Quantifying the perceptual value of lexical and non-lexical channels in speech

Quantifying the perceptual value of lexical and non-lexical channels in speech

Sarenne Wallbridge, Peter Bell, Catherine Lai

Centre for Speech Technology Research, University of Edinburgh

{s1301730, peter.bell, c.lai}@ed.ac.uk

Abstract

Speech is a fundamental means of communication that can be seen to provide two channels for transmitting information: the lexical channel of *which* words are said, and the non-lexical channel of *how* they are spoken. Both channels shape listener expectations of upcoming communication; however, directly quantifying their relative effect on expectations is challenging. Previous attempts require spoken variations of lexically-equivalent dialogue turns or conspicuous acoustic manipulations. This paper introduces a generalised paradigm to study the value of non-lexical information in dialogue across unconstrained lexical content. By quantifying the perceptual value of the non-lexical channel with both accuracy and entropy reduction, we show that non-lexical information produces a consistent effect on expectations of upcoming dialogue: even when it leads to poorer discriminative turn judgements than lexical content alone, it yields higher consensus among participants.

Index Terms: spoken dialogue, speech perception, prosody, discourse structure

1. Introduction

The human language system is often modelled as a predictive processor where expectations about the upcoming linguistic signal are conditioned on a host of contextual cues, including the previous signal [1]. These cognitive mechanisms have likely evolved to optimize our predictive capabilities for spoken interactions, a modality which can be framed as a multi-channel signal consisting of the lexical channel and non-lexical channels [2, 3]. Although there is plentiful evidence that both channels are used to encode and decode information, the relative effect of these channels on human expectations in dialogue is unclear [4, 5, 6]. In this work, we quantify the value of information in the lexical and non-lexical channels by how much it constrains human expectations regarding the upcoming dialogue turn. In particular, we address a previously-unanswered question of *how the non-lexical channel affects expectations when the lexical channel is uninformative*.

Attempts to disentangle channel effects often use acoustic manipulations such as low-pass filtering to delexicalise speech, and flattening pitch curves to remove changes in intonation [7, 8]. These modifications may be conspicuous and leave certain acoustic properties such as duration intact. In previous work, we proposed the turn discrimination paradigm, which instead disentangles channel effects using separate lexical and acoustic conditions [3]. Rather than examining a specific function of non-lexical information, this method quantifies the value of non-lexical information as how much it affects performance of the generic task of discriminating between upcoming turns. By sampling contexts and responses from a conversational corpus, we showed that people use prosodic cues to discriminate the true response from alternative prosodic realisations in natural dialogue. This experimental design used sets of lexically-equivalent responses to isolate variation between responses to their prosodic realisations. However, this severely limited our ability to quantify non-lexical channel value beyond short, often back-channel responses [9], or across variable lexical content.

In this work, we present a generalisation of the turn discrimination paradigm using a dialogue-based language model (LM) to select lexically diverse but similarly plausible turns, as well as a novel quantification of channel value as entropy reduction to account for the inherent optionality in upcoming dialogue. This augmented paradigm enables investigation into the effect of non-lexical information on dialogue acceptability perception for variable lexical content, and allows us to more fully address the question of how non-lexical information is used when the lexical channel is uninformative. We find that when the lexical channel is ambiguous, non-lexical information increases discriminative performance. However, when lexical content is informative, non-lexical information can worsen discrimination performance. Interestingly, it does so consistently: people tend to interpret non-lexical cues in similar ways even if this leads to incorrect judgements about the upcoming turn.

2. Background

2.1. Speech: multi-channel communication

As a communicative medium, speech encodes information in both lexical and non-lexical content. Non-lexical information includes features of a speaker's identity and environment. However, in this work, we are primarily interested in prosodic information and how it affects dialogue perception. Generally quantified by the acoustic correlates for intonation, intensity, and rhythm (F0, energy, and unit duration), prosody contributes to many important communicative functions such as marking novel information and topic shifts, conveying attitudinal reactions and uncertainty, and managing turn-taking [10].

There is experimental evidence for the interaction of lexical and prosodic information [11]; for example, [12] shows that lexical and non-lexical channels are used jointly to mediate information density. However, such studies often use carefully constructed stimuli or involve potentially conspicuous acoustic manipulation, a far cry from conversational speech which is rife with features like disfluencies and phonetic reduction [13].

Interest in the role of prosodic and lexical information in dialogue has a long history [14]. Still, only a small number of psycholinguistic studies have explored the use of prosody in dialogue and for good reason [8, 15, 16, 17]. Disentangling lexical and non-lexical effects is unsurprisingly difficult given the complicated relationship between prosody and communicative

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 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]*

intent, and the complexity of studying natural dialogue.

Most similar to the work at hand is [18] which investigated the roles of prosody and conversational context for turn-end estimation. Using button-press experiments in conditions where participants either receive written transcripts or the full speech signal, phrase-final prosodic cues were found to be relevant for turn-end projection. However, their results reflect the complexity of processing and studying dialogue. Although acoustic context facilitated turn-end estimation for short utterances, it produced an inhibitory effect on the accuracy of turn-end estimation for longer turns. Similarly, [19] found that listeners leverage acoustic discourse context to make predictions. However, when acoustic context was relevant, participants' turn-end judgements were produced earlier, but not more precisely.

2.2. Optionality in communication

While the studies above provide evidence that both the lexical and non-lexical channels affect dialogue perception, conclusions don't converge neatly. The majority of these studies [8, 15, 17, 18, 19] quantified channel value with accuracy-based metrics which don't account for the expected variability of the upcoming signal–its optionality. As text and speech generation approaches human-like naturalness, better understanding of appropriate production variability is increasingly important. We propose quantifying channel value in terms of entropy reduction to account for optionality in upcoming dialogue.

The informational value of a message is often framed in terms of how surprising it is [20, 1, 21]. Highly surprising messages are informative but may be difficult to process. Conversely, if the future signal is perfectly predictable, there is little value in expending effort to communicate at all. Recent work has demonstrated that, rather than being more probable, monological text that is perceived as natural encodes an amount of information that is close to the expected information content of natural language [22]. In other words, humans expect the upcoming linguistic signal to be within a certain interval of surprisal. The degree of predictability in spoken dialogue is likely to be more complicated: it involves multiple parties with potentially different goals, and variation can occur across both the lexical and non-lexical channels. Enactment studies demonstrate the increased complexity of production variability in speech: different speakers produce different prosodic realisations of the same lexical content [23].

3. Experimental design

3.1. Generalised turn discrimination paradigm

To quantify the perceptual value of lexical and non-lexical channels in constraining expectations of upcoming communication, we use the task of turn-acceptability presented in our previous work [3]: given conversational turns as context, participants are asked to rate the plausibility of potential continuations. Stimuli are presented as transcripts (lexical condition) or as speech recordings (acoustic condition). The value of non-lexical information is quantified by rating differences between conditions.

As described in the Introduction, the requirement of lexically-equivalent response sets in the original paradigm was a severe limitation. However, large LMs are known to align with aspects of human perception of monologues [24, 25, 26]. Along with others [27, 28], we have recently shown that dialogue-based LM scores also correlate with human turn acceptability judgements [29]. Building on these recent findings, we remove the lexical-equivalence constraint by using a dialogue-based LM to sample plausible responses (model details are below). Figure 1 displays a stimulus sampled from our model; responses are similarly plausible but lexically diverse. This generalised methodology allows us to investigate the effect of non-lexical information on dialogue acceptability perception across unconstrained lexical content and variable responses.

3.2. Task & stimuli design

Each stimuli consists of 3 contiguous turns of a conversation and five potential responses. Participants are instructed that one response is the true continuation for this conversation, then asked to score the plausibility of each response on a scale from 1 ('Very Unlikely') to 4 ('Very Likely'). We describe the stimuli construction and conditions below.

Data We construct stimuli from the Switchboard Telephone Corpus [30] which consists of over 2,400 dual-channel conversations between 542 speakers. 642 of these conversations were annotated post-hoc with information such as dialogue acts, information status, and prosodic features (Switchboard NXT [31]). We use these conversations as validation and test sets. We segment all conversations into turns using the associated word timings. First, all words spoken contiguously by a speaker are joined into segments. We remove completely overlapping segments before joining contiguous speaker segments into turns.

Dialogue language model We use a state-of-the-art response selection model to sample plausible responses (cf. [29]). The architecture is a BERT cross-encoder post-trained and fine-tuned on Switchboard [32]. Our model is implemented in PyTorch. We use `bert-base-uncased` from the Transformers library as our base model and follow the training procedure from the original paper. Post-training augments the standard masked LM task with an utterance relevance classification task; response selection is the fine-tuning objective. Both train stages apply early stopping using our validation set of Switchboard. We post-train and fine-tune for 9 and 17 epochs, respectively.

Transcript preprocessing We clean the transcripts to train our dialogue-based LM and present stimuli to participants. In particular, characters specific to the Switchboard transcription guidelines[1] (mispronunciations, pronunciation variants, partial words, coinages) and non-speech sounds including vocalised noises and laughter are removed. We treat these as transcriptions of the non-lexical channel, rather than lexical content. However, the importance of filled pauses such as "uh" and "um" is widely accepted [33, 34]. We opt for the more conservative channel split and retain them in transcripts.

Audio preprocessing Overlapping speech is a prominent feature of conversation. Although Switchboard recordings are dual-channel (one per speaker), certain conversations contain

[1] `https://isip.piconepress.com/projects/switchboard`

significant channel bleed. To maintain overlapping turns in our stimuli, we de-bleed all conversations: projections of the power spectra of both speaker channels are subtracted from each other before recomposing the individual channels into waveforms.

Stimuli construction We randomly select a context as 3 contiguous speaker turns from the NXT conversations. We randomly sample 1000 unique responses to score with our dialogue LM. The five responses for this context consist of its true response, along with the top-scoring responses. Random sampling allows us to explore the effect of non-lexical information across a range of linguistic functions. However, we perform broad filtering to ensure stimuli contain enough information but are not too long for the behavioural study. The final context turn must contain $3-50$ tokens with an audio length of $2-10$ s. Potential responses must be of length $0.25-10$ s, preceded by a pause within $-2-2$ s, 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 total[2].

Conditions In both lexical and acoustic conditions, participants receive transcripts of the context. In the lexical condition, participants also receive transcripts of all potential responses; in the acoustic condition, responses are presented in audio format with each response spliced onto the final context turn. Following [3], pause length is treated as a feature of utterance design. Each response is thus extracted with its preceding pause which is used to join it with the final context turn.

3.3. The online task & stimuli sets

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 one condition. 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). Participants were 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 resp.)

Stimuli sets Our primary research question is whether participants use non-lexical information to guide expectations about an upcoming turn when the lexical channel is uninformative but diverse. To ensure that lexical content is uninformative, we create a subset of stimuli where participants did not reliably score the true response highest in the lexical condition (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 ambiguous stimuli.

3.4. Metrics

We use several metrics to quantify the perceptual value of acoustic information. Accuracy is easy to interpret, however, human acceptability judgements have been shown to be probabilistic [35, 24, 29]. As such, it is possible for multiple responses to be considered plausible. To account for this optionality, we employ entropy reduction to examine the convergence of participants' judgements at both response- and question-levels. Entropy has previously been used to quantify the potential value of non-lexical component of spoken dialogue [36].

Accuracy Accuracy reflects the frequency with which the true response was rated highest. We also weight this frequency by the proportion of score mass assigned to the true response by each participant (weighted accuracy).

Ordinal Entropy We measure entropy-per-response using a variant of cumulative paired ϕ-entropy to account for the ordinal nature of scores [37]. Standard entropy $H(S)$ is a function of categorical label probabilities. Ordinal entropy $H_{Ord}(S)$ is instead a function of the cumulative probability for each score, thus reflecting dispersion among scores [38].

Permutation Entropy Permutation entropy measures entropy at the stimulus-level. Ordinal Pattern Analysis (OPA) is often used to quantify the complexity of time-series data by converting it to a sequence of ordinal patterns before computing standard Shannon entropy $H(S)$ across pattern frequencies [39]. We convert participant scores across responses to rank patterns. This quantifies agreement at the stimulus-level.

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

$$H_{ord}(S) = -\sum_{i=1}^{n} \Big(p_{\leq k} \log p_{\leq k} - (1 - p_{\leq k}) \log(1 - p_{\leq k}) \Big) \tag{2}$$

For score counts S_i over scores $i \in 1, ..., n$, we denote $p_i = P(S_i)$ and $p_{\leq k} = \sum_{i=1}^{k} p_i$

4. Results & Discussion

To test whether participants use non-lexical information to guide their expectations when the lexical channel is uninformative but unconstrained, we compare scoring behaviour in the lexical and acoustic conditions across the ambiguous stimuli. Chance performance is estimated by shuffling each set of participant ratings 100 times to maintain score distributions.

Additionally, we analyze the effect of condition on ordinal entropy 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 [40]. To investigate potential cognitive load differences, we include *response length* (in seconds) as a predictor. Following [27], we include the *mean surprisal* of the response conditioned on the context, and 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 [31], as well as effects for context, and context-response identifiers are included to control for stimuli variation. We include interaction terms with response length, target, mean surprisal, and dialogue acts to see if effects varied with the acoustic/lexical condition. We see non-zero variance estimates for the group-level effects, i.e., these factors do account for variation in the entropy.

We use the emmeans package [41] to compute estimated marginal means and 95% Highest Posterior Density Regions, i.e. Credible Intervals (CIs), to examine effects of predictors.

4.1. Accuracy

As can be seen in Table 1, accuracy for *ambiguous* stimuli in the lexical condition is close to chance. In the acoustic condition,

[2]We publish our stimuli: `https://sarenne.github.io/is-2023`

Table 1: *Evaluations of ambiguous stimuli across conditions, and the mean per-question difference between them (standard, weighted accuracy (acc, $acc_{\{W\}}$), ordinal, permutation entropy (H_{Ord}, H_{Perm}); all normalised). Differences are all significant using a directional Wilcoxon rank sum test ($p < 0.002$).*

	Condition			
Metric	***Chance***	**Lexical**	**Acoustic**	**Difference**
acc	0.07 ± 0.01	0.08	0.25	0.16
$acc_{\{W\}}$	0.05 ± 0.01	0.05	0.17	0.11
H_{Ord}	0.66 ± 0.01	0.54	0.48	-0.06
H_{Perm}	0.94 ± 0.01	0.90	0.86	-0.04

Table 2: *Accuracy metrics for check questions*

	Condition		
Metric	**Lexical**	**Acoustic**	**Difference**
acc	0.96	0.96	0.00
$acc_{\{W\}}$	0.91	0.74	-0.17

it is significantly higher. Increased accuracy provides strong evidence that participants leverage non-lexical cues to constrain their expectations about the upcoming dialogue turn.

Table 3 contains results for the *remaining* stimuli–where participants could discriminate the true response relatively accurately from the lexical channel alone. People judge these stimuli more accurately. However, their accuracy drops in the acoustic condition. Surprisingly, people are less apt at selecting the true turn when provided with non-lexical information.

Accuracy metrics for the check questions are shown in Table 2. Weighted accuracy decreases by -0.17 in the acoustic condition, reflecting less decisive scores for check questions. This is surprising as multi-modal communication is suggested to offer greater communicative power than single-channel communication [42, 43]. We hypothesise that acoustic stimuli place a higher cognitive load on participants. As such, these results likely understate non-lexical channel value.

4.2. Entropy Reduction

Given that the upcoming communicative signal is not perfectly predictable, 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 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 3 shows reduced accuracy between conditions for the *remaining* stimuli, score entropies decrease and to similar degrees as found across the lexically-ambiguous stimuli. The non-lexical channel seems to produce consistent perceptual effects, regardless of how informative lexical content is for turn discrimination.

Higher entropy in the lexical condition is reflected by our regression model, but the difference varies depending on other factors. In particular, we see an interaction between condition and response length: a positive slope estimate in the acoustic condition (0.02, CI=(0.01, 0.03)), and a flat slope for the lexical condition (-0.001, CI=(-0.01, 0.01)). That is, participants show higher agreement in the acoustic condition than the lexical condition for short responses, but the difference shrinks with utterance length. This further suggests potential differences in cognitive load between conditions.

We calculate estimated marginal mean ordinal entropy for true and false responses (by condition), while averaging over other predictors, to see their effect. We see lower entropy for true responses overall. Similarly, entropy is reduced in the acoustic condition compared to the lexical condition. However, the difference between conditions is greater for true responses (-0.13, CI=(-0.26,-0.001) vs. (-0.10, CI=(-0.20, 0.001)). This suggests that true responses have acoustic features that participants make use of for this task.

Table 3: *Evaluations of remaining stimuli. Mean per-question differences are all significant using a directional Wilcoxon rank sum test at $p < 0.0001$ except permutation entropy ($p < 0.02$).*

	Condition			
Metric	***Chance***	**Lexical**	**Acoustic**	**Difference**
acc	0.14 ± 0.02	0.54	0.36	-0.18
$acc_{\{W\}}$	0.10 ± 0.01	0.38	0.25	-0.12
H_{Ord}	0.70 ± 0.01	0.55	0.48	-0.06
H_{Perm}	0.96 ± 0.01	0.89	0.86	-0.03

Interestingly, surprisal affects ordinal entropy differently between conditions. The estimated effect is likely positive in the acoustic condition (0.08, CI=(-0.001, 0.16)), i.e., more lexically surprising responses result in less agreement. However, the effect in the lexical condition peaks around zero (0.00, CI=(-0.07, 0.08)). The latter estimate is likely a result of the ambiguous stimuli selection, but also again indicates that acoustic information changes participant expectations.

5. Conclusions

The generalised turn-discrimination paradigm presented here enabled our analysis of how lexical and non-lexical channels are used jointly for a much broader set of language than was previously possible. Our results provide firm evidence that non-lexical information constrains expectations of spoken dialogue–people can leverage non-lexical cues to discriminate true continuations from false candidates when the lexical channel is uninformative. However, when the lexical content is informative for the discriminative task, acoustic information can hinder performance. Although surprising, similar results were found by [19] who showed that listeners respond earlier but not necessarily more accurately in turn-end detection tasks when context is informative. Quantifying channel value as entropy reduction provides a novel perspective on channel value: even when it leads to incorrect discriminative judgements, the non-lexical channel affects expectations in consistent ways.

We believe our methodology has implications both for learning perceptually-motivated representations of spoken communication and for speech generation where the degree of acceptable production variability across both lexical and non-lexical channels in dialogue is relatively unexplored. Here, we investigated a small number of factors that could affect non-lexical channel value and found evidence of complex interactions with acoustics; in future work, we hope to develop a more formal conditional quantification of channel value. 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 [44].

Acknowledgements We'd like to thank Erfan Loweimi and Cassia Valentini Botinhao for help with channel bleed removal.

6. References

[1] J. Hale, "A probabilistic earley parser as a psycholinguistic model," in *NAACL*, 2001.

[2] M. H. Christiansen and N. Chater, *The Now-or-Never bottleneck: A fundamental constraint on language*. Cambridge University Press, 2015, vol. 39.

[3] S. Wallbridge, P. Bell, and C. Lai, "It's not what you said, it's how you said it: discriminative perception of speech as a multichannel communication system," *Interspeech*, 2021.

[4] V. M. Silva, J. Holler, A. Ozyurek, and S. G. Roberts, "Multimodality and the origin of a novel communication system in face-to-face interaction," *Royal Society Open Science*, vol. 7, no. 1, 2020.

[5] M. Pfau, R. L. Holbert, S. J. Zubric, N. H. Pasha, and W. K. Lin, "Role and Influence of Communication Modality in the Process of Resistance to Persuasion," *Media Psychology*, vol. 2, no. 1, pp. 1–33, 2000.

[6] P. Cohen, "The pragmatics of referring and the modality of communication," *Computational Linguistics*, vol. 10, no. 2, pp. 97–146, 1984.

[7] J. P. de Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, vol. 82, pp. 515 – 535, 2006.

[8] S. Bögels and F. Torreira, "Listeners use intonational phrase boundaries to project turn ends in spoken interaction," *Journal of Phonetics*, vol. 52, pp. 46–57, 2015.

[9] N. G. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, pp. 1177–1207, 2000.

[10] N. G. Ward, *Prosodic Patterns in English Conversation*. Cambridge University Press, 2019.

[11] S. Gahl and S. M. Garnsey, "Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation," *Language*, vol. 80, pp. 748 – 775, 2004.

[12] T. Bögel and A. Turk, "Frequency effects and prosodic boundary strength," in *ICPhS*, 2019, pp. 1014–1018.

[13] K. Johnson, "Massive reduction in conversational american english," 2004.

[14] A. W. Black, "Predicting the intonation of discourse segments from examples in dialogue speech," in *Computing Prosody*, 1997.

[15] C. Y. Tzeng, L. L. Namy, and L. C. Nygaard, "Communicative context affects use of referential prosody," *Cognitive science*, 2019.

[16] C. Lai, "What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue," in *Interspeech*, 2010.

[17] J. E. F. Tree and P. J. A. Meijer, "Untrained speakers' use of prosody in syntactic disambiguation and listeners' interpretations," *Psychological Research*, vol. 63, pp. 1–13, 2000.

[18] S. Bögels and F. Torreira, "Turn-end estimation in conversational turn-taking: The roles of context and prosody," *Discourse Processes*, vol. 58, pp. 903 – 924, 2021.

[19] R. E. Corps, M. J. Pickering, and C. Gambi, "Predicting turn-ends in discourse context," *Language, Cognition and Neuroscience*, vol. 34, pp. 615 – 627, 2018.

[20] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, 1948.

[21] R. Levy, "Expectation-based syntactic comprehension," *Cognition*, vol. 106, no. 3, pp. 1126–1177, 2008.

[22] C. Meister, G. Wiher, T. Pimentel, and R. Cotterell, "High probability or low information? The probability–quality paradox in language generation," in *ACL*, 2022, pp. 36–45.

[23] H. Mixdorff, J. Cole, and S. Shattuck-Hufnagel, "Prosodic similarity–evidence from an imitation study," in *Speech Prosody 2012*, 2012.

[24] J. H. Lau, A. Clark, and S. Lappin, "Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge," *Cognitive Science*, vol. 41, no. 5, pp. 1202–1241, 2017.

[25] C. Meister, T. Pimentel, P. Haller, L. Jäger, R. Cotterell, and R. Levy, "Revisiting the uniform information density hypothesis," in *EMNLP*, 2021.

[26] E. G. Wilcox, J. Gauthier, J. Hu, P. Qian, and R. P. Levy, "On the predictive power of neural language models for human real-time comprehension behavior," in *CogSci*, 2020, p. 1707–1713.

[27] S. Wallbridge, P. Bell, and C. Lai, "Investigating perception of spoken dialogue acceptability through surprisal," in *Interspeech*, 2022.

[28] M. Giulianelli and R. Fernández, "Analysing human strategies of information transmission as a function of discourse context," in *CoNLL*, 2021, pp. 647–660.

[29] S. Wallbridge, P. Bell, and C. Lai, "Do dialogue representations align with perception? an empirical study," in *EACL*. Association for Computational Linguistics, May 2023, pp. 2696–2713.

[30] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," vol. 1. IEEE Computer Society, 1992, pp. 517–520.

[31] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. I. Beaver, "The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language Resources and Evaluation*, vol. 44, pp. 387–419, 2010.

[32] J. Han, T. Hong, B. Kim, Y. Ko, and J. Seo, "Fine-grained post-training for improving retrieval-based dialogue systems," in *NAACL*, 2021.

[33] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[34] S. H. Fraundorf and D. G. Watson, "The disfluent discourse: Effects of filled pauses on recall." *Journal of memory and language*, vol. 65 2, pp. 161–175, 2011.

[35] N. Chater, J. B. Tenenbaum, and A. L. Yuille, "Probabilistic models of cognition: Conceptual foundations," *Trends in Cognitive Sciences*, vol. 10, pp. 287–291, 2006.

[36] N. Ward and B. Walker, "Estimating the potential of signal and interlocutor-track information for language modeling," in *Interspeech*, 2009.

[37] I. Klein, B. Mangold, and M. Doll, "Cumulative paired ϕ-entropy," *Entropy*, vol. 18, p. 248, 2016.

[38] R. R. Yager, "Dissonance: a measure of variability for ordinal random variables," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, pp. 39–53, 2001.

[39] D. Cuesta-Frau, A. Molina-Picó, B. Vargas, and P. González, "Permutation entropy: Enhancing discriminating power by using relative frequencies vector of ordinal patterns instead of their shannon entropy," *Entropy*, vol. 21, no. 10, p. 1013, 2019.

[40] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[41] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2023, r package version 1.8.4-1. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[42] M. Fröhlich, C. Sievers, S. W. Townsend, T. Gruber, and C. P. van Schaik, "Multimodal communication and language origins: integrating gestures and vocalizations," *Biological Reviews*, vol. 94, 2019.

[43] E. A. Hebets and D. R. Papaj, "Complex signal function: developing a framework of testable hypotheses," *Behavioral Ecology and Sociobiology*, vol. 57, pp. 197–214, 2004.

[44] A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: considerations for TTS," in *Speech Prosody*, 2008, pp. 661–665.