
Yuki Saito\textsuperscript{1} \hspace{1cm} Shinnosuke Takamichi\textsuperscript{1} \hspace{1cm} Eiji Iimori\textsuperscript{1} \hspace{1cm} Kentaro Tachibana\textsuperscript{2} \hspace{1cm} Hiroshi Saruwatari\textsuperscript{1}

\textsuperscript{1}The University of Tokyo, Japan, \textsuperscript{2}LINE Corp. Japan
Explain ChatGPT-EDSS.

ChatGPT-EDSS is a ChatGPT-based technology for controlling the speaking style of Empathetic Dialogue Speech Synthesis (EDSS), a text-to-speech for a conversational agent that can talk with humans empathetically. It estimates three context words: intention, emotion, and speaking style, from chat history and conditions the EDSS model by BERT embedding vectors of the three words to predict proper prosody of synthetic speech.
The research highlights of ChatGPT-EDSS are as follows:

- It investigated how to introduce ChatGPT into speech synthesis.
- It presented the prompt design to obtain useful contexts words and analyzed the obtained context words.
- From the results of EDSS experiments, it demonstrated that ChatGPT-EDSS achieved the comparable quality of synthetic speech to EDSS models conditioned on human-annotated emotion labels and deeply learned contextual embedding vectors.
Introduction: ChatGPT & EDSS

- **ChatGPT**: cutting-edge chatbot based on Large Language Model (LLM)
  - Various creative applications (e.g., writing novels & lyrics)
  - Superior **reading comprehension** (e.g., estimating personality\(^1\) / sentiment\(^2\))

- **EDSS\(^3\)**: Dialogue Speech Synthesis that can **empathize** with human
  - e.g., chit-chat between teacher/student\(^3\) & phone call in customer center\(^4\)
  - Prediction & control of appropriate speaking styles using **dialogue context**
    - Emotion label of speaker's/listener's utterance ... need for laborious annotation
    - Deeply learned chat history embedding vector\(^5\) ... low interpretability for humans
Key idea: Using ChatGPT as AI dialogue advisor for EDSS

Human speaker

Hi, teacher!

Oh, did you get a good score?

Bingo!!

AI listener (EDSS model)

Speaker: Hi, teacher!

Listener: Oh, did you get a good score?

Speaker: Bingo!!

Listener: Congrats!!

Ask advise on proper speaking style

Answer how to respond to the speaker, considering the chat history

Speak with vibrant style like blessing.
Overview of ChatGPT-EDSS

1. Collecting ChatGPT context words

Text prompt for ChatGPT

A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score on a test."

1. Speaker: *Hello!*
2. Listener: *Oh, did you get a good score?*
3. Speaker: *Bingo! I improved my scores!*
4. Listener: *Congratulations!*

Answer three words representing intention, emotion, and speaking style for each line in the conversation.

2. Training EDSS model conditioned by the context words

Answer from ChatGPT

1. Greeting, Joy, Clear
2. Question, Trust, Polite
3. Report, Joy, Vibrant
4. Blessing, Joy, Vibrant

Dialogue context

EDSS model

Congratulations!
A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score."

1. Student: Hello!
2. Teacher: Oh, did you get a good score?
3. Student: Bingo! I improved my scores!
4. Teacher: Congratulations!

Answer three words representing intention, emotion, and speaking style for each line in the conversation. The answer format should be "[Line number]. [Intention word] & [Emotion word] & [Speaking style word]", and not include the original lines. For example:

1. [word 1-1] & [word 1-2] & [word 1-3]
2. [word 2-1] & [word 2-2] & [word 2-3]
3. [word 3-1] & [word 3-2] & [word 3-3]
4. [word 4-1] & [word 4-2] & [word 4-3]

Select emotion and speaking style words from { neutral, joy, anticipation, anger, disgust, sadness, surprise, fear, trust } and { cute, cool, quiet, polite, intellectual, honest, clear, gentle, gravelly, vibrant }, respectively.

We collected context words in Japanese using Japanese prompts, but this slide translates it in English.
A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score."

1. Student: Hello!
2. Teacher: Oh, did you get a good score?
3. Student: Bingo! I improved my scores!
4. Teacher: Congratulations!

Answer three words representing intention, emotion, and speaking style for each line in the conversation. The answer format should be "[Line number]. [Intention word] & [Emotion word] & [Speaking style word]", and not include the original lines. For example:

1. [word 1-1] & [word 1-2] & [word 1-3]
2. [word 2-1] & [word 2-2] & [word 2-3]
3. [word 3-1] & [word 3-2] & [word 3-3]
4. [word 4-1] & [word 4-2] & [word 4-3]

Select emotion and speaking style words from { neutral, joy, anticipation, anger, disgust, sadness, surprise, fear, trust } and { cute, cool, quiet, polite, intellectual, honest, clear, gentle, gravelly, vibrant }, respectively.
1. Collecting ChatGPT context words: Prompt to ChatGPT

A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score."

1. Student: Hello!
2. Teacher: Oh, did you get a good score?
3. Student: Bingo! I improved my scores!
4. Teacher: Congratulations!

Answer three words representing intention, emotion, and speaking style for each line in the conversation. The answer format should be "[Line number]. [Intention word] & [Emotion word] & [Speaking style word]", and not include the original lines. For example:

1. [word 1-1] & [word 1-2] & [word 1-3]
2. [word 2-1] & [word 2-2] & [word 2-3]
3. [word 3-1] & [word 3-2] & [word 3-3]
4. [word 4-1] & [word 4-2] & [word 4-3]

Select emotion and speaking style words from { neutral, joy, anticipation, anger, disgust, sadness, surprise, fear, trust } and { cute, cool, quiet, polite, intellectual, honest, clear, gentle, gravelly, vibrant }, respectively.
A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score."

1. Student: Hello!
2. Teacher: Oh, did you get a good score?
3. Student: Bingo! I improved my scores!
4. Teacher: Congratulations!

Answer three words representing intention, emotion, and speaking style for each line in the conversation. The answer format should be "[Line number]. [Intention word] & [Emotion word] & [Speaking style word]", and not include the original lines. For example:

1. [word 1-1] & [word 1-2] & [word 1-3]
2. [word 2-1] & [word 2-2] & [word 2-3]
3. [word 3-1] & [word 3-2] & [word 3-3]
4. [word 4-1] & [word 4-2] & [word 4-3]

Select emotion and speaking style words from \{ neutral, joy, anticipation, anger, disgust, sadness, surprise, fear, trust \} and \{ cute, cool, quiet, polite, intellectual, honest, clear, gentle, gravelly, vibrant \}, respectively.

③ Request for generating context words
1. Collecting ChatGPT context words: ChatGPT's output

1. Informative & Neutral & Polite
2. Inquisitive & Anticipation & Polite
3. Excited & Joy & Clear
4. Congratulatory & Joy & Polite

Observations: tends to...

- Stop answering when the dialogue lines are too long (e.g., > 10 turns).
  - We split long dialogue lines into multiple batches including max. 5 turns.
    e.g., (1, 2, 3, 4, 5, 6, 7 ...) → (1, 2, 3, 4, 5), (3, 4, 5, 6, 7), ...
- Output context words that do not follow the instruction written in the prompt.
  - Using a word out of pre-defined choices for the emotion and speaking style
  - Splitting the three words by a different delimiter (e.g., commas)
- Generate different context words for the same dialogue line.

We also discuss the reliability of ChatGPT's answer in this talk.
Overview of ChatGPT-EDSS

1. Collecting ChatGPT context words

Text prompt for ChatGPT
A speaker has a conversation with a listener. The listener empathetically responds to the speaker under the situation: "The listener prizes the speaker who got a good score on a test."
1. Speaker: Hello!
2. Listener: Oh, did you get a good score?
3. Speaker: Bingo! I improved my scores!
4. Listener: Congratulations!

Answer three words representing intention, emotion, and speaking style for each line in the conversation.

2. Training EDSS model conditioned by the context words

Answer from ChatGPT
1. Greeting, Joy, Clear
2. Question, Trust, Polite
3. Report, Joy, Vibrant
4. Blessing, Joy, Vibrant

BERT & Linear
Dialogue context
EDSS model
Congratulations!
2. Training EDSS model conditioned by the context words

- **Related work**
  - Controlling speaking style by natural language description (e.g., PromptTTS[^5])
  - Predicting speaking style from text to be spoken by TTS (e.g., TP-GSTs[^6])
  - Training expressive TTS based on weakly supervised learning[^7]
Experiments & Discussions
(analysis of ChatGPT answers & EDSS evaluation)
Experimental setup for ChatGPT context word collection

- # of involved human workers: 31 for ...
  1. Copying & pasting the ChatGPT prompt & answers using Google Sheets
  2. (If necessary) Resending the prompt when ChatGPT failed to answer correctly
  3. Filling in the reliability of obtained ChatGPT answers by an integer betw. 1 ~ 5

1 & 2 can be fully automated with OpenAI API (available from Mar. 2, 2023).
### Analysis of ChatGPT-derived context words

- **Dataset:** STUDIES[^3] teacher's 3,365 utterances (5 hours)
  - GT emotion label: annotated by the corpus developers (i.e., human labelers)

<table>
<thead>
<tr>
<th>GT emotion label</th>
<th>Avg. reliability score</th>
<th>Intention</th>
<th>Emotion</th>
<th>Style</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Most frequent word</td>
<td># of unique words</td>
<td>Most frequent word</td>
</tr>
<tr>
<td>Neutral</td>
<td>3.95</td>
<td>Question</td>
<td>206</td>
<td>Anticipation</td>
</tr>
<tr>
<td>Happy</td>
<td>4.04</td>
<td>Blessing</td>
<td>76</td>
<td>Happiness</td>
</tr>
<tr>
<td>Angry</td>
<td>3.66</td>
<td>Empathy</td>
<td>17</td>
<td>Trust</td>
</tr>
<tr>
<td>Sad</td>
<td>4.03</td>
<td>Empathy</td>
<td>49</td>
<td>Sadness</td>
</tr>
</tbody>
</table>

## Analysis of ChatGPT-derived context words

- **Dataset:** STUDIES\(^3\) teacher's 3,365 utterances (5 hours)
  - GT emotion label: annotated by the corpus developers (i.e., human labelers)

<table>
<thead>
<tr>
<th>GT emotion label</th>
<th>Avg. reliability score</th>
<th>Intention</th>
<th>Emotion</th>
<th>Style</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Most frequent word</td>
<td># of unique words</td>
<td>Most frequent word</td>
</tr>
<tr>
<td>Neutral</td>
<td>3.95</td>
<td>Question</td>
<td>206</td>
<td>Anticipation</td>
</tr>
<tr>
<td>Happy</td>
<td>4.04</td>
<td>Blessing</td>
<td>76</td>
<td>Happiness</td>
</tr>
<tr>
<td>Angry</td>
<td>3.66</td>
<td>Empathy</td>
<td>17</td>
<td>Trust</td>
</tr>
<tr>
<td>Sad</td>
<td>4.03</td>
<td>Empathy</td>
<td>49</td>
<td>Sadness</td>
</tr>
</tbody>
</table>

All avg. reliability scores > 3.6 → generally reliable answers!
### Analysis of ChatGPT-derived context words

- **Dataset:** STUDIES[^3] teacher's 3,365 utterances (5 hours)
  - GT emotion label: annotated by the corpus developers (i.e., human labelers)

<table>
<thead>
<tr>
<th>GT emotion label</th>
<th>Avg. reliability score</th>
<th>Intention</th>
<th>Emotion</th>
<th>Style</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Most frequent word</td>
<td># of unique words</td>
<td>Most frequent word</td>
</tr>
<tr>
<td>Neutral</td>
<td>3.95</td>
<td>Question</td>
<td>206</td>
<td>Anticipation</td>
</tr>
<tr>
<td>Happy</td>
<td>4.04</td>
<td>Blessing</td>
<td>76</td>
<td>Happiness</td>
</tr>
<tr>
<td>Angry</td>
<td>3.66</td>
<td>Empathy</td>
<td>17</td>
<td>Trust</td>
</tr>
<tr>
<td>Sad</td>
<td>4.03</td>
<td>Empathy</td>
<td>49</td>
<td>Sadness</td>
</tr>
</tbody>
</table>

ChatGPT can understand the concept of "empathetic" dialogue!

[^3]: Link to original dataset or study.
### Analysis of ChatGPT-derived context words

- **Dataset**: STUDIES\textsuperscript{[3]} teacher's 3,365 utterances (5 hours)
  - GT emotion label: annotated by the corpus developers (i.e., human labelers)

<table>
<thead>
<tr>
<th>GT emotion label</th>
<th>Avg. reliability score</th>
<th>Intention</th>
<th>Emotion</th>
<th>Style</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Most frequent word</td>
<td># of unique words</td>
<td>Most frequent word</td>
</tr>
<tr>
<td>Neutral</td>
<td>3.95</td>
<td>Question</td>
<td>206</td>
<td>Anticipation</td>
</tr>
<tr>
<td>Happy</td>
<td>4.04</td>
<td>Blessing</td>
<td>76</td>
<td>Happiness</td>
</tr>
<tr>
<td>Angry</td>
<td>3.66</td>
<td>Empathy</td>
<td>17</td>
<td>Trust</td>
</tr>
<tr>
<td>Sad</td>
<td>4.03</td>
<td>Empathy</td>
<td>49</td>
<td>Sadness</td>
</tr>
</tbody>
</table>

ChatGPT's answers are too diverse, despite pre-defining the word candidates.
Experimental setup for EDSS evaluation

- **Baseline: Dialogue-history-aware EDSS from our previous work**\(^3\)
  - Comparing the use of 1) GT emotion, 2) CCE-derived context embedding, & 3) ChatGPT-derived context embedding (IES) as a conditional feature

- **Other conditions (see our paper for details)**
  - EDSS model: FastSpeech 2\(^8\) + HiFi-GAN vocoder\(^9\)
  - # of training/validation/test data = 2,209/221/221 utterances
Subjective evaluations of synthetic speech

- 2 MOS tests involving 100 listeners
  - Criteria: naturalness & speaking-style similarity

<table>
<thead>
<tr>
<th>Conditional feature</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT emo.  CCE IES</td>
<td>Naturalness</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>3.43 ± 0.14</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>3.54 ± 0.14</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>3.52 ± 0.14</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>3.52 ± 0.14</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>3.49 ± 0.14</td>
</tr>
</tbody>
</table>

50 listeners for each test (30 answers per listener)
Subjective evaluations of synthetic speech

- **2 MOS tests involving 100 listeners**
  - Criteria: naturalness & speaking-style similarity

<table>
<thead>
<tr>
<th>Conditional feature</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT emo. CCE IES</td>
<td>Naturalness</td>
</tr>
<tr>
<td>✓</td>
<td>3.43 ± 0.14</td>
</tr>
<tr>
<td>✓</td>
<td>3.54 ± 0.14</td>
</tr>
<tr>
<td>✓</td>
<td>3.52 ± 0.14</td>
</tr>
<tr>
<td>✓</td>
<td>3.52 ± 0.14</td>
</tr>
<tr>
<td>✓</td>
<td>3.43 ± 0.14</td>
</tr>
<tr>
<td>✓</td>
<td>3.49 ± 0.14</td>
</tr>
</tbody>
</table>

ChatGPT-derived context words achieve speech quality comparable to human-annotated emotion label & DNN-derived context embedding!

50 listeners for each test (30 answers per listener)
Subjective evaluations of synthetic speech

- 2 MOS tests involving 100 listeners
  - Criteria: naturalness & speaking-style similarity

<table>
<thead>
<tr>
<th>Conditional feature</th>
<th>GT emo.</th>
<th>CCE</th>
<th>IES</th>
<th>MOS Naturalness</th>
<th>MOS Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.43 ± 0.14</td>
<td>3.20 ± 0.15</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.54 ± 0.14</td>
<td>3.24 ± 0.14</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.52 ± 0.14</td>
<td>3.19 ± 0.15</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.52 ± 0.14</td>
<td>3.21 ± 0.14</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.43 ± 0.14</td>
<td>3.24 ± 0.14</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.49 ± 0.14</td>
<td>3.20 ± 0.14</td>
</tr>
</tbody>
</table>

The reliability scores are not related to the speech quality improvement. (Perhaps main reason is the rich diversity of ChatGPT's answers?)

50 listeners for each test (30 answers per listener)
## Speech samples

<table>
<thead>
<tr>
<th>GT emo.</th>
<th>CCE</th>
<th>IES</th>
<th>Naturalness</th>
<th>Similarity</th>
<th>Happy</th>
<th>Angry</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.43 ± 0.14</td>
<td>3.20 ± 0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
<td>3.54 ± 0.14</td>
<td>3.24 ± 0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.52 ± 0.14</td>
<td>3.19 ± 0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.52 ± 0.14</td>
<td>3.21 ± 0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.43 ± 0.14</td>
<td>3.24 ± 0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3.49 ± 0.14</td>
<td>3.20 ± 0.14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary of this talk

Summarize the contributions by ChatGPT-EDSS.

The research highlights of ChatGPT-EDSS are as follows:

- It investigated how to introduce ChatGPT into speech synthesis.
- It presented the prompt design to obtain useful contexts words and analyzed the obtained context words.
- From the results of EDSS experiments, it demonstrated that ChatGPT-EDSS achieved the comparable quality of synthetic speech to EDSS models conditioned on human-annotated emotion labels and deeply learned contextual embedding vectors.
Future work

Mention future work of ChatGPT-EDSS.

The future work includes:

- Investigating the effect of the dialogue domain in ChatGPT-EDSS.
- Examining whether ChatGPT’s hallucination occurs in our method.

Thank you for your attention! 😊
Do you have any questions or comments?


