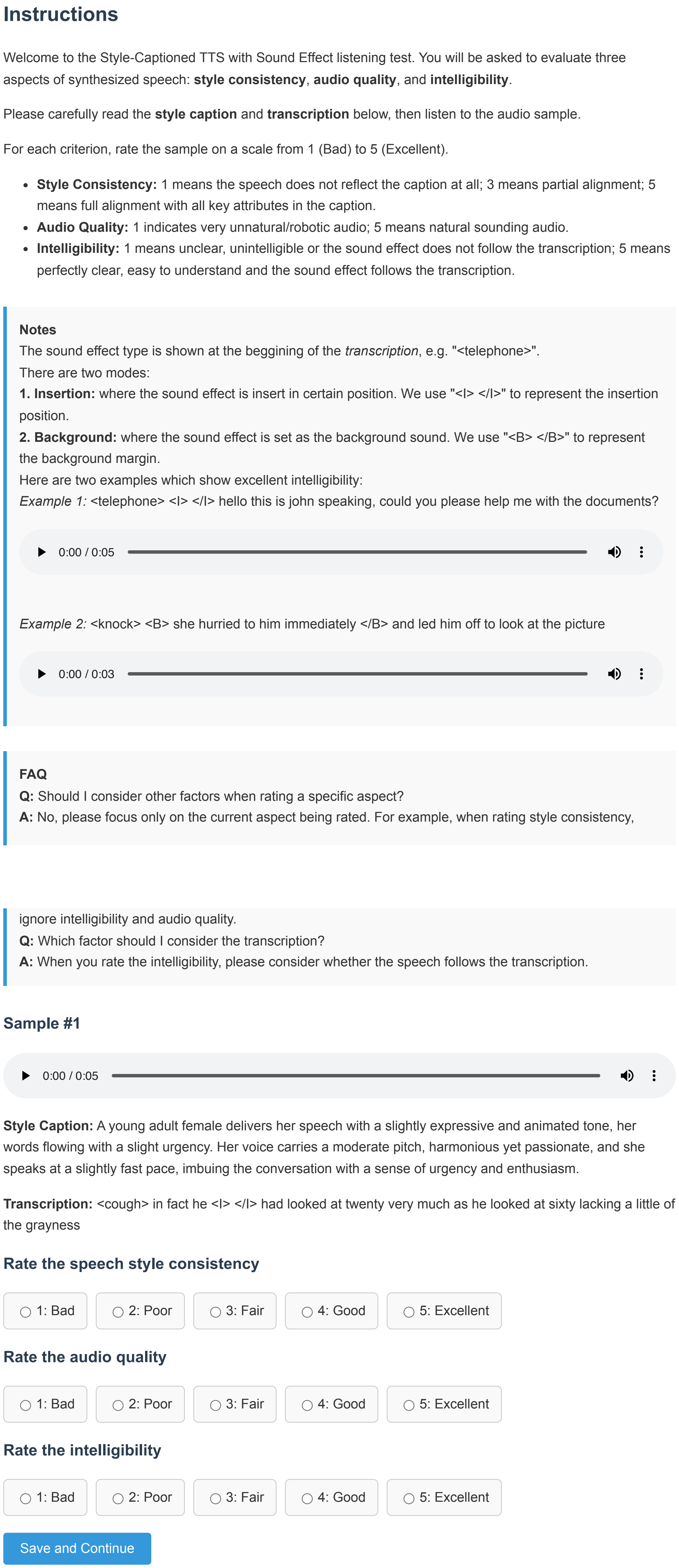

Instructions

Welcome to the Style-Captioned TTS with Sound Effect listening test. You will be asked to evaluate three aspects of synthesized speech: **style consistency**, **audio quality**, and **intelligibility**.

Please carefully read the **style caption** and **transcription** below, then listen to the audio sample.

For each criterion, rate the sample on a scale from 1 (Bad) to 5 (Excellent).

- **Style Consistency:** 1 means the speech does not reflect the caption at all; 3 means partial alignment; 5 means full alignment with all key attributes in the caption.
- **Audio Quality:** 1 indicates very unnatural/robotic audio; 5 means natural sounding audio.
- **Intelligibility:** 1 means unclear, unintelligible or the sound effect does not follow the transcription; 5 means perfectly clear, easy to understand and the sound effect follows the transcription.

> **Notes**
> The sound effect type is shown at the begging of the *transcription*, e.g. "<telephone>".
> There are two modes:
> **1. Insertion:** where the sound effect is insert in certain position. We use "<I> </I>" to represent the insertion position.
> **2. Background:** where the sound effect is set as the background sound. We use "<B> </B>" to represent the background margin.
> Here are two examples which show excellent intelligibility:
> *Example 1:* <telephone> <I> </I> hello this is john speaking, could you please help me with the documents?
>
> ▶ 0:00 / 0:05
>
> *Example 2:* <knock> <B> she hurried to him immediately </B> and led him off to look at the picture
>
> ▶ 0:00 / 0:03

> **FAQ**
> **Q:** Should I consider other factors when rating a specific aspect?
> **A:** No, please focus only on the current aspect being rated. For example, when rating style consistency, ignore intelligibility and audio quality.
> **Q:** Which factor should I consider the transcription?
> **A:** When you rate the intelligibility, please consider whether the speech follows the transcription.

Sample #1

▶ 0:00 / 0:05

Style Caption: A young adult female delivers her speech with a slightly expressive and animated tone, her words flowing with a slight urgency. Her voice carries a moderate pitch, harmonious yet passionate, and she speaks at a slightly fast pace, imbuing the conversation with a sense of urgency and enthusiasm.

Transcription: <cough> in fact he <I> </I> had looked at twenty very much as he looked at sixty lacking a little of the grayness

Rate the speech style consistency

○ 1: Bad ○ 2: Poor ○ 3: Fair ○ 4: Good ○ 5: Excellent

Rate the audio quality

○ 1: Bad ○ 2: Poor ○ 3: Fair ○ 4: Good ○ 5: Excellent

Rate the intelligibility

○ 1: Bad ○ 2: Poor ○ 3: Fair ○ 4: Good ○ 5: Excellent