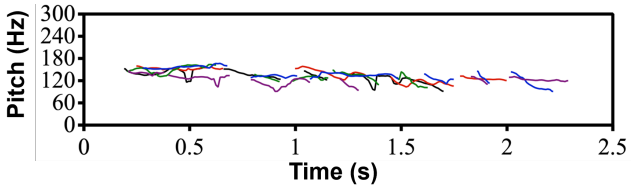
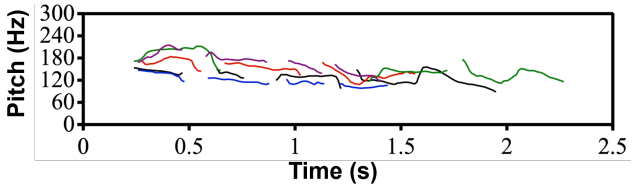
ter compared to VITS. For naturalness, we observe a consistently improving MOS-N of MQTTS as the model capacity grows. It demonstrates different scaling properties: higher model capacity brings naturalness to MQTTS, but diversity to VITS. Comparing the same parameter size (100M) for both VITS and MQTTS, MQTTS wins out in all metrics except MCD, which we explained earlier. This suggests MQTTS is generally better than VITS, given enough resources. Additionally, we observed overfitting for both 100M and 200M of MQTTS, but with a higher severity for the 200M version. This explains the little improvement from 100M to 200M and suggests that a larger training corpus is needed for further improvement.

Error Analysis. Despite the better average performance of MQTTS in Table 2, we find that it suffers from lower sampling robustness compared to non-autoregressive systems. This is reasonable as higher diversity inevitably comes with a higher risk of sampling poor syntheses. We observe unnaturally prolonged vowels in some samples with speaker reference speech of a slower speaking style, which is seldom the case for VITS. In addition, samples that start with a poor recording environment often result in bad syntheses. Deletion errors, which we consider more undesirable than substitution, are also more prevalent in MQTTS; it contributes for 8.4 out of 22.3% WER in MQTTS-100M, but only 6.8 out of 24.8% for VITS-100M. We conjecture that as intermittent pauses are not annotated explicitly, and thus it encourages MQTTS to produce silence even if attending to a certain phoneme. However, if the phones are successfully produced, they often sound more natural and intelligible than those in the syntheses of VITS. We observe these errors gradually rectified as the model capacity increases (from 40M to 200M), suggesting that more data and larger models can eventually resolve these issues.



(a) VITS.



(b) MQTTS.

Figure 4: Pitch contour for the utterance: "How much variation is there?" from two models within the same speaker.

5.3 Audio Prompt and SNR

To better understand the effect of the audio prompts on the synthesis, we made the audio prompts white noise drawn from different standard deviation σ (then encoded by Q_E). Figure 5 presents the relationship between σ and the signal-to-noise ratio (SNR). We use the WADA-SNR algorithm (**?**) for SNR estimation. From Figure 5, it is clear that the SNR drops as σ increases, confirming that the model is guided by the prompt. Using σ smaller than 10^{-4} effectively increases the SNR compared to not using any audio prompt. All our other experiments are done using $\sigma = 10^{-5}$. We also noticed that VITS has a lower SNR. Perceptually we can hear a small fuzzy noise universally across the syntheses. We conjecture that VITS tries to also model and sample from the environmental noise, which is extremely difficult. The unsuccessful modeling makes it synthesize only a single type of noise.

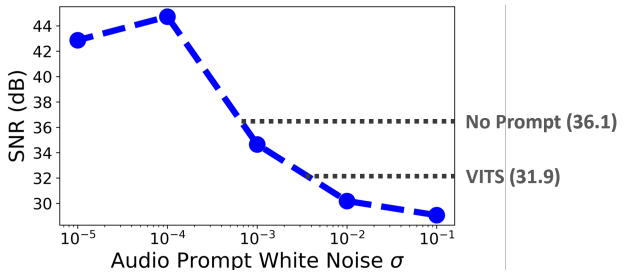


Figure 5: Comparison of SNR with different level of noise as audio prompt. SNR is calculated with 1472 syntheses.

6 Conclusion

On real-world speech, our empirical results indicate multiple discrete codes are preferable to mel-spectrograms for autoregressive synthesizers. And with suitable modeling, MQTTS achieves better performance compared to the non-autoregressive synthesizer. Nonetheless, a sizable gap still exists between our best-performing syntheses and human speech. We believe that bridging this gap is crucial to the development of human-level communication for AI. Acquiring more data is one straightforward solution. In this regard, we are interested in combining and leveraging ASR models to transcribe real-world speech corpora.On the other hand, better modeling can also be designed to mitigate the issues we mentioned in the error analysis. For instance, silence detection can be used in the decoding process to prevent phoneme transitions before the phonation, mitigating deletion errors. Additionally, we plan to further compare and incorporate self-supervised discrete speech and prosody representations with our learned codebooks.

Acknowledgments

We are grateful to Amazon Alexa for the support of this research. We thank Sid Dalmia and Soumi Maiti for the discussions and feedback.

Minima voluptates veniam est delectus quis, non vel voluptatum explicabo reiciendis ad perspiciatis aut veniam sunt saepe?Ad in accusamus, quaerat error expedita ab accusantium iure tenetur?Ex exercitationem ipsa accusamus, tempore dolor suscipit voluptatem error repellendus eveniet quis, molestias asperiores assumenda esse, dolores delectus

exercitationem natus nobis iusto vel eum pariatur necessitatibus quibusdam cum.Nostrum autem molestias quis reiciendis aliquam corrupti et