

A APPENDIX

A.1 DATA PROCESSING

From LibriLight, we extract audio clips of up to 30s. For each samples in the batch, we take a random crop of up to 10s starting and ending based on the time-aligned grapheme sequences from pre-decoded CTC model outputs. These outputs are obtained from the large wav2vec2 model pre-trained on the LibriLight and fine-tuned on the LibriSpeech dataset (Baevski et al., 2020). We infer word and chunk level timestamps of whisper transcripts by aligning them with the time-aligned grapheme sequences by the wav2vec model. These timestamp information is used in training and inference to simulate streaming.

A.2 TEST SETS FOR ADDITIONAL STUDY

For the experiments below, results are reported for small-scale test sets: (1) [D-off] Offline Test Set: 94 samples, each lasting 3-10s. (2) [D-mixed] Mixed Test Set: no filtering on the target duration. Each chunk contains 3-10 word tokens.

A.3 ADDITIONAL EXPERIMENTS

We train some variations of our base model:

- [M1] Base model, as described in the main paper
- [M2] Numbers of acoustic codes in each group are [2, 3, 4, 8] (versus [4, 4, 4, 5]). With less codes predicted in high-level groups, we expect better quality in high-level (early) codes, which are important to generate better low-level (later) codes.
- [M3] Different enrollment speech features are utilized in each head. Specifically, the speech encoder outputs 64 features, with each of the 4 heads accessing 16 features. This approach aims to allow each head to learn specific enrollment speech features independently, rather than sharing them.
- [M4] Only the grapheme decoding has access to the transcript. The other acoustic code decodings can only observe the generated graphemes. There are 1, 4, 4, 8 codebooks in each group, respectively, to allow this condition. This design ensures that acoustic tokens are generated solely based on the grapheme sequence.
- [M5] We fine-tune our base model [M1] and limit the text context at training time as we do at inference time when using semantic guidance.
- [M6] We pre-train our base model [M1] while limiting the text content at training time as we do at inference time using semantic guidance.

We report results in Table 7. To our surprise, prioritizing early codebook groups [M2] does not improve the CER score, although more capacity is given to predict high-level codes, which are crucial for content accuracy. The [M3] model shows a significant degradation in the CER score but a promising SS score, despite the low CER score. We observe a slight performance decrease in the [M4] model compared to the base model [M1], indicating that access to both the transcript and the predicted grapheme sequence are necessary for the prediction of acoustic codes.

Table 7: Additional Experiments. Results reported for the D-mixed dataset

ID	Model	CER	WER	SS	DNSMOS
[M1]	Base	2.6	3.6	64.7	3.9
[M2]	Prioritizing High-level	3.4	4.1	63.3	3.9
[M3]	Separating Enrollment Features	8.1	10.6	64.9	3.8
[M4]	Only Grapheme Seeing Transcript	3.2	4.2	63.9	3.9

A.4 MORE ABLATION STUDY & ANALYSIS

In this section, we report results on the D-off test set. We provide results for baseline models on this test set in Table 8.

Table 8: Baseline results for the test set used used for offline inference ablation study & analysis.

Model	CER	WER	SS	DNSMOS
XTTS v2	2.0	2.9	59.0	3.95
YourTTS	3.5	3.7	47.2	3.82

Guidance λ Table 9 presents the effects of different guidance values. Guidance at any level appears to benefit content accuracy; however, high guidance values may negatively impact the score. Additionally, we observe that high guidance values make it more challenging for the generation to complete full sentences ($\lambda = \infty$ in Table 10).

k in top-k sampling for semantic tokens Table 10 shows the effect of limiting the top-k candidates when sampling the semantic tokens with different values of λ . Overall, there is no clear conclusion on whether sampling benefits more from a larger or smaller number of candidates.

Table 9: Effect of guidance λ , $k^{(g)} = 5$

λ	CER	WER	SS	DNSMOS
0	3.1	3.5	61.0	3.85
1	2.4	2.9	60.7	3.85
2	2.6	3.2	60.9	3.85
3	2.5	2.9	60.8	3.85
4	2.6	3.1	60.9	3.85
5	2.7	3.0	61.2	3.85
6	2.9	3.0	61.1	3.85

Table 10: Effect of k in top-k sampling, dataset [D-off]. $\lambda = 0$ means no guidance, $\lambda = \infty$ means hard guidance

Model	λ	CER	WER	SS	DNSMOS
$k^{(g)} = 5$	0	2.8	3.5	60.9	3.85
	1	2.7	3.1	59.2	3.86
	2	2.5	3.1	60.9	3.86
	3	2.6	3.2	61.0	3.85
	∞	6.8	46.1	59.7	3.85
$k^{(g)} = 3$	0	2.8	3.7	60.7	3.85
	2	2.7	2.9	61.2	3.85
$k^{(g)} = 2$	0	2.9	3.3	61.1	3.85
	10	2.5	3.2	61.1	3.84

A.5 STREAMING-AWARE TRAINING

In the model discussed in the main paper, training is conducted with access to the full context, while online inference is performed with a restricted context, potentially resulting in a mismatch between training and inference conditions. As a result, the model does not perform well in extreme cases when the chunk has only one word, or the model sees one chunk ahead. In this section, we present the results of aligning training with inference by incorporating random context dropout during the training process.

Fine-tuning for the streaming scenario To simulate streaming conditions during training, we implement a dynamic attention masking strategy using Algorithms 2 and 3. This approach modifies the key-value mask to create a visible window centered around the query’s closest text position. The algorithms introduce controlled noise by randomly adjusting the start and end points of this visible window, effectively masking out different regions of the text input. This randomized masking simulates the noise and partial information availability characteristic of streaming scenarios.

Our masking strategy employs a “window range” parameter that creates a context window around each text token. This window ensures that each query has access to some context from the text, preventing complete information loss.

Algorithm 2: Dynamic Cross-Attention Text-Dropout Mask

Input: kv_mask, key_pos_id, seq_len, window_range (r_1, r_2)
Output: mask
kv_len \leftarrow Sum(kv_mask, dim=-1);
query_pos \leftarrow CreateQueryPositions(seq_len);
closest_text_pos \leftarrow arg min(|query_pos - key_pos_id|);
window_start \leftarrow Floor(Rand() \cdot (closest_text_pos - r_1)).clamp(min=0);
window_end \leftarrow closest_text_pos + r_2 + ;
Floor(Rand() \cdot (kv_len - closest_text_pos - r_2)).clamp(min=0));
window_end \leftarrow Min(window_end, kv_len - 1);
key_pos \leftarrow CreateKeyPositions(kv_mask.shape);
mask \leftarrow (key_pos \geq window_start) \wedge (key_pos \leq window_end);
return mask

Algorithm 3: Applying Cross-Attention Mask and Computing Attention Weights

Input: pre_w (results of QK^T) kv_mask, key_pos_id, seq_len, (r_1, r_2) window_range
Output: w (attention weights)
window_mask \leftarrow Algorithm 2(kv_mask, key_pos_id, seq_len, (r_1, r_2));
kv_mask \leftarrow CombineMasks(kv_mask, window_mask);
mask_values \leftarrow CreateMaskValues(kv_mask, pre_w.shape);
pre_w \leftarrow pre_w + mask_values;
w \leftarrow Softmax(pre_w, dim=-1);
return w

We conducted ablation studies to assess the effectiveness of this masked fine-tuning strategy. Tables 11 and 12 present the results of these studies, where we fine-tuned our base model [M1] using the masking techniques described in Algorithms 2 and 3. These results demonstrate the impact of our dynamic masking approach on the model’s performance in streaming scenarios.

Pre-training for the streaming scenario Similar to fine-tuning for the streaming scenario, we conducted ablation studies to assess the effectiveness of Algorithms 2 and 3.

Tables 13 and 14 present the results of these studies, where instead of first pre-training [M1] and then fine-tuning to produce [M5], we pretrain [M1] using Algorithms 2 and 3, producing [M6]. These results demonstrate the impact of our dynamic masking approach on the model’s performance in streaming scenarios.

A.6 AREAS FOR IMPROVEMENT

Separate grapheme token prediction from acoustic token prediction Prediction of the grapheme sequence given the word sequence should not be so challenging; however, in many cases, we observe that the model does not produce the correct grapheme sequence. We hypothesize that error in acoustic codes may affect the accuracy of grapheme prediction, which in turns adversely affect the acoustic codes. By making graphemes not depending on previously decoded acoustic codes, we can potentially improve the accuracy of predicting them. Similarly, high-level codes can be made independent of low-level codes to avoid being affected by their errors.

Hard guidance may work better for transformer Hard guidance avoids sampling the wrong candidate for the next grapheme; however, in some cases, the probability for these guiding tokens are low. In state space models, choosing a low probability candidate may hurt more than in transformers, since we can only “force” the input but not the internal state.

A.7 ATTENTION VISUALIZATION

We provide cross-attention visualization for a short speech (Figure 3) and a long speech (Figure 4). The first six rows are attention visualization for the first six shared layers. Each of the last six rows presents four plots for codebook groups. For the model reported in the main paper, these four groups

Table 11: [M5] Results with different text chunk lengths (online, 44 samples, offline, 94 samples). offline: $n_p=10, n_f=2$; online: $n_p=4, n_f=2$

l_{\min}	l_{\max}	WER	CER	SS
1	1	4.4 / 5.3	3.8 / 4.3	57.0 / 64.8
1	3	2.3 / 3.6	2.1 / 3.0	57.2 / 64.5
2	2	2.9 / 3.5	3.0 / 3.0	56.9 / 64.7
2	4	2.7 / 3.5	2.7 / 2.9	57.2 / 65.3
3	7	4.7 / 7.4	4.1 / 5.8	57.3 / 64.9

Table 12: [M5] Results with different numbers of text chunks

n_p	n_f	WER	CER	SS
1	1	3.5 / 4.6	3.2 / 3.1	56.0 / 64.9
10	1	3.1 / 4.7	3.1 / 3.6	56.4 / 64.6
2	2	2.7 / 4.5	2.6 / 3.8	56.5 / 64.5
10	2	2.7 / 4.9	2.7 / 3.5	57.2 / 65.1
10	4	2.9 / 3.3	2.7 / 2.7	56.5 / 64.4

Table 13: [M6] Results with different text chunk lengths (offline, 94 samples, online, 44 samples). offline: $n_p=10, n_f=2$; online: $n_p=4, n_f=2$

l_{\min}	l_{\max}	WER	CER	SS
1	1	5.4 / 6.8	4.9 / 5.1	59.9 / 68.4
1	3	4.0 / 4.7	3.2 / 3.9	60.3 / 67.8
2	2	3.3 / 3.6	2.9 / 2.5	61.1 / 69.4
2	4	3.3 / 5.1	2.9 / 3.8	61.1 / 68.9
3	7	4.4 / 8.4	3.4 / 5.9	60.7 / 68.4

Table 14: [M6] Results with different numbers of text chunks (offline, online)

n_p	n_f	WER	CER	SS
1	1	3.3 / 4.98	3.5 / 3.5	60.6 / 68.0
10	1	4.0 / 4.98	3.7 / 3.98	60.6 / 67.6
2	2	3.2 / 5.1	2.7 / 3.8	60.6 / 69.0
10	2	3.4 / 5.5	3.2 / 4.2	60.33 / 69.0
10	4	3.3 / 5.4	2.8 / 3.8	60.5 / 69.2

contain 4, 4, 4, 5 codebooks, respectively. In the first layer, it is observed that the speech frames tend to align with the word tokens that share similar positional indices. Specifically, the initial frame aligns with the first word in a chunk since they have the same positional indices. However, because a speech chunk contains significantly more frames than there are word tokens in a text chunk, most alignment occurs within the first few frames of the speech chunk. As we go deeper into the model layers, the alignment extends across the entire speech chunk. The alignment is also observed to be less noisy in the first group of codebooks, suggesting that word tokens hold greater significance for this group compared to others.

A.8 HIGH CER SAMPLES

Table 15 and 16 list samples with highest CER scores from the subjective evaluation set for short and long utterances. It is observed that when the model occasionally makes pronunciation mistakes, especially on hard words, it can mostly avoid problems caused by misalignment such as word repetition or early finishing / hallucinating.

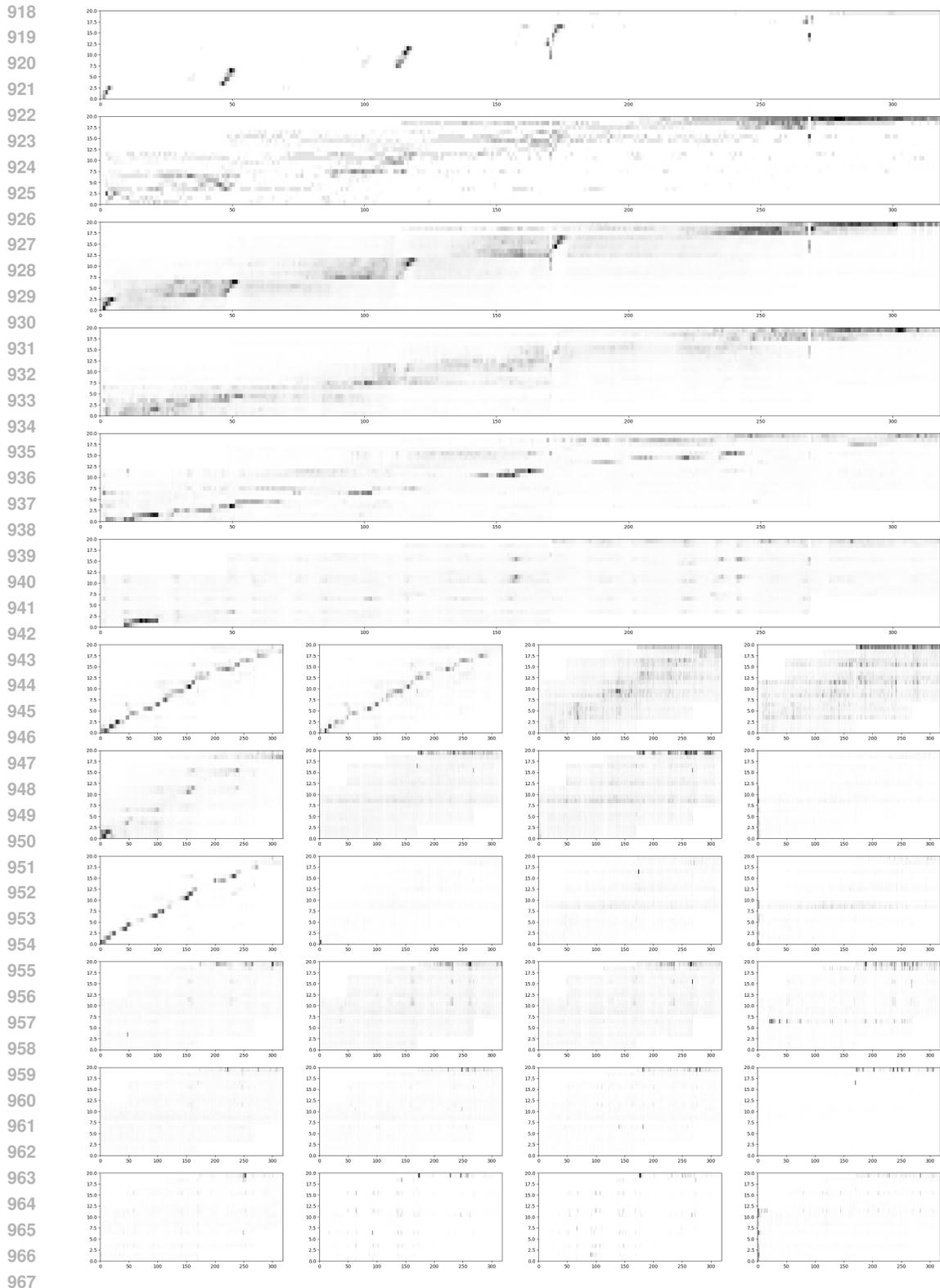


Figure 3: Cross-attention visualization for “There is even / a white row of / beehives in the / orchard under the walnut / trees”. There are 12 rows for each mamba layer, in which each in the first 6 rows has only one head and each in the last 6 rows has four heads, each predicting 4, 4, 4, 5 codes (total 1 grapheme token + 16 acoustic codes) in a frame, respectively. In each plot, the x-axis represents 318 audio frames generated and the y-axis represents 21 word tokens in the transcript.

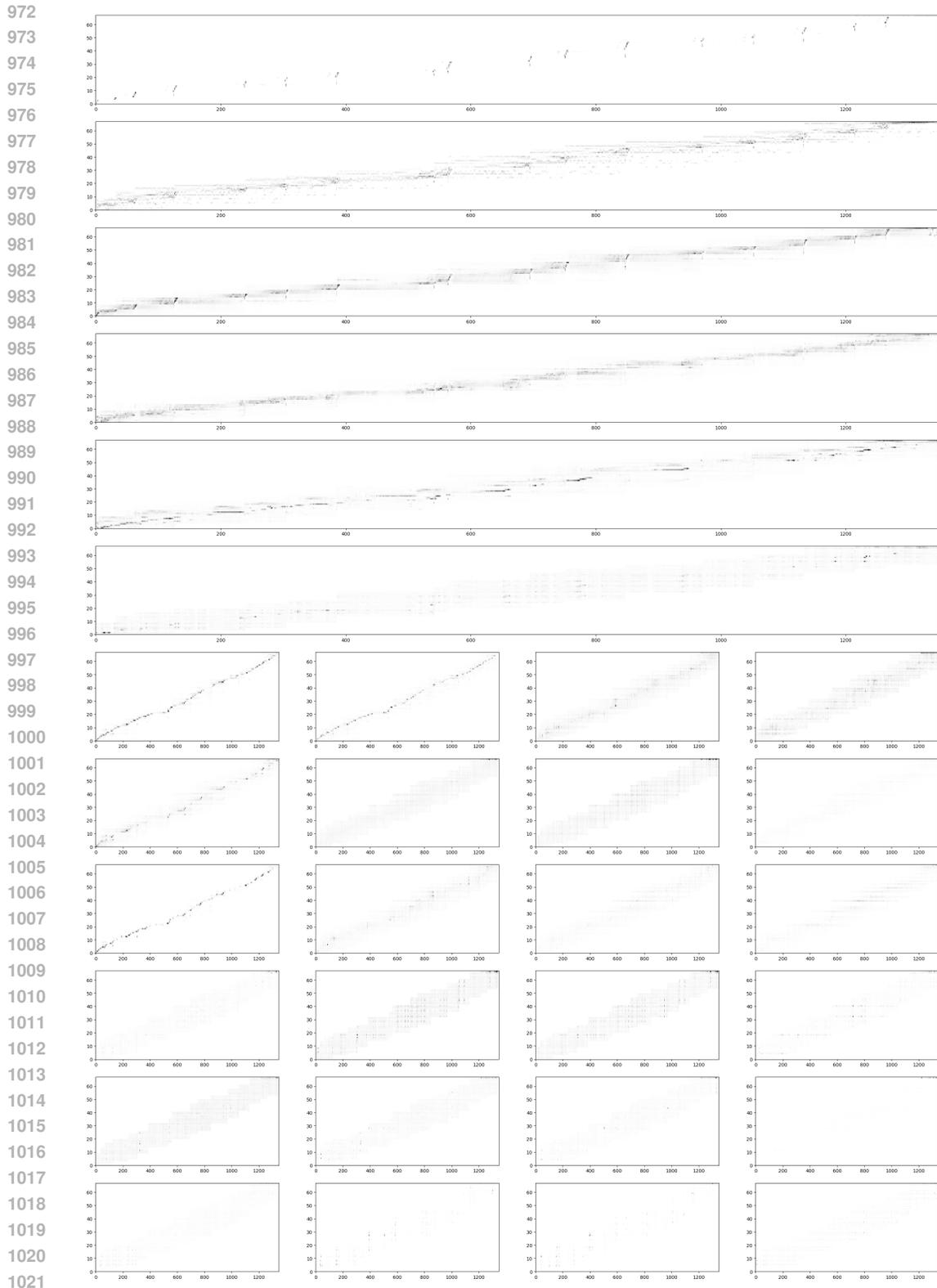


Figure 4: Cross-attention visualization for “When you go / out of /the house into the / flower garden, there you / feel again the / order and fine / arrangement manifest all over / the great / farm, in the fencing / and hedging, / in the windbreaks and / sheds, in the symmetrical / pasture ponds / planted with scrub / willows to give shade / to the cattle / in fly-time.”

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

Generated	Reference	CER
he keeps that the shot not command mats first rate and lord does	he keeps the thou shalt not commandments first rate hen lord does	18.5
but mormonism died but every taint of grief served but to unite the people	but mormonism died not every added pang of grief served but to unite the people	13.9
was it the bible oshed whispered bill harmon	was that the bible osh whispered bill harmon	11.4
all angelleaks folks are baking for it and all emittes twenty cousins	all angeliques folks are baking for it and all amitys twenty cousins	10.3
im afraid as when quite answer the purpose said his mamma smiling especially the last yet we must think of something	im afraid those wouldnt quite answer the purpose said his mama smiling especially the last yet we must think of something	8.3
the beggas plea the politicians sceptre and the drummes is ablest assistant	the beggars plea the politicians scepter and the drummers ablest assistant	8.1
every word fell distinctly in perfect harmony and eloquence upon lewis exors ears	every word fell distinctly in perfect harmony and eloquence upon louis xiv's ears	6.3

Table 15: Samples with highest CER scores from 58 short samples

Generated	Reference	CER
the invention is in universal use to day alike for direct and for alternating current and as well in the equipment of large buildings as in the inocion distribution [] of the most extensive central station metaress	the invention is in universal use today alike for direct and for alternating current and as well in the equipment of large buildings as in the distribution system of the most extensive central station networks	10.0
yet sometimes in the pauses of his work the young man frowned and looked up the ground with an intent intent which suggested that even twenty one might have its problems it	yet sometimes in the pauses of his work the young man frowned and looked at the ground with an intentness which suggested that even twentyone might have its problems	6.7
he hoped there would be stew for dinner turnips and carrots and bruised potatoes and fat mutton pieces to be laddled out in thick peppered flurfished sauce suffered into you his belly counselled him	he hoped there would be stew for dinner turnips and carrots and bruised potatoes and fat mutton pieces to be ladled out in thick peppered flourfattened sauce stuff it into you his belly counseled him	6.5
i sent the hurons he said speaking to the mo ends yonder yonders open sky through the tree tops and we are getting too nigh their encampment sagamore you will take the hillside to the right huncas will bend along the brck to the left while i will try the trail	i sent the hurons he said speaking to the mohicans yonder is open sky through the treetops and we are getting too nigh their encampment sagamore you will take the hillside to the right hunkus will bend along the brook to the left while i will try the trail	6.3

Table 16: Samples with highest CER scores from 30 samples