A APPENDIX

A. Objective Metrics

We design a set of objective metrics to evaluate the prosody similarity between TTS wave and recording with the same transcripts, including pitch, intensity, duration and pause. Specifically, as shown in Figure 5, we use an internal vocoder analysis tool to generate the frame level f0 (fundamental frequency) and intensity of the given TTS wave and recording. It is worth mentioning that in order to get more accurate results, we perform linear interpolation and median filtering on f0 to obtain a continuous and smooth f0 curve. Given the transcription text, we can obtain the phoneme boundary, phoneme duration and pause duration between adjacent phonemes by an internal force align tool. In addition, we use an internal syllabification tool to obtain the syllable boundary of phoneme sequences.

Combing frame-level f0, intensity and the phoneme boundary, we can get phoneme-level f0 and intensity by averaging the frame-level features inside a phoneme. The duration of all the phonemes inside the syllable is added to get the duration of the syllable. The pause duration after the last phoneme of the syllable is treated as the pause duration of the syllable [2]. When the phone-level f0, intensity and syllable duration, pause features are ready, we calculate the pearson correlation and RMSE of the above prosody features between TTS wave and recording as the objective metrics. The pearson correlation is calculated by Eq.(7), where x and y are the prosody feature vector of TTS wave and recording respectively, E_x and E_y are the mean of the feature vector of x and y. RMSE is calculated by Eq.(8), where N is the vector length of x and y. As we mentioned in section 4.1, these objective metrics are used as an additional reference combined with subjective evaluation for our model selection.

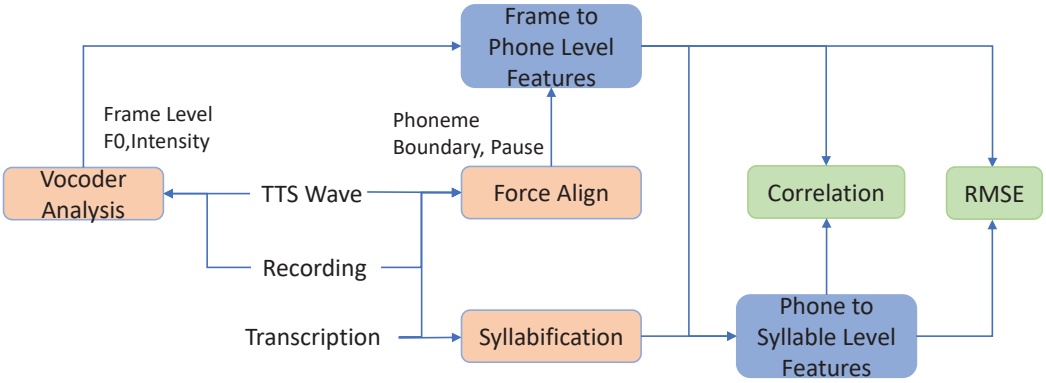

Figure 5: Objective Metrics.

$$pearson = \frac{\sum(x - E_x)(y - E_y)}{\sqrt{\sum(x - E_x)^2 \sum(y - E_y)^2}} \quad (7)$$

$$rmse = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \quad (8)$$

B. Selection of Memory Length

We have to select the best parameter setting of memory length for both encoder and decoder. As the number of different comparison pairs is large and the difference among these models is not obvious enough, we use the objective metrics described in Appendix A to perform the evaluation. Notice that the result here is based on the ContextSpeech model without the integration of text-based contextual encoder.

Encoder Memory Length To select the memory length of encoder, we first set the decoder memory length as 0, and then compare model performance under different encoder memory lengths (from

[2]From our experiments and observation, syllable-level features are more appropriate for the comparison of duration and pause related prosody feature.

32 to 512). The results are presented in Table 5. The distribution of the best metric values seems not concentrate on one model, since not all carried information in memory is useful for current sentence synthesis. We can see that the model with 128 as the encoder memory length achieves 3 best metrics, which is better than other models. 128 is also a number close to the average input length of 1 sentence in our data set. Therefore, we select 128 as the final setting of encoder memory length in our ContextSpeech model.

Table 5: Objective Metrics Result of Different Encoder Memory Length.

	Correlation				RMSE			
Metrics	Pitch	Intensity	Duration	Pause	Pitch	Intensity	Duration	Pause
32-0	0.674	0.852	0.760	**0.897**	25.204	11.620	46.576	**56.917**
64-0	0.676	0.849	0.775	0.887	25.474	11.796	46.692	58.051
128-0	**0.678**	0.856	**0.786**	0.884	**24.820**	11.475	46.723	57.437
256-0	0.662	0.856	0.776	0.890	24.966	**11.377**	46.620	58.307
512-0	0.669	**0.858**	0.770	0.892	25.134	11.477	**45.958**	58.904

Decoder Memory Length. Given the memory length of encoder is 128, we compare models with different memory length in the decoder, from 16 to 512, and the result are shown in Table 6. We can see that the model with decoder memory length as 64 achieves the best overall performance among the parameter tuning range [16, 32, 64, 128, 256, 512]. Thus, we select 64 as the decoder memory length in the final ContextSpeech model. The results also reveal that longer memory will not produce better model performance, which can be illustrated form two respects: 1) We just need a small piece of speech at the end of last sentence to help the current sentence synthesis to the greatest degree. 2) Directly leveraging speech information by concatenation with previous decoder memory cannot make the most use of it, we need a more complex module to handle longer decoder memory. We will do more investigation on these observations in the future work.

Table 6: Objective Metrics Result of Different Decoder Memory Length.

	Correlation				RMSE			
Metrics	Pitch	Intensity	Duration	Pause	Pitch	Intensity	Duration	Pause
128-16	0.674	0.843	0.740	0.889	25.973	11.969	48.107	58.418
128-32	0.686	**0.849**	0.759	0.891	**24.920**	11.777	47.862	57.390
128-64	**0.707**	**0.849**	**0.786**	**0.894**	24.980	**11.562**	**45.693**	**55.876**
128-128	0.666	0.845	0.761	0.890	25.476	11.821	48.118	58.084
128-256	0.664	0.846	0.763	0.891	26.703	11.835	47.421	57.824
128-512	0.677	0.843	0.761	0.891	25.677	11.906	47.227	57.961

C. Definition of Metrics in Paragraph MOS

As we mentioned in paper, in the paragraph MOS, 25 native speakers listen to each audio and give a score in 10-point scale according to the overall performance and each specific metric, including naturalness, pleasantness, speech pause, stress, intonation, emotion, style matchiness and listening effort. The definition of each metric can be found in the following items. The price of this test is 0.1 dollar per case per judge.

- **Overall impression**. How is your overall impression on this content reading, considering the inside and cross sentences? Consider if the voice is clear, natural, expressive, easy to understand and pleasant to listen to.
- **Naturalness**. How nature is this content reading, considering the inside and cross sentence prosody?
- **Pleasantness**. If the voice sounds comfortable and pleasant reading this content
- **Speech pause**. If the break between words and the silence between sentences are appropriate?
- **Stress**. If the degree of emphasis is natural and correct?
- **Intonation**. If the melody and variation in the pitch level fits the sentence?
- **Emotion**. If the emotion is expressive and suitable for the content?

- **Style matchiness**. How much is the voice suitable for reading the content from the speaking style?
- **Listening effort**. How easy is it to focus on this voice and get information?

D. Samples

https://contextspeech.github.io/demo/

E. Detail Steps of Some Equations

- **ConformerBlock**
  ConformerBlock in Eq.(1) is presented in Fig.1 and the process can also be formatted with the following equations:

$$H^{'}_{t,n} = ConvM(H^{n}_{t}) + H^{n}_{t} \tag{9}$$

$$H^{''}_{t,n} = MHSA(H^{'}_{t,n}) + H^{'}_{t,n} \tag{10}$$

$$ConformerBlock(H^{n}_{t}) = ConvFFN(H^{''}_{t,n}) + H^{''}_{t,n} \tag{11}$$

- **Linearized Self-Attention with Permute-based Relative Position Encoding**
  Combine permutation operation in Eq.(6) into linearized self-attention, Eq.(4-5) can be rewritten as Eq.(12-13)

$$\mathcal{A}(Q_i, K, V) = \Big(\sum_{j=1}^{L} (r_i P^i_B \phi(Q_i))^T (r^{-j} P^j_B \phi(K_j)) V_j \Big) / \Big(\sum_{j=1}^{L} (r_i P^i_B \phi(Q_i))^T (r^{-j} P^j_B \phi(K_j)) \Big) \tag{12}$$

$$= \Big(r_i P^i_B \phi(Q_i))^T \sum_{j=1}^{L} (r^{-j} P^j_B \phi(K_j)) V_j \Big) / \Big((r_i P^i_B \phi(Q_i))^T \sum_{j=1}^{L} (r^{-j} P^j_B \phi(K_j)) \Big) \tag{13}$$

  As the r is set as 1 in our model setting, the Eq.(13) can be simplified as:

$$\mathcal{A}(Q_i, K, V) = \Big(P^i_B \phi(Q_i))^T \sum_{j=1}^{L} (P^j_B \phi(K_j)) V_j \Big) / \Big((P^i_B \phi(Q_i))^T \sum_{j=1}^{L} (P^j_B \phi(K_j)) \Big) \tag{14}$$