Appendix: The Subjective evaluation

In the evaluation, each set of experimental configurations for each speaker is 30 synthesized speeches evaluated by 10 different listeners, and each person gives the evaluation results separately. The subjective evaluation of speech quality and speaker similarity adopts a 5-point system, and the evaluation score can only be an integer between 1-5. The meaning of each score is as follows:

Speech quality

```
1: Can’t understand, can only understand a few words
2: Some key words are unclear, and the pauses and pronunciation make peop
le feel uncomfortable
3: Generally understandable and acceptable. The rhythmic pause is not goo
d enoug
4: Natural, clear and understandable, good hearing, willing to accept
5: Broadcasting level, unable to distinguish between human voice and synt
hesized voice
```

Speech similiarity

Compared with tacotron based text2mel model

```
1: This is definitely not the same person, even the gender is different.
2: It's not like the same person, most of them don't resemble, there are
individual details.
3: It may be the same person, but I can't be sure.
4: Should be the same person, although there are some differences
5: This must be the same person, the tone and the way of speaking are the
 same person
```

The influence of prosody selection

To highlight the role of prosody selection, speaker similarity here uses more stringent standards.

```
1: This is definitely not the same person, even the gender is different.
2: It may be the same person, but I can't be sure.
3: Should be the same person, although there are some differences
4: High probability be the same person. Not only heard as one person, but
 the tone and intonation are also somewhat similar.
5: Must be the same person, their voice, intonation and speaking style ar
e very similar.
```