Quantifying the redundancy between prosody and text

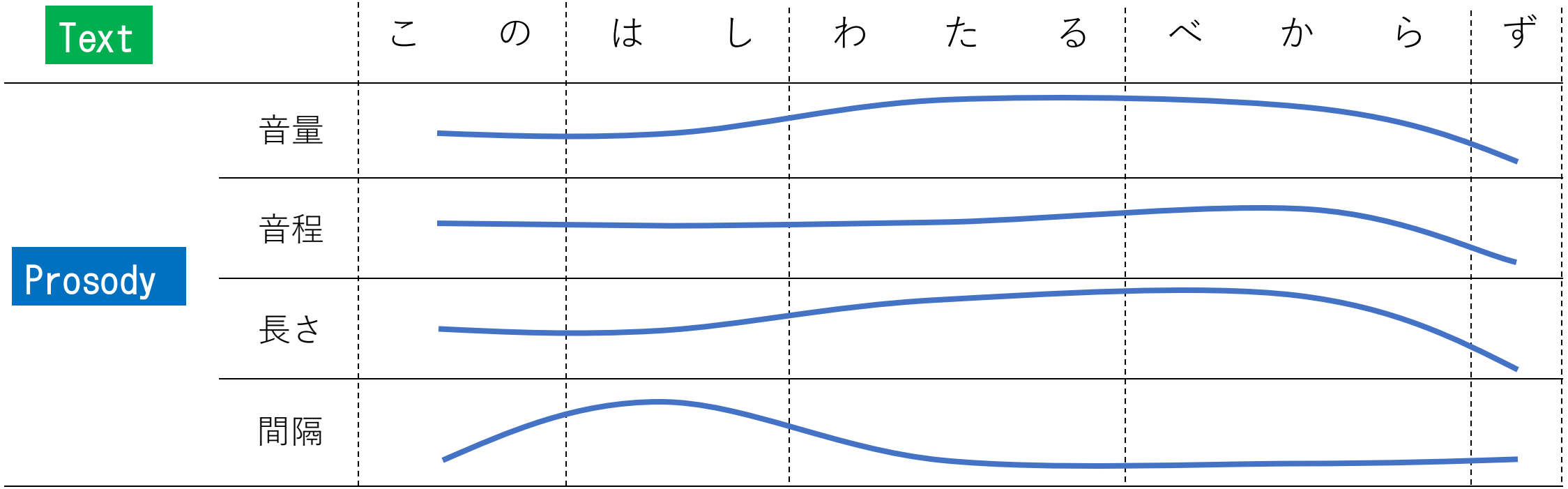
Lukas Wolf[ʕ] **Tiago Pimentel**[ð,ʕ] **Evelina Fedorenko**[ɱ] **Ryan Cotterell**[ʕ]
Alex Warstadt[ʕ] **Ethan Gotlieb Wilcox**[ʕ] **Tamar I. Regev**[ɱ]
[ʕ]ETH Zürich [ɱ]MIT [ð]University of Cambridge
{wolflu, ryan.cotterell, warstadt, ethan.wilcox}@ethz.ch
tp472@cam.ac.uk {evelina9, tamarr}@mit.edu

(EMNLP2023 main paper)

読み手：東京大学宮尾研究室 学術専門職員 神藤駿介

研究概要

自然言語の Text と Prosody（韻律）が内包する情報の冗長度を定量化する



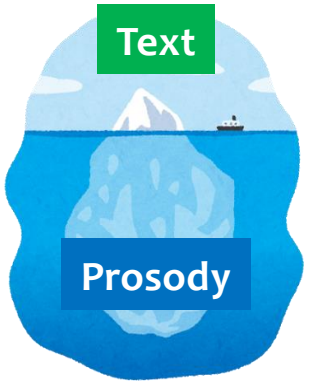
※単語分割はテキトーです

Text と Prosody の相互情報量によって定量化
自己回帰LM、双方向LMを活用することで文脈の影響も調査

モチベーション

- 「韻律の種類によっては冗長性がある」とする先行研究［1, 2, 3］
  - Pitch や Duration は surprisal（$-\log p(\mathrm{w}|context)$）と相関する
- 一方で、冗長性が小さい（＝韻律に固有な情報がある）ケースも考えられる
  - 皮肉の表明，統語的曖昧性の解消［4］，疑問形への変換，…
  - Pitch や Duration 以外の韻律的特徴との関連性は？
- **本研究：どんな韻律的特徴がどれくらい冗長なのかを網羅的・定量的に調査**
  - 冗長であるならば、音声にそんなに気を払わずにNLPしても良いのでは？
  - 逆に冗長でないとすれば、Text だけでNLPするのでは不十分なのでは？

[1] Matthew Aylett and Alice Turk. 2006. “Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei.” The Journal of the Acoustical Society of America.
[2] Scott Seyfarth. 2014. ”Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. “Cognition.
[3] Kevin Tang and Jason A. Shaw. 2021. “Prosody leaks into the memories of words.” Cognition.
[4] Trang Tran et al. 2018. “Parsing Speech: A Neural Approach to Integrating Lexical and Acoustic-Prosodic Information.” Proceedings of NAACL-HLT.



Text情報は自然言語の
氷山の一角かもしれない…

冗長度の定量化

事前準備（notation）

- Σ：alphabet の集合
- w $(\in \Sigma^*)$：text
- W：text の確率変数
- p_t $(\in \mathbb{R}^d)$：ある単語が発せられた時刻 t における prosody
- P_t：prosody の確率変数

冗長度の定量化

Text と Prosody の 相互情報量 によって両者の冗長性を定量化

$$\mathrm{MI}(\mathbf{P}_t;\mathbf{W}) = \sum_{\mathbf{w}\in\Sigma^*}\int_{\mathbb{R}^d} p(\mathbf{p}_t,\mathbf{w})\log\frac{p(\mathbf{p}_t,\mathbf{w})}{p(\mathbf{p}_t)\,p(\mathbf{w})}\mathrm{d}\mathbf{p}_t \quad (1)$$

エントロピーの差に変形し見通しを良くする

$$\mathrm{MI}(\mathbf{P}_t;\mathbf{W}) = \mathrm{H}(\mathbf{P}_t) - \mathrm{H}(\mathbf{P}_t \mid \mathbf{W}) \quad (2)$$

⚠ 実は離散分布と連続分布との間（mixed-pair）のMIにおいては一般に成立しない

以下の仮定を認めると成立する

Good mixed-pair assumption [1]

$p(p_t|w)$ は、全ての $w \in \Sigma^*$ において $p(p_t)$ に関して絶対連続

[1] Aleksandr Beknazaryan et al. 2019. On mutual information estimation for mixed- pair random variables. Statistics & Probability Letters.

冗長度の定量化

Text と Prosody の 相互情報量 によって両者の冗長性を定量化

$$\mathrm{MI}(\mathbf{P}_t; \mathbf{W}) = \mathrm{H}(\mathbf{P}_t) - \mathrm{H}(\mathbf{P}_t \mid \mathbf{W}) \quad (3a)$$
$$\approx \mathrm{H}_\theta(\mathbf{P}_t) - \mathrm{H}_\theta(\mathbf{P}_t \mid \mathbf{W}) \quad (3b)$$

クロスエントロピーで近似［1］

$$H_\theta(X) = -\sum p(x) \log p_\theta(x)$$

$p(x)$ … 真の分布
$p_\theta(x)$ … 推定分布

$$\mathrm{H}_\theta(\mathbf{P}_t) \approx \frac{1}{N} \sum_{n=1}^{N} \log \frac{1}{p_\theta(\mathbf{p}_t^{(n)})} \quad (4a)$$

$$\mathrm{H}_\theta(\mathbf{P}_t \mid \mathbf{W}) \approx \frac{1}{N} \sum_{n=1}^{N} \log \frac{1}{p_\theta(\mathbf{p}_t^{(n)} \mid \mathbf{w}^{(n)})} \quad (4b)$$

クロスエントロピーをモンテカルロ近似

$n \to \infty$ で一致

※近似の理論的背景は[1]を参照（神藤は勉強中…）

[1] David McAllester and Karl Stratos. “Formal Limitations on the Measurement of Mutual Information”. PMLR 2020.

冗長度の定量化

Text と Prosody の 相互情報量 によって両者の冗長性を定量化

$$\mathrm{MI}(\mathbf{P}_t; \mathbf{W}) = \mathrm{H}(\mathbf{P}_t) - \mathrm{H}(\mathbf{P}_t \mid \mathbf{W}) \quad (3a)$$
$$\approx \mathrm{H}_\theta(\mathbf{P}_t) - \mathrm{H}_\theta(\mathbf{P}_t \mid \mathbf{W}) \quad (3b)$$

クロスエントロピーで近似［1］

$$H_\theta(X) = -\sum p(x) \log p_\theta(x)$$

$p(x)$ … 真の分布
$p_\theta(x)$ … 推定分布

$$\mathrm{H}_\theta(\mathbf{P}_t) \approx \frac{1}{N} \sum_{n=1}^{N} \log \frac{1}{p_\theta(\mathbf{p}_t^{(n)})} \quad (4a)$$

$$\mathrm{H}_\theta(\mathbf{P}_t \mid \mathbf{W}) \approx \frac{1}{N} \sum_{n=1}^{N} \log \frac{1}{p_\theta(\mathbf{p}_t^{(n)} \mid \mathbf{w}^{(n)})} \quad (4b)$$

クロスエントロピーをモンテカルロ近似

$n \to \infty$ で一致

※近似の理論的背景は[1]を参照（神藤は勉強中…）

$p_\theta(\mathbf{p}_t)$ と $p_\theta(\mathbf{p}_t \mid \mathbf{w})$ を推定すれば相互情報量（の近似値）を計算できる！

[1] David McAllester and Karl Stratos. “Formal Limitations on the Measurement of Mutual Information”. PMLR 2020.

確率分布の推定

Prosody の確率分布：Gaussian Kernel によるカーネル密度推定

$$p_{\boldsymbol{\theta}}(\mathbf{p}_t) = \frac{1}{N_{\mathrm{trn}}\,\mathrm{h}} \sum_{n=1}^{N_{\mathrm{trn}}} \mathrm{K}\left(\mathbf{p}_t, \mathbf{p}_t^{(n)}, \Sigma_{\mathcal{D}_{\mathrm{trn}}}, \mathrm{h}\right)$$

$$\mathcal{D}_{\mathrm{trn}} = \{(\mathbf{p}_t^{(n)}, \mathbf{w}^{(n)})\}_{n=1}^{N_{\mathrm{trn}}} \sim p(\mathbf{p}_t, \mathbf{w})$$

Prosody の条件付確率分布：

$$p_{\boldsymbol{\theta}}(\mathbf{p}_t \mid \mathbf{w}) = \mathcal{Z}(\mathbf{p}_t; \boldsymbol{\phi}) \qquad (5b)$$

- パラメータ ϕ をもつ確率分布 $\mathcal{Z}$
- $\mathcal{Z}$ ： ガウス分布 or ガンマ分布
  - Prosody の種類によって変える

$$\boldsymbol{\phi} = \mathrm{LM}_{\boldsymbol{\theta}}(\mathbf{w}) \qquad (5a)$$

- パラメータ ϕ はLMを用いて推定
  - uncontextualized/contextualized LM で比較 → 文脈の影響を調査（後述）

実験設定：Prosody の特徴量

F0 Contours が8次元実数ベクトル・他は1次元の実数

- Energy（音量）：　バンドパスフィルタ → 振幅の対数
- Duration：　各単語が発話されている時間
- Pause：　単語と単語の間の無音時間
- F0 Contours（音程）：　平均だと粗い → *離散コサイン変換の係数で表現
- Prominence（強調）：　Energy, duration, F0 を組み合わせた特徴量
- Relative Prominence：　過去３単語の平均からの相対的な変化
  - 発話全体の中でどのくらい強調されているかが重要 → 相対的な変化が重要

*離散コサイン変換(DCT):系列をコサイン波の線形和で表現。各係数をその系列の特徴量とみなす。

実験設定：Text の表現

$$\phi = \mathrm{LM}_{\theta}(\mathbf{w}) \tag{5a}$$
$$p_{\theta}(\mathbf{p}_t \mid \mathbf{w}) = \mathcal{Z}(\mathbf{p}_t; \phi) \tag{5b}$$

- パラメータ ϕ はLMを用いて推定
  - uncontextualized/contextualized LM で比較 → 文脈の影響を調査

Non-Contextual Estimator (Current word)

- fasttext で埋め込み → MLP でパラメータ ϕ を推定

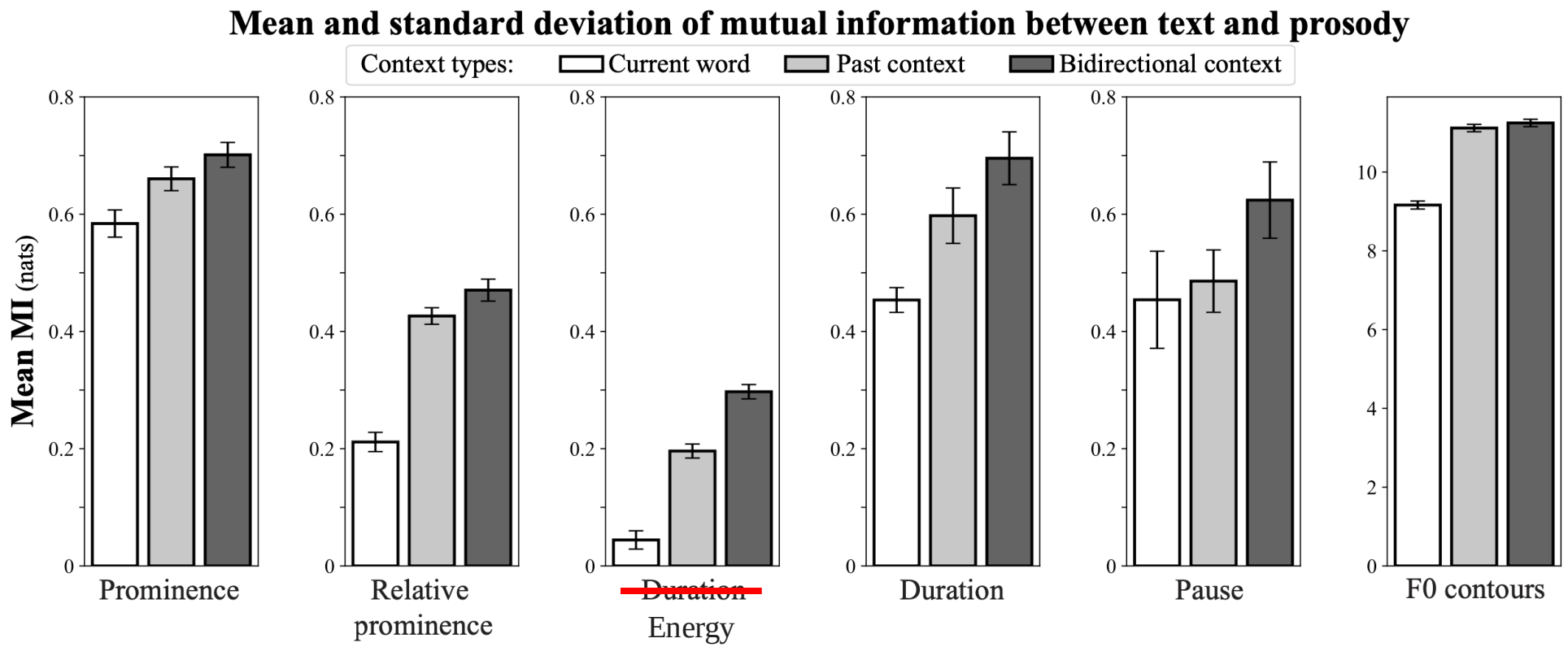
Contextual Estimator (Past context)

- 自己回帰モデル（GPT-2）+ 線形層 で fine-tuning してパラメータ ϕ を推定

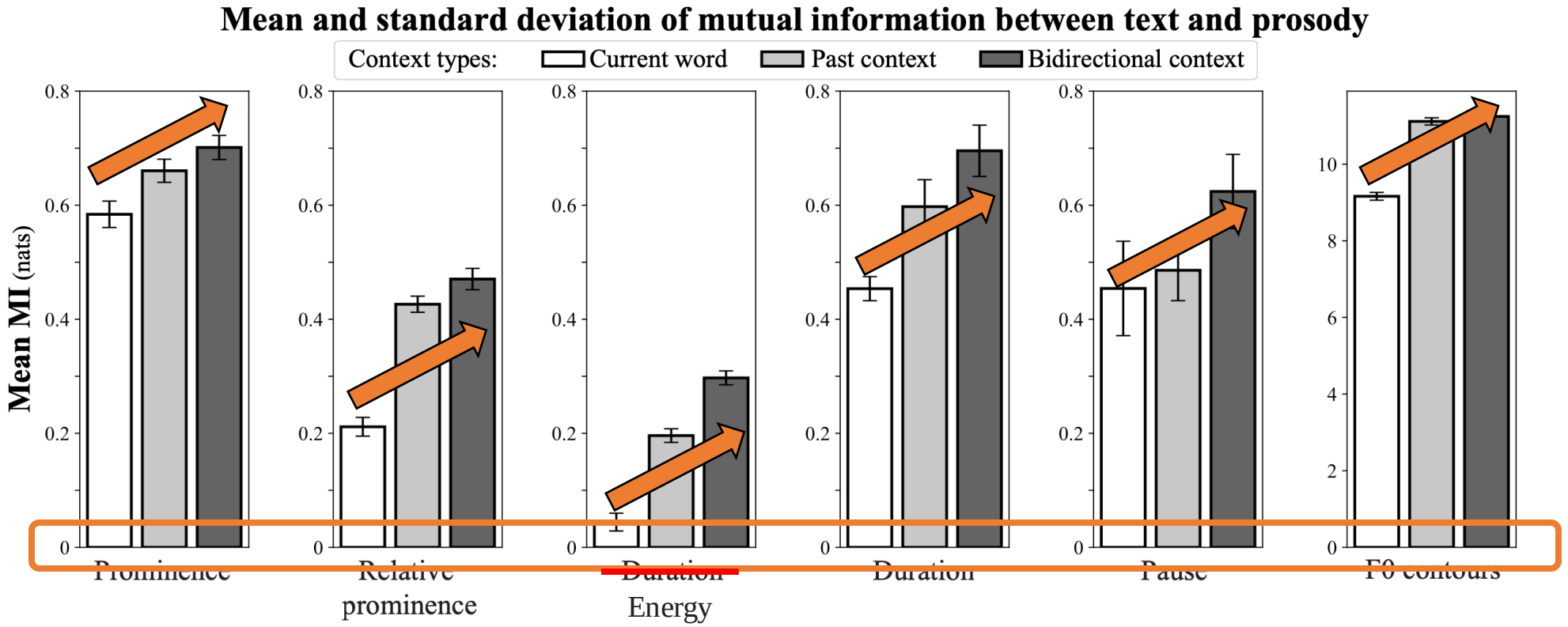
Contextual Estimator (Bidirectional context)

- BERT系モデル + 線形層 で fine-tuning してパラメータ ϕ を推定

実験結果

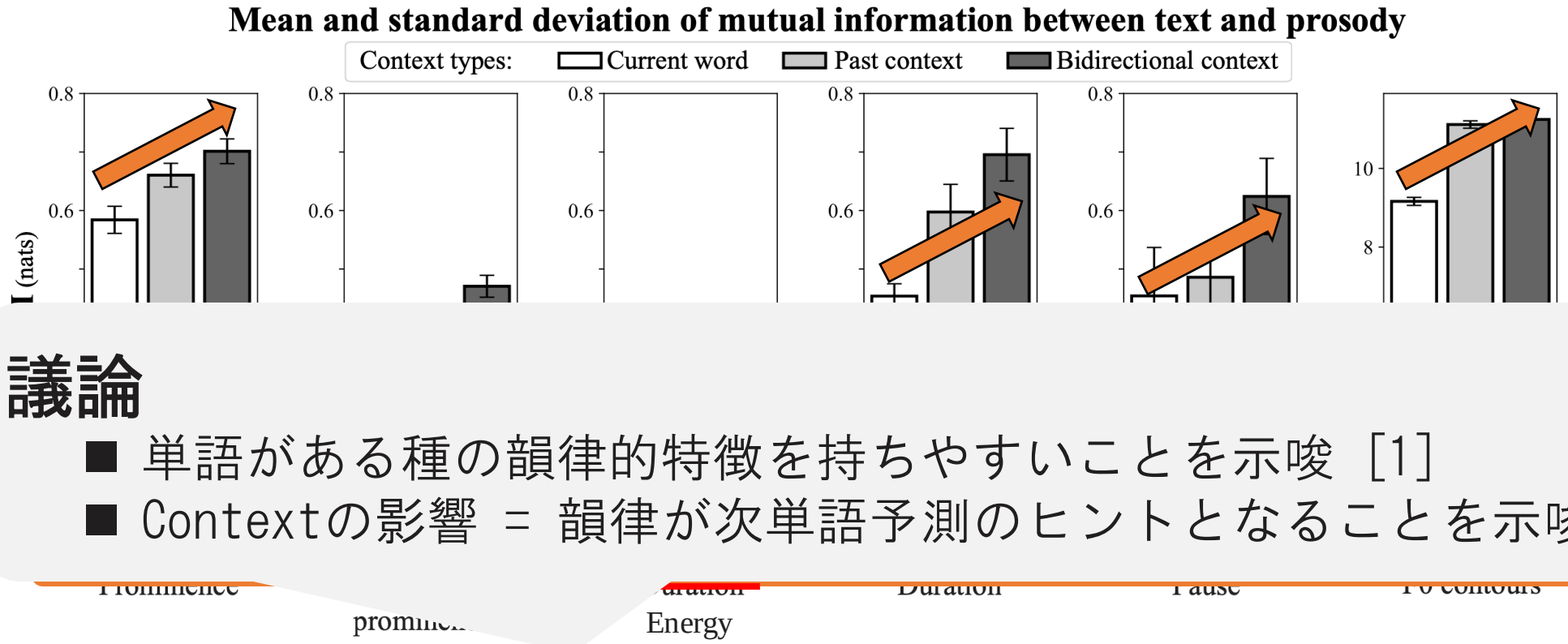

Mean and standard deviation of mutual information between text and prosody
Context types:
Current word
Past context
Bidirectional context
Mean MI (nats)
Prominence
Relative prominence
Duration
Energy
Duration
Pause
F0 contours

実験結果



- 全ての Prosody 特徴量で相互情報量が正 → 冗長性はある
- Context によって冗長度が増える傾向

実験結果



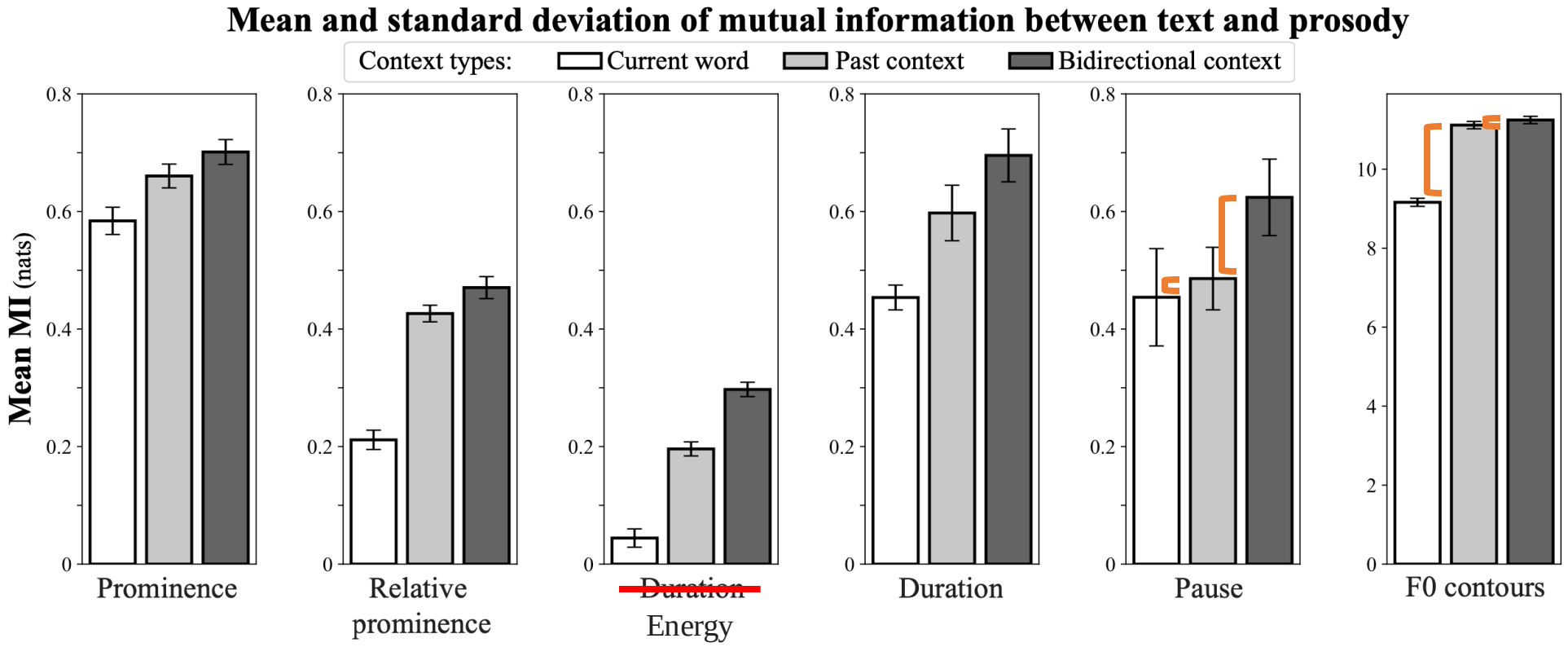
議論

- 単語がある種の韻律的特徴を持ちやすいことを示唆 [1]
- Contextの影響 = 韻律が次単語予測のヒントとなることを示唆

- **全ての Prosody 特徴量で相互情報量が正 → 冗長性はある**
- **Context によって冗長度が増える傾向**

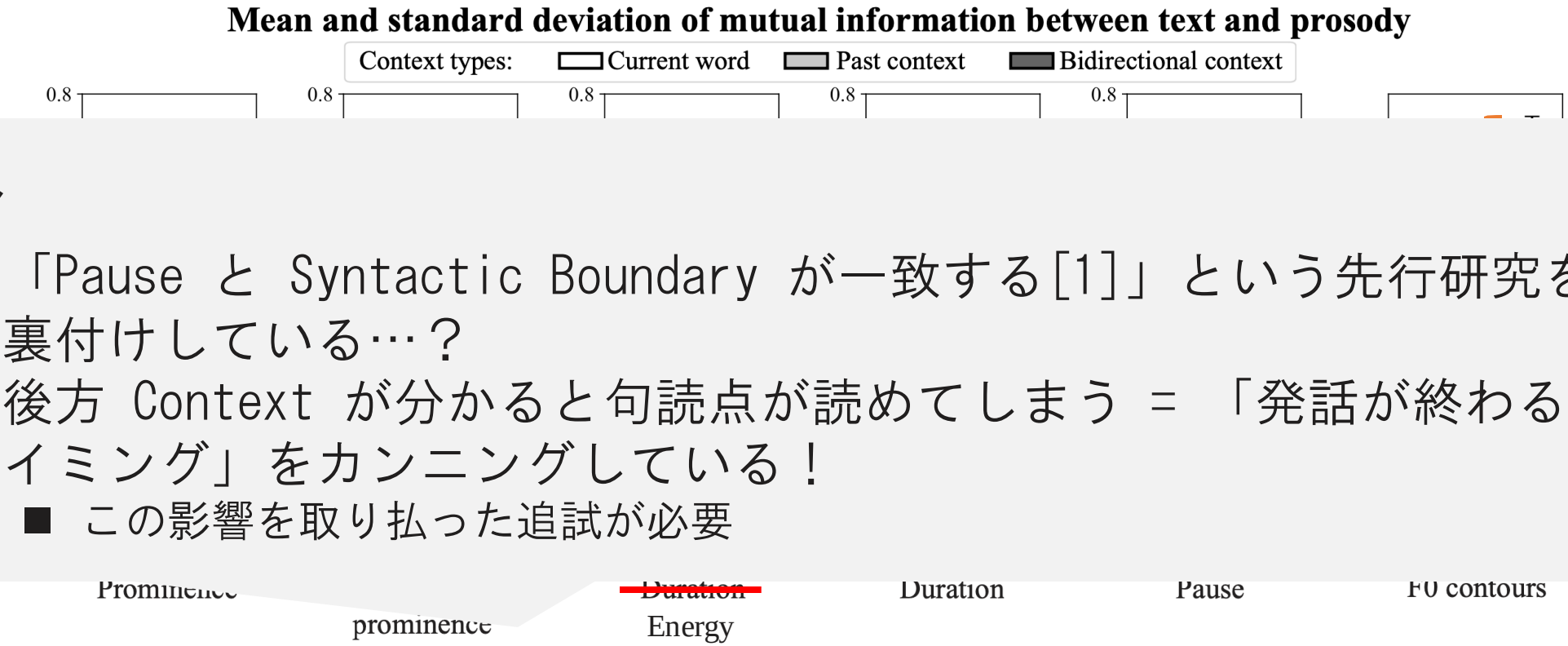
[1] Kevin Tang and Jason A. Shaw. 2021. “Prosody leaks into the memories of words.” Cognition.

実験結果



前方・後方のどちらの Context が大事かは特徴量ごとに差がある
例：F0 contours（音程）は前方、Pause は後方

実験結果



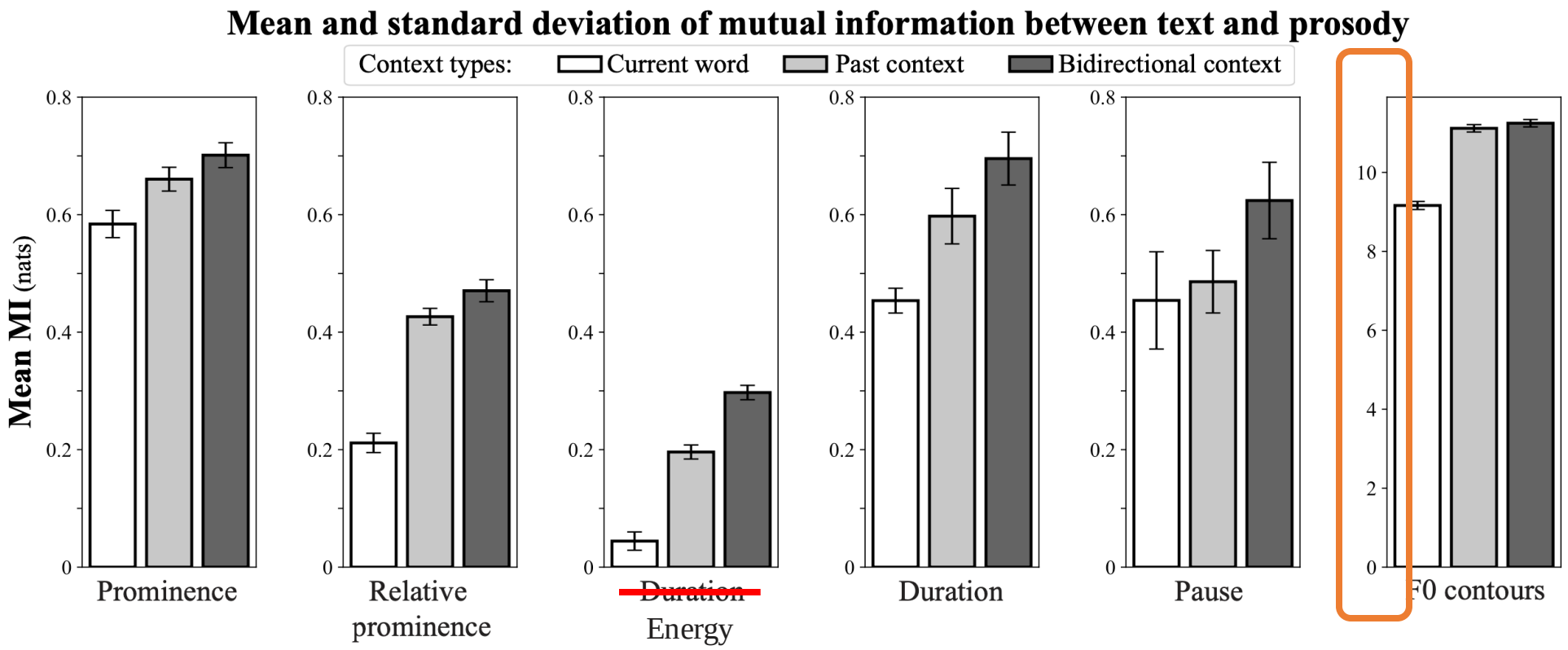
議論

- 「Pause と Syntactic Boundary が一致する[1]」という先行研究を裏付けしている…？
- 後方 Context が分かると句読点が読めてしまう = 「発話が終わるタイミング」をカンニングしている！
  - この影響を取り払った追試が必要

前方・後方のどちらの Context が大事かは特徴量ごとに差がある
例：F0 contours（音程）は前方、Pause は後方

[1] Peter R Hawkins. 1971. “The syntactic location of hesitation pauses.” Language and Speech

実験結果



F0 Contours の相互情報量がかなり大きい
→ これだけ8次元の特徴量 → 情報量・エントロピーが大きい

Limitations

- クロスエントロピーによる近似がゆるすぎるかもしれない
  - もっと大規模データ・強いLMを使うことで真の分布にできるだけ近づけてゆきたい
- 英語の電子書籍読み上げ（LibriTTS dataset）のみでの実験となっている
  - Cf. 冒頭の「この橋」vs「この端」
  - 多言語に限らずそもそも Tokenizer にかなり依存しそう
- 具体的にどんな単語で冗長度が高まるのかはわからない
  - 特に Prosody の予測精度が高い単語を集めてみれば何か分かるかも
  - とはいえ網羅的に定量評価していることが本研究のポイント

まとめとコメント

- 自然言語の Text と Prosody（韻律）が内包する情報の冗長度を定量化
  - 冗長性はある
  - 韻律特徴量によって傾向に違いがある
- 冗長度を定量化しているだけで「冗長だからいらない」というわけではもちろん無い
- 相互情報量の upper bound が明示されていると良かった。。（同一分布で測れば分かるはず？）
  - 相互情報量の値の大きさの解釈がちょっとよく分からないね