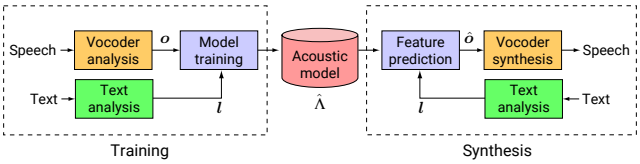

Figure 6: Outline of statistical parametric speech synthesis.

Zen, Heiga, Tokuda, Keiichi, and Kitamura, Tadashi. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic features. *Comput. Speech Lang.*, 21(1):153–173, 2007.

Zen, Heiga, Tokuda, Keiichi, and Black, Alan W. Statistical parametric speech synthesis. *Speech Commn.*, 51(11):1039–1064, 2009.

Zen, Heiga, Senior, Andrew, and Schuster, Mike. Statistical parametric speech synthesis using deep neural networks. In *Proc. ICASSP*, pp. 7962–7966, 2013.

Zen, Heiga, Agiomyrgiannakis, Yannis, Egberts, Niels, Henderson, Fergus, and Szczepaniak, Przemysław. Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices. In *Interspeech*, 2016. URL `https://arxiv.org/abs/1606.06061`.

A TEXT-TO-SPEECH BACKGROUND

The goal of TTS synthesis is to render naturally sounding speech signals given a text to be synthesized. Human speech production process first translates a text (or concept) into movements of muscles associated with articulators and speech production-related organs. Then using air-flow from lung, vocal source excitation signals, which contain both periodic (by vocal cord vibration) and aperiodic (by turbulent noise) components, are generated. By filtering the vocal source excitation signals by time-varying vocal tract transfer functions controlled by the articulators, their frequency characteristics are modulated. Finally, the generated speech signals are emitted. The aim of TTS is to mimic this process by computers in some way.

TTS can be viewed as a sequence-to-sequence mapping problem; from a sequence of discrete symbols (text) to a real-valued time series (speech signals). A typical TTS pipeline has two parts; 1) text analysis and 2) speech synthesis. The text analysis part typically includes a number of natural language processing (NLP) steps, such as sentence segmentation, word segmentation, text normalization, part-of-speech (POS) tagging, and grapheme-to-phoneme (G2P) conversion. It takes a word sequence as input and outputs a phoneme sequence with a variety of linguistic contexts. The speech synthesis part takes the context-dependent phoneme sequence as its input and outputs a synthesized speech waveform. This part typically includes prosody prediction and speech waveform generation.

There are two main approaches to realize the speech synthesis part; non-parametric, example-based approach known as concatenative speech synthesis (Moulines & Charpentier, 1990; Sagisaka et al., 1992; Hunt & Black, 1996), and parametric, model-based approach known as statistical parametric speech synthesis (Yoshimura, 2002; Zen et al., 2009). The concatenative approach builds up the utterance from units of recorded speech, whereas the statistical parametric approach uses a generative model to synthesize the speech. The statistical parametric approach first extracts a sequence of vocoder parameters (Dudley, 1939) $\mathbf{o} = \{\mathbf{o}_1, \ldots, \mathbf{o}_N\}$ from speech signals $\mathbf{x} = \{x_1, \ldots, x_T\}$ and linguistic features $\mathbf{l}$ from the text W, where N and T correspond to the numbers of vocoder parameter vectors and speech signals. Typically a vocoder parameter vector $\mathbf{o}_n$ is extracted at every 5 milliseconds. It often includes cepstra (Imai & Furuichi, 1988) or line spectral pairs (Itakura, 1975), which represent vocal tract transfer function, and fundamental frequency (F_0) and aperiodicity (Kawahara et al., 2001), which represent characteristics of vocal source excitation signals. Then a set of generative models, such as hidden Markov models (HMMs) (Yoshimura, 2002), feed-forward neural networks (Zen et al., 2013), and recurrent neural networks (Tuerk & Robinson, 1993; Karaali et al., 1997; Fan et al., 2014), is trained from the extracted vocoder parameters and linguistic features