- Gaussian process assumption; The conventional generative models are based on Gaussian process (Itakura & Saito, 1970; Imai & Furuichi, 1988; Poritz, 1982; Juang & Rabiner, 1985; Kameoka et al., 2010; Tokuda & Zen, 2015; 2016). From the source-filter model of speech production (Chiba & Kajiyama, 1942; Fant, 1970) point of view, this is equivalent to assuming that a vocal source excitation signal is a sample from a Gaussian distribution (Itakura & Saito, 1970; Imai & Furuichi, 1988; Poritz, 1982; Juang & Rabiner, 1985; Tokuda & Zen, 2015; Kameoka et al., 2010; Tokuda & Zen, 2016). Together with the linear assumption above, it results in assuming that speech signals are normally distributed. However, distributions of real speech signals can be significantly different from Gaussian.

Although these assumptions are convenient, samples from these generative models tend to be noisy and lose important details to make these audio signals sounding natural.

WaveNet, which was described in Section 2, has none of the above-mentioned assumptions. It incorporates almost no prior knowledge about audio signals, except the choice of the receptive field and μ-law encoding of the signal. It can also be viewed as a non-linear causal filter for quantized signals. Although such non-linear filter can represent complicated signals while preserving the details, designing such filters is usually difficult (Peltonen et al., 2001). WaveNets give a way to train them from data.

B Details of TTS Experiment

The HMM-driven unit selection and WaveNet TTS systems were built from speech at 16 kHz sampling. Although LSTM-RNNs were trained from speech at 22.05 kHz sampling, speech at 16 kHz sampling was synthesized at runtime using a resampling functionality in the Vocaine vocoder (Agiomyrgiannakis, 2015). Both the LSTM-RNN-based statistical parametric and HMM-driven unit selection speech synthesizers were built from the speech datasets in the 16-bit linear PCM, whereas the WaveNet-based ones were trained from the same speech datasets in the 8-bit μ-law encoding.

The linguistic features include phone, syllable, word, phrase, and utterance-level features (Zen, 2006) (e.g. phone identities, syllable stress, the number of syllables in a word, and position of the current syllable in a phrase) with additional frame position and phone duration features (Zen et al., 2013). These features were derived and associated with speech every 5 milliseconds by phone-level forced alignment at the training stage. We used LSTM-RNN-based phone duration and autoregressive CNN-based $\log F_0$ prediction models. They were trained so as to minimize the mean squared errors (MSE). It is important to note that no post-processing was applied to the audio signals generated from the WaveNets.

The subjective listening tests were blind and crowdsourced. 100 sentences not included in the training data were used for evaluation. Each subject could evaluate up to 8 and 63 stimuli for North American English and Mandarin Chinese, respectively. Test stimuli were randomly chosen and presented for each subject. In the paired comparison test, each pair of speech samples was the same text synthesized by the different models. In the MOS test, each stimulus was presented to subjects in isolation. Each pair was evaluated by eight subjects in the paired comparison test, and each stimulus was evaluated by eight subjects in the MOS test. The subjects were paid and native speakers performing the task. Those ratings (about 40%) where headphones were not used were excluded when computing the preference and mean opinion scores. Table 2 shows the full details of the paired comparison test shown in Fig. 5.