

ChatGPT-EDSS: ChatGPT 由来の Context Word Embedding から学習される共感的対話音声合成モデル

○齋藤 佑樹¹, 高道 慎之介¹, 飯森 栄治¹, 橘 健太郎², 猿渡 洋¹
(1: 東京大学, 2: LINE 株式会社)

概要: ChatGPT を活用して対話の文脈情報を推測 & 合成音声の韻律を制御可能な共感的対話音声合成

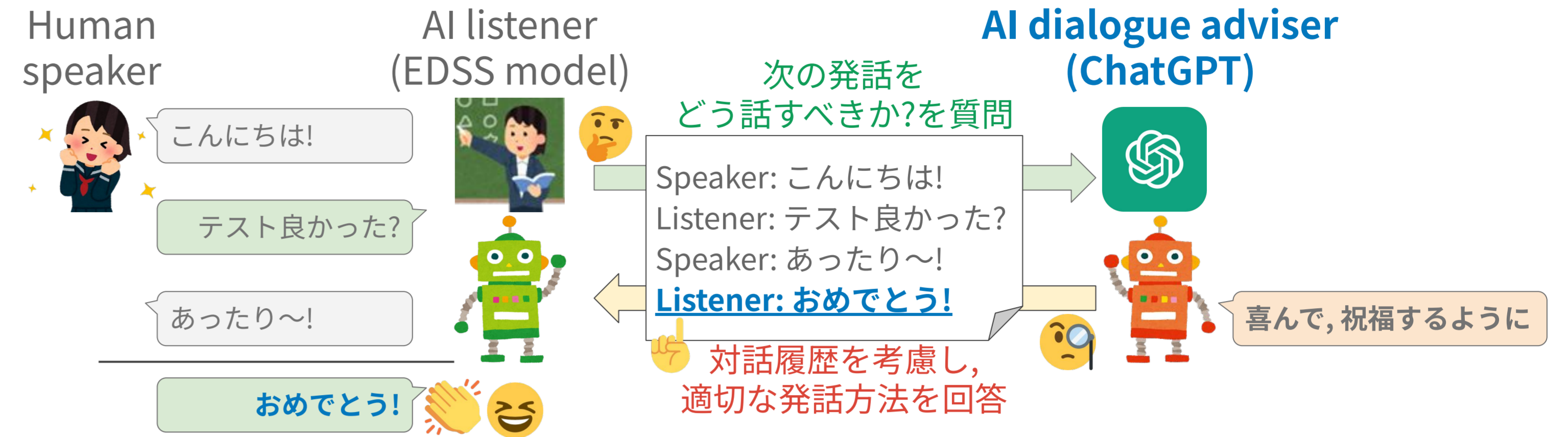
● 共感的対話音声合成 (Empathetic Dialogue Speech Synthesis: EDSS)

- 対話相手に共感するような韻律での対話音声合成^[1]
 - 対話状況の例: 生徒と塾講師の雑談^[1], 顧客とオペレータの通話^[2]
- 対話文脈情報により, 適切な発話スタイルを予測・制御
 - 話者の感情ラベル → 人手でのアノテーションが必要 😞
 - DNN 由来の対話文脈ベクトル^[3] → 人間にとって解釈困難 😞

● ChatGPT: 種々のNLPタスクでブレークスルーを達成

- 小説執筆, 作詞, パーソナリティ推定^[4], 感情推定^[5], etc.
- この優れた文章読解能力を音声対話研究に活用できないか?

● ChatGPT-EDSS: ChatGPT を AI 対話アドバイザーとして活用



ChatGPT による指示に基づき, 合成音声の韻律を制御可能な EDSS モデルを実現

提案法: ChatGPT を活用した文脈語の収集 & EDSS モデル学習

ChatGPT 文脈語収集

講師と生徒の会話があります。生徒は塾に通っており、講師はその塾で働いています。会話のシチュエーションは「生徒がテストで良い点数を取って、嬉しそうにしているので、講師が生徒を褒める」です。

1. 女子生徒「先生、こんにちは!」
2. 講師「こんにちは。何かいいことあったでしょ。」
3. 女子生徒「わかります? 物理のテスト、すごく良かったです!」
4. 講師「おめでとう! 難しいって言いながら、頑張っていたから。」
5. 女子生徒「それもあつけど、先生が「出るよ」って言ったところ、全部出たんですよ! ひっくり!」

それぞれの台詞の意図&感情&発話スタイルを一つの単語 (例えば、1. 祝福&喜び&喜んで、2. 報告&喜び&元気、3. 問いかけ&信頼&可愛い、4. 疑問&期待&クール、5. 疑問&驚き&穏やかな) で答えてください。但し、感情は「平静、期待、怒り、嫌悪、悲しみ、驚き、恐れ、信頼、喜び」、発話スタイルは「可愛い、クール、落ち着いた、丁寧、知的、誠実、爽やか、穏やか、渋い、生き生きした」の中から絶対に1つずつ選択してください。他の言葉は使わないでください。回答はExcelにコピーできるような形式にした上で絶対に日本語で書いてください。英語や中国語では書かないでください。回答にはオリジナルの台詞と話者を絶対に含まないでください。

1. 祝福 & 喜び & **喜んで**
2. 問いかけ & 信頼 & 可愛い
3. 報告 & 喜び & **元気**
4. 祝福 & 期待 & クール
5. 報告 & 驚き & 穏やか

事前に指定したカテゴリ外の語での回答

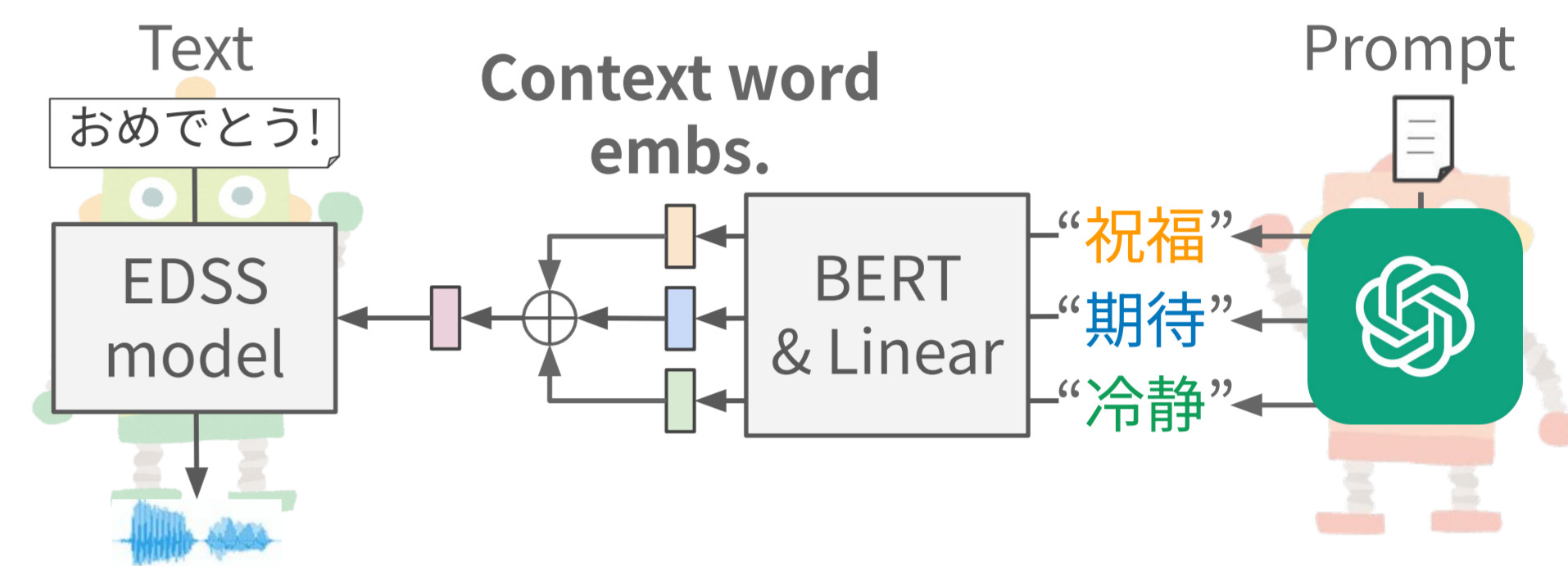
● データ: STUDIES コーパス^[1]

- 長い対話: 最大5ターンの複数バッチに分割
 - 例: (1, 2, 3, 4, 5, 6, 7, ...)
 - (1, 2, 3, 4, 5), (3, 4, 5, 6, 7), ...
- 長すぎる対話で ChatGPT が回答途中で停止する問題を緩和

● ChatGPT への入力プロンプト

- ① 対話状況の説明: 話者の役割, 対話の前提
- ② 対話台本の記述: “[ターンID][話者][発話内容]”の列
- ③ 文脈語生成の指示: 対話意図 (Intention)^[6] (感情 (Emotion)^[7]) 発話スタイル (Style) 事前にカテゴリを指定

EDSS モデル学習



● ChatGPT 文脈語埋め込みでの条件付

- { 発話意図, 感情, 発話スタイル } 語を BERT で埋め込みベクトル化して加算 → 文脈語による合成音声の韻律制御を実現
 - c.f. 自然言語記述に基づく合成音声制御^[8]
- ChatGPT を, 相互作用可能な対話文脈ベクトル抽出器として活用
 - 対話文脈を考慮した Text-Predicted Global Style Tokens (TP-GSTs)^[9] としても解釈可能
 - ただし, 文脈語推定結果には誤りを含む (一種の hallucination)

ChatGPT 文脈語収集結果の分析と EDSS 実験

文脈語収集結果の分析

● 文脈語の収集条件

- 作業数: 31名
 - Excel 上に記入されたプロンプト (A列) を ChatGPT 質問欄にコピー
 - ChatGPT の回答を B列にコピー
 - 得られた回答の信頼性を 1~5 でスコアリング (スコアリング以外は OpenAI API で自動化可能)
- ChatGPT への再回答依頼基準
 - 分脈語生成の途中で停止した場合
 - 回答が分脈語以外 (話者名や元の対話文) を含む場合
 - 日本語以外 (英語・中国語など) で回答された場合 (区切り文字が統一されていないこともあったが, 許容)

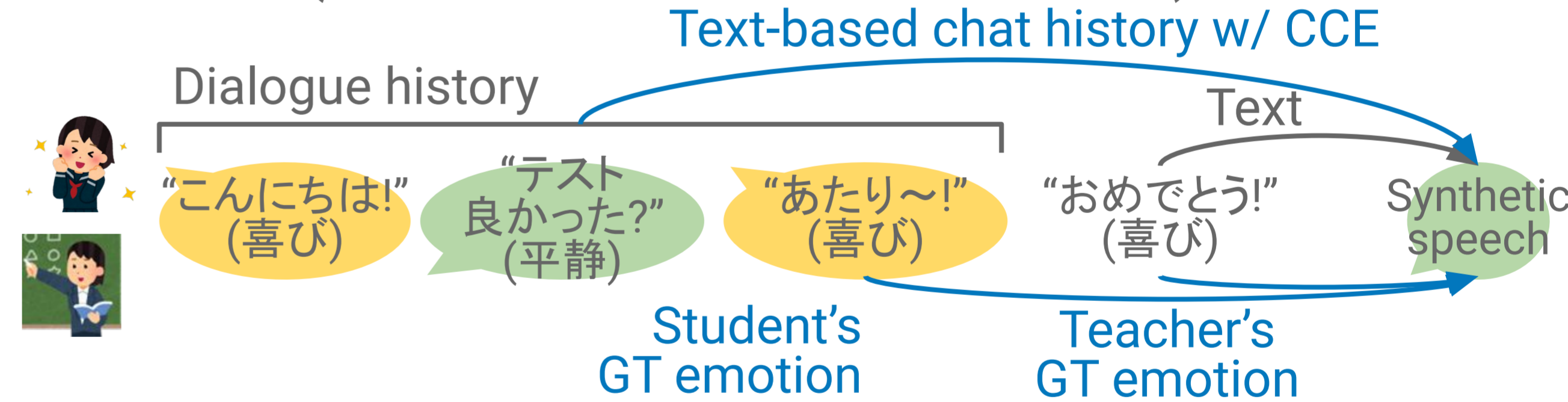
	A	B	C
1	講師と生徒の会話があります。生徒は塾に通っており、講師はその塾で働いています。講師は、生徒に共感するように喋ります。会話のシチュエーションは「生徒がテストで良い点数を取って、嬉しそうにしているので、講師が生徒を褒める」です。	ChatGPT 回答	信頼性スコア
2	1. 女子生徒「先生、こんにちは!」	祝福 & 喜び & 喜んで	1
3	2. 講師「こんにちは。何かいいことあったでしょ。」	問いかけ & 信頼 & 可愛い	4
4	3. 女子生徒「わかります? 物理のテスト、すごく良かったです!」	報告 & 喜び & 元気	5
5	4. 講師「おめでとう! 難しいって言いながら、頑張っていたから。」	祝福 & 期待 & クール	4
6	5. 女子生徒「それもあつけど、先生が「出るよ」って言ったところ、全部出たんですよ! ひっくり!」	報告 & 驚き & 穏やか	4

1. 女子生徒: 祝福、喜び、喜んで
2. 講師: 可愛い、喜び、元気
3. 女子生徒: 問いかけ、信頼、可愛い
4. 講師: 期待、喜び、クール
5. 女子生徒: 驚き、喜び、穏やか



EDSS 実験

● 実験設定 (基本構造は先行研究^[1]に基づく)

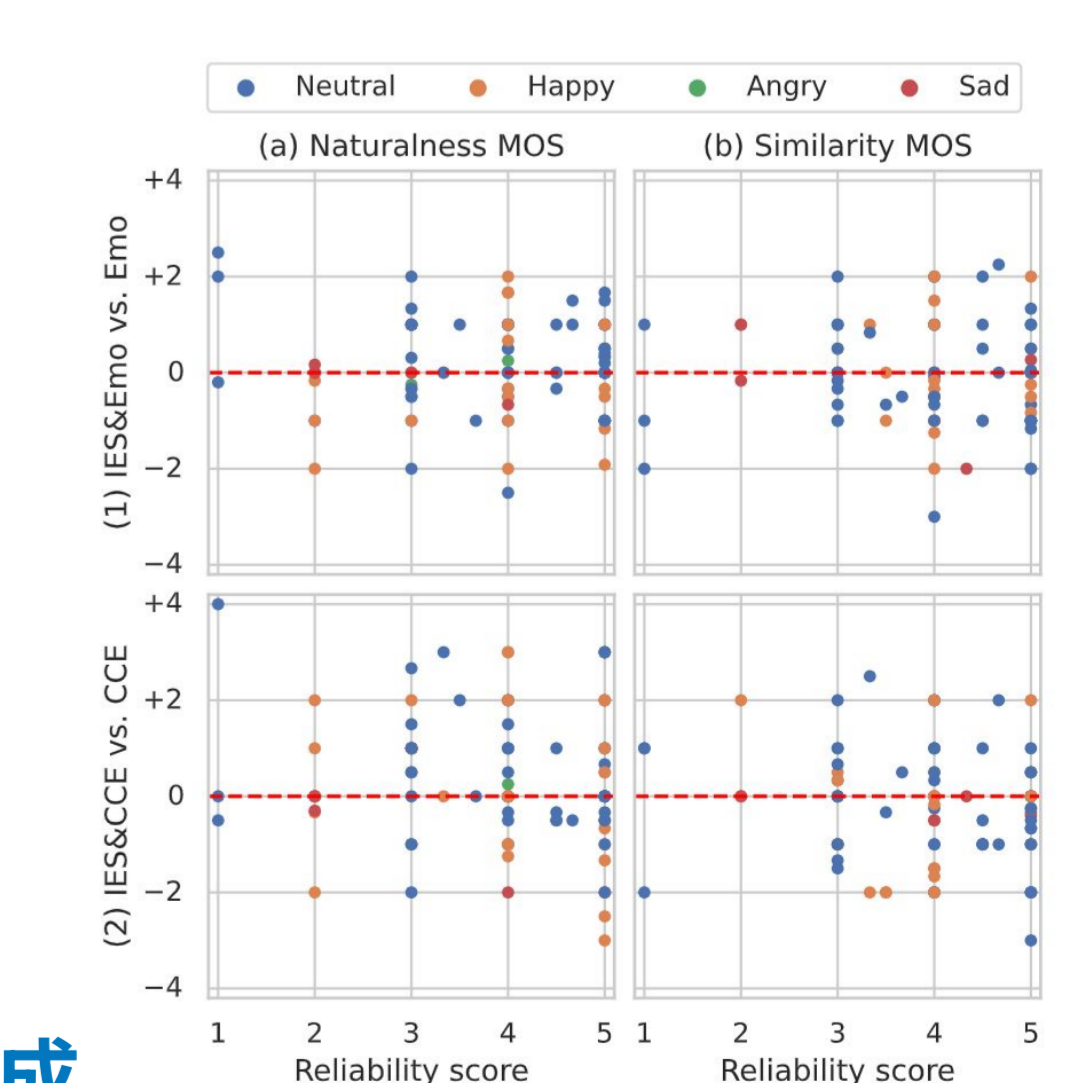


EDSS モデル	FastSpeech 2 ^[10] + HiFi-GAN ^[11]
Conversational Context Encoder (CCE) ^[3]	BERT-based ^[12]
感情ラベルの入力形式	One-hot vector → lookup emb.
学習 / 検証 / 評価データ (講師)	2,209 / 221 / 221 発話

● 合成音声の Mean Opinion Score (MOS) 評価

- FastSpeech 2 の条件付け特徴量を比較
 - GT 感情, CCE, ChatGPT 分脈語 (IES)
- 音声の自然性 (Nat.) と発話スタイル再現性 (Sim.) を評価

条件付け特徴量	MOS	
	Nat.	Sim.
GT感情	3.43	3.20
CCE	3.54	3.24
IES	3.52	3.19
GT感情 + CCE	3.52	3.21
GT感情 + IES	3.43	3.24
CCE + IES	3.49	3.20



- 結果まとめ
 - ChatGPT 分脈語単体でも, GT感情, CCEと同程度の品質を達成
 - 回答の信頼性スコアと合成音声品質には相関なし (↑ 図)

● STUDIES 講師の発話に対して得られた分脈語の分析

- GT感情: STUDIES 対話文に付与されている Ground-Truth の感情ラベル

GT感情	発話意図 (Intention)		感情 (Emotion)		発話スタイル (Style)	
	信頼性スコアの平均	ユニークな分脈語数	最頻出語	ユニークな分脈語数	最頻出語	ユニークな分脈語数
平静	3.95	206	問いかけ	130	落ち着いた	42
喜び	4.04	76	祝福	35	穏やか	19
怒り	3.66	17	共感	17	丁寧	8
悲しみ	4.03	49	共感	53	丁寧	19

概ね信頼できる | 共感を理解 | 感情も理解 | 穏やかな話者と推定

非常に多様な分脈語表現

Reference

[1] Y. Saito et al., INTERSPEECH, 2022. [4] H. Rao et al., arXiv:2302.02083, 2023. [7] T. S. Polzin et al., ITRW on Speech and Emotion, 2000. [10] Y. Ren et al., ICLR, 2021.
 [2] Y. Saito et al., INTERSPEECH, 2023. (accepted) [5] C. Qin et al., arXiv:2302.06476, 2023. [8] Z. Guo et al., ICASSP, 2023. [11] J. Kong., NeurIPS, 2020.
 [3] H. Guo et al., SLT, 2021. [6] N. Hojo et al., INTERSPEECH, 2019. [9] D. Stanton et al., SLT, 2018. [12] J. Devlin et al., NAACL-HLT, 2019.