



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Quantifying the perceptual value of lexical and non-lexical channels of spoken dialogue

Sarenne Wallbridge

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2024

Abstract

Spoken conversation is one of the most widely-used means of communication that exists. Still, the cognitive mechanisms we employ to decode information from the speech signal are not well understood. Predictive theories of language processing, which explain language comprehension as a function of the alignment between our predictions about the upcoming signal and its actual realization, have been widely adopted across the fields of psycholinguistics, cognitive science, and linguistics. However, empirical support for such theories stems primarily from written language comprehension.

This thesis is a study of how predictive processing may be involved in the comprehension of spoken dialogue. We focus on two features that distinguish spoken interaction from written sentences and how these features may alter comprehension behaviours. First, speech is a much richer signal than text; information can be communicated through both the lexical channel of *which* words are said, and the non-lexical channel of *how* those words are spoken. We examine how these channels contribute, both independently and jointly, to predictions about upcoming turns in dialogue. Second, making predictions in spoken dialogue involves reasoning about abstract aspects of language such as pragmatics that may not be evident in written sentences. As such, we shift focus away from the content of predictions to instead investigate their genesis—specifically, how do the lexical and non-lexical channels constrain the shape of expectations regarding the upcoming signal? To answer this question, we propose Perceptual Information Value which quantifies the value of a channel in terms of how it alters expectations, as well as a behavioural paradigm to measure it in spoken dialogue.

We begin by investigating predictions of the lexical content of dialogue. Our results show that humans generate expectations at the level of dialogue turns and that their predictions contain inherent variability. We argue that this variability is important component of how people process more realistic forms of language-use, such as spoken dialogue. Leveraging large language models as purely predictive processing mechanisms, we demonstrate a degree of alignment between human and model predictions; however, it is highly sensitive to the model architecture and training objective. In particular, model predictions do not contain the variability that we observe in human predictions. Next, we use this foundation to study how predictions manifest in spoken dialogue where messages can be distributed across both lexical and non-lexical channels. Our experiments show that expectations about spoken dialogue turns are a function of both lexical and non-lexical channels, as well as their joint expression.

Importantly, we find that access to channels can constrain expectations meaningfully, even if it yields less accurate predictions. We therefore argue that the perceptual value of information in a channel lies not in its effect on predictive accuracy but more generally in its capacity to shape expectations.

Lay Summary

Talking to each other is one of the most common ways we communicate. But even holding a mundane conversation is quite a complicated act. In a matter of seconds, we listen to our conversational partner, discern what they've said, decide what message we would like to respond with, and determine when we can start speaking ourselves. We still don't understand exactly *how* people are able to hold conversations so easily. Cognitive scientists have proposed that a key mechanism behind understanding language is by making predictions about what might come next. When our predictions are confirmed, we can process the language smoothly; when we are surprised by what comes next, it takes more effort. Although this idea is well-supported by research, most of the evidence comes from studies of how people read sentences.

Predicting what could come next in spoken conversation is far more complicated than predicting the next word in a written sentence. We focus on two key differences. First, our speech conveys information not only through *which* words we choose to say (the lexical channel) but also *how* we choose to say them (the non-lexical channel). Second, there are many more possibilities for what could happen next in a conversation compared to a written sentence. This thesis asks how the lexical and non-lexical channels of speech affect the predictions that people might make in spoken conversations.

We begin by studying whether people can make predictions in conversations at all—after reading part of a conversation, can people pick the true upcoming turn from a set of potential turns? Our results show that people can make accurate predictions when they read dialogue, but not perfectly. Instead, people can predict many potential upcoming turns and we believe that this foresight is a crucial component of language comprehension. We then conduct the same experiment with spoken dialogues—how do people's predictions change when they can hear the turns instead of reading them? What about when they can hear part of the preceding conversation? We find that people make predictions about both what words someone could say and how they might say them. Interestingly, hearing more of the conversation doesn't always make people more accurate but it does lead to more agreement about which turns are likely. This suggests that people have similar ideas about how conversations tend to unfold.

Alongside studying human predictions, we consider predictions from modern language models that process and generate text and speech. Although there are some similarities, models don't show the same degree of flexibility as people in predicting

different trajectories in a conversation and struggle to integrate the lexical and non-lexical channels. Studying more realistic forms of language—and how both people and models process them—reveals the importance of not simply considering whether a correct prediction can be made, but examining the range of continuations that may be predicted. These findings could help us assess the underlying causes when models, or even people, struggle to understand language.

Acknowledgements

The past 4 years have been tremendous and thanking everyone who made them so feels nearly impossible—but this is a start.

I've been incredibly fortunate to have not one, but two wonderful supervisors who provided me with the tools to formulate my own research questions, along with the freedom and support to investigate them. To Catherine Lai, I am especially grateful for your thoughtful guidance, compassion, and unwavering trust. To Peter Bell, your endless enthusiasm fuelled me through all of the highs and lows of my PhD. Together you gave me an incredible breath and depth of knowledge to learn from, and the confidence to pursue research questions that once felt beyond my reach. To Steve Renals, thank you for believing in my research proposal from the beginning. And to my examiners, Sharon Goldwater and Okko Räsänen—I thoroughly enjoyed discussing my work with researchers of your calibre; this thesis is undoubtedly stronger for it. Sharon, it was particularly special to have you examine my doctoral work after starting my research journey with you as an undergraduate student.

I was privileged to be a part of both the CSTR and the wider ILCC communities. Being surrounded by researchers from every corner of language research broadened my thinking along countless axes. My work benefitted from every hallway discussion, reading group, abstract session (and pub night). A heartfelt thank you to Raquel Fernández and the DMG for welcoming me to the University of Amsterdam and to Mario Giulianelli for facilitating; I look forward to discussing language, cognition, and music in many more places around the world.

An incredibly special thanks goes to my fellow students in Edinburgh, in particular Johannah O'Mahony, Dan Wells, Atli Sigurgeirsson, Irene Winther, Jacob Webber, Adaeze Adigwe, Sander Bijl de Vroe, Elizabeth Neilsen, and Ramon Sanabria. Thank you for the countless city walks, office chats, Holyrood runs, beach dips, pub evenings, and even late nights in the Forum. Though a PhD can seem like a solitary experience, it hasn't been for me. The friendships I've made here will last a lifetime.

To my family, I am forever grateful. Nigel, Carrol, Andrew, and Elise; you've been a constant source of encouragement, comfort, inspiration, and patience (not least with the ever-present wind on my end of the phone).

Finally, I must finish with a short love letter to the city of Edinburgh beyond the University. To the Jazz Bar crew—Jonny, Aki, Graeme, and the rest: though I've started to attend more jazz gigs than late slots, you've marked my life in the most

beautiful ways. No basement will ever mean more to me than this one. To the climbing community—especially Tommy, Max, and Eilidh—that taught me strength in so many senses of the word and has taken me to incredible corners of the earth, I can't wait for the next adventure. To the museums and the pubs, to the ever-changing array of coffee shops, to the Craggs and the sea and the mountains of Scotland, thank you. Learning the seasons of this place has been an utter joy.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Sarenne Wallbridge)

Contents

1	Introduction	1
1.1	Perception as prediction	3
1.2	Speech as a multi-channel system	5
1.3	Perceptual information value	9
1.4	Thesis overview & contributions	12
1.4.1	Part I: Lexical component of dialogue	13
1.4.2	Part II: Perceptual value of the channels	14
1.4.3	Part III: Multi-channel representations	16
1.5	Contributions	18
2	Background	20
2.1	Prosody	20
2.1.1	Definitions	20
2.1.2	Form of prosody	21
2.1.3	Functions of prosody	24
2.2	Machine representations of language	28
2.2.1	Learning mechanisms	29
2.2.2	Representing text	30
2.2.3	Representing speech	31
2.3	Predictive theories of language perception	33
2.3.1	Surprisal theory	33
2.3.2	Information Density	34
2.3.3	Estimating surprisal with large language models	35
I	Expectations in single-channel communication	39
	Part I Introduction	40

3	Examining expectations in written dialogue through surprisal	41
3.1	Introduction	41
3.2	Related work	42
3.2.1	Acceptability as a perceptual probe	42
3.3	Experimental design	44
3.3.1	The turn acceptability task	44
3.3.2	Data	44
3.3.3	Language model surprisal	45
3.3.4	Behavioural study and Participants	46
3.3.5	Surprisal on a turn level	47
3.4	Results & discussion	49
3.4.1	Dialogue turn acceptability	49
3.4.2	Quantifying predictive power of surprisal	51
3.5	Conclusions	54
4	Machine expectations in written dialogue	56
4.1	Introduction	56
4.2	Related work	57
4.2.1	Language modelling for dialogue	57
4.3	Experimental setup	59
4.3.1	Data	59
4.3.2	Models	59
4.4	Results & discussion	61
4.4.1	Correlation with plausibility perception	61
4.4.2	Discriminative performance	62
4.4.3	Model ablation for response selection	65
4.4.4	Relationship between response selection and predictive power	68
4.4.5	Discussion	69
4.5	Conclusions	70
II	Expectations in multi-channel communication	73
	Part II Introduction	74
5	Perceptual information value of the non-lexical channel	76
5.1	Related work	76

5.2	Turn discrimination paradigm	78
5.2.1	The discriminative task	78
5.2.2	Conditions	79
5.3	Experiment design	81
5.3.1	Data & processing	82
5.3.2	Stimuli construction	83
5.3.3	Online experiment	84
5.3.4	Evaluation metrics	84
5.4	Results & discussion	85
5.4.1	Perceptual judgements	85
5.4.2	Certainty analysis	88
5.4.3	Acoustic analysis	89
5.5	Conclusions	91
6	Perceptual information value across multiple channels	94
6.1	Introduction	94
6.2	Related work	95
6.3	Generalised turn discrimination paradigm & experiment design	96
6.3.1	Data & processing	96
6.3.2	Stimuli construction	98
6.3.3	Online experiment	102
6.4	Perceptual information value: evaluation metrics	102
6.5	Results	106
6.5.1	Evaluation methods	106
6.5.2	Participant performance	107
6.5.3	Applying entropy metrics to lexically equivalent stimuli	110
6.6	Conclusions	111
III	Multi-channel representations	115
	Part III Introduction	116
7	Systematicity in prosody independent of lexical content	117
7.1	Introduction	117
7.2	Background	119
7.2.1	Self-supervised learning	119

7.2.2	Representations of prosody	120
7.3	Methodology	121
7.3.1	Masked Prosody Model	122
7.3.2	Downstream tasks	124
7.4	Experimental setup	127
7.5	Experiments	128
7.5.1	Effects of mask and codebook size	128
7.5.2	Probe performance	129
7.5.3	Layer evaluation	131
7.6	Discussion & conclusions	132
8	Acoustic-conditioned turn discrimination	134
8.1	Introduction	134
8.2	Related work	135
8.2.1	SSL speech representations	135
8.2.2	Combining acoustic and lexical representations	137
8.3	Methodology	139
8.3.1	Experimental setup	139
8.4	Experiment I: individual responses	141
8.4.1	Results	141
8.5	Experiment II: variable responses	143
8.5.1	Results	144
8.6	Comparing to human judgements	151
8.7	Acoustic analysis	153
8.8	Discussion & conclusions	154
9	Conclusions	158
9.1	Directions for future work	160
IV	Appendices	164
A	Machine expectations in written dialogue	165
B	Acoustic-conditioned turn discrimination	170
	Bibliography	172

Chapter 1

Introduction

More so than any other species, humans are incredibly resourceful at putting to use almost any available medium in their quest to communicate with others (Scott-Phillips, 2014). Think about the last time you were at a dinner party: a silent nudge or a glance flashed to a friend can convey more information than the conversation happening at the table. A well-timed arch of the eyebrows from the speaker can completely invert the meaning of the words they are producing. If you are addressing someone on the other side of the room, you may subconsciously enunciate and raise the volume of your voice when speaking the most crucial words—“Time to *go*?”. Even an ordinary scenario like this one exemplifies our ability to harness all kinds of mediums—tactile, visual, auditory—to make our communication more efficient, nuanced, and robust. Each of these media can be viewed as a distinct communication channel with its own capacity for transmitting different types of information.

Although there is little agreement on how information is distributed across sets of channels in human communication, there is a wealth of empirical evidence that communicative productions are a function of the channels available in a particular communicative setting. Processing richer, multi-channel communicative signals could be expected to be more effortful; however, many studies suggest that multi-channel communication is more effective in a number of ways. For example, participants tasked with generating novel communication systems are more accurate and efficient when they can use multiple channels (Macuch Silva et al., 2020; Fröhlich et al., 2019), while Holler et al. (2018) find that questions accompanied by gestures receive faster responses. Studies of collaborative tasks consistently report higher success and faster rates of task completion when participants can make use of multiple communicative channels (Moll, 2013).

Access to multiple channels is theorized to make communication more efficient through both increased expressivity—the breadth of information that we can convey—and robustness to noise—the likelihood that our message will be understood. Different channels will be more appropriate for conveying certain types of information, while the joint use of channels can make a message more expressive through interactions of their content, or less susceptible to noise through redundancy of what is encoded in them (Scott-Phillips, 2014). Consider again the examples in the first paragraph of this section; each depicts one of these efficiency gains. Even so, current computational models of the processes and information involved in how language comprehension have been almost exclusively limited to the study of linguistic signals consisting of a single channel—text—and concerned with how we process individual words or sentences isolated from a wider discourse context. Neither of these features reflects the type of language that humans have almost certainly evolved to process: direct communicative interaction (Levinson and Holler, 2014).

This thesis presents a study of how people comprehend one of the most common forms of communication: spoken dialogue. We are far from the first to advocate the importance of studying realistic forms of language use for understanding the processes that support language comprehension. For example, Pickering and Garrod (2004) argue that language processing theories developed only to explain the comprehension of written sentences will inevitably fall short due to fundamental differences between sentences and dialogue. However, the set of tools available for representing language that is at researchers’ disposal today looks quite different to those that were available when many models and theories of human language comprehension were developed (for example, Surprisal Theory (Hale, 2001) or the Smooth Signal Redundancy hypothesis (Aylett and Turk, 2004); we discuss these in far greater detail in Section 2.3). Not only do neural language models provide us with much more complex characterizations of language that account for wider windows of context, but we also have the ability to collect human judgements of rich stimuli at much larger scales than has ever been possible. With platforms like Prolific Academic or Mechanical Turk, judgements that would have taken days or weeks to collect in a lab can now be obtained in hours.

In the proceeding chapters, we focus on two features that differentiate spoken dialogue from isolated written sentences. First is the importance of context in understanding communicative utterances. Dependencies between an utterance and its corresponding context, such as co-reference, have long been recognized as fundamental features of dialogue (Morgan, 1973; Sacks et al., 1978; Ginzburg, 2001). The effect

of such contextual dependence may not be apparent when studying the comprehension of isolated sentences. Second, speech is a much richer signal than text. Speakers can distribute their message across multiple channels—both the lexical channel of *which* words they choose to say, and the non-lexical channel of *how* those words are spoken. Building on predictive theories of language comprehension, we ask how these channels affect the ways in which people make predictions during the comprehension of spoken dialogue.

In the remainder of Chapter 1, we formally introduce the core components that underpin our exploration of spoken dialogue perception as a multichannel system. We begin by introducing the theoretical basis for studying language processing as a predictive process in Section 1.1, and then defining and motivating our multi-channel framing of speech in Section 1.2. Finally, we introduce Perceptual Information Value, a proposal for allocating value to information based on the degree to which it affects expectations of the upcoming communicative signal.

1.1 Perception as prediction

Many prominent theories of human language comprehension—discussed in detail in Chapter 2—are centered around the conceptualisation of perception as prediction (Friston et al., 2006; Clark, 2013; Christiansen and Chater, 2016, *inter alia*). These theories suggest that we are constantly monitoring and anticipating the trajectory of communication. Cross-modal eye-tracking studies provide clear demonstrations of anticipatory processing: when provided with an image and a spoken sentence, listeners often start to look at plausible objects for the main verb of a sentence as soon as they hear the verb but before an object has been spoken (Bromley et al., 1993). Generating accurate predictions about what may happen next allows us to interact with the world around us efficiently by minimizing the effort required to process what we encounter (Friston, 2003). Our expectations about the upcoming communicative signal are influenced by the immediate linguistic context—such as the word that was most recently spoken and our interpretation of a speaker’s underlying goals—and mediated by our prior beliefs which include our understanding of how language tends to be structured, as well as pragmatic inferences such as our relationship with our conversational partner (Griffiths et al., 2010).

Both production and comprehension of communication provide ample evidence of latent predictive processing (Pickering and Garrod, 2004). An example that is partic-

ularly relevant to this thesis can be found in spoken conversation where the pauses between speaker changes average a mere 200ms across languages (Stivers et al., 2009; Stokoe, 2018). This is far less time than would be required to allow our conversational partner to finish speaking, plan a response, and then begin executing it. Rather, predictive theories suggest that we engage in several forward-looking processes: predicting what the speaker is aiming to convey, planning our response, and projecting both when they will finish and when we should launch the execution of our reply; as shown in the proceeding chapters, it is not uncommon for the latter to precede the former (Sacks et al., 1978; Levinson and Torreira, 2015; Garrod and Pickering, 2015). We tend to manage these processes extremely successfully and notice mismanagement very quickly.

As is often the case in scientific writing, terminology is saturated and often overlaps between fields. When researchers refer to “predictions” in studies of human perception, we are often interested in the act of assigning a probability to a future event—“what is the probability that word ‘blue’ will occur next?”, for example. If the corresponding probability is less than 1, it follows that the prediction itself belongs to a set containing alternative predictions. Throughout this thesis, we refer to this conceptual distribution of potential predictions and their associated probabilities as *expectations* for a given context. Note that this stands in contrast to the mathematical expectation function $E[X]$ which produces a singular value by integrating over all possible values of the random variable X .

Alternative accounts of language perception. The large body of empirical evidence for predictive processing in language comprehension makes it a useful basis for this thesis. However, we note that prediction is unlikely to be the only mechanism driving language comprehension. Predictive accounts can trace their roots back centuries to philosophers like Kant and von Helmholtz (1924) who proposed that human visual perception involved “unconscious inference”¹. The original von Helmholtz article written in German was published even earlier, in 1867, but predictive processing has only moved to the fore of psycholinguistics relatively recently. Other mechanisms posited to support language comprehension include integration—connecting new incoming information to existing knowledge or established concepts (Kintsch and Van Dijk, 1978)—and good-enough processing, which focuses on the role of underspec-

¹In their pre-emptive work framing perception using generative models titled “The Helmholtz Machine”, Dayan et al. (1995) paid direct homage to these ideas. Swanson (2016) provides a thorough timeline of how predictive accounts of perception emerged.

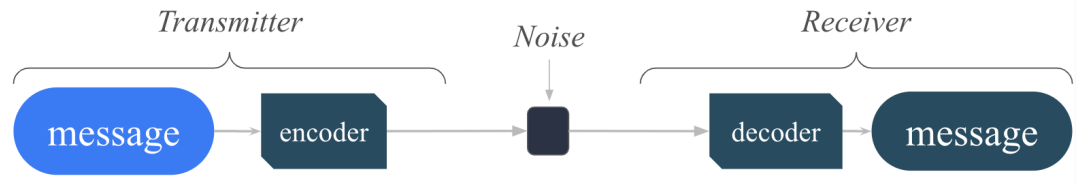
ification and incompleteness in language processing (Sanford and Sturt, 2002; Ferreira et al., 2002). Recent works have proposed that such accounts are compatible with, or at least complementary to, predictive processing, however the question of when, where, and who applies predictive processing is still open (Ferreira and Chantavarin, 2018; McConnell and Blumenthal-Dramé, 2021).

Machine learning. The perception-as-prediction paradigm may have stemmed from Cognitive Science but has garnered attention in related fields of psycholinguistics, psychology, and, perhaps more surprisingly, machine learning, where parallels have been drawn between the act of prediction in language comprehension and some self-supervised learning mechanisms (Tian et al., 2020; Leong and Linzen, 2024). These learning objectives underpin a large majority of the neural language representations developed over the last decade (Mikolov et al., 2013; Devlin et al., 2019). As opposed to classical supervised machine learning methods which are a function of externally generated labels, a self-supervised method uses its own input as a training signal. Representations are learned by tasking a model to predict parts of the input signal that have been obscured or corrupted. For data with a temporal axis such as language, a popular variant of self-supervision involves making predictions about the upcoming signal (Radford et al., 2019). The shared aspect of prediction has catalysed the study of human language perception through the application of such models. We provide a much deeper discussion of this learning mechanism and examples of such representations of text and speech in Chapter 2.

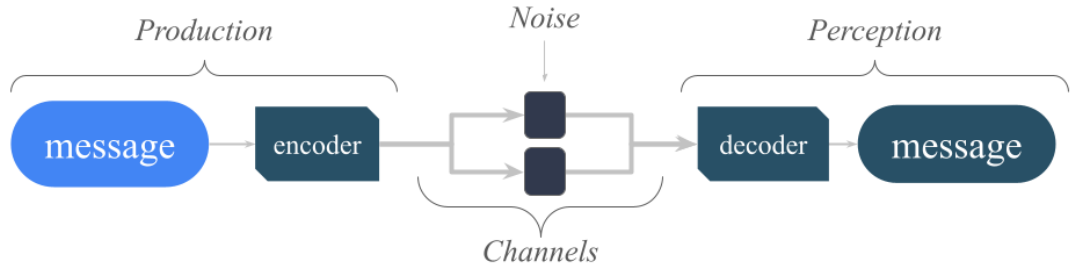
In sum, although it may not be the only process at play during language comprehension, the perception-as-prediction framing provides us with an opportunity to examine how predictive processes may be employed when people process spoken communication as well as the alignment between human perception and the predictive mechanisms that power many self-supervised machine learning methods.

1.2 Speech as a multi-channel system

When considering communication as an act of explicit information transmission, one of the most established models is Shannon’s noisy channel depicted in Figure 1.1a (Shannon, 1948). Under this framework, communication is a system in which people use lossy encoding and decoding processes to transmit information through a noisy



(a) An illustration of Shannon's classical noisy channel communication model.



(b) The multi-channel extension (depicted with two channels). Rather than flowing through a single channel, the producer can distribute their message across both channels. Different information distribution strategies can be leveraged to make multi-channel communication more expressive, efficient, and robust.

Figure 1.1: Variants of the noisy channel communication model.

channel. Though originally developed in the context of telegraph communication where the communicative signal consisted of electrical pulses through a channel of copper wiring and noise referred to electrical attenuation, Shannon himself applied this framework to written communication. It has since been used to investigate an extensive range of linguistic and psycholinguistic phenomena (Levy, 2008b; Bergen et al., 2012; Futrell and Gibson, 2017; Futrell et al., 2020). However, most applications maintain the original depiction with a single channel of information transmission.

As we have discussed in the preceding paragraphs, communicative productions are a function of the channels available to the transmitter; by distributing the components of their message across the channels available to them, the transmitter can make their communication more efficient, expressive, and robust. The multi-channel extension that we propose, depicted in Figure 1.1b, allows us to examine the strategies people may use to distribute information across speech. In particular, we consider how channels may be used in isolation and in conjunction with each other. We detail our proposed channel definitions as well as a theoretical framework for talking about channel interactions below.

Defining channels. The generic framework in Figure 1.1b doesn’t place constraints on either the number or definition of channels. Although sensory modalities seem like an obvious starting point for this, the term “channel” is used rather loosely in studies of human communication. For example, perceptual studies often consider gaze and facial expression to be separate sources of information (Pecchinenda and Petrucci, 2016; Wu et al., 2019). Multi-modal studies of language in machine learning research have gained significant attention² and use similarly fluid definitions of modality that are often driven by practical constraints rather than adherence to sensory modalities (Parcalabescu et al., 2021; Baltrušaitis et al., 2018). As such, we propose our own channels. In this thesis, we frame speech as a bi-channel system where information can be transmitted through the lexical channel of *which words* are said, and the non-lexical channel of *how* they are spoken.

The **lexical** channel of speech can be defined as the words that were said. This yields a rather simple representation: the transcript. The **non-lexical** channel, however, lacks such a neat definition. Alongside its lexical content, speech carries a host of information that is encoded relatively inadvertently by the speaker—features such as speaking rate and accent can be used to infer characteristics of a speaker’s identity such as their age, their mood, or the social groups that they may belong to (Eckert, 1989; Ladefoged et al., 2006; Skoog Waller et al., 2015)³. Non-lexical content can also be used to contribute to the message of an utterance more purposefully, particularly through changes to the intonation, loudness, and rhythm of speech. In this thesis, we are primarily interested in these more explicit features. Generally captured via their acoustic correlates (fundamental frequency, intensity, and duration, respectively), variations in the intonation, loudness, and rhythm of speech define the *form* of prosody (Ladd and Cutler, 1983; Wagner and Watson, 2010). It is widely accepted to play an important role in a range of communicative functions in English—from marking novel information and directing interlocutor attention to resolving lexical and syntactic ambiguities (Hirschberg and Pierrehumbert, 1986; Clark and Brennan, 1991; Ben-David et al., 2016; Cervone et al., 2018). However, the degree to which prosody can be dis-

²This is motivated in part by both increased compute capacity which allows processing of high-dimensional data, as well as philosophical arguments that language understanding cannot be achieved by learning from text alone and instead requires grounding in other modalities (Bisk et al., 2020; Bender and Koller, 2020). Similarly, much of the work in this thesis is only possible due to the advances in methods for processing rich linguistic signals.

³Many socio-cultural implications are woven into how different people speak the same language. This is exemplified in George Bernard Shaw’s *Pygmalion* which depicts the phonetician Higgins’s quest to “change Eliza into a whole new person by creating new speech for her”.

entangled from other aspects of non-lexical information is unclear (Xu and Armony, 2021; Williams, 2022). As such, we refer to this type of information as “non-lexical”. Prosody is discussed in much greater depth in Chapter 2

This definition of channels allows us to investigate how the perception of spoken language may differ compared to written language and offers insights into methods for representing speech automatically. In particular, they enable us to leverage and investigate the powerful existing machine representations of lexical content that are being increasingly used in and outside of academic research, and second, they provide an excellent opportunity to answer questions about how to represent prosody in perceptually-motivated ways. The problem of representing lexical information has dominated the field of speech recognition and spoken language understanding for decades. In contrast, the automatic extraction of prosodic information has seen much slower progress. We explore several reasons for this in the proceeding chapters of this thesis but in particular, non-lexical information is encoded at various timescales and across multiple acoustic features. This makes learning a vocabulary of discrete units—a fundamental step for many current automatic representation learning methods—far less straightforward than for its lexical counterpart.

While we’ve argued that this framing can provide valuable insights, it has drawbacks. The line we’ve drawn between lexical and non-lexical information becomes increasingly blurred when studying speech produced spontaneously compared to prepared or read speech: which channel should laughter or a sigh belong to? What about filled pauses like “um”? The degree to which it is even possible to disentangle how lexical and non-lexical information are produced and interpreted is debatable. We address potential pitfalls of our channel definitions like these throughout this thesis.

Partial Information Decomposition. To conceptualize how the lexical and non-lexical channels of speech are used to convey information between speakers, we make use of the Partial Information Decomposition (PID) framework which offers a formal description of information in multivariate settings (Williams and Beer, 2010). Under this framework, the information that two (or more) sources (e.g. $\{Z_1, Z_2\}$) contain about a target variable Y is decomposed into three non-negative terms that contribute to the total information about Y in the system:

- *Unique information* $U_{\{Z_1, Z_2\}}$ that one source provides but the other does not.
- *Redundant information* R that is contained in both sources.

- *Synergistic information S* that is the joint product of both sources; it is only available when they are observed together.

To illustrate how these terms may manifest in speech, consider a conversation with a friend in a café; they greet you with “It’s absolutely freezing out there”. You can easily interpret the information carried across the lexical channel of this utterance (U_{lexical}). However, you can also infer that your interlocutor is in a cheery mood, even as they describe dreich weather that is typical of Scotland in March ($U_{\text{non-lexical}}$). People have been shown to be relatively adept at inferring affective information such as emotional state from non-lexical information regardless of the underlying lexical content (Livingstone and Russo, 2018). Once your interlocutor has sat down, you may ask them “Would you like a coffee?”. In English, rising pitch at the end of an utterance is a common international feature of questions (Ohala, 1983). The lexical content of this utterance already indicates that you are asking a question, so a rise in pitch at the end of the word “coffee” is redundant (R). However, you are impatient for a warm drink and instead ask more informally “You want coffee?”. The inference that you are seeking information is now a synergistic product of your pitch-rise and lexical content (S).

The PID framework provides us with a vocabulary for exploring the impact of both the lexical and non-lexical channels, as well as their interactions, in language comprehension.

1.3 Perceptual information value

Along with the type of processing that occurs, we argue that useful models of human language comprehension should also offer explanations about what information is valuable⁴ during comprehension. Predictive accounts of language processing already propose a formal valuation function; value is assigned to information in the upcoming signal based on the degree to which it deviates from expectations. This has an elegant mathematical definition known as conditional surprisal (Shannon, 1948): the negative log-probability of the t^{th} unit in a communicative signal u_t conditioned on its prior context $\mathbf{u}_{<t} = (u_1, \dots, u_{t-1})$:

$$I(u_t | \mathbf{u}_{<t}) = -\log_2 p(u_t | \mathbf{u}_{<t}). \quad (1.1)$$

⁴In this context, the term “value” is often used interchangeably with “salience” (Jaeger and Weatherholtz, 2016).

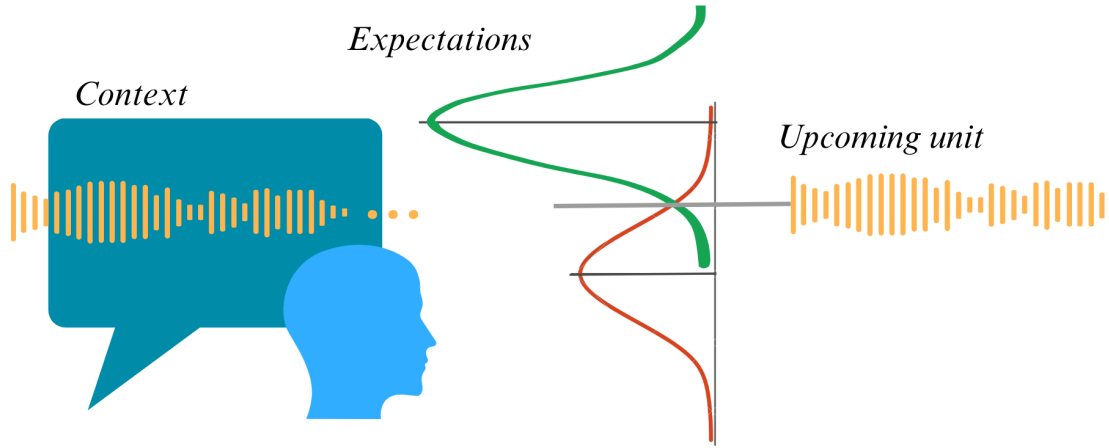


Figure 1.2: Illustration of how the same upcoming unit can obtain the same surprisal, even under different sets of expectations.

From this perspective, perfectly predictable signals (i.e., with zero surprisal) yield no information and have no value, while highly surprising signals are more informative and thus more valuable.

There are several drawbacks to applying the surprisal valuation function to the perception of spoken communication. Some of these are practical difficulties—computing the probability of a unit u_t as per Equation (1.1) requires that u_t belongs to a fixed-size vocabulary. This is not a problem when the unit of interest is, for example, an alphabetic character in English which can be drawn from a set of 27 units (including a whitespace character). However, moving beyond such well-defined vocabularies means that computing probabilities directly becomes intractable. Even computing surprisal for orthographic words requires some finagling (we discuss the details of this in Part I). Evaluating surprisal of more abstract units like conversational turns that lack any semblance of a fixed-size vocabulary requires careful consideration. Other drawbacks are more conceptual; as depicted in Figure 1.2, a particular linguistic unit can be equally likely, and thus obtain the same surprisal value, under quite different sets of expectations. As such, surprisal values don't provide information regarding *why* a particular upcoming signal may be surprising.

In this thesis, we present an alternative perspective to study the role of prediction in language comprehension. We propose attributing value to information based on the degree to which it affects our expectations about the upcoming communicative signal and term this value-assignment *Perceptual Information Value (PIV)*. As we move towards studying the comprehension of more natural forms of language-use that are

embedded in a wider discourse context, expectations are likely to be made at higher levels of abstraction, and thus cover a much broader range of potential continuations than predictions regarding the upcoming word in a written sentence (Christiansen and Chater, 2016). We suggest that understanding the *shape* and *source* of these expectations is an important component of understanding human language comprehension. Focusing on the expectations themselves also offers the potential to explain why a particular upcoming signal is surprising.

Investigating expectations at higher levels of abstraction than words is still uncharted territory and this ambition comes at a trade-off with formality. We use the lay term “shape” to refer to the characteristics of the distribution of possible continuations that could be predicted in a given linguistic context. For example, when two people meet on the street, their discourse usually follows a neat pattern: (turn 1: “*Hello*”, turn 2: “*Hello*”, turn 3: “*How are you?*”, ...). The shape of their expectations after the first turn is likely to be a sharply peaked distribution, reflecting the conventionalisation of greetings (Laver, 1981). However, after the third turn in this sequence, the shape of expectations will likely relax as the range of predictable trajectories expands. Though we note that a distribution can be characterized in many ways, this thesis focuses particularly on the variance of expectations. The two expectations depicted in Figure 1.2 have different shapes in this regard.

Assigning value based on how it affects the variance of expectations is closely related to the notion of entropy (Shannon, 1948). Entropy $H(X)$ of a discrete random variable X can be defined as its expected surprisal:

$$H(X) = \mathbb{E}(I(X)) \quad (1.2)$$

When this equation is applied to the conditional definition of surprisal from Equation (1.1), we can define conditional entropy, where U_t is a random variable representing the potential upcoming units u_t from a vocabulary U with distribution $p(\cdot|\mathbf{u}_{<t})$, as follows:

$$H(U_t) = \mathbb{E}_{u_t \sim p(\cdot|\mathbf{u}_{<t})}(I(u_t|\mathbf{u}_{<t})) \quad (1.3)$$

Conditional entropy represents the uncertainty associated with predicting an upcoming unit at time t , offering a measure of variance across expectations. Conditional entropy has already been applied in the study of language comprehension; Wilson and Carroll (1954) proposed that the value of a word should reflect the reduction in conditional entropy that results from its addition to the end of a sentence fragment. This definition has been examined by several studies that compare word-level reading

times to operationalizations of entropy reduction—these include word-level entropy assessments such as successor entropy (Aurnhammer and Frank, 2019) and conditional entropy (Pimentel et al., 2023), as well as entropy reductions over syntactic sentence parses (Hale, 2006; Linzen and Jaeger, 2014). We expect that the importance of variance across expectations will only increase as we study human comprehension at more abstract linguistic levels; however, we still lack methods for representing such expectations or defining a vocabulary of units for them. As such, we use the term “shape” in our general research goals to maintain flexibility beyond the mathematical formality of entropy. This enables us to explore these questions more broadly.

The PIV studies presented in this thesis are centred around the question: “How do the lexical and non-lexical channels contribute to the shape of expectations regarding upcoming spoken dialogue?”. By comparing the shape of expectations that arise when participants can condition their judgements on different contexts, we can lay the foundations for defining the “source” of expectations. We investigate different contexts by altering which communicative channels can be used to condition expectation. Doing so allows us to begin exploring the unique, redundant, and synergistic contributions of the lexical and non-lexical channels to the shape of expectations in spoken dialogue. These studies provide evidence that people can make predictions at the level of dialogue turns, a much more abstract event space than written words. In particular, they highlight that people can maintain a range of potential continuations as plausible and that both channels of context constrains expectations about the upcoming communicative signal, even if doing so doesn’t make expectations more accurate. As such, we argue that PIV should reflect how linguistic information alters the shape of expectations rather than whether it alters the accuracy of expectations.

1.4 Thesis overview & contributions

Having presented the broad research agenda of this thesis, we now outline the process undertaken to achieve it. The following six experimental chapters are organised into three parts: I. Lexical representations of dialogue, II. Perceptual value of the channels, and III. Multi-channel representations.

In Part I, we present studies of the lexical component of spoken dialogue. Beginning in this simplified version of the research domain allows us to establish whether people engage in prediction at the level of dialogue turns, and to understand the extent to which these expectations are reflected in existing machine representations of writ-

ten dialogue. In Part II, we capitalise on our findings regarding both human perception and language models to present a paradigm for quantifying the perceptual value of both channels of spoken dialogue by isolating their effects on human expectations of upcoming turns. Finally, Part III consists of a proposal for applying neural models for the same task of scoring dialogue turns.

1.4.1 Part I: Lexical component of dialogue

We begin characterizing the *shape* of expectations in communication by focusing exclusively on the predictions people make about lexical content; we consider both written dialogue and the information unique to the lexical channel of spoken dialogue. Expectations about lexical information have been studied at word- and sentence- levels in monological data but there is no existing task to quantify comprehension behaviour in dialogues. We therefore propose a behavioural experimental paradigm to quantify expectations at the level of dialogue turns through acceptability ratings. This lays the foundations we use to analyze expectations in the full multi-channel system of spoken dialogue.

Using this paradigm, we ask a number of sub-questions about the shape of expectations:

RQ1.1: “How accurate are human expectations of the lexical content of dialogue?”

Through our behavioural study, we show that people can make predictions about upcoming dialogue turns in both written dialogue and the transcripts of spoken dialogue based on fixed amounts of dialogue history. We find people can make accurate judgements in both domains, though they do so less often when stimuli are constructed from spoken dialogue transcripts. However, we consistently find multiple candidate turns to be judged as plausible continuations. We term the variation in plausible continuations as *optionality*. The studies presented in Part I leave open the question of to what degree optionality is affected by a lack of context; however, we argue that optionality is not simply a product of insufficient contextualization but rather a feature of human expectations. We return to the concept of optionality and its sensitivity to different features of context in Part II.

RQ1.2: “To what extent do neural estimates of surprisal align with dialogue acceptability judgements?” We also test the degree to which machine representations of lexical content and their associated estimates of surprisal align with human dialogue

perception. Such alignment has previously been quantified by a language model’s capacity to predict psychometric behaviour; however, studies in this area have been limited to examining alignment at the level of individual words. To study alignment at the utterance-level where computing surprisal in its classical form is intractable, we explore a number of methods for estimating utterance predictability by aggregating word-level estimates. Our results show that widely-used machine representations align with human expectations, but only to a certain degree. Through architecture ablation studies and qualitative analysis of human and machine errors, we find that this alignment is strongly contingent on the model being capable of complex utterance-level aggregation and fine-grained contextualisation. This provides insights into how prediction may function in humans, however our findings also point towards fundamental differences between human perception of dialogue and the representations that can be learned through self-supervised objectives. In particular, SSL representations that can accurately discriminate true upcoming turns do not display human-like optionality in how they score responses.

Much of the content of these chapters draws from the following publications:

Wallbridge, S., Bell P., and Lai, C. (2022) *Investigating perception of spoken dialogue acceptability through surprisal*. In Proceedings of Inter-speech, (4506–4510)

Wallbridge, S., Bell P., and Lai, C. (2023) *Do dialogue representations align with perception? An empirical study*. Long paper in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (2696–2713)

1.4.2 Part II: Perceptual value of the channels

The lexical-only studies presented in Part I demonstrate that people can leverage context to discriminate between different upcoming lexical turns in dialogue, validating the turn discrimination paradigm as a means to study the effect of information on expectations about the upcoming communicative signal. In Part II, we turn our focus to both the *shape* and *source* of expectations in the full speech signal of dialogue. For this, we propose an extension of the turn discrimination task. Rather than rating individual turns, participants are asked to rate the plausibility of sets of potential upcoming turns based on either lexical context or a combination of lexical and non-lexical context. This paradigm offers a method for comparing the unique and synergistic in-

formation between channels by quantifying their relative contributions to constraining expectations about the prosodic realisation and lexical content of the upcoming turn in a dialogue.

The research questions that we address in Part II are as follows:

RQ2.1: “How do the different channels of speech affect expectations about the channels in upcoming spoken dialogue?” We present the *turn acceptability paradigm* to explore the Perceptual Information Value of the lexical and non-lexical channels in spoken dialogue. We begin in a highly constrained set-up of the paradigm where stimuli consist of sets of lexically-equivalent responses; our goal is to isolate the effects of information in both the lexical and non-lexical channels on expectations of prosodic realisations. The results highlight the value of non-lexical channel, of both the response and context, in constraining and modifying predictions; people become more accurate at discriminating the true turn from negative turns when they are provided with non-lexical context compared to only lexical context. Additional lexical context has little effect on people’s expectations, indicating that the observed optionality is not a consequence of lack of context but rather an important property of human expectations. We then extend the paradigm to stimuli with variable lexical content. Based on the alignment between language models and human judgements found in Part I, we use neural models of dialogue to sample turns with similarly plausible lexical content. This set-up highlights the synergistic interactions between lexical and non-lexical information, demonstrating that expectations are a joint product of information in both channels. We find that the effect of non-lexical information on participant accuracy is dependent on the lexical content of the response; although non-lexical context is helpful for discriminating the true response when the underlying lexical content is very ambiguous, when discrimination is possible from the lexical content alone, we find that non-lexical context can hinder participant accuracy.

RQ2.2: “How should the Perceptual Information Value of channels be quantified?” Our behavioural results indicate that discriminative performance measured by accuracy only reflects a certain impact of contextual information. Instead, we propose quantifying the PIV of channels through measures such as ordinal and permutation entropy that capture changes to participant agreement and certainty. These metrics offer a much more nuanced picture of how expectations may be constructed in spoken dialogue. We find that, regardless of its impact on the accuracy of participants,

including non-lexical context produces consistent effects on response scores across participants—entropy-based analysis shows that people agree more about how to rate turn plausibility, even if they may be incorrect. Our results highlight the importance of understanding not only the source of expectations (i.e., how each channel affects participant expectations) but also accounting for their shape (i.e., expectations display optionality as multiple upcoming communicative trajectories can be acceptable continuations).

The experiments described in these chapters are published in the following papers:

Wallbridge, S., Bell P., and Lai, C. (2021) *It's not what you said, it's how you said it: discriminative perception of speech as a multi-channel communication system.* In Proceedings of Interspeech, (2386–2390)

Wallbridge, S., Bell P., and Lai, C. (2023) *Quantifying the perceptual value of lexical and non-lexical channels in speech.* In Proceedings of Interspeech (2708–2712)

1.4.3 Part III: Multi-channel representations

Our results in Part II demonstrate that people make use of both lexical and non-lexical information to constrain their expectations regarding the upcoming speech signal and provide evidence that the channels function both independently and in conjunction with each other in this regard. In the final part of this thesis, we present a set of preliminary experiments studying whether the predictions of neural models are altered in similar ways to human expectations by the inclusion of different channels in their input. We focus on presenting an experimental design to answer this question that presents the prosodic and lexical information as completely isolated input channels. Doing so offers an alternative perspective on the value of lexical and non-lexical information. In this section, we pose the following research questions:

RQ3.1: “At what timescale do the acoustic correlates of prosody exhibit predictable structure?” We begin Part III with a study of information unique to the non-lexical channel. The behavioural experiment presented in Chapter 5 demonstrates that non-lexical information contributes to our expectations of the upcoming speech signal synergistically with lexical content and that variation in the non-lexical realisation of upcoming speech is a valuable cue for discriminating between potential upcoming turns. However, this leaves open the question of whether these effects are the result of the

unique and synergistic contributions of information in the lexical and non-lexical channels. As such, we explore the predictive power of prosody isolated from lexical content using neural models: does prosodic context enable prosodic prediction? By applying self-supervised learning to acoustic correlates of prosody, we show that prosodic features display predictable structure independently of lexical content across multiple timescales. The value of resulting representations for different downstream tasks depends on their temporal resolution. Along with evidence of non-lexical prosodic structure across temporal granularities, a major contribution of this chapter is a self-supervised method for generating representations of prosodic realisation that are, by design, independent of lexical content.

RQ3.2: *“To what extent are predictive models capable of leveraging non-lexical information for turn discrimination?”* Finally, we present a set of proof-of-concept experiments asking whether predictive neural models behave similarly to human participants for the turn-discrimination task. Mirroring the behavioural study presented in Chapter 5, we present a training set-up for models to discriminate between different prosodic realisations of upcoming turns. We use the representations of lexical and prosodic information that we validated in Chapter 4 and Chapter 7, respectively, and test model variants based on training data and input representations which contain different degrees of context, prosodic representations, and cues to lexical content of responses. None of these set-ups come close to human performance, highlighting the difficulty of upcoming turn discrimination and pointing towards fruitful directions for future work in this area. We believe that these future directions could allow us to further probe both human expectations as well as representations of speech in spoken dialogue—for example, what information is most useful for approximating human-like expectations in natural spoken communication? And do models encode it?

The experiments described in Chapter 7 are submitted as a pre-print to be available on Arxiv:

Wallbridge, S.*, Minixhofer, C.*, Bell P., and Lai, C. (2024) *Systematicity in prosody beyond lexical content: a study of self-supervised learning*. (* for joint authorship)

1.5 Contributions

This thesis makes a number of contributions towards advancing the study of human language processing, as well as the degree to which modern neural language models align with this type of perceptual behaviour. We highlight the following primary contributions:

- We present a number of novel behavioural paradigms based on turn acceptability judgements for studying how people make predictions regarding upcoming spoken dialogue. Compared to classical psycholinguistic stimuli which are heavily controlled and often acted, we create stimuli from existing dialogues to study predictions in realistic language-use. Integrating Large Language Models (LLMs) into the stimuli construction process allows us to study a broader range of communicative phenomena that occur in true spoken dialogue.
- Evidence that people can make predictions about upcoming spoken dialogue turns. Though numerous studies of how people read sentences support predictive theories of language comprehension, the behavioural studies in this thesis provide strong evidence that predictive processing also plays a role in the comprehension of a far more realistic form of language.
- Rather than considering how the upcoming signal is related to expectations (e.g., its surprisal), we propose studying how expectations are generated in the first place. We propose Perceptual Information Value as an alternative perspective to surprisal, where value is attributed to information based on how it alters expectations. Using metrics like entropy reduction and agreement, we argue that information in the communicative signal can be a valuable component of the comprehension process regardless of its effect on predictive accuracy.
- By framing speech as a multichannel communication system, we demonstrate that the lexical and non-lexical channels of speech function both in isolation and in conjunction with each other to constrain people's expectations about the upcoming turn in a spoken dialogue. We show that the effect of non-lexical information on participants' discrimination ability is dependent on the corresponding lexical channel; however, when quantified by changes to participant agreement, the non-lexical channel produces consistent effects on participants scoring behaviour.

- Our studies of expectations in spoken dialogue highlight the inherent variability (a feature which we define as *optionality*) in human predictions about the upcoming linguistic signal. We present evidence for optionality across both the lexical and non-lexical channels of speech and argue that understanding optionality will become increasingly important as we expand studies of comprehension behaviors to more realistic forms of language-use like spoken dialogue.
- We present evidence for alignment between human expectations about upcoming dialogue turns and the predictions of automatic neural networks; however, we also demonstrate the sensitivity of this relationship to features of the model architecture and training set-up.

Chapter 2

Background

In this chapter, we review literature related to the use of prosody in human communication and prediction in language comprehension. We also review neural approaches to automatic language processing and connections to human language processing that inspired and facilitated our work.

2.1 Prosody

2.1.1 Definitions

The term *prosody* refers to the systematic variation of timing, loudness, and pitch of speech. In simpler terms, it can be described as “*how*” an utterance is spoken beyond its phonetic segmental representation. Though it might seem straightforward, it is a concept characterized by nuance. On the one hand, prosody contributes to a wide spectrum of communicative functions—a topic we delve into shortly—and plays important and evolving roles in the landscape of language acquisition (Gleitman and Wanner, 1982; Pate and Goldwater, 2013; Kakouros et al., 2018); it is also intricately woven into our identities as speakers, both on an individual level (Meizel, 2020, Chapter 8 Lost Voices) and as members of various social or cultural groups. On the other hand, spoken language cannot exist without the suprasegmental dimensions of unit duration, frequency, and amplitude that define prosody (Bolinger, 1964). As such, defining prosody and which of its components are perceptually important has been a long-standing goal of linguistics (Cutler et al., 1997; Cutler and McQueen, 2014). In this section, we present an overview of important concepts that allow us to talk about prosody in the context of this thesis. It’s important to acknowledge that this thesis fo-

cuses exclusively on prosody in the English language; however, different languages use the components of prosody for different purposes (Vander Klok et al., 2018). For example, in tonal languages such as Thai, pitch changes serve to distinguish between different lexical units while in English, pitch changes generally do not affect the linguistic identity of phonetic sequences (Beckman and Pierrehumbert, 1986); in languages like Greek or Russian with less constrained word order, statements and polar questions are often distinguished through the placement of prominence in a sentence while speakers of Italian and English generally rely on syntactic structures to make this distinction and thus can be more flexible with where prominence is placed in a sentence (Ladd, 2008). These variations between languages underscore yet another facet of interplay between the lexical and non-lexical components of speech, and how they share the burden of imparting expressiveness and structure to language.

Given the broad set of impacts introduced already, it is perhaps understandable that the term “prosody” is interpreted and utilized in diverse ways by researchers across fields of speech research. Definitions of prosody tend to fall between two extremes—in terms of *form*, *function*, and the relationship between them (Wagner et al., 2005). Prosodic form can be expressed as the perceived variation in suprasegmental aspects of speech. In concrete terms, these constitute the timing, amplitude, and frequency spectrum of the acoustic signal. Variations to these features occur across a range of time scales (i.e., at the level of syllables, intonational phrases, and full utterances) and are generally agreed to form a hierarchical structure (Selkirk, 1984; Beckman, 1986). Prosodic function, on the other hand, refers to the communicative purpose of a particular suprasegmental realisation, including how prosody is used to mediate syntactic and conversational structure and convey affective and rhetorical information. Below, we review the most relevant aspects of both form and function.

2.1.2 Form of prosody

Representations of prosodic form range from objective schemes based on acoustic realisation and subjective annotations of perceptual effects.

2.1.2.1 Acoustic correlates

The most commonly studied acoustic correlates of prosody—F0, segment and pause duration, intensity, and spectral tilt—are the respective counterparts to the perceptual aspects of pitch, rhythm, loudness, and voice quality. These features can be extracted

algorithmically from the speech waveform and are often considered as an objective representation of prosody that is independent of the lexical content of speech. Each one is defined in turn below.

Pitch. The perceived fundamental frequency (F0) at which the vocal folds vibrate is referred to as pitch (Talkin and Kleijn, 1995). Numerous pitch tracking algorithms have been proposed¹ to manage complexities that arise from noise, artefacts, and harmonic features in speech recordings. However, perhaps the most fundamental challenge of pitch tracking is the lack of clear ground truth. Generally, pitch tracking algorithms are compared to F0 curves obtained through electroglottographic (EGG) recordings. By applying a low voltage across participants’ vocal apparatus, fundamental frequency can be measured relatively directly from vibrations of the vocal folds. Although these recordings can contain errors, and there is still a long chain of operations that separate the sound produced by the vocal folds and the resulting speech that we hear, EGGs are widely accepted as gold standard F0 curves (Pirker et al., 2011; Drugman et al., 2018). This also helps explain why the terms “pitch” (the perceptual quality) and “F0” (the inverse of the smallest period of a quasi-periodic signal) are often used interchangeably in speech processing research (Talkin and Kleijn, 1995). We provide further discussion of specific methods to represent F0 in Chapter 7 when we propose a structural representation of the acoustic correlates of prosody.

Rhythm. This term encompasses the duration of various speech units such as syllables, words, and phrases. The duration of pauses between units is also an informative characteristic of speech; for example, conversational partners rely heavily on pauses to coordinate who will speak at a given moment (Duncan, 1972). Unit durations are typically extracted automatically using tools that align an orthographic or phonetic transcription with the corresponding speech recording; a popular alignment tool is Montreal Forced Aligner (McAuliffe et al., 2017). Though automatic alignment is often an errorful process, humans also disagree on how transcripts and recordings align, particularly in spontaneous speech where the boundaries between words and phones are muddled by co-articulatory effects like reduction (Johnson, 2004; Rodd et al., 2021).

Loudness. The perceived loudness of speech is often approximated as the intensity of the sound signal—the logarithm of its short-term average energy (Toledano et al.,

¹Exemplified by the self-effacing title of the YAAPT pitch tracker: “yet another algorithm for pitch tracking” (Kasi, 2002).

2003). Energy can be easily extracted from either the raw audio as the amplitude of the waveform or from a spectrogram as the total magnitudes of frequencies present in a frame.

Voice quality. This concept refers to the different effects that arise from modifying the tension over the vocal folds when producing speech. For example, speakers can produce vocal fry, breathy voice, whispering, and shouting through changes to their vocal effort (Eckert and Laver, 1994). This is often quantified at the spectral level, by characterizing the distribution of energy across the frequency spectrum; for example, spectral tilt reflects attenuation patterns of harmonic energy.

These acoustic features are practically useful; for example, speech synthesis models integrate them explicitly to better control the resulting spoken realisation of text (Ren et al., 2021). However, acoustic representations of prosody do not always reflect what people find to be perceptually salient. Deviation between what is perceived and the true realisation of speech is evidenced by phenomena like the McGurk and Ganong effects (McGurk and MacDonald, 1976; Ganong, 1980). Incongruencies between auditory and visual cues, such as hearing “ba” while observing lip movements for “ga”, can lead to fused percepts (e.g., “da”), known as the McGurk effect. The Ganong effect shows that ambiguous sound segments (for example, a sound between /d/ and /t/) are perceived as different phonemes depending on the lexical context surrounding them—participants consistently resolve these ambiguities to perceive real words rather than non-words (e.g. “task” is heard, rather than “dask”). Prosodic perception is similarly affected by context; for example, Buxó-Lugo and Watson (2016) show the perceived strength of a boundary marked by prosody depends on the syntax of the underlying lexical content.

2.1.2.2 Prosodic annotation

An alternative strategy for representing prosody consists of perceptual annotation schemes. Intonational phonology has a long history of categorizing and segmenting the prosodic realisation of a phonetic sequence into discrete vocabularies of contours and units that people perceive as important (Pierrehumbert, 1980; Silverman et al., 1992; Ladd, 2000; Cole, 2015). These generally aim to capture hierarchically organised prosodic events across syllables and words into phrases and utterances.

At least for English, annotation schemes have been developed based on two perceptual effects: 1. *prominence*, the demarcation of salience across units, and 2. *phrasing*, how sequential units are grouped. Both effects are signalled by changes to the acoustic correlates described above—pitch, rhythm, loudness, and voice quality—in conjunction with the underlying phonetic sequence. A classic annotation scheme is the TOnes and Break Indices schema (ToBI) (Pierrehumbert and Hirschberg, 1990; Silverman et al., 1992). ToBI represents the pitch contour of an utterance as a sequence of tone markings (H* or L* for high and low tones) across words, and break indices (on a scale from [0-4]) to mark the boundaries between them. Importantly, it reflects how prosody contributes to hierarchical structure in speech. Though it was developed for Standard American English, ToBI has since been adapted to other varieties of English (e.g., GlaToBI for Glaswegian English (Mayo et al., 1997)) and other languages such as German (Grice et al., 1996) and Japanese (Venditti, 1997). However, labelling tones and boundaries is an expensive process that requires expert annotators. Other frameworks have been proposed to generate annotations less expensively. For example, SLAM (stylisation and labelling of speech melody) proposed by Obin et al. (2014) represents units of speech with five values automatically derived from an F0 estimate of the recording (initial F0, final F0, time of the main saliency position, and mean F0). More recently, the Rapid Prosody Transcription paradigm was proposed by Cole et al. (2017) to collect manual annotations on a much wider scale using a simplified binary classification for prosodic phenomenon at the word level.

Though these annotation schemes capture perceptual aspects of prosody, mapping specific prosodic cues to particular functions or downstream information is complicated by variation in both production and perception of prosody between speakers, contexts, and languages Cole (2015). For instance, on the perception side, Roy et al. (2017) showed that annotators rely on different acoustic cues to label prosodic events (through both ToBI and RPT paradigms), even among monolingual speakers of American English with no hearing or reading deficits. On the production side, despite a widespread association between utterances used for contradiction and specific prosodic contours, Lai (2012) found that speakers do not produce these contours consistently.

2.1.3 Functions of prosody

In this thesis, we are interested in how prosodic cues shape listeners' expectations about the upcoming speech signal, and how it may do so both in conjunction with

and in isolation from the corresponding lexical content of speech. These questions are motivated by a large body of evidence linking prosody to a wide variety of communicative functions in English. These include both affective and augmentative purposes (Taylor, 2009). Affective functions generally refer to the transmission of information which may be independent of the lexical content of speech, such as emotion, attitude, or other aspects of a speaker's mental state. Augmentative prosody, on the other hand, refers to the use of prosodic effects for facilitating language comprehension through more direct interactions with lexical information. We introduce some of the most relevant links between prosodic realisation and language comprehension here.

Prosody in conjunction with lexical content. Classical links between language comprehension and prosodic realisation include lexical recognition, syntactic disambiguation, and processing discourse structure (Pisoni, 1978; Cutler et al., 1997). In all of these cases, prosody contributes to comprehension through its *interaction* with lexical content. For example, the prosodic realisation of words has been linked to both word recognition and segmentation: English listeners leverage syllable stress to resolve ambiguity (e.g., in noisy conditions listeners are more likely to interpret the initial consonant in the following pair to produce the real (underlined) word in each pair (*DI*gress, *TI*gress) vs (*di*GRESS, *ti*GRESS)) (Cutler, 1987) while prosodic features can also be used to signal word boundaries (Turk, 2010).

Prosody is also a tool for highlighting salient information in wider discourse (Bolinger, 1978). Question-answer adjacency pairs are often used to exemplify this. Consider the following sentence:

I_1 was₂ outside₃ .

This lexical string is an appropriate response to a number of questions: e.g., “Who was outside?”, “Are you outside now?”, “Where were you?”. However, the placement of prominence in the spoken realisation, on word 1, 2, or 3, would depend on the preceding question itself. In other words, acceptable prosodic structures are dependent on the pragmatic context in which they occur. Prosodic prominence is known to affect the efficiency of spoken language processing—for example, listeners react more quickly to both mispronunciations and target phonemes when they occur as an accented word in a sentence compared to an unaccented word (Cutler and Foss, 1977; Cole and Jakimik, 1980). There is also evidence that prosodically-realised prominence affects listener predictions about upcoming speech: Cutler (1976) finds that listeners make

predictions about sentence-level prominence based on the intonation of the preceding context.

However, much of the evidence presented in the preceding paragraph comes from studies involving elicited or rehearsed speech which is significantly different to the speech investigated in this thesis. Speech used during conversation differs from read speech in acoustic characteristics such as decreased prosodic range, more frequent and longer pauses, and syllabic reduction (Howell and Kadi-Hanifi, 1991; Blaauw, 1994). In related work, we have found that listeners are sensitive to these differences: they can distinguish between read and spontaneous speech from single utterances with identical lexical content (Adigwe et al., 2024). Neat structures like the one depicted in Example 2.1.3 are not actually very common in natural spoken dialogue as ambiguity can often be resolved from contextual factors (Tzeng et al., 2019). Though the effect of prosody on the perception of natural spoken dialogue has been far less studied, there is evidence that prosody still has important effects on the comprehension of spoken dialogue. For example, Hirschberg and Nakatani (1996) find that inter-annotator reliability is far higher for segmenting both read and spontaneous speech into discourse segments when annotators have access to both speech and text, compared to annotations produced using only text. Similarly, Mehta and Cutler (1988) find that participants are faster at detecting target phonemes that occurred in accented words than in unaccented words of spontaneous speech, but not read speech.

One contribution of prosody to dialogue processing that has been comparatively well-studied is the management of turn-taking. The average adult has spent hundreds of hours engaged in the shared task of determining *who will speak when* with various conversational partners. Yet, the question of how this structure unfolds so quickly has been an important topic in psycholinguistics for decades (Duncan, 1972). Though conversational partners can interleave their turns with extremely short transitions², our capacity to plan and produce speech is comparatively slow. To reconcile this puzzle, Sacks et al. (1978) theorized that listeners make predictions about appropriate points to begin their own turn while their partner is still speaking. This predictive process is well-exemplified by Holler and Kendrick (2015) who show that, in multi-party conversations, the unaddressed listener often shifts their gaze from the current speaker to the next before the current speaker has finished. Though the underlying cues that enable such predictions are debated, non-lexical features of speech including pitch,

²A typical gap between turns ranging from 500ms to -100 ms (where a negative gap indicates an overlap) (Levinson and Torreira, 2015)

intensity, and modifications to duration are widely accepted as important components (Duncan, 1972; Schaffer, 1983; Gravano and Hirschberg, 2011; Skantze, 2020). Ford and Thompson (1996) highlight the importance of prosody in contributing to the pragmatic completeness of an utterance in the context of the conversation, which signals an appropriate turn change position. The use of the non-lexical cues in turn-taking is further exemplified by infants as young as 6 months old engaging in proto-conversations before they have acquired a linguistic vocabulary: the timing of turn-taking in these interactions approximates the timing of adult spoken conversation (Levinson, 2016).

Prosody in isolation of lexical content. Recall the example in the Chapter 1 of a conversation in a cafe where you could deduce that your conversational partner is cheerful, despite the fact that they are describing the Scottish winter weather. Prosody is also known to contribute to a listener’s interpretation of speech in addition to its lexical content. Common evidence includes the transmission of affective information like emotion or other paralinguistic factors such as the relationship between the speaker and listeners or a speaker’s uncertainty regarding a topic (Grandjean et al., 2006).

The link between emotion and prosodic realisation independent of lexical information has been evidenced by both automatic modelling and behavioural studies. For example, Cao et al. (2014) show that acted emotion can be classified from low-level acoustic features derived from intensity, loudness and F0, as well as from the more abstract representations of ToBI labels. The RAVDESS dataset collected by Livingstone and Russo (2018) contains over 1000 utterances spoken by a range of speakers across eight emotion categories (Calm, Happy, Sad, Angry, Fear, Surprise, Disgust, Neutral). Importantly, all productions are one of two neutral phrases (“Kids are talking by the door”, “Dogs are sitting by the door”), meaning that the emotional expression and lexical content are completely disjoint. Even in this difficult setting where lexical content bears no cues to the acted emotion, raters who were asked to classify the emotion obtained relatively high accuracy (62% of samples in the audio-only condition were correctly classified, averaged across all emotions and intensities). We address the relationship between prosodic realisation and emotion in Chapter 7.

Other studies have also demonstrated that the prosodic realisation of de-lexicalised speech (i.e., where lexical content has been obscured) also carries information that may be redundant with lexical content. For example, listeners can still identify major syntactic boundaries from natural speech where phonological information has been suppressed (Kreiman, 1982), or through hummed speech when such information is

not present (Collier and Hart, 1975).

Taken as a whole, the studies presented above indicate that speech comprehension is facilitated by the prosodic realisation of speech, both through the efficient transmission of information beyond lexical content as well as through structure that is coherent with it. However, these works paint a complex picture: conclusions from these studies regarding the value of prosody across spoken language comprehension tasks are nuanced and results are not always robust. As we have already alluded to in our discussion, an important potential cause could be the difficulty of studying the comprehension of truly natural language-use. Stimuli used in psycholinguistic studies often consist of carefully controlled stimuli presented in unnatural conditions without a wider linguistic context, physical noise, or other draws on a listeners attention that characterizes realistic communicative settings (e.g., Pols and Plomp, 1986; Nielsen et al., 2023). We discuss more existing evidence pertaining directly to the effect of prosody on comprehension in dialogue in Chapter 5.

2.2 Machine representations of language

In this thesis, we use machine representations of language from artificial neural networks both as a tool for designing behavioural experiments and for investigating the underlying processing that may occur in people during language comprehension.

Neural network basics. At their core, neural networks are a general-purpose mechanism for approximating the underlying function that generated a given set of data. The resulting statistical summaries can be powerful tools for generalising beyond the data used to train the network. Often, training data is in the form of (x, y) where x are the input data, and y are corresponding labels. Traditionally, neural networks have been broadly classified as discriminative or generative based on which facet of their data they are used to summarize. Discriminative networks learn decision boundaries between label classes by representing the conditional distribution of a label given the input data $P(y|x)$. Generative models, on the other hand, learn to represent the probability of x and y occurring together (the joint distribution $P(x, y)$) or simply the probability of the data itself (the marginal distribution $P(x)$). Unlike classical statistical techniques like logistic regression or Naive Bayes models which solve the same distributions an-

alytically, neural networks are trained through iterative optimization processes such as stochastic gradient descent (SGD). Though the basic building blocks have existed for decades (Rumelhart et al., 1986; Rosenblatt, 1958), neural networks have become pervasive throughout Natural Language Processing (NLP), especially with the introduction of Transformer architectures (Vaswani et al., 2017) which use self-attention to learn complex representations of their marginal distribution of their training data.

We can take a neural network to be composed of four core components: a model architecture, a learning objective, an optimization strategy, and a set of training data. This thesis focuses far more on the learning objectives and underlying training data. We discuss the foundations of learning objectives below, as well as how these mechanisms are applied to text and speech data.

2.2.1 Learning mechanisms

The process of iterative “learning” that characterizes neural networks requires a learning objective to optimize. Traditionally, objectives have been categorized as supervised and unsupervised learning schemes based on whether they use labels y : *supervised* learning involves learning some relationship between the input data x and the corresponding labels y (e.g., solving a classification problem) while *unsupervised* learning involves learning some structure from unlabelled data x (e.g., performing autoencoding (Kramer, 1991; Hinton and Zemel, 1993)). More recently, *self-supervised* learning (SSL) has emerged as a hugely popular variant that can be seen as a combination of the two. Rather than requiring external labels, learning objectives estimate the marginal distribution $P(x)$ by exploiting the input data as a training signal. Many self-supervised networks learn to reconstruct corrupted or obscured portions of the input data by learning structures within the data. These networks produce generic representations of their input that are useful for a range of downstream tasks (Bengio et al., 2013; Milde and Biemann, 2018). An important advantage of this strategy is that the networks can be fine-tuned to specific downstream tasks with a much smaller amount of labelled data (Asano et al., 2019). This mechanism was first proposed in the early 90s for image recognition tasks (Becker and Hinton, 1992; Bromley et al., 1993); however, its popularity in NLP has exploded much more recently in the form of pre-trained representations; by removing a need for explicit labels, self-supervised mechanisms can produce generally useful presentations by pre-training on massive amounts of unlabelled text (Mikolov et al., 2013).

The content of SSL-generated representations is driven by the corruption strategy used during training. In the context of this thesis, we are interested in encoding input with a temporal axis; to learn representations of input sequences, we can distinguish masked and auto-regressive corruption strategies. Masking involves obscuring portions of a particular training sample and tasking the model to reconstruct them (Vincent et al., 2008; Devlin et al., 2019); auto-regression is a specific instantiation of this paradigm where masking is constrained to the right-most portion of the training sample. In other words, masking the “future”, for time-series signals (Uria et al., 2016; Radford et al., 2019). This is particularly relevant in cognitive science as auto-regressive models can represent the predictive process that we’ve outlined in Section 2.3. Below, we discuss several SSL representations of both text and speech that are relevant throughout the coming experimental chapters.

2.2.2 Representing text

Primarily driven by the availability of vast swathes of textual data on the web, pre-trained representations of text from SSL language models³ have become important components in both research of language as well as user-facing products. To process textual data, input text is first split into a string of tokens drawn from a finite vocabulary. Self-supervised models are then tasked with learning to reconstruct sequences of tokens that have been corrupted. Tokens form the base units of language models and constructing the token vocabulary of a model is an important design decision; for example, a vocabulary could consist of alphabetic characters or predefined sets of words. Common vocabularies lie between the two where tokens consist of word segments known as subword tokens⁴ (Sennrich et al., 2016).

BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019) and **GPT2** (*Generative Pre-training Transformer*) (Radford et al., 2019) are popular variants of masked and auto-regressive modelling strategies, respectively. Both models consist of a Transformer architecture (Vaswani et al., 2017) which enables them to represent the words as a function of their context. For example, in the sentence “I *left* my book on the *left* side of the table.”, each occurrence of the word *left* would obtain a different representation. These models differ primarily in their learn-

³Though they are called language models, these almost always refer to models of *text*.

⁴Significant efforts have been dedicated to constructing optimal vocabularies for language models. Rust et al. (2022) provides an interesting perspective on the trade-off in vocabulary design between vocabulary size and the resulting receptive field of language models.

ing objectives. BERT is trained with two objectives: masked language modelling and next-sentence prediction. During masked language modelling, BERT receives input sequences where tokens have been randomly masked and is then tasked with reconstructing the original string. Because the model has access to the entire masked input sequence (i.e., tokens both before and after masked portions), its predictions can be conditioned on both past and future tokens; these are referred to as bi-directional. The next-sentence prediction task aims to capture even longer bi-directional dependencies by predicting whether a pair of sentences is consecutive or not. GPT2, on the other hand, is trained to reconstruct only the last token in the string sequence. By learning to predict the upcoming tokens from only past context, auto-regressive models are more similar to the incremental processing that humans engage in when they encounter language. We make use of both of these representations in Part I.

2.2.3 Representing speech

There is a long history of extracting structure from speech using automatic methods: for example, Levinson (1985) details the use of Hidden Markov Models and small neural networks to learn mappings between acoustic features and phoneme or word classes. However, speech differs significantly from text as an input signal for neural models. In particular, it is much more obvious how electronic text can be segmented into sequences made from a finite set of discrete tokens. Every instance of the word “speech” in this thesis will occur with the same spelling; however, if I were to read a chapter aloud, each occurrence would be produced with a different duration, intonation contour, and intensity. As such, producing a vocabulary of tokens for speech, similar to the vocabularies in language models of text, is much more difficult without pre-existing labels such as phonetic alignments. Instead, input to unsupervised models of speech generally consists of frames in a spectrogram or segments in raw audio that represent a fixed-length temporal window. The length of this window affects the type of structures that a model will reflect, and thus what kind of downstream tasks the resulting representations can be useful for.

Early applications of unsupervised learning to speech include the use of self-organizing map networks and vector quantization learning mechanisms over short windows of speech; these were used for clustering phonetic features and compressing speech recordings (Linde et al., 1980; Kohonen, 1982, 1988). Modern self-supervised mechanisms share similar general outlines to these works but, as is the case with SSL

models of text, involve learning from much larger datasets (e.g., Schneider et al., 2019; Baevski et al., 2020; Hsu et al., 2021; Zeghidour et al., 2022). We describe a few in detail below.

Unspeech (Milde and Biemann, 2018) was one of the first direct applications of text-like self-supervised pretraining to speech. In this model, windows of speech (0.5s) are encoded such that representations of speech segments that share similar contexts are close in embedding space. Milde and Biemann (2018) defined context as “factors of variance” that tend to change across utterances such as speaker identity, environment characteristics, and recording set-up. These are captured through discriminative training where pairs of segments are either adjacent in a speech recording or randomly sampled. **Wav2vec** takes a similar approach to the GPT2 model of text in that it is trained to make predictions about future units (Schneider et al., 2019). Rather than predicting from a vocabulary of tokens, the model is trained to distinguish an embedding of a speech sample z_{i+k} that occurs k steps in the future from timestep i from distractor embeddings. The embedding speech segments in this model span approximately 30ms. Another popular variant of pretrained speech representation is **HuBERT** (Hsu et al., 2021). To apply BERT-like pre-training where masked units of speech are reconstructed, the authors propose generating a fixed-size vocabulary of units prior to training using an offline clustering step. Each speech unit in HuBERT corresponds to approximately 25ms of audio. Though **Wav2vec2.0** (Baevski et al., 2020) shares a very similar name with wav2vec, this representation also uses a BERT-like masking strategy. Rather than predicting a learned codebook like HuBERT representations, wav2vec2.0 is trained to discriminate between true and false latent representations of masked segments. Windows of the raw waveform are first encoded into discrete units using a convolutional neural network followed by product quantization. A powerful transformer is then used to produce more contextualised latent representations.

Though UnSpeech was also evaluated with other tasks, both wav2vec and HuBERT were originally presented as techniques to improve automatic speech recognition (ASR). The fine-grained time scale at which they encode speech makes them well-suited to this task and both models have shown impressive capabilities for encoding phonetic information. However, it’s unclear what temporal granularity is most useful for other tasks which may involve longer-term structures conveyed through prosody. Much recent work has been dedicated to uncovering what other information pre-trained representations like HuBERT and wav2vec2.0 encode (e.g., the SUPERB benchmark includes tasks related to word recognition along with speaker identifica-

tion and paralinguistic information such as emotion recognition; Yang et al., 2021). Another approach to understanding the contents of SSL-based speech representations involves attempting to disentangle components of the speech signal such as “content”, “style”, and speaker attributes. Such works often use carefully tuned information bottlenecks or quantization to encourage representations to only encode information about a particular component (Qian et al., 2020; Williams, 2022; Qian et al., 2022). Disentangled representations have proved useful for constrained downstream applications like speaker verification (Williams and King, 2019) or speech recognition (Hsu et al., 2017), however they also raise questions about whether disentanglement of certain factors is achievable at all. Locatello et al. (2019) discusses the challenges of disentangled representations more generally, pointing out that real-world data often involves factors of variation that are inherently interdependent and that attempts to separate them can lead to the loss of nuanced information. These challenges are certainly applicable to speech research, where relationships between factors like content, speaker identity, prosody, and noise are similarly complex and interdependent. We discuss works that investigate whether prosodic cues are encoded in pretrained representations of speech in Part III, motivating our investigation into the effect of temporal granularity on representations of prosody presented in Chapter 7.

2.3 Predictive theories of language perception

As discussed in Chapter 1, this thesis examines the role of prediction in language comprehension: as a means for processing the incoming linguistic signal efficiently, a listener engages in anticipating possible trajectories of the incoming linguistic signal. The preceding section outlines Surprisal Theory, an influential formalism of this account that we use throughout the thesis. Surprisal Theory has formed the basis for other theories of how people optimize language comprehension; these are particularly relevant in Chapters 3, 5 and 6.

2.3.1 Surprisal theory

Perhaps the most prominent predictive account of language comprehension is Surprisal Theory (Hale, 2001; Levy, 2008a). It suggests that the effort required to process a segment of linguistic signal is directly related to how easily it can be predicted from its preceding context. This formalism draws on the information-theoretic concept of

surprisal which quantifies the amount of information conveyed by a unit as the uncertainty associated with its occurrence (Shannon, 1948)⁵. The standard definition of *conditional surprisal* is the negative log-probability of a unit in $\mathbf{u}_{1:N} = [u_1, \dots, u_N]$ conditioned on its prior context:

$$I(u_n | \mathbf{u}_{<n}) = -\log_2 p(u_n | \mathbf{u}_{<n}). \quad (2.1)$$

Surprisal theory draws a direct relationship between surprisal and processing effort—units that are less predictable, and thus more informative, are more difficult to process.

Empirical validation of Surprisal Theory spans many different aspects of language comprehension; these including the processing of syntactic and pronoun ambiguity (Levy, 2011; Demberg and Keller, 2008; Kehler and Rohde, 2013), sentence acceptability judgements (Lau et al., 2017; Meister et al., 2021), and reading times measured through self-paced reading measures and eye-tracking studies (e.g., Goodkind and Bicknell, 2018; Aurnhammer and Frank, 2019; Wilcox et al., 2020; Pimentel et al., 2023). Word-level surprisal has also been shown to correlate to more implicit measures of comprehension such as the N400 signal in EEG studies (Frank et al., 2015; Schrimpf et al., 2021) and Wilcox et al. (2023) have demonstrated that surprisal is predictive of reading times across 11 morphologically-diverse languages.

2.3.2 Information Density

Many theories of how language-use optimizes comprehension and production effort can be viewed through the lens of surprisal theory (Clark and Wilkes-Gibbs, 1986). Of particular relevance to this thesis are theories pertaining to the optimization of information density: language users prefer when the information content of a message is distributed as smoothly as possible across the communicative signal (Fenk and Fenk, 1980; Levy and Jaeger, 2007). The underlying intuition is that by avoiding large peaks and troughs in the information content of their message, speakers can increase the likelihood that their message will be understood. For example, more stable information density may reduce the cognitive load during comprehension—a listener may be unable to attend to the upcoming units $[u_{i+1}, u_{i+2}, \dots]$ if unit u_i is difficult to process (Jaeger and Levy, 2006). On the other hand, producing many redundant, uninformative units is not an efficient use of a listener’s processing effort. However, a preference

⁵Surprisal is closely related to entropy, however they differ in their functional interpretation. Entropy, which can be defined as $H(u_i | \mathbf{u}_{<n}) = -\sum p(u_n | \mathbf{u}_{<n}) \cdot \log_2 p(u_n | \mathbf{u}_{<n})$, is the weighted sum over the surprisal values of all possible units in a given context. In other words, it is the surprisal averaged across all potential continuations.

for smooth information content contours can also be seen as a consequence of the joint desire to increase predictability; if listeners can assume that speakers tend to spread information uniformly across their speech, they can constrain their predictions about upcoming speech more effectively.

In one of the first works to demonstrate this concept, Aylett and Turk (2004) proposed that speakers use prosodic effects to mediate the information density of their speech. The **Smooth Signal Redundancy (SSR)** hypothesis that they propose posits an inverse relationship between the predictability of a linguistic unit and the redundancy of its prosodic realisation. As evidence, they show that a large proportion of syllabic duration (a manifestation of prosodic prominence) can be explained by features of predictability. In other words, highly surprising syllables tend to be produced with longer duration. Similarly, the **Entropy Rate Constancy (ERC)** principle proposed by Genzel and Charniak (2002) posits that language users also use context to maintain stable information density during discourse. They show that the surprisal of sentences taken from written paragraphs tends to increase with the position of the sentence when computed in isolation. However, when conditioned on context, surprisal remains relatively constant. This effect has been demonstrated over other coherent contextual units including topic episodes and written dialogue transactions (Qian and Jaeger, 2011; Giulianelli et al., 2021).

These theories highlight both the role of available context and the set of available communicative channels in generating expectations about the upcoming communicative signal. We explore both factors in Parts I and II.

2.3.3 Estimating surprisal with large language models

Without direct access to the probability that a listener may assign to a particular linguistic unit, psycholinguists have obtained surprisal estimates using a variety of methods to investigate Surprisal Theory. These have ranged from statistical models over handcrafted features (Kehler and Rohde, 2013; Dubey et al., 2013), parsing algorithms (Hale, 2001), and models of lexical distributional statistics such as n-gram models and recurrent neural language models (McDonald and Shillcock, 2003; Wilcox et al., 2020). More recently, large-scale language models such as BERT and GPT2 have emerged as powerful estimators of surprisal for written words, reflected by their ability to predict different facets of human language comprehension behaviour. Such psychometric variables include self-paced and eye-tracked reading times of sentences and

short paragraphs (self-paced and gaze duration) (Keller, 2004; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Meister et al., 2021; Shain et al., 2024; Oh and Schuler, 2023b), acceptability judgements for sentences (Lawrence et al., 2000; Heilman et al., 2014; Lau et al., 2015, 2017; Warstadt et al., 2019), and brain response data (Frank et al., 2015; Schrimpf et al., 2021).

Large neural language models can condition their predictions on both the local context of their input and the wider distributional properties of language from their training data. Although they are often referred to as black boxes, they are less opaque (or at least opaque in different ways) than humans. Studying the degree to which they align with human language comprehension allows us to draw some inferences about processes that may be at play during human language comprehension, particularly the extent to which the types of predictions made by language models can explain human language comprehension behaviours. Here, we discuss some challenges that must be faced to expand the application of language model surprisal estimates to the study of human dialogue comprehension.

Estimating surprisal beyond individual word sequences. As the studies listed above demonstrate, a large proportion of empirical support for surprisal theory involves predicting comprehension behaviours associated with isolated linguistic stimuli, often at the level of individual words. In this thesis, we are interested in the predictability of longer sequences: upcoming dialogue turns. Advances in language modelling techniques over the past decade have allowed token-level probability estimates to be conditioned on much wider contexts than was previously possible⁶ (Ethayarajh, 2019; Liu et al., 2020b). Although predictive theories of language comprehension like the ones discussed in Section 2.3 have hypothesised how people distribute information across contextualised sequences, only a small number of works have investigated whether large language models and their capacity to contextualise surprisal estimates align with these theories (Vega and Ward, 2009; Doyle and Frank, 2015; Giulianelli et al., 2021). One potential cause is the difficulty of estimating surprisal associated with *sequences* of words. Computing utterance-surprisal directly would require computing probabilities over a high-dimensional, structured, and ultimately unbounded event space. Instead, a small number of previous works have produced utterance-level surprisal values by chaining token-level surprisal estimates together (Lau et al., 2017;

⁶BERT, proposed in 2019, used a maximum context window length of 512 tokens to condition its predictions. In 2023 alone, context window size has scaled from 4000 tokens (e.g., LLaMA (Touvron et al., 2023)) to 100000 (e.g., Claude-2 (Anthropic, 2023)).

Meister et al., 2021). For example, Meister et al. (2021) compares measures that capture different proposed pressures of information distribution such as minimizing the maximum surprisal across tokens, or the variance between tokens. However, both Lau et al. (2017) and Meister et al. (2021) apply these utterance-level measures to predict perceptual judgements based on sentences in isolation. In Chapters 3 and 4, we use similar token-level aggregations to investigate the perception and representation of contextualised language—dialogue turns in the context of the broader conversation.

Predictive power across perceptual tasks. The degree to which a model can predict different features of human language comprehension behaviour has been termed its *psychometric predictive power*. Multiple large-scale comparisons between the predictive power of different families of language models provide strong evidence that human language comprehension behaviours involve some degree of prediction (Wilcox et al., 2020; Goodkind and Bicknell, 2018; Richter and Chaves, 2020; Lau et al., 2017; Warstadt et al., 2019; Meister et al., 2021; Frank et al., 2015; Schrimpf et al., 2021). However, a growing body of work demonstrates that the predictive processes employed by language models only reflect certain components of human language processing. Such studies provide a richer picture of the types of predictive processes that may occur during human language comprehension, and under what circumstances they occur.

For example, predictive power for reading times has been shown to be a function of both model capacity and the amount of training data. Though Goodkind and Bicknell (2018) find a positive relationship between model quality quantified as next-word prediction performance (i.e., perplexity) and psychometric predictive power, it is dependent on model architecture. For example, beyond a certain model capacity, lower perplexity models yield surprisal estimates that are less predictive of human reading times (Oh and Schuler, 2023b). Similarly, surprisal estimates from models diverge from human-like expectations once their training data exceeds a certain size; Oh and Schuler (2023a) estimate the turning point to be at two billion tokens of training data. The authors of these two works suggest that the increased capacity of modern LM allows far greater degrees of memorization during training which results in underprediction of human reading times of infrequent nouns and adjectives, for example⁷.

More relevant to our own studies are findings regarding how the predictive power of language models changes across perceptual tasks. Although most of the studies

⁷Though a certain amount of model capacity and training data is required to accurately model human-like expectations, the implicit presumption in NLP that “bigger is better” does not hold for psychometric predictive power.

presented above focus on how people read written sentences, they already provide evidence that different comprehension tasks rely on different types of linguistic expectations. Wilcox et al. (2020) finds a dissociation between a model’s ability to predict human reading times, and its syntactic generalization capabilities which are more dependent on model architecture. Similar results were found by Meister et al. (2021) who find that BERT is highly predictive of acceptability judgements but describe it as “remarkably poor” for estimating reading times. Though these results do not yet paint a complete picture, they indicate that the degree of alignment between language model surprisal estimates and behaviours associated with the comprehension of contextualised language-use such as dialogue cannot be extrapolated from studies of sentence reading times. Each chapter in Part I investigates the relationship between perceptual judgements and language model surprisal estimates of dialogue from different perspectives, contributing to a more comprehensive understanding of how people make sense of more natural usages of language and motivating the use of language models in stimuli construction in Chapters 5 and 6.

Part I

Expectations in single-channel communication

Part I Introduction

The goal of Part I is to begin characterizing the *shape* of expectations in dialogue. In the following chapters, we take a step back from the proposed multi-channel perspective of spoken communication and begin by presenting both perceptual and modelling exercises using a single-channel communicative system—written dialogue. Dialogue differs significantly from the isolated words and sentences that many studies of language perception are based upon. In particular, context plays an important role in shaping the pragmatic, referential, and affective (among others) inferences that communicative partners rely on to make sense of their interactions (Ginzburg, 2001). Though the form and function of context differ in spoken and written language-use, context is important in both modalities. As such, our conclusions here form the basis for our studies of spoken communication in Part II.

Through the chapters in Part I, we extend the application of Surprisal Theory from isolated linguistic stimuli at the word-level to contextualized language use at the level of dialogue turns. This work is only possible thanks to the advances in language modelling introduced in Section 2.2.2 that allow probability estimates to be conditioned on far longer context than was previously possible.

Chapter 3

Examining expectations in written dialogue through surprisal

The content of this chapter is based on the following publication:

Wallbridge, S., Bell P., and Lai, C. (2022) *Investigating perception of spoken dialogue acceptability through surprisal*. In Proceedings of Interspeech, (4506-4510)

and addresses the following questions:

RQ: “How accurate are human expectations of the lexical content of dialogue?”

RQ: “To what extent do neural estimates of surprisal align with dialogue acceptability judgements?”

3.1 Introduction

Although monological texts have been the primary testing ground for psycholinguistic theories of language comprehension, the cognitive mechanisms for comprehension are tuned to natural, spontaneous language (Pickering and Garrod, 2004). To understand how people form expectations in dialogue, we present a novel task of *dialogue turn acceptability*. Given a context, participants are asked to judge the plausibility of a potential upcoming turn. These scores are generic and intuitive, making them an ideal window into predictive perception at the level of dialogue comprehension. To our knowledge, this is the first study of acceptability at this level of abstraction.

Importantly, this task is applicable across both written and spoken stimuli, laying the groundwork for our comparisons between the perception of these modalities in Part II. In this chapter, we only consider written stimuli; however, we begin to explore

differences between the shape of expectations between written and spoken language by studying how people make judgments when faced with written dialogues as well as transcripts of spoken dialogues. Although closely related, spoken and written communication is generated in fundamentally different conditions. As discussed in Section 2.1, the additional non-lexical channel of speech conveys novel information in its own right (Ward and Walker, 2009; Kim and Sumner, 2017; Ben-David et al., 2016) and interacts with the lexical channel, for example to mark novel content, disambiguate lexical information (Hirschberg and Pierrehumbert, 1986; Clark and Brennan, 1991), and moderate the distribution of information during communication (Aylett and Turk, 2004). Incrementality also asserts much stronger pressure in the spoken domain where utterance design in dialogue is often modelled as a parallel and predictive process (Sjerps et al., 2019). As such, lexical information is likely to be distributed differently across written and spoken signals.

Motivated by Surprisal Theory (see Chapter 2 for a detailed discussion), we leverage large language models to investigate the underlying predictive processes at play when judging the plausibility of dialogue turns. Comparing aggregates of token-level estimates from generative dialogue models to the acceptability scores obtained from human participants provides grounds to investigate the underlying processing that people may rely on when assessing dialogue turn plausibility.

3.2 Related work

In this section, we motivate our proposal of using acceptability judgments to investigate the use of predictive processes during the comprehension of dialogue.

3.2.1 Acceptability as a perceptual probe

Acceptability judgments have been used throughout both linguistics and psycholinguistics to test and develop theories of language and comprehension for decades (Hill, 1961). The fundamental premise of the task is simple: present participants with a stimulus—generally an utterance or “word set” to use the term of Coleman (1965)—and ask them to rate how acceptable it is. Though much of the research involving acceptability judgements falls into the fields of syntax and grammar, acceptability—or well-formedness—of a sequence is not synonymous with grammaticality. Ferreira and Henderson (1991) depict this with garden-path sentences. The classic example “the

horse raced past the barn fell” is syntactically valid but often judged as unacceptable. From these findings, they suggest that such ratings are related to processing difficulty, along with features related to frequency and plausibility. As such, acceptability judgments are widely accepted to be gradient (Parker, 2019; Lau et al., 2017; Huang and Ferreira, 2020).

Many of the acceptability judgments studies cited above are collected from “non-linguists”¹, however only recent technological advances have allowed the collection of large-scale acceptability datasets (Keller, 2000). These more modern corpora generally consist of sentences varying in their degree of well-formedness, either constructed automatically by introducing errors through automatic round-trip machine translation (i.e., translating a sentence to a target language, and then back to the original source language)—for example, Lau et al. (2015, 2017)—or written by linguists as per Warstadt et al. (2019) and Lawrence et al. (2000) who extracted sentences from textbooks and pre-existing linguistic literature. All of these datasets involve judgements about sentences *in isolation*. As discussed throughout this thesis, particularly in Chapter 2, comprehension and perception are heavily influenced by context. One of the only studies (at the time of writing) to collect contextualised acceptability judgments is from Bernardy et al. (2018). Using extracts drawn from Wikipedia articles, stimuli that have been corrupted are presented to participants both in isolation and in context (i.e., with their preceding and proceeding sentences). Their findings highlight the complexity of the perception of acceptability: context leads to increased acceptability ratings for the most ill-formed sentences while it diminishes ratings for the sentences judges as most acceptable in isolation. However, the use of round-trip machine translation to obtain different degrees of well-formedness is likely to constrain acceptability judgements to features of grammaticality.

Instead of considering synthetic stimuli, we propose studying realistic dialogue data, the perception of which differs from the comprehension of monological, isolated stimuli in a number of fundamental aspects (Pickering and Garrod, 2004). Dialogue has been described as a game where participants only “win” if both understand the dialogue (Lewis, 1969), or as a joint process where interlocutors collaborate over the course of conversation to build common ground (Clark and Wilkes-Gibbs, 1986). As such, context plays a much larger role in dialogue comprehension than features like syntactic correctness (Nieuwland and Berkum, 2006). As a concrete example, Fer-

¹Carden (1976) justified this controversially: “The linguist’s own intuitions are plainly untrustworthy”.

nandez and Ginzburg (2002) find that more than 11% of dialogue turns in the British National Corpus are non-sentential in isolation. Pragmatic context is particularly important. Conversational acts such as backchanneling or the use of adjacency and coordinate pairs tend to have low lexical information density, but are crucial for turn-taking and grounding (Clark and Wilkes-Gibbs, 1986; Kawahara et al., 2016). Our proposed dialogue-turn acceptability task expands the scope of acceptability judgements to include not only surface features such as syntax but also abstract features such as pragmatic coherence.

3.3 Experimental design

This section details the turn acceptability task itself, the procedure for constructing stimuli, and the collection of human judgments as well as surprisal estimates from neural language models tuned to dyadic interaction.

3.3.1 The turn acceptability task

We present a novel *dialogue continuation acceptability* judgement task where participants are presented with a segment of a dialogue c , followed by a potential upcoming turn r . To capture the gradient nature of acceptability judgments (Lau et al., 2017; Chater et al., 2006), participants rate how plausible r is in the context of c on a scale of 1 (“Very Unlikely”) – 5 (“Very Likely”). Using dialogue turns in context rather than isolated sentences as our unit of interest provides the opportunity to explore whether surprisal reflects utterance-level aspects of perception and to compare the alignment across both written and spoken dialogues. Although language users are accustomed to processing filled pauses, false starts, and corrections when listening to spoken interactions, these features can make reading conversations difficult. As such., participants are presented with an example of a spoken dialogue transcript in the stimulus presentation format accompanied by the corresponding audio before beginning the rating task.

3.3.2 Data

A major goal of this study is to study realistic language-use. As such, we focus on chit-chat dialogue. Spoken dialogue stimuli are constructed from the transcripts of the SWITCHBOARD Telephone Corpus (Godfrey et al., 1992) which consists of over 2,400 chit-chat style conversations between 542 participants covering 70 topics. The corpus

includes manual transcriptions and utterance segmentations. Through annotators are encouraged to segment based on phrase structure, segmentation is primarily based on silences: utterances are defined as speech padded by $\approx 0.5s$ of silence on either side while long utterances are split at shorter silences (Hamaker et al., 1998).

These telephone conversations are an ideal data source for this task for a number of reasons. First, speech is spontaneous. Compared to other types of dialogue such as interviews, turns in chit-chat style conversation are relatively short and provide a diverse set from which we can sample potential upcoming turns. Second, telephone conversations ensure that all information transmission is constrained to the speech modality. Third, potential effects of shared common ground based on previous interactions are controlled for by the fact that SWITCHBOARD participants have not met a priori (Vogels et al., 2020).

We construct written-dialogue stimuli using the DAILYDIALOG corpus (Li et al., 2017). It includes 13,100 written conversations intended to resemble conversations from “daily life” and thus provides a good match for the SWITCHBOARD transcripts. Dialogues in this corpus were extracted from web pages for English learners and, similar to SWITCHBOARD, span a broad range of topics from politics, to relationships.

3.3.3 Language model surprisal

We obtain surprisal estimates using the TurnGPT architecture from Ekstedt and Skantze (2020), an extension of the classical GPT-2 model described in further detail in Chapter 2 (Radford et al., 2019). Previous works which investigate the relationship between different language models and human comprehension behaviour consistently demonstrated that GPT-2 outperforms other comparable language model families including LSTMS, RNNs, and n-gram, as well as other state-of-the-art architectures such as TransformerXL (Wilcox et al., 2020; Meister et al., 2021; Golan et al., 2023). TurnGPT involves fine-tuning a pretrained GPT-2 model (117M parameters) with an augmented input consisting of three embeddings that better encode conversational structure: the original token, its position, and a binary speaker id tag that is appended to each token. We implement our model using the TurnGPT GitHub repository <https://github.com/ErikEkstedt/TurnGPT>.

To compare the relationship between surprisal estimates and human judgements across written dialogue and transcripts of spoken dialogue, the LM should be similarly effective at making predictions in both domains and the vocabulary should be fixed. We

Corpus	PPL
DAILYDIALOG	60.27
SWITCHBOARD	68.31

Table 3.1: Perplexity (PPL) in nats obtained by our TurnGPT language model on the validation sets for DAILYDIALOG and SWITCHBOARD.

take several steps to ensure comparability between spoken and written dialogue surprisal estimates. First, we use the GPT-2 BPE subword vocabulary, avoiding domain differences from out-of-vocabulary tokens (50259 tokens). Additionally, we remove punctuation except for turn-segmentation from DAILYDIALOG. Our TurnGPT model is trained from scratch on equal amounts of data from DAILYDIALOG and SWITCHBOARD ($\sim 4\text{M}$ tokens total). To account for the smaller sizes of these datasets, we use a slightly smaller architecture with 8 layers, 8 attention heads, and an embedding size of 256 compared to the originally published TurnGPT model with 12 layers, 12 attention heads, and an embedding size of 768. The final model is trained with early stopping to obtain the lowest cross-entropy on a validation set containing equal proportions of data from SWITCHBOARD and DAILYDIALOG. To further verify that our model was not over-parameterized, we ensured that a smaller model with 4 layers and 4 heads didn’t obtain lower perplexity on the validation set.

We compare the model’s predictive ability (i.e., perplexity) on corpus-specific validation sets, shown in Table 3.1. Surprisal estimates for both corpora used throughout our analysis are scaled by this modality ratio to adjust for inherent differences in predictability across corpora.

3.3.4 Behavioural study and Participants

Two related desiderata of this study are first, for stimuli to cover a wide range of realistic communicative functions (i.e., the stimuli set should consist of variable (*context, response*) pairs). Second, the amount of information in each context used to condition the scores from people and models should be comparable to enable comparisons between the written and spoken stimuli. To this end, stimuli are generated from each corpus by first obtaining contexts with a comparable quantity of information, operationalised as the cumulative per-token surprisal of a set of turns. This context surprisal measure is normalised by the corpus perplexity ratio (see Table 3.1) to draw closer

comparisons between written and spoken stimuli.

We sample 10 contexts from each modality across a range of cumulative token surprisal, with the aim of producing stimuli that were appropriate for a behavioural test (i.e., that provide enough information while not being too long), while controlling for the amount of information available in the contexts between corpora. We select upper and lower cumulative surprisal bounds for contexts manually from the upper and lower quartiles of the distributions of cumulative surprisal across 1–5 turns in both corpora, adjusted with the perplexity ratio. The resulting context set from DAILYDIALOG consists of between 1–3 turns (1.7 on average) and between 1–5 turns (2.9 on average) from SWITCHBOARD. Each context is used to create 10 (dialogue context c , upcoming turn r) stimuli—1 pair contains the true upcoming turn; 9 contain negative upcoming turns. To ensure that negative samples are not all immediately obvious, we sample negative turns across the range of conditional surprisals expected from true (c, r) pairs. For this, we fix the maximum conditional surprisal to the third quartile marker in the distribution of conditional surprisal across both corpora and sample uniformly below that. In total, we generate 100 stimuli for each modality².

We recruited 52 participants from Prolific Academic; all were native English speakers based in North America. Each participant was presented with 25 stimuli through a Qualtrics survey, taking 9 ± 2 minutes to complete. Attention check stimuli which were manually selected for being extremely plausible (i.e., true (c, r) pairs that include noun overlap) were interspersed throughout each survey. 20 participants obtained less than 75% overall accuracy on the check questions; their results were excluded from our analyses.

3.3.5 Surprisal on a turn level

Surprisal can be estimated at the token level as the log probability for token x_n under a language model. The majority of studies investigating surprisal estimates from language models involve tasks at the level of individual tokens; however, there is less consensus for how to compute surprisal at a sequence level. Recent works have investigated surprisal at the sentence- and phrase- levels, and their alignment with syntactic generalization ability (Wilcox et al., 2018). Of particular interest is the work by Meister et al. (2021) which studies the information distribution of a sequence using variations of both global and local aggregates of token-level surprisal. Using correlation analyses

²Examples of our stimuli can be found at <https://sarenne.github.io/is-2022/>

as well as linear regression models, they find that aggregates align to different degrees with sentence reading times and acceptability judgements of isolated sentences. In particular, utterance scores $S(\mathbf{u})$ for $\mathbf{u} = (u_0, u_1, \dots, u_N)$ that are linear or super-linear functions of their constituent token-level surprisal estimates ($s(u_t)$) are most predictive for both behavioural variables:

$$S(\mathbf{u}) = \frac{1}{N} \sum_{t=0}^N s(u_t)^\alpha \quad (3.1)$$

where $\alpha > 1$ produces a super-linear relationship. In their work, Meister et al. (2021) test α between 0.25 and 2, finding that values of 1 or slightly above are most predictive.

We obtain token-level surprisal estimates from our trained TurnGPT model for each potential upcoming response $\mathbf{r} = (r_0, \dots, r_N)$ in our set of stimuli and use them to compute a number of global and local operationalisations of turn surprisal; these are described in turn below.

Global metrics

Cumulative surprisal, $S_{total}(\mathbf{r}|\mathbf{c})$, is often used to model processing effort of an utterance (Meister et al., 2021).

$$S_{total}(\mathbf{r}|\mathbf{c}) = \sum_{n=1}^N [S(r_n|\mathbf{r}_{<n}, \mathbf{c})] \quad (3.2)$$

We also consider **average surprisal**, $S_{mean}(\mathbf{r}|\mathbf{c})$, per token to eliminate the influence of sentence length which will likely have different effects on plausibility scores compared to processing effort (Lau et al., 2017).

$$S_{mean}(\mathbf{r}|\mathbf{c}) = \frac{1}{N} \sum_{n=1}^N [S(r_n|\mathbf{r}_{<n}, \mathbf{c})] \quad (3.3)$$

To control for the inherent surprisal of an utterance, we also consider the difference between the conditional and isolated mean utterance surprisal. We term this **relative surprisal** $S_{relative}(\mathbf{r}|\mathbf{c})$.

$$S_{relative}(\mathbf{r}|\mathbf{c}) = S_{mean}(\mathbf{r}|\mathbf{c}) - S_{mean}(\mathbf{r}) \quad (3.4)$$

Psycholinguistic theories diverge on whether or not discourse comprehension involves a context-independent analysis before integrating wider discourse context (Nieuwland and Berkum, 2006). Though the isolated utterance surprisal, $S_{mean}(\mathbf{r})$, could theoretically be obtained by computing the surprisal of the utterance without

context, language models are trained to make predictions conditioned on context. Therefore, we compute $S_{mean}(\mathbf{r})$ as the average surprisal of response turn $\mathbf{r}$ conditioned on 100 randomly sampled contexts within the range of acceptable cumulative surprisal.

Local metrics

Theories such as the uniform information density (UID) hypothesis and smooth signal redundancy (SSR) hypothesis, both of which are discussed in Chapter 2, posit that language users prefer information content to be evenly distributed across the constituent units of a signal (Aylett and Turk, 2004; Levy and Jaeger, 2006). Motivated by such theories, we operationalise utterance surprisal locally, as function of the individual token surprisal. These definitions also provide a more granular view on potential differences in how lexical information is distributed between written and spoken language. We consider **maximum per-unit surprisal**, $S_{max}(\mathbf{r}|\mathbf{c})$, which has been used to capture points of extreme cognitive load (Lau et al., 2017).

$$S_{max}(\mathbf{r}|\mathbf{c}) = \max[S(r_n|\mathbf{r}_{<n}, \mathbf{c})] \quad (3.5)$$

as well as **surprisal variance** between words, $S_{var}(\mathbf{r}|\mathbf{c})$, as used by Meister et al. (2021).

$$S_{var}(\mathbf{r}|\mathbf{c}) = \frac{1}{N-1} \sum_{n=2}^N [S(r_n|\mathbf{r}_{<n}, \mathbf{c}) - S(r_{n-1}|\mathbf{r}_{<n-1}, \mathbf{c})]^2 \quad (3.6)$$

3.4 Results & discussion

Similar to the previous works, we examine the relationship between these definitions of surprisal and judgement scores using Spearman’s rank correlation ρ , as well as ordinal regression models (Lau et al., 2017; Meister et al., 2021).

3.4.1 Dialogue turn acceptability

As shown in Figure 3.1, participants were able to distinguish true turns from negatives samples; 95% and 90% of scores for true turns were either 4 or 5 for DAILYDIALOG and SWITCHBOARD, respectively. For each stimulus, we compute the mean (*context*, *response*) scores across participants. Using the highest scoring response per context, we find respective accuracies of 90% and 70%. Although people can discriminate between true and false upcoming turns relatively accurately, they are often happy to

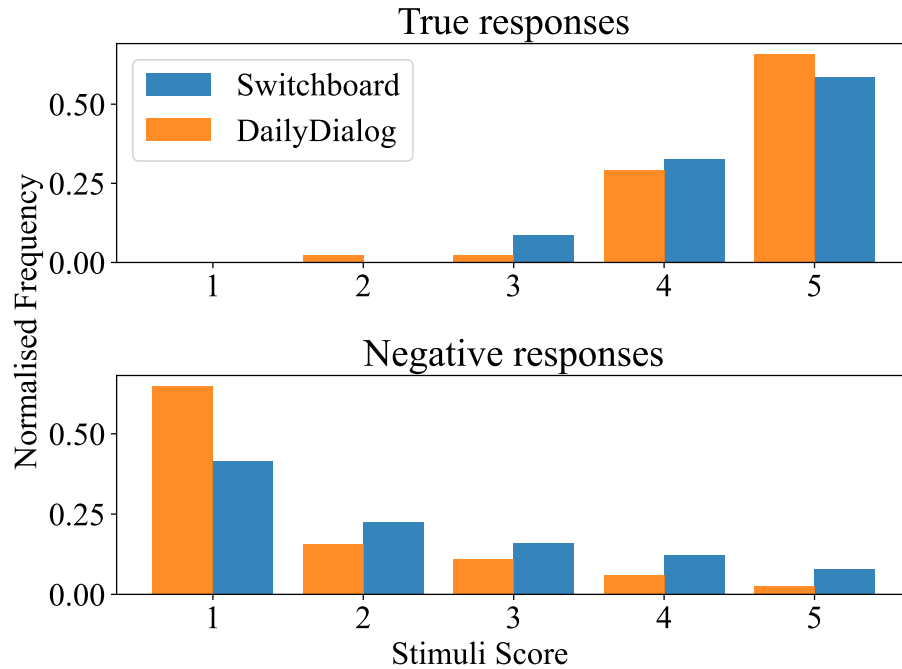


Figure 3.1: Distributions of scores assigned to the true and negative (context,response) pairs by participants across the SWITCHBOARD and DAILYDIALOG corpora.

accept more than one plausible response for a given context. This result has implications for the perceptual validity of language models that have been adapted to dyadic interaction where the widely used training paradigm of response selection involves discriminating a single true upcoming response from negative samples given a context. We address this point further in the upcoming Chapter 4.

Figure 3.1 also highlights differences between the spoken and written corpora, particularly that participants make less certain judgements for stimuli constructed from SWITCHBOARD transcripts. Although score distributions for true stimuli are similar across corpora, negative stimuli from SWITCHBOARD receive a wider range of scores – twice as many were rated as likely (either 3 or 4) in SWITCHBOARD, and participants were more likely to rate turns as “Very Unlikely” (1) in DAILYDIALOG stimuli. Because the informativeness of our stimuli context was controlled for, differences in discriminative ability and scoring certainty could suggest differences in the informative nature of basic turn units between the lexical content of spoken and written dialogue.

Surprisal Metric	DAILYDIALOG		SWITCHBOARD	
	ρ	p -val	ρ	p -val
<i>Total</i>	-0.341	0.001	-0.273	0.006
<i>Mean</i>	-0.350	<0.001	-0.299	0.003
<i>Relative</i>	-0.360	<0.001	-0.262	0.009
<i>Max</i>	-0.407	<0.001	-0.400	<0.001
<i>Variance</i>	-0.295	0.003	-0.217	0.038

Table 3.2: Spearman ρ correlation between different surprisal estimates and median human judgement scores across both DAILYDIALOG and SWITCHBOARD.

3.4.2 Quantifying predictive power of surprisal

Given that people could leverage contexts to accurately discriminate true upcoming turns from false ones, we use the plausibility scores to explore the relationship between human acceptability judgements and LM-estimated surprisal characteristics. If surprisal estimates, as we’ve operationalised them here, align with perception at the level of dialogue turns, we should expect responses with high surprisal to obtain lower plausibility scores, i.e., a negative relationship between surprisal and score.

Comparing measures

Results in Table 3.2 demonstrate weak but statistically significant negative correlations of median score with our operationalisations of surprisal. Slightly stronger correlation coefficients have been reported for related judgements of grammatical acceptability (Meister et al., 2021; Lau et al., 2017). These likely reflect differences in the reasoning required to make predictions about dialogue (e.g., pragmatic/interaction features) and the factors involved in making isolated syntactic judgements.

Interestingly, previous work using surprisal estimates from similar language models find global operationalisations to offer more explanatory power than local ones for reading times and isolated sentence acceptability judgements (Meister et al., 2021). Our results are more similar to those reported by Lau et al. (2017), showing maximum conditional surprisal per token to have the strongest correlation with judgement scores. Again, this variation may be explained by differences in the reasoning required for these behavioural tasks.

Removed feature	ELPD difference	Standard error
S_{total}	+1.9	0.4
$S_{relative}$	+0.7	0.9
S_{mean}	+0.3	1.2
S_{max}	-9.6	4.3
S_{var}	-2.1	1.9
<i>Corpus</i>	+0.3	1.2
<i>All surprisal</i>	-39.0	10.5

Table 3.3: Effect of ablating features from full model. Significant differences (i.e., where the resulting Expected Log Predictive Density difference is ≥ 2 SE) are shown in bold.

Combining measures: Ordinal logistic regression

Based on the significant but relatively weak correlations of individual surprisal measures with median scores, we examine whether combinations of surprisal measures could provide more predictive power. Since our perception experiment used a categorical rating scale, we fit multilevel cumulative ordinal regression models (logit link function, uninformative flat priors), using the R package `brms` (Bürkner and Vuorre, 2019; Bürkner, 2017). The goal of this type of regression is to determine thresholds for different score categories (the predictee) from a latent variable determined by a linear combination of the predictors. Our predictors include the five surprisal metrics and a corpus indicator. Additionally, we include the context surprisal as well as unigram and bigram overlap between the context and response (weighted by corpus frequency) to control for these potential sources of variation. Similarly, we include group-level effects (i.e., random intercepts) to control for participant and context identity, for which we consistently find non-zero variance associated with group members. Models are evaluated using leave-one-out cross-validation, estimating Expected Log Predictive Density (ELPD), i.e., log probabilities of outcomes given the predictors, summed over all the held-out data (Vehtari et al., 2017). Log predictive density scores provide an estimate of the model fit to the data, while the cross-validation penalizes cases of overfitting.

We perform ablation of the individual surprisal measures with respect to the model

Included Features	ELPD difference	Standard error
<i>global</i>	-8.4	4.4
<i>local</i>	+0.6	2.2
<i>local+relative</i>	+1.3	1.2
<i>local+mean</i>	+2.2	1.0
<i>(local+mean)*corpus</i>	+0.4	2.1

Table 3.4: Differences between models with all operationalizations of surprisal compared to those with only local or only global measures. **corpus* indicates that interaction between the corpus indicator and the surprisal measures for that model have been added to the model.

using all predictors to investigate their contributions to model fit. Table 3.3 shows the difference in ELPD with respect to the full model, as well as the associated standard error of the difference (SE). We take a model to have a significantly better fit when the ELPD difference is more than twice the SE. The ELPD associated with the full model is -1078.3 . Removing all surprisal measures (leaving only n-gram and group predictors) significantly decreases the fit, indicating that participants were not making decisions based only on direct lexical matching between context and turn. In general, removing local measures (particularly S_{max}) reduces the fit, while removing global measures improves it, though not all differences were significant. Removing the corpus indicator also potentially reduces model fit, though the change is within the error margin.

Table 3.4 shows the difference between the full model, models with local or global surprisal metrics, and combinations of the two. The model including only global features performs worse than the full model, suggesting that their inclusion (specifically S_{total}) is leading to overfitting. However, the difference is not significant. Our best-fitting model includes local (S_{max} , S_{var}) metrics with S_{mean} . We find that a combination of individually weak predictors results in the greatest predictive power for the perceived likelihood of a turn in context. The 95% confidence intervals for surprisal measure coefficient estimates and the corpus indicator all exclude zero (while context surprisal and n-gram measures confidence intervals include zero), supporting a non-zero contribution for these effects. The estimates indicate that stimuli with higher turn surprisals (S_{mean} , S_{max}) received lower ratings. However, contrary to the correlation analysis, higher word-to-word surprisal variability (S_{var}) contributes to higher scores

when accounting for the other metrics. We did not find any further improvement in model fit from including corpus interaction terms with the surprisal measures.

3.5 Conclusions

The results presented here are the first, to our knowledge, that probe human perception of plausibility at the level of dialogue turns. By examining how people rate the plausibility of true and false dialogue continuations, we find that true turn continuations often receive higher scores than distractor turns. However, people could entertain multiple continuations as plausible. The idea that the upcoming communicative signal is not perfectly predictable is intuitive; however, this study is one of the first attempts to quantify the degree of unpredictability in dyadic communication. We term this acceptable variability in expectations *optionality*. In Part II, we return to this concept and investigate its sensitivity to different features of context, arguing that it is an important feature of human language processing to account for when studying predictions of both people and models.

We then investigated whether expectations regarding the upcoming turn in dialogue differed between written dialogue and spoken dialogue transcripts using two corpora with matched domains. Both DAILYDIALOG and SWITCHBOARD consist of casual conversations covering a wide variety of topics. We find differences between people’s judgements between corpora: for example, they are less effective at discerning between true and negative turns in SWITCHBOARD. These differences suggest that information is distributed differently across lexical content in written and spoken dialogue, which we believe may be an effect of the different channels available in these modalities. We test this hypothesis more directly through the perceptual studies presented in Part II.

Finally, using the popular GPT2-style language model, we aggregate token-level estimates of surprisal to derive different operationalizations of utterance surprisal and explore their predictive power for the dialogue turn acceptability task. In some ways, our findings are complementary to the previous works discussed in this chapter—all operationalisations displayed a weak but statistically significant correlation with human judgements across written and spoken dialogue. This provides some evidence that purely predictive models reflect the processing required to make such acceptability judgements. We find that combinations of global and local operationalisations from TurnGPT provided the highest predictive power. Though individual operationalisations are more interpretable, the increased power of combined operationalisations

indicates that utterance-level surprisal requires higher degrees of complexity and flexibility than the relatively rigid aggregation strategies presented here provide, pointing to potential pitfalls of using token-level surprisal estimates from neural models as a model of utterance-level plausibility. We explore this further in Chapter 4 with different styles of language modelling.

Chapter 4

Machine expectations in written dialogue

This chapter presents the insights from works published in both

Wallbridge, S., Bell P., and Lai, C. (2022) *Investigating perception of spoken dialogue acceptability through surprisal*. In Proceedings of Interspeech, (4506-4510)

Wallbridge, S., Bell P., and Lai, C. (2023) *Do dialogue representations align with perception? An empirical study*. Long paper in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (2696–2713)

to answer the following question:

RQ: “*To what extent do machine representations of dialogue turns align with human expectations?*”

4.1 Introduction

Chapter 3 presented an utterance-level view of expectations and prediction in dialogue. We found that human predictions were often accurate; however, our results also showed expectations to be somewhat stochastic: people seem able to comfortably maintain multiple upcoming turns as plausible continuations at once. Prior to the recent popularization of generative language modelling, the majority of dialogue language modelling was based around the self-supervised task of response selection: selecting the true upcoming utterance in a dialogue from distractor samples (Wolf et al., 2019; Wu et al., 2020; Henderson et al., 2020; Liu et al., 2020a). Though this train-

ing objective has produced strong performance on a number of dialogue modelling tasks (Peng et al., 2020), it is unclear to what degree the resulting models capture the optionality found in human turn acceptability judgements.

In this chapter, we begin by framing the dialogue acceptability task as a discriminative one—how accurately can people perform the response selection task? This provides one of the first human benchmarks on this widely used evaluation for dialogue models. We use this benchmark to investigate the perceptual validity of the response selection objective by comparing the predictive power of models trained towards this objective and the autoregressive model introduced in Chapter 3, and examining the similarities between human and model predictions.

4.2 Related work

In this section, we discuss the use of large language models in the domain of dialogue processing. In doing so, we also motivate our choice of models for our experiments.

4.2.1 Language modelling for dialogue

Similarly to the development of psycholinguistic theories as discussed in the introduction of this chapter, the vast majority of NLP models have been developed primarily using monological data. However, interest in modelling interaction and dialogue is booming, fuelled by the development of commercial dialogue systems like Apple Siri¹ and Amazon Alexa². Studies related to dialogue perception are also gaining interest, focused on tasks such as dialogue act classification and dialogue coherence estimation (Shriberg et al., 1998; Tran, 2020; Cervone et al., 2018; Wu and Xiong, 2020; Kumar et al., 2021).

Response selection. Regardless of the downstream application, response selection has become a pervasive task in dialogue modelling, where it is used both as an evaluation metric and as a training signal (Wolf et al., 2019; Han et al., 2021). A major reason for this is that it doesn’t require annotated labels. Given some degree of dialogue history as an anchor, response selection involves selecting the upcoming response from a set of potential turns. This is a direct extension of quintessential next-sentence-prediction task introduced in BERT pretraining which makes a strong implicit

¹<https://www.apple.com/siri>

²<https://developer.amazon.com/alexa>

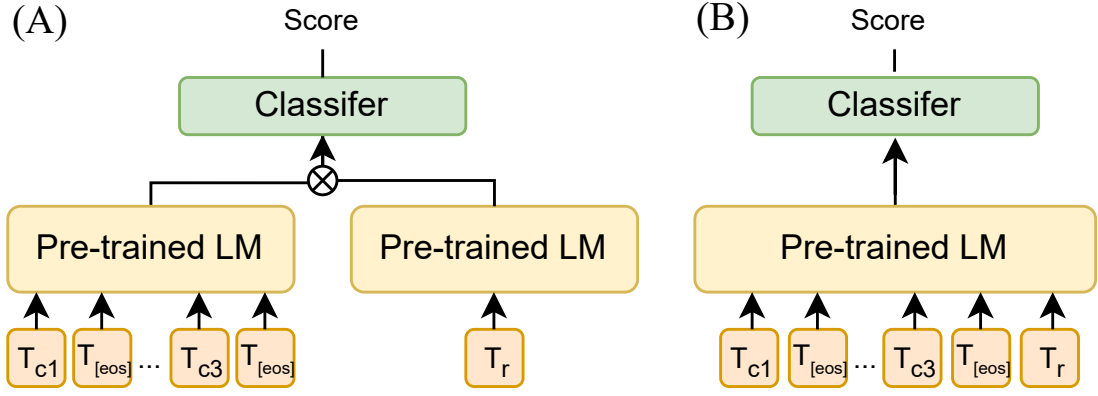


Figure 4.1: Depictions of architectures for (A) bi- and (B) cross-encoders. T denotes tokenized text and is structured following the format required for SBERT encoding. When BERT-based models are used, we prepend all input sequences with a $[CLS]$ token. Cross-encoder input also contains a $[SEP]$ token before and after T_r .

assumption that there is a single correct upcoming sentence (Devlin et al., 2019). In this work, we explore the effect of this assumption for making predictions regarding turns dialogue.

Response selection models, also known as retrieval-based dialogue models, can be broadly separated into two classes: *bi-encoder* models that learn independent representations of responses and their respective contexts, and *cross-encoders* which encode the context and response as a single representation before scoring them (Henderson et al., 2020; Zhou et al., 2016; Wu et al., 2020, 2017). These two variants are depicted in Figure 4.1. The latter often involve tuning a pre-trained language model using dialogue-specific objectives (Wolf et al., 2019; Han et al., 2021; Xu et al., 2020). Such methods aim to make use of the representation capacity of pre-trained language models, while also leveraging important and unique structural features of dialogue data.

By encoding contexts and responses in isolation, bi-encoders achieve less expensive training and inference, as representations can be cached. However, they enforce a strong assumption of independence between contexts and responses. The separate encoders can provide a weak notion of position (i.e., the same lexical content may be encoded differently as a response than as context). For example, the utterance “What do you think?” could be found in the context or be employed as a response. The contextualisation of cross-encoders comes at a high computational cost but captures richer interactions between context and responses. A notable exception includes the poly-encoders proposed by Humeau et al. (2020) which combine bi- and cross-encoding

strategies.

Given the importance of response selection as a self-supervised task in dialogue modelling, we use this chapter to explore to what extent this task aligns with the perception of spoken dialogue acceptability, and which aspects of model architecture are important for response selection performance.

4.3 Experimental setup

4.3.1 Data

Evaluation. To investigate the relationship between representations of dialogue and human perception, we evaluate how well different models can predict the human acceptability scores collected in Chapter 3. Recall that this dataset comprises of 100 (*context c*, *response r*) pairs extracted from the SWITCHBOARD Telephone Corpus with their associated median plausibility scores. Each of the 100 stimuli consists of a set of speaker turns as context **c** and an upcoming response **r** where **r** is either the true upcoming turn in the dialogue or a turn selected from elsewhere in the corpus. Each dialogue context **c** was presented in 10 pairs: one with the true upcoming turn, and 9 with negative upcoming turns.

Training. We extend our comparison between human judgements and a generative language model to retrieval-based models by training a range of models towards the response selection task on the SWITCHBOARD dataset. We split the corpus by conversation into training, validation, and test sets (80%, 10%, 10%). Sampling conversations was done randomly, however all conversations from which contexts in the Chapter 3 dataset were selected are kept in the test set. Transcripts are all lower-cased and speech-based annotations such as pronunciation markers and speech events are removed.

4.3.2 Models

We compare two of the most prominent styles of language modelling used in for dialogue applications – generative and retrieval-based. All of our models are implemented in PyTorch (Paszke et al., 2019) and pre-trained language models are obtained from the Transformers library (Wolf et al., 2020).

Generative language modelling. To test generative model capabilities, we employ the same TurnGPT architecture introduced in Chapter 3. For our experiments here, we train a fresh model by fine-tuning the pretrained GPT-2 base model from the Transformers library (`gpt2`) on our training portion of SWITCHBOARD using the augmented TurnGPT cross-entropy loss. Response-level scores are obtained using the token-level surprisal aggregates introduced in the same chapter (see Chapter 3 for a detailed description of the architecture and utterance-level aggregation strategies).

Retrieval-based language modelling. Compared to generative models that compute word-level surprisal estimates that are aggregated into response-level scores, retrieval-based models act as much more powerful aggregation strategies by producing response scores implicitly. We test a number of retrieval-based models, based primarily on the cross-encoder BERT-FP architecture proposed by Han et al. (2021), which obtains state-of-the-art response selection performance for written conversational benchmarks including the Ubuntu (Lowe et al., 2015), E-commerce (Zhang et al., 2018), and Douban (Wu et al., 2017) datasets. This model has 109M parameters, making it comparable to the 117M parameters of TurnGPT. To generate a score for a response, this model encodes the corresponding (*context c*, *response r*) pair as a single sequence using pre-trained BERT (Devlin et al., 2019). As is common in many sequence-level representation tasks, the resulting representation of the BERT [CLS] token for this sequence is fed through a single-layer classifier to produce a response selection score—the relevance of the response given the context. The “FP” suffix of BERT-FP denotes “Fine-grained Post-training”, the two training strategies applied to obtain response scores. Post-training involves augmenting the standard masked language modelling task with an utterance relevance classification task. The model is then fine-tuned towards response selection.

Training procedure. We finetune all models towards response selection using a binary cross entropy loss on our training portion of SWITCHBOARD. To produce a greater number of training samples, we apply the fine-BERT-FP tuning procedure of splitting each conversation into contiguous (*c*, *r*) segments. We fix the number of context turns to 3 to compare with the human judgements collected in Chapter 3. The response selection training set consists of an even split of positive and negative (*c*, *r*) pairs. Negative pairs are generated by sampling a random response turn from the training set. This procedure results in 267,562 samples for training the cross- and bi-encoders.

Model	Metric	ρ	p -value
<i>TurnGPT</i>	S_{total}	-0.302	0.002
	S_{mean}	-0.392	<0.001
	$S_{relative}$	-0.395	<0.001
	S_{max}	-0.463	<0.001
	S_{var}	-0.346	0.001
<i>BERT-FP</i>	RS Score	0.637	<0.001

Table 4.1: Spearman ρ correlations between human plausibility scores and LM response scores. Scores are computed using the range of turn-level surprisal estimates for TurnGPT, while scores for BERT-FP consist of the joint (context, response) response selection score.

We follow recent “BERTology” research to mitigate the pervasive issue of training instability when fine-tuning LLMs (Merchant et al., 2020; Mosbach et al., 2021). Compared to the original BERT fine-tuning approach presented by Devlin et al. (2019), this involves using smaller learning rates with bias correction to avoid vanishing gradients early in training and increasing the number of epochs. We use a batch size of 16 and make use of the Pytorch AdamW optimizer with an initial learning rate of $1e-5$. With early stopping, our best model in terms of validation loss is achieved at epoch 8.

4.4 Results & discussion

We begin by comparing the correlation between human acceptability judgements and scores obtained from generative and retrieval-based language models before examining their judgements from a discriminative perspective.

4.4.1 Correlation with plausibility perception

Table 4.1 show the correlations between model response scores and human plausibility judgements. Similar to previous works (Lau et al., 2017; Meister et al., 2021) and our own results presented in Chapter 3, we find a relatively weak correlation between TurnGPT surprisal metrics and plausibility scores, as well as variation in correlation

strength between surprisal metrics³.

These results demonstrate that language modelling strategy has a larger effect on correlation strength than surprisal metric: we find a much stronger correlation between turn-plausibility scores from humans and scores predicted by our retrieval-based cross-encoder BERT-FP than for any TurnGPT surprisal aggregate. This suggests that the response selection task used to fine-tune BERT produces a latent space that matches some aspects of dialogue acceptability.

4.4.2 Discriminative performance

To better understand the role of the response selection objective in the strong correlation between BERT-FP response scores and human acceptability judgements, we reframe the rating task as a discriminative one.

We convert the acceptability scoring task to response selection classification by comparing the response scores for each response to a given context and selecting the highest-scoring one across each scoring metric for both humans and models. Table 4.2 provides the response selection performance of our models, as well as a benchmark of human consensus performance for this task. We report both the standard response selection accuracy metric $R_{10}@1$ as well as F1 score. $R_{10}@K$ is a measure of the proportion stimuli for which the true response was correctly identified in the K top-scoring responses, while F1 denotes the highest linear classification accuracy that can be achieved across all (c, r) pairs based on a given metric.

The response selection paradigm is used widely throughout dialogue modelling. But, to our knowledge, this is one of the first investigations of how well humans can judge upcoming responses. Our results show that people can execute the response selection task relatively well using a small amount of context; however, their performance is not perfect in terms of $R_{10}@1$ accuracy.

Interestingly, we find that BERT-FP outperforms people at this task, achieving higher accuracy and F1 score⁴. This “superhuman” performance is further evidence

³We find slight differences between the relative correlation strengths of the surprisal operationalisations obtained from the TurnGPT models used here and in Chapter 3. Although S_{max} consistently achieves the strongest correlation, this instability provides further evidence that these aggregation strategies are not robust. Future work could explore how training affects this further.

⁴The original BERT-FP paper presents a two-stage training strategy for response selection (Han et al., 2021) First, the model is post-trained using the task of utterance relevance classification, then it is fine-tuned towards response selection. To facilitate our comparisons across architectures, we report experiments that involve only fine-tuning with no post-training. However, we find that following the full procedure of post-training before fine-tuning increases response selection accuracy $R_{10}@1$ by 3 absolute points ($R_{10}@[1, 2, 5] = 0.705, 0.840, 0.975$)

Model	Metric	F1	$R_{10}@1$
<i>Human</i>	Mean score	0.783	0.7
	Median score	0.800	0.8
<i>TurnGPT</i>	Max surprisal	0.435	0.3
	Mean surprisal	0.538	0.4
<i>BERT-FP</i>	Score	0.889	0.9

Table 4.2: Discriminative performance (response selection) of people, TurnGPT, and BERT-FP measured using F1 score and accuracy at the first position ($R_{10}@1$).

that the response-selection objective deviates from human perception of dialogue. We provide a some qualitative observations of deviations between model and human judgements below (see Error Analysis).

Although TurnGPT shows a weak to moderate correlation with human judgements, it performs very poorly in the response selection task framing. Different surprisal definitions had relatively large impacts on the correlation strengths reported in Table 4.1, but these were not reflected in the F1 scores. TurnGPT is trained using cross-entropy loss over next-token prediction (autoregressive pre-training), while BERT-FP is based on the masked language modelling paradigm – these differences in underlying pre-training could help explain performance deviations and would be interesting to explore in future work. We discuss these findings further in Section 4.4.5.

Error analysis: cross-encoders and humans

Neither BERT-FP nor human raters achieve perfect performance on the response selection proxy task. However, their mistakes are different. The cross-encoder succeeds on all three questions where humans did not rate the true response highest, and vice versa for the question mis-ranked by the model. We present a qualitative analysis of these stimuli here; the full stimuli can be found in Appendix A.

For all three stimuli that people mis-ranked, the true response obtained the second highest score and had a mean plausibility score greater than 3. In two of these cases, the highest-scoring response was a generic response (“mhm”, “yeah”), which can be plausible in a wide range of contexts. In the final case, the highest-scoring response was a question that shifted the topic of conversation. We noted this pattern in other stimuli: responses that initiated topic shifts – often questions – were often rated as

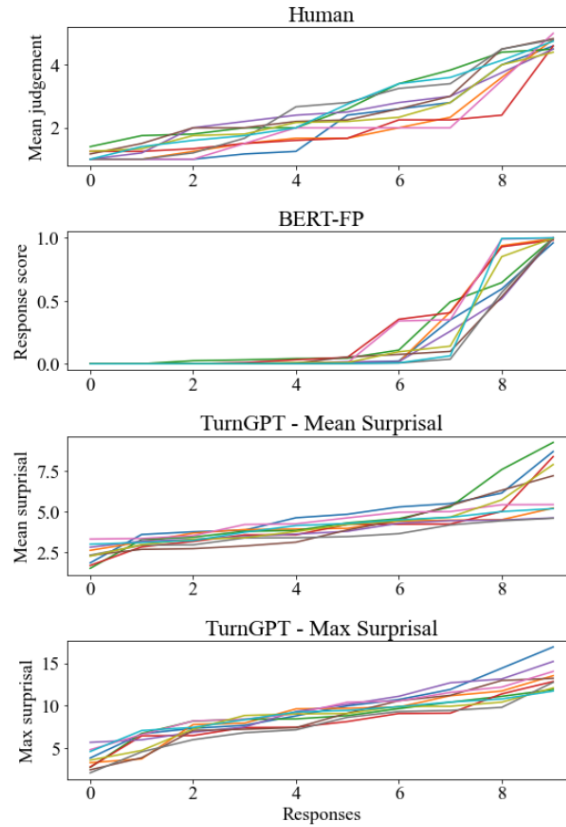


Figure 4.2: A comparison of scoring behaviour across participants and models. Each line depicts the set of scores (sorted in ascending order) obtained for a stimulus which consists of the 10 responses paired with a particular context.

somewhat likely by people but obtained low BERT-FP scores.

In comparison, for the stimuli mis-ranked by BERT-FP, the true response obtained the 4th highest score. All four of the mis-ranked stimuli included two or more context turns, suggesting that the amount of context is not necessarily a driving factor in mistakes for either the model or people.

We also examine the rank correlation (Spearman r) between human ratings and model scores on a per-stimuli basis. Coefficients range from 0.33 to 0.93 ($\mu = 0.65 \pm 0.18$). The lowest correlations are obtained for stimuli where the model only scores a single response highly while the rest obtain scores close to 0. The distributions of BERT-FP scores across stimuli are more peaked than ratings produced by people (see Figure 4.2). Surprisal metrics from TurnGPT show score distributions more similar to human judgements, further highlighting differences between these models.

4.4.3 Model ablation for response selection

Generative (TurnGPT) and retrieval-based (BERT-FP) models display significantly different propensities for psychometric predictive power and response selection performance. We use ablation studies to better understand which aspects of the retrieval-based model architecture are valuable for response selection. In particular, we consider:

1. Independence of context and response encoding: bi- versus cross-encoders.
2. Inclusion of dialogue structure: using *[eos]* tokens.
3. Importance of fine-tuning pre-trained models.

Model variants

We compare **cross- and bi-encoder networks**, depicted in Figure 4.1. The cross-encoder architecture is based on BERT-FP which encodes the joint context and response pair and then feeds the resulting sequence representation through a classification network (Han et al., 2021). We compare this cross-encoder architecture to a bi-encoder style counterpart, converting the BERT-FP model such that each context and response is encoded in isolation using the pre-trained BERT language model. The resulting context and response representations are concatenated and fed through the same classification network as the cross-encoder architecture. To this end, we also modify the final **classification networks**, comparing results obtained with the original single linear-layer classifier and with three fully-connected ReLU layers with dropout followed by a final sigmoid layer.

Additionally, we compute response scores based on different methods of **aggregating utterance representations** produced from BERT encoding. We select the BERT *[CLS]* token and SBERT sequence representations (Devlin et al., 2019; Reimers and Gurevych, 2019). SBERT representations are specifically designed to capture semantic similarity at the utterance-level and are produced by fine-tuning pooled representations of the constituent tokens from BERT (Reimers and Gurevych, 2019). For our experiments, we use `bert-base-uncased` and `all-MiniLM-L6-v2` from the Transformers library as base encoders for BERT and SBERT variants, respectively.

To maintain consistency with pre-training, joint (c, r) pairs are presented in the following format for BERT-based models:

$$([CLS], c_1, [eos], \dots, c_k, [eos], [SEP], r, [SEP]) \quad (4.1)$$

Encoding Strategy	$R_{10}@1$ accuracy			
	Linear classifier		Non-linear classifier	
	Frozen	Unfrozen	Frozen	Unfrozen
<i>Bi-encoder</i>	0.102	0.105	0.455	0.507
<i>Cross-encoder</i>	0.236	0.465	0.307	0.472
<i>Cross-encoder + [eos]</i>	0.216	0.611	0.306	0.612

Table 4.3: Response selection accuracy measured as $R_{10}@1$ for models based on SBERT representations. Note that our SBERT baseline (response selection based on euclidean distances between pre-trained SBERT representations of context and responses) achieves an $R_{10}@1$ of 0.480.

And as such for SBERT-based models:

$$(c_1, [eos], c_2, [eos], c_3, [eos], r) \quad (4.2)$$

where the $[eos]$ token represents the end of a speaker’s turn.

Finally, we test the importance of **unfreezing** the BERT model parameters by comparing classification scores obtained from training all model parameters compared to training only the classification component of the model.

Results

We present a simple baseline for response selection: minimizing the Euclidean distance between pre-trained SBERT representations of (c, r) pairs. This outperforms many of the ablated models discussed hereafter (see Table 4.3).

Independence of context and responses. Although bi-encoders achieve cheaper training and inference by encoding contexts and responses in isolation, they lack the rich contextualisation of cross-encoders. Previous work reports better response selection performance from cross-encoders for written dialogue (Urbanek et al., 2019). However, recent work has shown that models can make predictions about the acoustic realisation of upcoming conversational turns based on small amounts of context (Fuscone et al., 2020).

We, therefore, compare bi- and cross-encoding strategies to understand the value of such contextualisation for the lexical content of spoken dialogue.

Encoding Strategy	$R_{10}@1$ accuracy			
	Linear classifier		Non-linear classifier	
	Frozen	Unfrozen	Frozen	Unfrozen
<i>Bi-encoder</i>	0.103	0.104	0.212	0.316
<i>Cross-encoder + [eos]</i>	0.324	0.675	0.346	0.673

Table 4.4: Response selection accuracy measured as $R_{10}@1$ for models based on BERT [CLS] representations.

Tables 4.3 and 4.4 show that bi-encoders require a classifier with some degree of nonlinearity. Regardless of whether the underlying language model is frozen or not, bi-encoders with a linear classifier do not improve on chance performance. In comparison, the cross-encoder improves on chance even with a frozen language model and a linear classifier. This is the case for both SBERT- and BERT-based models, indicating that masked language model pre-training extracts information that is of some value for predicting upcoming dialogue turns.

When non-linearity is introduced in the classifier, bi-encoding outperforms cross-encoding by quite a margin on top of a frozen SBERT model. This may reflect similarities between response selection and the SBERT training objective of contrastive sentence similarity (Reimers and Gurevych, 2019). We verify that this result was not caused by the special [eos] turn delimiter that wasn’t seen during cross-encoder pre-training by training the same model without this token added to the input.

Regardless of the underlying LM, the performance gap between unfrozen bi- and cross-encoders indicates that the lack of contextualisation may cap bi-encoder performance. This difference confirms that fine-grained contextualisation is important for dialogue representations.

Dialogue structure: the importance of [eos]. The effect of special tokens like [eos] on the behaviour of large LMs is an active research area. Some works have found such tokens to have detrimental effects on the generalisability of generative models (Newman et al., 2020), but many recent models for dialogue representation rely on these tokens to capture structural characteristics of dialogue (Gu et al., 2020; Wolf et al., 2019).

Comparing the cross-encoder performance with and without [eos] in Table 4.3 shows that explicit turn information is extremely valuable for response selection.

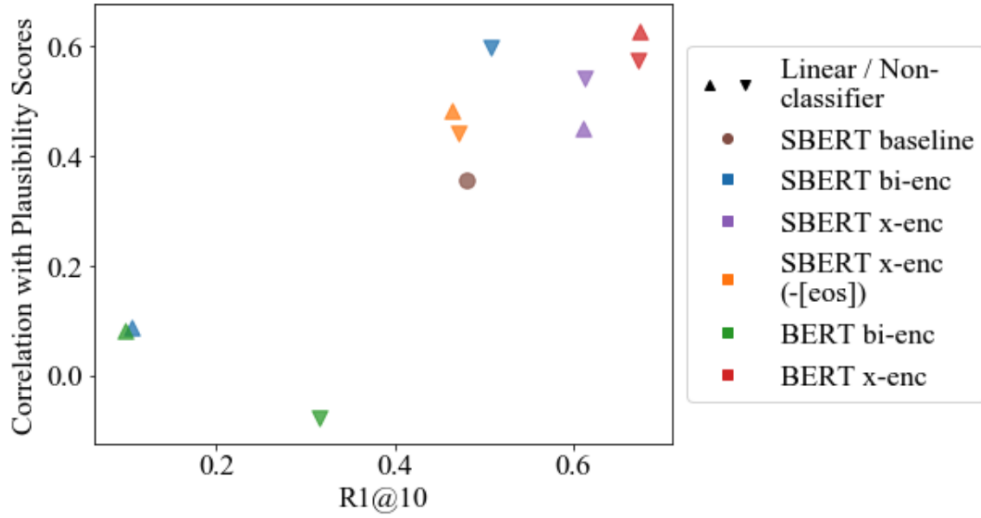


Figure 4.3: The relationship between response selection accuracy and psychometric predictive power for the variants of bi- and cross-encoders developed throughout this chapter.

Fine-tuning. The importance of pre-trained language models in NLP can’t be understated; however, spoken dialogue is fundamentally different to the monological text used for pre-training. Similar to previous work investigating the importance of fine-tuning with dialogue data, we find that pre-trained BERT representations do contain some useful information for response selection, but that fine-tuning towards the target task is still important (Noble and Maraev, 2021). In addition, we find that fine-tuning is much more beneficial when dialogue structure information is provided explicitly via the *[eos]* token.

4.4.4 Relationship between response selection and predictive power

Given the range of response selection results from our ablations, we consider the relationship between response selection performance and correlation with human plausibility judgements (PPP). We find a strong correlation between the two: $\rho = 0.84$. This relationship is marginally stronger if a larger range of response selection accuracies is considered; comparing response selection performance measured as $\sum_{k \in \{1,2,5\}} [R10@k]$ and PPP produces a correlation of $\rho = 0.86$. To ensure this correlation is not dominated by human response selection ability, we also compute this correlation across only negative responses and find a similarly strong correlation between model response selection performance and human scores ($\rho = 0.82$).

The strong correlation suggests that response selection performance can be used as a coarse proxy for how well a model may correlate with human plausibility judgements. However, response selection performance patterns (SBERT cross-encoding outperforming bi-encoding, for example) are not reflected in this relationship. Although the modelling capacity of the classifier (i.e., linear versus non-linear) did not have much effect on the response selection performance of cross-encoders (see pairs of points grouped around $R_{10}@1$ of 0.47, 0.61, 0.67 in Figure 4.3), the impact on PPP is larger.

4.4.5 Discussion

Using the task of upcoming dialogue turn prediction, we presented a benchmark for human performance on the widely-used task of response selection: people are able to do this task, but not perfectly. To the best of our knowledge, this is the first perceptual validation of response selection for spoken dialogue transcripts. Urbanek et al. (2019) similarly measure human response selection using stimuli from written dialogues grounded in a virtual world. However, they use sets of 20 responses for each stimulus and note that this large number made it difficult to obtain reliable judgements from participants. The rating paradigm proposed here avoids these issues, provides additional information about judgement certainty, and is removed from the number of negative samples.

Human performance showed that a fundamental assumption of response selection—that a single “correct” upcoming turn can be predicted from previous context—does not reflect the reality of perception with respect to dialogue acceptability. This result supports findings in cognitive science regarding the perception of linguistic acceptability: rather than being deterministic, this judgement is intrinsically probabilistic (Lau et al., 2017; Chater et al., 2006).

We then used this quantification of dialogue acceptability perception to establish the psychometric predictive power of different language model families, finding that response selection performance can be used as a coarse proxy for alignment with this aspect of perception, but only to a certain extent. Strong correlation of BERT-FP scores and human plausibility ratings suggests that, as a training objective, the discriminative response selection task does align with some aspects of dialogue acceptability perception. However, this model achieved “superhuman” response selection performance. Qualitative differences in response scoring detailed in the error analysis also indicate

that this discriminative training paradigm doesn't capture the open-ended nature of dialogue acceptability perception. On the other hand, scores generated from TurnGPT produced a moderate correlation with plausibility judgements, but performed poorly at response selection. Urbanek et al. (2019) and Wu and Xiong (2020) both report similar findings regarding the response selection performance of generative models compared to retrieval-based models. We conducted preliminary analysis of the differences between autoregressive and masked LMs, however given that language model architecture is known to significantly affect predictive power for other perceptual tasks (Wilcox et al., 2020), we believe this to be a fruitful avenue for further investigation into what predictive mechanisms are most explanatory for human judgements.

Finally, to investigate the strong correlation between the cross-encoder-style BERT-FP scores and perceptual judgements of upcoming dialogue turns, we ablated aspects of the model. In particular, we examined the importance of contextualisation in dialogue representations, the value of including turn-taking structure, and whether pre-training is beneficial. Our bi-encoder models provided a small amount of dialogue structure information, however, even our best-performing bi-encoder only marginally outperforms our baseline of simply measuring the distance between SBERT embeddings for response selection. This may suggest that this method of encoding dialogue structure is too weak. Cross-encoding, on the other hand, provides much more granular contextualisation. We found cross-encoders to be much less sensitive to the capacity of the classifier than bi-encoders. Our cross-encoder outperformed the SBERT baseline and bi-encoders by a large margin when dialogue structure is explicitly encoded in the input (as *[eos]* tokens). Inclusion of dialogue structure information and fine-tuning towards response selection boosted response selection accuracy by over 30 absolute percentage points, highlighting the value of turn structure information for response selection models.

4.5 Conclusions

This study has allowed us to draw further conclusions about the nature of expectations in human dyadic communication, as well as the degree to which existing machine representations of dialogue turns align with this component of perception.

Perception Compared to previous studies of acceptability perception which tend to focus on grammatical acceptability, our proposed dialogue turn acceptability task

places a much greater emphasis on higher-level reasoning about semantic and pragmatic factors. The analysis of human dialogue turn plausibility judgements presented in the preceding chapter shows such acceptability judgements to be similarly probabilistic. Framing the dialogue turn acceptability judgement task instead as a discrimination task provides one of the first benchmarks on human response selection ability as well as a novel perspective on this probabilistic feature of perception: the accuracy levels across participants clearly demonstrate that context constrains expectations about the upcoming turn, however only to a certain extent. Participants’ expectations are not always correct and often multiple upcoming turns can be considered acceptable for a given context.

Alignment of machine representations The degree to which language models align with human language perception—their *psychometric predictive power*—has so far been constrained to the perception of monological discourse in the literature. In this study, we extend the evaluation to the perception of dialogues by asking whether different styles of language modelling are predictive of dialogue turn acceptability judgements. Though we find a strong correlation between plausibility scores from humans and models, there are two major caveats.

First, models vary widely in their predictive power. We find retrieval-based models to far outperform their autoregressive counterparts in terms of correlation strength. Our ablation studies demonstrate that features of cross-encoders which facilitate fine-grained contextualisation are important for alignment with human acceptability judgements. However, it remains unclear whether the divergence between autoregressive and retrieval-based models is a function of the different underlying modelling strategies or the power of aggregation functions used to compute utterance-level scores. The aggregation strategies used to compute utterance-level scores from TurnGPT are far more rigid than the implicit aggregation occurring in retrieval-based models. Along with the increased predictive power from combinations of aggregations shown in Chapter 3, the marked difference in the predictive power of these models further supports the conclusion that utterance-level surprisal benefits from modelling strategies that aim to explicitly encode information at higher levels of abstraction than individual token prediction. In related work, we have similarly shown that quantifying the predictability of an utterance relative to a set of plausible alternative utterances is a stronger method for predicting dialogue turn acceptability judgements than token-level surprisal aggregates (Giulianelli et al., 2023). Additionally, we show that this predictability measure

is complementary to surprisal for predicting other psychometric measures such as eye-tracked reading times.

Second, even for the retrieval-based models that achieved the strongest correlation with human scores, our qualitative analysis revealed fundamental differences between human acceptability judgements and model scores. Models and people make very different “mistakes” in the discriminative task. They also seem to maintain different degrees of optionality in their predictions—although scores from TurnGPT were less strongly correlated with human judgements, the distribution of scores across stimuli was a closer match to human scores than those from retrieval-based models. Though discriminative training objectives and evaluation metrics are valuable tools, they ignore the degree of optionality in predictions, which we are finding to be an important component in human judgments of contextualised language-use.

Part II

Expectations in multi-channel communication

Part II Introduction

Part I provided insights into the shape of expectations *for the lexical content* of dialogue: although conversational context constrains the space of possible trajectories, people maintain a range of potential continuations and their anticipations do not always align with how the conversational actually develops. Compared to surprisal theory, which considers how likely an individual continuation may be, studying human language comprehension in this way provides an alternative perspective on predictive language processing by characterizing the distribution of expectations themselves.

In Part II, we propose measuring the perceptual value of lexical and non-lexical channels by how much they alter human expectations regarding upcoming spoken dialogue. As discussed in Chapter 2, prosody is widely accepted to contribute to a number of important communicative functions during spoken interaction (Cohen, 1984). However, the majority of works studying the perception of prosody rely on heavily constrained or contrived stimuli and examine its role in specific communicative functions, making it difficult to draw generalizable conclusions. On the other end of the spectrum, speech synthesis research takes a much coarser approach to studying the effect of prosody, often evaluating the “naturalness” of synthetic speech and almost exclusively considering speech samples in isolation⁵. The nebulous definition of “naturalness” similarly hinders the scope of conclusions that can be drawn regarding the role of prosody (Shirali-Shahreza and Penn, 2018). The lack of context in such evaluations also disregards ample evidence that the subprocesses used to make sense of real language are not apparent during the processing of isolated stimuli (Marslen-Wilson and Welsh, 1978; Willems and Peelen, 2021; Gutierrez et al., 2021; Brown et al., 2022). We aim to strike a balance between examining the perception of prosody with respect to narrow communicative functions and the broad concept of “naturalness” while using natural speech embedded in its communicative context.

To quantify the added value of prosodic information in constraining expectations in natural spoken dialogue, we propose the *turn discrimination paradigm*. Inspired by recent developments in self-supervised contrastive machine learning, we modify the turn acceptability task from Part I to model participants’ use of lexical and non-lexical channels as a function of their ability to discriminate between upcoming utterances. We begin in Chapter 5 by introducing our behavioural paradigm in a highly

⁵Works by Clark et al. (2019); O’Mahony et al. (2021) have begun to demonstrate the value, and complexity, of much-needed evaluation methods for generated speech in context.

constrained experimental setup to isolate the contribution of prosodic information to perceptual judgments. In Chapter 6, we use the models analysed in Part II to present a more general version of the paradigm that allows us to quantify the effect of prosodic information over a much broader set of language and propose several metrics for quantifying the perceptual information value of the two channels.

Chapter 5

Perceptual information value of the non-lexical channel

In this chapter, we discuss the work published in

Wallbridge, S., Bell P., and Lai, C. (2021) *It's not what you said, it's how you said it: discriminative perception of speech as a multi-channel communication system*. In Proceedings of Interspeech, (2386–2390)

where we present our first steps towards answering both of the following questions:

RQ: “*How do the lexical and non-lexical channels of speech affect expectations about upcoming spoken dialogue?*”

RQ: “*How should the Perceptual Information Value of channels be quantified?*”

We motivate, introduce, and validate the turn discrimination paradigm which allows us to explore how different aspects of context affect the shape of expectations in spoken dialogue. In particular, we focus on isolating the effect of non-lexical information by comparing participant judgements when they are given access to spoken and written context.

5.1 Related work

Given what has been discussed so far in this thesis regarding the predictive nature of language comprehension, the idea that the prosodic realisation of previous turns impacts our expectations about what message a speaker may be likely to convey in the next turn of dialogue is rather intuitive. However, this effect has yet to be demonstrated and quantified directly. In Section 2.1, we introduced evidence for the effect of prosody

on spoken language comprehension; we now discuss some of the existing evidence—both behavioural and modelling—for the effects of prosody on the comprehension of spoken *dialogue*, and how such studies motivate the design of our experimental paradigm.

A well-studied aspect of prosody in dialogue is *entrainment*, i.e., the tendency of conversational partners to become more similar to each other (cf. alignment or convergence). This phenomenon has been studied on both lexical and non-lexical dimensions and has been found to have significant cognitive and social effects, facilitating comprehension and prediction of upcoming speech, as well as the perception of rapport, naturalness, and communicative success (Stephens et al., 2010; Levitan et al., 2011; Lubold et al., 2015). These findings provide evidence that the non-lexical channel carries valuable information about the semantic/social context. Most research in this area has examined behavioural similarity over entire conversations; however, studies have shown that alignment changes can produce perceptual effects by conditioning on very localised information: Lubold et al. (2015) finds significant impact on human perception of synthetic dialogue when generated speech is conditioning on the realisation of the previous speaker’s turn. Previous turn features have also been found to be predictive of acoustic properties of upcoming turns (Fuscone et al., 2020). We extend this direction through our application of the turn discrimination paradigm by measuring the effect and utility of varying degrees and types of information on human perception.

Beyond perceptual studies, numerous works have shown that non-lexical features and discourse context can be exploited for modelling different aspects of spoken discourse structure. For example, acoustic correlates of prosody can help identify paragraph structure (Lai et al., 2020), rhetorical relations (Kleinhans et al., 2017; Murray et al., 2006), turn-taking behaviour (Gravano and Hirschberg, 2011; Roddy et al., 2018), and dialogue acts (Shriberg et al., 1998; Tran, 2020). Alongside these relatively well-defined functions, prosody has also been shown to affect the perception of dialogue at a more generic level of *discourse coherence*. This concept can be broadly defined as characterizing how logical and consistent the internal structure of a multi-utterance segment is (Cervone et al., 2018), and is thus more closely related to our framing of perception as prediction. Discourse coherence has primarily been used to assess the quality of texts, most notably by representing documents as distributions of entities appearing in adjacent sentences (Barzilay and Lapata, 2008), however, it has recently been used in both the evaluation and training objectives for generative models of written and spoken dialogue (Li and Jurafsky, 2017; Malkin et al., 2022).

Such studies provide evidence that additional features such as dialogue acts, some of which may capture prosodic information, outperform purely lexical models (Cervone and Riccardi, 2020; Mesgar et al., 2020). These works demonstrate that modelling varying degrees of non-lexical context, both in isolation and in conjunction with lexical content, can enhance the automatic modelling of human speech understanding. In other words, they provide further evidence that we should expect prosodic cues to modify human expectations regarding upcoming dialogue turns.

The works discussed here demonstrate that local information affects different aspects of spoken dialogue perception, however, any number of prosodic realisations can still be likely when the context is underspecified (Hodari et al., 2020). To design a task that captures this probabilistic aspect of expectations across the communicative channels in speech, we draw inspiration from discriminative training schemes in self-supervised machine learning. Generally, such schemes involve training a model to discriminate between target and distractor samples. These methods have been extremely successful primarily because they leverage contextual information as a training signal rather than requiring explicit labels (Milde and Biemann, 2018). Using a discriminative task enables us to investigate the factors that drive expectations of prosody while avoiding reliance on human-generated labels.

5.2 Turn discrimination paradigm

The turn discrimination paradigm enables investigation into how listeners use both lexical and non-lexical contexts to generate their expectations about the upcoming communicative signal. In more practical terms, we define the future communicative signal as the upcoming turn in a dialogue. We examine expectations over both its lexical content and prosodic realisation, and how they are shaped by both lexical and prosodic context. In the preceding section, we describe the core components of the turn discrimination paradigm: the task itself and the conditions across which participant responses are compared.

5.2.1 The discriminative task

The stimuli for this task involve a context made up of previous turns in a conversation ($C = [T\{-k\}, \dots, T0]$), and a set of target response turns ($\{T1\}$), one of which is the true response for this conversation. The remaining distractor responses are sampled

Channel condition	Description
<i>Lexical only</i>	Context and response turns are presented as only transcripts
<i>T0 Lexical</i>	Context is presented as a transcript; response turns (T1) are presented as recordings
<i>T0 Prosody</i>	Context is presented as a transcript; response turns (T1) are presented as recordings joined with the recording of the final context turn (T0)

Table 5.1: Descriptions of how stimuli presentation varies in each of the different channel conditions of the turn discrimination paradigm.

from elsewhere in the corpus; the following chapters in part II involve experiments that employ different strategies for sampling the distractor responses.

Regardless of the sampling strategy, participants are explicitly instructed that one of the responses in the set of potential continuations is the correct response that occurred in the dialogue context and then asked to rate each response by how likely it is to be the true continuation, giving a score from 1 (‘Very Unlikely’) to 4 (‘Very Likely’) as depicted in the example of stimuli presentation in Figure 5.1. We choose a rating task as opposed to ranking to capture a more nuanced perspective on participant expectations such as relative certainty.

5.2.2 Conditions

To compare the effect of lexical and non-lexical information on participant responses, we present stimuli in a number of conditions where different information is present. For this, we modify both the *amount* of context and the *channels* of context (i.e., whether the context consists solely of lexical content or of the full speech recording).

In all conditions, participants receive the transcript of the context turns. In the *lexical-only* condition, all potential responses are presented as transcripts. Otherwise, responses are instead presented in an audio format. Participants either receive the audio of each T1 response (*T0 lexical*), or recordings of each response spliced onto the recording of the final context turn (*T0 Prosody*). The pause—or lack thereof—between speakers’ turns is an important feature of natural dialogue. It is known to signal intention, emphasis, and perhaps syntactic complexity (Esposito et al., 2008; Rühlemann

Consider the following segment of a phone conversation (you will hear the sections in underlined bold):

...

lawyers sue if you put a ten thousand dollar cap on the amount of money that a lawyer could make from a law suit against a doctor

yeah

they wouldn't sue a one of them

well that's probably true too

you know

but people are very conscious of how much money they can get real fast these days

well sure but you know when you think about reasonableness now i believe that a law suit should have some foundation in th[e]- the tort theory you know what tort theory is

and so

no

...

Only one of the following recordings is the **true** segment in this conversation. Assign each audio segment a score of how likely it is to be the **true** segment.

(You can listen to the recordings as many times as you need to make your judgements.)

	Very unlikely	Somewhat Unlikely	Somewhat likely	Very likely
▶ 0:00 / 0:03 ———— 🔊 ⋮	☐	☐	☐	☐
▶ 0:00 / 0:03 ———— 🔊 ⋮	☐	☐	☐	☐
▶ 0:00 / 0:03 ———— 🔊 ⋮	☐	☐	☐	☐
▶ 0:00 / 0:03 ———— 🔊 ⋮	☐	☐	☐	☐

Figure 5.1: A screenshot of the stimuli presented to participants in *T0 Prosody* condition, where the context consists of transcripts of past and future turns, as well as the acoustic realisation of the preceeding turn T0 (underlined, in blue bubble).

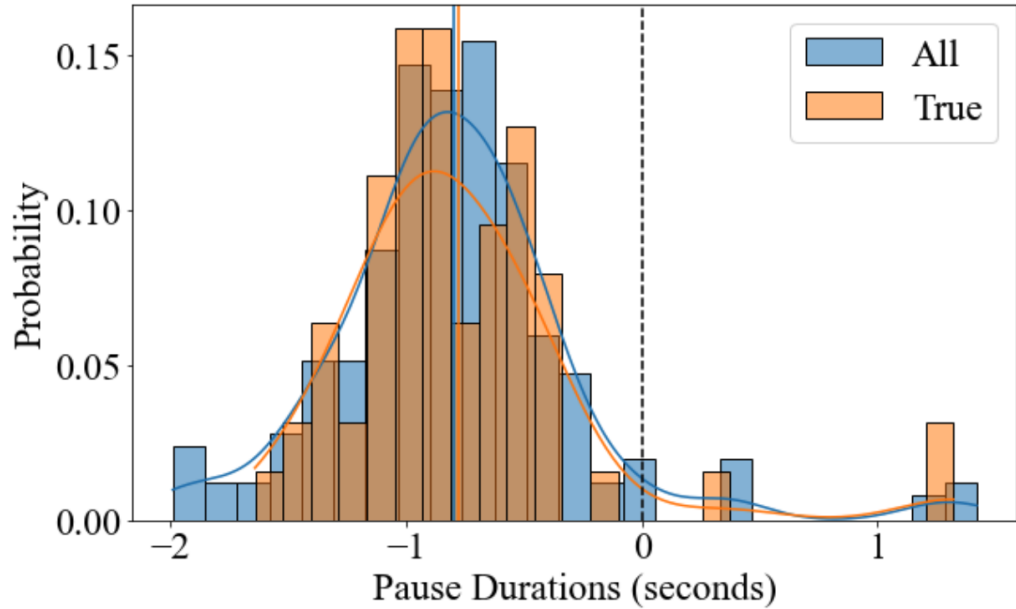


Figure 5.2: The distributions of pause lengths for each response in our stimuli set, showing the subset of true responses. As expected, there is no difference in the distributions of all responses and this true response subset.

et al., 2011; Gorisch et al., 2012; Bögels et al., 2015; Stokoe, 2018)¹. To construct our stimuli, we chose to model pauses as a feature under the control of the speaker; each response was thus extracted along with its preceding pause which was used to join it to the context audio. Note that this defines pauses as both positive and negative lengths (i.e., overlapping). Although the average pause between conversational turns is often quoted as 200ms (Ten Bosch et al., 2005), Figure 5.2 demonstrates that the responses in our stimuli are heavily skewed to overlapping turns. Overlap is a well-known characteristic of backchannel responses, further highlighting the proportion of such instances in our stimuli set (Levinson and Torreira, 2015; Sacks et al., 1978).

5.3 Experiment design

In this chapter, we focus on isolating the effects of the non-lexical channel and present the turn discrimination paradigm in a constrained setup where each of the potential

¹Outside of academic research, their importance is also reflected by the transition from “dialogue” to “conversations” in filmmaking. Replacing an orderly series of turns with overlapping speech was a revolution in scripting more realistic conversations (Jaekle, 2013). Although it was already regular practice in Broadway stage comedies, its introduction is often credited to Orson Welles in the 1940s with films like *Citizen Kane* (Welles et al., 1998).

responses in a stimulus has the same lexical form. The only variation that participants can leverage for discrimination is in the prosodic realisation of responses.

This constraint illuminates the unique information in the non-lexical channel and enables an exploratory study into the characteristics of communicative context, across both lexical and non-lexical channels, that affect human expectations of prosodic realisation.

5.3.1 Data & processing

Experiments were carried out on the Switchboard Telephone Corpus (Godfrey et al., 1992) which we introduce and motivate in Chapter 3. Briefly, the dataset consists of over 2,400 conversations between pairs of strangers across America who have been supplied with a prompt to discuss. There are 542 participants in total.

The subsequent 260 hours of speech have been manually transcribed and split into segments.

Transcription. The original Switchboard transcription guidelines contain a number of dataset-specific annotations for features including pronunciation variants, partial words, non-speech noises such as laughter and coughing, and disfluencies including false-starts, speech errors, and filled pauses (Hamaker et al., 1998). Although at first glance, the distinction between lexical and non-lexical channels seems clear, the waters become quickly muddled in the presence of such features (Markl, 2023). We treat those listed above as part of the non-lexical channel and remove them from the transcripts. However, filled pauses such as “uh” and “um” have been widely accepted as important cues in conversation (Clark and Tree, 2002; Fraundorf and Watson, 2011). To avoid our evaluation of prosodic information being skewed by these specific cues, we opt for the more conservative channel split and retain filled pauses in our transcripts.

Take the following example of a transcription excerpt:

“We w[ent]- went to the [prace/place] where we uh we [laughter] found out about_1 this”

where “w[ent]-” is a partial word, “[prace/place]” shows a mispronunciation of the word “place”, and “about_1” is a pronunciation variant of the word “about”. The cleaned transcript would read as:

“We went went to the place where we uh we found out about this”

Segmentation. Unlike written text where sentences can be identified systematically, utterance segmentation of spontaneous dialogue is characterized by overlaps, disfluencies, and restarts. Carletta (1996) suggests that it is one of the most challenging elements of discourse annotation. Discourse segments are often defined based on the pragmatic structure of discourse, making the unit of interest subjective and often task-dependent. Switchboard segmentation is the result of multiple segmentation schemes, both manual and automatic, which split both speaker-sides of each conversation into utterances (Deshmukh et al., 1998; Calhoun et al., 2010). We use these segmentations as the resulting utterances contain less overlap and are generally shorter than conjoined speaker turns which allows for a greater number of lexically-equivalent utterances.

5.3.2 Stimuli construction

An individual stimulus in this task consists of a context and 4 lexically equivalent responses, one of which is the true continuation. To generate such stimuli, all speaker utterances were first filtered to ensure that the resulting stimuli would be of suitable length and quality for a behavioural study: a response must have a preceding turn containing between 4–25 word tokens and the absolute length of its preceding pause must be < 2 seconds. Acceptable response turns were grouped by lexical content, discarding utterances with less than 4 occurrences. The longest utterance with at least 4 occurrences was 5 tokens long. From these sets, we sampled 50 stimuli to obtain a uniform distribution of response token lengths (i.e., between 1–5 tokens).

Given that the responses in each stimulus were equivalent in their lexical content, the performance in the lexical-only channel condition will be at chance level. We present participants with the two acoustic channel conditions (*T0 Lexical* and *T0 Prosody*; detailed in Table 5.1) with different amounts of non-lexical information presented. In addition, we examine the effect of different degrees of lexical context by presenting different numbers of previous and following dialogue turns in both channel conditions; we consider a single previous turn (1,0), 4 preceding turns (4,0), 7 preceding turns (7,0), and 4 preceding and 3 following turns (4,3).

The final experimental set-up thus consisted of 50 stimuli across the 8 conditions, resulting in 400 questions in total.

5.3.3 Online experiment

67 participants were recruited using Prolific Academic, all were native English speakers and based in North America. Participants were presented with 20 stimuli through a Qualtrics survey; the average time to complete the survey was ~ 25 minutes. Each participant received a random selection of stimuli from the full set of context and format conditions, with the constraint that a participant could only receive a particular sample in one condition. Participants weren't allowed to complete multiple surveys, ensuring independent responses from different participants for each question. Additional check questions were interspersed throughout each survey as a basic attention check: select the gender of the response speaker in the previous question. The responses of participants who obtained less than 70% overall accuracy on the set of check questions were removed from our analyses.

5.3.4 Evaluation metrics

The metrics we use to compare responses are based on discriminative performance—how well participants can distinguish the true upcoming response from the negative samples. In this study, we use accuracy and a cross-entropy-based metric.

Accuracy. A measure of the frequency with which the true sample was rated highest across participants. We normalise by the number of samples sharing the highest score. For each set of ratings, an accuracy score was computed as:

$$accuracy = \begin{cases} \frac{1}{|\text{max-scoring responses}|} & \text{if correct response has max score} \\ 0 & \text{otherwise} \end{cases} \quad (5.1)$$

Cross Entropy $H(P, Q)$. A measure of deviation between probability distributions P and Q .

$$H(P, Q) = - \sum p(x) \log(q(x)) \quad (5.2)$$

We apply cross entropy by converting the “ideal” scoring (i.e., where the true response receives the maximum score (4) and all other samples, the minimum score (1)) and participant scores to probability distributions. Given that a wide range of prosodic realisations can still be likely, particularly in circumstances where the context is under-specified (Hodari et al., 2020), this metric provides a more nuanced perspective on changes to participant ratings by capturing information about certainty. We compared

Channel condition	Context Condition			
	(1,0)	(4,0)	(7,0)	(4,3)
<i>T0 Lexical</i>	0.346	0.329*	0.349*	0.358
<i>T0 Prosody</i>	0.400	0.444*	0.468*	0.378*
<i>Random</i>	0.250	0.250	0.250	0.250

Table 5.2: Mean participant performance as measured by **accuracy** as well as the random baseline of chance performance. Statistical significance between conditions is computed using a Welch’s *t*-test; pairs of conditions that obtain a difference greater than 0 with a *p*-value > 0.5 are denoted by matching symbols (*, *, •).

the set of ratings across each response set of our stimuli to a vector representing the correct scoring—where the correct response received the maximum score (4) and all other samples, the minimum score (1). A cubic mapping was applied to the two score vectors before they were converted to probability distributions P and Q to produce peakier distributions.

5.4 Results & discussion

We compare all combinations of context and channel conditions to the lexical-only baseline which, given the constraint of lexical equivalence, would be equivalent to chance performance from participants. We model two representations of this **lexical baseline** condition:

1. Accuracy of the uniform probability of selecting the true sample.
2. Cross entropy obtained by shuffling each set of human ratings 50 times, i.e., maintaining the distribution of scores.

The significance of score differences is computed for 95% confidence intervals around estimated means.

5.4.1 Perceptual judgements

The results of Chapter 3 demonstrated differences between the expectations regarding upcoming turns in written dialogue compared to transcripts of spoken dialogue, reflected in participants’ ability to discriminate the true upcoming turn from distractors.

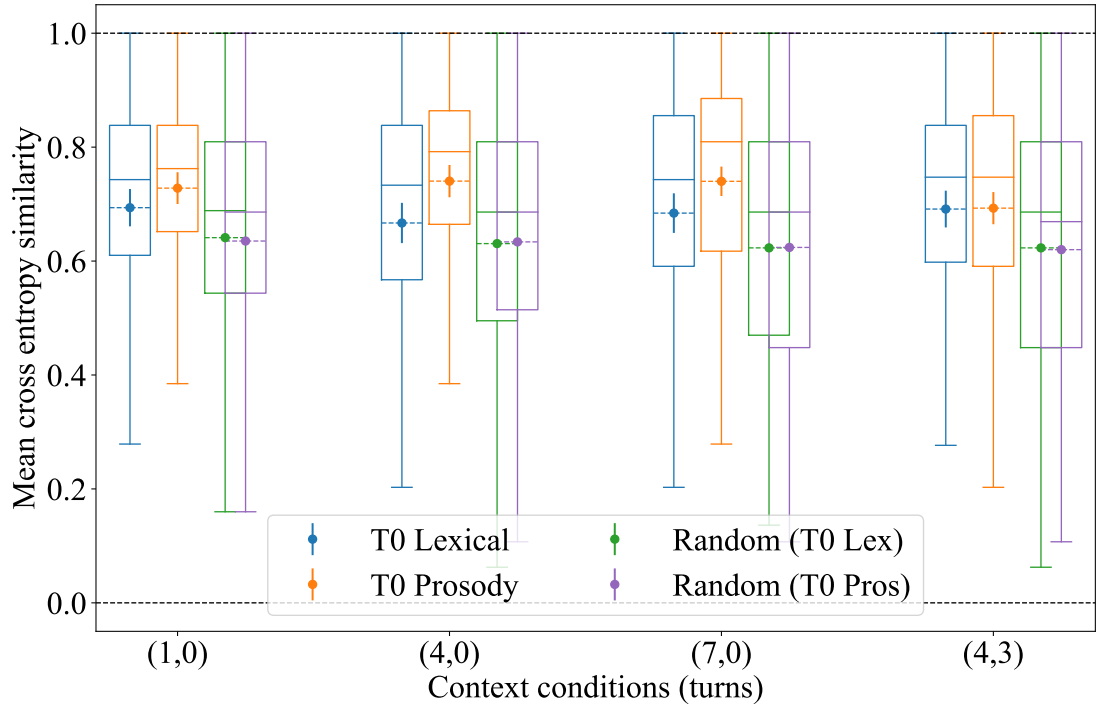


Figure 5.3: Average **cross entropy similarity** across context conditions. Box plots correspond to responses in both formats as well as the random baseline where distributions were computed by shuffling the scores from both formats. Higher values indicate that participant judgements are closer to the “ideal” scoring.

We presented a potential explanation: spoken dialogue transcripts are a lossy representation of the messages actually produced. Accuracy should thus increase when participants have access to fuller representations of the message and its corresponding context. Both accuracy and cross-entropy performance, displayed in Table 5.2 and Figure 5.3, respectively, demonstrate that participants are able to discriminate the true prosodic realisation at a rate that is significantly greater than if they were only provided with the response transcript (i.e., all variations were equally likely). This result holds over both channel conditions—with a mean accuracy of 0.423 and 0.345 when participants did and did not have access to prosodic context, respectively—and across all context conditions. Thus, we provide firm evidence for the role of non-lexical information in expectations: even with only the transcript of a preceding speaker turn to condition their judgements on (*T0 Lexical*), people can make effective predictions regarding non-lexical aspects of upcoming dialogue turns.

Previous works have drawn conflicting conclusions regarding the role of prosody in

potential predictive components of language comprehension: (Cutler, 1976) suggests that listeners use preceding intonation contour to predict upcoming stress while (Mens and Povel, 1986) suggests that the temporal structure of accents in spoken utterances—rhythm—doesn't affect processing of the accented items.

We begin with the first channel condition. With access to response audio but only transcripts of the previous conversational turns (T0 Lexical), participants achieve a mean accuracy of 0.345 across all context conditions. This is significantly better than if they only had the response text (i.e., all variations equally likely), which is also reflected in cross-entropy performance. Even with only the transcript of a preceding speaker turn (first column, first row of Table 5.1), people can effectively discriminate between different prosodic realisations of the same response. Interestingly, mean participant performance across the context conditions was relatively consistent, showing no significant differences. The fact that additional lexical context didn't significantly alter participant discriminative ability indicates that information in the lexical channel may have little value in constraining expectations of upcoming prosodic realisation.

In the second channel condition (T0 Prosody), participants have access to the audio of the response as well as the immediately preceding turn. Similar to the T0 Lexical format, accuracy and cross-entropy performance both reflect a significantly better ability to discriminate the turn response from discriminators compared to the random baselines, with a mean accuracy of 0.422. This provides further evidence for the value of the non-lexical channel. Except for the “future” context condition (4,3), additional text context yields slight performance improvements. This trend is similar to the T0 Lexical condition; the fact that additional contextual information doesn't affect human performance could indicate that either participants were not making use of the information in the additional context, or that the information it provides isn't useful. We discuss this below.

When we compare discrimination ability between the two channel conditions, both evaluation metrics consistently show better discriminative ability for T0 Prosody compared to T0 Lexical. That is, access to prosodic realisation of context constrains dialogue expectations beyond the lexical context. In all context conditions, the addition of prosodic context provides a greater boost to performance than future lexical context. The differences in performance between the channel conditions are significant when participants can condition their scores on transcripts of previous speaker turns ((4,0), and (7,0)). However, we do not find significant differences for extreme context conditions (1,0) and (4,3). The non-significance in the (1,0) condition could simply be due to

insufficient context. However, the fact that participants do significantly worse in condition (4,3) where they receive context from before and after the target response is more surprising. Previous work found that future information is more predictive of macro acoustic features of dialogue markers than past information; Nath and Ward (2022) use a number of acoustic features as representations of local prosody both preceding and proceeding discourse markers like “really” and “uh-huh” to predict the mean and maximum log energy, fundamental frequency, cepstral flux (as a measure of lengthening and reduction), and harmonic ratio of the marker. They find that combining past and future context is more predictive than past context alone. Discourse markers are known to play specific roles in spoken conversation, and carry a high proportion of information in their prosody compared to other dialogue acts with more lexical content (Fraser, 1999). Given that our stimuli contain a high proportion of such responses, we may expect a similar trend. If instead, information about the past and future state of a conversation is equally useful for this task, performance in context condition (4,3) should be similar to the (7,0) condition as participants have access to the same number of turn transcriptions; if the future information has no value, we could expect similar performance to (4,0). However, cross-entropy performance of participants in context condition (4,3) is significantly worse than both (4,0), (7,0) when participants can hear the context turn. A potential explanation for this finding are the cognitive load or interference involved in this condition: people may struggle to integrate information spread across both textual and acoustic formats as well as preceding and proceeding context.

5.4.2 Certainty analysis

Though discrimination performance increased slightly with the amount of textual context in the T0 Prosody condition, the differences are not significant. This consistency could indicate that either people are not using this additional context, or that they are but it is not reflected in their discrimination performance.

To probe whether participants used the additional lexical context, we examine the distribution of scores with respect to certainty across context conditions. Table 5.3 shows slight differences in the degree of certainty expressed by participants across channel and context conditions. Participants became more certain of their scores as additional context was provided in both T0 Lexical and T0 Prosody, with the trend being more pronounced in the latter. This indicates that participants’ expectations were affected by additional context, even if the performance across conditions was not

Channel condition	Context condition	% Certain (1,4)	% Uncertain (2,3)
<i>T0 Lexical</i>	(1,0)	39.22	60.78
	(4,0)	41.19	58.81
	(7,0)	41.13	58.87
	(4,3)	43.79	56.21
<i>T0 Prosody</i>	(1,0)	39.08	60.92
	(4,0)	43.79	56.21
	(7,0)	44.57	55.43
	(4,3)	43.67	56.33

Table 5.3: Proportion of certain versus uncertain ratings across different channel and context conditions.

significantly different. Interestingly, participants were slightly more certain in the text-only condition with future context, even though they performed closer to the random baseline in that condition.

5.4.3 Acoustic analysis

Both discriminative metrics proposed here converge on the conclusion that participants leverage prosodic information to constrain their expectations of the upcoming communicative signal. The next question is then, how are listeners using prosodic information to drive their expectations?

One hypothesis is that similar-sounding responses are scored similarly. We test this hypothesis at the level of macro acoustic features, using a range of features commonly used to characterise prosody in dialogue (Lai, 2009; Fuscone et al., 2020). These include aggregate statistics (mean, standard deviation, slope, min, max, and range) for F0 and intensity across the response, as well as duration features which include the length of the pause preceding the response as well as its overall duration. F0 features are based on a pitch track obtained using the SWIPE pitch estimation method implemented in `pysptk.swipe` (Camacho and Harris, 2008). The F0 contour of each utterance in our acoustic analysis is normalised for its labelled gender, where the average F0 for each gender is based on the manually-labelled gender of all utterances used in our stimuli set. The voiced segments of the normalised contour are converted to semitones before applying aggregation functions; slope estimates are obtained from the coefficient of

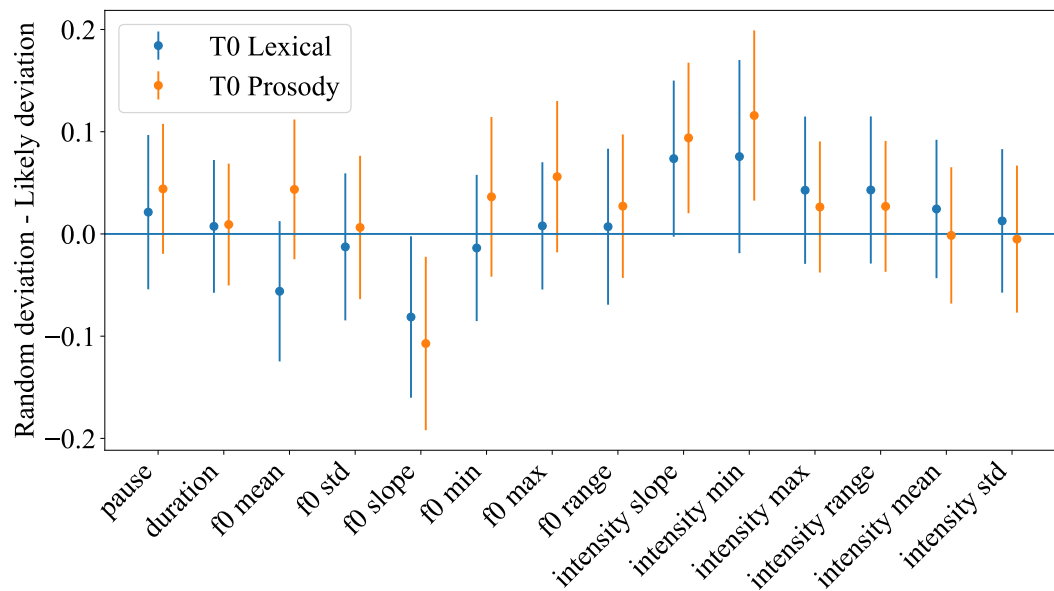


Figure 5.4: Comparing the average deviation between responses scored as *likely* and all responses in a stimulus for macro acoustic features; error bars correspond to 95% confidence intervals for the estimated mean across stimuli. A difference of 0 means the deviation between *likely* responses in a stimuli is the same as the deviation across all responses for that stimuli. Deviations across both channel conditions are shown.

the line of best fit. Intensity features are obtained from the `surfboard` python package which includes automatic computation of the aggregates (Lenain et al., 2020).

We consider whether responses that are scored similarly are also similar under each of these acoustic dimensions. For each set of stimuli ratings where ≥ 1 response is rated as likely (i.e., a score of either 3 or 4), we measure the mean pairwise deviation between the acoustic features of each response. This is the deviation we would expect if participants scored samples at random. We can compare this to the mean deviation across responses scored as likely. The difference in deviations is scaled by the random mean to compare relative differences between features. Under this metric, a difference of 0 implies that *likely* responses are no more related than responses rated randomly. Differences > 0 or < 0 , however, indicate that samples rated as likely are more (or less) similar in acoustic space than expected by chance.

Figure 5.4 shows that, on the whole, samples rated as *likely* tend to be slightly more similar in acoustic feature space than chance. However, we don't see significant differences between the relationship in acoustic space for the ratings obtained in the different channel conditions (i.e., T0 Lexical and T0 Prosody). Rather, these results highlight that the relationship between score similarity and acoustic similarity for these features can't be easily mapped. Similar complexities have been previously reported; for example, Fuscone et al. (2020) finds that different features of conversational context vary in their ability to predict aspects of the acoustic realisation of an upcoming turn.

Overall, this analysis indicates that we need to allow for variation in what constitutes "likely" prosody in a specific context and develop more flexible ways of describing what that context includes.

5.5 Conclusions

Perception In Chapter 3, we found that people's predictions about the upcoming lexical content in dialogue are constrained quite accurately by lexical context. The results presented in here provide firm evidence that people also build expectations about the prosodic realisation of upcoming turns and that these are a function of both lexical and non-lexical information. Our main conclusions, discussed further below, are as follows:

1. People can discriminate between true and false prosodic realisations of upcoming turns significantly better than chance.

2. Spoken context—the synergistic product of lexical and non-lexical channels—is more useful for increased discrimination performance than additional lexical context.
3. People can perform the discrimination task using only very local context.
4. Macro acoustic features of the prosodic realisations alone aren't sufficient to explain discriminative performance.

We find small amounts of local prosodic context has a larger impact than additional lexical context on participants' ability to discriminate the true response from distractors. Discriminative performance in both the T0 Prosody and T0 Lexical channel conditions (i.e., where participants do and do not have access to prosodic context) remained relatively consistent regardless of how much additional textual context was provided. Differences between the distribution of scores in Section 5.4.2 suggest that extended text contexts do in fact shape participant expectations, even if they do not always lead to more accurate expectations. This observation suggests that accuracy-based metrics are not sufficient for quantifying the perceptual value of information. Rather, quantification of perceptual information value should reflect optionality in expectations: numerous upcoming productions can be likely in a given context.

Even in this constrained set-up, our analysis of macro acoustic features highlights a complex relationship between the effects of lexical and prosodic information on expectations in spoken dialogue. In the set of acoustic features that we examine here, we do not find a clear relationship between acoustic realisation and plausibility rating of responses with the same underlying lexical content. Similarly, we found that additional lexical context had different effects on performance depending on whether prosodic context was also present. We found that performance increased with lexical context in the T0 Prosody condition, but not in the T0 Lexical condition. This could suggest that discriminative information in the lexical component of context is easier for people to leverage when they have access to enough non-lexical information; however, this is an open question that we leave for future work.

Paradigm The requirement for lexically-equivalent sets of responses allowed us to establish the utility of this paradigm for studying predictions based on the unique information in the lexical channels, and the synergistic information in the joint realisation of lexical and non-lexical information. However, it limited our stimuli set to short responses (recall that the longest response of high enough frequency was 5 words), many

of which were backchannels (e.g., “yeah” and “really”) which play specific roles in the management of spoken conversation by allowing the listener to provide feedback to the speaker (Ward and Tsukahara, 2000). Short responses like these are known to admit wide prosodic variation depending on their context (Benus et al., 2007; Lai, 2010). In Chapter 6, we remove the constraint of lexical equivalence to test whether these conclusions hold across a broader set of language and to study expectations of prosodic realisations and lexical content in conjunction with each other.

However, this lexically-equivalent paradigm offers valuable insights for the evaluation of text-to-speech (TTS) systems where lexical content is pre-determined. The goals of current speech generation systems have moved beyond producing intelligible speech towards the production of speech that is appropriate in a given context. However, evaluation methods have not kept up. In one of the first targeted studies of TTS evaluation in context, Clark et al. (2019) notes that traditional methods of evaluating TTS using naturalness ratings are not sufficient for this new objective. In a similar vein, O’Mahony et al. (2021) provides additional empirical evidence for the additional complexity of contextualised evaluation. In their paradigm, participants were presented with a context sentence followed by either a felicitous or infelicitous synthesized response. Although participants found canonical renditions to be more appropriate both in isolation and in context, all renditions were found to be more appropriate in context than in isolation. Developing contextually appropriate synthetic speech will almost certainly benefit from understanding the degree of optionality in expectations of prosodic realisations, and what aspects of context are responsible for licensing different prosodic realisations.

Chapter 6

Perceptual information value across multiple channels

This chapter consists of work presented in the following publication:

Wallbridge, S., Bell P., and Lai, C. (2023) *Quantifying the perceptual value of lexical and non-lexical channels in speech*. In Proceedings of Interspeech (2708–2712).

It involves a generalisation of the turn discrimination paradigm introduced in Chapter 5 as well as novel methods to quantify perceptual information value, both of which allow us to more fully answer the question:

RQ: “How do the different channels of speech affect expectations about the channels in upcoming spoken dialogue?”

And draw more concrete conclusions to the following question:

RQ: “How should the Perceptual Information Value of channels be quantified?”

6.1 Introduction

The constrained application of the turn discrimination paradigm presented in Chapter 5 clearly demonstrates that prosody of an utterance can shape human expectations regarding upcoming spoken communication. However, the requirement for lexically identical responses limited our analysis to short utterances often consisting of backchannel responses. In this chapter, our first goal is to relax this constraint—rather than using stimuli with lexically equivalent responses, we construct response sets where the lexical realisations are similarly plausible for the given lexical context. Building on our findings from Chapter 4, we use language models to achieve this. The

generalised version of the paradigm presented in this chapter provides the grounds to refine our definition of perceptual information value, particularly though exploring what metrics are appropriate for quantifying PIV.

6.2 Related work

In the previous chapter, we presented related works that evidenced how modifications to prosodic realisation can alter the perception of speech with the same lexical content, however, many effects of prosodic realisation are a function its *joint use* with lexical content. Wider evidence for interactions between lexical and prosodic information was presented in Chapter 2; here, we delve into the details of related studies that motivate our experimental design.

Previous studies that have demonstrated interactions between prosodic and lexical information often rely on carefully constructed, isolated stimuli produced in experimental conditions. For example, example, Bögel and Turk (2019) have recently shown that lexical features such as syntactic- and word-frequency impact the placement and strength of prosodic word and phrase boundaries in English. However, these materials are a far cry from conversational speech, which abounds with disfluencies, phonetic reduction, etc (Johnson, 2004). Others involve acoustic manipulation such as flattening pitch curves to remove changes in intonation or low-pass filtering to delexicalise speech (Bögels and Torreira, 2015; de Ruiter et al., 2006). These modifications may be conspicuous and leave certain acoustic properties such as duration intact. Our paradigm of sampling contexts and responses from existing conversational data circumvents both of these issues.

Although there has long been interest in the role of prosodic and lexical information in dialogue (Black, 1997), only a relatively small number of both perceptual and production experiments have explored the effect of prosody on the comprehension of spoken interaction. This is perhaps unsurprising given the difficulty of studying language perception in natural scenarios. However, it means that we do not yet have a cohesive picture of how lexical and non-lexical channels affect the perception of spoken dialogue. For example, both Tree and Meijer (2000) and Tzeng et al. (2019) show that informative prosody can be produced and used by listeners for syntactic and semantic disambiguation in dialogue. However, when the communicative need to do so is not present, speakers do not produce informative prosodic cues, and listeners may not use them even if they are present. Lai (2010) shows that intonation of cue words

like “yeah”, “right” and “really” affect their interpretation in dialogue contexts but that the effects depend on the lexical identity of the cue word.

Very similar to the work at hand is Bögels and Torreira (2021) which investigates the roles of prosody and conversational context for turn-end estimation. Using button-press experiments in two conditions where participants receive only written transcripts or the full speech signal, they find that phrase-final prosodic cues are relevant for turn-end projection. However, the subtleties of their results further reflect the complexity of processing—and studying—dialogue. Although acoustic context facilitates turn-end estimation for short utterances, they find an inhibitory effect of acoustic context on the accuracy of turn-end estimation for longer turns. Corps et al. (2018) also find that while listeners make use of acoustic discourse context to make turn-end predictions, access to relevant acoustic context yielded judgements which were produced earlier but not necessarily more precisely. In this study, we hope to contribute to understanding these seemingly surprising conclusions through an evaluation that accounts for the notion of optionality in human expectations.

6.3 Generalised turn discrimination paradigm & experiment design

We extend the turn discrimination paradigm to unbounded lexical content using findings in Part I to automatically sample plausible lexical turns using language models adapted to dialogue. The goal of this sampling was to achieve the same conditions as the lexically-constrained setting where all responses in a given stimuli are similarly plausible. This involved modifying the methodology for constructing stimuli, however, many other elements of the experiment design were unchanged. In particular, the underlying task described in Section 5.2.1 of rating potential upcoming responses remained the same, as did the conditions outlined in Table 5.1 where we compare the lexical-only condition, and fullest prosodic condition (*T0 Prosody*).

Details of the components that differ from the paradigm outlined in Chapter 5 are detailed below.

6.3.1 Data & processing

Stimuli for this experiment were constructed from the Switchboard Telephone Corpus. 642 conversations were annotated post-hoc with labels including dialogue acts, infor-

mation status, and prosodic features to form the Switchboard NXT subset (Calhoun et al., 2010). We use these conversations as validation and test sets (50% for each) for developing the language model to sample similarly plausible responses. All of our final stimuli are therefore sampled from this subset.

Turn segmentation

As described in Section 5.2.1, the potential responses for the lexically-equivalent set-up were sampled from the utterance-level segmentation provided by Switchboard. These constitute entire turns or turn fragments and thus provide a higher chance of responses occurring multiple times. In this experimental design, we segment all conversations into turn units based on speaker changes in conversation. We define speaker changes using the manually corrected word timings provided in Switchboard. We compute the timestamps for turn segmentation by first joining all words spoken contiguously by a speaker into segments. Completely overlapping segments are removed (i.e., overlapping frame sequences surrounded by mutual silence) before joining contiguous speaker segments into final turns. Though these interjections are important features of spoken conversation, in the context of training our dialogue LM, this segmentation strategy produces turn units that are likely to be more semantically coherent as they are not interleaved with overlaps.

Transcript preprocessing

We clean the transcripts to train our dialogue-based LM and present stimuli to participants in the same way as described in Section 5.3. In particular, characters specific to the Switchboard transcription guidelines¹ (mispronunciations, pronunciation variants, partial words, coinages) and non-speech sounds including vocalised noises and laughter are removed. Again, we retain filled pauses in transcripts.

Audio preprocessing

Although Switchboard recordings are dual-channel (i.e., one per speaker), we note that certain conversations contain significant channel bleed. To maintain overlapping turns in our stimuli while minimizing the chance of providing participants with incidental cues to the identity of the conversational partner, we de-bleed all conversations: pro-

¹<https://isip.piconepress.com/projects/switchboard>

C1 A "I think their pitching's deep enough that it doesn't matter um"
C2 B "it may be"
C3 A "I uh I like them a lot I I think they're going to go all the way um and and I I mean by saying
 I like them I like their chances I I actually don't like the dodgers I'm a giants fan"

R1 B "right"
R2 B "yeah well they've they've built a new baseball stadium downtown um it's opening next year"
R3 B "uh I mean of course this year is a good year to be a a chicago bulls fan I guess because
 they're doing pretty good"
R4 B "I mean they were terrible they could not pitch you know they couldn't even take hand-offs it
 was terrible but um"
R5 B "who the heck is gonna root for tampa bay right"

Figure 6.1: An example of plausible responses (R[1:5]) for speaker B sampled from our dialogue language model based on context turns C[1,2,3].

jections of the power spectra of both speaker channels are subtracted from each other before recomposing the individual channels into waveforms.

6.3.2 Stimuli construction

Although the requirement of lexically-equivalent response sets in the paradigm of Chapter 5 enabled strict isolation of prosodic information, it only provided partial answers to the question of how prosody influences people's expectations of upcoming spoken dialogue. First, a side effect of requiring frequently-occurring responses limited the stimuli to frequent, short, and often back-channel responses which are known to possess particular prosodic functions (Benus et al., 2007): does this finding generalise across other types of language when the lexical content is non-discriminatory? Second, it leaves open the question of how expectations are shaped when lexical information *is* discriminatory: although the findings from Chapter 5 make clear that people can leverage prosodic cues, do they do so in the presence of other information?

In Part I, we found that language models fine-tuned to dialogue data can achieve relatively strong correlations with human judgements of plausibility for upcoming turns. Though they were caveated, these results motivate the use of such models for creating sets of response which are similarly plausible in a given context without being lexically equivalent. Figure 6.1 displays an example of responses that are similarly plausible in the given context while being lexically diverse. This stimulus was creating by selecting responses that were scored highly by the dialogue LM trained for this study. This methodology offers us a chance to investigate the effect of non-lexical information on dialogue acceptability perception across unconstrained lexical content and variable

responses.

Dialogue language model

We select the model that was the most predictive of human acceptability scores in the experiments presented in Chapter 4. The model architecture is a BERT-based cross-encoder post-trained and fine-tuned on Switchboard proposed by Han et al. (2021), which we implement in PyTorch. We use `bert-base-uncased` from the Transformers library as our base model and follow the training procedure outlined in the original work. Post-training augments the standard masked language modelling task with an utterance relevance classification task and applies this joint loss to short (context-response) pairs; response selection is used as the fine-tuning objective. Full details of the model architecture and training procedure can be found in Chapter 4. For this model, we apply early stopping to both train stages using our validation set of Switchboard. We post-train and fine-tune for 9 and 17 epochs, respectively.

Stimuli construction

Following the training procedure, contexts are defined as 3 contiguous speaker turns from the NXT conversations. For each context, we randomly sample 1000 unique responses to score with our dialogue LM. The five responses for a particular context consist of its true response, along with the top-scoring responses. Random sampling of both contexts and responses allows us to explore the effect of non-lexical information across a range of linguistic functions. However, we perform broad filtering to ensure that stimuli are suitable—a trade-off between being informative enough but not too long—for the behavioural study. As such, we set some conditions for both contexts and responses. The final context turn must contain 3-50 tokens with an audio length of 2-10s. Potential responses must be of length 0.25-10s, preceded by a pause within –2-2s, and should not be from the same speaker as the context. The five first and last turns from all conversations are removed as the lexical content of greetings and farewells is highly conventionalised. We construct 120 stimuli in total².

²We have published our stimuli and their corresponding ratings: <https://sarenne.github.io/is-2023>

...

yeah

because he was around with a crowd of kids and he just really missed being able to be by himself and having some peace and quiet or

even at that age they need that time we all need that time

...

One of these responses is the true upcoming response in this conversation. Score how likely you think each potential response is in the context of this conversation.

(Please note: you can assign the same score to any number of responses and there is no requirement to score any response as 'Very Likely')

	Very unlikely	Somewhat Unlikely	Somewhat likely	Very likely
yeah	☐	☐	☐	☐
right it's not so much that they come to you every minute but it's that you are there when they need you	☐	☐	☐	☐
yeah we all do that	☐	☐	☐	☐
yeah i'll be able to spend the time with my daughter	☐	☐	☐	☐
just because they're grandparents just	☐	☐	☐	☐

Figure 6.2: A screenshot of the stimuli presented to participants through the online Qualtrics platform in the lexical-only condition.

Consider the following segment of a phone conversation and potential responses. In each audio clip, you will hear the turn in bold along with a potential response.

One of the responses is the true upcoming response in this conversation.

Based on *both the audio and the previous messages*, **score how likely you think the response is in the context of this conversation.**

(Please note: you can assign the same score to any number of responses and there is no requirement to score any response as 'Very Likely'. You can listen to the audio as many times as you need to make your decision.)

...

uhuh

i went to england and i really liked it there too

it's so expensive over there

...

	Very unlikely	Somewhat Unlikely	Somewhat likely	Very likely
	☐	☐	☐	☐
	☐	☐	☐	☐
	☐	☐	☐	☐
	☐	☐	☐	☐
	☐	☐	☐	☐

Figure 6.3: A screenshot of the stimuli presented to participants through the online Qualtrics platform in the *T0 Prosody* condition.

6.3.3 Online experiment

Mirroring the experiment described in Section 5.3, participants were recruited from Prolific Academic. We selected participants from North America for whom English was their first language to increase familiarity with the accents of speakers in Switchboard. Each participant received stimuli from either the lexical-only or acoustic (*T0 Prosody*) conditions, examples of which are pictured in Figure 6.2 and Figure 6.3 respectively. Running a within-participant design where participants are presented with a stimulus in the lexical condition followed by the speech condition carries the risk of priming participants to attend to the lexical information. Instead, we aimed to collect enough responses per stimuli to mitigate the effect of a between-participant design to then consider how perception *tends* to be affected by additional non-lexical information.

Manually constructed stimuli where only the true response was acceptable were interspersed throughout each survey as attention checks. Results of participants who obtained less than 80% accuracy on these questions were removed (24% and 26% of participants in the lexical and acoustic conditions). Each participant was presented with 20 random stimuli and the same five check questions. Average durations were 18.5 ± 8 (lexical) and 45.5 ± 26 (acoustic) minutes. In total, we collected 1200 responses: 5×120 stimuli in both conditions (45 and 50 participants in the lexical and acoustic conditions, respectively).

6.4 Perceptual information value: evaluation metrics

Having established the value of the turn discrimination paradigm in Chapter 5, we turn our focus to how it can be used to quantify the perceptual information value of non-lexical information. Our studies in both Chapter 3 and Chapter 5 showed people to be relatively adept at predicting different components of the upcoming signal. However, they also demonstrated that expectations are probabilistic and don't always align with reality. Accuracy-based metrics, which ignore the optionality in human expectations highlighted in our results, may thus be too coarse for measuring the value of information. Instead, we propose quantifying the value of information with metrics that are independent of the true upcoming signal and ask how much a piece of information *constrains expectations* more generally. In particular, we propose using entropy reduction across the scores assigned to responses to measure how channel conditions

affect participant agreement. Earlier works have used entropy reduction to quantify the potential value of non-lexical components of spoken dialogue for speech recognition: using a variant of Shannon’s guessing game where participants were asked to predict the upcoming word in speech or text, Ward and Walker (2009) demonstrated that people require fewer guesses—reflected by a reduction in entropy—before selecting the true word when they had access to the speech signal. However, their evaluation is still based on the true upcoming signal. Our study is the first of our knowledge to consider the effect of information in terms of expectation constraints.

In this study, we evaluate changes to participant responses using metrics that qualify both the general discriminative performance and entropy of responses among participants, both at the level of individual responses and stimuli.

Accuracy acc this measure reflects the frequency with which the true response was rated highest; it is computed in the same manner as in Section 5.3³. We also compute *cross entropy* which accounts for differences in certainty. Recall that this metric reflects the deviation between the “ideal” scoring distribution, and participant scores converted to a probability distribution; see Section 5.3 for full details.

Ordinal Entropy H_{Ord} First, we consider entropy at the level of individual responses. For this, we use a variant of cumulative paired ϕ -entropy that accounts for the ordinal nature of scores (Klein et al., 2016).

Note that for score counts S_i across scores $i \in 1, \dots, n$, we denote $p_i = P(S_i)$ and $p_{\leq k} = \sum_{i=1}^k p_i$.

$$H(S) = - \sum_{i=1}^n p_i \log p_i \quad (6.1)$$

$$H_{Ord}(S) = - \sum_{i=1}^n \left(p_{\leq k} \log p_{\leq k} - (1 - p_{\leq k}) \log(1 - p_{\leq k}) \right) \quad (6.2)$$

Standard entropy $H(S)$, defined in Equation (6.1), is a function of categorical label probabilities. As denoted in Equation (6.2), Ordinal entropy $H_{Ord}(S)$ is instead a function of the cumulative probability for each score to reflect the dispersion among scores (Yager, 2001). As displayed in Figure 6.4, there is an interesting distinction between

³We use a different method to compute accuracy than the results published in (Wallbridge et al., 2023) to facilitate comparisons with Chapter 5.

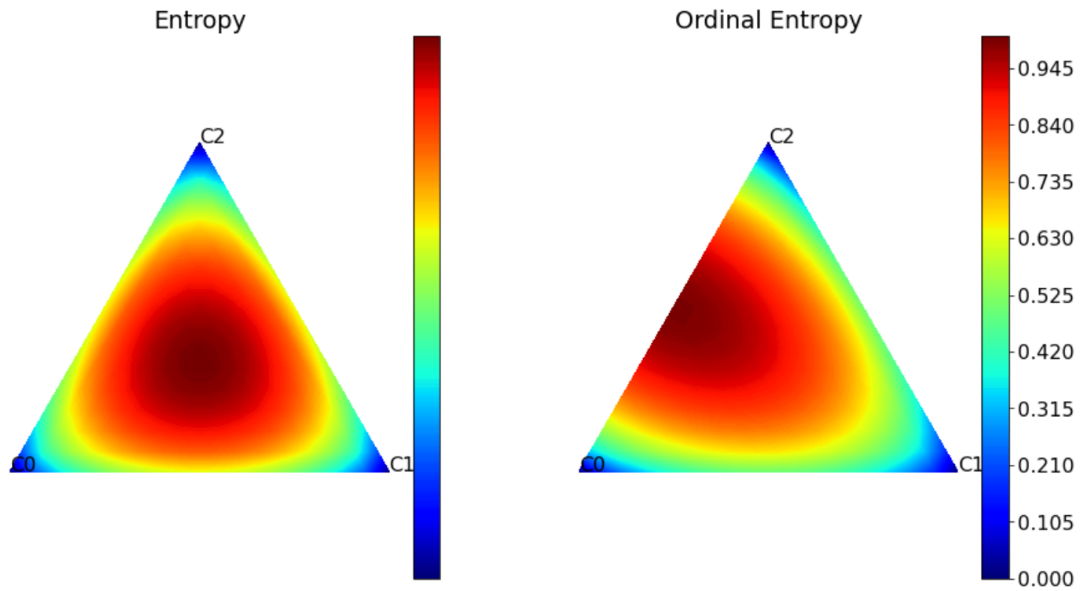


Figure 6.4: A visualisation of how ordinal entropy and entropy differ, using an example of three ordered classes. Although minimum entropy (normalised) is achieved at the corners of both metric plots (i.e., when each category carries the highest probability possible), the maxima of these metrics are produced in different conditions. Maximum standard entropy occurs when probability mass is evenly distributed between categories while maximum ordinal entropy is obtained when probability is evenly distributed between the most distal ordinal classes.

ordinal and Shannon entropy. The latter is maximized when all categories occur with the same probability. For example, this would occur in our experiment when two of the five participants assign a score ‘1’, and the remaining participants each assign a different score. Ordinal entropy, however, is maximized by contradictory information (Klein et al., 2016). This would occur when the extreme scores are equally likely e.g., for five participants, two assign ‘1’ and three assign ‘4’).

Permutation Entropy H_{perm} Our stimuli involve presenting multiple responses at once. As such, we additionally examine entropy at the stimulus-level – how do different conditions affect scoring behaviour of the responses in a given stimulus relative to each other? For this, we take inspiration from Kolmogorov-Sinai entropy which has been used to quantify the theoretical complexity of time-series data. It is often approximated through Ordinal Pattern Analysis (OPA): time-series data is converted into sequences of ordinal patterns, then standard Shannon entropy $H(S)$ (shown in Equation (6.1)) can be computed over pattern frequencies (Cuesta-Frau et al., 2019). To compute H_{perm} , we convert participant scores across responses to rank patterns and compute standard Shannon entropy over these patterns. This quantifies agreement at the stimulus-level by considering how scoring behaviour may change across sets of responses. Because OPA approaches consider the relative ordered relations between time-series values rather than the absolute values themselves, it can be seen as a form of quantization (Fouda et al., 2017), however comparing scoring behaviour across response sets captures another way in which conditions could constrain expectations compared to ordinal entropy.

Consider the follow two examples of scores that could be obtained from two participants A and B across the 5 responses ($R_{[1,...,5]}$) for one of our stimuli:

Example 1	Participant A	$[R_1:2, R_2:2, R_3:1, R_4:2, R_5:3]$
	Participant B	$[R_1:3, R_2:3, R_3:2, R_4:3, R_5:4]$
Example 2	Participant A	$[R_1:3, R_2:2, R_3:2, R_4:2, R_5:4]$
	Participant B	$[R_1:2, R_2:3, R_3:1, R_4:3, R_5:3]$

Though both examples produce the same same ordinal entropy for each response, the rank patterns are equivalent in the first example but not in the second, yielding different permutation entropy values.

6.5 Results

To answer the research questions for this chapter, our goal was to establish an experimental setting where we could examine participant expectations about an upcoming turn in spoken dialogue when the contents of the lexical channel are plausible but diverse. Although the dialogue language model used to sample turns has demonstrated a capacity to align in some ways with human perception, it is far from perfect. As such, we sample a subset of stimuli where we can be more certain that lexical content is uninformative—the **ambiguous stimuli** set. The condition for selection is based on the participant responses in the lexical condition: where participants did not reliably score the true response highest (i.e., no more than two of the five participants ranked the true response as the only highest-scoring). From the 120 stimuli, this produced a subset of 63. We group the other 57 stimuli as the **remaining stimuli** set.

6.5.1 Evaluation methods

To evaluate the effect of non-lexical information, we compare scoring behaviour across the channel conditions and to chance scoring behaviour. We estimate chance by shuffling each set of participant ratings 100 times to maintain score distributions.

Additionally, we analyze the effect of condition on ordinal entropy and mean scores while controlling for other factors with Bayesian multilevel regression models. As entropy values are bounded and continuous, we scale them to $[0, 1]$ and use Zero-One Inflated Beta Regression. Models were fit using `brms` in R (Bürkner, 2017). We use the `emmeans` package (Lenth, 2023) to compute estimated marginal means and 95% Highest Posterior Density Regions (i.e. Credible Intervals (CIs)) to examine the effects of predictors. To investigate potential cognitive load differences, we include *response length* (in seconds) as a predictor. Following the results from Chapter 4 where we found the plausibility of a response to be related to its LM-estimated surprisal, we include the mean surprisal of the response conditioned on the context from TurnGPT as described in Chapter 4. We also include an indicator for whether the response was the *true* continuation to see if participants treated true and false continuations differently. Group-level effects (i.e., random effects) for context and response dialogue acts from the Switchboard NXT annotations Calhoun et al. (2010), as well as effects for context, and context-response identifiers are included to control for stimuli variation. We also include interaction terms with response length, target, mean surprisal, and dialogue acts to see if the effects varied between the acoustic and lexical conditions. We see

Metric	Condition			
	Chance	Lexical	Acoustic	Difference
$acc \uparrow$	0.23 ± 0.01	0.27	0.40	0.13
$cross\ entropy \downarrow$	0.44 ± 0.02	0.43	0.44	-0.01
$H_{Ord} \downarrow$	0.66 ± 0.01	0.54	0.48	-0.06
$H_{Perm} \downarrow$	0.94 ± 0.01	0.90	0.86	-0.04

Table 6.1: Evaluations for the ambiguous stimuli set across conditions, and the mean per-question difference between them (standard and cross-entropy accuracy (acc , $cross\ entropy$), ordinal, permutation entropy (H_{Ord} , H_{Perm}); all normalised). Differences are all significant using a directional Wilcoxon rank sum test ($p < 0.002$).

Metric	Condition			
	Chance	Lexical	Acoustic	Difference
$acc \uparrow$	0.21 ± 0.03	0.65	0.53	-0.12
$cross\ entropy \downarrow$	0.45 ± 0.02	0.23	0.25	0.02
$H_{Ord} \downarrow$	0.70 ± 0.01	0.55	0.48	-0.06
$H_{Perm} \downarrow$	0.96 ± 0.01	0.89	0.86	-0.03

Table 6.2: Evaluations for the remaining stimuli set. Mean per-question differences are all significant using a directional Wilcoxon rank sum test at $p < 0.0001$ except permutation entropy ($p < 0.02$).

non-zero variance estimates for the group-level effects. i.e., these factors do account for variation in the entropy.

6.5.2 Participant performance

Accuracy

As can be seen in Table 6.1, accuracy for *ambiguous* stimuli in the lexical condition is close to chance. In the acoustic condition, it is significantly higher. We find statistically significant mean per-question differences between channel conditions. Increased accuracy provides strong evidence that participants leverage non-lexical cues to constrain their expectations about the upcoming dialogue turn and that these cues are informative

Metric	Condition		
	Lexical	Acoustic	Difference
<i>acc</i> ↑	0.97	0.96	0.01
<i>cross entropy</i> ↓	0.07	0.11	0.04

Table 6.3: Accuracy-based metrics for the check questions included in each participant survey.

regarding the trajectory of spoken dialogue.

Similarly, Table 6.2 contains results for the *remaining* stimuli—where participants could discriminate the true response relatively accurately from the lexical channel alone. Participants score these stimuli more accurately than the ambiguous stimuli in both the lexical and acoustic conditions. Surprisingly, people are less adept at selecting the true turn when provided with non-lexical information for this set of stimuli.

Finally, accuracy metrics for the check questions are shown in Table 6.3. Standard accuracy remains consistent in both conditions, however, cross-entropy is worse in the acoustic condition, indicating that participant scoring is less decisive in this condition. This is surprising in light of the evidence we discuss in Chapter 1 that the interactions and redundancy between channels in multi-modal communication increase communicative power over single-channel communication (Hebets and Papaj, 2004; Fröhlich et al., 2019). Higher cross-entropy may reflect a higher cognitive load for participants in the acoustic condition; we find further evidence for this hypothesis in the proceeding evaluation. As such, the results presented here likely understate non-lexical channel value.

Entropy reduction

As our results have suggested in Chapters 4 and 5, accuracy-based metrics do not reflect the optionality in human expectations about upcoming communication. As such, we also quantify the effect of non-lexical information as entropy reduction over participant scores. Both entropy-based metrics decrease in the acoustic condition for the *ambiguous* stimuli—participants agree more in the acoustic condition (Table 6.1). This suggests that the non-lexical channel provides additional cues for what may come next and that participants interpret them similarly. Crucially, although Table 6.2 shows reduced accuracy between conditions for the *remaining* stimuli, score entropies decrease

and to similar degrees as found across the lexically ambiguous stimuli. Entropy reduction is thus not simply a consequence of increased accuracy. Rather, the non-lexical channel seems to produce consistent perceptual effects, regardless of how informative lexical content is for turn discrimination.

Regression Models

The regression models described above allow us to investigate how our evaluation metrics interact with other features of our stimuli; we detail each below.

Response length To investigate potential differences between conditions in terms of cognitive load, we explore whether response length affects the predictive power of our models. Higher entropy in the lexical condition compared to the acoustic condition is reflected by our ordinal entropy regression model, but we also see an interaction of condition with response length. Specifically, we find a positive slope estimate in the audio condition (0.022 , $CI = (0.010, 0.033)$), contrasting with a flatter slope for the lexical condition (-0.001 , $CI = (-0.012, 0.011)$). That is, participants show higher agreement in the audio condition for short responses compared to the lexical condition, but the difference shrinks for longer utterances. This provides further evidence of differences in the cognitive load required for each channel condition. We find no interaction effect for response length for the mean score model with estimates for conditions centred at zero (difference = 0.01 , $CI = (-0.00, 0.01)$).

Target We examine whether participants treated true and false continuations differently by including a corresponding binary label and calculating estimated marginal mean ordinal entropy for target and non-target responses while averaging over other predictors. Here we see that non-targets have higher ordinal entropy than target responses in the audio condition, but the credible interval spans zero (audio difference = 0.05 , $CI = (-0.0311, 0.12)$). We see the same trend in the lexical condition, with a smaller difference (lexical difference = 0.02 , $CI = (-0.05, 0.083)$). However, we do see a non-zero difference between the audio and lexical condition for target responses (difference = -0.13 , $CI = (-0.257, -0.001)$), though not as clearly for non-target responses (difference = -0.10 , $CI = (-0.203, 0.001)$). In other words, responses have lower ordinal entropy in the acoustic condition, but the difference between conditions is more pronounced for the true responses than for false candidates.

The mean score model displays this difference more clearly, where target responses obtain higher ratings than non-target responses in both conditions. For the acoustic condition, we find target score = 0.846, $CI = (0.794, 0.893)$; non-target score = 0.694, $CI = (0.630, 0.751)$; for the lexical condition, we find target score = 0.755, $CI = (0.694, 0.815)$; non-target score = 0.666, $CI = (0.595, 0.727)$. Again, target responses are rated higher in the acoustic condition than in the lexical condition (difference = 0.09, $CI = (0.03, 0.144)$), while the difference appears smaller for non-target responses (difference = 0.02, $CI = (-0.03, 0.09)$). This indicates that target responses have acoustic properties that participants use in the rating task, but this does not always translate into participant agreement.

Surprisal Following the work described in Chapter 4 where we found the plausibility of a response to be related to its associated surprisal, we include the mean surprisal of the response conditioned on the context. This predictor also produces a difference in slope between the acoustic and lexical conditions. Estimated slopes in the audio condition are generally positive (audio median = 0.081, $CI = (-0.001, 0.159)$), while estimates for the lexical condition are centered around zero (lexical median = 0.001, $CI = (-0.068, 0.334)$). Responses that are more lexically surprising tend to display less agreement in the acoustic condition, but there is no clear effect of lexical surprisal on ordinal entropy in the lexical condition. This may be a result of the ambiguous stimuli selection, but also again indicates that acoustic information changes participant expectations.

We also find the effect of mean surprisal on the mean score model to depend on the channel condition (condition difference = -0.05 , $CI = (-0.10, -0.00)$). Surprisal has an overall negative effect on the scores, but the effect is stronger in the audio condition.

The effect of surprisal and target status on mean plausibility scores are in line with our findings presented in Chapter 3. However, we also see differences depending on how their effect on dialogue expectations is quantified: predictors of higher plausibility ratings may not lead to higher agreement.

6.5.3 Applying entropy metrics to lexically equivalent stimuli

In Chapter 5, our primary goal was to validate the turn discrimination paradigm—to what extent could people perform this task? Though we evaluated how participant certainty changed between the *T0 Lexical* and *T0 Prosody* conditions, our evaluation was

Channel condition	Context Condition			
	(1,0)	(4,0)	(7,0)	(4,3)
<i>T0 Lexical</i>	0.523	0.549	0.543	0.532
<i>T0 Prosody</i>	0.531	0.529	0.573	0.544

Table 6.4: Mean participant performance as measured by **ordinal entropy** for the data collected in Chapter 5.

primarily centred on participant accuracy. Here, we report how participant agreement changed. However, we note that the dataset collected in this chapter contains fewer participant responses per stimulus. With fewer responses than ordinal categories, permutation entropy cannot be computed. We report ordinal entropy in Table 6.4 though note that the fewer participant responses also affect metric reliability. To test the significance of our values, we estimate “chance” ordinal entropy by shuffling the responses for each stimulus 100 times. As shown in Figure 6.5, all values across both the *T0 Lexical* and *T0 Prosody* conditions are significant (at $p < 0.05$) compared to the mean ordinal entropy expected by chance; however, none of the differences between the *T0 Lexical* and *T0 Prosody* are significant. The size of the dataset from Chapter 5 limits the scope for applying consensus-based metrics; however, the data collected in this chapter has demonstrated the value of such analysis.

6.6 Conclusions

The generalised turn-discrimination paradigm presented in this chapter enables an analysis of the shape of lexical and non-lexical expectations and the forces that act on them over a much broader set of language than was previously possible. The results from both of our turn-discrimination studies provide firm evidence that non-lexical information constrains expectations of spoken dialogue—people leverage non-lexical cues to discriminate true continuations from false candidates when the lexical content is equally plausible but variable. However, when the lexical content is informative for the discriminative task, acoustic information can lead to less accurate expectations. In other words, expectations are a product of the synergistic information that arises between the lexical and non-lexical channels in spoken dialogue. This finding is particularly interesting in the context of previous works which present varying conclusions

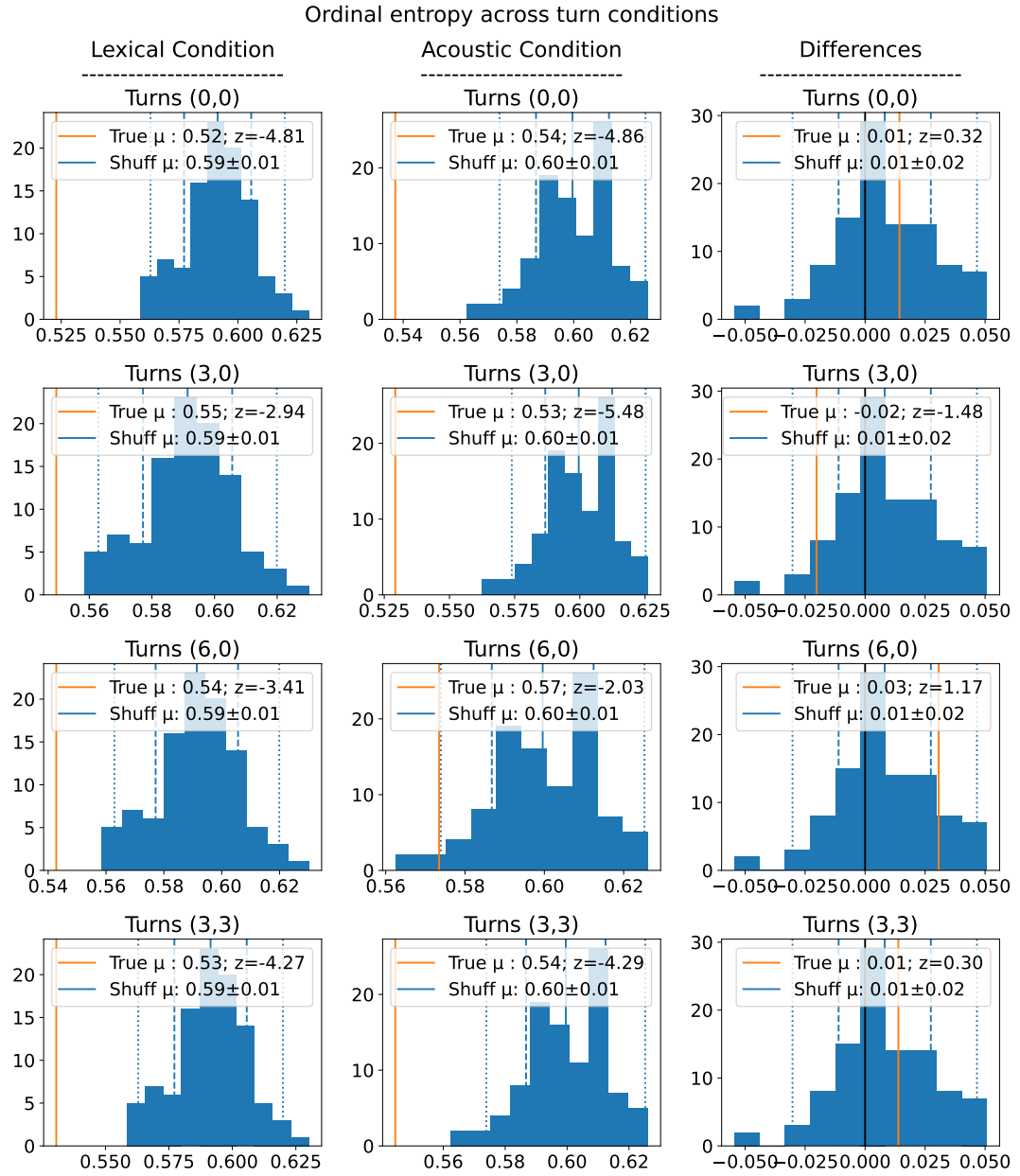


Figure 6.5: Comparing the significance of ordinal entropy values across conditions. The blue distribution reflects the mean ordinal entropy that could be expected by chance while the orange line shows the true ordinal entropy obtained in each turn and channel condition. The final column shows the difference between the acoustic *T0 Prosody* and lexical *T0 Lexical* conditions, none of which are significant.

about listeners' use of prosodic cues. Although surprising, these results join a small set of works demonstrating high agreement between listener judgements of speech, even if they are incorrect. Using spontaneous dialogue where segmental information has been suppressed through spectral inversion and low-pass filtering Kreiman (1982) asked listeners to identify local sentence and paragraph (i.e., topic) boundaries. Though participant accuracy was relatively low, agreement was high. In a comparison similar to ours between speech+text, and text-only conditions, Hirschberg and Nakatani (1996) find "markedly higher inter-labeler reliability" for segmenting discourse in the speech+text condition. Corps et al. (2018) similarly showed that listeners respond earlier but not necessarily more accurately in turn-end detection tasks when they receive informative contextual cues. As we highlighted in Chapter 5 through an analysis of participant certainty and in Chapter 4 in comparing human and model acceptability scores, accuracy-based metrics are not sufficient for quantifying the perceptual value of information. Quantifying the perceptual value of the non-lexical channel instead as entropy reduction across participant perceptual judgements provides a novel perspective on channel value: even though it may lead to incorrect discriminative judgements, people tend to adjust their expectations in similar ways when provided with non-lexical context. These results suggest that people share knowledge regarding expected prosodic structure in dialogue.

We investigated factors including response length and surprisal that could affect non-lexical channel value and found evidence of complex interactions between lexical and non-lexical information. Future work could develop a more formal quantification of channel value that explicitly accounts for interactions between channels (e.g., between lexical content and prosodic realisation). For example, different speech acts have been shown to exhibit greater prosodic variation, potentially indicating that the information mass is skewed more heavily towards the non-lexical channel for certain speech acts (Syrdal and Kim, 2008). Though we sampled similarly plausible turns from our dialogue language model with only broad filtering for use as behavioural stimuli, more targeted response sampling could also provide insights into the interactions between channels. In future work, we would also like to consider comparing the acceptability of responses in a no-context condition. This acts as a different baseline for the intrinsic acceptability of an utterance and sheds light on what may constitute a "generic" prosodic realisation. This would likely require careful piloting as the task of judging plausibility without a context is far less intuitive. O'Mahony et al. (2021) shows that the wording of questions (i.e., "naturalness" versus "appropriateness") af-

fects participants' judgements.

Our methodology also touches on the field of text-to-speech synthesis (TTS) which could both benefit from the evaluation that this paradigm enables as well as be used to improve this paradigm. On one hand, this paradigm provides an alternative method for evaluating synthesized speech; we discussed this possibility in Section 5.5. On the other hand, the fact that we could only include one audio context turn is a limitation of both studies presented in Chapters 5 and 6. It would also be desirable to provide participants with longer-term non-lexical information (e.g., about speaker turn-taking or entrainment patterns) to identify the different contributions of each channel to expectations. Exposing the listeners to more audio context could be approached through synthetic speech, though it would require a high degree of controllability which has not yet been established for current TTS systems (Xie et al., 2022; Sigurgeirsson and King, 2023).

Part III

Multi-channel representations

Part III Introduction

We have found that humans are relatively adept at making predictions about the upcoming communicative signal in dialogue; however, their expectations are probabilistic and consistently reflect *optionality*—the scope of trajectories which may be deemed as plausible continuations. Turn discrimination judgements show that lexical information is useful in constraining upcoming expectations of both upcoming lexical content and prosodic form, and that prosodic context provides further constraints.

Here, we use self-supervised mechanisms as a tool to investigate the predictable structures in and between these channels that may contribute to human prediction. We begin in Chapter 7 by studying the information unique to the non-lexical channel of speech: what forms of predictable structures are exhibited in the acoustic correlates of prosody in and of themselves? This study addresses a question that we introduced in Section 2.2.3 about appropriate temporal granularities for modelling speech. Though prosodic effects are largely achieved through the joint realisation of acoustic cues and lexical content, we discussed evidence that prosody can affect speech perception even when lexical content is suppressed in Section 2.1. We believe ours to be the first exploration into the effect of temporal and feature granularity on the predictability of prosodic structure independent of lexical content. We demonstrate that self-supervised methods can be applied to F0 and energy features, indicating that they do exhibit predictable structures. In Chapter 8, we utilize these learned representations along with the lexical representations which we investigated in Chapter 4 in a series of proof-of-concept experiments into how artificial neural networks that rely on predictable structures in their input solve the turn discrimination task.

Chapter 7

Systematicity in prosody independent of lexical content

The content of this chapter is based on a collaboration with Christoph Minixhofer¹. In relation to this thesis, our broader goal was to introduce and validate a methodology for producing generic representations of prosodic information independently of lexical content. In doing so, we address the following research question:

RQ: “*At what timescale do the acoustic correlates of prosody exhibit predictable structure?*”

7.1 Introduction

Throughout this thesis, we have discussed predictive theories of language comprehension, as well as empirical evidence for prediction in the comprehension of written language at length (Hale, 2001). We now shift our focus to a key underlying assumption in these theories: the ability to generate such expectations relies on the presence of predictable structure in the linguistic signals, i.e., the *systematicity* of the signal. Similarly to comprehension behaviors for written language, there is plentiful evidence that listeners exploit systematicity to predict features of upcoming speech. Predictions range from the informational content of speech (Brown et al., 2015) and the position of prosodic focus (Ip and Cutler, 2017) to the length of upcoming linguistic units including sentences (Grosjean, 1983) and conversational turns (Ekstedt and Skantze, 2022).

¹Statement of contributions: I proposed the overarching idea behind the masked prosody model and utterance-level evaluations, as well as contributed to the MPM architecture design and processing of the underlying data for BURNC; CM implemented the MPM model and all experiments. While I led the analysis and writing of the manuscript, CM also contributed to these components of the study.

Our results in Part II provide further evidence, showing that people make predictions about both upcoming prosodic realisations and lexical content. However, returning to our description of PID from Section 1.3, it is unclear how much of this predictive capacity is contingent on the structures unique to the lexical channel, those unique to the non-lexical channel, or on patterns across their joint realization. In particular, the extent to which information *independent* of lexical content contributes to predictions has been relatively unexplored. In this chapter, we investigate systematicity in specific components of the non-lexical channel—relative pitch (F0) and loudness (energy). Through these experiments, we develop a method for representing prosodic realisation isolated from lexical information and demonstrate that the acoustic correlates of prosody exhibit structure themselves. This, in turn, provides further insight into the processes listeners may employ to generate expectations of upcoming speech.

We address two major challenges in studying the structure of prosody. First, prosodic effects are realised at multiple temporal granularities, making it difficult to define a multipurpose timescale for modelling prosody. Appropriate modelling dimensions for phonetic content have been proposed in numerous works (Paliwal et al., 2010; Baeviski et al., 2020), however, the shape of a prosodic “unit” that is independent of communicative function is a persistent debate in linguistics (cf. Cutler and McQueen, 2014). Second, prosodic and lexical content are not independent. Prosody functions primarily *in conjunction* with lexical content: for example, Lai (2010) shows that prosodic contours interact with, rather than override, the interpretation of the underlying lexical identity of cue words. Similarly, Tang and Shaw (2020) finds that, in Mandarin, phonetic details associated with the average prosodic prominence of a word in discourse can be predicted from its lexical identity. However, prosody also contributes communicative information independently of lexical content. For example, people leverage prosodic features to learn word boundaries in artificial language (Matzinger et al., 2021) and can be primed to different syntactic disambiguation strategies by prosodic differences in de-lexicalised audio (Mills, 2020). Moreover, pitch contours are useful cues for automatically detecting reported speech (Cervone et al., 2015), emotions (Paeschke et al., 1999), and rhetorical relations (Murray et al., 2006). These works provide evidence that the acoustic correlates of prosody exhibit systematicity in their own right. As a solution to both problems, we propose leveraging Self-Supervised Learning (SSL), a powerful mechanism that exploits complex structures across timescales, to explore the degree of systematicity inherent to acoustic correlates of prosody.

We introduce a novel Masked Prosody Model (MPM) which encodes correlates of pitch, loudness and voice activity by learning to reconstruct corrupted sequences of these input features. Using MPM as a tool for studying the structures in the acoustic correlates of prosody, we break the overarching research question of this chapter into two sub-questions: 1. *Can we identify structural predictability in prosody independent of lexical content using SSL representations?* and 2. *How well-suited are the resulting representations for predicting perceptual labels of speech features at different temporal granularities?* Our results demonstrate SSL models can uncover systematicity exhibited by the acoustic features of prosody. Using linear probes and a range of perceptual tasks that involve temporal dependencies at different granularities, we first investigate the effect of corruption strategy (altering masking and codebook properties) on task performance. Our results show that different tasks benefit from different masking strategies, driven primarily by the timescale of the prosodic prediction objective. Additionally, we use these probes to compare SSL representations to more a constrained encoding of hierarchical prosodic structure, demonstrating that the complex contextualisation of SSL is beneficial for perceptual tasks that involve more abstract reasoning such as emotion recognition. Finally, we use more powerful probe architectures to compare the value of generic structures learned through SSL and task-specific structures encoded by the probes themselves. These experiments validate that the MPM encodes useful aspects of prosody independently of lexical content, while also highlighting the complexity of structure in the acoustic correlates of prosody.

7.2 Background

In this section, we motivate our use of self-supervised learning as a valuable tool for studying the acoustic correlates of prosody and review existing methods for representing prosodic information.

7.2.1 Self-supervised learning

As described in Chapter 2, self-supervised learning (SSL) mechanisms encode the implicit structure of their input data by learning to reverse a corruption function—e.g., removing future, past, or intermediate context. Without a need for external labels, SSL models can exploit large amounts of unlabeled data. The resulting representations may be used as inputs to models trained on much smaller sets of labelled data or the

SSL model itself may be fine-tuned (Donahue et al., 2014; Gunasekar et al., 2017). Transformers trained with SSL produce highly contextualised representations of both global and local structures and have been successfully applied to represent, segment, and study text (Devlin et al., 2019), speech (Baeovski et al., 2020) and a host of other signals including protein sequences (Ofer et al., 2021), and brain activity scans (Bellier et al., 2023). However, the nature of these models provides little information about which input structures these mechanisms exploit during learning. As such, it is unclear how much prosodic information is maintained in representations (de Seyssel et al., 2023; Li et al., 2023).

7.2.2 Representations of prosody

Representations of prosodic realisation have been proposed for both the automatic synthesis and understanding of speech. To our knowledge, no works have used them to explore the systematicity in non-lexical prosodic information and the majority of such representations are not suitable for such an investigation. They do, however, motivate the current study. We discuss the most relevant representations below.

Many recently-developed automatic representations of prosody have been designed for the purposes of speech synthesis where a long-standing goal has been to control the prosodic realisation of text (and speaker) inputs (Klatt, 1987). Such encodings often use a subtractive definition of prosody: the variation remaining once phonetic, speaker, and channel information are removed (Skerry-Ryan et al., 2018). A popular strategy for learning such representations is through a reference encoder—during training, it is tasked with producing a compressed representation of the target speech signal Wang et al. (2018). Disentanglement of these components is achieved using carefully tuned bottlenecks or training data design (Ioannides et al., 2023). While the resulting representations have proved useful for controlling speech generation (Huang et al., 2023) and for related tasks like emotion recognition (Ioannides et al., 2023), the degree of achieved (and achievable) disentanglement is unclear. For example, popular representations learned through reference encoding such as Global Style Tokens can be sensitive to speaker and lexical perturbation (Skerry-Ryan et al., 2018; Sigurgeirsson and King, 2023). As such, it is difficult to draw conclusions about prosodic contributions independently from lexical content from such representations.

Representations of prosodic information based instead on the acoustic correlates of prosody are independent of lexical and phonetic content by design, but much less

well-studied. One of the few examples of such representations was proposed by Suni et al. (2017) and applies Continuous Wavelet Transform (CWT) analysis to F0, energy, and duration features for automatic annotation of local prosodic events. CWT generates hierarchical representations from scalograms obtained by convolving the original signal with wavelets across different frequencies. The convolution acts as a measure of similarity between a wavelet and the original signal; thus, the wavelet frequencies each represent a hierarchical level. Using a handcrafted method to convert these representations into labels for perceived prominence and boundary strength at the word-level, the resulting labels achieve comparable performance to supervised methods for both tasks on the Boston University Radio News Corpus, a corpus of six speakers reading news stories. They have also been shown to be useful for generating prosodic variance in TTS that is more faithful to reference speech (Suni et al., 2020). These results both highlight the importance of imbuing hierarchical structure in representations of prosody, a feature which has long been postulated by linguistic theories (Selkirk, 1984; Beckman, 1986). Other work has applied relatively simple statistical methods for learning patterns over word-level pitch and intensity features to detect prominent words based on the predictability of prosodic features (Kakouros and Räsänen, 2015). This study trains n-gram models on the quantized values for pitch and intensity features taken from a corpus of acted infant-directed speech and although the representations require gold-standard word alignments, they demonstrate that even simple predictive models can be trained from pitch and intensity features of speech. Learned representations of F0 and de-lexicalised speech have also been shown to be useful to control re-synthesis of speech (Polyak et al., 2021), for music information retrieval using singing (Noufi and Verma, 2020), and for classification tasks such as persuasion and sarcasm detection (Weston et al., 2021). These works demonstrate that prosodic correlates display some form of systematic structure but leave open the question of at what time scale. We address this question directly.

7.3 Methodology

Inspired by masked language models for text which have been shown to capture information at different timescales, we introduce a Masked Prosody Model which produces representations of the correlates of pitch, loudness and voice activity by learning to reconstruct corrupted sequences of these values during pre-training (Devlin et al., 2019).

Using linear probes, we investigate the effects of corruption strategies, in terms

of mask size and feature granularity, on the utility of resulting representations for a range of downstream tasks. While defining units of text is more intuitive than speech, experiments on written language find that mask size and tokenization strategy have been shown to play important roles in the quality of representations (Joshi et al., 2020; Levine et al., 2020; Ankner et al., 2023; Mielke et al., 2021). Similarly, we expect the mask size and granularity of acoustic features to impact the complexity and value of self-supervised representation for downstream task performance. Additionally, we compare MPM representations to CWT representations that encode purely hierarchical structures to investigate the value of broader forms of structure. Using more powerful conformer probes, we also compare the utility of generic structures learned through SSL to task-specific structures learned through the probes themselves.

In the following sections, we introduce the MPM architecture and our set of downstream tasks.

7.3.1 Masked Prosody Model

The architecture of the Masked Prosody Model is shown in Figure 7.1; the model is primarily made of Conformer blocks (Gulati et al., 2020) which are well-suited to the high resolution and continuous nature of the acoustic inputs. These blocks connect three separate input and output sequences corresponding to each of the three prosodic features.

Pitch (F0) and voice activity features are extracted using the WORLD vocoder (Morise et al., 2016) while energy is computed as the Root-Mean-Square (RMS) of each Mel Spectrogram frame. As discussed in Section 2.1, pitch tracking is a difficult task with numerous proposed algorithms but no consensus on a universally applicable method. We discuss some popular methods here:

- **Praat** (Boersma and Van Heuven, 2001) is one of the most popular speech analysis programs. Its default F0 tracking method is based on autocorrelation which computes the similarity of a single signal against itself at different relative delays. The maximum autocorrelation value is obtained for the lag at which the signal repeats itself most accurately. Local maxima are then selected as F0 hypotheses.
- **RAPT** Talkin and Kleijn (1995) generates F0 candidates by extracting local maxima of the normalized cross-correlation function which is calculated over

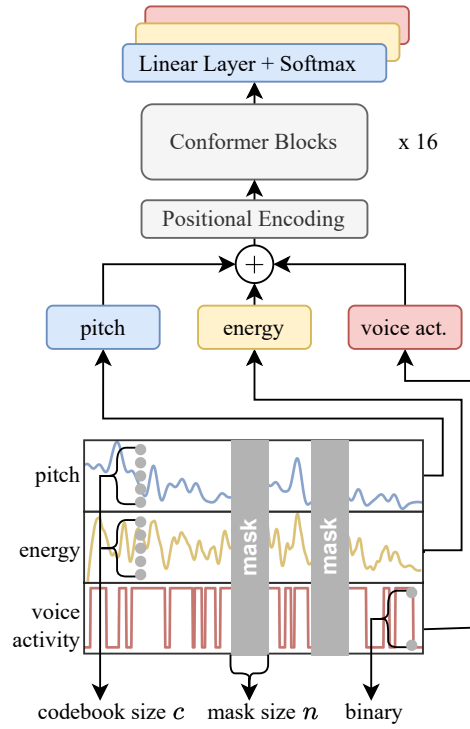


Figure 7.1: Architecture of the Masked Prosody Model.

voiced speech. RAPT incorporates a DTW smoothing strategy to minimize spurious jumps and F0 halving or doubling in F0 trajectories. This can be counter-productive when processing noisy speech.

- **WORLD** (Morise et al., 2016) F0 is estimated with the DIO method (Morise et al., 2009). This involves applying a range of low-pass filters to generate F0 candidates (if the filtered signal contains only F0, it should produce a sine wave). The most reliable candidate is selected as the F0 estimate (where reliability is based on the deviation of the 4 intervals of the respective waveform (intervals between peaks, dips, and 0-crossings)).

We select the WORLD extraction method because DIO achieves low Gross and Fine Pitch Errors using EGG recordings as gold truth (compared to methods including Harvest, RAPT, SWIPE) (Drugman et al., 2018).

All features are extracted at a resolution of 10 ms and normalised (min-max normalisation) across the audio segment. Normalisation provides relative contours which are known to be perceptually important (Rosenberg, 2010; Kimball and Cole, 2016). While it removes information about e.g., the relative feature values for individual

speakers, it makes our method applicable to unseen speakers and thus more useful for our experiments in Chapter 8. Before masking, features are quantized into respective codebooks P , E and V of size c . During training, random segments of the aligned feature sequences, each with a mask length of m , are masked until $(50 \pm 5)\%$ of the input signal has been obscured. The model is trained to reconstruct each feature with an independent Categorical Cross Entropy loss. We define the training objective as the sum of the loss for each feature, normalized by $1/\log c$.

To evaluate model representations, we extract features from intermediary Conformer block layers as input for downstream tasks. When aggregating at word or utterance levels, we compute the mean and maximum features for the target unit and concatenate them. Representations are evaluated using both linear probing (Alain and Bengio, 2017), in which a linear classifier is trained, and a more powerful Conformer classification model with 2 layers.

7.3.2 Downstream tasks

To answer the research questions of this chapter, we investigate the perceptual utility of the MPM representations trained using different corruption functions and compare them to other representations of prosody. To do so, we ask whether they can predict human-generated annotations of speech. Because the speech signal is perceived as a combination of lexical and non-lexical information, there are very few datasets of perceptual labels that are completely unaffected by the underlying semantic contribution of lexical information (an example of such a dataset could be asking participants to label prominence for speech in an unfamiliar language). Instead, we select annotation tasks that are known to be at least partially driven by prosodic features. Importantly, we choose tasks that depend on prosodic information across different temporal ranges to evaluate the extent to which the MPM representations encode such dependencies.

We assess whether representations capture short-term dependencies through fine-grained segmentation performance at phone and word levels. Although prosodic structures are unlikely to be useful for phone segmentation, prosodic features are theorized to contribute to word recognition by marking boundaries (Turk, 2010). Prosody is also known to support the linguistic organisation of speech above words through the demarcation of structure and salience (Cole, 2015; Calhoun et al., 2021; de Seyssel et al., 2023). As such, we test prominence and break detection. Finally, we evaluate whether MPM features capture utterance-level systematicity through the task of

emotion classification, as affective information can be conveyed through prosody at the utterance level (Batliner et al., 2010). The relationship between emotion recognition and prosodic information has been demonstrated by similar studies that make use of handcrafted features like ToBI (Cao et al., 2014), as well as representations of delexicalised speech from artificial neural networks (Weston et al., 2021).

We note that though prosody is known to play a role in these tasks, the absolute performance for all selected tasks would benefit from access to the full speech signal.

Fine-grained segmentation. We evaluate **phone and word segmentation** using the **TIMIT** dataset, a set of high-quality recordings from 630 speakers with American English dialects, each reading ten phonetically-rich sentences with corresponding manual segmentations at phone and word levels (Garofolo, 1993). The segmentation tasks involve correctly identifying a frame labelled as a segment boundary within a small tolerance window of contiguous frames. Though the larger frame rate of MPM representations means our tolerance window is wider, self-supervised representations of the full speech signal report F1 scores of 85.47 and 25.45 for phone and word segmentation tasks (Bhati et al., 2022). Though we know of no directly comparable studies using only prosodic features for such fine-grained segmentation of speech, infant studies (Räsänen et al., 2018; Ludusan et al., 2022) and artificial language learning experiments (Matzinger et al., 2021) suggest that prosody is a useful cue for word segmentation but less so for phone segmentation. Compared to the performance that can be achieved from the input features alone, we do not expect the complex contextualization of SSL to contribute to phone segmentation. We only expect marginal gains for the word segmentation task.

Prosodic event detection. We use the Boston University Radio News Corpus **BURNC** to evaluate **prominence and break detection**. The Lab News portion of this corpus consists of 6 professional news speakers reading 4 news stories, each annotated for prosodic phrasing, phrasal prominence, and boundary tones (Ostendorf et al., 1995). In this corpus, prominence is annotated with ToBI pitch accent types while boundaries between words are labelled for strength on a scale $\{0, \dots, 4\}$ (Silverman et al., 1992).

Prominence is often defined in terms of the perceived strength of a speech unit (either a syllable or word) with respect to the surrounding speech (Cole et al., 2010). Relative “strength” is realised acoustically through changes to duration, intensity, and

F0 features—these are often referred to as signal-based factors (Jasmin et al., 2021). However phrasal prominence, as annotated in BURNC is also a function of expectation-based factors such as pragmatic focus in English. For example, listeners often assign prominence to words that introduce novel or valuable information (e.g., that are not anaphorically recoverable). As such, although prosodic factors in isolation are not sufficient for accurately identifying perceived prominence, we expect them to at least reflect the acoustic realisations of prominence.

Following previous works including the original CWT feature implementation, we convert both tasks to binary classification; a word is considered prominent if any constituent syllable carries an accent ($\{H^*, L^*, L^*+H, L+H^*, H+, !H^*\}$) while a boundary is marked as either of the strongest break indexes ($\{3,4\}$) (Suni et al., 2017; Kalinli and Narayanan, 2007; Rosenberg, 2009). For both tasks, combinations of pitch, energy, and duration features have been shown to outperform individual features (Ludusan and Dupoux, 2014; Suni et al., 2017). Using pitch and energy features comparable to ours, Suni et al. (2017) report F1 performances of 0.85 and 0.59 for prominence and boundary detection tasks, respectively. We note that these results are not directly comparable as this paper makes use of a handcrafted classification method and the BURNC dataset doesn't have standard splits for training and evaluation. As such, we reimplement this method to perform more direct comparisons, as detailed in Section 7.4.

Emotion classification. We use the **RAVDESS** dataset (Ryerson Audio-Visual Database of Emotional Speech and Song Dataset), a corpus of 1,440 utterances spoken by 24 speakers in eight emotions, to evaluate emotion classification (Livingstone and Russo, 2018). The speech portion of this multimodal dataset consists of 1,440 utterances spoken by 24 actors (12 female, 12 male) in eight emotions ($\{\text{Calm, Happy, Sad, Angry, Fear, Surprise, Disgust, Neutral}\}$). Emotion classification is a complex task even for humans; we select RAVDESS as emotion labels were validated with a wide battery of listening tests with raters asked to classify the emotion, its strength, and its genuineness. Accuracy in the audio-only setting was moderate (0.62 across all emotions and intensities) with moderate inter-annotator agreement; however, these varied between categories and intensity levels. Although they only test across 4 of the 8 emotion classes, Vu et al. (2022) find an F1 score of 47.11 for RAVDESS using a similar classification network based on GeMAP features. This feature set contains 66 features related to pitch and intensity, as well as other voice production features like jitter, shimmer, and spectral slope (Eyben et al., 2016). Handcrafted prosodic features

(Cao et al., 2014) as well as machine representations of delexicalised speech (Weston et al., 2021) have also proved useful for emotion classification on other datasets.

Another useful feature of this dataset is that all spoken productions are one of two lexically-neutral phases (“Kids are talking by the door” and “Dogs are sitting by the door”). Though our proposed representations are independent of lexical content, this dataset feature would enable closer comparisons with other methods that encode the full speech signal as lexical content is not informative.

7.4 Experimental setup

Corruption strategies. We define corruption strategies as combinations of mask (m) and codebook (c) sizes with $m \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ and $c \in \{4, 8, 16, 32, 64, 128, 512\}$. We expect the value of self-supervision to be bounded by these parameters: small masks may be predicted trivially from the continuous nature of acoustic features, while larger masks will become impossible to predict. Motivated by Joshi et al. (2020), who found that masking randomly-sized contiguous spans of text consistently outperforms masking individual tokens, we additionally propose *random masking* as a generally useful corruption strategy. This involves sampling mask size m uniformly across the set above for each batch to encourage representations that encode dependencies across timescales

Baseline and MPM representations. To examine the value of different encoding strategies, we compare CWT and MPM representations to the unprocessed input features. This allows a comparison between the value of systematicity (i.e., the self-supervised objective of MPM) and hierarchical structure (i.e., encoded by CWT) for downstream tasks. In their original work, Suni et al. (2020) apply a classification mechanism based on lines of min/max amplitude to create a continuous label for each word. To make a fairer comparison across representations, we instead apply the CWT processing to each of our extracted input features (pitch, energy, and voice activity) using 10 scales, concatenate the resulting vectors, and feed them into the same classifiers used for MPM features.

MPM Training We train an MPM under each corruption strategy on the full LibriTTS dataset, comprising over 500 hours of audiobook data (Zen et al., 2019). Each model is trained for 10k steps (≈ 8 epochs, 30 minutes of TPUv3 runtime) with a batch

size of 256, over utterances truncated to a maximum length of 6 seconds. Unless otherwise stated, we extract the representations from layer 8 as representations from intermediate layers of Transformers have been found to be robust across tasks (Yosinski et al., 2014). To allow fair comparison between our representations and the original input features, we don’t perform hyperparameter tuning or architecture adaptations for individual downstream tasks.

Classification probes and training. We train both linear and Conformer (2 Conformer blocks, 5M parameters) classifier probes on each of the downstream tasks for 1000 steps with a batch size of 32. We use the AdamW optimizer and a linear learning rate schedule with a maximum learning rate of 4×10^{-5} and 100 warmup steps. BURNC does not have a standard split for training and testing; given the small size of the dataset, we create a single held-out test set following the approach in Suni et al. (2017) of using speaker f2b for testing only. We augment this test set with 20% of data from all other speakers to make the evaluation set larger and more robust. TIMIT and RAVDESS both provide standard train and test splits for evaluation; for consistency with the BURNC training set up, we use the established test set as a single held-out evaluation set.

7.5 Experiments

7.5.1 Effects of mask and codebook size

The losses on the test sets (averaged and normalized across all tasks) are shown in Figure 7.2 (a). All tasks benefit from a granularity $c \geq 16$. Though RAVDESS classification is more robust to different bin sizes, $c = 128$ results in the lowest evaluation loss across tasks. The TIMIT segmentation tasks are particularly sensitive to the larger bin sizes ($\{512, 1024\}$) while smaller bin sizes are more detrimental for the information structure tasks in BURNC. As expected, the smallest mask sizes ($\{1, 2\}$) generally produce below-average losses across all tasks; models trained with small mask sizes can capitalize on frame-level continuity but may not result in useful learning. Each task displays a different preference for mask size. Prominence and boundary detection and phone segmentation tasks are relatively robust to mask sizes, showing similar losses for masks between 4-128 frames. Boundary detection, however, is more sensitive. Coarse bins are particularly detrimental for this task, perhaps indicating that fine details, likely

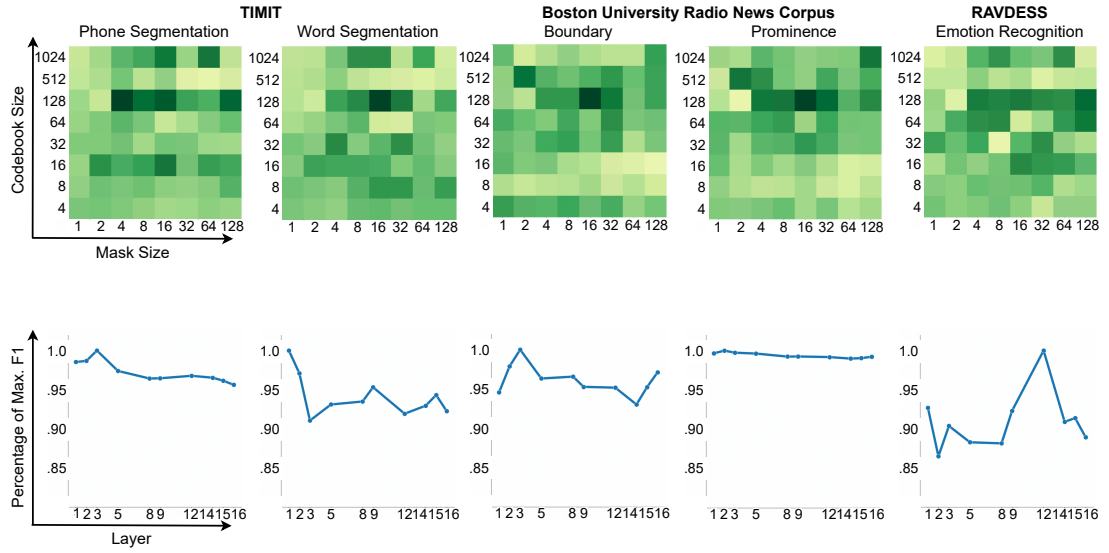


Figure 7.2: (a) The losses on the evaluation set (darker→lower) for the different tasks by codebook and mask size. (b) The F1 scores by layer for each task; Scores are computed using a model with a random mask size and codebook size of 128.

of energy, are important for this task. Word segmentation is more narrowly constrained to mask sizes in the range of 8-32. Interestingly, similar window dimensions are used in representations of the full speech signal (e.g., wav2vec2 Baevski et al., 2020).

7.5.2 Probe performance

In Tables 7.1 and 7.2, we compare the performances resulting from the specific (m, c) strategies that produced the lowest evaluation losses per task as well as the random masking strategy; these are (4,128) for TIMIT phone segmentation, (16,128) for TIMIT word segmentation and both BURNC tasks, and (128,128) for RAVDESS. We train each of these models for an additional 100k steps. Additionally, we compare these to the classification performances achieved using the input features themselves and the CWT method from Suni et al. (2017). These comparisons provide insight into whether other forms of structure exist and at what timescales they may be useful.

Using a linear probe classifier facilitates comparisons between corruption strategies (m, c) and results shown in Table 7.1 confirm that the lowest loss strategy for each task also displays strong F1 performances. In this condition, all MPM representations outperform the input and CWT features and often by a large margin. Similar to MPM results, CWT features provide the most benefit over input features for emotion recog-

Encoding Strategy		TIMIT		BURNC		RAVDESS
MPM Mask m	Codebook c	Phone	Word	Boundary	Prominence	Emotion
4	128	0.46 $\pm$ 0.002	0.14 $\pm$ 0.010	0.53 $\pm$ 0.010	0.71 $\pm$ 0.006	0.24 $\pm$ 0.008
16	128	0.45 $\pm$ 0.002	0.16 $\pm$ 0.002	0.56 $\pm$ 0.001	0.73 $\pm$ 0.006	0.28 $\pm$ 0.020
128	128	0.44 $\pm$ 0.002	0.14 $\pm$ 0.004	0.54 $\pm$ 0.009	0.72 $\pm$ 0.004	0.32 $\pm$ 0.006
Random	128	0.43 $\pm$ 0.002	0.14 $\pm$ 0.002	0.53 $\pm$ 0.012	0.73 $\pm$ 0.004	0.34 $\pm$ 0.008
Input Features		0.42 $\pm$ 0.002	0.13 $\pm$ 0.004	0.38 $\pm$ 0.006	0.60 $\pm$ 0.001	0.07 $\pm$ 0.040
CWT Features		0.44 $\pm$ 0.003	0.15 $\pm$ 0.005	0.40 $\pm$ 0.002	0.62 $\pm$ 0.003	0.16 $\pm$ 0.003

Table 7.1: F1 scores and corresponding 95% confidence intervals obtained from the linear probe on the test set for each task. The table includes probe performance achieved using representations from MPMs trained with a range of corruption strategies, CWT representations, and the original input features. The highest score for each task is shown in bold.

tion, however, MPM representations still far outperform them. This suggests that while the hierarchical encoding provided by CWT is useful for this task, the additional predictable structures extracted by SSL may be more valuable for tasks dependent on longer-term structure. MPM and CWT representations only slightly outperform the input features for both TIMIT tasks and the phrase boundary detection BURNC task; this is in line with our expectation that such tasks don’t require wide contextualization.

TIMIT and BURNC boundary detection tasks enjoy similar benefits from SSL encoding across MPM mask sizes $m = \{4, 16, 128\}$. However, greater F1 differences across mask sizes for emotion classification suggest that some information relevant to this task is only present when large mask sizes are encountered during training. For tasks that are expected to rely more on longer-ranging temporal context—prominence detection in BURNC and the RAVDESS emotion classification—the random mask model performs best overall. The random-mask strategy therefore may encourage encoding of dependencies across a wider range of timescales, although some fine-grained information may be lost.

We test whether the self-supervision of MPM provides value over task-specific contextualisation using a more powerful Conformer probe. Results in Table 7.2 provide further evidence that random masking may not maintain local structures most important for TIMIT tasks. For the TIMIT segmentation tasks, both CWT and input features outperform MPM representations obtained from random masking. Similarly, CWT features outperform the MPM features for BURNC boundary detection. The gen-

Encoding Strategy	TIMIT		BURNC		RAVDESS
	Phone	Word	Boundary	Prominence	Emotion
Random MPM	0.48 ± 0.008	0.23 ± 0.005	0.64 ± 0.008	0.75 ± 0.007	0.43 ± 0.010
Input Features	0.52 ± 0.014	0.26 ± 0.013	0.45 ± 0.007	0.62 ± 0.008	0.36 ± 0.018
CWT Features	0.63 ± 0.004	0.40 ± 0.004	0.67 ± 0.009	0.72 ± 0.006	0.39 ± 0.010

Table 7.2: F1 scores and corresponding 95% confidence intervals obtained from the conformer probe on the respective test sets across all tasks. The table includes probe performance achieved using representations from MPMs trained the random corruption strategy, CWT representations, and the original input features. The highest score for each task is shown in bold.

eral encoding of systematicity from SSL still provides additional value for tasks where context is potentially more valuable—prominence and emotion classification.

Overall, the linear probe suggests that MPM representations are generally useful features. However, training a conformer with task-specific contextualisation demonstrates that fine-grained tasks such as the TIMIT segmentations can be performed more effectively using structures encoded by CWT. MPM representations are, however, much more valuable for performing RAVDESS emotion recognition, suggesting that the hierarchical structure encoded by CWT may not be sufficiently expressive for this task.

7.5.3 Layer evaluation

Substantial work has investigated which types of linguistic representation can be extracted from different layers within a neural network, particularly for the Transformer architecture which forms the basis for many self-supervised networks (Rogers et al., 2020). Across both text and speech, network layers have been shown to follow an “autoencoder” pattern, where the earliest and latest layers are most similar to the input signal and the mid-layers are the most abstract (Hewitt and Manning, 2019; Pasad et al., 2021; Li et al., 2023). For example, syntactic information is most prominent in the intermediate layers while word order is more prominent in the earliest layers (Lin et al., 2019). Lin et al. (2023) finds that representations from earlier layers of SSL speech models are better suited for reconstructing pitch and energy features than later ones. To evaluate how layer depth affects downstream tasks when using the MPM, we use features from layers $\{2, 3, 5, 8, 9, 12, 14, 15\}$ of our random masking model as

input to the linear classifier probes outlined previously. Figure 7.2 (b) shows that earlier layers yield the best performance for all tasks except emotion recognition where performance increases up to layer 12 before dropping. This supports the results from Tables 7.1 and 7.2 that emotion recognition benefits from the longer contextualisation and abstraction in later layers.

7.6 Discussion & conclusions

We have found that self-supervised representations of the acoustic correlates of prosody are useful for predicting a range of perceptual labels known to be at least partial functions of prosodic effects. Along with providing a means to encode prosodic information independently of lexical information, the results of this study indicate that prosody exhibits predictable structure—*systematicity*—independently of lexical content. This systematicity may offer an explanation to our observations in Chapters 5 and 6 that people make predictions about upcoming prosodic realisation and that predictions based on spoken context are more accurate than those conditioned only on lexical context.

Our exploration of model training schemes reveals the multi-scale nature of this systematicity. We find that tasks, specifically chosen to be dependent on prosodic structure across different temporal granularities are better performed by MPM trained with different mask sizes: segmentation tasks benefit from small mask sizes while larger masks and mixed-size masking (i.e., the random strategy) are more effective at learning representations useful for more complex classifications of prominence and emotion. Compared to CWT representations that encode more constrained structural properties of prosodic correlates, we find that the additional flexibility of SSL is particularly valuable for emotion and prominence prediction which likely rely on longer-term, complex dependencies. Our proposed random-mask strategy proves valuable for all of the tested tasks; they capture longer-term structures while maintaining fine-grained structural information.

To the best of our knowledge, we are the first to examine whether longer-term structure of prosodic acoustic correlates can be encoded through SSL mechanisms. However, given that the functions of prosody are tightly coupled with lexical content, it is unsurprising that performance from MPM representations is not competitive with methods that leverage the full speech signal (cf. de Seyssel et al. (2023)). Instead, this work provides insights into prosodic structure independent of lexical content, as well

as further evidence that prosody contributes to these perceptual tasks. Future work could investigate the relative importance of pitch and energy features in this model, as well as more explicit inclusion of duration information which is known to be perceptually important (Aylett and Turk, 2004). Our method also provides a means to explore potential differences in the prosodic systematicity of read and spontaneous speech. We believe that comparing MPM representations with SSL-based representations of the full speech signal could be a fruitful avenue for future works as it may provide insight into both what features of speech are important for particular tasks, as well as what information is encoded within such speech representations.

Chapter 8

Acoustic-conditioned turn discrimination

This chapter consists of an exploratory study into whether neural models can integrate lexical and non-lexical information to perform the turn-discrimination task presented to people in Part II. As such, we pose the following research question:

RQ: *“To what extent are predictive models capable of leveraging non-lexical information for turn discrimination?”*

8.1 Introduction

Our behavioural studies presented in Part II show that people generate expectations regarding both the lexical content and prosodic realisation of an upcoming conversational turn, and condition their expectations on both lexical and non-lexical context. Access to these channels affects both turn-discrimination accuracy and agreement across participants. In this chapter, we present a proof-of-concept study asking whether models respond in similar ways.

The experimental design is motivated by our findings from previous chapters of the thesis. In particular, we modify the response-selection training objective, shown in Chapter 4 to produce representations that align to a certain degree with human expectations of lexical content, to mirror the behavioural studies from Part II. Models are trained to discriminate between true and false (*context*, *response*) pairs where the *context* consists of a lexical encoding and the *response* is represented by the prosodic features presented in Chapter 7 which were found to be predictive of several perceptual labels of speech. Importantly, our experimental design isolates lexical and non-lexical

channel information to provide an alternative method for quantifying the relative contributions of these channels in shaping expectations about upcoming speech.

While the results obtained in the proof-of-concept setup presented here do not allow us to answer our research question fully, they point to potentially fruitful directions for future work.

8.2 Related work

The question of whether neural models of speech are affected by prosodic realisation in ways that mirror human speech comprehension is only starting to be addressed in the wider research field. There is growing evidence that SSL models of speech do encode certain aspects of prosodic realisation (Ekstedt and Skantze, 2022; de Seyssel et al., 2023; Lin et al., 2023; de Seyssel et al., 2023), and that language models of text can integrate prosodic features to alter their performance on downstream tasks (Tran, 2020; Nielsen et al., 2023; Murray et al., 2005; Moon et al., 2018; Kaffle et al., 2019). Although none of these studies consider prediction at the level of dialogue turns, they serve as important groundwork for our investigation.

8.2.1 SSL speech representations

In Section 2.2.3, we introduced some popular SSL-based representations of speech and highlighted that their original evaluations were heavily focused on encoding information that is relevant for speech recognition such as phonetic content. Below, we discuss recent works that propose using different downstream tasks to evaluate whether these representations also encode prosodic information.

de Seyssel et al. (2023) propose a benchmark to evaluate whether neural models of speech encode prosodic information related to information structure. By modifying the placement of pauses in utterances of read speech, they test whether models can distinguish strong and weak prosodic boundaries (“protosyntax” task), and pauses occurring between and within words (“lexical” task). For each pair of modified and unmodified stimuli, the likelihood of generating each stimulus under the speech model is computed¹. Importantly, model performance is compared to human responses. Interestingly, they report quite large differences in the accuracy of lexical and protosyntax

¹In this work, the likelihood of a stimulus is defined as the product of chained probabilities for each unit $[u_1, \dots, u_T]$ in the encoded sequence of speech: $P(u_1, \dots, u_T) = \prod_{i=1}^T P(u_i | u_1, \dots, u_{i-1})$.

tasks across both models and humans: humans are far more accurate than models for the protosyntax task, however the simpler model in their test setup outperforms its pre-trained counterpart; while models perform similarly for the lexical task, they outperform humans by a large margin in terms of accuracy. These results provide evidence that SSL models do encode some structural information conveyed through prosodic realisation (i.e., duration of pauses), but they do so quite differently to people. In particular, super-human performance suggests a strong divergence between what a model learns as likely and what people find to be “natural”. Additionally, the underlying mechanics of the model have a strong, and sometimes unexpected impact on performance.

Other benchmarks include de Seyssel et al. (2023) who consider whether models resynthesize emphasis accurately, and Lin et al. (2023) who explore whether models encode prosodic information that is valuable for pragmatic functions. Lin et al. (2023) propose five tasks: along with casting sentiment analysis, sarcasm detection, and persuasion prediction into classification tasks, they assess reconstruction and prediction of pitch and energy features (“pseudo-tasks”). They compare performance across a suite of SSL models. The input for classification tasks consists of mean-pooling the SSL representation of a stimulus along the time axis. For the pseudo-tasks, a linear model is used to predict a pitch or energy feature from a frame-level SSL representation. Perhaps unsurprisingly given that they are trained with some form of predictive objective, most models outperform the acoustic baseline of log mel filterbanks for the pseudo-tasks. As demonstrated by a generally strong text-only baseline using RoBERTA features, lexical information is useful for all of the classification tasks. For sentiment and persuasiveness classification, models do not consistently outperform this baseline. However, these results provide further evidence for the marked differences *between* models across tasks. For example, certain models do not beat the acoustic baseline for sarcasm detection while others improve on the previous audio-only state-of-the-art results.

A final example is from Ekstedt and Skantze (2022) who investigate how well turn-taking activity can be predicted in diadic interactions from the representations of SSL speech models. Using a speech model that encodes the combined waveform of two speakers as input, they demonstrate that their model can be trained to predict upcoming turn-taking behaviour. Importantly, their analyses show that these predictions are sensitive to acoustic manipulations of the original speech that alter its prosodic realisation (e.g., flattening F0 and intensity, altering the duration of phones, and obscuring phonetic information). The model is most sensitive to the removal of phonetic infor-

mation through low-pass filtering; however, flattening intensity and F0 contours also negatively impact turn-end prediction performance. They find F0 to be the most reliable cue when the completeness of a turn is ambiguous from lexical content. These analyses involved only modifying the test data, therefore divergence between the training and test data may affect the results, but they suggest that predictive models can leverage prosody to model upcoming speech more accurately.

Taken together, these studies demonstrate that neural pre-trained representations of speech can be used to predict several perceptual labels which are driven in part by prosody, suggesting that at least some information pertaining to prosodic effects is contained in the representations derived from self-supervised learning mechanisms. However, these findings don't provide explicit insights into the relative contributions of prosodic and lexical information, or their product. Overall, results suggest that these models encode prosodic features quite differently to people and that the learning mechanisms driving the encoder have strong impacts on which aspects of prosody are encoded.

8.2.2 Combining acoustic and lexical representations

Other works have investigated whether automatic systems can leverage acoustic features alongside lexical information for a number of downstream tasks; these include parsing (Tran, 2020; Nielsen et al., 2023), summarizing meetings (Murray et al., 2005), named entity recognition (Moon et al., 2018), and predicting word importance (Kafle et al., 2019), among others. Together, these works provide evidence that models *can* make use of isolated information sources in conjunction with each other, and thus motivate our exploration in whether statistical models can learn appropriate upcoming prosodic realisations from lexical context. However, they also highlight that the value of prosodic information depends on the task and type of stimuli. We discuss a small number of these works below.

In a series of works summarized in their thesis, Tran (2020) proposes a methodology to integrate prosodic features with lexical representations for constituency parsing and dialogue act recognition of speech. They propose an encoder that functions at the word-level; a word $w_i = [e_i; \phi_i; s_i]$ is represented as its lexical identity e_i , its duration and preceding pause ϕ_i , and a representation of its energy and pitch features that is learned jointly with the overall encoder s_i . Though the representations of prosody are task-specific, prosodic information produces performance gains, albeit

modest ones, for both tasks. Since then, more powerful models of lexical content have outperformed these models using only lexical content (Jamshid Lou et al., 2019); however, the conclusions drawn by Tran (2020) indicate that the value of prosody depends on features of the input. For dependency parsing, prosody produced greater benefits for longer and more disfluent sentences; while for constituency parsing, including prosodic information produced significantly greater error reduction for ASR transcripts compared to gold-standard transcripts. Using a similar modelling setup, Nielsen et al. (2021) demonstrates the value of prosody from another perspective. The parsing task tested by Tran (2020) involves generating a parse tree for individual sentences. Instead, Nielsen et al. (2021) consider whether prosodic features are useful for parsing speech that hasn't been segmented into sentences. Prosodic information produces additional benefits when sentence segmentation and parsing tasks are performed jointly.

Other works show similar findings regarding how the value of prosody depends on its wider context. Using word importance annotations from Switchboard, Kafle et al. (2019) investigate whether the addition of prosodic information increases a model's predictive power over lexical content alone. Their method involves learning attention-based edits to word embeddings (pretrained GLoVE (Pennington et al., 2014)) based on acoustic features. Although the word importance labels in Switchboard were obtained from annotations of the transcript rather than the audio², Kafle et al. (2019) find that combining lexical and acoustic features achieves better performance than either modality alone. Much of this increased performance seems to be centred on words that fall outside of the GLoVE embedding vocabulary (out-of-vocabulary OOV), indicating that the value of prosodic realisation is dependent on the underlying lexical content.

Overall, the fact that different models can benefit from relatively low-level, localised prosodic features reflects the evidence for its contribution to human speech comprehension. However, it is hard to know if the limited benefits are a function of the evaluation tasks themselves—e.g., single sentence parsing of transcripts is likely achievable from powerful enough lexical representations (Nielsen et al., 2023), and the word importance labels discussed above were obtained from judgements of text only (Kafle et al., 2019). Human communication is far more complex: we perform many complicated, and often simultaneous, tasks during language comprehension. The benefits of prosody may thus be obscured when considering specific, constrained tasks.

²Annotators were asked to rate importance based on how their understanding of the utterance would be affected if the word were masked.

We believe that the turn discrimination task presented in Part II provides another opportunity to highlight the value of prosody in realistic language comprehensions.

8.3 Methodology

As a proof-of-concept study, we mirror the task presented in Chapter 5 and train a response selection model to leverage representations of conversational context to discriminate between different prosodic realisations of potential responses with the same lexical content. Our goal is to present a model that is suitably parameterized to perform this task.

Lexical representations It is well-established (by others as well as ourselves in Part I) that utterance-level lexical representations such as SBERT are useful for response selection. As such, these representations are good candidates for containing information that is useful in predicting other features of the upcoming turn.

Prosodic representation In Chapter 7, we demonstrated that prosodic representations, based both on MPM and CWT features, encode information that is useful for utterance-level tasks like prominence detection and emotion classification. As such, we have good reason to expect these prosodic representations to provide information valuable for turn discrimination.

8.3.1 Experimental setup

We train a model following the response selection setup introduced in Chapter 4 which consists of a response selection cross-encoder architecture: a model encodes (*context*, *response*) pairs (both congruent and negative pairs) as a single embedding which is used to produce a corresponding score. The training objective is to minimize the binary cross entropy across (*c*, *r*) scores—in other words, the model is encouraged to produce high scores for congruent pairs and low scores for negative pairs.

We use a held-out portion of the training data for validation and early stopping where best validation performance is evaluated as the mean of the top-*k* accuracy scores up to $k = 3(\frac{1}{3} \sum_{i=1}^3 R_{k=i})$. To mitigate the effects of the relatively unstable validation performance that we obtained during training, we evaluate ensembles of models by aggregating the predictions from the model checkpoint at the epoch where the best

validation performance was achieved and at ± 2 epochs. We take the mean of their predictions as the final (c, r) score. Final evaluations are done using another held-out set. Additionally, we evaluate performance on the stimuli that were presented to participants in Chapter 5; we refer to this set as IS'21.

Following Han et al. (2021), the training set is constructed such that each positive (c, r) pair has a single corresponding negative instance. Negative instances are created by randomly pairing the context with a different response. Evaluation sets use 3 negative samples for every true one to match the stimuli presented to people in Chapter 5.

Model architecture We use a 4-layer MLP (GeLU activation with dropout) followed by a single linear classification layer as a PoC model—a total of approximately 70000 trainable parameters.

Input $(context, response)$ pairs are represented by concatenating SBERT representation of context and a pooled prosodic representation of response. We found SBERT representations of conversational turns were useful for discriminating the next upcoming turn in Chapter 4 while pooled CWT and MPM representations were useful for utterance-level emotion classification in Chapter 7. We experiment with CWT-based representations first before comparing the two methods.

SBERT features are the product of pooling BERT representations that are fine-tuned to be useful utterance-level representations; the context representation is 384 dimensions. MPM and CWT features are extracted in the same way as in the previous chapter: CWT features are produced from convolving 10 wavelets over pitch and intensity features normalised at the response level while MPM features are extracted directly from the model described in Chapter 7. Aggregate prosodic representations are simply a concatenation of the mean and max of the frame-level CWT features across the response utterance. We used the same simple aggregation strategy for BURNC prominence and break detection where words were represented as the mean of their frame-level CWT or MPM features, and in RAVDESS where full utterances were represented as the concatenation of mean and max operations over the frame-level features in Chapter 7.

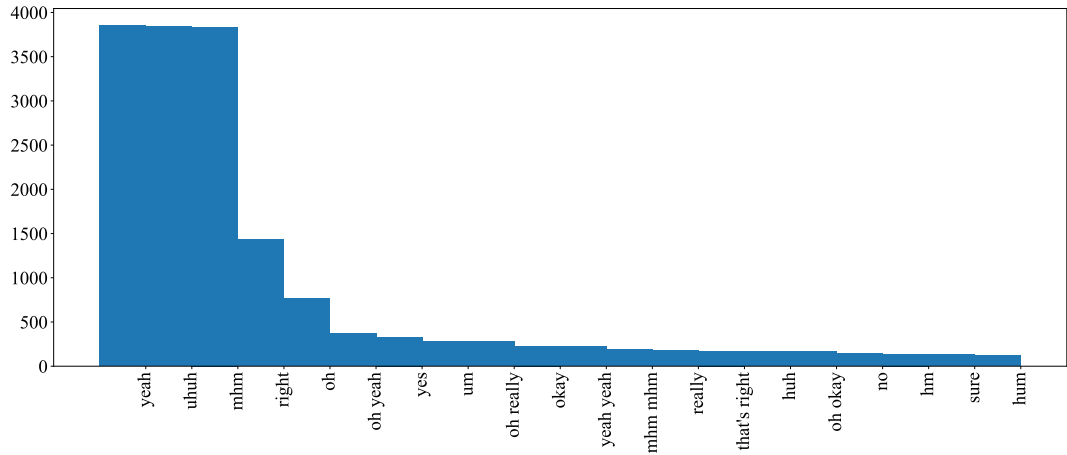


Figure 8.1: Frequency distribution of the most frequent dialogue responses in our segmentation of SWITCHBOARD utterances.

8.4 Experiment I: individual responses

We begin by training models on samples consisting of single lexical responses. Although it strongly limits the amount of data, the constrained problem setup allows us to investigate to what extent this model architecture can leverage relatively conventionalised prosodic variation from its $(context, response)$ inputs. Additionally, we examine how model performance and learning evolve with respect to the size of the training sets.

The most frequent responses in the SWITCHBOARD utterance set are shown in Figure 8.1. We consider the following responses $\{“yeah”, “no”, “oh really”, “mhm”\}$, a mix of frequent responses which allow us to verify appropriate learning dynamics and a range of underlying semantics.

For each set of responses, we split the total number of true samples into training and validation sets before generating negative samples in each respective set. We train models using different proportions of the train set; we consider $\{150, 250, 500, 1000, 4000\}$ true samples per set (as shown in Figure 8.1, only “yeah” and “mhm” have sufficient frequency for the largest two dataset sizes).

8.4.1 Results

Table 8.1 reports the best $R_{k=1}$ accuracy achieved for each response and dataset size over the validation set, the corresponding epoch at which the performance was achieved, and the generalised performance on the held-out test set. Through our be-

Datasize		4000			1000			250			150		
Word	Run	test	val	epoch	test	val	epoch	test	val	epoch	test	val	epoch
“yeah”	1	0.262	0.304	17	0.282	0.304	9	0.280	0.234	65	0.267	0.244	93
	2	0.277	0.275	82	0.220	0.249	40	0.236	0.283	20	0.264	0.278	41
“mhm”	1	0.292	0.292	6	0.260	0.266	26	0.237	0.224	16	nan	nan	0
	2	nan	nan	0	0.253	0.260	2	0.255	0.237	1	0.195	0.229	24
“oh really”	1	–	–	–	–	–	–	0.214	0.214	33	0.321	0.179	37
	2	–	–	–	–	–	–	0.179	0.143	19	0.214	0.214	45
“no”	1	–	–	–	–	–	–	–	–	–	0.133	0.286	36
	2	–	–	–	–	–	–	–	–	–	0.133	0.286	31

Table 8.1: Turn discrimination performance: best $R_{k=1}$ ensembled validation accuracy and the corresponding epoch across responses and dataset sizes. The performance for two model initialisations is shown for each training set. Note that accuracy at chance is 0.25. Best test accuracy per response are shown in **bold**. Note that “nan” corresponds to runs where the best performance was achieved in the first epoch of training, indicating that the model didn’t learn to generalise.

haviour studies in Part II and the “superhuman” performance achieved by lexical representations in Chapter 4, we have argued that accuracy is not the most useful metric for analysing whether models encode perceptually relevant metrics; however, we use it here as a broad evaluation of training dynamics.

Learning dynamics

All models could overfit to their training set, reflected by near-perfect performance, both in terms of training loss and F1 scores, by the end of training. Models trained from smaller datasets converge after ≈ 100 epochs while larger datasets require > 150 epochs.

Effect of training set size across responses

As we might expect, generalisation of the models improves slightly with training set size as reflected in better performance (e.g., “mhm” shows a steady increase in $R_{k=1}$ from 0.195 to 0.292 on the held out test set as training size increases from 150 to 4000 positive samples), as well as smaller differences in performance between the validation and test sets. However, these are not consistent across response sets or model initialisations.

Accuracy across response sets

Although we cannot make direct comparisons between the validation sets for different responses, it is interesting to note that “*oh really*” achieves the lowest validation performance across responses where top performance is never greater than chance level, suggesting that it may be the most challenging set for distinguishing the true prosodic realisation from the negative examples. Conversely, “*yeah*” consistently achieves relatively high validation and test set accuracy across dataset sizes. Differences in the effect of prosodic realisation on discrimination performance are supported by works such as Lai (2010) who find that the effect of prosodic realisation on the interpretation of discourse markers including “*yeah*” differs based on lexical content.

Even in this extremely difficult instantiation of the turn discrimination task, and constrained training setup where the amount of data is severely limited, $R_{k=1}$ accuracies for both of the held-out sets (validation and test) tend to be slightly above chance performance of 0.25, especially for “*yeah*” and “*mhm*” responses where larger training sets are available. Recall that human performance across varied responses was 0.346, where people could hear the full response. These results suggest that the relatively simple representation of prosodic realisation used in this PoC may contain information that is useful in conjunction with utterance-level representations of the previous turn for discriminating between true and negative prosodic realisations; however, additional information may be required to increase the discriminative performance.

8.5 Experiment II: variable responses

We make incremental changes to the task setup to produce conditions closer to those in the human turn-discrimination paradigm by expanding the complexity of the training data. In doing so, we aim to evaluate the impact of different types of information and training schemes on the overall performance; modifications are described below.

Training a model with multiple lexical responses. We test two strategies for building a training set that contains responses with varying lexical content: TOP-4 or MIX. For TOP-4, we maintain a balanced proportion of lexical responses in the training set by selecting the four most frequent responses from the set of SWITCHBOARD utterances displayed in Figure 8.1: {“*yeah*”, “*uhuh*”, “*mhm*”, “*right*”}. For the MIX dataset, half of the training data consists of a balanced set of the top four responses

while the other half is sampled from the remaining 20 most frequent responses. Note that negative responses for each $(context, response)$ pair still have the same lexical content as the true response. This provides a more varied training set while maintaining a fixed set of lexical identities.

Varying the amount of lexical context. In the human experiments, we tested how people’s acceptability judgements were affected by the amount of lexical context they were provided with. To replicate this, the SBERT representation of context consists of either $\{1, 4, 7\}$ turns preceding the response.

Including information about the lexical identity of the response. We include an input feature that maps to the lexical identity of the response to allow the model to capture the effect of prosodic realisation in conjunction with lexical identity (Lai, 2010).

Altering the number of negative samples. We test the impact of training with more negative samples which allows us to expand the size of the training set in a different way than sampling more positive samples.

Altering prosodic representations. We compare the performance of models trained using CWT and MPM representations of the prosodic realisation of responses.

Along with the held-out evaluation sets, we make use of the IS’21set constructed in Chapter 5 which has associated human judgements. Chance performance for these models is $R_{k=\{1,2,3\}} = 0.25, 0.50, 0.75$.

8.5.1 Results

Dataset size

Table 8.2 shows that dataset size does have slight impacts on the accuracy that a model can achieve on held-out test sets: the best ensembled validation and test set accuracies are found for the largest training sets. As expected, the gap between the validation and test set performances decreases as dataset size increases. This gap is similar across both M and L dataset sizes but larger for S , indicating that this dataset size may be insufficient to allow generalisation.

Dataset	Epoch	$R_{k=1}$ across evaluation sets			
		val (<i>indv.</i>)	val	test (<i>indv.</i>)	test
1000 S	16	0.270	0.283	0.223	0.223
4000 M	94	0.318	0.302	0.280	0.273
8000 L	59	0.280	0.305	0.280	0.280

Table 8.2: Effect of dataset size on the learning dynamics (i.e., the epoch at which best validation performance was obtained) and corresponding validation performance for the turn discrimination task. Models are trained on the TOP-4 set of frequent responses (context = 1, no lexical id). Results show performance evaluated with a single model (“*indv.*”), as well as ensembled performance. The training condition leading to the highest accuracy in each evaluation set is shown in bold.

Context effects

Recall that human experiments in Chapter 5 showed no significant relationship between participant accuracy and the amount of context available when context was presented in a purely lexical format. Here, we evaluate whether the model shows a similar pattern or whether additional context can serve as a cue for turn discrimination. We may expect different trends as the model has access to a less informative prosodic realisation than human participants. Longer contexts are encoded with SBERT using a special token to separate turns, as was done in Chapter 4.

The effect of different numbers of context turns on model accuracy is shown in Table 8.3a for the M dataset size. The differences between validation and test set performance in Table 8.2 suggests that training with the M dataset size obtains comparable generalisation ability as the L set; both S and L training performance are included in Appendix B Tables B.1 and B.2 for completeness. For accuracy on the held-out evaluation sets (both test and validation sets), the best performance is achieved in the shortest context condition. This is particularly true for the M dataset size, where the differences in test and validation performance between context conditions is greatest. Is’21 accuracy is unstable across context conditions. Though the 4000 dataset shows a potentially positive relationship with the amount of context, this may be a feature of different model initialisations. Table 8.3a shows a large variation between the two initialisations that were tested for each context length condition. Performance tends to

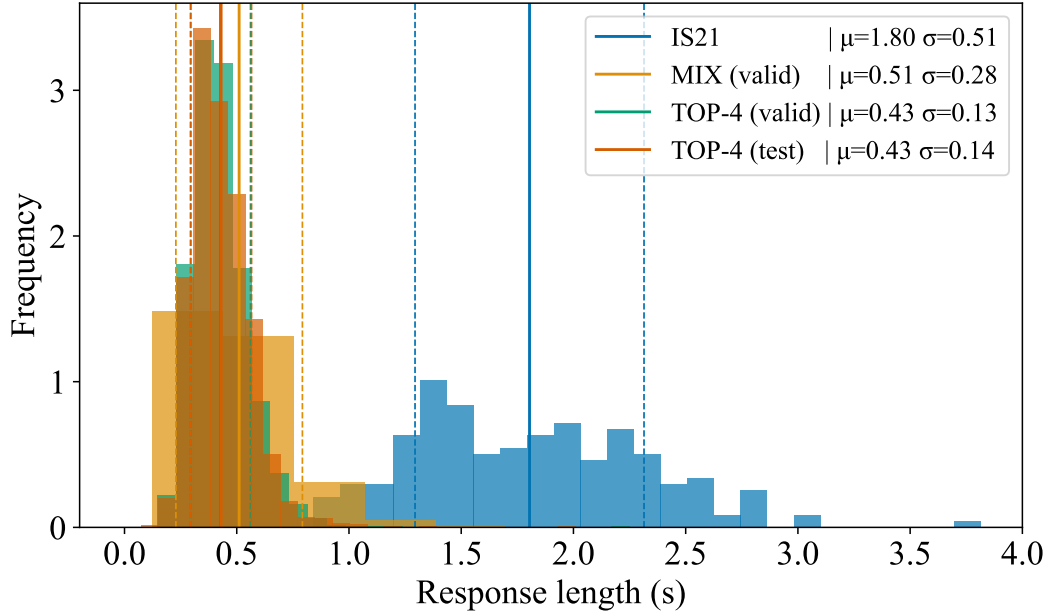


Figure 8.2: Distribution of response length in seconds across the validation sets.

be lower than chance for the Is’21 set. This may be expected in this training setup as Is’21 is quite different to the training data. More concerning is that the test accuracy scores are not consistently above chance, highlighting the difficulty of this task.

In summary, accuracy scores on held-out sets across context conditions are not consistently above chance, highlighting the difficulty of this task. Although we find some model initialisations obtain $R_{k=1}$ accuracy scores that are close to human performance on the Is’21 ($\{0.346, 0.329, 0.349\}$ for $\{1, 4, 7\}$ turns, respectively), these results are not obtained reliably. One explanation could be the difference between the TOP-4 training condition and the Is’21 evaluation set which contains a much broader set of lexical responses. Although the lexical content of responses in each of the Is’21 stimuli was fixed, participants were also aware of this lexical content. We address both of these differences below.

MIX training data & lexical identity

The Is’21 evaluation set contains a much broader set of responses than the TOP-4 training data: Figure 8.2 demonstrates how the evaluation sets vary in terms of the length of responses. This could explain the gaps between the accuracies achieved on the test and Is’21 sets (though the direction of the gap doesn’t hold for all model initialisations) displayed in Table 8.3a. We expect that the wider set of lexical responses

Context turns	Epoch	val	test	Is'21		
		$R_{k=1}$	$R_{k=1}$	$R_{k=1}$	$R_{k=2}$	$R_{k=3}$
1	94	0.302	0.273	0.180	0.400	0.620
1	94	0.302	0.273	0.180	0.400	0.620
4	37	0.297	0.258	0.240	0.560	0.840
4	4	0.260	0.238	0.180	0.420	0.700
7	100	0.255	0.228	0.300	0.500	0.800
7	21	0.268	0.250	0.320	0.620	0.800

(a) No lexical identity in input

Context turns	Epoch	val	test	Is'21		
		$R_{k=1}$	$R_{k=1}$	$R_{k=1}$	$R_{k=2}$	$R_{k=3}$
1	13	0.285	0.235	0.200	0.420	0.720
4	37	0.285	0.283	0.280	0.540	0.840
7	61	0.288	0.263	0.200	0.440	0.740
7	32	0.273	0.285	0.260	0.580	0.780

(b) Lexical identity in input

Table 8.3: Turn discrimination performance for models trained on the TOP-4 (M) dataset. Multiple runs for each context condition are evaluated; we report the number of context turns included as input, the epoch at which the best validation performance was obtained, and the ensembled performance on the validation, test, and Is'21 sets. The training condition that produces the highest accuracy in each validation condition is bolded.

in the MIX dataset will make the discriminative task more difficult to learn; however, it increases the similarity between the training and Is'21 evaluation sets. As such, Is'21 performance may be higher for this training condition (or at least closer to the held-out sets).

Additionally, we test the impact of including information about the lexical content of the response. Because the set of responses across all training sets is finite and relatively small, we append a scalar *lexical identity feature* to each input vector to identify the lexical content of the response.

Tables 8.3a and 8.4a show model performance without a lexical identity feature while Tables 8.3b and 8.4b show results when the lexical identity feature is included. The lexical identity feature seems to facilitate training in the MIX condition, reflected

Context turns	Epoch	val	test	Is'21		
		$R_{k=1}$	$R_{k=1}$	$R_{k=1}$	$R_{k=2}$	$R_{k=3}$
1	0	0.275	0.210	0.200	0.420	0.700
1	0	0.275	0.205	0.160	0.420	0.680
4	0	0.263	0.225	0.160	0.480	0.680
4	32	0.275	0.285	0.340	0.480	0.820
4	3	0.275	0.230	0.160	0.460	0.660
7	17	0.248	0.245	0.200	0.500	0.800
7	102	0.248	0.250	0.260	0.480	0.720

(a) No lexical identity in input.

Context turns	Epoch	val	test	Is'21		
		$R_{k=1}$	$R_{k=1}$	$R_{k=1}$	$R_{k=2}$	$R_{k=3}$
1	4	0.285	0.245	0.200	0.540	0.760
1	32	0.258	0.278	0.340	0.520	0.780
4	4	0.293	0.248	0.180	0.560	0.760
4	5	0.253	0.245	0.260	0.440	0.720
7	5	0.265	0.245	0.240	0.600	0.760
7	1	0.290	0.235	0.200	0.560	0.780

(b) Lexical identity in input

Table 8.4: Turn discrimination performance for models trained on the MIX (M) dataset. Note that runs where validation performance didn't improve after the first epoch (i.e., the best Epoch was 0) are greyed out. The training condition that produces the highest accuracy in each validation condition is bolded.

by fewer model initialisations failing to learn (i.e., obtain best validation performance at epoch 0) compared to conditions without a lexical identity feature. We may expect MIX models to produce better $R_{k=1,2,3}$ on IS'21 than TOP-4 alone; however, we find a similarly broad range of accuracy scores across model initialisations and context conditions across these two datasets.

When the input is augmented with the lexical identity feature, accuracies on the TOP-4 held-out evaluation sets don't change significantly ($R_{k=1}$ for the test set improves slightly with the lexical identity feature but this doesn't hold for $k = [2,3]$). This is perhaps surprising as we may expect the training task to be easier with the lexical identity feature. It's difficult to hypothesize about the impact on IS'21 evaluation, as many of the responses in this set are still unseen during training. However, the lexical identity feature does seem useful in the MIX training conditions: fewer models fail (i.e., generalise best before training has started) and validation performance generally increases. Even so, the performance across the test and IS'21 sets does not change significantly.

Overall, we find that the additional lexical identity feature seems to facilitate training in the MIX condition but not TOP-4. This confirms our expectations that MIX is a more difficult dataset to learn from than the TOP-4 set, and that capturing interactions between lexical identity and prosodic realisation is a useful cue. However, models trained in with the MIX data condition only obtain slightly higher accuracy on the IS'21 set and do not do so consistently.

Number of negatives

The effect of increasing dataset sizes is displayed in Table 8.2; however, these sizes are still heavily constrained by the number of frequent responses in SWITCHBOARD. As an additional investigation into whether the limited amount of training data has capped performance, we increase the number of negative responses sampled per true (*context, response*). Contrastive learning paradigms often involve training with far greater ratios of negative samples for each positive training sample than the balanced split used in our experiments (van den Oord et al., 2018), however this requirement has recently been called into question (see (Mitrovic et al., 2020)). Table 8.5 displays the evaluation results for different sizes of datasets generated with 9 negatives per positive sample. Although the model trained on the smaller dataset (4000 samples) produces high $R_{k=\{1,2,3\}}$ accuracies for IS'21, this increased performance is not consistent across the held-out evaluation sets, and the performances are not significantly greater than to

Dataset size	Epoch	val	test	Is'21		
		$R_{k=1}$	$R_{k=1}$	$R_{k=1}$	$R_{k=2}$	$R_{k=3}$
4000	111	0.268	0.243	0.380	0.760	0.900
20000	49	0.253	0.270	0.220	0.600	0.780

Table 8.5: Turn discrimination performance for models trained on the TOP-4 training set with an increased proportion of negative samples. We use a ratio of 9:1 negative to true (*context*, *response*) pairs in the training set, but use the original evaluation sets with a 3:1 ratio. Models are trained with 4 context turns.

those achieved using our standard negative, positive sample ratio in Table 8.3b.

Though we can only draw speculative conclusions from this experiment, we find some performance increases from increasing the size of the training set. The improvements are unstable; however, they may support our earlier findings that the amount of data is holding back performance on this task.

MPM prosodic representations

The Masked Prosody Model (MPM) representations introduced in Chapter 7 achieved higher performance on emotion recognition at the utterance-level than CWT features. As such, we use them here as an alternative prosodic representation of the response.

To do so, we use the same method described in the preceding chapter to generate a response representation: after normalising the pitch and intensity features within the minimum and maximum bounds of the pretrained MPM model, the acoustic inputs are quantized and encoded into frame-level features. We extract features from the 8th intermediary layer of the model. The response representation is then generated by concatenating the maximum and mean of these frame-level features. Compared to CWT features which are computed using 10 wavelets, MPM represents features in a 512-dimensional space; as such, we adjust the learning rate to ensure that the model can converge with higher-dimensional input feature space.

Table 8.6 shows the results of training models in the MIX data condition with appended lexical identity features. We do not find significant differences between these models and those trained with CWT features in the same training condition; although validation and test performance is slightly lower for the MPM features, the $R_{k=[1,2,3]}$ scores on the Is'21 set are generally higher with MPM input features.

Context turns	Epoch	val	test	Is'21		
		$R_{k=1}$	$R_{k=1}$	$R_{k=1}$	$R_{k=2}$	$R_{k=3}$
1	93	0.225	0.230	0.340	0.640	0.760
4	15	0.245	0.265	0.240	0.520	0.760
7	43	0.248	0.238	0.280	0.580	0.820

Table 8.6: Models trained on the MIX (M) dataset with MPM representations and lexical identity features.

Overall, the MPM prosodic representations don't result in significantly better performance than CWT features.

8.6 Comparing to human judgements

We have not found a response selection model to discriminate between true and false responses reliably. However, as we have argued throughout this thesis, accuracy only reflects a small component of how information can affect human expectations. As such, we also take a more qualitative approach to comparing predictions from models and human participants:

1. Do they make similar mistakes to people (i.e., do they find the same questions easy)?
2. Do they tend to rate responses similarly?

To avoid selecting models that have overfit to the test set, we randomly select one model run from each of the TOP-4 and MIX training conditions where the model was trained with 4 context turns and the lexical identity feature.

As can be seen in Section 8.6, the stimuli in which the true response is more obvious are not the same for human listeners and our models. Surprisingly, the TOP-4 model only correctly predicts three of the ten “easiest” stimuli for participants, while the MIX model only predicts one. There is a slight trend suggesting that the responses that were correctly predicted by both models are short, however, this is not consistent.

We also consider whether models and people tend to rank the responses similarly, regardless of the true response, by computing Spearman rank correlations between the response scores from the models and the average score from participants. We find no

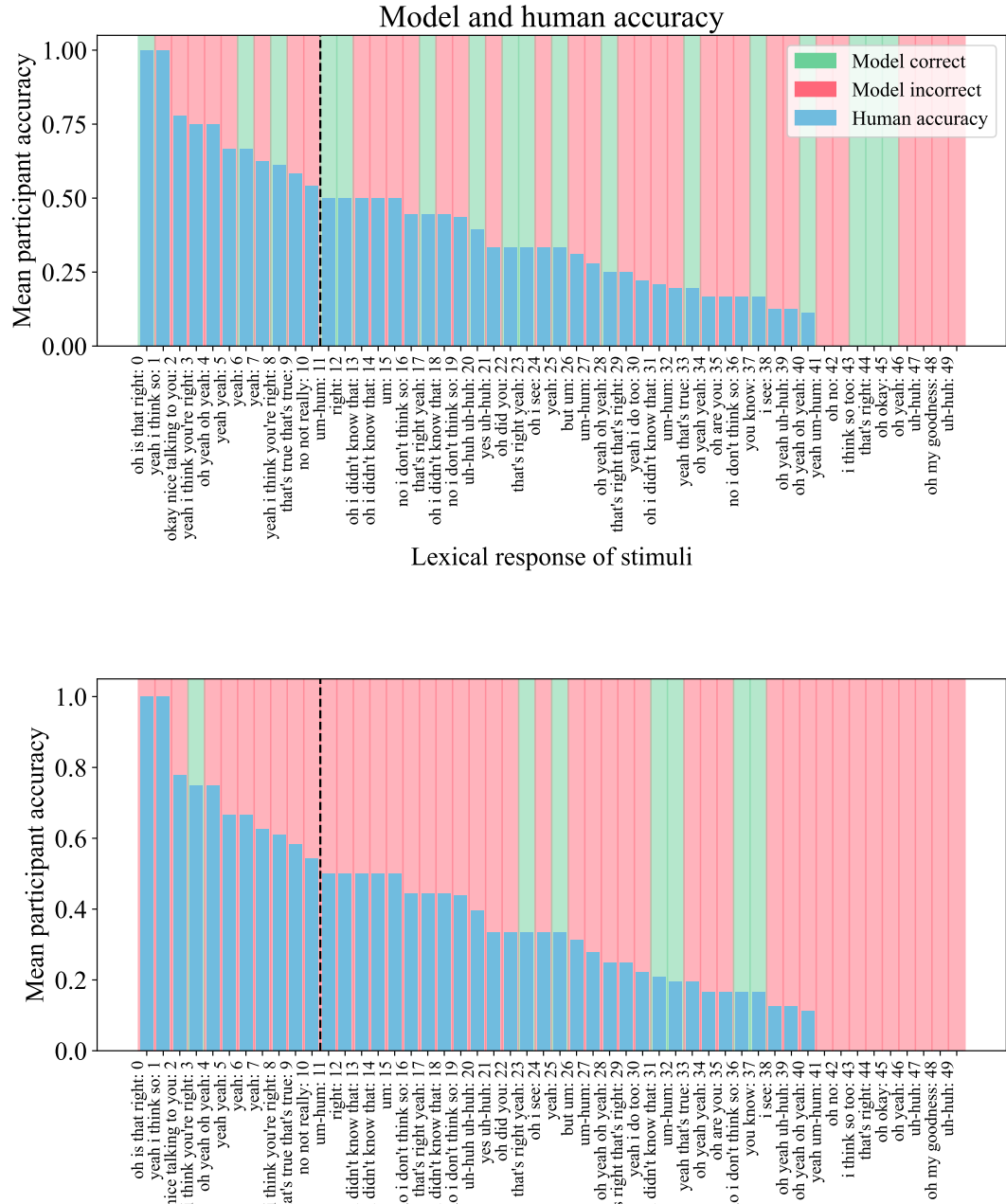


Figure 8.3: The accuracy across individual stimuli for both human participants and (a) TOP-4 (b) MIX models. The dotted line marks the top 10 questions across which participants were most reliably correct.

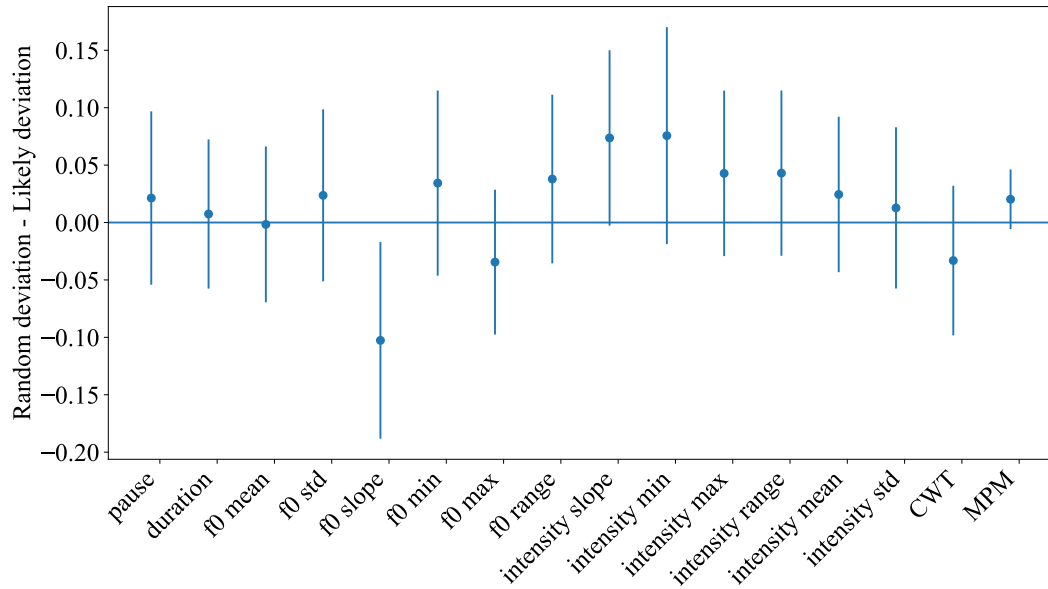


Figure 8.4: Similarity in acoustic feature space for samples scored similarly compared to acoustic similarity expected by chance. As was done in Chapter 5, we compare the average deviation between responses scored as *likely* and all responses in a stimulus; error bars correspond to 95% confidence intervals for the estimated mean across stimuli.

correlation; the MIX model obtains $\rho = -0.11$ (p value=0.98) across all responses and an average correlation of $\rho = -0.02$ across individual stimuli, while the TOP-4 model we find similar results $\rho = -0.12$ (p value=0.09) across all responses and $\rho = 0.05$ over stimuli. The lack of relationship confirms that models do not reflect human-like behaviour.

8.7 Acoustic analysis

As an alternative explanation for why our models don't learn to discriminate between prosodic variants, we return to Chapter 5 where we asked whether responses that were scored as similarly plausible in a context (i.e., rated as likely or very likely) tended to be closer in terms of acoustic realisation. To answer this question, we compared the deviation between sets of responses rated similarly and the deviation between the full set of responses across macro-acoustic features. The feature deviation across all responses is the deviation we could expect if ratings were completely random. We perform the same analysis using both the CWT and MPM response representations,

measuring deviation as the pairwise cosine distance between response representations.

Figure 8.4 shows the difference between the deviation of the samples rated as likely and the expected deviation: a difference of 0 indicates that “likely” responses are no more similar than random, while a deviation > 0 suggests that “likely” samples are more similar than would be expected by chance. Similar to the macro acoustic features originally presented in Chapter 5, we find that likely responses are not significantly more similar than would be expected by chance. In other words, responses judged to be likely continuations are not necessarily similar in their acoustic realisations. Given that in Chapter 7, we found that the prosodic spaces parameterized by CWT and MPM features contain perceptually relevant information, the present analysis provides even stronger evidence that the plausibility of prosodic realisations is not driven purely by their acoustic similarity. This complicates the discrimination task dramatically.

8.8 Discussion & conclusions

In Chapter 5, we asked participants to rate the acceptability of upcoming turns in a dialogue, based on variation in the prosodic realisations of the turns. We found that people can discriminate between true and negative turns; however, their judgements are not always accurate and people often find multiple potential turns acceptable. In the current study, we evaluated a proof-of-concept setup testing first, whether neural models would learn to score true (*context, response*) pairs more highly than false pairs, and second, whether their scores aligned with human judgements when given access to representations of the lexical content of contexts and the prosodic realisation of responses. While we find that models lag significantly behind people in this regard, our results across the training setups and our additional analyses provide a number of suggestions for why this might be the case; we discuss them and potential future directions in turn below.

Amount of data. This is the most obvious improvement. Using responses with the same lexical content provided important control for our preliminary studies but strongly constrained the amount of available training data. The results have provided weak evidence that the amount of data has limited the performance observed in these training setups; however, it has also placed limitations on the input representations that could be used. Given that the lexical identity cue did not seem to have much of an effect on our results, one direction for future work could be to test much larger training

setups using less stringent constraints on the lexical content of responses to better understand the data requirements for learning the complex relationship between context and response.

Lexical representations. We used representations of lexical content that we have previously found to be useful for predicting the upcoming lexical content of turns in dialogue (Chapter 4). Though the frozen SBERT model could obtain moderate accuracy, our results in Chapter 4 demonstrated that fine-tuning with the response selection task had a large effect on the downstream accuracy. It is possible that the information required to predict upcoming lexical content differs from what is useful for predicting upcoming prosodic realisations, however, it seems more likely that this information is too difficult to distil from frozen SBERT representations.

Prosodic representations. We selected the MPM and CWT prosodic representations used in our model inputs based on our findings that these representations contain information that is useful for predicting perceptually salient features independently of lexical content (Chapter 7). However, it is possible these prosodic representations were not sufficiently informative for the turn-selection task. One potential issue is that the MPM features were produced from a model pre-trained on a corpus of read speech. As discussed in Section 2.1, the acoustic properties of read and conversational speech differ and listeners can distinguish them. Tran (2020) shows that task-specific prosodic representations learned from read speech don't transfer robustly to spontaneous speech (while vice versa works much better). Additionally, the MPM features were trained on 6 second spans of speech. We therefore can't expect the model to learn about potential dependencies across utterances or between speaker turns (Cole, 2015). Developing a model that could learn these structures would require considering how to expose the model to these structures with a limited window of context. This may require compressing the speech signal to enable the encoding of longer sequences; this strategy has proved valuable in modelling the full speech signal (Zeghidour et al., 2021). Additionally, neither of the prosodic representations tested here contain information regarding the pause between speaker turns which is known to be an important cue to turn-taking (Sacks et al., 1978). Future experiments could consider including this information explicitly as an additional feature or implicitly by integrating it into the prosodic representation more carefully. Another feature to consider is the relatively coarse pooling strategy used to aggregate prosodic features. Mean-pooling

at the frame-level is a common strategy for producing utterance-level representations from SSL speech features; however, there are already investigations into improving the lossy nature of mean-pooling (Lin et al., 2023; Kakouros et al., 2023). Many of the acoustic-lexical tasks discussed in Section 8.2 involve encoding prosodic features at the word-level and learning them specifically for individual downstream tasks. Though the responses in our experiments only consisted of a couple of words, this will become more important when considering longer, more varied responses.

Reflecting human-like optionality. Along with examining the effect of training conditions on how accurately a model could learn to discriminate true and false continuations, we also considered whether our models scored responses similarly to humans. None of the tested variants seem to align with human judgements. However, the acoustic analysis presented in Section 8.7 revealed that responses with similar acceptability scores in the Is’21 evaluation set were not necessarily acoustically similar, even in the MPM and CWT acoustic feature spaces. This serves as a potential explanation for why we did not find stronger discriminative performance from neural models trained to distinguish true and negative (c, r) pairs and motivates potential future work. Understanding why this is the case could provide insights into both how prosodic features should be represented in wider dialogue contexts. On the representation side, we can make two hypotheses: perhaps the information required to distinguish true and false continuations in dialogue is not present in the acoustic representations tested here; perhaps it is present but the relationship between the context and the acoustic representation is complex and requires more training examples. Testing these hypotheses could shed light on the PIV of prosodic information for upcoming dialogue turns—multiple varied prosodic realisations can be perceived as plausible in a given context but we don’t yet understand what characteristics of an utterance lead to a particular score.

Isolated channels. Keeping the lexical and non-lexical channels isolated was an important component of our experimental design as it paves the way to different methods of quantifying the relative contributions of these channels. For example, if future iterations of this experiment achieve more reliable results, the framework could be used to quantify the amount of redundant information between channels. One method for this could be to inject noise into the representations of individual channels. For example, Nielsen (2024) demonstrates redundancy between prosodic and lexical information for

sentence parsing by adding varying amounts of noise to the target speech before extracting lexical and prosodic features. Another method is to measure the relative resistance to compression³ between channels through quantization techniques. The ratio of compression-resistance between channels is likely to vary dramatically across downstream tasks—consider the value of non-lexical information for prominence prediction vs keyword spotting. An alternative research direction is to test representations from pre-trained SSL models on the task of turn discrimination. This is particularly relevant in the current setup where the lexical content of responses is fixed and therefore not discriminative (though it is likely to be a useful cue to integrate).

³The idea of compression resistance is partially inspired by Covell et al. (1998) who perform non-uniform time compression of speech based on the degree to which segments of speech resist a change in speaking rate, a feature which they term “audio-tension”.

Chapter 9

Conclusions

In this thesis, we have explored how information in the lexical and non-lexical channels of spoken dialogue drives expectations regarding the realisation and content of potential upcoming dialogue turns. Compared to the classical notion of surprisal that quantifies how expected an upcoming linguistic unit may be in a particular context (Shannon, 1948; Hale, 2001; Levy, 2008a), we've instead examined how context shapes expectations themselves by defining the *Perceptual Information Value* (PIV) of a linguistic signal as the degree to which it constrains expectations. This alternative perspective provides a deeper understanding of the shape expectations involved in the comprehension of realistic language. We have used PIV to explore the relative contributions of the lexical and non-lexical components of speech (*which* words are said and *how* those words are spoken, respectively) on expectations.

Dialogue is a much more complex form of language-use than the sets of isolated sentences that have served as the foundations for many studies of language processing (Hale, 2001; Levy, 2011; Demberg and Keller, 2008; Kehler and Rohde, 2013). As such, we begin in Part I by introducing the task of dialogue turn acceptability for examining predictions of *lexical* content in dialogue. In this paradigm, participants must reason about more abstract pragmatic factors than previously-used acceptability tasks that examine isolated sentences (Coleman, 1965; Parker, 2019). Participant accuracy in our task clearly demonstrates that people are capable of making predictions at this level of abstraction and that lexical context constrains their expectations; however, participants often rate multiple upcoming turns as acceptable. We term the stochastic nature of their predictions as *optionality* and, throughout the remainder of the thesis, argue that it is an important component of language processing to account for in behavioural studies and models of language. Additionally, we find that under certain

architectural and training setups, modern neural models of dialogue align to a degree with human judgements. However, qualitative analysis of the optionality inherent in human and model scores points towards fundamental differences between predictions made by people and modern language models.

In Part II, we presented an extension of the turn acceptability behavioural paradigm to compare the PIV of lexical and non-lexical information. We investigate how expectations change when participants have access to either lexical information or a combination of lexical and non-lexical information. Our results have shown that people use both channels of information to develop expectations about the upcoming lexical content as well as its non-lexical realisation. The channels affect peoples' predictions both in isolation and in conjunction with each other. Importantly, our results show that access to different channels can produce consistent effects on expectations across participants even if it yields less accurate predictions. As such, we argue that Perceptual Information Value should be quantified to reflect the capacity to alter expectations rather than simply altering their accuracy.

Based on our findings about the predictions people can make about upcoming dialogue turns, Part III consists of two studies that apply self-supervised learning mechanisms as tools to investigate the predictable structures that enable human prediction. In Chapter 7, we find evidence of predictable structure in the acoustic correlates of prosody isolated from lexical content across multiple timescales. By varying the length of masking used to train these SSL models, we show that the value of the resulting representations for predicting different perceptual annotations of speech is dependent on the temporal resolution of masking. Through these experiments, we validate a method for producing generic representations of prosody that are independent of lexical content. Finally, Chapter 8 presents an exploratory study into whether predictive neural models can learn to perform the turn discrimination task using representations of the lexical and prosodic content of dialogue turns. Though we do not find evidence that models can learn to perform this task in our preliminary experimental setups, we believe this could be a valuable task for further probing both how people generate expectations in more complex forms of language-use like dialogue as well as whether machine representations of spoken language encode this information.

9.1 Directions for future work

An obvious limitation of the current studies was that the data used for the majority of our behavioural studies came from one corpus. Switchboard was a very suitable choice for our purposes in many ways—using telephone speech ensures that the information conveyed between conversational partners is transmitted across the lexical and non-lexical channels as opposed to recordings of face-to-face conversations where the visual modality also plays a role, and conversations between pairs of strangers means that there is less chance of pre-established common ground between conversational partners. However, the relatively small size of this corpus (at least by modern standards) limited both the stimuli we could present to participants in Chapter 5 as well as the amount of data available for training models in Chapter 8. Applying the turn discrimination paradigm using different datasets could relieve some of these constraints while also providing insights into how spoken communication differs across domains and languages. For example, the purpose of prosody is suggested to differ in adult- and child-directed speech (Baron, 1990; Räsänen et al., 2018), and speakers place prominence differently when producing read and spontaneous speech (Yuan et al., 2005). Similarly, the relationship between prosodic cues and communicative functions varies across languages (Ladd and Arvaniti, 2023). For example, Calhoun et al. (2021) show that speakers of Samoan and English weight syntactic and prosodic cues differently when identifying sentence prominence. As such, it would be interesting to compare the results presented for the Switchboard dataset to speech in different styles and languages.

Another important consideration that we did not touch on directly in this thesis but that merits further investigation is the question of individual differences across participants. In Chapter 8, we proposed quantifying PIV as entropy reduction across scoring. Our metrics demonstrated that, even when they became less accurate in terms of discriminating the true upcoming turn, participants tended to agree more when presented with non-lexical context. However, there is ample evidence that people vary in how they weigh different features of speech during comprehension Idemaru et al. (2012); Tremblay et al. (2018). This is part of what makes prosodic annotation so difficult (Roy et al., 2017). We took care to recruit participants that had relatively similar linguistic “backgrounds” (i.e., they shared the same first language and general location); however, future work could examine how performance varied across participants in more detail.

Though this thesis has presented a preliminary definition for PIV, there are many directions to consider for future work to apply, develop, and formalise this notion further. We believe there is potential to combine the turn discrimination paradigm with modern speech generation models, both for investigating human perception of more naturalistic language-use like dialogue and for evaluating the output generated by such models. We chose to implement the first PIV experiments presented in this thesis using natural (i.e., human-generated) data, however generative models also offer the possibility to produce stimuli that are difficult or impossible to extract from people. For example, our PIV experiments presented in Part II demonstrate that people generate expectations at the level of dialogue turns, but our posthoc acoustic analysis revealed that the relationship between context and response is complex—we didn’t find evidence that scoring behaviour was driven by acoustic similarity of responses (across macro acoustic features in Chapter 5 and features learned from pitch and intensity measures in Chapter 8). A deeper understanding of which aspects of a response are important for constraining scoring and conversely, which aspects can support greater optionality, would provide a new understanding of how people develop expectations. As they continue to improve, generative models could serve as a useful tool for generating different realisations with greater control (Hodari, 2022). They also offer methods to produce stimuli that control for confounding factors such as speaker identity; another avenue for future work that we considered was applying voice conversion as a means to present more non-lexical context to participants. In our current experiments, we were limited to only presenting the spoken realisation of the preceding contextual turn to avoid revealing the target speaker identity, however as voice conversion models continue to improve, synthesizing conversations to a fixed set of speakers would allow us to run experiments with broader spoken contexts that may relieve some of the cognitive load effects that we posited in Chapter 5 (Xie et al., 2022; Baas et al., 2023).

As we discussed in the conclusion of Chapter 5, the turn discrimination paradigm could serve as a valuable evaluation paradigm for the appropriateness of generated speech in a given context; however, our results have highlighted the importance of optionality in the people’s expectations regarding upcoming speech. As such, another interesting direction for future work is to compare the diversity of speech deemed appropriate by people, and the diversity of speech produced by generative models. Using distributions of generated output rather than an individual sample is already cropping up across a range of NLP applications. For example, in related work, we propose measuring the predictability of a particular continuation as a function of its

similarity to continuations generated by language models (Giulianelli et al., 2023). We find that predictability estimated in this way correlates more strongly with human language comprehension behaviours than classical surprisal computed from the same model. Guerreiro et al. (2023) finds that selecting a translation from a set of LM-produced generations decreases the rate of machine translation hallucination threefold. Comparing human judgements and model outputs in this way could illuminate which aspects of speech are easier for models to encode, but also serve as a mechanism to investigating *why* a potential upcoming linguistic unit is surprising or not. Analyses of when and where optionality manifests have been relatively constrained to language production tasks like re-enactment or imitation (Mixdorff et al., 2012; Avila and Ward, 2023)); rating the acceptability of utterances could provide a far less expensive route to these analyses.

Finally, we have dedicated a lot of space in this thesis to framing PIV in relation to surprisal. However, we have mostly refrained from discussing one of the core components of surprisal theory: the link to processing effort in language comprehension (Hale, 2001). Researchers have drawn a direct link between the two, where processing effort has been measured through behaviours including reading times, (Keller, 2004; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Meister et al., 2021; Shain et al., 2024; Oh and Schuler, 2023b), acceptability judgements, (Lawrence et al., 2000; Heilman et al., 2014; Lau et al., 2015, 2017; Warstadt et al., 2019), and brain activity (Frank et al., 2015; Schrimpf et al., 2021). Yet, as we have pointed out, the comprehension of realistic language forms like dialogue is likely to involve predictions with much greater variability than what can be observed from studying isolated, written stimuli. This raises an intriguing question: does the shape of expectations also affect processing effort? As Figure 1.2 in the very first chapter depicts, a unit may be equally surprising under different expectations; however, the shape of these different expectation distributions may also affect the processing effort required. A potential set of experiments would be to observe how the predictive power of surprisal for language comprehension behaviour changes if characterisations of expectations are considered. The conditional entropy studies discussed in Section 1.3 already provide evidence that this feature of expectations is predictive of some language comprehension behaviours; for example, Pimentel et al. (2023) show that the expected surprisal of an upcoming word is predictive of its reading time. However, these works have been limited to reading written sentences or passages, and have only considered one operationalization of “shape”. Integrating PIV into models of human comprehension effort is an entire re-

search agenda in and of itself, but one that we believe will be fruitful as we continue to acquire tools to progress the study of human language processing beyond written sentences to more realistic forms of language.

Part IV

Appendices

Appendix A

Machine expectations in written dialogue

Table A.1, Table A.2, Table A.3, Table A.4 contain examples of the stimuli presented in Chapter 4 along with associated scores from both human participants and the BERT-FP language model. The contexts that make up our stimuli consist of a variable number of conversational turns (1-4); each is paired with ten potential upcoming responses, one of which is the true response in the particular conversation. Participants were presented with a single (context, response) pair and asked to rate how plausible the response was given the context.

Tables Table A.1, Table A.2, Table A.3 show human misrankings while Table Table A.4 contains the stimuli misranked by BERT-FP.

Table A.1: Example of response selections stimuli (*sw3512_14*) with human and BERT-FP scores.

Table (A) contains the set of conversational turns provided as context. Table (B) contains the set of 10 responses—these include the true upcoming turn for the given context and the 9 negative samples. Each response is displayed with its true response label, the median plausibility score received from participants and its BERT-FP response score; the highest scores are displayed in bold. Responses are sorted by the median plausibility score from participants.

(A) *Context* Speaker A : that’s right
 Speaker B : and then it’s just so noisy that you can’t visit
 Speaker A : oh i know
 Speaker B : know and normally when i’m eating out i you know
 with people and i wanna sit and talk i mean

(B) Responses	Human	BERT-FP	Transcript
Negative	4.50	0.57	mhm
True	4.40	0.99	that’s the half the fun is the conversation right
Negative	3.83	0.60	dear
Negative	3.40	0.15	true you know and you have to start think- ing about is it is it worth spending the money to go see it or shall i just wait
Negative	2.60	0.12	hello
Negative	2.00	0.17	but a lot of the stuff they do really you know evidently is pretty easy but i’ve just never
Negative	2.00	0.00	was lonely and she needed company for her mother and so she opened a nursing home and initially started with eight ladies
Negative	1.80	0.02	it’s uh it’s about an eight or nine hour drive really i make it in two days because i i don’t push it
Negative	1.75	0.03	yeah when he played Danny Boy it just al- most brought tears to your eyes because he can make that flute sing
Negative	1.40	0.01	yeah it’s raining out here and i just steam cleaned my carpet today and i really don’t wanna let the dog in

Table A.2: Example of response selections stimuli (*sw4040_20*) with human and BERT-FP scores.

(A) *Context* Speaker A : say you went back to school
 Speaker B : the wife went back to school
 Speaker A : i see uhuh
 Speaker B : he's been there for almost a year now

(B) Responses	Human	BERT-FP	Transcript
Negative	4.83	0.51	yeah
True	4.50	0.94	um gee i think where where does she go to school
Negative	3.00	0.06	what kind of things have you read
Negative	2.60	0.40	but that that's if
Negative	2.25	0.04	um she was referred to me by a couple of people and she turned out to be wonderful i couldn't have asked for anything better i don't think
Negative	2.20	0.01	it's so much easier to sit there and besides i can be doing other things and still listening to the news
Negative	2.00	0.01	you you say show music like Broadway musical type show music
Negative	2.00	0.02	and you they find everything i mean they find out everything about you they want to know your you know where you live what you do what you know and some of the questions
Negative	1.50	0.07	the starter was Bosch American so
Negative	1.17	0.00	that's there for direct yeah the direct sun beating on it yeah that's right

Table A.3: Example of response selections stimuli (*sw3959_32*) with human and BERT-FP scores.

(A) *Context* Speaker A : uhuh oh my
 Speaker B : down in Houston for um several years seven years
 and then uh my son is a CPA another has a business
 degree

(B) <i>Responses</i>	Human	BERT-FP	Transcript
Negative	4.80	0.63	before that did he go somewhere else or
True	3.60	0.93	my
Negative	2.33	0.00	that uh lots lots of new uh newsmen were created during that so it will be interesting to see what happens
Negative	2.00	0.00	you know with um the US funding Israel and you had the um Soviet Union funding the Arab countries
Negative	1.67	0.58	um yeah
Negative	1.67	0.02	uh what kind of lawn and garden work do you wind up doing
Negative	1.50	0.08	something that we're looking forward to
Negative	1.25	0.00	painted on T-shirts or sweat shirts at all
Negative	1.00	0.01	i do know of a way to get around the com- puter generated calls
Negative	1.00	0.00	we try to stay away from those things which might have uh salmonella in them

Table A.4: Example of response selections stimuli (*sw2303_59*) with human and BERT-FP scores.

(A) *Context* Speaker A : so they're going to be
 Speaker B : that's another and that's another interesting question
 should judges be elected or appointed

(B) <i>Responses</i>	Human	BERT-FP	Transcript
True	4.60	0.18	that's true that's true well
Negative	4.00	0.50	hm
Negative	2.80	0.90	idea
Negative	2.60	0.01	anyway what do you think we've gained from the space flights
Negative	2.40	0.49	makes you wish they had uh still had inden- tured servitude for this sort of thing
Negative	1.25	0.12	mean like those you know twenty thousand dollar toilets
Negative	1.17	0.14	well the the you know those little arms are supposed to twist almost any
Negative	1.00	0.00	yeah it's nothing but woods up here
Negative	1.00	0.00	it's uh it's about an eight or nine hour drive really i make it in two days because i i don't push it
Negative	1.00	0.00	i i i used to exercise at night and i found that you don't come home tired you come home with a new found energy so you

Appendix B

Acoustic-conditioned turn discrimination

Context turns	Epoch	val $R_{k=1}$	test $R_{k=1}$	$IS21R_{k=1}$	$IS21R_{k=2}$	$IS21R_{k=3}$
1	6	0.290	0.235	0.220	0.460	0.720
1	16	0.283	0.223	0.280	0.440	0.760
4	15	0.283	0.238	0.220	0.440	0.700
7	10	0.288	0.238	0.220	0.420	0.720
7	27	0.283	0.233	0.240	0.400	0.660

Table B.1: TOP-4 training set with dataset size 1000 (S) and no lexical identity feature. Table 8.2 indicates that the Small dataset size may not be sufficient for generalisable performance; we include it here for completeness.

Context turns	Epoch	val $R_{k=1}$	test $R_{k=1}$	$IS21R_{k=1}$	$IS21R_{k=2}$	$IS21R_{k=3}$
1	59	0.305	0.280	0.300	0.520	0.700
4	46	0.295	0.258	0.180	0.560	0.780
7	37	0.278	0.263	0.220	0.440	0.740

Table B.2: TOP-4 training set with the L dataset size (8000) and no lexical identity. Table 8.2 suggested that both M and L dataset sizes achieved similar generalisation as quantified by the performance difference between validation and test sets. This table confirms this result. Similar to Table 8.3a, we find the best held out performance on for the smallest number of context turn, however unlike Table 8.3a, $R_{\{k=1\}}$ on Is’21 is also best with shorter context.

Bibliography

- Adigwe, A., Wallbridge, S., and King, S. (2024). What makes conversational speech: Investigating acoustic-prosodic cues in speech perception. In *Proceedings of Interspeech*.
- Alain, G. and Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. *Proceedings of the International Conference on Learning Representations*.
- Ankner, Z., Saphra, N., Blalock, D., Frankle, J., and Leavitt, M. L. (2023). Dynamic Masking Rate Schedules for MLM Pretraining. *ArXiv (arXiv:2305.15096)*.
- Anthropic (2023). Model card and evaluations for claude models. Technical report, Anthropic.
- Asano, Y. M., Rupprecht, C., and Vedaldi, A. (2019). A critical analysis of self-supervision, or what we can learn from a single image. In *Proceedings of the International Conference on Learning Representations*.
- Aurnhammer, C. and Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.
- Avila, J. E. and Ward, N. G. (2023). Towards Cross-Language Prosody Transfer for Dialog. In *Proceedings of Interspeech*, pages 2143–2147.
- Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.
- Baas, M., Nieker, B. V., and Kamper, H. (2023). Voice conversion with just nearest neighbors. In *Proceedings of Interspeech*.

- Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.
- Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.
- Baron, N. S. (1990). *Pigeon-Birds and Rhyming Words: The Role of Parents in Language Learning. Language in Education: Theory and Practice*. ERIC.
- Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.
- Batliner, A., Seppi, D., Steidl, S., and Schuller, B. (2010). Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human-Computer Interaction*, 2010(1):782802.
- Becker, S. and Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163.
- Beckman, M. (1986). *Stress And Non-Stress Accent*. De Gruyter.
- Beckman, M. E. and Pierrehumbert, J. B. (1986). Intonational structure in japanese and english. *Phonology*, 3:255–309.
- Bellier, L., Llorens, A., Marciano, D., Gunduz, A., Schalk, G., Brunner, P., and Knight, R. T. (2023). Music can be reconstructed from human auditory cortex activity using nonlinear decoding models. *PLOS Biology*.
- Ben-David, B. M., Multani, N., Shakuf, V., Rudzicz, F., and van Lieshout, P. H. (2016). Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research*, 59(1):72–89.
- Bender, E. M. and Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

- Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.
- Benus, S., Gravano, A., and Hirschberg, J. B. (2007). The prosody of backchannels in American English. In *Proceedings of the International Congress of Phonetic Sciences*.
- Bergen, L., Levy, R., and Gibson, E. (2012). Verb omission errors: Evidence of rational processing of noisy language inputs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.
- Bernardy, J.-P., Lappin, S., and Lau, J. H. (2018). The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Melbourne, Australia. Association for Computational Linguistics.
- Bhati, S., Villalba, J., Żelasko, P., Moro-Velazquez, L., and Dehak, N. (2022). Un-supervised speech segmentation and variable rate representation learning using segmental contrastive predictive coding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2002–2014.
- Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., and Turian, J. (2020). Experience grounds language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 8718–8735.
- Blaauw, E. (1994). The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech communication*, 14(4):359–375.
- Black, A. W. (1997). Predicting the intonation of discourse segments from examples in dialogue speech. In *Computing Prosody*.
- Boersma, P. and Van Heuven, V. (2001). Speak and unspeak with praat. *Glott International*, 5(9/10):341–347.
- Bögel, T. and Turk, A. (2019). Frequency effects and prosodic boundary strength. In *Proceedings of the International Congress of Phonetic Sciences*, pages 1014–1018.

- Bögels, S., Kendrick, K. H., and Levinson, S. C. (2015). Never say no...How the brain interprets the pregnant pause in conversation. *PLoS ONE*, 10(12).
- Bögels, S. and Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.
- Bögels, S. and Torreira, F. (2021). Turn-end estimation in conversational turn-taking: The roles of context and prosody. *Discourse Processes*, 58:903 – 924.
- Bolinger, D. (1964). Around the edge of language: Intonation. *Harvard educational review*, 34(2):282–296.
- Bolinger, D. (1978). Intonation across languages. *Universals of human language*, 2(471-524).
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *Proceedings of Advances in Neural Information Processing Systems*, 6:737–744.
- Brown, A. C., Childers, E., Bowen, E. F., Zuckerberg, G. A., and Granger, R. (2022). Phonemes in continuous speech are better recognized in context than in isolation. *Frontiers in Communication*, 7:865587.
- Brown, M., Salverda, A. P., Dilley, L. C., and Tanenhaus, M. K. (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of Experimental Psychology: Human Perception and Performance*, 41(2):306.
- Buxó-Lugo, A. and Watson, D. G. (2016). Evidence for the influence of syntax on prosodic parsing. *Journal of memory and language*, 90:1–13.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.
- Bürkner, P.-C. and Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.
- Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. I. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*.

- Calhoun, S., Wollum, E., and Kruse Va'ai, E. (2021). Prosodic prominence and focus: Expectation affects interpretation in samoan and english. *Language and Speech*, 64(2):346–380.
- Camacho, A. and Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3):1638–1652.
- Cao, H., Štefan Beňuš, Gur, R. C., Verma, R., and Nenkova, A. (2014). Prosodic cues for emotion: analysis with discrete characterization of intonation. In *Proceedings of Speech Prosody*, pages 130–134.
- Carden, G. (1976). Syntactic and semantic data: replication results. *Language in Society*, 5(1):99–104.
- Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.
- Cervone, A., Lai, C., Pareti, S., and Bell, P. (2015). Towards automatic detection of reported speech in dialogue using prosodic cues. In *Proceedings of Interspeech*, volume 15, pages 3061–3065.
- Cervone, A. and Riccardi, G. (2020). Is this dialogue coherent? learning from dialogue acts and entities. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 162–174.
- Cervone, A., Stepanov, E. A., and Riccardi, G. (2018). Coherence models for dialogue. In *Proceedings of Interspeech*, pages 1011–1015.
- Chater, N., Tenenbaum, J. B., and Yuille, A. L. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10:287–291.
- Christiansen, M. H. and Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39 62:1–72.
- Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.
- Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In *Perspectives on socially shared cognition.*, pages 127–149. American Psychological Association.

- Clark, H. H. and Tree, J. E. F. (2002). Using “uh” and “um” in spontaneous speaking. *Cognition*, 84:73–111.
- Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1):1–39.
- Clark, R., Silen, H., Kenter, T., and Leith, R. (2019). Evaluating Long-form Text-to-Speech: Comparing the Ratings of Sentences and Paragraphs. In *Proceedings of the Workshop on Speech Synthesis*, pages 99–104.
- Cohen, P. (1984). The pragmatics of referring and the modality of communication. *Computational Linguistics*, 10(2):97–146.
- Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1-2):1–31.
- Cole, J., Mahrt, T., and Roy, J. (2017). Crowd-sourcing prosodic annotation. *Computer Speech & Language*, 45:300–325.
- Cole, J., Mo, Y., and Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2):425–452.
- Cole, R. A. and Jakimik, J. (1980). How are syllables used to recognize words? *The Journal of the Acoustical Society of America*, 67(3):965–970.
- Coleman, E. (1965). Responses to a scale of grammaticalness. *Journal of verbal learning and verbal behavior*, 4(6):521–527.
- Collier, R. and Hart, J. (1975). The role of intonation in speech perception. In *Structure and Process in Speech Perception: Proceedings of the Symposium on Dynamic Aspects of Speech Perception held at IPO*, pages 107–123. Springer.
- Corps, R. E., Pickering, M. J., and Gambi, C. (2018). Predicting turn-ends in discourse context. *Language, Cognition and Neuroscience*, 34:615 – 627.
- Covell, M., Withgott, M., and Slaney, M. (1998). Mach1: Nonuniform time-scale modification of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 349–352. IEEE.

- Cuesta-Frau, D., Molina-Picó, A., Vargas, B., and González, P. (2019). Permutation entropy: Enhancing discriminating power by using relative frequencies vector of ordinal patterns instead of their shannon entropy. *Entropy*, 21(10):1013.
- Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics*, 20:55–60.
- Cutler, A. (1987). Components of prosodic effects in speech recognition. In *Proceedings of the International Congress of Phonetic Sciences*, pages 84–87.
- Cutler, A., Dahan, D., and Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech*, 40(2):141–201.
- Cutler, A. and Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and speech*, 20(1):1–10.
- Cutler, A. and McQueen, J. M. (2014). How prosody is both mandatory and optional. In Caspers, J., Chen, Y., Heeren, W., Pacilly, J., Schiller, N. O., and Zanten, E. V., editors, *Above and Beyond the Segments: Experimental linguistics and phonetics*. Benjamins.
- Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, 7(5):889–904.
- de Ruiter, J. P., Mitterer, H., and Enfield, N. J. (2006). Projecting the end of a speaker’s turn: A cognitive cornerstone of conversation. *Language*, 82:515 – 535.
- de Seyssel, M., D’Avirro, A., Williams, A., and Dupoux, E. (2023). Emphasess: a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. *arXiv preprint arXiv:2312.14069*.
- de Seyssel, M., Lavechin, M., Titeux, H., Thomas, A., Virlet, G., Revilla, A. S., Wisniewski, G., Ludusan, B., and Dupoux, E. (2023). ProsAudit, a prosodic benchmark for self-supervised speech models. In *Proceedings of Interspeech*, pages 2963–2967.
- Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.
- Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., and Picone, J. (1998). Resegmentation of switchboard. In *Proceedings of the International Conference on Spoken Language Processing*. Sydney.

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.
- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proceedings of Machine Learning Research*.
- Doyle, G. and Frank, M. (2015). Shared common ground influences information density in microblog texts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1587–1596, Denver, Colorado. Association for Computational Linguistics.
- Drugman, T., Huybrechts, G., Klimkov, V., and Moinet, A. (2018). Traditional machine learning for pitch detection. *IEEE Signal Processing Letters*, 25(11):1745–1749.
- Dubey, A., Keller, F., and Strut, P. (2013). Probabilistic modeling of discourse-aware sentence processing: Computational models of natural language. *Topics in cognitive science (Print)*, 5(3):425–451.
- Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283.
- Eckert, H. and Laver, J. (1994). Menschen und ihre stimmen. aspekte der vokalen kommunikation. beltz psychologie-verlags-union [people and their voices: aspects of vocal communication].
- Eckert, P. (1989). *Jocks and burnouts: Social categories and identity in the high school*. Teachers college press.
- Ekstedt, E. and Skantze, G. (2020). TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics, Empirical Methods in Natural Language Processing*, pages 2981–2990. Association for Computational Linguistics.

- Ekstedt, E. and Skantze, G. (2022). How much does prosody help turn-taking? investigations using voice activity projection models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551.
- Esposito, A., Stejskal, V., and Smékal, Z. (2008). Cognitive role of speech pauses and algorithmic considerations for their processing. *International Journal of Pattern Recognition and Artificial Intelligence*, 22(05):1073–1088.
- Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of Empirical Methods in Natural Language Processing*.
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.
- Fenk, A. and Fenk, G. (1980). Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß? *Zeitschrift für Experimentelle und Angewandte Psychologie*, 27(3):400–414.
- Fernandez, R. and Ginzburg, J. (2002). Non-sentential utterances in dialogue: A: Corpus-based study. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 15–26, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Ferreira, F., Bailey, K. G., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1):11–15.
- Ferreira, F. and Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current directions in psychological science*, 27(6):443–448.
- Ferreira, F. and Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30(6):725–745.
- Ford, C. E. and Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184.

- Fouda, J. A. E., Koepf, W., and Jacquir, S. (2017). The ordinal Kolmogorov-Sinai entropy: A generalized approximation. *Communications in Nonlinear Science and Numerical Simulation*, 46:103–115.
- Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.
- Fraser, B. (1999). What are discourse markers? *Journal of pragmatics*, 31(7):931–952.
- Fraundorf, S. H. and Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of memory and language*, 65 2:161–175.
- Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352.
- Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of physiology-Paris*, 100(1-3):70–87.
- Fröhlich, M., Sievers, C., Townsend, S. W., Gruber, T., and van Schaik, C. P. (2019). Multimodal communication and language origins: integrating gestures and vocalizations. *Biological Reviews*, 94(5):1809–1829.
- Fuscone, S., Favre, B., and Prévot, L. (2020). Neural representations of dialogical history for improving upcoming turn acoustic parameters prediction. In *Proceedings of Interspeech*, volume 2020-Octob, pages 4203–4207.
- Futrell, R. and Gibson, E. (2017). L2 processing as noisy channel language comprehension. *Bilingualism: Language and Cognition*, 20(4):683–684.
- Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.
- Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of experimental psychology: Human perception and performance*, 6(1):110.
- Garofolo, J. S. (1993). TIMIT acoustic phonetic continuous speech corpus. *Linguistic Data Consortium*.

- Garrod, S. and Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in psychology*, 6(751):1–12.
- Genzel, D. and Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 199–206.
- Ginzburg, J. (2001). Fragmenting meaning: Clarification ellipsis and nominal anaphora. In *Computing Meaning: Volume 2*, pages 247–270. Springer.
- Giulianelli, M., Sinclair, A., and Fernández, R. (2021). Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Giulianelli, M., Wallbridge, S., and Fernández, R. (2023). Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653.
- Gleitman, L. and Wanner, E. (1982). *Language Acquisition: The State of the Art*. Cambridge University Press, Cambridge UK.
- Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing-Volume 1*, pages 517–520.
- Golan, T., Siegelman, M., Kriegeskorte, N., and Baldassano, C. (2023). Testing the limits of natural language models for predicting human language judgements. *Nature Machine Intelligence*, 5(9):952–964.
- Goodkind, A. and Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.
- Gorisch, J., Wells, B., and Brown, G. J. (2012). Pitch Contour Matching and Interactional Alignment across Turns: An Acoustic Investigation. *Language and Speech*, 55(1):57–76.

- Grandjean, D., Bänziger, T., and Scherer, K. R. (2006). Intonation as an interface between language and affect. *Progress in brain research*, 156:235–247.
- Gravano, A. and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.
- Grice, M., Reyelt, M., Benzmueller, R., Mayer, J., and Batliner, A. (1996). Consistency in transcription and labelling of german intonation with gtobi. In *Proceedings of the International Conference on Spoken Language Processing*, volume 3, pages 1716–1719. IEEE.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364.
- Grosjean, F. (1983). How long is the sentence? Prediction and prosody in the online processing of language. In *Linguistics*.
- Gu, J., Li, T., Liu, Q., Ling, Z., Su, Z., Wei, S., and Zhu, X. (2020). Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In d’Aquin, M., Dietze, S., Hauff, C., Curry, E., and Cudré-Mauroux, P., editors, *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 2041–2044. ACM.
- Guerreiro, N. M., Voita, E., and Martins, A. (2023). Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.
- Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech*.
- Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017). Implicit regularization in matrix factorization. *Proceedings of the Annual Conference on Neural Information Processing Systems*.

- Gutierrez, E., Oplustil-Gallegos, P., and Lai, C. (2021). Location, Location: Enhancing the Evaluation of Text-to-Speech synthesis using the Rapid Prosody Transcription Paradigm. In *Proceedings of the Speech Synthesis Workshop*, pages 25–30.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics*.
- Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive science*, 30(4):643–672.
- Hamaker, J., Zeng, Y., and Picone, J. (1998). Rules and guidelines for transcription and segmentation of the switchboard large vocabulary conversational speech recognition corpus. Technical report, Mississippi State University.
- Han, J., Hong, T., Kim, B., Ko, Y., and Seo, J. (2021). Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558. Association for Computational Linguistics.
- Hebets, E. A. and Papaj, D. R. (2004). Complex signal function: developing a framework of testable hypotheses. *Behavioral Ecology and Sociobiology*, 57:197–214.
- Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., and Tetreault, J. (2014). Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, Baltimore, Maryland. Association for Computational Linguistics.
- Henderson, M., Casanueva, I., Mrkšić, N., Su, P.-H., Wen, T.-H., and Vulić, I. (2020). ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics (Empirical Methods in Natural Language Processing)*, pages 2161–2174. Association for Computational Linguistics.
- Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of*

- the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4129–4138. Association for Computational Linguistics.
- Hill, A. A. (1961). Grammaticality. *Word*, 17(1):1–10.
- Hinton, G. E. and Zemel, R. (1993). Autoencoders, minimum description length and helmholtz free energy. *Proceedings of Advances in Neural Information Processing Systems*, 6.
- Hirschberg, J. and Nakatani, C. H. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 286–293.
- Hirschberg, J. and Pierrehumbert, J. (1986). The intonational structuring of discourse. In *24th Annual Meeting of the Association for Computational Linguistics*, pages 136–144.
- Hodari, Z. (2022). *Synthesising prosody with insufficient context*. PhD thesis, University of Edinburgh.
- Hodari, Z., Lai, C., and King, S. (2020). Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of F0. In *Proceedings of Speech Prosody*, pages 965–969.
- Holler, J. and Kendrick, K. H. (2015). Unaddressed participants’ gaze in multi-person interaction: optimizing reciprocity. *Frontiers in psychology*, 6(98):1–14.
- Holler, J., Kendrick, K. H., and Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychonomic bulletin & review*, 25(5):1900–1908.
- Howell, P. and Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. *Speech communication*, 10(2):163–169.
- Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

- Hsu, W.-N., Zhang, Y., and Glass, J. (2017). Unsupervised learning of disentangled and interpretable representations from sequential data. *Proceedings of Advances in Neural Information Processing Systems*, 30.
- Huang, R., Zhang, C., Ren, Y., Zhao, Z., and Yu, D. (2023). Prosody-TTS: Improving prosody with masked autoencoder and conditional diffusion model for expressive text-to-speech. In *Findings of the Association for Computational Linguistics*.
- Huang, Y. and Ferreira, F. (2020). The application of signal detection theory to acceptability judgments. *Frontiers in Psychology*, 11:73.
- Humeau, S., Shuster, K., Lachaux, M., and Weston, J. (2020). Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.
- Idemaru, K., Holt, L. L., and Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of japanese stop lengths. *The Journal of the Acoustical Society of America*, 132(6):3950–3964.
- Ioannides, G., Owen, M., Fletcher, A., Rozgic, V., and Wang, C. (2023). Towards paralinguistic-only speech representations for end-to-end speech emotion recognition. In *Proceedings of Interspeech*.
- Ip, M. and Cutler, A. (2017). Intonation facilitates prediction of focus even in the presence of lexical tones. In *Proceedings of Interspeech*.
- Jaekle, J. (2013). *Film dialogue*. Columbia University Press.
- Jaeger, T. and Levy, R. (2006). Speakers optimize information density through syntactic reduction. *Proceedings of Advances in Neural Information Processing Systems*, 19.
- Jaeger, T. F. and Weatherholtz, K. (2016). What the heck is salience? how predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology*, 7:1115.
- Jamshid Lou, P., Wang, Y., and Johnson, M. (2019). Neural constituency parsing of speech transcripts. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2756–2765. Association for Computational Linguistics.

- Jasmin, K., Tierney, A., and Chisom Obasih, L. H. (2021). Short-term perceptual re-weighting in suprasegmental categorization. Technical report, Birkbeck, University of London.
- Johnson, K. (2004). Massive reduction in conversational american english. In *Spontaneous speech: Data and analysis. Proceedings of the 10th international symposium*, pages 29–54. Citeseer.
- Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Span-BERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*.
- Kafle, S., Alm, C. O., and Huenerfauth, M. (2019). Fusion Strategy for Prosodic and Lexical Representations of Word Importance. In *Proceedings of Interspeech*, pages 1313–1317.
- Kakouros, S. and Räsänen, O. (2015). Automatic detection of sentence prominence in speech using predictability of word-level acoustic features. In *Proceedings of Interspeech*.
- Kakouros, S., Salminen, N., and Räsänen, O. (2018). Making predictable unpredictable with style-behavioral and electrophysiological evidence for the critical role of prosodic expectations in the perception of prominence in speech. *Neuropsychologia*, 109:181–199.
- Kakouros, S., Stafylakis, T., Mošner, L., and Burget, L. (2023). Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.
- Kalinli, O. and Narayanan, S. S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In *Proceedings of Interspeech*, pages 1941–1944.
- Kasi, K. (2002). *Yet another algorithm for pitch tracking: (YAAPT)*. PhD thesis, Old Dominion University.
- Kawahara, T., Yamaguchi, T., Inoue, K., Takanashi, K., and Ward, N. G. (2016). Prediction and generation of backchannel form for attentive listening systems. In

- Morgan, N., editor, *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, pages 2890–2894. ISCA.
- Kehler, A. and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2):1–37.
- Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD thesis, University of Edinburgh.
- Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 317–324.
- Kim, S. K. and Sumner, M. (2017). Beyond lexical meaning: The effect of emotional prosody on spoken word recognition. *The Journal of the Acoustical Society of America*, 142(1):EL49–EL55.
- Kimball, A. E. and Cole, J. (2016). Pitch contour shape matters in memory. In *Proceedings of the International Conference on Speech Prosody*, volume 8, pages 1171–75.
- Kintsch, W. and Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5):363.
- Klatt, D. H. (1987). Review of text-to-speech conversion for english. *The Journal of the Acoustical Society of America*, 82(3):737–793.
- Klein, I., Mangold, B., and Doll, M. (2016). Cumulative paired ϕ -entropy. *Entropy*, 18:248.
- Kleinhans, J., Farrús, M., Gravano, A., Pérez, J. M., Lai, C., and Wanner, L. (2017). Using prosody to classify discourse relations. In *Proceedings of Interspeech*.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69.
- Kohonen, T. (1988). The “neural” phonetic typewriter. *Computer*, 21(3):11–22.
- Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *Proceedings of the Annual Meeting of the American Institute of Chemical Engineers*, 37(2):233–243.

- Kreiman, J. (1982). Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, 10(2):163–175.
- Kumar, A., Narayanan Sundararaman, M., and Vepa, J. (2021). What BERT based language model learns in spoken transcripts: An empirical study. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 322–336. Association for Computational Linguistics.
- Ladd, D. R. (2000). Bruce, pierrehumbert, and the elements of intonational phonology. In *Prosody: Theory and experiment: Studies presented to Gösta Bruce*, pages 37–50. Springer.
- Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.
- Ladd, D. R. and Arvaniti, A. (2023). Prosodic prominence across languages. *Annual Review of Linguistics*, 9(1):171–193.
- Ladd, D. R. and Cutler, A. (1983). *Prosody: Models and Measurements*, chapter Models and Measurements in the Study of Prosody. Springer Berlin Heidelberg.
- Ladefoged, P., Johnson, K., and Ladefoged, P. (2006). *A course in phonetics*, volume 3. Thomson Wadsworth Boston.
- Lai, C. (2009). Perceiving surprise on cue words: Prosody and semantics interact on right and really. In *Tenth Annual Conference of the International Speech Communication Association*.
- Lai, C. (2010). What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue. In *Proceedings of Interspeech*, pages 1413–1416.
- Lai, C. (2012). Response types and the prosody of declaratives. In *Proceedings of Speech Prosody*.
- Lai, C., Farrús, M., and Moore, J. D. (2020). Integrating lexical and prosodic features for automatic paragraph segmentation. *Speech Communication*, 121:44–57.
- Lau, J. H., Clark, A., and Lappin, S. (2015). Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1618–1628.

- Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.
- Laver, J. (1981). Linguistic routines and politeness in greeting and parting. *Conversational routine*, 289304.
- Lawrence, S., Giles, C. L., and Fong, S. (2000). Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge & Data Engineering*, 12(01):126–140.
- Lenain, R., Weston, J., Shivkumar, A., and Fristed, E. (2020). Surfboard: Audio Feature Extraction for Modern Machine Learning. In *Proceedings of Interspeech*, pages 2917–2921.
- Lenth, R. V. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.4-1.
- Leong, C. S.-Y. and Linzen, T. (2024). Testing learning hypotheses using neural networks by manipulating learning data. *arXiv preprint arXiv:2407.04593*.
- Levine, Y., Lenz, B., Lieber, O., Abend, O., Leyton-Brown, K., Tennenholtz, M., and Shoham, Y. (2020). PMI-Masking: Principled masking of correlated spans. In *Proceedings of the International Conference on Learning Representations*.
- Levinson, S. C. (2016). Turn-taking in human communication—origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14.
- Levinson, S. C. and Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130302.
- Levinson, S. C. and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731.
- Levinson, S. E. (1985). Structural methods in automatic speech recognition. *Proceedings of the IEEE*, 73(11):1625–1650.
- Levitan, R., Gravano, A., and Hirschberg, J. (2011). Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 113–117.

- Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.
- Levy, R. (2008b). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Honolulu, Hawaii. Association for Computational Linguistics.
- Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065. Association for Computational Linguistics.
- Levy, R. and Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In Schölkopf, B., Platt, J. C., and Hofmann, T., editors, *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems*, pages 849–856. MIT Press.
- Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Proceedings of Advances in Neural Information Processing Systems*, 19:849.
- Lewis, D. K. (1969). *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell.
- Li, J. and Jurafsky, D. (2017). Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209. Association for Computational Linguistics.
- Li, Y., Mohamied, Y., Bell, P., and Lai, C. (2023). Exploration of a self-supervised speech model: A study on emotional corpora. In *IEEE Spoken Language Technology Workshop*, pages 868–875. IEEE.
- Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 986–995.
- Lin, G.-T., Feng, C.-L., Huang, W.-P., Tseng, Y., Lin, T.-H., Li, C.-A., Lee, H.-y., and Ward, N. G. (2023). On the utility of self-supervised models for prosody-related

- tasks. In *Proceedings of the IEEE Spoken Language Technology Workshop*, pages 1104–1111. IEEE.
- Lin, Y., Tan, Y. C., and Frank, R. (2019). Open sesame: Getting inside BERT’s linguistic knowledge. In Linzen, T., Chrupała, G., Belinkov, Y., and Hupkes, D., editors, *Proceedings of the BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.
- Linde, Y., Buzo, A., and Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Transactions on communications*, 28(1):84–95.
- Linzen, T. and Jaeger, T. F. (2014). Investigating the role of entropy in sentence processing. In *Proceedings of the fifth workshop on cognitive modeling and computational linguistics*, pages 10–18.
- Liu, Q., Chen, Y., Chen, B., Lou, J.-G., Chen, Z., Zhou, B., and Zhang, D. (2020a). You impress me: Dialogue generation via mutual persona perception. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427. Association for Computational Linguistics.
- Liu, Q., Kusner, M. J., and Blunsom, P. (2020b). A survey on contextual embeddings. *ArXiv*, abs/2003.07278.
- Livingstone, S. R. and Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*.
- Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning*, pages 7247–7283.
- Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294. Association for Computational Linguistics.

- Lubold, N., Pon-Barry, H., and Walker, E. (2015). Naturalness and rapport in a pitch adaptive learning companion. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 103–110. IEEE.
- Ludusan, B., Cristia, A., Mazuka, R., and Dupoux, E. (2022). How much does prosody help word segmentation? A simulation study on infant-directed speech. *Cognition*.
- Ludusan, B. and Dupoux, E. (2014). Towards low-resource prosodic boundary detection. In *Proceedings of the Workshop on Spoken Language Technologies for Under-resourced languages*.
- Macuch Silva, V., Holler, J., Ozyurek, A., and Roberts, S. G. (2020). Multimodality and the origin of a novel communication system in face-to-face interaction. *Royal Society open science*, 7(1):182056.
- Malkin, N., Wang, Z., and Jojic, N. (2022). Coherence boosting: When your pretrained language model is not paying enough attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8214–8236. Association for Computational Linguistics.
- Markl, N. (2023). *Language variation, automatic speech recognition and algorithmic bias*. PhD thesis, The University of Edinburgh.
- Marslen-Wilson, W. D. and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1):29–63.
- Matzinger, T., Ritt, N., and Fitch, W. T. (2021). The influence of different prosodic cues on word segmentation. *Frontiers in Psychology*.
- Mayo, C., Aylett, M., and Ladd, D. R. (1997). Prosodic transcription of glasgow english: an evaluation study of glatobi. In *Intonation: Theory, Models and Applications*.
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of Interspeech*, volume 2017, pages 498–502.
- McConnell, K. and Blumenthal-Dramé, A. (2021). Usage-based individual differences in the probabilistic processing of multi-word sequences. *Frontiers in Communication*, 6:703351.

- McDonald, S. A. and Shillcock, R. C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research*, 43(16):1735–1751.
- McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.
- Mehta, G. and Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech*, 31:135–156.
- Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., and Levy, R. (2021). Revisiting the Uniform Information Density hypothesis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 963–980. Association for Computational Linguistics.
- Meizel, K. (2020). *Multivocality: Singing on the Borders of Identity*. Oxford University Press, USA.
- Mens, L. H. and Povel, D.-J. (1986). Evidence against a predictive role for rhythm in speech perception. *The Quarterly Journal of Experimental Psychology Section A*, 38(2):177–192.
- Merchant, A., Rahimtoroghi, E., Pavlick, E., and Tenney, I. (2020). What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44. Association for Computational Linguistics.
- Mesgar, M., Bücker, S., and Gurevych, I. (2020). Dialogue coherence assessment without explicit dialogue act labels. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1439–1450.
- Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., et al. (2021). Between words and characters: a brief history of open-vocabulary modeling and tokenization in NLP. *arXiv:2112.10508*.
- Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*.

- Milde, B. and Biemann, C. (2018). Unspeech: Unsupervised speech context embeddings. In *Proceedings of Interspeech*, volume 2018-Sept, pages 2693–2697.
- Mills, J. (2020). Delexicalised auditory priming of implicit prosody. In *Proceedings of Speech Prosody*.
- Mitrovic, J., McWilliams, B., and Rey, M. (2020). Less can be more in contrastive learning. In *Proceedings of the Conference on Neural Information Processing Systems; “I Can’t Believe It’s Not Better!” Workshop*.
- Mixdorff, H., Cole, J., and Shattuck-Hufnagel, S. (2012). Prosodic similarity – evidence from an imitation study. In *Proceedings of Speech Prosody*, pages 571–574.
- Moll, J. (2013). *The influence of Modality Combinations on Communication in Collaborative Virtual Environments*. PhD thesis, KTH Royal Institute of Technology.
- Moon, S., Neves, L., and Carvalho, V. (2018). Multimodal named entity recognition for short social media posts. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 852–860. Association for Computational Linguistics.
- Morgan, J. (1973). *Sentence Fragments and the Notion ‘Sentence’*. University of Illinois Press.
- Morise, M., Kawahara, H., and Katayose, H. (2009). Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society.
- Morise, M., Yokomori, F., and Ozawa, K. (2016). World: a vocoder-based high-quality speech synthesis system for real-time applications. *Institute of Electronics, Information and Communication Engineers; Transactions*.
- Mosbach, M., Andriushchenko, M., and Klakow, D. (2021). On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.
- Murray, G., Renals, S., and Carletta, J. (2005). Extractive summarization of meeting recordings. In *Proceedings of Interspeech*, pages 593–596.

- Murray, G., Taboada, M., and Renals, S. (2006). Prosodic correlates of rhetorical relations. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 1–7. Association for Computational Linguistics.
- Nath, A. and Ward, N. G. (2022). On the predictability of the prosody of dialog markers from the prosody of the local context. In *Proceedings of Speech Prosody*, pages 664–668.
- Newman, B., Hewitt, J., Liang, P., and Manning, C. D. (2020). The EOS decision and length extrapolation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 276–291. Association for Computational Linguistics.
- Nielsen, E. (2024). *Effects of prosody on natural language processing*. PhD thesis, The University of Edinburgh.
- Nielsen, E., Steedman, M., and Goldwater, S. (2021). Prosodic segmentation for parsing spoken dialogue. *arXiv preprint arXiv:2105.12667*.
- Nielsen, E., Steedman, M., and Goldwater, S. (2023). Prosodic features improve sentence segmentation and parsing in english. In *Proceedings of Interspeech*.
- Nieuwland, M. S. and Berkum, J. J. A. V. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18:1098–1111.
- Noble, B. and Maraev, V. (2021). Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics*, pages 166–172. Association for Computational Linguistics.
- Noufi, C. and Verma, P. (2020). Self-supervised learning of context-aware pitch prosody representations. *arXiv:2007.09060*.
- Obin, N., Beliao, J., Veaux, C., and Lacheret, A. (2014). Slam: Automatic stylization and labelling of speech melody. In *Proceedings of Speech Prosody*, page 246.
- Ofer, D., Brandes, N., and Linial, M. (2021). The language of proteins: Nlp, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, 19:1750–1758.

- Oh, B.-D. and Schuler, W. (2023a). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics; Empirical Methods in Natural Language Processing*, pages 1915–1921, Singapore. Association for Computational Linguistics.
- Oh, B.-D. and Schuler, W. (2023b). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.
- Ohala, J. J. (1983). Cross-language use of pitch: an ethological view. *Phonetica*, 40(1):1–18.
- O’Mahony, J., Oplustil-Gallegos, P., Lai, C., and King, S. (2021). Factors affecting the evaluation of synthetic speech in context. *Proceedings of the Speech Synthesis Workshop*.
- Ostendorf, M., Price, P. J., and Shattuck-Hufnagel, S. (1995). The boston university radio news corpus. *Linguistic Data Consortium*, pages 1–19.
- Paeschke, A., Kienast, M., Sendlmeier, W. F., et al. (1999). F0-contours in emotional speech. In *Proceedings of the International Congress of Phonetic Sciences*, volume 2, pages 929–932.
- Paliwal, K. K., Lyons, J. G., and Wójcicki, K. K. (2010). Preference for 20-40 ms window duration in speech analysis. In *2010 4th International Conference on Signal Processing and Communication Systems*, pages 1–4. IEEE.
- Parcalabescu, L., Trost, N., and Frank, A. (2021). What is multimodality? In *Proceedings of the Workshop on Multimodal Semantic Representations*, pages 1–10.
- Parker, D. (2019). Two minds are not always better than one: Modeling evidence for a single sentence analyzer. *Glossa*.
- Pasad, A., Chou, J.-C., and Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 914–921. IEEE.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.

- (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of Advances in Neural Information Processing Systems*, pages 8024–8035.
- Pate, J. K. and Goldwater, S. (2013). Unsupervised dependency parsing with acoustic cues. *Transactions of the Association for Computational Linguistics*, 1:63–74.
- Pecchinenda, A. and Petrucci, M. (2016). Emotion unchained: Facial expression modulates gaze cueing under cognitive load. *PLoS One*, 11(12):e0168111.
- Peng, B., Zhu, C., Li, C., Li, X., Li, J., Zeng, M., and Gao, J. (2020). Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics; Empirical Methods in Natural Language Processing*, pages 172–182.
- Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. Association for Computational Linguistics.
- Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27 2:169 – 190.
- Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In Philip R. Cohen, Jerry Morgan, M. E. P., editor, *Intentions in Communication*. The MIT Press.
- Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology.
- Pimentel, T., Meister, C., Wilcox, E. G., Levy, R. P., and Cotterell, R. (2023). On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.
- Pirker, G., Wohlmayr, M., Petrik, S., and Pernkopf, F. (2011). A pitch tracking corpus with evaluation on multipitch tracking scenario. In *Proceedings of Interspeech*, pages 1509–1512.
- Pisoni, D. B. (1978). Speech perception. *Handbook of learning and cognitive processes*, 6:167–233.

- Pols, L. and Plomp, R. (1986). How to make more efficient use of the fact that the speech signal is dynamic and redundant. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 1963–1966. IEEE.
- Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.-N., Mohamed, A., and Dupoux, E. (2021). Speech resynthesis from discrete disentangled self-supervised representations. In *Proceedings of Interspeech*.
- Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M., and Cox, D. (2020). Unsupervised speech decomposition via triple information bottleneck. In *Proceedings of the International Conference on Machine Learning*, pages 7836–7846. PMLR.
- Qian, K., Zhang, Y., Gao, H., Ni, J., Lai, C.-I., Cox, D., Hasegawa-Johnson, M., and Chang, S. (2022). Contentvec: An improved self-supervised speech representation by disentangling speakers. In *International Conference on Machine Learning*, pages 18003–18017. Proceedings of Machine Learning Research.
- Qian, T. and Jaeger, T. F. (2011). Topic shift in efficient discourse production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.
- Räsänen, O., Kakouros, S., and Soderstrom, M. (2018). Is infant-directed speech interesting because it is surprising?—Linking properties of IDS to statistical learning and attention at the prosodic level. *Cognition*.
- Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2021). Fast-speech 2: Fast and high-quality end-to-end text to speech. In *Proceedings of the International Conference on Learning Representations*.
- Richter, S. and Chaves, R. (2020). Investigating the role of verb frequency in factive and manner-of-speaking islands. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society*. cognitivesciencesociety.org.

- Rodd, J., Decuyper, C., Bosker, H. R., and Ten Bosch, L. (2021). A tool for efficient and accurate segmentation of speech data: announcing ponss. *Behavior research methods*, 53:744–756.
- Roddy, M., Skantze, G., and Harte, N. (2018). Multimodal continuous turn-taking prediction using multiscale RNNs. In *Proceedings of the International Conference on Multimodal Interaction*, pages 186–190.
- Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Rosenberg, A. (2009). *Automatic detection and classification of prosodic events*. PhD thesis, Columbia University.
- Rosenberg, A. (2010). Classification of prosodic events using quantized contour modeling. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 721–724.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.
- Roy, J., Cole, J., and Mahrt, T. (2017). Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology*, 8(1).
- Rühlemann, C., Bagoutdinov, A., and O’Donnell, M. B. (2011). Windows on the mind. *International Journal of Corpus Linguistics*, 16(2):198–230.
- Rumelhart, D. E., Hinton, G. E., McClelland, J. L., et al. (1986). A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(45-76):26.
- Rust, P., Lotz, J. F., Bugliarello, E., Salesky, E., de Lhoneux, M., and Elliott, D. (2022). Language modelling with pixels. *arXiv preprint arXiv:2207.06991*.
- Sacks, H., Schegloff, E. A., and Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

- Sanford, A. J. and Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, 6(9):382–386.
- Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics*, 11(3):243–257.
- Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.
- Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).
- Scott-Phillips, T. (2014). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Macmillan International Higher Education.
- Selkirk, E. O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press.
- Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Shain, C., Meister, C., Pimentel, T., Cotterell, R., and Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
- Shirali-Shahreza, S. and Penn, G. (2018). Mos naturalness and the quest for human-like speech. In *IEEE Spoken Language Technology Workshop*, pages 346–352. IEEE.
- Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., Taylor, P., Ries, K., Martin, R., and Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492.

- Sigurgeirsson, A. T. and King, S. (2023). Do Prosody Transfer Models Transfer Prosody? In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B., and Hirschberg, J. (1992). Tobi: A standard for labelling english prosody. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 867–870.
- Sjerps, M. J., Decuyper, C., and Meyer, A. S. (2019). Initiation of utterance planning in response to pre-recorded and “live” utterances. *Quarterly Journal of Experimental Psychology*, 73:357 – 374.
- Skantze, G. (2020). Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, page 101178.
- Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R., Clark, R., and Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In *Proceedings of the International Conference on Machine Learning*.
- Skoog Waller, S., Eriksson, M., and Sörqvist, P. (2015). Can you hear my age? influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in psychology*, 6:978.
- Stephens, G. J., Silbert, L. J., and Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences of the United States of America*, 107(32):14425–14430.
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.
- Stokoe, E. (2018). *Talk: The science of conversation*. Hachette UK.
- Suni, A., Kakouros, S., Vainio, M., and Simko, J. (2020). Prosodic prominence and boundaries in sequence-to-sequence speech synthesis. In *Proceedings of Speech Prosody*.

- Suni, A., Simko, J., Aalto, D., and Vainio, M. (2017). Hierarchical representation and estimation of prosody using continuous wavelet transform. *Comput. Speech Lang.*
- Swanson, L. (2016). The predictive processing paradigm has roots in kant. *Frontiers in Systems Neuroscience*, 10.
- Syrdal, A. K. and Kim, Y.-J. (2008). Dialog speech acts and prosody: considerations for TTS. In *Proceedings of Speech Prosody*, pages 661–665.
- Talkin, D. and Kleijn, W. B. (1995). A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, 495:518.
- Tang, K. and Shaw, J. A. (2020). Prosody leaks into the memories of words. *Cognition*, 210.
- Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.
- Ten Bosch, L., Oostdijk, N., and Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1-2):80–86.
- Tian, Y., Krishnan, D., and Isola, P. (2020). Contrastive multiview coding. In *Proceedings of the European Conference of Computer Vision*, pages 776–794. Springer.
- Toledano, D. T., Gómez, L. A. H., and Grande, L. V. (2003). Automatic phonetic segmentation. *IEEE transactions on speech and audio processing*, 11(6):617–625.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. Technical report, Meta AI.
- Tran, T. (2020). *Neural Models for Integrating Prosody in Spoken Language Understanding*. PhD thesis, University of Washington.
- Tree, J. E. F. and Meijer, P. J. A. (2000). Untrained speakers’ use of prosody in syntactic disambiguation and listeners’ interpretations. *Psychological Research*, 63:1–13.
- Tremblay, A., Broersma, M., and Coughlin, C. E. (2018). The functional weight of a prosodic cue in the native language predicts the learning of speech segmentation in a second language. *Bilingualism: Language and Cognition*, 21(3):640–652.

- Turk, A. (2010). Does prosodic constituency signal relative predictability? a smooth signal redundancy hypothesis. *Laboratory phonology*, 1(2):227–262.
- Tzeng, C. Y., Namy, L. L., and Nygaard, L. C. (2019). Communicative context affects use of referential prosody. *Cognitive science*, 43 11.
- Urbanek, J., Fan, A., Karamcheti, S., Jain, S., Humeau, S., Dinan, E., Rocktäschel, T., Kiela, D., Szlam, A., and Weston, J. (2019). Learning to speak and act in a fantasy text adventure game. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 673–683. Association for Computational Linguistics.
- Uria, B., Côté, M.-A., Gregor, K., Murray, I., and Larochelle, H. (2016). Neural autoregressive distribution estimation. *Journal of Machine Learning Research*, 17(205):1–37.
- van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Vander Klok, J., Goad, H., and Wagner, M. (2018). Prosodic focus in english vs. french: A scope account. *Glossa: a journal of general linguistics*, 3(1).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems*, pages 5998–6008.
- Vega, A. and Ward, N. G. (2009). Looking for entropy rate constancy in spoken dialog. Technical report, University of Texas at El Paso.
- Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27(5):1413–1432.
- Venditti, J. J. (1997). Japanese tobi labelling guidelines. Technical report, Ohio State University. Department of Linguistics.
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine learning*, pages 1096–1103.

- Vogels, J., Howcroft, D. M., Tourtouri, E., and Demberg, V. (2020). How speakers adapt object descriptions to listeners under load. *Language, Cognition and Neuroscience*, 35(1):78–92.
- von Helmholtz, H. (1924). *Helmholtz’s treatise on physiological optics (Trans. from the 3rd German ed.)*. the Optical Society of America.
- Vu, L., Phan, R. C.-W., Han, L. W., and Phung, D. (2022). Improved speech emotion recognition based on music-related audio features. In *Proceedings of the European Signal Processing Conference*, pages 120–124.
- Wagner, M. et al. (2005). *Prosody and recursion*. PhD thesis, Massachusetts Institute of Technology.
- Wagner, M. and Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945.
- Wallbridge, S., Bell, P., and Lai, C. (2023). Quantifying the perceptual value of lexical and non-lexical channels in speech. In *Proceedings of Interspeech*.
- Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., and Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the International Conference on Machine Learning*, pages 5180–5189. PMLR.
- Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.
- Ward, N. G. and Walker, B. H. (2009). Estimating the potential of signal and interlocutor-track information for language modeling. *Proceedings of Interspeech*.
- Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Welles, O., Bogdanovich, P., Rosenbaum, J., et al. (1998). *This is Orson Welles*. Da Capo Press New York.
- Weston, J., Lenain, R., Meepegama, U., and Fristed, E. (2021). Learning de-identified representations of prosody from raw audio. In *Proceedings of the International Conference on Machine Learning*, pages 11134–11145. PMLR.

- Wilcox, E., Levy, R., Morita, T., and Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In *Proceedings of BlackboxNLP@EMNLP; Empirical Methods in Natural Language Processing*.
- Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., and Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713. Cognitive Science Society.
- Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., and Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.
- Willems, R. M. and Peelen, M. V. (2021). How context changes the neural basis of perception and language. *IScience*, 24(5).
- Williams, J. (2022). *Learning disentangled speech representations*. PhD thesis, University of Edinburgh.
- Williams, J. and King, S. (2019). Disentangling style factors from speaker representations. In *Proceedings of Interspeech*, pages 3945–3949.
- Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *ArXiv*, abs/1004.2515.
- Wilson, K. and Carroll, J. B. (1954). Applications of entropy measures to problems of sequential structure. *Psycholinguistics: A survey of theory and research*, pages 103–110.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2019). TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *ArXiv*, abs/1901.08149.

- Wu, C.-S., Hoi, S. C., Socher, R., and Xiong, C. (2020). TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 917–929, Online. Association for Computational Linguistics.
- Wu, C.-S. and Xiong, C. (2020). Probing task-oriented dialogue representation from language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5036–5051. Association for Computational Linguistics.
- Wu, S., Du, Z., Li, W., Huang, D., and Wang, Y. (2019). Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. *Proceedings of the International Conference on Multimodal Interaction*.
- Wu, Y., Wu, W., Xing, C., Zhou, M., and Li, Z. (2017). Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.
- Xie, Q., Yang, S., Lei, Y., Xie, L., and Su, D. (2022). End-to-end voice conversion with information perturbation. In *Proceedings of the International Symposium on Chinese Spoken Language Processing*, pages 91–95.
- Xu, H. and Armony, J. L. (2021). Influence of emotional prosody, content, and repetition on memory recognition of speaker identity. *Quarterly Journal of Experimental Psychology*, 74(7):1185–1201.
- Xu, R., Tao, C., Jiang, D., Zhao, X., Zhao, D., and Yan, R. (2020). Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues. *ArXiv preprint*, abs/2009.06265.
- Yager, R. R. (2001). Dissonance: a measure of variability for ordinal random variables. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9:39–53.
- Yang, S., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., tik Lee, K., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., and yi Lee, H. (2021).

- SUPERB: Speech Processing Universal PERFORMANCE Benchmark. In *Proceedings of Interspeech*, pages 1194–1198.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Proceedings of the Annual Conference on Neural Information Processing Systems*.
- Yuan, J., Brenier, J. M., and Jurafsky, D. (2005). Pitch accent prediction: effects of genre and speaker. In *Proceedings of Interspeech*, pages 1409–1412.
- Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. (2021). Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.
- Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. (2022). Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.
- Zen, H., Dang, V.-T., Clark, R. A. J., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. (2019). Libritts: A corpus derived from librispeech for text-to-speech. In *Proceedings of Interspeech*.
- Zhang, Z., Li, J., Zhu, P., Zhao, H., and Liu, G. (2018). Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Zhou, X., Dong, D., Wu, H., Zhao, S., Yu, D., Tian, H., Liu, X., and Yan, R. (2016). Multi-view response selection for human-computer conversation. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381. Association for Computational Linguistics.