The time scale of redundancy between prosody and linguistic context

Anonymous ACL submission

Abstract

In spoken language, speakers transmit information not only using words, but also via a rich array of non-verbal signals, which include prosody—the auditory features of speech. However, previous studies have shown that prosodic features exhibit significant redundancy with both past and future words. Here, we examine the time scale of this relationship: How many words in the past (or future) contribute to predicting prosody? We find that this scale differs for past and future words. Prosody's redundancy with past words extends across approximately 3-8 words, whereas redundancy with future words is limited to just 1-2 words. These findings indicate that the prosody-future relationship reflects local word dependencies or short-scale processes such as next word prediction, while the prosody-past relationship unfolds over a longer time scale. The latter suggests that prosody serves to emphasize earlier information that may be challenging for listeners to process given limited cognitive resources in real-time communication. Our results highlight the role of prosody in shaping efficient communication.[1]

1 Introduction

Auditory features of speech such as pitch, loudness, and tempo—collectively termed prosody—play a crucial role in conveying meaning. Prosody influences sentence-level interpretation, encoding both linguistic and para-linguistic cues relevant to the communicative context (Cole, 2015; Wagner and Watson, 2010; Breen et al., 2010). For example, prosody can signal phrase boundaries, emphasize key elements, transform statements into questions, and express sarcasm, excitement, or doubt. However, some of the cues transmitted by prosody are redundant with the information encoded in the words themselves (Wolf et al., 2023).[2] This raises the question: Why is prosody used if its information is recoverable from the words themselves?

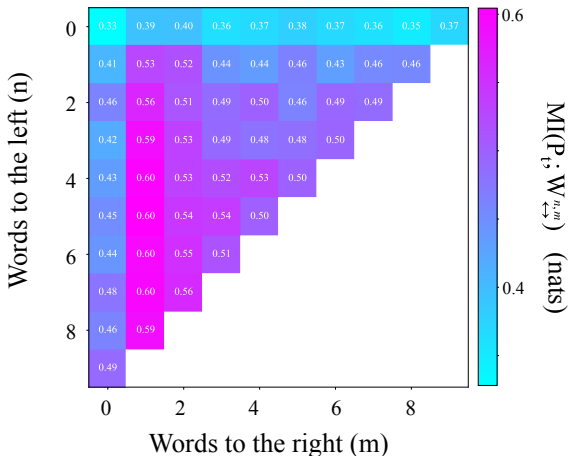

Figure 1: The redundancy between prosody P_t and linguistic context $\mathbf{W}^{n,m}_{\leftrightarrow}$—quantified as their mutual information—as a function of the number of words contained in the past (n) and/or future (m) linguistic context. The values are averaged across 6 prosodic features (see Fig. 3 for the values of each of these features separately).

One possibility is that prosody carries information that is recoverable from *long-term past context*, but not from short-term past context (see Hypothesis 1 below). Such long-term linguistic information may be challenging for listeners to maintain during real-time communication due to limited human cognitive capacities (e.g., working memory; Gibson, 1998; Lewis et al., 2006; Futrell et al., 2020) and prosody may therefore be used by speakers from an audience-design perspective[3] (Clark and Murphy, 1982). For instance, it might be possible to infer which word is most important in the current sentence given long-term linguistic context, but speakers still choose to emphasize the most

[1] Code will be added upon publication.

[2] In both (Wolf et al., 2023) and our study, text is used as a proxy to measure information present in the words themselves, also referred to as "segmental information" in the phonology literature. We thus use these terms interchangeably. We also use the term 'linguistic context' to describe a word together with the words around it.

[3] 'Audience design' is sometimes termed 'listener-oriented' or 'intelligibility-oriented' pressures.

important word using prosody to help the listener access this information more easily. This explanation implies that the information conveyed by prosody, although redundant with long-term past linguistic information, is locally unique.

Apart from redundancy with past context, however, prosody is also redundant with *future* words (Wolf et al., 2023). Since at any point in time a listener only has access to past context, if prosody is primarily shaped by audience-design, then its redundancy with the future likely reflects its role in helping listeners prepare for, or predict, upcoming words. Assuming prediction mechanisms are mostly tuned for local, incremental comprehension,[4] we hypothesize that the redundancy between prosody and future context has a shorter scale than with past context (see Hypothesis 2 below).

To summarize our predictions, we thus formulate two main hypotheses:

Hypothesis 1. ***Long-scale past redundancy***. *The redundancy between prosody and past linguistic context unfolds across a time scale that is longer than two words.*

Hypothesis 2. ***Short-scale future redundancy***. *The redundancy between prosody and future linguistic context unfolds across a time scale that is short relative to the past.*

To test these hypotheses, we build on the approach of Wolf et al. (2023), quantifying the redundancy between prosody and linguistic context as mutual information, which we estimate using pre-trained language models and a dataset of people reading aloud English audiobooks. We extend their approach to investigate the time scale of this redundancy by limiting the past and future linguistic context available to our models. We vary context lengths parametrically, from 1 to 10 words for both past and future contexts, and analyze how the redundancy changes across time for several prosodic features: pitch, loudness, duration, pause, and prominence. Our main results, averaged across these features, can be seen in Fig. 1. These results confirm that prosody is only redundant with short-term future context (up to 2 words), but with relatively longer-term past context (up to 8 words)—in agreement with our hypotheses. This advances our understanding of the role of prosody in natural spoken communication.

[4]This is suggested by recent work on surprisal theory, which shows that limiting the context of language models used to estimate surprisals improves predictive power over reading times (Kuribayashi et al., 2022).

2 Prosodic Features

Prosodic information can be conveyed via multiple acoustic features of speech, which typically manifest a high degree of co-variation. We next present the features we study here. We independently examine the scale of redundancy with linguistic context for each of these prosodic features.

Pitch Pitch is the perceptual dimension over which listeners can order sounds on a scale from low to high. The acoustic correlate giving rise to this perception is the periodicity of sound signals; pitch is thus often measured as the fundamental frequency (f_0) of the sound. Pitch is (arguably) the hallmark feature of prosody, with its contours, having been extensively studied and characterized (Pierrehumbert, 1980; Silverman et al., 1992; Jun, 2006). These contours carry contextual information that can signal a wealth of information, including, in stress-accent languages like English, emphasizing specific words, signaling boundaries, speech act type, and speaker's intent (like interrogation, sarcasm, and the affective state of the speaker). Some typical pitch curves are the rise of pitch towards the last word of a question (e.g. in yes/no questions in American English), rise and then fall of pitch on a specific word to emphasize it, and a fall toward the end of phrases.

Loudness Loudness is the perceptual dimension over which listeners can order sounds on a scale from quiet to loud. The acoustic correlate giving rise to this perception is sound pressure, being measured as the intensity of acoustic energy. The loudness of speech can be used to transmit information such as emphasize important words or convey emotion. The correlation between pitch and loudness is partly explained by vocal production constraints; producing speech with higher energy is helpful for raising and better controlling the fundamental frequency. However, loudness variations may also convey independent information from pitch.

Duration A word's duration is the difference between its offset (end) and onset (start) times. The relationship between word duration and linguistic information has long been studied as a signature of efficiency in communication such that more predictable words are reduced to a shorter duration (Jurafsky et al., 1998; Bell et al., 2009; Seyfarth, 2014; Coupé et al., 2019; Pimentel et al., 2021). Further, elongating a word is a common way to

emphasize it, or signal prosodic boundary. Duration is thus also highly correlated with pitch and loudness in natural speech, but it can also be used independently to convey meaning, or to compress words of low information content.

Pause A word's pause is the time difference between its offset (end) time and the next word's onset (start), being another way to emphasize an important word in context, or to signal phrase boundaries (Hawkins, 1971). In contrast to phrase boundaries, within the phrase speech tends to be 'connected' such that there are usually no pauses between words; i.e., most pauses are of zero seconds.

Prominence Prosodic prominence is a term that describes how salient a linguistic entity—in our case, a single word—is perceived relative to the words surrounding it in an utterance (Terken and Hermes, 2000). Unlike the previously described prosodic features, prominence is a higher-level percept in the sense that it is not elicited by a single acoustic dimension. The perception of prominence is affected by multiple acoustic features (Cole et al., 2010)—elongating a word's duration, increasing the speech energy or modulating the f_0 contour of a specific word can all make this word be perceived as more prominent in context. Although other acoustic features, like timbre, can also affect prominence, and factors like word frequency influence it as well (Cole et al., 2010), a combination of duration, loudness and pitch has been proposed as an effective acoustic measure to quantify prosodic prominence (Talman et al., 2019), which we use here.

3 Redundancy between Prosody and Linguistic Context

This paper concerns the time scale of the redundancy between prosody and linguistic context, where 'linguistic context' here refers to the segmental information of an utterance, represented in our experiments as text. In this section, we first explain how this redundancy can be formalized as a mutual information, following Wolf et al. (2023). We then expand on this framework by discussing how context-length manipulations allow us to investigate the time scale aspect of this redundancy.

3.1 Redundancy as Mutual Information

Let P_t be a prosody-valued random variable, which takes values $\mathrm{p}_t \in \mathbb{R}$. Further, let $\mathbf{W}$ be a words-valued random variable, which takes values $\mathbf{w} \in \mathcal{W}^*$, where $\mathcal{W}$ is a language's lexicon. We follow Wolf et al. (2023) in formalizing the redundancy between prosody and linguistic context as the mutual information: $\mathrm{MI}(\mathrm{P}_t; \mathbf{W})$. Under a few technical assumptions (e.g., the good mixed-pair assumption, see Wolf et al., 2023 for details), we can write this value as:

$$\mathrm{MI}(\mathrm{P}_t; \mathbf{W}) = \mathrm{H}(\mathrm{P}_t) - \mathrm{H}(\mathrm{P}_t \mid \mathbf{W}) \quad (1)$$

In this equation, **unconditional entropy** $\mathrm{H}(\mathrm{P}_t)$ serves as a baseline, representing how much uncertainty there is about P_t. In turn, **conditional entropy** $\mathrm{H}(\mathrm{P}_t \mid \mathbf{W})$ represents how much uncertainty remains about P_t once we know the context $\mathbf{W}$. Their difference then represents how much information $\mathbf{W}$ contains about P_t (and *vice versa*).

We are now left with the problem of estimating these entropies. While these values are unknown, we only require two things to estimate them: a corpus of prosodic values coupled to linguistic contexts, sampled from the ground-truth distribution,

$$\mathcal{D}_{\text{tst}} = \{\mathrm{p}'_t, \mathbf{w}'\}_{n=1}^{N}, \;\; \mathrm{p}'_t, \mathbf{w}' \sim p(\mathrm{p}_t, \mathbf{w})$$

and models $p_{\boldsymbol{\theta}}$ of distributions $p(\mathrm{p}_t)$ and $p(\mathrm{p}_t \mid \mathbf{w})$. We can then use a cross-entropy upper-bound (Pimentel et al., 2019) to estimate these entropies:

$$\begin{aligned} \mathrm{H}(\mathrm{P}_t \mid \mathbf{W}) &\leq \mathrm{H}_{\boldsymbol{\theta}}(\mathrm{P}_t \mid \mathbf{W}) \quad (2) \\ &\approx \frac{1}{|\mathcal{D}_{\text{tst}}|} \sum_{\mathrm{p}'_t, \mathbf{w}' \in \mathcal{D}_{\text{tst}}} \log \frac{1}{p_{\boldsymbol{\theta}}(\mathrm{p}'_t \mid \mathbf{w}')} \end{aligned}$$

where $p_{\boldsymbol{\theta}}(\mathrm{p}_t \mid \mathbf{w})$ is replaced with $p_{\boldsymbol{\theta}}(\mathrm{p}_t)$ when estimating $\mathrm{H}(\mathrm{P}_t)$. We describe our dataset $\mathcal{D}_{\text{tst}}$ and how to estimate $p_{\boldsymbol{\theta}}$ in Section §4.

3.2 Manipulating Context Length

To analyze the time scale of the redundancy between prosody and linguistic context, we will estimate $\mathrm{MI}(\mathrm{P}_t; \mathbf{W})$ while systematically manipulating the amount of context, i.e. the number of words, in $\mathbf{W}$. We thus quantify 'time' in units of words, as opposed to seconds, acknowledging the discrepancy between these concepts due to varying duration of words and speaking rates. To this end, we define $\mathbf{W}_{\overleftrightarrow{n,m}}$ as the linguistic context comprising n words before and m words after word W_t, including the word itself:

$$\mathbf{w}_{\overleftrightarrow{n,m}} = \langle \mathrm{w}_{t-n}, \cdots, \mathrm{w}_t, \cdots \mathrm{w}_{t+m} \rangle \quad (3)$$

Thus, for instance, $\mathbf{W}_{\overleftrightarrow{0,0}}$ corresponds to the word W_t by itself, and $\mathbf{W}_{\overleftrightarrow{3,6}}$ corresponds to the word

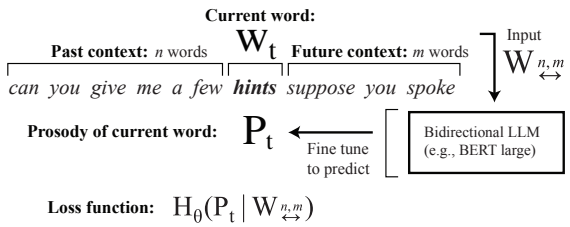

Figure 2: Estimation procedure for $\mathrm{H}(\mathrm{P}_t \mid \mathbf{W}_{\overleftrightarrow{n,m}})$. A span of words $\mathbf{W}_{\overleftrightarrow{n,m}}$ which includes word W_t is used as input to a model which predicts that word's prosody P_t. The loss function that the model minimizes estimates the conditional entropy.

W_t with 3 words in its past context and 6 words in its future context (see Fig. 2).

Given this definition, we can then explore the time scale we are interested in by estimating the mutual information $\mathrm{MI}(\mathrm{P}_t, \mathbf{W}_{\overleftrightarrow{n,m}})$ while varying n and m. This amounts to estimating $\mathrm{H}(\mathrm{P}_t)$ and $\mathrm{H}(\mathrm{P}_t \mid \mathbf{W}_{\overleftrightarrow{n,m}})$. The unconditional entropy does not depend on context; thus, to estimate it, we only need to compute a relatively simple prior over the domain of each prosodic feature. On the other hand, estimating the conditional entropy requires computing a family of conditional distributions for each prosodic feature, with one distribution for each n, m combination. In other words, we need a model $p_{\boldsymbol{\theta}}(\mathrm{p}_t \mid \mathbf{w}_{\overleftrightarrow{n,m}})$ that works for any n, m pair. We elaborate on this model in Section §4.2.

Importantly, the $\mathrm{MI}(\mathrm{P}_t; \mathbf{W}_{\overleftrightarrow{n,m}})$ is a monotonically increasing function of both m and n; larger linguistic contexts must contain at least as much (but maybe more) information about prosody than smaller contexts.[5] However, at some value of m and some value of n, the MI might reach a plateau. We consider the "time scale" of the redundancy between prosody and linguistic context to be the value from which increasing context does not significantly increase the MI.

4 Methods

We now detail our dataset and modeling choices.

4.1 Dataset

Our data-extraction process follows Wolf et al.'s (2023) proposed and publicly available pipeline (see their paper for more details).

[5]We note that—while the MI is monotonically increasing in theory—it is not necessarily the case that this underlying monotonicity will be reflected by our estimation methods.

Speech Data We use the LibriTTS spoken language corpus [6] (Zen et al., 2019),[7] which contains public domain audiobook materials (audio and text) recorded by volunteer narrators. This dataset contains 585 hours of English speech data at a 24kHz sampling rate, and includes recordings from 2,456 speakers reading aloud books which are paired with the corresponding transcripts. We filtered out texts from LibriTTS that contained less than three words (such as book and chapter titles) and we eliminated punctuation marks, since these can be very informative regarding prosody, and are not explicitly present in spoken communication.

Prosody Feature Extraction The procedure for extracting prosodic features starts with aligning the audio and text using Montreal Forced Aligner (MFA) (McAuliffe et al., 2017). Given this alignment, both the duration and pause of each word can be easily computed from the words' offset and onset times.[8] Duration was then normalized relative to the number of syllables in the word, to reflect duration per syllable. To extract pitch, loudness, and prominence, we rely on algorithms provided by Suni et al. (2017). These algorithms return the pitch curve, or fundamental frequency (f_0), of a speech signal, which is z-scored per speaker, to remove inter-speaker differences. For each word, we focused on the pitch curve in an interval up to $250ms$ (or across the word's duration, if shorter than $250ms$) around the word's primary syllable (identified using CELEX (Baayen et al., 1996)). These curves were then averaged, resulting in a single mean pitch value per word. Suni et al.'s (2017) algorithms also return continuous energy curves, which we average per word. For prominence, we use a mean value per word which was released and validated by (Talman et al., 2019) for this dataset. In short, this prominence value reflects the steepest time-frequency variation in a signal combining duration, energy and f_0. Apart from the absolute average prominence value per each word, we also computed their relative prominence, subtracting the mean prominence value of the three preceding words from the current word's value; this emphasizes local changes in prominence.

[6]The LibriTTS corpus is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0)

[7]This dataset is derived from LibriSpeech audiobooks corpus (Panayotov et al., 2015), which is itself derived from LibriVox (Kearns, 2014).

[8]We note that about 89.4% of the words in our dataset have a pause of 0 seconds.

Splitting the Data to Train, Validation and Test The dataset was divided into separate train, validation and test sets, using splits of the dataset provided by Talman et al. (2019). For training, we used a data split (termed train-360) containing 904 speakers, 11,262 sentences and 2,076,289 words. For validation, we used a data split (termed dev) containing 40 speakers, 5,726 sentences and 99,200 words. We had access to a data split (termed test) containing 39 speakers, 4,821 sentences and 90,063 words, as well as to another data split (termed train-100) containing 247 speakers, 33,041 sentences and 570,592 words. We therefore used the latter split (train-100) as the test set, except for prominence (absolute and relative) where we added this split to the test split, which led to more stable results.

4.2 Estimating the Cross-Entropies

We now explain how we model the probability distributions $p_{\boldsymbol{\theta}}(\mathrm{p}_t)$ and $p_{\boldsymbol{\theta}}(\mathrm{p}_t \mid \mathbf{w}_{\overleftrightarrow{n,m}})$, which serve to estimate the unconditional and conditional cross-entropies, respectively.

Modeling $p_{\boldsymbol{\theta}}(\mathrm{p}_t)$ To model this unconditional distribution over prosodic values, we simply follow Wolf et al. (2023) in using a Gaussian kernel density estimator (KDE). Given a training set $\mathcal{D}_{\text{trn}}$, sampled from $p(\mathrm{p}_t, \mathbf{w})$, this model is defined as:

$$p_{\boldsymbol{\theta}}(\mathrm{p}_t) = \frac{1}{|\mathcal{D}_{\text{trn}}|} \sum_{\mathrm{p}'_t \in \mathcal{D}_{\text{trn}}} \mathcal{N}(\mathrm{p}_t; \mu = \mathrm{p}'_t, \sigma = \widehat{\sigma}) \quad (4)$$

where $\mathcal{N}$ are Gaussian distributions, each centered at a prosodic value $\mu = \mathrm{p}'_t$ and all having the same variance $\sigma = \widehat{\sigma}$. We choose $\widehat{\sigma}$ that achieves the highest likelihood on our validation set.

Modeling $p_{\boldsymbol{\theta}}(\mathrm{p}_t \mid \mathbf{w}_{\overleftrightarrow{n,m}})$ To model this conditional distribution, we again follow Wolf et al. in finetuning a language model (LM), with an added linear layer on top, to predict the parameters of a conditional distribution over p_t. Unlike Wolf et al., however, we limit our models input to include only $\mathbf{w}_{\overleftrightarrow{n,m}}$ instead of the entire $\mathbf{w}$. We assume the conditional distribution over prosody to follow a parametric distribution $\mathcal{Z}$, and use a LM to predict this distribution's parameters $\widehat{\boldsymbol{\phi}}$:[9]

$$\widehat{\boldsymbol{\phi}} = \mathrm{LM}(\mathbf{w}_{\overleftrightarrow{n,m}}) \quad (5)$$

$$p_{\boldsymbol{\theta}}(\mathrm{p}_t \mid \mathbf{w}_{\overleftrightarrow{n,m}}) = \mathcal{Z}(\mathrm{p}_t; \boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}) \quad (6)$$

We finetune this model by minimizing its cross-entropy on a training set $\mathcal{D}_{\text{trn}}$; which amounts to minimizing the right-hand side of Eq. (2). As the cross-entropy is an upper bound on the entropy, the lower its value (and consequently the better our model) the tighter the estimate we get for the entropy. Notably, we estimate all n, m combinations using a single finetuned LM. During training, we sampled inputs of varying lengths, spanning 1 to 10 words.[10] For each sample, the model then predicts the prosody of each of the words in this span in parallel; in a 7-word span, thus, the first word's prosody is predicted as $p_{\boldsymbol{\theta}}(\mathrm{p}_t \mid \mathbf{w}_{\overleftrightarrow{0,6}})$ and the 5th word's prosody is predicted as $p_{\boldsymbol{\theta}}(\mathrm{p}_t \mid \mathbf{w}_{\overleftrightarrow{4,2}})$.[11]

5 Results

Fig. 1 displays the average mutual information (MI) across the 6 prosodic features tested here (pitch, ludness, duration, absolute prominence, relative prominence, pause), and Fig. 3 displays those results for each feature separately. See App. A for the values of unconditional entropies that comprise those MIs. Our results' overall trend confirms Hypothesis 1, the long-scale past redundancy hypothesis. When averaged across all linguistic features, the mutual information between prosody and past linguistic context (first column in Fig. 1) increases as a function of the number of words available, up to about 5-8 words, after which it plateaus. Our results also support Hypothesis 2, the short-scale future redundancy hypothesis. When averaged across all linguistic features, the mutual information between prosody and future linguistic context (first row in Fig. 1) increases only up to about one or two words and then plateaus. Furthermore, the MI with the past is higher than the MI with the future, a trend that becomes larger for longer spans of words.

When examining the mutual information with linguistic contexts containing both past and future words (i.e., $n > 0$ and $m > 0$; the n-th column and m-th row in Fig. 1), we observed interesting interactions. Specifically, a combination of about one word in the future and about 5-8 words in the

[9] We evaluate models with Gaussian, Gamma and Laplace distributions, choosing the distribution that leads to the lowest cross-entropy on a validation set. Parameters $\boldsymbol{\phi}$ are, e.g., the mean and standard deviation for a Gaussian.

[10] These inputs were obtained by first sampling an item from the dataset, and then randomly cutting it into the desired length (between 1-10 words).

[11] For each prosodic feature, we tested several models, namely BERT, BERT-large and RoBERTa-large, and selected the model that gave the best results. In most cases, this was BERT-large, except for pauses and pitch where it was BERT. An early stopping criterion was applied, such that if the loss did not decrease for 3 epochs the model stopped training.

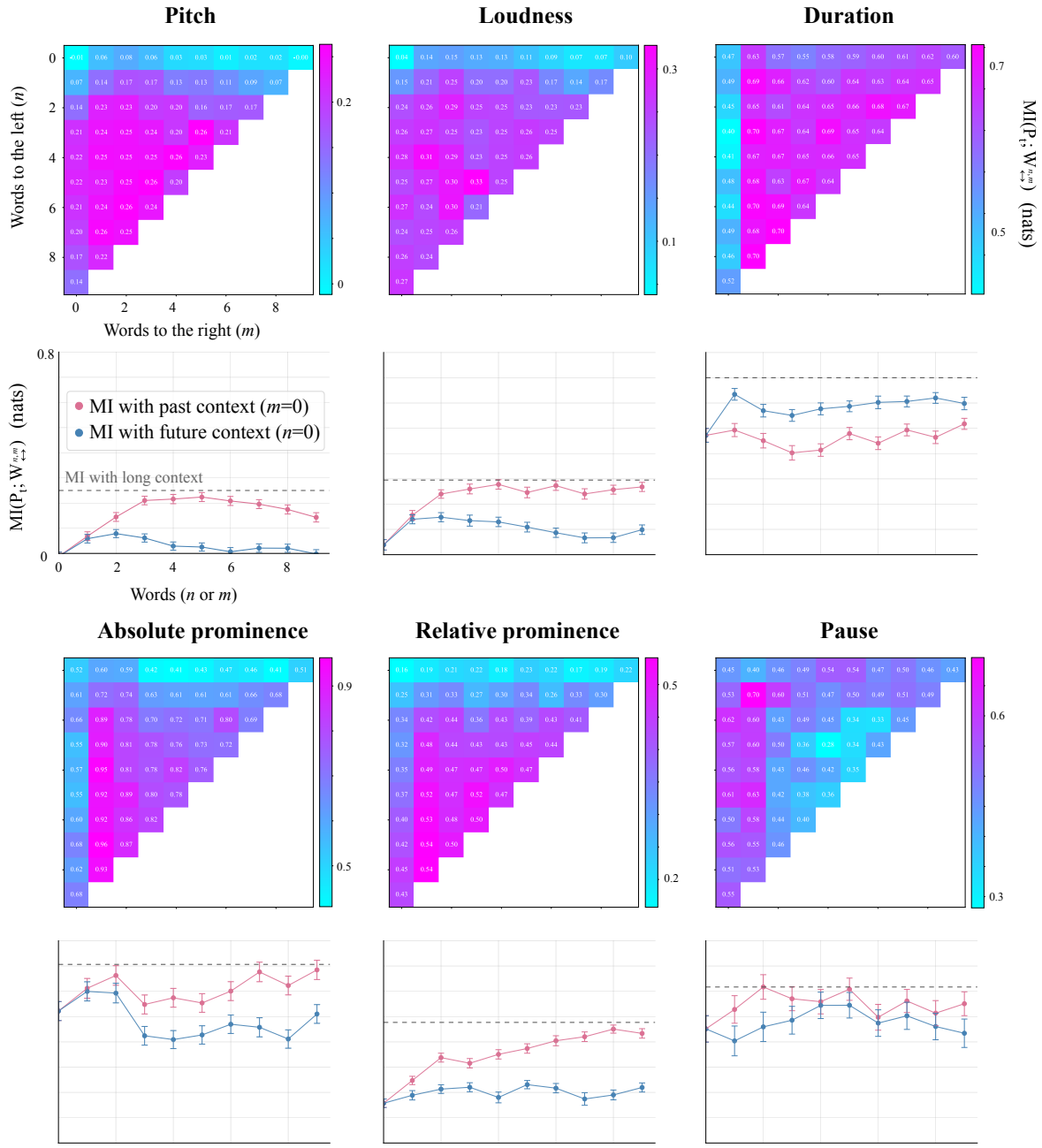

Figure 3: For each of the 6 tested prosodic features, two plots are presented. The upper plots are similar to Fig. 1, and display the redundancy, quantified as mutual information between a prosodic feature at a given word p_t and the linguistic context $w_{\overleftrightarrow{n,m}}$, which includes the word itself, n words before, and m words after it. The lower plots display just the first column (corresponding to linguistic context which includes the word itself and a gradually increasing number of past words, red curve) and the first row (corresponding to lingusitic context which includes the word itself and a gradually increasing number of past words, blue curve) in the upper plots. Wherever numbers and units are not displayed, they correspond to the units displayed for Absolute Prominence. The MI values correspond to the mean across all train data. Error bars correspond to standard errors of the mean. Dashed horizontal line corresponds to the MI with the longest available context. See App. B for the distributions of these features from our dataset.

past already contains most of the information about prosodic features. Notably, this n, m combination led to higher mutual information than even other combinations with larger context (i.e., with n', m', $n' \geq n$, and $m' \geq m$). While this is theoretically impossible (adding more context can never decrease mutual information), this is likely due to our models' training procedure not being able to ignore unhelpful contributions of long-scale contexts.

Interestingly, each individual prosodic feature shows a slightly different pattern (see Fig. 3). The long-scale past hypothesis (Hypothesis 1) is supported for most prosodic features individually, but not for duration and pauses; for prominence, pitch and loudness, the MI with past context (red curves in the lower plot per each feature in Fig. 3) increases up to at least 3 words.[12] For pauses though, the past MI saturates after only two words, and for duration, the MI with past alone does not seem to significantly rise above the $0, 0$ point, indicating the past context does not add information beyond the word identity itself. However, for all prosodic features, even duration, when looking at all n, m combinations, redundancy is largest around 1-2 words in the future and 3-8 words in the past—thus supporting Hypothesis 1.

The short-scale future hypothesis (Hypothesis 2) is also supported for most individual features, except for duration and pause. For prominence, pitch and loudness, the time scale of the future MI is shorter than the scale of the past MI, saturating somewhere between 1 and 3 words (blue curves in the lower plot per each feature in Fig. 3). For pauses though, the future MI saturates after about 4 words, which is longer than for the past—although these MI estimates seem noisy. For duration, it saturates after 1 word, which is more than the past since the past curve shows no increasing trend. Notably however, the MI with the past is larger than with the future in all features except for duration, and, even for duration, $\mathrm{MI}(\mathrm{P}_t, \mathbf{W}_{\overleftrightarrow{n,1}})$ is larger than $\mathrm{MI}(\mathrm{P}_t, \mathbf{W}_{\overleftrightarrow{1,m}})$. We thus conclude that all these individual feature results' support Hypothesis 2.

6 Discussion

This work aimed to estimate the time scale of the redundancy between prosodic and linguistic information. We built on an existing published pipeline (Wolf et al., 2023), but extended it to include a systematic modulation of context length, from 1 to 10 words. We will make our extension of this pipeline publicly available for researchers wishing to extend our analysis, potentially exploring the time scale of redundancy between other communication channels of interest.

Overall, we confirm our two hypotheses (see Section §1): For most prosodic features we tested (apart from pause and duration, see §6.4), redundancy with past linguistic context unfolds across a long time scale (of about 3-8 words), while redundancy with future words is shorter-scale (concentrated on 1-2 future words). We next discuss the implications of these results with respect to previous literature.

6.1 The Time Scale of Prosodic Information

Sentence comprehension is constrained by cognitive demands such as attention and working memory, leading listeners to maintain a lossy representation of past linguistic context (Gibson, 1998; Vasishth et al., 2010; Futrell et al., 2020; Kuribayashi et al., 2022). While the precise number of words maintained in working memory probably depends on many factors, estimates suggest a range of about 3-5 words (Cowan, 2010). In the brain, language-selective neural populations integrate linguistic information across distinct time scales; these scales are quantified to span between 1-6 words (Jain and Huth 2018; Regev et al. 2024).

Despite the well-studied dynamics of linguistic processing, the time scale at which *prosody* interacts with linguistic information had been previously underexplored. Intonation units, a meaningful organizational unit of prosodic information, follow a rhythmic structure of about 1 Hz (Inbar et al., 2020); meaning that each unit is about 1 second long and therefore contains about 3-4 words, given an average speech rate of about 200 words per minute (Yuan et al., 2006). This, together with the span of working memory (mentioned above), suggests a natural alignment between prosodic structure and the cognitive constraints of linguistic processing.

Here, we provide a quantification of the time scale at which redundancy between prosody and linguistic context operates. Our findings suggest that this redundancy spans around 3-8 past words, a scale comparable to both linguistic working memory limitations and prosodic segmentation. These results thus highlight a possible role of prosody in optimizing comprehension.

[12] For absolute prominence, the curve is a little noisy but saturates at around 7 words.

6.2 Prosody as an Audience-Design Tool in Communication

A longstanding debate in linguistics concerns the extent to which language production is shaped by audience design, with speakers actively tailoring their utterances to facilitate listener comprehension. Evidence suggests that syntactic choices are not strongly adapted for listener needs but rather reflect the speaker's own constraints (Ferreira, 2008; Morgan and Ferreira, 2022). This apparent lack of audience design in syntax may stem from the rigid structural constraints imposed by linguistic systems. In contrast, prosody may offer greater flexibility, allowing speakers to dynamically modulate pitch, loudness, and rhythm in real time. This flexibility suggests that prosody may play a larger role in audience design (Clark et al., 2025), serving as an additional communicative channel that enhances intelligibility. In fact, prior work shows a trade-off between a word's duration and its information content (Jurafsky et al., 1998; Bell et al., 2009; Coupé et al., 2019; Pimentel et al., 2021), a trade-off which is typically interpreted as arising to facilitate listeners comprehension by smoothing the amount of information they receive per second (known as the uniform information hypothesis; Fenk and Fenk, 1980; Genzel and Charniak, 2002; Levy and Jaeger, 2007). Our findings support this audience-design view by revealing that the redundancy between prosody and past linguistic context extends over long time scales, suggesting that prosody may serve as a "reminder", helping listeners access information from the long-scale past.

6.3 The Relationship Between Prosody and Future Words

Our findings indicate that redundancy between prosody and linguistic information is weaker for future words than for past words. However, prosody still exhibits a strong relationship with the immediately upcoming word or two. This short-range relationship could stem from motor constraints on prosody production, or from local linguistic dependencies like fixed expressions. Another possibility is that prosodic planning occurs at the level of entire sentences and therefore observed redundancy with the next word reflects broader contour structuring. Finally, prosody may actively signal upcoming words through cues such as duration, pauses, or pitch changes, aiding listener expectations. Future work should explore these potential mechanisms to better understand prosody's role in forward-looking processing. Notably, prior research has shown that a word's duration correlates with its predictability given future context (Bell et al., 2009). Further, even reading times—a setting in which the subject is assumed to not know what future words are—correlate with features of future words (such as frequency, predictability and entropy; Roark et al., 2009; Angele et al., 2015; van Schijndel and Schuler, 2017).

6.4 Pause and Duration Do not Follow Hypotheses 1 and 2

Compared to the other prosodic features, pause and duration stand out in their relatively short time scale of redundancy with both past and future words, as well as a relatively strong redundancy with future words. This may suggest that pausing after a word and elongating it, are mainly served in order to prepare for the next upcoming word and facilitate its processing by slowing down the rhythm of speech. Alternatively, as discussed above, pauses are most common at the end of sentences. Similarly, sentence-final words are also typically elongated (Seifart et al., 2021; Paschen et al., 2022). The high predictability of both these features with future context could be partially due to our models predicting a sentence-final vs sentence-middle distinction, which could itself be used to roughly predict the value of these prosodic features.

7 Conclusion

Our findings reveal a fundamental asymmetry in the time scale of redundancy between prosody and linguistic context: while prosody exhibits redundancy with both past and future words, this relationship extends across a longer span for past words (3–8 words) than for future words (1–2 words). This suggests that prosody's relationship with future words primarily reflects short-term effects such as next-word prediction, local word dependencies, or other production factors—future work should try to distinguish those explanations. In contrast, its relationship with past words operates over a broader scale, potentially serving to reinforce or highlight information that may be cognitively demanding for listeners to remember in real-time communication. These results provide new insights into the role of prosody in spoken language.

Limitations

Our study has several limitations that should be considered when interpreting the results.

Data-related Limitations. The first set of limitations relates to the dataset used. Our dataset consists of audiobooks, which do not necessarily reflect natural prosody in real time communication, potentially affecting the generalizability of our findings. Redundancy may be higher in audiobooks than in spontaneous speech, because the text is written with the assumption that it must convey all necessary information without relying on prosody. We address this concern to some extent by removing punctuation marks, which serve as a substitute for prosody in written text. Another dataset-related limitation is the sample size. Larger datasets may be required for more stable estimations, especially given that we compute 55 different values (for $\mathbf{w}_{\overleftrightarrow{n,m}}$, n and m from 1 to 10), each based on different subsets of the data, effectively reducing the number of samples available for each estimate. This sparsity likely contributes to some of the observed noise in our results.

Estimation-related Limitations. The second set of limitations has to do with the estimation procedure. The mutual information we compute approximates the true value, and is constrained by the quality of the models we use $p_{\boldsymbol{\theta}}(\mathrm{p}_t \mid \mathbf{w}_{\overleftrightarrow{n,m}})$. One of our modeling assumptions is the functional form of the conditional distribution of prosody given a linguistic context (namely, Gaussian, Gamma or Laplace distributions depending on the prosodic feature). However, this parametric assumption may limit the model's performance and future work should explore alternative conditional distributions which may improve results. This assumption is particularly violated for features that manifest different distributions; pause, for instance, takes a 0 value in 89.4% of the data and may therefore be better modeled by a zero-inflated distribution. Indeed, our results for pause seem particularly noisy. Additionally, to estimate $\mathbf{w}_{\overleftrightarrow{n,m}}$, we provided the models with short segments of 1 to 10 words. However, these large language models were not pretrained on such short segments but rather on longer spans of text, which might have impacted their efficiency in extracting the information from short segments. While finetuning likely helped mitigate this issue, it remains a potential limitation. Furthermore, we train a single model for all combinations of n, m, which does not guarantee that the value is optimal for each combination separately. Finally, we observed cases where the mutual information decayed for longer contexts, which contradicts expectations from information theory, as additional context can never reduce information. This phenomenon likely stems from issues in training the models, which could be biased toward under-utilizing the available context for longer spans. Future work should address these limitations to refine our understanding of redundancy between prosody and linguistic information.

Ethics Statement

We foresee no potential ethical concerns or risks associated with this study, although we acknowledge the inherent risks in using any AI system.

References

Bernhard Angele, Elizabeth R. Schotter, Timothy J. Slattery, Tara L. Tenenbaum, Klinton Bicknell, and Keith Rayner. 2015. Do successor effects in reading reflect lexical parafoveal processing? Evidence from corpus-based and experimental eye movement data. *Journal of Memory and Language*, 79-80:76–96.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. The celex lexical database (cd-rom).

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111.

Mara Breen, Evelina Fedorenko, Michael Wagner, and Edward Gibson. 2010. Acoustic correlates of information structure. *Language and cognitive processes*, 25(7-9):1044–1098.

Herbert H Clark and Gregory L Murphy. 1982. Audience design in meaning and reference. In *Advances in psychology*, volume 9, pages 287–299. Elsevier.

Thomas H Clark, Moshe Poliak, Tamar I Regev, AJ Haskins, Edward Gibson, and Caroline Robertson. 2025. The relationship between surprisal, prosody, and backchannels in conversation reflects intelligibility-oriented pressures.

Jennifer Cole. 2015. Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1-2):1–31.

Jennifer Cole, Yoonsook Mo, and Mark Hasegawa-Johnson. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2):425–452.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

Nelson Cowan. 2010. The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science*, 19(1):51–57.

August Fenk and Gertraud Fenk. 1980. Konstanz im Kurzzeitgedächtnis - Konstanz im sprachlichen Informationsfluß? *Zeitschrift für Experimentelle und Angewandte Psychologie*, 27(3):400–414.

Victor S Ferreira. 2008. Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and motivation*, 49:209–246.

Richard Futrell, Edward Gibson, and Roger P Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3):e12814.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Peter R Hawkins. 1971. The syntactic location of hesitation pauses. *Language and Speech*, 14(3):277–288.

Maya Inbar, Eitan Grossman, and Ayelet N Landau. 2020. Sequences of intonation units form a˜ 1 hz rhythm. *Scientific reports*, 10(1):15846.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31.

Sun-Ah Jun. 2006. *Prosodic typology: The phonology of intonation and phrasing*, volume 1. Oxford University Press.

Daniel Jurafsky, Alan Bell, Eric Fosler-Lussier, Cynthia Girand, and William D Raymond. 1998. Reduction of english function words in switchboard. In *ICSLP*.

Jodi Kearns. 2014. Librivox: Free public domain audiobooks. *Reference Reviews*, 28(1):7–8.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Roger P. Levy and Tim Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, pages 849–856.

Richard L Lewis, Shravan Vasishth, and Julie A Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, 10(10):447–454.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Adam M Morgan and Victor S Ferreira. 2022. Still no evidence for audience design in syntax: Resumptive pronouns are not the exception. *Journal of Memory and Language*, 127:104368.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Ludger Paschen, Susanne Fuchs, and Frank Seifart. 2022. Final lengthening and vowel length in 25 languages. *Journal of Phonetics*, 94:101179.

Janet Breckenridge Pierrehumbert. 1980. *The phonology and phonetics of English intonation*. PhD Thesis, Massachusetts Institute of Technology.

Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tamar I Regev, Colton Casto, Eghbal A Hosseini, Markus Adamek, Anthony L Ritaccio, Jon T Willie, Peter Brunner, and Evelina Fedorenko. 2024. Neural populations in the language network differ in the size of their temporal receptive windows. *Nature Human Behaviour*, 8(10):1924–1942. Equal first authors: Regev and Casto.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Frank Seifart, Jan Strunk, Swintha Danielsen, Iren Hartmann, Brigitte Pakendorf, Søren Wichmann, Alena Witzlack-Makarevich, Nikolaus P Himmelmann, and Balthasar Bickel. 2021. The extent and degree of utterance-final word lengthening in spontaneous speech from 10 languages. *Linguistics Vanguard*, 7(1).

Scott Seyfarth. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.

Kim Silverman, Mary Beckman, John Pitrelli, Mori Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *Second international conference on spoken language processing*.

Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. 2019. Predicting prosodic prominence from text with pretrained contextualized word representations. *arXiv preprint arXiv:1908.02262*.

Jacques Terken and Dik Hermes. 2000. The perception of prosodic prominence. In *Prosody: Theory and experiment: Studies presented to Gösta Bruce*, pages 89–127. Springer.

Marten van Schijndel and William Schuler. 2017. Approximations of predictive entropy correlate with reading times. In *Proceedings of the Cognitive Science Society*, pages 1260–1265.

Shravan Vasishth, Katja Suckow, Richard L Lewis, and Sabine Kern. 2010. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4):533–567.

Michael Wagner and Duane G Watson. 2010. Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. 2023. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.

Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2006. Towards an integrated understanding of speaking rate in conversation. In *Ninth International Conference on Spoken Language Processing*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

A Unconditional Entropies

Prosodic Feature	Unconditional Entropy
Absolute Prominence	0.536
Relative Prominence	1.355
Energy	0.815
Duration	-0.920
Pause	-5.193
f_0	3.469

Table 1: Unconditional entropies of each prosodic feature.

B Prosodic Features' Histograms

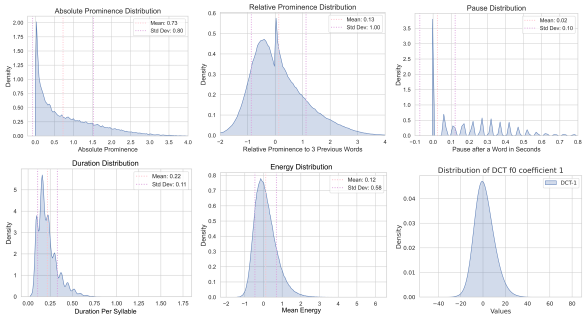

Figure 4: Histogram of prosodic features. (top-left) Absolute prominence; (top-center) relative prominence; (top-right) pause; (bottom-left) duration; (bottom-center) energy; (bottom-right) pitch