, Sia Furler, Greg Anderson, Babe the Blue Ox, my sister, Paul McCartney, my other sister, the books themselves, my parents, my cats and the Oxford comma.

Preface

∗ḱléwos n̥-g^{wh}í-t-om

(Kuhn, 1853; Schmitt, 1967)

Table of Contents

Abstract . . . . . **vii**

1. Introduction **1**

2. Background **5**
2.1 Language Perception in the Presence of Noise 6
2.1.1 Noise across Linguistic Domains 8
2.1.2 Managing Noise 12
2.1.3 Likelihood and Prior Expectations 16
2.1.4 The Noisy Channel Model 18
2.2 Expecting Noise 21
2.2.1 Error Identification Signal 21
2.2.2 Context-Dependent Likelihoods 22
2.2.3 Expecting Noise in Music 23
2.2.4 This Thesis 29

3. Methods **31**
3.1 Participants 31
3.2 Poliak, Kimura, and Gibson (2024) 31
3.2.1 Materials 32
3.2.2 Forced Choice Procedure 32
3.2.3 Likelihood Metric 33
3.2.4 Prior Metric 33
3.3 Free Response Procedure 34
3.4 Contextual Likelihood Metrics 34
3.5 Contextual Prior Metrics 36

4. Results **39**
4.1 Attention Check 39

4.2 Inferences . . . 40
4.2.1 Forced Choice vs. Free Response . . . 40
4.2.2 General vs. Contextual Metrics . . . 42
4.3 Effect of Song Familiarity . . . 44
4.4 Prior Results . . . 45
4.5 Model Parsimony . . . 45

5. Discussion . . . 49
5.1 The Noisy Channel Processing Framework . . . 49
5.1.1 The Intended Message . . . 49
5.1.2 The Perceived Message . . . 50
5.1.3 Contextualizing Modalities . . . 51
5.2 Distributions on Perception . . . 51
5.2.1 Expectability as EIS Variance . . . 52

6. Conclusion . . . 55
6.1 Future Directions . . . 56
6.1.1 Direct Cross-Modal Comparisons . . . 57
6.1.2 Individual Differences in Adaptive Capacity . . . 58
6.2 Computational Linguistics and the Convergence of Formal Systems 59

7. Appendix . . . 61

Cited Works . . . 67

Abstract

JUST as visual perception adjusts to lighting conditions or auditory perception compensates for background noise, linguistic interpretation mechanisms recalibrate based on communicative context. This recalibration is described and evidenced as both a quantitative adjustment and a qualitative shift in how communicative intent maps to acoustic signals and in how acoustic signals map to linguistic representations. This thesis extends recent findings by accurately modeling how listeners adapt their phonological expectations when processing *sung* speech, allowing me to compare the unique benefits and challenges of producing and perceiving language in the communicative context of music. In doing so, I introduce new methodological tools for future researchers, and shed light on broader questions of Language as a process wholly determined by communicative intent. I also replicate the findings of Poliak et al. (2024) on predictors of song lyric (mis)interpretation, and demonstrate that a model that accounts for the context-specific peculiarities of sung speech significantly improves these predictors. I conclude discussing what these results tell us about how and when the human language processing systems use expectations to flexibly adjust to varying environments within the Noisy Channel Processing model, as well as the relevance of this perspective to understanding more general semiotic endeavors.

Dedication

This thesis is dedicated to the children of Gaza.

Introduction

WHEN we listen to song lyrics, what we think we hear is often not what was actually sung. This phenomenon, known as a "mondegreen," offers a fascinating window into how the human brain processes language under noisy conditions. While lyric misinterpretations are sometimes funny, they also reveal important principles about language processing that extend far beyond music. This thesis investigates how listeners make sense of speech in challenging acoustic environments, with a specific focus on how the unique properties of sung speech affect perception and interpretation of lyrics.

Song lyrics present distinctive perceptual challenges: vowels are often centralized during singing, consonants are articulated differently, musical accompaniment masks portions of the signal, and melodic constraints alter natural prosody. These factors create predictable patterns of noise that may require or be facilitated by specialized processing mechanisms. By studying how listeners interpret lyrics under these conditions, we can better understand the flexible, adaptive nature of human language processing.

This research contributes to our understanding of context-dependent language processing by testing whether listeners employ specialized mechanisms for different linguistic contexts with predictable noise patterns. While traditional models of speech perception often assume uniform processing across contexts, evidence suggests that the brain adapts its strategies based on contextual expectations. The Noisy Channel Model (NCM) provides a framework for understanding this process, which provides an explanation for how listeners combine prior linguistic knowledge with perceptual input to derive the best possible interpretation of a noisy message.

This thesis investigates two primary hypotheses. First, I hypothesize that par-

ticipants will consistently identify lyrics more accurately during forced-choice tasks than during free response tasks. This prediction reflects methodological concerns about how lyric interpretation has been studied. While forced-choice paradigms are experimentally convenient and defensible, they may artificially constrain responses and inflate accuracy by providing a limited set of options, one of which will always be correct. Free response tasks more closely approximate natural listening conditions and may reveal different patterns of perception and interpretation.

Second, I hypothesize that listeners' linguistic processing of noise in contextually modulated speech is better explained by specialized mechanisms than by generalized mechanisms. Specifically, listeners use the contextually-derived fact that they are listening to music to anticipate specific types of linguistic noise, and as a result, the Noisy Channel Model makes more accurate predictions when parameterized by context-relevant metrics than when parameterized by generally applicable metrics. This prediction is borne out of evidence that listeners adapt their perceptual strategies based on context-specific expectations.

To test these hypotheses, I conducted an experiment with 70 English-speaking monolingual participants who listened to short clips from 37 English-language songs. Unlike previous studies that relied solely on forced-choice methods, participants completed both free response tasks (typing what they heard without prompting) and forced-choice tasks (selecting from four options). This design allows direct comparison between the two response formats within the same participant population while allowing validation of new results.

I developed context-specific metrics for both parameters of the Noisy Channel Model: *prior probability* and *likelihood*. To measure a lyric's prior probability, I compared three large language models with varying degrees of specialization in song lyrics. For likelihood estimation, I created a feature-based distance metric that accounts for the specific patterns of phonetic variation in sung speech, in contrast to the coarser Levenshtein distance used in previous research.

My findings ultimately demonstrate that listeners use specialized processing mechanisms tailored to the unique constraints of sung speech, rather than applying the same general strategies used for spoken language overall. The staggering improvement in descriptive accuracy when using context-specific metrics provides compelling evidence that the human language noise processing system flexibly adapts to varying communicative environments.

This thesis is organized as follows: Chapter 2 provides a comprehensive review of language perception in noisy environments, exploring how noise manifests across linguistic domains and how listeners manage to derive meaning in these noisy contexts. The chapter introduces the Noisy Channel Model and examines the unique characteristics of sung speech that affect perception. Chapter 3 contains my experimental methods, including participant selection, stimuli, experimental procedures, and the development of context-specific metrics. Chapter 4 presents my results, comparing performance across task types and evaluating the predictive power of different metrics' combinations. Chapter 5 discusses the implications of these findings for theories of language processing, identifies limitations, and suggests directions for future research. Before the bibliography is a short review of some mathematical concepts related to this thesis.

Background

How can we make sense of unclear signals? Much research has addressed and formalized the ways in which the human linguistic system is capable of successfully deriving sense. These frameworks can often be understood as an explanation of how linguistic input is processed from one abstract domain into another. In other words, theories of linguistic competence can be viewed as stipulating the necessary conditions of how linguistic "signal" is processed from one abstract domain into another. Depending onto which domains it relates, this processing is sometimes called "realization," "derivation," "assimilation," "inflection," "compounding," "merge," "composition," "type-shifting," "inference," "style-shifting," etc.

Many modern media, e.g. the text before you, are meant to facilitate accurate reconstruction of its author's original communicative intent.[1] However, things aren't always as *clear* in the real world. Both internal cognitive and external environmental factors can diminish the accuracy with which communicative intent is ultimately transmitted, processed, and understood by a message's recipient (i.e the "listener"). I use the term *noise* to refer to any of these factors. We should, then, also endeavor to describe the effect of noise on this processing, and what happens when such errors occur. For example, how does a mistake in one domain of linguistic processing percolate into others? When such mistakes occur, what happens next? Does one simply give up? When we don't simply throw up our hands and walk away when errors occur, *how do we recover from them?*

This thesis addresses how listeners recover from such errors. Our broad fundamental premise, called *rational integration* (Gibson et al., 2013), is that a listener always "selects" an interpretation of someone's intended message as whichever

[1] Notice that text's transmission of communicative intent is imperfect as well. Recent extensions to this medium can disambiguate some of these errors 😈!!

message is most likely to be correct, given everything that the listener knows (about what was said, about the world, about our languages, etc.). The Noisy Channel Model (NCM) notes that the maximally likely message is always equal to the message with the highest *likelihood* and *prior probability* of being correct.[2] If we can construct an accurate approximation of these values, then rational integration predicts that we should be able to model what happens when communication falters with complete accuracy. Specifically, it predicts that all interpretations, including misinterpretations, have high values for at least one of these two factors. The best way to construct this approximation is an open question addressed in this thesis. Specifically, I interrogate whether listeners' expectations of noise are affected by modal factors, which we can observe by comparing approximations of these values when we account for mode-specific noise distributions versus when we account for only the types of noise that speakers are exposed to in general. While recent work has shown that general approximations of these probabilities make not insignificantly accurate predictions about how noise is processed in the mode of song lyrics, I address whether a parameterization of the NCM that accounts for listener expectations in musical settings better models and predicts how they reconcile noisy messages.

2.1 Language Perception in the Presence of Noise

Our working definition of *noise* is interference in the linguistic communication channel. So, noise means any interference in a speaker's communication with a listener, such as before production, after perception, or in transit between. Because there are so many aspects of the communication process where noise can interfere, noise must exist across all linguistic domains, in various forms. For example, comprehenders frequently encounter noisy input, ranging from syntactic ambiguities and mispronunciations during production (Poppels & Levy, 2016) to environmental sounds and one's own attentional lapses during perception (Poliak et al., 2024). Speakers' linguistic systems can also unintentionally impart noise onto their own intended messages before they're even produced, such as when phonological categorization interferes with accurate perception and reproduction (Liberman et al., 1957). In general, recovering and understanding an intended message becomes increasingly challenging when the message is distorted by noise (Shannon, 1949).

[2] I.e. the message for which the product of these two factors is highest.

Many linguistic theories, e.g. most morphological, syntactic, semantic and pragmatic frameworks, are forced to overlook the pervasive nature of noise by assuming that the listener perceives an idealized, noise-free input. They do so in one of two ways: While theories of linguistic performance tend to focus on those factors that facilitate communication rather than those that dampen it, theories of linguistic competence retreat from such questions altogether by assuming that the abstract mental principles that govern production and perception can only relate to discrete units of meaning. For example, when a morphologist asserts that a certain set of affixes must combine with a stem before another set of affixes, they are not simply stating a necessary condition for competent affixation without regard for the nature of the input and how it is perceived. They are in fact inherently assuming something very important about the input, which is that discrete affixes *exist*. This is not to assert any flaw in the methodology of linguistic theories of competence, it is merely to point out that statements of the type "in order to make a correct C, a speaker must combine a and b according to the rule f" always must inherently assume the input, a and b, are discrete elements of f's domain, and ones that are independent of the rules applied previously to generate a and b in the first place. This assumption may even be true, but regardless, it's one we need to remain aware of. This assumption is inherently embedded when we begin our linguistic analyses from discrete items (e.g. phonemes, morphemes, constituents, denotations, entities), and proceed from there to model how these elements interact with one another. In the cases when this assumption does not hold, such as when noise has corrupted the telegraphing of discrete representations, we ought to also have a theory of what happens next. When we do admit a notion of imperfect communication, we often assume that at most only one type of noise is present at a time (e.g. phonological, lexical, syntactic or semantic ambiguity), without regard for how a collection different species of noise may interact with one another at the same time.

Recent research has began to investigate the role noise plays in explaining and shaping language processing across different linguistic domains, from the fine-grained details of phonetics and phonology to the higher-order complexities of semantic interpretation. Related bodies of work investigate the many varied strategies for repairing a noisy signal that exist at each of these levels of linguistic analysis, some of which are reviewed in the next sections. Later sections review the Noisy Channel Hypothesis (Levy, 2008), which is an abstraction of a repair technique that is hypothesized to exist across all of these domains simultaneously

based on the premises of rational integration. But for now, before we can understand how a noisy message is repaired, we first explore how noise manifests across various levels of analysis in the first place, examining how it degrades both the production and perception of language.

2.1.1 Noise across Linguistic Domains

The phonetic and phonological levels of linguistic analysis aim to describe the production and perception of individual speech sounds. Noise at this level often manifests as deletions, distortions, or other misperceptions of these fundamental units. Unexpected distributions perceived across a listener's space of phonemes, such as those we experience when we listen to someone speak with a different accent than our own, are a common type of phonetic and often phonological noise (Hallé et al., 2004).

Research shows that even seemingly subtle phonetic shifts, such as those introduced by singing instead of speaking, can greatly impact intelligibility. For instance, singers tend to centralize their vowels, making their pronunciation less precise since these vowels are more difficult to delineate and distinguish from one another (Benolken & Swanson, 1990), both during production and perception (Smith & Scott, 1980), and especially at higher pitches (Hollien et al., 2000). Additional types of phonetic and phonological noise generated by singing, such as de-emphasizing plosives (Vurma et al., 2023), pitch-modulated perceptibility of vowels in certain contexts, centralized vowels (Hollien et al., 2000; Collister & Huron, 2008), prosodic divergences (Johnson et al., 2014), and musically-induced procedural lethargy (Wickham, 2013), are discussed in the section below.

Noise is not limited to production. Noise, specifically cognitive noise, can also affect perception. This is because we are understanding noise to mean *anything* that contributes to the difference between a speaker's intended message and a listener's best guess at what that intended message was. At the lexical level, noise can involve misidentifying, dropping, or substituting entire words. This phenomenon was first notably highlighted by Wright (1954) in the context of song lyrics. In her essay "The Death of Lady Mondegreen," Wright describes the experience of discovering that the lyric she had remembered since her childhood, *"And Lady Mondegreen,"* was incorrect. While she had remembered the lyrics of the Scottish ballad "The Bonnie Earl o' Moray" to be

"Ye Highlands and ye Lowlands,
Oh, where hae ye been?
They hae slain the Earl Amurray,
And Lady Mondegreen."

the final line in fact sings "*And* laid him on the green." Wright first termed this phenomenon of lexically misrepresenting a song lyric a *mondegreen*.

Syntactic processing is also susceptible to noise, which leads to potential misinterpretations of grammatical structures when noise is present. A key area of research within this domain involves examining how and when listeners reinterpret utterances that are pragmatically implausible but syntactically well-formed. Researchers, when comparing utterances with identical syntactic structure but unequally plausible meanings, found that listeners are likely to apply syntactic noise repair strategies when they would yield more pragmatically plausible interpretations (Gibson et al., 2013; Chen et al., 2023). Concretely, they showed that a listener is more likely to interpret a sentence like

"The mother gave her candle the daughter"

as meaning

$$\mapsto [\![\text{The mother gave her candle } \textit{to} \text{ the daughter}]\!]$$

than they are to interpret the sentence

"The mother gave her daughter the candle"

as meaning

$$* \mapsto [\![\text{The mother gave her daughter } \textit{to} \text{ the candle}]\!].$$

Here we take an "edit" to mean any operation that takes one string to a different but highly similar string, i.e. element deletions, element insertions, and element swaps. Regardless of the domain of repair, listeners are especially likely to reinterpret such utterances if the correction involves fewer edits, and particularly if it involves deletions rather than insertions (Gibson et al., 2013). This pattern aligns with the Bayesian "size principle," which suggests that simpler explanations are generally preferred. This pattern also aligns with empirical observations of how often each of these three types of edits occurs during production, and with the observation that the types of environmental noise that affect spo-

ken human language are more prone to interrupt or dampen the signal than they are to insert something into it. This particularly emphasizes, then, the effect of semantic noise demonstrated above: listeners which would prefer to repair semantically improbable denotations even if the repair is an insertion. Poppels & Levy (2016) took a closer look at noise caused when they swapped elements in an utterance with one another, intentionally creating a type of noise called *exchange errors*. They found, by similar methods as Gibson et al. (2013), that this sensitivity to repairing with certain types of edits versus others also appeared to be *structure-sensitive*, meaning that comprehenders are more likely to expect exchanges (swaps) between certain types of elements (e.g. prepositions) than others (e.g. nouns), where they would prefer to repair by means of deletion or insertion. This implies that comprehenders' syntax noise models (e.g. their approximation of the probability that an input has been affected by certain types of noise versus others) aren't simply string-based evaluations of how many elements may have been swapped, deleted, or inserted within an utterance, but rather must also incorporate probabilistic knowledge about syntactic constituency and common error-inducing syntactic noise patterns in language.

These findings suggest that listeners tailor their noise repair strategies at the levels of syntax and the syntax-semantics interface by means of rational probable inference, and that speakers appear to select repairs to noisy inputs by probabilistic evaluation of which type(s) of noise are likely to have interfered with perception. This conclusion was replicated and concretely extended into the domain of pragmatics and semantics by Chen et al. (2023). They found that the rate at which listeners make these syntactic repairs (resulting in an unfaithful interpretation) decreases when listeners were first primed with pragmatic information that supports the faithful but implausible semantic interpretation of a utterance. For example, first telling the listener that "This sentient candle has cursed the mother, and she must sacrifice someone's firstborn child to the candle if she wants to break the curse" might make the listener more likely to demonstrate a faithful interpretation of the first sentence above with a double-object construction. They found this rate also decreased when listeners believed that higher levels of syntactic noise (e.g. higher levels of producing unintended syntactic errors) were present when the utterance was first produced, independently of the utterance's actual content. They quantified the level of noise that the listeners would estimate an utterance to have undergone as the number of syntactic edits necessary to make an utterance's interpretation match the implausible yet faithful interpretation. Both of these effects suggest that contextual awareness plays a part in how noisy-channel

repairs are applied at the semantic level as well, and that listeners apply fewer noise repairs to some domain when they have reason to believe that the source of the noise stems from another. For example, when a listener is first primed with a pragmatic context that supports an implausible faithful semantic interpretation, listeners assume that the cause of any noise detected upon hearing "The mother gave her candle the daughter" is more attributable to *dissonance* between their global pragmatic expectations and their local contextual understanding, rather than to syntactic deletion as it would have otherwise been without the pragmatic priming.

This second effect, where listeners apply less syntactic repair as the amount of perceived syntactic noise increases, was modulated by the type of syntactic noise being evaluated. Listeners were more likely to apply semantic repair to derive plausible, non-literal interpretations when the speaker swapping the locations of the theme and the recipient is the only type of syntactic noise that could have been applied to the intended message to yield the perceived message. Inversely, listeners were less likely to apply semantic repair, and instead apply syntactic repair, when the necessary syntactic noise was the speaker inserting a preposition. Finally, listeners were least likely to apply semantic repair over syntactic repair when the necessary syntactic noise would have been the speaker deleting a preposition (as in the first example above). In sum, Chen et al. found that listeners are more likely to not use syntactic repair and to instead parse an utterance's syntax literally 1) if the literal interpretation is plausible, 2) if they were primed with a supporting pragmatic context, or 3) if the utterance requires repairing uncommon types of noise, like swaps, instead of more common types of noise, like deletions and insertions. They took these findings to mean that listeners use the local discourse context as well as global models of noise to integrate their *expectations* of noise into their repair strategies.

All of these findings suggest that listeners, by some mechanism, integrate contextual information with their latent linguistic knowledge to arrive at the most locally plausible interpretation, across many, perhaps all, levels of analysis. Without realizing, listeners demonstrate that they actively engage in rational inference, combining their perception with a sophisticated understanding of language and the world to decipher the intended message. This process involves integrating multiple sources of information, including local and global knowledge of linguistic production and perception, and local and global knowledge of how various types of noise may affect each.

2.1.2 Managing Noise

Effectively managing sensory input is crucial for successful language comprehension, especially in the presence of noise (Cherry, 1953; Norris et al., 2000). In our ongoing quest to understand understanding, much research has attempted to shed light on the various strategies utilized by listeners to filter noise and extract the intended message from the speech signal. This subsection reviews the literature on how the human brain efficiently manages sensory input, particularly in the context of noisy speech perception.

Many cognitive processes are categorized by those studying them as either "top-down," "bottom-up," or some composition of both (Clark, 2013).[3]A bottom-up cognitive process is one that is applied to sensory inputs across a specific domain (Kintsch, 2005), such as the pain you feel when you get a splinter. A top-down cognitive process is simply a bottom-up process wherein first "a person can choose at will" which sensory inputs will be ignored before applying the bottom-up process to them (Theeuwes, 2010), such as choosing to attend to the speech of an interlocutor (in a bottom-up fashion) and consciously suppressing the speech of someone else in the background. Equivalently, it's a bottom-up cognitive process that is selected and steered by an explicit "memory or knowledge component" (Kintsch, 2005).

This framework can be applied to the processes that attempt to remove noise from the auditory signal as well. Different lines of inquiry suggest that both top-

[3] One controversial claim that has gained ground across diverse disciplines in recent memory is that *all* cognitive processing can be described as fundamentally bottom-up, but that the operation crucial to differentiate top-down from bottom-up qualia, namely that of "deciding" which percepts are attended to, is an often helpful yet perhaps illusory humanistic abstraction. In this view, "deciding" is posited instead as the process of *predicting* which percepts would induce action most closely matching the expectations generated by the memory/knowledge component when those percepts are attended to. The bottom-up setting is then converted to the specific case wherein all percepts are always attended to, just to different degrees at different times. In turn, this view does not require us to assume the existence of a unique conscious or rational deciding entity that's independent from this memory/knowledge component (Augustine of Hippo (397-400); Spinoza (1677); Hume (1739-1740); Kant (1781); Schopenhauer (1818); Libet et al. (1983); Dennett (1991); Johnson (1997); Matuschak (2012); Clark (2013), *inter alia*). Aside from two of the fore-mentioned citations, one interesting question this perspective can't always precisely answer is how the memory/knowledge component generates these predictions in the first place. How could such a system efficiently make predictions over such a large action space? These concerns may be gracefully sidestepped in the classical humanistic view with whims of instinct, souls, or conscious rational deduction, which can range from contradictory to self-referential.

down and bottom-up processes are used to filter noise from the speech signal. For example, bottom-up strategies may help resolve more fine-grained, predictable, low-entropy types of noise, while top-down strategies may be called upon to help resolve more complex, significant, unpredictable, or high-entropy types of noise.[4] I explore these cases below.

The ability to passively attend to relevant information amidst background noise is a fundamental bottom-up aspect of human linguistic cognition, as evidenced by the well-known "Cocktail Party Effect" (Cherry, 1953). One's ability to hear their own name being spoken by someone across the room in the midst of a noisy and crowded cocktail party demonstrates the brain's capacity to focus on a specific sound source, such as a speaker's voice, while filtering out irrelevant auditory input. This selective attention mechanism, which the listener is not always perceptually aware of, is crucial for successful speech perception in noisy contexts.

Research using dichotic listening tasks, where different messages are presented to each ear, supports the existence of top-down processes in pruning certain types of noise from the signal. Work like Cherry (1953) has also demonstrated that listeners can effectively attend to one ear while ignoring the other. This suggests a top-down mechanism for segregating and selecting relevant auditory streams based on low-level acoustic features, in addition to the empirically supported bottom-up

[4] One kind of event has higher entropy than another if the first type of event occurs in a more uniform, consistently unpredictable way. For example, consider a fair die. This die has high entropy, because all of the possible outcomes have equal probabilities, making it (un)predictable in a consistent way. As an example of lower entropy, consider a weighted coin as well. This unfair coin lands on tails 5 out of the 6 times it's flipped, while the die lands on each of its 6 faces uniformly randomly. The entropy of rolling the die is higher than the entropy of flipping the coin, because the outcome of rolling the die is less predictable, compared to the more predictable outcome of flipping the coin. The die has six equally likely outcomes, leading to higher uncertainty and hence higher entropy. In contrast, the coin, being weighted to favor tails, has a more predictable outcome, resulting in lower entropy than the die.

In linguistic terms, for example, the phonemic signal has lower entropy than the syntactic signal. Specifically, noise caused by syntactic ambiguities during t seconds of speech has higher entropy than noise caused by phonemic ambiguities during t seconds of speech. The latter type of noise has more opportunities to occur in t seconds than the former, making the *amount* of noise more accurately predictable per a fixed length of time. Concretely, we expect the amounts of phonemic noise between two equal length but otherwise arbitrary utterances to be more similar than the similarity in the amounts of syntactic noise between the same utterances, all else being equal. However, the *probability* that phonemic noise occur is not necessarily any different that the *probability* that syntactic noise occur, just as the probability that our die lands on a given face is the same as the probability that our coin lands on heads.

cocktail party effect.

For example, Cherry found that listeners have relatively little comprehension difficulty when asked to separate disparate audio streams, such as two different speakers' voices or two distinct utterances entirely. In these contexts, there is a large amount of noise that the listener must remove from the signal. However in these contexts, the listeners were aware of the noise, and perhaps because they were able to predict it well, had better performance. Interestingly, he also found that listeners have *more* difficulty in some contexts where the noise is made weaker. For example, he found that when the two ears receive the exact same stimuli, but with a small delay in the speech entering one ear after the other, comprehension plummets. If the delay were 0 across both ears (i.e. if the stimuli were synchronous), or if the delay between the stimuli reaching one ear and reaching the other was large (greater than 2 to 6 seconds), then listeners' ability to account for the noise returned. However these results must be taken with a grain of salt in light of the quality of the recording equipment available at the time:

> *"The reversed speech was identified as having* 'something queer about it' *by a few listeners, but was thought to be normal speech by others."* (Cherry (1953), p. 978)

Cherry also found that when the same signal switched between both ears at a regular interval, comprehension also decreased. He found that comprehension was worst when the same signal alternated between ears approximately 13 times each second, and that if the alternation was slower than this, or faster than this, intelligibility increased across participants. When participants were not able to reproduce the words from the alternating stimuli, they were usually still able to correctly identify the language as English, or pick out a few words. When the participants were able to account for the noise induced by the rapid alternations between ears, they reported that they were able to do so by "listen[ing] as though to both ears simultaneously." However, unlike when the alternations slowed, the participants "varied in their ability [to correctly identify the alternating stimuli] considerably" when the alternations accelerated past 13 per second (Cherry (1953), p. 979).

Even when listening to speech in an ideal environment, the physical signals that actually reach the two ears are quite different from each other. For example, differences in signal timing and intensity between the two ears are what allow us to understand the direction from which speech arrives, and account for certain

types of noise. The delay in sound wave propagations reaching one ear a few milliseconds before the other ear actually helps the listener to decode certain parts of the signal (Hirsh, 1948).

These observations along with the subjective reports of the listeners themselves suggest that the listeners in Cherry's experiment were able to filter the alternation noise by recruiting the cognitive circuitry that typically is active when processing speech from both ears to account for the differences in the two signals. That is, the successful participants were able to transition from a constrained, top-down process of attending to one ear and then to the other, to a more relaxed bottom-up process of attending to both ears simultaneously. This is similar to how in the first experiment, listeners were more easily able to account for those types of noise that they were already accustomed to accounting for, i.e. when the two ears are receiving markedly different signals at a given moment. Here, the top-down task of ignoring the irrelevant audio stream became easier as the streams became more different. When the noise instead took the form of the same signal with a very small time delay, the listeners had no experience or mechanism to rely upon, since there is not strong pressure to develop a resource-greedy bottom-up cognitive process that parses near-identical audio streams but with roughly 80 milliseconds of delay. This type of noise, even though the amount of information-theoretic noise in the signal is in one sense less than the amount of noise caused by a longer delay, is not something that the listeners would often encounter in a natural setting. In sum, both the quantity and quality of noise affect speech processing accuracy.

Another prominent example of top-down processing is the phenomenon of phonemic restoration. In phonemic restoration, listeners perceptually 'fill in" missing or distorted sounds based on their knowledge of words and sentence structure (Ganong, 1980). Predictions about sentence structure and meaning can even guide the interpretation of lower-level phonetic information in the presence of acoustic distortions, but low-level phonetic features are still the most consistent predictor of phonemic (mis)interpretation in settings like these (Samuel, 1981). This ability to reconstruct missing phonetic information demonstrates a combination of top-down and bottom-up influences of prior phonemic, morphological, and syntactic knowledge on speech perception.

The concept of predictive coding provides a strong theoretical framework for understanding how the brain alternates between using top-down and bottom-up strategies to manage sensory input, including noisy speech. Predictive coding

models propose that the brain continuously generates predictions about incoming sensory information based on its internal models of the world. These predictions are compared with the actual sensory input, and any discrepancy between the predicted and actual input generates a prediction error signal (Clark, 2013).

This prediction error signal must then serve some sort of role in updating internal models of noise prediction and filtering, leading to more accurate predictions in the future. In the context of speech perception, predictive coding suggests that the brain anticipates upcoming sounds based on linguistic context and prior knowledge. This predictive processing reduces the processing load by effectively "explaining away" expected sensory input in a bottom-up fashion, allowing the brain to allocate resources to schedule processing of unexpected or unpredictable information.

Regardless on one's stance on the precise role of prediction in cognition, the findings reviewed in this section demonstrate that a strict information-theoretic notion of "noise" is not enough to account for why certain messages are more difficult to interpret correctly than others. We see that listeners apply certain strategies tailored to deal with certain types of noise, rather than only applying a static, passive, bottom-up, or context-oblivious "filter" over the audio input. This means that simple explanations simply do not serve our desires to understand how noise is dealt with in general. Instead, we observe that listeners employ a dynamic battery of noise-removing strategies, and most adeptly employ bottom-up processes to those types of noise that they are likely to have experience experiencing. This thesis provides evidence that directly supports this claim by showing that listeners employ noise removal strategies that are tailored specifically to the types of noise that occur in song lyrics.

2.1.3 Likelihood and Prior Expectations

As I discussed in the last section both local, discourse-specific and global, exemplar-based statistical considerations are necessary to give an accurate prediction of how a listener interprets a noisy message. I'll now consider a concrete example of how we might resolve a specific instance of noise interfering with communication, given this background. Then I explore how the details of this process can be represented using the probabilistic tools covered in the appendix. We'll discuss this specific example instead of other types of noise because the pragmatic domain of lin-

guistic processing is more integrated within phenomenal consciousness than other domains (Mazzone, 2013). As such you, the reader, are able to more accurately evaluate my account of how this noise is parsed, and generalize appropriately, since you have relevant conscious experiences (which we might call "top-down processing" in later sections) to compare against. If we instead considered, say, resolving phonetic noise, a process of which you are not consciously aware, my gestures towards the processes that your subconscious uses to resolve this noise couldn't be meaningfully evaluated.

Say you and I are sharing a romantic evening at the circus, the animal act is nearing its finale, and we're currently watching a 6,000 kg African savanna elephant walk in circles and such. I lean in close to your ear and whisper the words "I'm uncomfortable with the elephant in the room." You now must select between two options: the literal interpretation or the idiomatic one. This is an example of pragmatic noise since the signal is ambiguous as to which pragmatic frame is correct. Your ultimate interpretation would 1) depend on the global understanding that what I said matches an idiom which you know has some specific meaning, and 2) depend on the local, contextually nuanced understanding that my affinity for rearing illegal elephant herds has recently put a bit of a wrench in our relationship, which we perhaps need to discuss. This global and local knowledge work in tandem and likely resolve to the idiomatic interpretation: both pieces of information suggest that there could be a difficult topic that the two of us need to broach. The literal interpretation of my statement does suggest alternative interpretations, e.g. that the way Jumbo has been staring directly at me with a troublesome look of deceit in his eyes is making me genuinely uneasy. However your context-specific knowledge of my illicit activities contradict any interpretation wherein I am uncomfortable with elephants. There are of course more than these two possible interpretations, e.g. perhaps I'm trying to make a joke of some sort. However this same process of integrating different types of expectations and knowledge can of course be carried out over many possibilities as well, leading you to eventually settle on one. Thus in this example, you use the global and local information available to you to determine which interpretation is most congruent with both. In other words, you determine the interpretation that is most likely to be true if you assume that both your local and global understandings of the world are also true.

2.1.4 The Noisy Channel Model

According to the Noisy Channel Model (NCM) of linguistic processing (Levy, 2008), to help parse noisy messages, a listener uses a mental approximation of a probability function $\Pr : \mathcal{M} \to [0, 1]$ over all possible messages $\mathcal{M}$ (if you would like a brief review of the mathematical constructs used in this thesis, please pay a visit to the appendix before proceeding to this section). This function assigns a score between 0 and 1 to all messages, based on how likely a speaker would be to intend to communicate each. The NCM then models a listener's interpretation of a noisy version of this message by using the listener's mental approximation of this probability function Pr. First I explain the basic premises of this model, and then I'll continue to a description of how it accounts for noise:

1. First, the speaker selects some intended message $I \in \mathcal{M}$ as the message they want to communicate.
$$I \sim \Pr$$

2. The speaker produces their message I to the listener, while both are within the ambient context c.

3. The listener first perceives this production as some message $m_p \in \mathcal{M}$. The perceived message m_p is not necessarily the same message as the speaker's *intended* message I, since noise may have affected the message. The message m_p can also be thought of as the message that the listener would have interpreted, if they had known for certain that no noise had affected the message in any way.

4. The listener then uses this perception m_p to "calculate" which message is most likely to have been the speaker's intended message I. They do this by evaluating candidate messages m_i by which is most likely to have been said, given their original perception m_p:
$$I' = \arg\max_{m_i \in \mathcal{M}} \Pr[m_i | m_p] \approx I$$
The listener determines their own interpretation I' as whichever message m_i is most likely to have been the speaker's intended message, given the listener's perceived message m_p, according to some inherent ability they have to estimate such likelihoods. Alternatively phrased, the listener "reconstructs" their interpretation I' of a perceived message m_p as whichever

intended message m_i would most likely result in the given perceived message m_p.

This simplified version of the NCM which we have described so far contains an issue. Given a message m_p perceived by the listener, it is always the case that $\Pr[m_p|m_p] = 1$. Since the probability of an event cannot be greater than 1, we also know that $m_p \in \arg\max_{m_i \in \mathcal{M}} \Pr[m_i|m_p]$, meaning that given m_p, the message most likely to have been intended should *always* be m_p itself. This is an issue because it would suggest that given a perceived message, that perceived message itself is always one of the best and most likely interpretations. However, linguistic communication is imperfect and always takes place within the context of noise. If a listener hears a very strange message, but that message is very similar to a much more expected or plausible message, we should expect that they attempt to somehow account for the possibility that noise corrupted the original message. In this case, we should expect that the listener would likely interpret the very similar yet more plausible alternative as what the speaker originally intended, and assume that the minor discrepancy was caused by environmental noise of some kind. This consideration is formalized within the NCM by incorporating a noise-adding function n:

$$n : c \times \mathcal{M} \to \mathcal{M}$$

This function n takes some message $m \in \mathcal{M}$ and adds noise to it based on the context of the utterance, c. This then gives us the new noisy message $n_c(m) = m'$. It's possible, even likely, but not guaranteed, that the message is unchanged. The exact effect of the context c and the distribution Pr aren't known precisely to the listener (nor to the speaker), but they do their best to maintain an accurate approximation of both. Accounting for noise, we can now re-write our earlier expressions as

$$n_c(I) = m_p$$

$$\Pr[m_p|m_i] = \Pr[n_c(m_i) = m_p]$$

$$I' = \arg\max_{m_i \in \mathcal{M}} \Pr[m_i \cap (n_c(m_i) = m_p)|m_p]$$

We now select the intended message I that both maximizes the probability it was chosen in the first place and then was corrupted into our perceived message m_p. This now aligns with our expectation that a literal interpretation might be discarded in favor of one that is both more likely overall and likely to have resulted in the perceived message after noise was added.

1. By Bayes' Theorem, this calculation is equivalent to the expression

$$I' = \arg\max_{m_i \in \mathcal{M}} \frac{\Pr[m_p | m_i \cap (n_c(m_i) = m_p)] \cdot \Pr[m_i \cap (n_c(m_i) = m_p)]}{\Pr[m_p]}.$$

2. Since our perceived message m_p is fixed, so is the denominator $\Pr[m_p]$, and we can remove it from the calculation of the maximizing argument.

$$I' = \arg\max_{m_i \in \mathcal{M}} \Pr[m_p | m_i \cap (n_c(m_i) = m_p)] \cdot \Pr[m_i \cap (n_c(m_i) = m_p)]$$

3. We are given a message m_i and told that $n_c(m_i) = m_p$ when noise was applied, so in this case the first term evaluates to at least 1.

$$m_i \cap (n_c(m_i) = m_p) \subseteq m_p$$
$$\Pr[m_p | m_i \cap (n_c(m_i) = m_p)] \geq \Pr[m_p | m_p] = 1$$

   Since the probability of an event cannot be more than 1, this must be exactly 1, as the distribution of the environmental noise and the distribution of intended messages are independent of each other. We can remove this first factor completely.

$$I' = \arg\max_{m_i \in \mathcal{M}} \Pr[m_i \cap (n_c(m_i) = m_p)] \cdot 1$$
$$I' = \arg\max_{m_i \in \mathcal{M}} \Pr[n_c(m_i) = m_p | m_i] \cdot \Pr[m_i]$$

Thus, in this model, given a fixed perceived message m_p, there are two essential things that a listener must evaluate to determine their interpretation of a noisy message:

1. $\Pr[m_i]$: The interpretation's overall prior probability, i.e. a way of scoring how "likely" any message is to be said "in general", and

2. $\Pr[n_c(m_i) = m_p | m_i]$: The likelihood that applying noise to the interpretation, $n_c(m_i)$, would yield the perceived message m_p, i.e. the "distance" between $n_c(m_i)$ and m_p.

So, a listener must be able to efficiently approximate two probability distributions to some level of accuracy each time they make these noisy inferences: the

probability of a specific message $m \in \mathcal{M}$ being intended, $\Pr[m]$, and specifically the probability of noise in the current context, n_c, acting in a certain way on a known intended message, $\Pr[n_c(m) = m'|m]$. Concretely, this thesis investigates if this is the case: Are our mental approximations of n defined in general as some function

$$n' : \mathcal{M} \to \mathcal{M},$$

or do we take the context c into account as well to approximate n's preimage?

$$n' : \mathcal{M} \to c \times \mathcal{M}$$

2.2 Expecting Noise

2.2.1 Error Identification Signal

Levy (2008) describes an *Error Identification Signal* (EIS), which is an online (bottom-up) cognitive "signal" accessible by the linguistic system. This is an adaptation of the psychological concept of saliency which is itself related to expected Bayesian surprise. The EIS at time t is higher when the distribution of likely *interpretations* shifts rapidly after a new subsection of the message is processed,

$$\text{EIS}_t = D_{KL}([\Pr[m_{[0,t)i}]|m_{[0,t)p}]||[\Pr[m_{[0,t)i}]|m_{[0,t-1)p}]),$$

where $m_{[0,t)p/i}$ is the subsection of a message that has been received (either as the perceived message $m_{[0,t)p}$ or the intended message $m_{[0,t)i}$) before time t. Levy predicts that the linguistic system allocates working memory in proportion to this signal.

The strength of this signal is directly proportional to the *expected variance* of the noise over a given time interval with respect to t. This means that contexts where noise is more predictable and self-similar, such as in music, have a lower expected EIS than another context where the noise is more varied and less predictable, even if both contexts have the same amount of noise present.

Coincidentally, this is analogous to how I measure *surprisal* later in this thesis,

which can be thought of as measuring the mean strength of this signal, over all masks of a lyric interpretation, of the predictions for the masked element. There is more on this below.

2.2.2 Context-Dependent Likelihoods

There are certain contexts in which noise becomes more predictable. Let's consider a few, which help contextualize why music might also be such a context:

Speaker Familiarity

Research shows a listener more successfully interprets what someone says as their exposure to that speaker's speech increases, without consciously choosing to do so. For example, Clarke & Garrett (2004) found that processing speed for a foreign-accented speaker rebounds to near-native levels after $\approx$ 30 to 60 seconds of ordinary listening. This capability to learn to account for idiolect-specific noise patterns demonstrates that listeners are able to attune their noise models (i.e. their internal approximations of n) to specific contexts, e.g. speaker contexts, in a bottom-up fashion.

Foreign Languages

A listener typically more successfully interprets what someone says in a certain language as their own proficiency speaking that language increases. Bradlow & Bent (2008) found that this increase doesn't even need to be more significant than a short exposure, and that sentence recognition for a non-native speaker improves across a short session. They also found that training with several talkers *generalizes* the gain in abilities to new speakers. We can model this within the NCM as the listener's ability to account for noise increasing. That is, before learning this language, the listener has no information about how they would likely perceive some intended message. However, with time they become better at predicting $\Pr[m_{p'}|m_i]$, via observing pairs of assumed speaker intentions and perceptions (bottom-up), and/or by learning explicit rules concerning how to parse perceived messages into representations of the speaker's intention (top-down). This capability to learn to account for language-specific noise patterns demonstrates that listeners are able to attune their noise models to specific contexts, e.g. language

contexts.

However, consider the listener who has not yet begun to learn anything about this language, but hears a message m'_p in that language.

$$I' = \arg\max_{m_i} \Pr[m_{p'} \sim p' | m_i, (p' \neq p)]$$

Here, $m_{p'}$ isn't sampled from the usual distribution p, but rather from a different distribution of message perceptions specific to the unknown language p'. The NCM predicts that in this case, the listener still evaluates the speaker's intended message the same way that they would evaluate a message in a language they actually speak.

$$\stackrel{?}{=} \arg\max_{m_i} \Pr[m_{p'} | m_i] \cdot \Pr[m_i].$$

We of course expect that a listener in this situation would likely recognize that they're listening to a language that they don't speak, and leave the noisy message "undefined" accordingly, or otherwise rely on external information to approximate m_i as whichever m_i gives the maximum value of $\Pr[m_i]$ alone,

$$\stackrel{?}{=} \begin{cases} \text{undefined} \\ \arg\max_{m_i} \Pr[m_i] \end{cases}$$

e.g. using prosodic or non-linguistic cues to approximate the speaker's intended message. Crucially in both cases, the listener somehow incorporates information about the *type* of noise that they are exposed to, and uses this information to modify their noise compensation strategy. Instead of evaluating which message in their own language is closest to $m_{p'}$, we expect the listener to fail to assign the speaker's intention beyond perhaps estimating the prior. This capability to account for noise differentially in unfamiliar languages, and disregard the likelihood when it's irrelevant, demonstrates that listeners are able to attune their noise models to specific contexts in a top-down fashion based on their beliefs and expectations about the noise themselves.

2.2.3 Expecting Noise in Music

Modal Effects on Sung Speech

Vowel production in singing differs significantly from vowel production in spoken speech. Sung vowels tend to be produced as more centralized compared to spoken

vowels, meaning the tongue position is closer to the center of the mouth (Hollien et al., 2000; Collister & Huron, 2008). This phenomenon, known as centralization, is especially pronounced at higher pitches, leading to perceptual confusion between vowels like /i/ (as in "see") and /ɪ/ (as in "sit") (Benolken & Swanson, 1990). Vurma et al. (2023) showed that singing causes vowels to be emphasized more than voiceless plosives compared to when speaking, and that this effect also predicts overall intelligibility. They characterized the change in emphasis as the change in a phoneme's intensity, which was quantified by duration, intensity, and Δpitch. Investigating plosives specifically, they found that "the recognition of /k/ is the most sensitive and /t/ is the least sensitive to the burst intensity, and that too intense a burst for /p/ decreases its intelligibility."

Pitch also plays a critical role in vowel intelligibility. As the fundamental frequency of sung vowels increases, they become increasingly difficult to discriminate (Smith & Scott, 1980). Particularly, when the fundamental frequency exceeds the *perceived* first formant of the vowel, intelligibility declines drastically (Condit-Schultz & Huron, 2015). This effect was first examined by Benolken & Swanson (1990), who found that

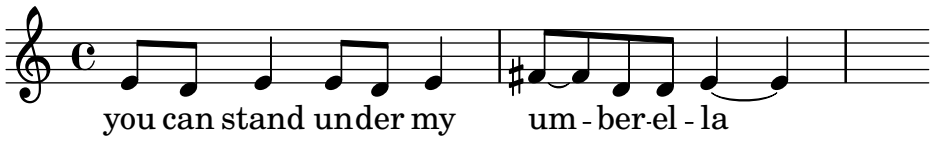
> *"American English sung vowels become increasingly difficult to discriminate as the fundamental frequency is increased [...] As the fundamental frequency increases, there is an increase in the frequency of both the first formant produced by the singer and the perceptual formant of the listeners. Since the formant frequency of the soprano rises more rapidly than the listeners' perceptual formant, a vowel is misinterpreted as one with a higher first-formant frequency than that of the intended vowel."*

This effect is especially problematic for soprano singers who often sing in this high-pitch range (Sundberg, 1975).

Consonants are also affected by singing. Collister & Huron (2008) show that voiced stops (like /b/, /d/, /g/) and nasals (like /m/, /n/) have consistently higher difficulty being recognized in sung speech compared to liquids (like /l/, /ɹ/), both types of consonants "having twice the error rate of liquids" when being perceived. They also found evidence that when sung, "listeners had much more difficulty with voiced consonants than voiceless ones, which, in the context of singing, might seem counterintuitive." This is counterintuitive since in spoken language, voiced consonants are generally considered easier to perceive than

voiceless consonants, and since *during production*, singing causes voiced elements like vowels to be emphasized more than certain voiceless elements, e.g. plosives (Vurma et al., 2023). This difficulty with voiced consonants might be attributed to the hyperarticulation of certain consonants by singers, where the mouth movements are exaggerated to ensure clarity. This hyperarticulation of certain sounds and not others leads, counterintuitively, to distortions in the perception of vowel quality.

There are also types of prosodic noise in sung speech, because the rhythmic and melodic features of music also affect intelligibility (Johnson et al., 2014). For example, *melisma*, where a single syllable is sung over multiple notes, has been shown to reduce intelligibility. This is because when the stress patterns of words clash with the musical rhythm, comprehension becomes more challenging. In sung contexts, in addition to the typical way that stress is indicated with duration and intensity, most musical traditions also have perceptual emphasis associated with notes that fall on certain beats of a rhythm. For example, in the chorus of the famed composition "Umbrella" by Rihanna & Jay-Z (2007), the word "umbrella" is sung over two on-beats, with an off-beat after each. In many Western musical traditions, a note beginning on an on-beat indicates additional emphasis for that note. The word being sung, "umbrella," is part of the larger phrase "you can stand under my umbrella," which is in *nearly* perfect iambic meter. Words in iambic meter can be faithfully produced in lyrics nicely, since each on-beat of a song is emphasized, and unstressed syllables can be placed in non-emphasized off-beats between stressed syllables. This aligns well with how in iambic meter, every other syllable is stressed.

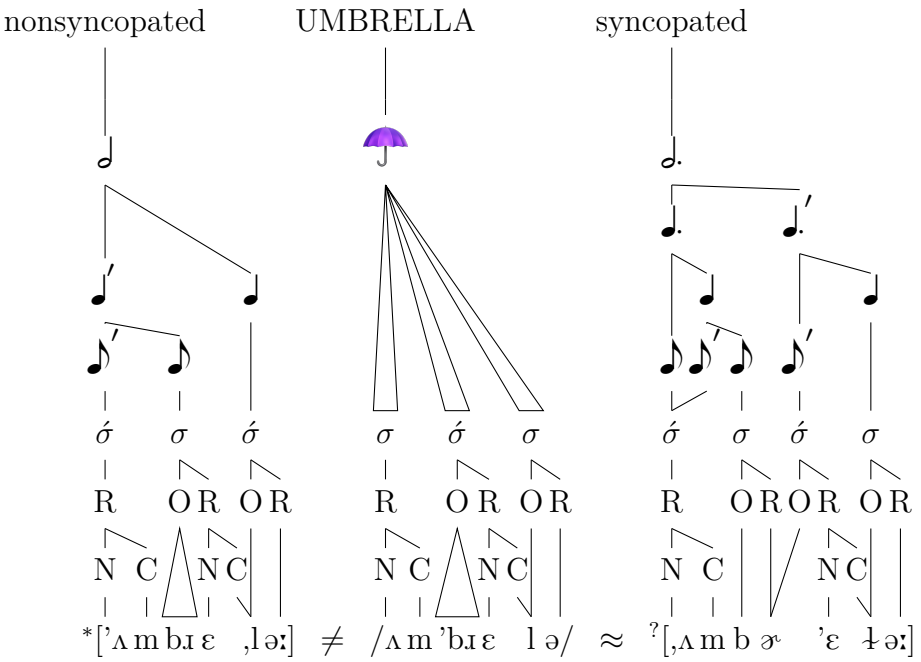


The only break in the iambic meter, in this line's spoken form, would be between the syllables "stand" and "un-," both of which would typically be stressed in iambic meter. However, this is handled gracefully by using both emphasis on the on-beats *and* note duration to indicate stress: the stressed syllables "you" and

"un-" fall on on-beats, and the stressed syllables "stand" and "my" fall on notes with twice the duration of the preceding note. However, this doesn't continue to work when we get to the word "umbrella," since if we were to continue this rhythmic pattern, the syllable we would want to stress would actually fall on a short eighth note, as well as on an off-beat.

To repair this incorrectly unstressed syllable, Rihanna introduces a *syncopated* version of the chorus' rhythmic motif,[5] which results in each syllable receiving more appropriate emphasis. Syncopation means that we're slightly displacing the emphasis from the typical on-beat, in this case to a half beat ($\frac{1}{8}$th of a measure) earlier than before. This transition in emphasis happens immediately before the word "umbrella," which results in both "my" and "um-" falling on sequential off-beats when syncopation starts. This allows "umbrella" to be repaired and segmented into four syllables, with a syllabified [bɝ] adding to the overall duration of the underlying syllable /-bɹɛ-/, and the rest of the syllable, [ɛɬ], being sung and stressed on the on-beat. By singing the three syllable word over four beats, Rihanna is simultaneously able to introduce a new rhythmic motif while using the rhythm's structure and duration to indicate the correct stress in the word "umbrella." This repair strategy also aligns well with the broader syncopation of the song, and establishes the memorable, repetitive hook of "'ella, 'ella, 'ella..." Below, we see on the left how an attempt to faithfully match this word to the rhythmic constraints without applying any sort of repair might look, as well as a construction of the alternative four-syllable production on the right. The apostrophes and note duration indicate how primary and secondary stress is derived.

[5] Alternatively, we could also view this transition as returning back to a non-syncopated version of the rhythm, and derive the same conclusions. The crucial part is that when we transition from one rhythm to the other at this point, there's a beat with longer than typical duration that can shift the stress pattern's alignment with the rhythm.



In general, the modified articulation and perception of vowels and consonants in singing makes the acoustic signal less clear. These general categories of differences are often types of noise that listeners encounter in non-sung speech as well. Even pitch-related factors have the possibility of affecting day-to-day speech. However, certain types of noise, such as melisma, typically only interfere with the speech signal in sung speech, and have consistent or predictable effects. Listeners are likely able to recruit pre-existing cognitive circuits to account for these types of noise in sung speech, but it is not clear if they are applied as-is, or if they are *fine-tuned*, as it were, to the specific ways in which these issues tend to manifest in lyrics.

Mondegreens

Mondegreens are likely affected by differences in the perception of lyrics and speech. Because mondegreens only occur in lyrics, these differences contribute to a predictable distinction between the types of noise that can lead to mistakes in each domain (Collister & Huron, 2008).

These misperceptions are also explainable by the "noisy channel" of singing

(Poliak et al., 2024), where listeners attempt to infer the intended message by combining the perceived signal with their prior expectations (Gibson et al., 2013). When listening to song, the acoustic signal is degraded by the unique articulatory properties of sung speech and the presence of instrumental accompaniment, in addition to the many sources of noise that affect both spoken and sung speech (see section **1.3.1**).

This means that when trying to understand lyrics, listeners consider both the acoustic properties of the sung words, and their prior knowledge about language, music, and the context of the song. For example, since singing affects the acoustic production of certain sounds in predictable ways, we would expect a listener to incorporate this knowledge into their lyric interpretation process.

We know that a song's genre can predict what *proportion* of its lyrics are intelligible (Condit-Schultz & Huron, 2015). However it is unclear the degree to which this is because certain genres more than others contain more noise, and the effect being caused by the *type* of noise found in certain singing styles being more typical and easier to parse than others. For example, genres that prioritize vocal clarity, such as Jazz, might employ less melisma and prioritize rhythmic alignment with word stress, leading to fewer errors related to these factors. Melisma is the process of shifting stress on a word from its original placement so that it better aligns with the rhythmic context in which the word's produced. Conversely, genres like Classical music, which often emphasize vocal technique over lyrical clarity, might exhibit a higher occurrence of errors due to melisma and complex rhythmic settings.

Impact on Lyric Production and Perception

Several factors contribute to the "noise" in song lyrics perception, such as the modified articulation of vowels and consonants in singing, the presence of instrumental accompaniment which masks or interferes with the vocal track, repetition, rhyme, and rhythmic alignment with word stress, and the use of uncommon or archaic vocabulary in song lyrics.

There is already evidence that we adapt to the specific types of noise found in music on a lexical level. For example, Squires (2019) presented participants with written stimuli, some of which contained non-standard conjugation, for example

"A man don't have to die." She found that when told that these stimuli came from music lyrics, participants read them at a more consistent speed than when no such information was provided, although they did not read these stimuli significantly faster than equivalent stimuli with standard auxiliaries. She concluded that this "information caused participants to orient to the sentences differently, which partially—but not straightforwardly—mitigated surprisal at nonstandard *don't*." This would suggest that although we are not particularly better or worse (i.e. faster or slower) at processing the semantic content of music lyrics than we are at processing the semantic content of speech in general, we may have specific cognitive mechanisms that, when primed to do so, process the specific types of noise in music lyrics in a more consistent manner than an otherwise underspecified approach to estimating which parts of the signal are meaningful and which are not.

2.2.4 This Thesis

In this thesis I investigate the following hypotheses:

I Participants consistently identify lyrics more accurately during forced-choice tasks than during free response tasks.

II Listeners' linguistic processing of contextually-modulated speech is better explained by specialized mechanisms, including specialized phonological, lexical, and pragmatic mechanisms, than generalized mechanisms. Specifically, listeners use contextual information to anticipate specific types of linguistic noise, and as a result, the *Noisy Channel Model* makes more accurate predictions when parameterized by context-relevant metrics than when parameterized by generally applicable metrics. Specifically, I hypothesize that either

   i operationalizing the likelihood-distance between two strings of phonemes by accounting for the specific features from which they're composed, and/or

   ii operationalizing the prior probability of an utterance by examining the internal state of a fine-tuned large language model upon processing that utterance

   will more successfully parameterize the Noisy Channel Model by means of accounting for context-sensitive noise.

These hypotheses emerge from the literature reviewed in this chapter and address gaps in our understanding of noise processing in linguistic communication. Hypothesis I is motivated by methodological concerns regarding how lyric interpretation has been studied. The forced-choice paradigm used by Poliak et al. (2024) provides participants with predefined options, potentially constraining their responses and inflating accuracy rates compared to natural listening conditions. In real-world situations, listeners must identify lyrics without predefined options, making free response tasks more ecologically valid. Moreover, the Noisy Channel Model suggests that providing options could artificially boost the influence of prior probabilities by narrowing the interpretation space. This hypothesis allows us to examine whether the forced-choice methodology provides an accurate representation of how listeners naturally process and interpret lyrics.

Hypothesis II extends from the large body of evidence on context-specific adaptations in speech perception. The unique phonetic properties of sung speech, such as vowel centralization (Hollien et al., 2000; Collister & Huron, 2008), modifications in consonant articulation (Vurma et al., 2023), and melodic constraints (Johnson et al., 2014), create both predictable and unpredictable patterns of noise that differ from spoken speech. Prior research by Cherry (1953), Clark (2013), and others demonstrates that listeners adapt their noise-filtering strategies based on contextual information. Squires (2019) showed that people process non-standard language differently when they know it comes from lyrics, already hinting at specialized processing mechanisms for this context. This hypothesis examines whether listeners develop specialized processing mechanisms for different contexts with specific, predictable noise patterns, rather than applying general noise processing strategies uniformly across all contexts.

I aim to resolve these hypotheses in order to advance our understanding of how the human language processing system flexibly adapts to varying communicative environments by integrating context-specific knowledge with perceptual input.

Methods

3.1 Participants

SEVENTY participants were recruited through Prolific, an online research portal, who identified as English-speaking monolinguals from the U.S.A and had access to a computer. These 70 participants were sampled from the pool of over 48,000 possible participants on Prolific's platform that satisfy these criteria. This constraint was placed on the pool of all possible participants to ensure that my results would be comparable with those of Poliak et al. (2024), who used the same constraint to filter their participants in order to ensure to the greatest extent possible that differences among participants' linguistic systems did not confound or obscure results. Although this should be seen as a methodological weakness, as most people have some level of experience with more than one language, I choose to prioritize consistent methodological reproduction over controlling for the effect of monolingualism on song lyric comprehension.

I was not able to filter the 50 participants that participated in Poliak et al. (2024)'s study from my own, but it is highly unlikely that any of my participants were also participants in their study. To estimate a lower bound, if we assume random sampling for both studies from a pool of 48,000 possible participants, we would expect that approximately 0.07 people participated in both studies. The participants were also required to use headphones with their computer during their participation in both Poliak et al. (2024)'s study and in mine.

3.2 Poliak, Kimura, and Gibson (2024)

Many aspects of this thesis' methodology, particularly the metrics that are not adapted to specific noise contexts, are adapted from Poliak et al. (2024).

3.2.1 Materials

Participants were presented with 37 short audio clips. Each clip was from a unique song in English, lasted for ten seconds or less, began with at least a $\frac{1}{2}$ second of instrumentals before the lyrics, was normalized to have the same average volume across stimuli, and was modified to have a fade-in of 0.3 seconds "to avoid jarring participants with unexpected music." Poliak et al. (2024), who created these stimuli, state that "We chose songs from a wide range of genres and avoided famous songs (all the songs that we selected had less than 500,000 plays on Spotify at the time of selection)." While I agree that subjectively, the stimuli reflect a relatively wide gamut of genres, many of them come from songs that are significantly more popular than claimed (e.g. one stimulus is from *Don't Stop Believing* by Journey, which, as of March 2025, has over 2.3 million streams on Spotify). However, this discrepancy is not concerning for our purposes, since I note later that we don't observe any significant difference in how participants respond to songs with which they are familiar versus songs with which they are not. This discrepancy also does not concern our interpretation of the results of Poliak et al. (2024), as they also failed to find a significant effect of familiarity on lyric interpretation accuracy.

3.2.2 Forced Choice Procedure

In both Poliak et al. (2024)'s study and in mine, participants were presented with each of the 37 stimuli in a randomized order. In my study, they were first presented with the free response procedure for those stimuli, described below. After participants heard each audio excerpt, they were presented with four possible transcriptions of the lyric they had just heard, and clicked which possible transcription they thought was correct before proceeding. The exact same four possible options were given for each stimulus to each participant across Poliak et al. (2024)'s study and mine. Further, the order in which a single stimulus' 4 possible options were presented was randomized across every participant and every stimulus. Exactly one of the options was always the "correct" transcription (i.e. matching the lyric published by that song's author's record label), while the other three were mondegreens. As an attention check, 5 of the 37 stimuli were paired with 3 highly dissimilar mondegreens, while the other 32 stimuli's mondegreens were created by Poliak et al. (2024) for the purpose of this research, and "differed only a little from the true lyrics while still being grammatical".

In both studies, the participants were then asked to indicate whether they had heard the song before proceeding to the next stimulus. Controlling for this,

I later use the Noisy Channel model to predict participants' lyric interpretations based on both their prior probability and acoustic similarity to the correct lyrics, comparing the predictive power of general versus context-specific metrics across both forced-choice and free response tasks.

3.2.3 Likelihood Metric

To measure the likelihood of a lyric's interpretation, i.e. the probability of hearing one lyric given the other is what was actually sung, Poliak et al. (2024) used the *Levenshtein distance* (Levenshtein, 1965) between the two lyrics' phonological representations. The Levenshtein distance measures the *minimal* number of insertions, deletions, or substitutions that are necessary to convert one string into another. For example, each of the following pairs of strings are a Levenshtein distance of 1 from one another:

1. Substitution: /kæt/ → /kɪt/ ⟨cat⟩ → ⟨kit⟩
2. Deletion: /taɪm/ → /taɪ/ ⟨time⟩ → ⟨tie⟩
3. Insertion: /meɪk/ → /meɪks/ ⟨make⟩ → ⟨makes⟩

This simple metric provides a quantitative measure of phonological similarity between two strings, with smaller distances indicating greater acoustic similarity and therefore higher likelihood of one being perceived as the other.

3.2.4 Prior Metric

To measure the prior probability of a lyric interpretation, Poliak et al. (2024) used a version of Google's bert-base-uncased large language model (Devlin et al., 2019). Variants of this model architecture have gained popularity in various disciplines, as it can be easily modified for specific tasks "without substantial task-specific architecture modifications." When it was released, this model "obtain[ed] new state-of-the-art results on eleven natural language processing tasks" (Kauf & Ivanova, 2023).

Poliak et al. (2024) *fine-tuned* this model on the genius.com lyric database (Kreiner, 2023), and used the fine-tuned model to calculate the *surprisal* of each of their responses (Misra, 2022). I say more about fine-tuning below. Conceptually, the surprisal measures how unpredictable each part of the response was, given its

context. Higher surprisal values indicate that the language model found the text that it was fed to be more unexpected, less predictable, or *higher entropy* compared the text it had seen before. In this thesis I investigate what happens when we use fine-tuning to steer the model to base its judgements of what's "expected" either exclusively on general patterns, predominantly on patterns in song lyrics, or somewhere in between.

3.3 Free Response Procedure

The forced choice procedure, discussed above, always presented participants with exactly one correct answer. This may not have been an accurate proxy for the quality of answers and quantity of correct answers that participants may have arrived at otherwise. In addition to this task, my participants also completed a free response task for each stimulus. The free response task was before the forced choice task for every stimulus. After listening to the stimulus once, participants were asked to type the words they'd heard in the song. They were then offered the chance to listen to the same stimulus twice more, and were asked to write what they'd heard after each. No stimulus was heard more than three times by the same participant. I converted each text string response to this and the forced-choice task to corresponding phonological representations using the CLTS transcription system (Anderson et al., 2018), and converted these phonological representations into phonemic representations as described below.

3.4 Contextual Likelihood Metrics

In addition to Levenshtein distance, discussed above, this thesis investigates and compares the predictive powers of other likelihood metrics. Using Levenshtein distance to quantify how "far" one speech signal is from another requires us to assume a few things:

1. First, we must assume that the basic units of sound that we use to communicate when speaking are the same as the basic units of sound that we use to communicate linguistic intent when singing. This assumption contradicts, for example, the consistent cross-linguistic pattern of certain types of consonants being exaggerated or certain vowels merging when singing. As pitch increases, especially when F_0 is perceived as passing F_1,[6] the spectral

[6] These refer to the fundamental frequency bands that are acoustically exhibited and relate

properties of vowels converge, making them acoustically similar and harder to distinguish. This convergence results in a perceptual merging of vowels, effectively reducing the number of distinct vowel sounds at higher pitches (Sundberg, 1975; Smith & Scott, 1980; Benolken & Swanson, 1990; Hollien et al., 2000; Deme, 2015, 2017).

2. Secondly, even if we do in fact use the same phonetic inventory when producing and perceiving lyrics and when producing and perceiving speech, we must still assume that the level of perceptual similarity between two distinct phonemes when we speak them is approximately the same as the perceptual similarity of those two phonemes when we sing them.

To interrogate these assumptions about the accuracy and applicability of phonetic Levenshtein distance to sung speech signals, I introduce an alternative metric we can use to measure these raw distances in addition to Levenshtein distance:

1. Contextually-Weighted Phoneme Distance
   Instead of simply counting the number of insertions, deletions, or substitutions necessary to convert one string of phonemes into another, we consider the specific varieties of phoneme insertions, deletions, or substitutions that are prevalent in lyric perception. For example, since vowels are more physically emphasized in production (in terms of duration, intensity, and Δ pitch) and more perceptually salient (in terms of identification task accuracy) than plosives in sung speech compared to in spoken speech (Vurma et al., 2023), deleting a plosive is weighted less than deleting a vowel, and substituting a plosive for a vowel is weighted less than substituting a vowel for a plosive.

2. FeatureContextDistance
   At a more granular level than phonemes, I introduce a method of taking a representation of an utterance as a sequence of ternary vectors of phonetic features, and measuring the edit distance between two utterances' feature-wise representations. I used the SoundVectors Python library (Rubehn, 2024) to convert the phonetic representations of my data to sequences of vectorized features. I then used the FeatureContextDistance algorithm to measure the distances between each correct lyric and each of its associated responses, weighted for the context described above. This algorithm works by considering two sequences of feature vectors, and returning the distance between them. I'll give a description of its naïve and inefficient,

to vowel categorization.

yet easier to discuss and equivalent, implementation:

To start off with, we are given two sequences of phonemes, and want to know their features' distance. Say that one string has more phonemes than the other, and that the lengths of these sequences (i.e. the numbers of phonemes in each) are $n \geq k$. We initialize the distance at $n - k$. For "each" of the $\binom{n}{k}$ choices of indices at which a new phoneme can be inserted that results in the two resulting strings being of the same length, we then calculate the Levenshtein distance between two ternary row vectors length n that correspond to a single *feature*, and sum this distance over all 39 features that represent each phoneme. We assign the cost of swapping a -1 with a 1 as double that of swapping a 0 with a 1 or a 0 with a -1 to capture the ternary nature of our feature string alphabet. After doing this for every possible choice of phoneme insertion indices (and glossing over the heuristics and dynamic optimizations which make this factorially large search space resolvable in a matter of days instead of decades), we return the lowest value (sum of feature-wise Levenshtein distances + the innate length difference initialized between the two strings) as their distance. Concretely, this distance corresponds to an optimal choice of where to insert empty "phonemes" into a shorter string to match the length of the longer string such that the average distance between each feature's string-wise vector is minimized. The magnitude of this optimum is equivalent to the distance given by the algorithm, e.g. between the correct lyric and a participant's response.

3.5 Contextual Prior Metrics

To evaluate how predictive power is affected by contextualizing our predictors, I consider three unique prior metrics. In addition to Poliak et al.'s fine-tuned large language model (2024), I used the generalized, non-fine-tuned base version of this model. I also trained a second fine-tuned model myself. Compared to Poliak et al.'s model, this third model was twice as accurate at guessing the missing elements of lyrics from the genius.com lyric dataset (Figure 3.1). This allows us to construct a spectrum from generalization to contextualization: at one extreme is the base model that is not specialized for lyrics whatsoever, and at the other is my model that's highly adept at predicting how probable a given lyric that it hasn't seen before is.

Fine-tuning could be thought of as analogous to specialized language acquisition (Kuperberg & and, 2016), although they are not the same thing. Imagine a student who has gained general proficiency in English through extensive reading, consuming a diverse range of texts, like novels, newspapers, and academic articles, as well as Twitter feeds, neolithic cave markings, and Pokémon cards. This is like to the "base" language model, which has been trained on a broad corpus of text. Now imagine that that student spends several months exclusively studying, say, poetry, or legal documents, or song lyrics. Over time, they would develop specialized knowledge of the vocabulary, structures, and conventions specific to that genre. Fine-tuning works similarly: we take a pre-trained language model with general linguistic knowledge and train it further on a specialized corpus (in this case, genius.com). This additional training helps the model recognize and internalize patterns specific to lyrics, like their unique vocabulary, syntax, and semantic tendencies. It might also learn something about phonological patterns in music, such as which words might rhyme with others if they are often seen in tandem at the end of complementary phrases. The degree of specialization can be controlled by adjusting how much additional training on lyrics we provide, allowing us to create models with varying levels of lyric-specific knowledge while maintaining their general linguistic capabilities. Although the base model already makes significantly accurate predictions (Poliak et al., 2024), this spectrum of specialization lets us investigate how different degrees of contextual knowledge affect the model's ability to predict lyrics' interpretations.

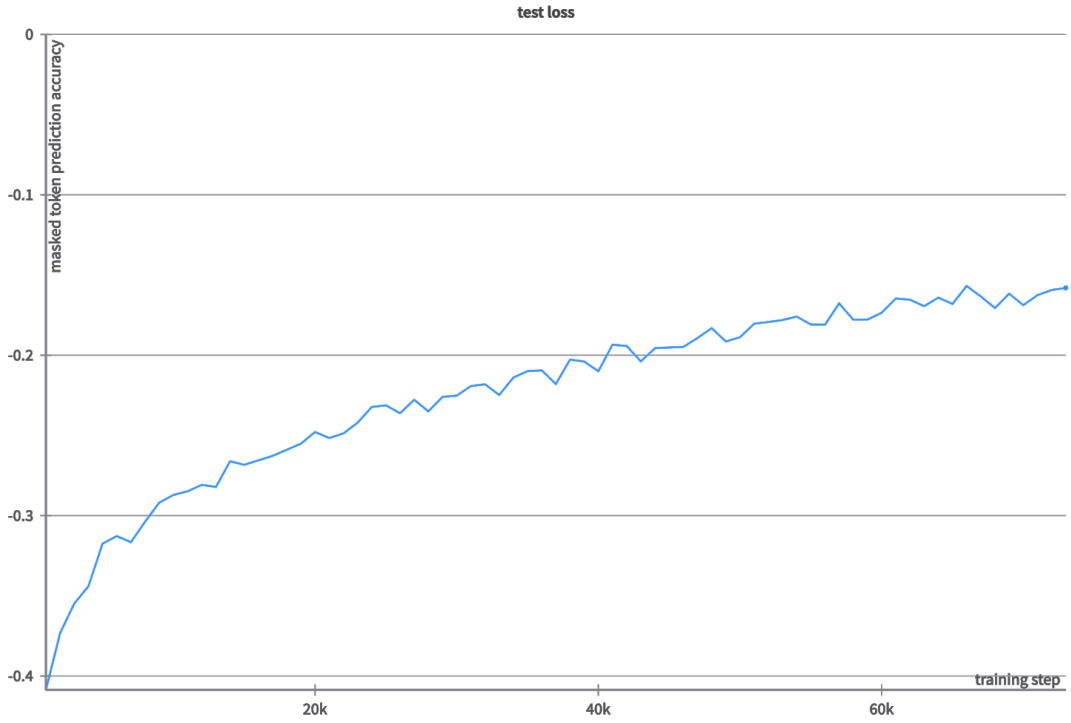


Figure 3.1: This plot shows the how the language model's accuracy at predicting the content of song lyrics increases as we continue to train it specifically on the genius.com database. The model used for the most highly contextually-tailored metric is represented on the right of the plot. The graph never reaches 0, which would indicate perfect accuracy at identifying the masked tokens in unseen song lyrics, but grows significantly more accurate as it's trained. The general model's accuracy is seen at the beginning of training on the left of the graph, and the accuracy of the intermediate model trained by Poliak et al. (2024) is equivalent to -0.29 along the y-axis on this scale.

Results

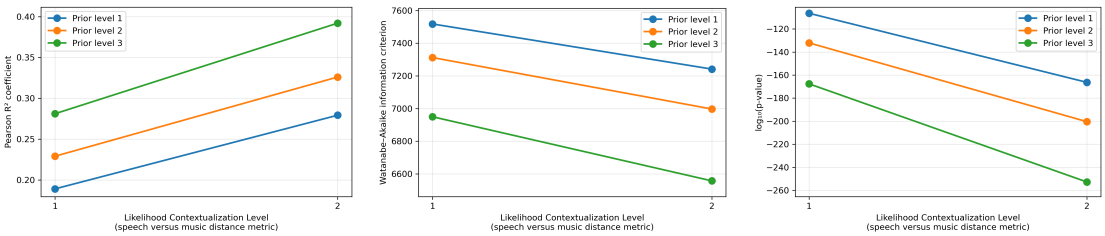


Figure 4.2: The orange points on the left axes of each plot represent the fitness of the metrics used by Poliak et al. (2024). In each plot, the points on the right represent the fitness of the new, contextualized likelihood metric, and the other colors represent the fitnesses of the new metrics of prior probability (with blue being the least contextualized and green the most). Making either predictor of language interpretability (prior or likelihood) more contextually-relevant yields significant improvements in the NCM's general predictive power without sacrificing model parsimony.

4.1 Attention Check

EVERY single participant chose the correct forced-choice option for each of the 5 catch items. This perfect performance on the attention check items indicates that all participants were actively engaged with the task and attending to the stimuli. This should make us more confident that errors observed in the experimental trials reflected genuine cognitive processes rather than inattention or task disengagement, and that the addition of a free-response task in addition to the forced choice task did not add significant fatigue effects.

4.2 Inferences

This section presents the key findings from my analysis of lyric interpretation. I first replicate and validate the results from Poliak et al. (2024), then extend the analysis to compare performance between forced-choice and free response tasks, evaluate the efficacy of different metrics for modeling lyric interpretation, and examine the effect of song familiarity on performance.

4.2.1 Forced Choice vs. Free Response

In addition to the forced-choice task used by Poliak et al. (2024), discussed below, my study incorporated a free response task where participants typed what they heard without being given predefined options. This allowed me to examine whether the constraints of the forced-choice paradigm affected participants' lyric interpretations.

My data reveal significant differences between the two response formats. Participants were consistently more accurate in the forced-choice task (mean accuracy ≈45%) compared to the free response task (mean accuracy ≈11 %), $p \leq 0.001$. This substantial difference suggests that having options to choose from fundamentally alters the perception and/or decision process involved in lyric interpretation.

Further, qualitative analysis of the free responses revealed greater variability in interpretations than what was captured in the forced-choice options. While forced-choice responses naturally clustered at the provided options, free responses showed more diverse phonological variations, with participants often generating interpretations that combined elements from multiple forced-choice options or creating entirely novel interpretations not represented in the forced-choice set.

These findings suggest that the forced-choice paradigm, while methodologically convenient, may artificially constrain our understanding of lyric perception by limiting the range of possible interpretations. The free response data provide a more ecologically valid picture of how listeners naturally interpret sung speech in the absence of explicit alternatives.

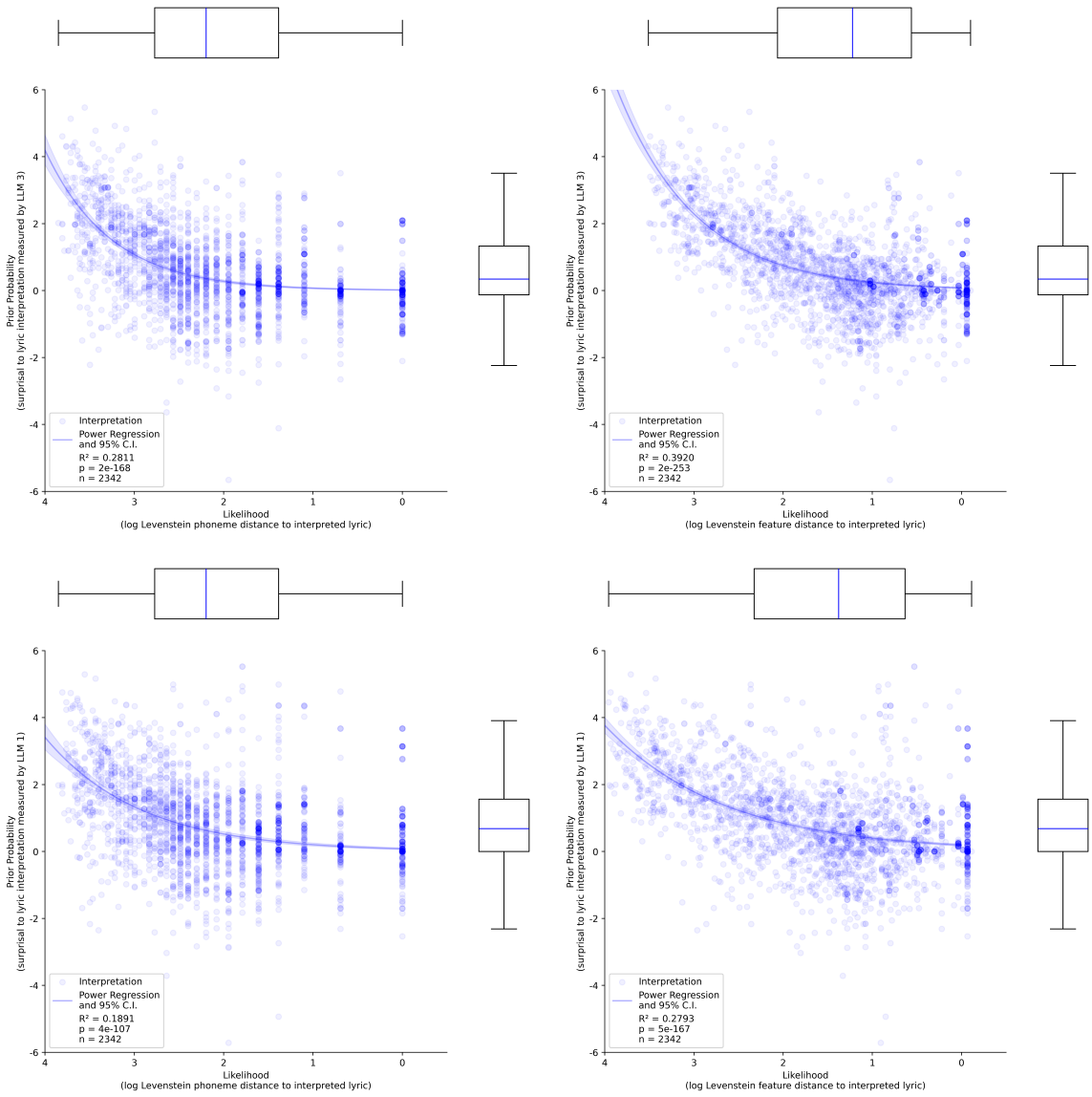


Figure 4.3: Scatterplots of our experimental data operationalized by general (left) and contextual (right) likelihood metrics and by general (bottom) and contextual (top) metrics of prior probability. Since either metric is more accurate when contextualized, the NCM better explains communication errors when we assume the context of these errors is integrated into our repair strategies.

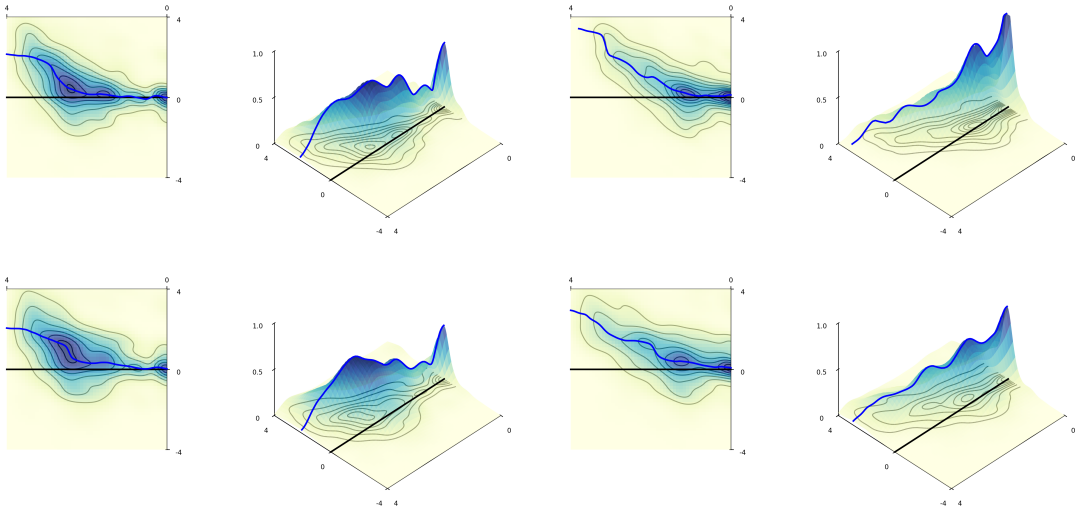


Figure 4.4: Surprisals (y-axes) and distances (x-axes), from Figure 4.3, directly represented as probability distributions rather than scatterplots over the space of possible interpretations. Each contour line represents a 10% increase in relative probability. Across all models, we observe that these two axes do not represent independent distributions, but rather that there is a strong relationship between our two variables throughout. Specifically, we notice that the interpretations that are selected appear to be sampled from a distribution with strong negative covariance between likelihood and probability. Further, increasingly context-specific models make more robust, accurate, and theoretically consistent predictions about how a noisy utterance is interpreted. As we move from the generally applicable metrics at the bottom left, to the context specific metrics in the top right, the contour lines grow tighter *and* in number, our probability distributions form sharper edges along their ridges, and they demonstrate less variance. The top row uses the most accurately tuned version of our surprisal model, and the rightmost column uses the most fine-grained contextual feature distance metric.

4.2.2 General vs. Contextual Metrics

The main contribution of this thesis is justification for the development and evaluation of context-specific metrics for both prior probability and likelihood in specific linguistic contexts. Figures 4.4 and 4.3 present a comprehensive comparison of how these different metrics performed in modeling participants' responses in the context of lyric interpretation.

Figure 4.3 shows scatterplots of our experimental data using different combinations of metrics. The x-axis represents the likelihood metrics (standard Levenshtein distance to the left of our contextually weighted feature distance), while the y-axis represents prior probability metrics (comparing the base LLM with two versions of fine-tuned models). Points represent individual response options, with clustered interpretations shown additively in darker colors.

As we move from the general metrics (bottom left panels) to more context-specific metrics (top right panels), we observe a clearer separation between selected and non-selected options, indicating that the context-specific metrics provide better discrimination between plausible and implausible interpretations.

Figure 4.4 transforms these data into probability distributions, with contour lines representing 10% increases in relative probability. The visualization reveals a few important patterns:

1. Across all model combinations, there is a strong negative correlation between surprisal (y-axis) and distance (x-axis), indicating that these two dimensions are not independent but interact in systematic ways. This aligns with the NCM's prediction that listeners integrate prior probabilities and likelihoods when interpreting noisy input.

2. As the metrics become more context-specific (moving from bottom left to top right panels), the probability distributions become more sharply defined, with tighter and more numerous contour lines. This suggests that context-specific metrics capture more precise patterns in participants' responses.

3. The most context-specific models (top right panel) show the clearest separation between regions of high and low probability, indicating that these metrics most effectively discriminate between selected and non-selected interpretations.

Quantitatively, the context-specific metrics outperformed the general metrics in predicting participants' responses. The model using the most context-specific prior metric (the most fine-tuned LLM) and the most context-specific likelihood metric (FeatureContextDistance) achieved an R^2 of 0.392, compared to 0.189 for the model using general metrics. This represents a 20% improvement in explanatory power, demonstrating the substantial benefit of incorporating contextual in-

formation into our predictive models.

In plain English: When we listen to singing, our brains don't just use all-purpose language tools—they break out the specialized singing equipment! Models that knew about the singing context were twice as good at predicting how people would respond. It's like bringing a specialized toolkit rather than a Swiss Army knife to a concert.

These findings strongly support my second hypothesis, that listeners' linguistic processing of contextually-modulated speech is better explained by specialized mechanisms than generalized mechanisms. The superior performance of context-specific metrics suggests that listeners actively adapt their perceptual and interpretive processes to the specific constraints and characteristics of sung speech.

4.3 Effect of Song Familiarity

I also examined whether familiarity with a song influenced participants' ability to accurately interpret its lyrics. During the experiment, participants indicated whether they had heard each song before, allowing us to compare performance on familiar versus unfamiliar songs.

Contrary to what might be intuitively expected, song familiarity did not significantly impact lyric interpretation accuracy whatsoever in either the forced-choice task or the free response task. Participants showed similar performance patterns regardless of whether they reported prior exposure to the song.

This finding is consistent with Poliak et al. (2024), who also found no significant effect of familiarity. The lack of a familiarity effect suggests that lyric interpretation may rely more on general linguistic and perceptual processes than on memory for specific songs. It also indicates that the Noisy Channel Model's effectiveness in predicting interpretation patterns is robust across both novel and familiar musical stimuli.

However, it is worth noting that my measure of familiarity was based on a yes/no self-report and did not assess the depth or recency of exposure. Future research could benefit from more nuanced measures of familiarity, potentially revealing effects that were not captured in my and Poliak et al. (2024)'s studies.

4.4 Prior Results

I successfully replicated the main findings of Poliak et al. (2024), as shown in Figure 4.5. The left panels present my reproduction of their results, while the right panels show their original findings. In both studies, there is a positive correlation between the frequency with which participants selected a particular lyric interpretation and both 1) that interpretation's prior probability and 2) its likelihood given the actual lyric.

In accordance with the NCM, participants' selections were influenced by both the inherent plausibility of a lyric option (operationalized as the prior probability assigned by LLM surprisal) and the acoustic similarity between the option and the correct lyric (operationalized by the phonetic Levenshtein distance). This supports the theoretical prediction that lyric interpretation involves a rational integration of both prior expectations (e.g. expectations of semantic content) and perceptual evidence (e.g. phonemic percepts).

The replication of these patterns provides a solid foundation for the subsequent analyses in this thesis, as it confirms that our experimental paradigm is capturing the same cognitive processes observed in the original study. The consistency of these findings across different participant groups also strengthens the case for the generalizability of the NCM in explaining lyric perception.

4.5 Model Parsimony

When evaluating predictive models, it is important to consider not only their accuracy but also their parsimony, or how efficiently they explain the data relative to their complexity. Figure 4.2 presents three different measures of model fitness across our different metric combinations: R^2 (left), AIC (Watanabe-Akaike Information Criterion, middle), and p-values (right).

The orange points on the left side of each plot represent the model using the metrics from Poliak et al. (2024), while the colored points on the right represent our new context-specific metrics. Several important patterns emerge from this analysis:

First, all three measures consistently show that the context-specific metrics outperform the general metrics. The R^2 values increase substantially (indicating

better fit), the WAIC values decrease (indicating better model performance while accounting for model complexity), and the p-values become more significant (indicating stronger statistical evidence for the model's validity). All three models have significant p-values, which indicates that the general architecture of these models is valid regardless, and is related to the absolute quantity of datapoints that I was able to collect (2342). Interestingly, each of the changes independently had roughly the same effect size on model fit: changing either metric, but not the other, accounted for an additional 10% of the variation seen in the data that was unexplained by a fully general model.

Second, the improvement in model performance is achieved without an increase in model complexity. Both the original and context-specific models use the same number of parameters (two main predictors: prior probability[7] and likelihood), yet the context-specific metrics provide significantly better fit. This means that the improvement comes from better measurement of the relevant constructs rather than from adding complexity to the model structure. The WAIC analysis is particularly informative because it penalizes models for complexity, helping to avoid overfitting. The consistent decrease in WAIC values for the context-specific metrics indicates that their improved performance represents a genuine enhancement in explanatory power rather than just fitting noise in the data.

Third, both components of the model, prior probability and likelihood, benefit from contextualization. The different colored points represent various combinations of prior metrics with the new likelihood metric. The consistent upward trend across all colors indicates that the context-specific likelihood metric improves model performance regardless of which prior metric is used. Similarly, within each likelihood metric, the progression from blue to orange to green shows that increasingly context-specific prior metrics also enhance performance.

These findings reinforce my conclusion that context-specific metrics more accurately capture the cognitive processes involved in speech interpretation in a particular context. The parsimony of these improvements suggests that listeners' perceptual systems are indeed attuned to the specific phonological and statistical patterns that characterize sung speech, allowing them to more efficiently process this modality compared to using generally-applicable linguistic mechanisms.

[7] Technically, this predictor is the result of a large language model that itself has hundreds of millions of parameters, but crucially, this (large) number is constant across all three versions of the LLM.

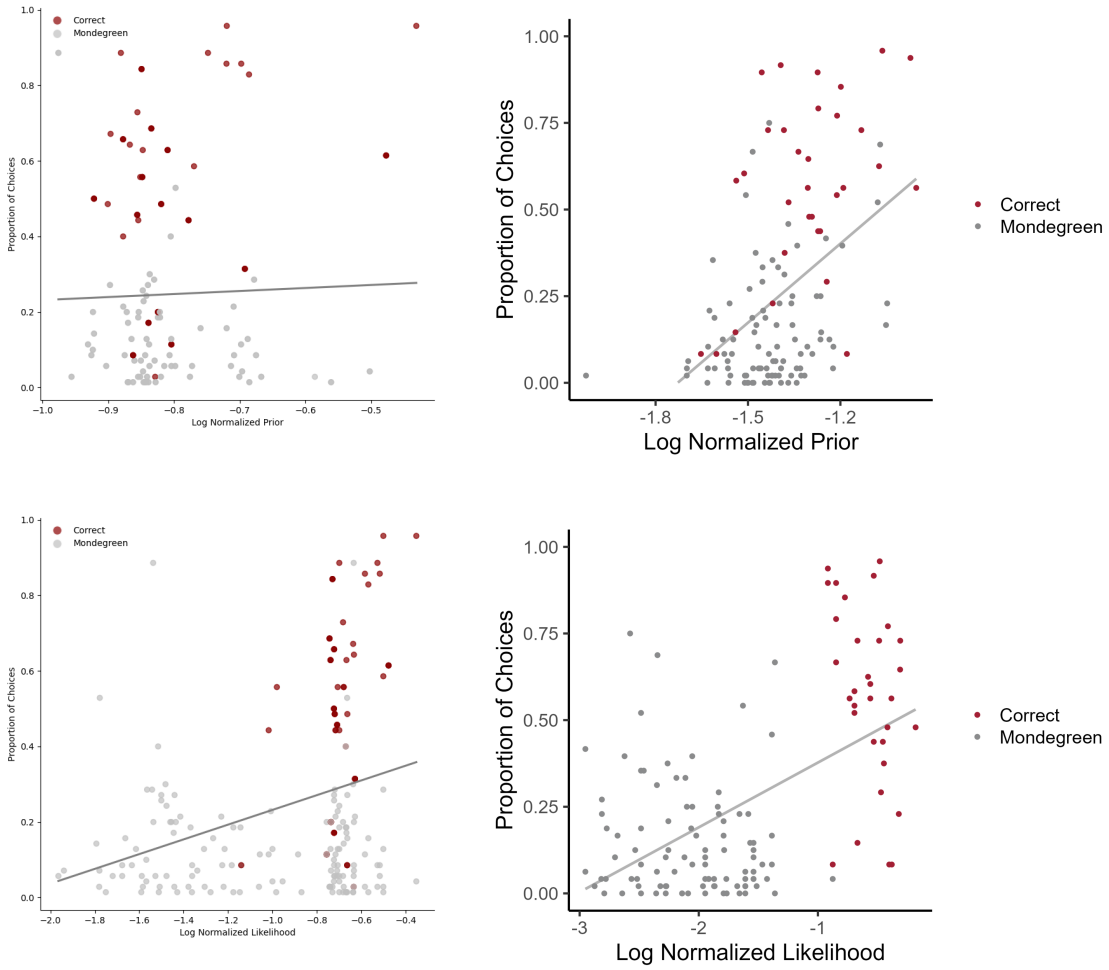


Figure 4.5: Reproduction (left) of the main result of Poliak et al. (2024) (right). When normalized to our satisfaction, a choice's increase in either prior probability or in likelihood positively correlates with that choice being selected more often by participants in a forced-choice task. It's unclear whether the difference in the slope is meaningful in the first graph. Poliak et al.'s 2024 official description of their normalization methods would have led to all of the red points lying vertically, with the same value along the x-axis, on their graphs. Since this is not the case, I instead attempted to adapt the normalization procedure provided in their codebase, which differs in order of operations from that reported in their paper, but does not place the red points along the same vertical. This ad-hoc adaptation of their normalization procedure difference may have contributed to this difference in slope. All other data reported in this thesis (i.e. everything except the forced-choice data above) were normalized by performing log normalization before value normalization, rather than normalizing first before applying a logarithmic transformation, which is able to conserve crucial interdependencies in the dataset. When I processed the data above in this standard way, I found that the relationship between proportion and likelihood (bottom) remained a significant (but weaker) positive correlation, while the relationship between proportion and prior (top) became an insignificant negative correlation.

Discussion

In this section, I discuss how my findings on lyric interpretation inform and extend probabilistic models of language processing. The results presented in Chapter 3 demonstrate that incorporating contextual information into both likelihood and prior metrics significantly improves our ability to predict listeners' interpretations of sung speech. These findings could have implications for language processing across different communicative contexts and modalities.

5.1 The Noisy Channel Processing Framework

My results provide strong empirical support for the NCM (Levy, 2008) as a framework for understanding speech perception across different modalities. The successful replication of Poliak et al. (2024)'s findings confirms that the NCM applies to sung speech, while my extensions demonstrate that the model's predictive power can be enhanced through contextual parameterization informed by concepts of formal linguistic theories, such as phonetic feature saliency. This suggests that the NCM captures fundamental cognitive processes that operate across diverse communicative contexts, rather than being limited to a specific domain processing.

5.1.1 The Intended Message

The concept of the "intended message" in the NCM becomes particularly interesting in the context of sung speech, where the artist's intended lyrics may be deliberately obscured by artistic choices and unintentional intermediate effects. My findings suggest that listeners attempt to infer intentions even when these are difficult to discern, and that they do so by integrating prior knowledge with perceptual evidence in a statistically optimal manner.

The significantly higher accuracy in forced-choice tasks compared to free response tasks indicates that when alternative interpretations are available, listeners are better able to approximate the intended message. This accounts for how the space of possible intended messages is able to be efficiently constrained by external factors, even when the signal strength is itself unchanged. This should help us better understand why it seems, perceptually, that meaning suddenly springs forth in the mind fully formed, rather than always having the sensation of deriving meaning be the result of a convoluted and sequential application of whichever rules may be applicable.

This finding could have implications for theories of communication more broadly. It suggests that the inferential process described by the NCM does not operate in isolation but can be influenced by contextual factors that shape the hypothesis space. In natural communication, these factors might include shared knowledge, discourse context, or gestural cues that help constrain the range of possible interpretations. The difference between forced-choice and free response accuracy suggests that constraints on the hypothesis space might help successful communication in noisy environments.

5.1.2 The Perceived Message

My analysis of free response data reveals that the perceived message is far more variable than might be suggested by forced-choice paradigms alone. This variability is not random but structured, appearing to follow principles predicted by the NCM, where acoustically similar and contextually probable interpretations are more commonly perceived.

The feature-based distance metric I developed captures this structured variability more effectively than traditional Levenshtein distance, suggesting that listeners' perceptions are sensitive to fine-grained phonetic features rather than only treating phonemes as discrete, atomic units when "measuring" the distance between them. This aligns with research on spoken speech perception but had not previously been demonstrated for sung speech.

This finding reinforces theories of speech perception that emphasize the role of phonetic features in shaping perception (Hallé et al., 2004; Liberman et al., 1957). It suggests that when processing sung speech, listeners are particularly at-

tuned to certain feature distinctions that are preserved in singing (such as manner of articulation) while being especially tolerant of variations in features that are typically distorted by singing (such as vowel height). This *selective* sensitivity to phonetic features is best explained as an adaptation to the specific constraints of sung speech, allowing listeners to extract meaningful linguistic information despite the distortions introduced by musical context.

5.1.3 Contextualizing Modalities

To reiterate, the starkly improved performance of context-specific metrics evidences how listeners adapt their processing strategies to specific modalities, such as sung speech. This is likely not a binary switch between processing modes but rather a continuous adjustment of processing strategies based on contextual cues: listeners may be employing specialized adaptations for specific linguistic contexts. This indicates that the NCM's parameters, i.e. the factors that determine successful communication, should not be treated as fixed across all communication modalities but should be calibrated to the specific constraints and characteristics of each.

This perspective aligns with dynamic theories of language processing that view perception as an active, adaptive process rather than a passive decoding operation (Clark, 2013). My findings suggest that listeners maintain multiple versions of the same predictive mechanism for the noisy channel, which may be activated by contextual cues. In the case of sung speech, these parameters appear to include something that resembles specialized phonetic distance metrics that account for the unique distortions introduced by singing, as well as adjusted expectations about the distinctive probabilities over various lexical, syntactic, and semantic structures that reflect the distinctive patterns in song.

5.2 Distributions on Perception

The probability distributions visualized in Figure 4.4 offer insight into how prior probabilities and acoustic likelihoods interact to shape perception. The clear *non-linear* negative correlation between surprisal and phonetic distance supports a *multiplicative*, rather than additive, integration of these factors, as predicted by the NCM.

This interaction effect is particularly notable because it suggests that listeners do not simply weigh prior probability and acoustic evidence independently but rather integrate them in a way that reflects their statistical dependencies. When a lyric is highly probable based on context but acoustically distant from the perceived signal, or acoustically similar but contextually improbable, listeners still may select it as their interpretation! However, interpretations that "optimize" both factors, being both contextually probable and acoustically similar, are dramatically more likely to be selected. To underscore this justification for the fundamental premises of the NCM, interpretations from my dataset are so highly clustered near the correct response that they consistently occupy the densest (darkest-colored) neighborhood of the negative log-transformed graphs in Figure 4.4.

5.2.1 Expectability as EIS Variance

A key idea is that what might be considered "noise" in one modality (e.g., repeated high-pitched beeps interrupting spoken speech) may be treated as typical variance in another (e.g., repeated high-pitched beeps interrupting and cutting through lyrics). The superior performance of the contextually weighted feature distance metric suggests that listeners recalibrate their expectations about phonetic variability based on the modality.

This recalibration appears to be sophisticated and feature-specific, with listeners expecting different patterns of variation for different phonetic features. For example, vowel centralization in singing, a deviation that would be noise in spoken speech, is treated as expected variance in sung speech. This suggests that listeners maintain separate distributions of expected phonetic variation for different communication contexts, allowing for more accurate and efficient interpretation across diverse settings.

My findings support Levy's (2008) concept of an Error Identification Signal (EIS) as a mechanism for allocating cognitive resources during language processing. The sharper probability distributions observed with context-specific metrics suggest that in familiar linguistic contexts, listeners may generate more precise expectations, allowing for more efficient allocation of processing resources. In other words, this suggests that listeners are sensitive to the statistical regularities specific to sung speech. This sensitivity allows them to generate more accurate expectations about likely interpretations, thereby reducing the cognitive load as-

sociated with processing within this modality.

The relationship between expectability and processing efficiency has implications for theories of prediction in language comprehension (Anderson et al., 2018). My results suggest that when listeners can accurately predict the types of distortions likely to occur in a specific modality, they can more efficiently allocate attention and processing resources. This may explain why experienced listeners of certain musical genres report less difficulty understanding lyrics than novices, as their exposure to genre-specific patterns of distortion may allow them to form more precise expectations and thus process the input more efficiently.

In a similar vein, this perspective may actually also help explain why mondegreens, which seem to embed themselves deeper as experience and exposure to them increases, so frequently occur in the first place. In contexts like music that are culturally embedded within certain highly-regular communicative structures such as pitch temperament, instrumentation, timber, or mode,[8] listeners may develop overly specific but potentially misleading priors about what constitutes "normal" variation. When these priors become entrenched through repeated exposure to the exact same misheard interpretation, in a context where we are accustomed to deriving structural meaning beyond the words themselves in a systematized way, we may effectively treat the mondegreen as the norm rather than the exception. By engaging in a self-reinforcing cycle where initial misperceptions can become progressively resistant to correction in the right context, the efficiency of context-specific processing comes with the trade-off of occasionally converging on misinterpretations, especially when emotional *resonance* or personal satisfaction with a particular piece of music's interpretation further strengthens these over-specified priors. But I'm just spitballing.

[8] Here I refer to "mode" in the music theoretical sense, as in a tonality or collection of pitches with a certain culturally-embued set of meanings, rather than in the sense of the various contextually-determined modes investigated.

Conclusion

This thesis has examined how listeners navigate the perceptual labyrinth of sung speech. By interrogating the Noisy Channel Model (NCM) through the lens of modality-specific adaptations, I've delineated a process that balances contextual expectations with perceptual realities to derive coherent interpretations. Several substantive insights have fallen out of this.

First, the contextualization of prior probability and likelihood metrics significantly enhances our capacity to predict lyric interpretation. Standard phonological measures like Levenshtein distance, while sometimes serviceable for spoken speech, inadequately capture the phonetic quirks endemic to sung speech. Vowel centralization, prosodic disruption, and consonant articulation shifts fundamentally transform how signals map to interpretations in this domain. The introduction of contextually-calibrated feature-based distance metrics yielded more accurate predictions, suggesting that listeners hone their noise compensation strategies to accommodate modality-specific distortions. This refinement does not manifest as a crude statistical correlation but as a sophisticated cognitive calibration, with listeners reconfiguring their interpretive biases to align with the expected noise distribution of the ambient communicative environment.

Second, the methodological contrast between forced-choice and free-response paradigms gave significant differences in terms of the magnitude of the results they suggest. Forced-choice tasks, while experimentally tractable, artificially constrain the possibility space and potentially obscure the natural interpretive processes that we use when actually practicing language. The free-response paradigm revealed a more diverse array of interpretations, with participants generating variants that went far, far, so so far beyond the bounds of pre-fabricated alternatives. This methodological insight carries implications beyond sung speech, and suggests that studies of linguistic interpretation in general might benefit from employing

tasks that more faithfully mirror naturalistic communication contexts.

Third, my findings reinforce the Bayesian underpinnings of the NCM, while simultaneously complicating its application. The integration of likelihood and prior probability remains the fulcrum of interpretation, but these parameters mutate across communicative contexts in ways that fixed metrics simply cannot capture. Alternatively, these fixed metrics may be fixed, but they have a larger domain than previously assumed. This insight beckons a more flexible conception of the model, one which acknowledges the probabilistic infrastructure of language processing while accommodating the context-sensitive calibration of its components.

Linguistic processing is not, by anyone's account, a monolithic apparatus uniformly applied across contexts, but rather a flexible system that dynamically modulates its interpretive strategies based on contextual cues. This modularity enables listeners to efficiently process linguistic input across diverse communicative environments, from cocktail parties to opera houses, without requiring independent cognitive mechanisms for each. In a broader theoretical landscape, these findings harmonize with emergent perspectives on predictive processing that emphasize the brain's role in continuously generating and refining predictions about sensory input. When processing lyrics, listeners aren't passively decoding acoustic signals but actively constructing interpretations by integrating prior knowledge with perceptual evidence, with both components calibrated to the specific demands of musical contexts. This positions the NCM as a psychologically valid model of human interpretive processes, one that captures both the probabilistic foundations of language processing and its context-sensitive implementation, rather than simply as a computational convenience.

This adaptive capacity enables consistently robust and recoverable interpretation under suboptimal conditions, reminding us that this is a linguistic system optimized not for ideal laboratory environments but for the messy, multimodal realities of everyday communication.

6.1 Future Directions

Although outperforming previous models twice over, mine still cannot account for 60% of the variation observed in my dataset. These findings open avenues for future research on and within the Noisy Channel Model, and contextual adaptations in speech perception. While linguistic theories must traditionally assume

that listeners process language in a modality-agnostic way by perceiving the exact intended message, these results allow us to extend frameworks like the NCM to more accurate and nuanced understandings of how listeners actually process language in different contexts. While the NCM has proven remarkably effective in explaining language perception across various contexts, several important questions remain unanswered. Future research might address:

6.1.1 Direct Cross-Modal Comparisons

A direct extension of my work would be a study that directly compares how the same strings of language are interpreted across spoken and sung modalities. This controlled comparison would allow for precise measurement of how modality-specific adaptations emerge and operate. By presenting identical linguistic content in both spoken and sung forms to the same participants, one could isolate the specific perceptual adjustments listeners make when switching between modalities.

Such a study could systematically vary acoustic properties known to differ between spoken and sung speech (e.g., vowel centralization, consonant articulation, prosodic features) while controlling for other variables. This would enable researchers to build more detailed models of how linguistic context shapes the parameters of the NCM, particularly how the likelihood function is calibrated for different communication contexts.

I predict that participants would show systematically different patterns of misperception between the two modalities, with errors in the sung condition more closely aligned with the known articulatory constraints of singing. These differences would likely be captured more effectively by context-specific metrics than by general metrics, which would further the approach developed in this thesis.

One hurdle that such a study would have to overcome is that it would have to develop a set of stimuli, where each stimulus is a pair of sung and spoken versions of the same string. The researchers would have to control for speaker-specific effects in the stimuli, likely by using a range of different speakers for each stimulus. For such a study to be generalizable to how listeners typically perceive sung music, researchers would also have to produce a set of sung stimuli that are stylistically similar enough to typical distributions on sung music. This means that researchers would have to produce a large set of song clips, sung by a large variety of singers, that are stylistically varied across many different genres, in-

stead of re-using pieces of music that trained professional musicians have already performed. Alternatively, researchers could track down each of the 37 artists used in the stimuli set from this thesis, and request that they speak the lyrics instead of sing them so they may be paired with the sung clips. This seems difficult as well, as many have passed away.

6.1.2 Individual Differences in Adaptive Capacity

Listeners likely vary in their ability to adapt to different communication contexts. Although it would require substantial amounts of data collection, future work could explore what factors affect a listener's ability to efficiently recalibrate their perceptual systems when processing sung speech. Individual differences in cross-contextual interpretive ability might also help explain the variability in lyric comprehension that isn't explained by the current model.

For example, how quickly do listeners adapt their perceptual systems when switching between different linguistic contexts? My current research treats modality-specific adaptation as a stable state, but in reality, listeners constantly move between different communication contexts. Studies employing rapid switching between modalities could reveal the temporal dynamics of this adaptation and identify effects on adaptation speed.

Context Effect Discovery

This thesis used previous research to develop a context-specific noisy channel model of speech perception that made more accurate predictions. This process could be carried out in reverse: future work could collect intelligibility data in a different context (other than singing), and then perform computational regressions to determine which simple changes to the noisy channel model give significantly better predictions of speech intelligibility. Ostensibly, these changes might mirror the real-world effect of whichever context we've chosen on speech transmission. Following work could then experimentally (in)validate these hypotheses.

This methodological innovation could allow researchers to use the NCM as a tool for discovering previously unidentified context effects on speech perception. It could be particularly valuable for investigating understudied communicative contexts such as speech in reverberant environments, speech produced during physical exertion, or speech produced among a large crowd. By identifying which

parameters of the NCM require adjustment to accurately model these contexts, researchers could generate novel hypotheses about the specific distortions introduced by these conditions and how listeners adapt to them.

6.2 Computational Linguistics and the Convergence of Formal Systems

The findings of this thesis represent a point of convergence between the linguistic and computational sciences, fields that have historically maintained a symbiotic yet distant relationship. The NCM, with its mathematical scaffolding and probabilistic framework, demonstrates how formal computational models can render visible linguistic phenomena that otherwise remain obscured by non-constructive and descriptive approaches alone. When approaching language as a system partially configured by exposure to linguistic data, rather than as one exclusively theorized in terms of innate capabilities, we gain unique leverage to discern, for example, precisely how and when language users selectively integrate or disregard information from the ambient signal.

This perspective operates at a different level than the Chomskyan distinction between competence and performance. While traditional generative accounts focus on the abstract knowledge systems "necessarily" underlying language production, the NCM addresses how those systems interact with varying and inconsistent environmental factors. In this sense, what we observe is not merely a performance "deficit" or "incompetence," but a highly calibrated performance *adaptation* on which we can rely when these necessarily-present systems fail us. The regularities in how we make mistakes, and our ability to explain mistakes' regularities in terms of *enhancing* our communicative abilities, indicates that theory can lend much more nuance than a binary distinction between competent and incompetent linguistic behavior. The remarkably precise contextual adjustments evidenced in my results suggest something of a systematized knowledge of expected noise distributions across communicative contexts, which we might call *modal* competence.

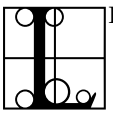
The promise of computational approaches to linguistics lies not in supplanting traditional theoretical frameworks but in formalizing, quantifying, validating, and extending their claims by actually constructing some of the structures they describe. I'd like to think this thesis exemplifies how computational techniques can validate linguistically motivated hypotheses about perceptual adaptation in

ways that are not possible with descriptive approaches or traditional methods of data analysis alone. Neither mathematical formalism nor linguistic theory is subordinate to the other, but rather mutually enriching.

Conversely, linguistics offers the computational sciences a wide domain of complexity, and has been a source of inspiration for the development of many of the most successful advances in the field. The modality-specific adaptations documented in this thesis represent precisely the kind of complex yet structured behavior that pushes computational models beyond narrow, contextually impoverished regimes. Rather than reducing language to a simplistic input-output system, this work demonstrates how computational models can capture the complex relationship between signal, noise, and knowledge.

While listeners continuously apply processes of probabilistic inference to extract meaning from noisy signals, with each guess always informed by the one prior, so too do researchers repeatedly refine, review, and rectify our formal models to better approximate these very processes. The abundance of recursive phenomena at the edge of chaos is sublime.

Appendix: Probability Review



ET's briefly revisit the some of the elements of probability theory that are relevant to this thesis. Probability theory is a box of tools often used to reason about *events*. An event is simply an element of some *event space*.

Before we define event spaces, we need to define *sample spaces*. If X is a set, and the elements of X are all the possible outcomes of a single event, then we call X a sample space. If we want to specify that x is one of the outcomes in the sample space X, we can write $x \in X$.

An *event space* E can be thought of as all the possible combinations of the outcomes in the sample space.[9] In other words, while the sample space only contains individual outcomes, a single event can contain multiple outcomes. For example, an event x can have the form "a or b or c" (alternatively written $a \cup b \cup c$), where $a, b, c \in X$. This means that the event x is considered to have "happened" if either a, b, or c is the outcome of the event. Outcomes and events are typically written in minuscule, and their spaces in majuscule, as we see here.

In addition to a sample space and an event space, the final tool in our box is a *probability function*. A probability function Pr assigns a value between 0 and 1 to each element of the event space (including to each of the event space that is also an elements of the sample space). In order to give a coherent definition of probability, the probability of all outcomes must also sum to 1. Formally,

1. The sample space X is a set of outcomes $a \in X$.

2. The event space E is $E \subseteq \mathcal{P}(X)$.

[9] This is the same as the power set in the discrete case. Going forward in this section I'm going to pretend all probability mass functions have discrete and definite domains.

3. The probability function Pr is $\Pr : E \to [0, 1]$,
   such that $\Pr[\varnothing] = 0$ and $\sum_{x \in X} \Pr[x] = 1$.

Beyond these reasoning tools, we have orthographic representations of ways they can be used. The notation $\Pr[x|y]$ represents a measure of *conditional* probability, and is read as "the probability of x given y." This refers to the likelihood that the event x also occurred if we already know that y has definitely occurred. Another way of posing this is as asking what the probability is that the outcome of the event was also in x if we know that it was in y. If the events x and y do not contain any of the same outcomes, then we can simply evaluate $\Pr[x|y]$ as 0.

$$\begin{aligned}(x \cap y = \varnothing) &\Leftrightarrow \Pr[x|y] = 0 \\ &\Leftrightarrow \Pr[y|x] = 0\end{aligned}$$

This is because if we assume that y occurred, then regardless of which specific outcome within y occurred, we know that none of the outcomes within x occurred because their intersection is empty. If two events are in an implicational relationship, then they must share some outcome.

If there are outcomes that are contained in both events, then we calculate $\Pr[x|y]$ as

$$(x \cap y = s \neq \varnothing) \Rightarrow \Pr[x|y] = \frac{\Pr[s]}{\Pr[y]}.$$

In this case, we first calculate $\Pr[s]$, which is the probability that the outcome is in both in event x and in event y, and then divide by the probability of event y.[10] As a sanity check, we can observe that this case is actually a generalization of the first, where

$$x \cap y = s = \varnothing \Rightarrow \Pr[s] = 0 = \frac{\Pr[s]}{\Pr[y]}$$

Two events x and y are said to be *independent* of one another if the chance of one happening is unrelated to the chance of the other happening. Formally, x and y are independent of each other if and only if

[10] We take this second step to account for the fact that we want to be able to know that if some outcome in y really did occur, that for a fixed value of $\Pr[s]$, we can ensure that as the probability of y grows, making our assumption less informative, the chance that the outcome is also in s scales proportionately.

$$\Pr[x \cap y] = \Pr[x] \cdot \Pr[y].$$

This helpful property allows us to re-write conditional probabilities on independent events like so:

$$\begin{aligned}\Pr[x|y] &= \frac{\Pr[s]}{\Pr[y]} \\ &= \frac{\Pr[x \cap y]}{\Pr[y]} \\ &= \frac{\Pr[x] \cdot \Pr[y]}{\Pr[y]} \\ &= \Pr[x]\end{aligned}$$

Using the same method we can also calculate that $\Pr[y|x] = \Pr[y]$ when x and y are independent. While the insight that unrelated events shouldn't affect each other may feel obvious, statistical independence is an extremely helpful tool for probabilistic analysis when it can be applied.[11]

In the general case, if we don't want to assume that two events are independent of one another, we can calculate the probability of events x and y cooccurring by first calculating the probability that x occurs given y, and scaling that by the probability that y occurred in the first place.

$$\begin{aligned}\Pr[x \cap y] &= \Pr[x|y] \cdot \Pr[y] \\ &= \Pr[y|x] \cdot \Pr[x]\end{aligned}$$

This is again a generalization of our earlier calculation with independent events. In that case, since the events are independent, we were able to assume that $\Pr[x|y] = \Pr[x]$. Combining the facts that $\Pr[x|y] = \frac{\Pr[x \cap y]}{\Pr[y]}$ and that $\Pr[x \cap y] = \Pr[y|x] \cdot \Pr[x]$, we arrive at a general formula for computing $\Pr[x|y]$ without assuming independence, namely Bayes' theorem:

$$\Pr[x|y] = \frac{\Pr[y|x] \cdot \Pr[x]}{\Pr[y]}$$

[11] As an aside, probabilistic independence can be generalized to a third event as well: Events x and y are independent given event z if $\Pr[x \cap y|z] = \Pr[x|z] \cdot \Pr[y|z]$.

When we write $x \sim \Pr$, we read it as "x is sampled from the distribution of Pr." This means that x is a randomly sampled element of the domain of Pr, in this case is E, weighted by the probability assigned to each event in E by Pr. We're choosing an event out of all possible events, but probable events are more likely to be chosen than improbable ones. That means that when we sample x from the distribution of Pr, we expect x to be more likely under Pr than a truly randomly selected event from the event space E.

$$\mathbb{E}_{X\sim\Pr}[\Pr(X)] \geq {}^{12}\mathbb{E}_{Y\sim U}[\Pr(Y)]$$

Finally, I review the *maximizing argument* operator, written arg max. This operator tells you which value maximizes the result of some calculation. Concretely, it takes any set and any function parameterized by that set, and returns the element of that set for which the function evaluates to the highest value when using that element as parameter. Here's an unnecessarily precise definition:[13]

$$\begin{aligned}\arg\max_{S} :&(S \times f) \to S \\ &\text{such that } \forall x \in S, y \in D \subseteq S, f : D \to (I, \geq), \\ &\arg\max_{s\in S} f(s) \mapsto x \Leftrightarrow \forall y \in D : \lim_{i\to x} f\,(i) \geq_I \lim_{i\to y} f\,(i)\end{aligned}$$

To get a concrete intuition for what this means, let's consider some specific, arbitrary example. Say our function is a map from the weight of an object, to how far we expect to be able to throw that object. Suppose that you cannot throw heavy objects very far, because you lack the strength. Suppose that you cannot throw light objects very far either, because air resistance quickly counteracts their low inertia. The *maximizing argument* of this function, then, is the weight of the object you can throw the farthest, somewhere between these extremes. Any lighter or heavier object will not be thrown as far. On the other hand, the *maximum* of this function is the distance that that object can be thrown. The arg max is in the pre-image of the maximum.

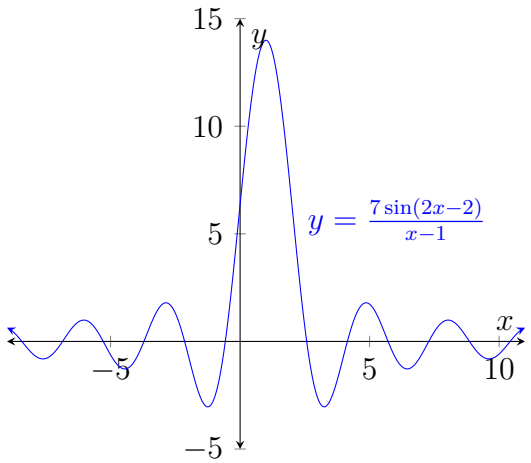
We'll evaluate the maximizing argument of the real-valued function defined by the formula

$$f(x) = \frac{7\sin(2x-2)}{x-1}.$$

[12] This is a strict inequality unless Pr is the uniform distribution.

[13] The image I of the function f must also admit a partial order $\geq$ in order for a "maximum" subset of I to be defined.

Our goal here is to determine the value of $\arg\max_{x\in\mathbb{R}} f(x)$, i.e. which $x \in \mathbb{R}$ approaches the largest value possible when using that x to calculate $\frac{7\sin(2x-2)}{x-1}$.



Using the graph of this function to make an estimate, we see that a unique solution for this value appears to be 1, since the value of y is greatest when the value of x is close to 1.

$$\arg\max_{x\in\mathbb{R}} \frac{7\sin(2x-2)}{x-1} = 1$$

It might also be helpful later to convince yourself that $\arg\max_{x\in\mathbb{R}} \frac{7\sin(2x-2)}{x-1}\cdot\mathbf{c}$ is also equal to 1 for all values of $\mathbf{c} \in \mathbb{R}_+$, since scaling an image by a positive value doesn't affect the locations of its maxima. Regardless of how much you vertically compress or stretch the graph above, the location of its highest peak doesn't change. This is why we are able to ignore the quotient when applying Bayes' Theorem in this thesis to approximate the maximizing arguments of probabilistic functions. Once we have these probabilistic tools under our belt, we can begin to explore the specifics of how they have been used to construct a mathematical model of language interpretation in noisy conditions.

Cited Works

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., & List, J.-M. (2018). A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznan Linguistic Meeting*, *4*(1), 21–53. https://pressto.amu.edu.pl/index.php/yplm/article/view/yplm-2018-0002

Augustine of Hippo (397-400). *Confessiones*. Book VII.

Benolken, M. S., & Swanson, C. E. (1990). The effect of pitch-related changes on the perception of sung vowels. *The Journal of the Acoustical Society of America*, *87*(4), 1781–1785.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Chen, S., Nathaniel, S., Ryskin, R., & Gibson, E. (2023). The effect of context on noisy-channel sentence comprehension. *Cognition*, *238*, 105503.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, *25*(5), 975–979.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *Journal of the Acoustical Society of America*, *116*(6), 3647–3658.

Collister, L. B., & Huron, D. (2008). Comparison of word intelligibility in spoken and sung phrases. *Empirical Musicology Review*, *3*(3), 109–125.

Condit-Schultz, N., & Huron, D. (2015). Catching the lyrics: Intelligibility in twelve song genres. *Music Perception: An Interdisciplinary Journal*, *32*(5), 470–483.

Deme, A. (2015). *Phonetic Analysis of Sung Vowels*. Ph.D. thesis, Eötvös Loránd University, Budapest, Hungary.

Deme, A. (2017). The identification of high-pitched sung vowels in sense and nonsense words by professional singers and untrained listeners. *Journal of Voice*, *31*(2), 252.e1–252.e14.

Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little, Brown and Co.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive psychology*, *47*(2), 164–203.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of experimental psychology. Human perception and performance*, *6 1*, 110–25. https://api.semanticscholar.org/CorpusID:14340223

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Hallé, P. A., Chang, Y., & Best, C. T. (2004). Identification and discrimination of mandarin chinese tones by mandarin chinese vs. french listeners. *J. Phonetics*, *32*, 395–421. https://api.semanticscholar.org/CorpusID:1790144

Hirsh, I. (1948). The influence of interaural phase on interaural summation and inhibition. *The Journal of the Acoustical Society of America*, *20*, 536–544.

Hollien, H., Mendes-Schwartz, A. P., & Nielsen, K. (2000). Perceptual confusions of high-pitched sung vowels. *Journal of Voice*, *14*(2), 287–298.

Hume, D. (1739-1740). *A Treatise of Human Nature*. Book I, Part IV, Section VI.

Johnson, K. (1997). *Speech Perception without Speaker Normalization: An Exemplar Model*, chap. 8, (pp. 363 – 389).

Johnson, R. B., Huron, D., & Collister, L. (2014). Music and lyrics interactions and their influence on recognition of sung words: An investigation of word frequency, rhyme, metric stress, vocal timbre, melisma, and repetition priming. *Empirical Musicology Review*, *9*(1), 2–20.

Kant, I. (1781). *Kritik der reinen Vernunft*. Transcendental Dialectic, Book II, Chapter II.

Kauf, C., & Ivanova, A. (2023). A better way to do masked language model scoring. *arXiv preprint arXiv:2305.10588*.

Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: The ci perspective. *Discourse Processes*, *39*(2-3), 125–128. https://doi.org/10.1080/0163853X.2005.9651676

Koelsch, S., Vuust, P., & Friston, K. (2019). Predictive processes and the peculiar case of music. *Trends in cognitive sciences*, *23*(1), 63–77.

Koroteev, M. V. (2021). Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Kreiner, B. (2023). Genius lyrics dataset. https://huggingface.co/datasets/brunokreiner/genius-lyrics.

Kuhn, A. (1853). Ueber die durch nasale erweiterten verbalstämme. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen*, *2*(5), 392–398. http://www.jstor.org/stable/40844305

Kuperberg, G. R., & and, T. F. J. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. PMID: 27135040. https://doi.org/10.1080/23273798.2015.1102299

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, *10*, 707–710. https://api.semanticscholar.org/CorpusID:60827152

Levy, R. (2008). A noisy-channel model of human sentence comprehension under uncertain input. *Proceedings of the 2008 conference on empirical methods in natural language processing*, (pp. 234–243).

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54 5*, 358–68. https://api.semanticscholar.org/CorpusID:10117886

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). the unconscious initiation of a freely voluntary act. *Brain*, *106*(3), 623–642.

Liu, Y., Ryskin, R., Futrell, R., & Gibson, E. (2020). Structural frequency effects in noisy-channel comprehension. Presentation at the Penn Linguistics Conference.

Matuschak, A. (2012). Top-to-bottom; bottom-to-top.

Mazzone, M. (2013). Attention to the speaker. the conscious assessment of utterance interpretations in working memory. *Language & Communication*, *33*(2), 106–114. https://www.sciencedirect.com/science/article/pii/S0271530913000025

Misra, K. (2022). minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*(3), 299–325.

Poliak, M., Kimura, H., & Gibson, E. (2024). Mis-heard lyrics: an ecologically-valid test of noisy channel processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46.

Poppels, T., & Levy, R. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. In *CogSci*.

Rihanna, & Jay-Z (2007). Umbrella. Audio recording. Written by Christopher Stewart, Terius Nash, Kuk Harrell, and Shawn Carter. Produced by Tricky Stewart. Released by Def Jam Recordings on March 29, 2007. https://www.discogs.com/release/983310-Rihanna-Feat-Jay-Z-Umbrella

Rubehn, A. (2024). Generating phonological feature vectors with soundvectors and clts. *Computer-Assisted Language Comparison in Practice*, *7*(2), 59–67. https://ojs3.uni-passau.de/index.php/calcip/article/view/343

Samuel, A. (1981). The role of bottom-up confirmation in the phonemic restoration illusion. *Journal of experimental psychology. Human perception and performance*, *7*, 1124–31.

Schmitt, R. (1967). *Dichtung und Dichtersprache in indogermanischer Zeit*. Wiesbaden: Harrassowitz.

Schopenhauer, A. (1818). *Die Welt als Wille und Vorstellung*, vol. 1.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10–21.

Smith, L. A., & Scott, B. L. (1980). Increasing the intelligibility of sung vowels. *The Journal of the Acoustical Society of America*, *67*(5), 1795–1797.

Spinoza, B. (1677). *Ethica, ordine geometrico demonstrata*. Part II, Propositions 48-49.

Squires, L. (2019). Genre and linguistic expectation shift: Evidence from pop song lyrics. *Language in Society*, *48*(1), 1–30.

Sundberg, J. (1975). Formant technique in a professional female singer. *Acta Acustica united with Acustica*, *32*(2), 89–96. https://www.ingentaconnect.com/content/dav/aaua/1975/00000032/00000002/art00006

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta Psychologica*, *135*(2), 77–99. https://www.sciencedirect.com/science/article/pii/S0001691810000429

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vurma, A., Meister, E., Meister, L., Ross, J., Raju, M., Kala, V., & Dede, T. (2023). The intensities of vowels and plosive bursts and their impact on text intelligibility in singing. *The Journal of the Acoustical Society of America*, *154*, 2653–2664.

Weidema, J. L., Roncaglia-Denissen, M. P., & Honing, H. (2016). Top–down modulation on the perception and categorization of identical pitch contours in speech and music. *Frontiers in Psychology*, *7*, 817.

Wickham, E. (2013). From speech to song: A response to johnson, huron and collister on the interaction of music and lyrics. *Empirical Musicology Review*, *9*, 25.

Wright, S. (1954). The death of lady mondegreen. *Harper's Magazine*, *209*(1254), 48–51.

This study was supported in part by the Gillespie Family Student Research Fund.