We propose an **incremental text-to-speech (TTS) synthesis method** that synthesizes speech in small linguistic units for streaming applications.

Our method uses **pseudo lookahead** generated with a language model as **future context** to **address the tradeoff between naturalness and waiting time**.

**Background**: Incremental TTS for streaming application (e.g., simultaneous speech translation)  
- **Sentence-level TTS**: Using full sentence to generate output speech  
- **Incremental TTS**: Synthesizing speech in small linguistic units  
- **Incremental TTS suffers from tradeoff between quality and latency**  
  Synthesizing each segment independently [1] ⇒ Lower latency but lower quality  
  Waiting for k words (lookahead-k policy) [2] ⇒ Higher quality but need waiting time  

We propose an incremental TTS method **generated pseudo lookahead as future context**  
Leveraging human's incremental reading (reading while predicting framework)  
⇒ Achieving higher naturalness without waiting for future input words

**Method**  
1) **Incremental synthesis procedure**  
   - Generating pseudo lookahead with **GPT2** [3] from observed segment  
   - Current segment, past observed segment, and pseudo lookahead are fed to encoder  
   - Encoded past observed segment and pseudo lookahead are fed to contextual encoder  
   - Encoded current segment and contextual embedding are combined and fed to decoder  

2) **TTS model architecture and training**  
   - Tacotron2 [4]-based neural TTS model  
   - Training encoder, decoder, and context encoder in an **end-to-end manner**  
   - Training model with ground-truth lookahead  

**Evaluation**  
1) **Experimental conditions**  
   - Corpus  
   - Number of words in each input segment: 3 (training), 2 (inference)  
   - Number of words in pseudo lookahead: 5

2) Incremental TTS systems  
   - Independent [1]  
   - Unicontext  
   - Bicontext  
   - Bicontext (fine-tuned)  
   - Bicontext (truth) [2]

   Waiting for future input words

3) Evaluation metrics  
   - **Objective evaluation**: Calculated error rates with speech recognition model  
   - **Subjective evaluation**: Mean opinion score (MOS) test with 40 native English speakers

4) Results  
   - Bicontext > Unicontext: Demonstrating the effectiveness of pseudo lookahead  
   - Bicontext (fine-tuned) ≫ Bicontext (truth): Equivalent to waiting for future words

5) Discussion  
   - GPT2 prediction is much better than random  
   - Random sampling or Maximum likelihood?  
   - Maximum likelihood (k = 1) is better  
   - How fine-tuning affects contextual embedding?  

![Cosine similarity between e_{truth} and e_{pseudo}](chart.png)

**Cosine similarity between e_{truth} and e_{pseudo}**

**Reference**  
[1] Yanagita+, 2019  
[3] Radford et al., 2019  
[5] Ito et al., 2017