Investigating the Robustness of Sequence-to-Sequence TTS models to Imperfectly-Transcribed Training Data

THE UNIVERSITY of EDINBURGH

Jason Fong, Pilar Oplustil Gallegos, Zack Hodari, Simon King

1. Motivation: Can seq2seq TTS handle transcription errors?

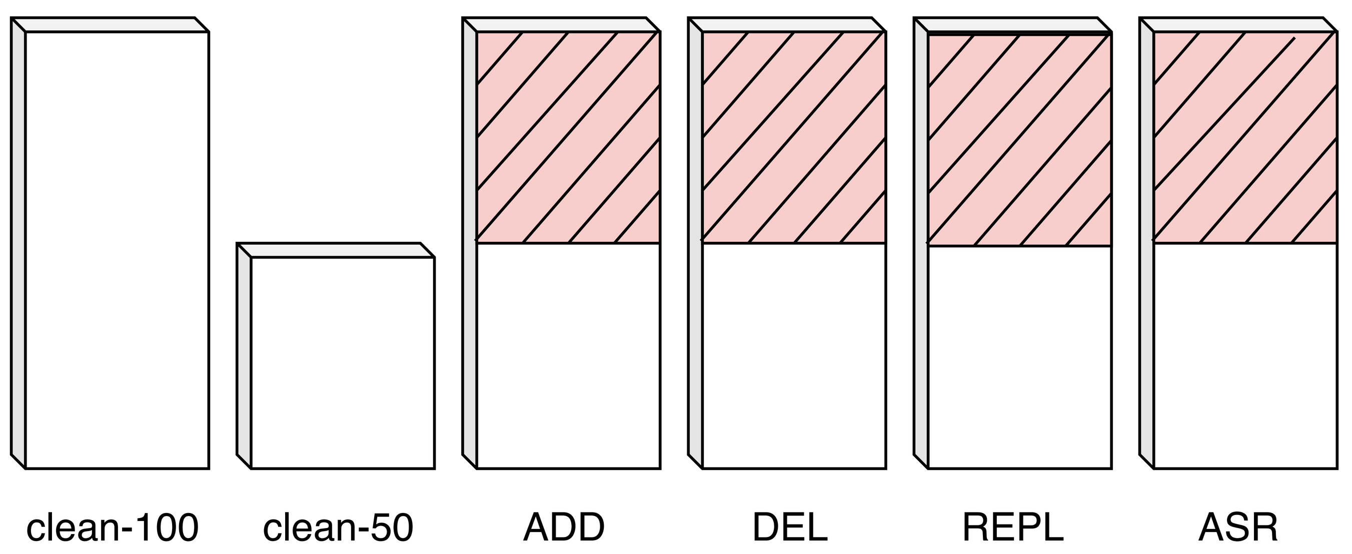
- Seq2seq can generate high quality speech
    - But needs large amounts of data
- Could use found data (i.e. audiobooks)
    - But, then transcription errors are common
- Previous approaches typically excluded such data
- Does seq2seq TTS need such data cleaning?
- **Goal:** Investigate robustness of seq2seq TTS to transcription errors
- **Method:** Train on corrupted transcripts

2. Simulating transcription errors to create our training sets

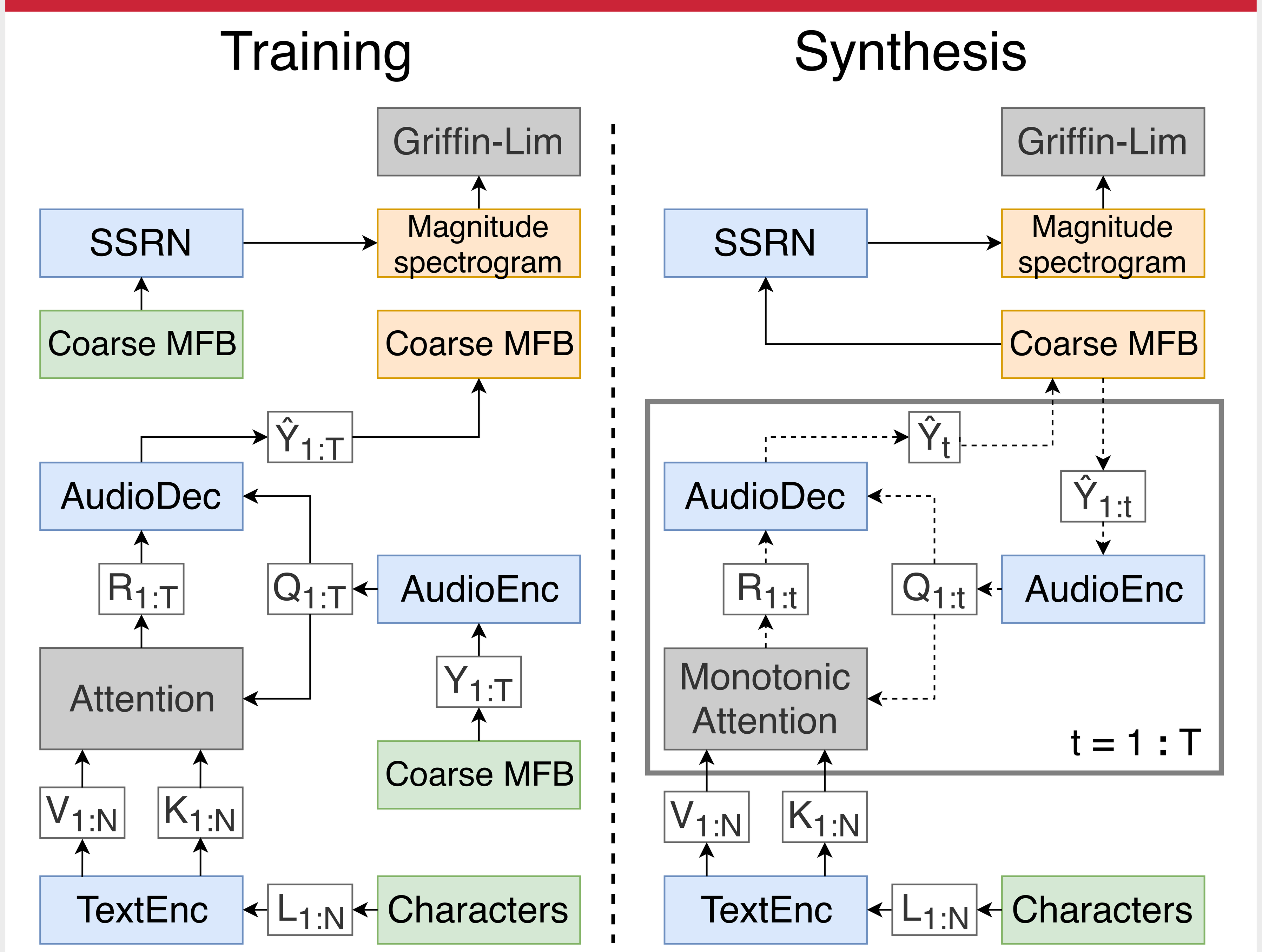
To create corrupted transcripts to train on, we artificially corrupted 50% of the training transcript in 1 of 4 ways (LJSpeech):

Corruption type	Result of corruption
Clean	In being comparatively modern
1. Addition	In being region comparatively sailed modern
2. Deletion	_ being _____________ modern
3. Replacement	Eg being strengthening modern
4. ASR (34.8%WER)	Indie comparatively modeled

We also trained on clean transcripts consisting of 50% and 100% of the original dataset. Thus we had the following training sets:



3. Model architecture: Fully convolutional + autoregressive

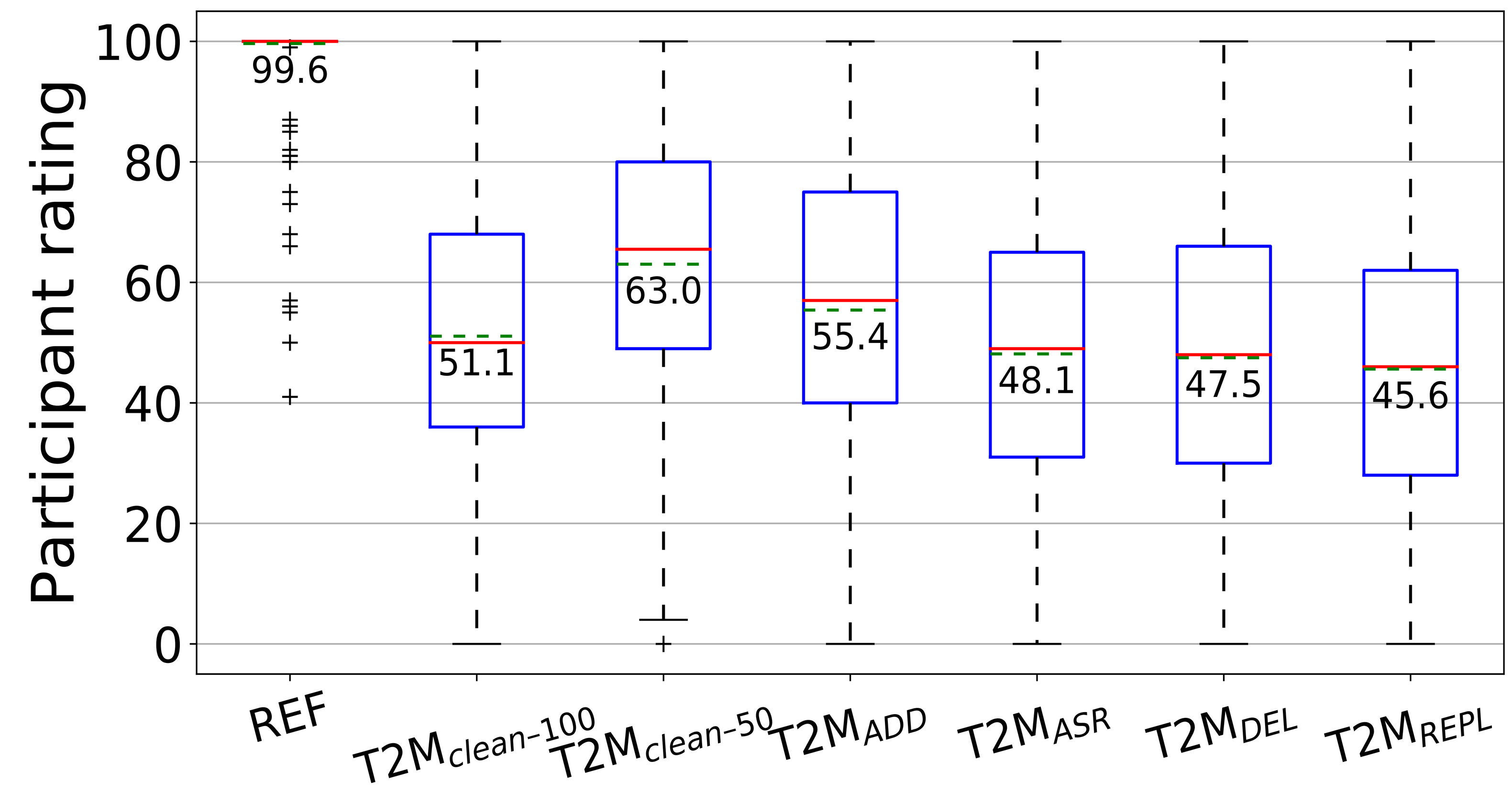


- Seq2seq model: Deep-Convolutional TTS (DCTTS) [1]
- Input is characters and output is coarse mel-spectrogram
- No monotonic attention prior when training because monotonicity must be violated to handle transcription errors

[1] Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention (Tachibana et al., ICASSP 2018)

4. Results: Naturalness and frequency of mispronunciations

MUSHRA-like listening test to evaluate quality of synthesised speech after training on corrupted transcripts



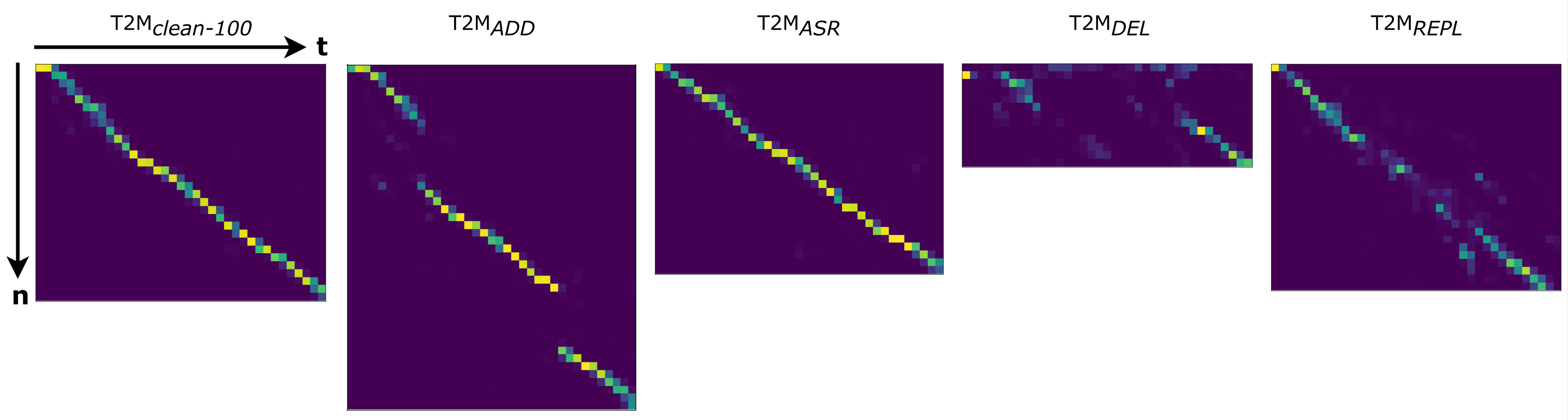
We also counted mispronunciations per system over the test set.

T2M clean–100	T2M clean–50	T2M ADD	T2M ASR	T2M DEL	T2M REPL
45	45	51	34	63	68

Interestingly, the ASR system had the fewest errors, so its lower quality is likely due to overall degradation of its acoustic model.

5. Analysis: How attention dealt with different corruption types

Attention matrices for sentence *"In being comparatively modern"*



- Attention aligns text that has corresponding audio due to teacher-forcing
- Attention skips over extraneous text in the input during training *(ADD)*
- Attention is robust to acoustically-plausible text corruptions *(ASR)*
- **But** attention **not** robust to missing text (audio output that is not explained by any corresponding input) so it attends to all input timesteps with roughly uniform probability *(DEL and REPL)*

6. Conclusion

Takeaways:

- Seq2seq TTS models with attention robust only to certain transcription error types
- Training on transcripts produced by high error rate ASR actually works to some extent
    - Transcribing audio-only data using a low error rate ASR system could be a viable proposition

Further work:

- Make seq2seq models robust to *all* error types
    - Add internal mechanism to detect transcription errors?
- Reproduce results on transcripts with real-world imperfections