3.2 TEXT-TO-SPEECH

For the second experiment we looked at TTS. We used the same single-speaker speech databases from which Google's North American English and Mandarin Chinese TTS systems are built. The North American English dataset contains 24.6 hours of speech data, and the Mandarin Chinese dataset contains 34.8 hours; both were spoken by professional female speakers.

WaveNets for the TTS task were locally conditioned on *linguistic features* which were derived from input texts. We also trained WaveNets conditioned on the logarithmic fundamental frequency ($\log F_0$) values in addition to the linguistic features. External models predicting $\log F_0$ values and phone durations from linguistic features were also trained for each language. The receptive field size of the WaveNets was 240 milliseconds. As example-based and model-based speech synthesis baselines, hidden Markov model (HMM)-driven unit selection concatenative (Gonzalvo et al., 2016) and long short-term memory recurrent neural network (LSTM-RNN)-based statistical parametric (Zen et al., 2016) speech synthesizers were built. Since the same datasets and linguistic features were used to train both the baselines and WaveNets, these speech synthesizers could be fairly compared.

To evaluate the performance of WaveNets for the TTS task, subjective paired comparison tests and mean opinion score (MOS) tests were conducted. In the paired comparison tests, after listening to each pair of samples, the subjects were asked to choose which they preferred, though they could choose "neutral" if they did not have any preference. In the MOS tests, after listening to each stimulus, the subjects were asked to rate the naturalness of the stimulus in a five-point Likert scale score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). Please refer to Appendix B for details.

Fig. 5 shows a selection of the subjective paired comparison test results (see Appendix B for the complete table). It can be seen from the results that WaveNet outperformed the baseline statistical parametric and concatenative speech synthesizers in both languages. We found that WaveNet conditioned on linguistic features could synthesize speech samples with natural segmental quality but sometimes it had unnatural prosody by stressing wrong words in a sentence. This could be due to the long-term dependency of F_0 contours: the size of the receptive field of the WaveNet, 240 milliseconds, was not long enough to capture such long-term dependency. WaveNet conditioned on both linguistic features and F_0 values did not have this problem: the external F_0 prediction model runs at a lower frequency (200 Hz) so it can learn long-range dependencies that exist in F_0 contours.

Table 1 show the MOS test results. It can be seen from the table that WaveNets achieved 5-scale MOSs in naturalness above 4.0, which were significantly better than those from the baseline systems. They were the highest ever reported MOS values with these training datasets and test sentences. The gap in the MOSs from the best synthetic speech to the natural ones decreased from 0.69 to 0.34 (51%) in US English and 0.42 to 0.13 (69%) in Mandarin Chinese.

	Subjective 5-scale MOS in naturalness	
Speech samples	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

Table 1: Subjective 5-scale mean opinion scores of speech samples from LSTM-RNN-based statistical parametric, HMM-driven unit selection concatenative, and proposed WaveNet-based speech synthesizers, 8-bit μ-law encoded natural speech, and 16-bit linear pulse-code modulation (PCM) natural speech. WaveNet improved the previous state of the art significantly, reducing the gap between natural speech and best previous model by more than 50%.

3.3 MUSIC

For out third set of experiments we trained WaveNets to model two music datasets: