as

$$\hat{\Lambda} = \arg\max_{\Lambda} p\left(\mathbf{o} \mid \mathbf{l}, \Lambda\right), \tag{4}$$

where Λ denotes the set of parameters of the generative model. At the synthesis stage, the most probable vocoder parameters are generated given linguistic features extracted from a text to be synthesized as

$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}} p(\mathbf{o} \mid \mathbf{l}, \hat{\Lambda}). \tag{5}$$

Then a speech waveform is reconstructed from $\hat{\mathbf{o}}$ using a vocoder. The statistical parametric approach offers various advantages over the concatenative one such as small footprint and flexibility to change its voice characteristics. However, its subjective naturalness is often significantly worse than that of the concatenative approach; synthesized speech often sounds muffled and has artifacts. Zen et al. (2009) reported three major factors that can degrade the subjective naturalness; quality of vocoders, accuracy of generative models, and effect of oversmoothing. The first factor causes the artifacts and the second and third factors lead to the muffleness in the synthesized speech. There have been a number of attempts to address these issues individually, such as developing high-quality vocoders (Kawahara et al., 1999; Agiomyrgiannakis, 2015; Morise et al., 2016), improving the accuracy of generative models (Zen et al., 2007; 2013; Fan et al., 2014; Uria et al., 2015), and compensating the oversmoothing effect (Toda & Tokuda, 2007; Takamichi et al., 2016). Zen et al. (2016) showed that state-of-the-art statistical parametric speech syntheziers matched state-of-the-art concatenative ones in some languages. However, its vocoded sound quality is still a major issue.

Extracting vocoder parameters can be viewed as estimation of a generative model parameters given speech signals (Itakura & Saito, 1970; Imai & Furuichi, 1988). For example, linear predictive analysis (Itakura & Saito, 1970), which has been used in speech coding, assumes that the generative model of speech signals is a linear auto-regressive (AR) zero-mean Gaussian process;

$$x_t = \sum_{p=1}^{P} a_p x_{t-p} + \epsilon_t \tag{6}$$

$$\epsilon_t \sim \mathcal{N}(0, G^2) \tag{7}$$

where a_p is a p-th order linear predictive coefficient (LPC) and G^2 is a variance of modeling error. These parameters are estimated based on the maximum likelihood (ML) criterion. In this sense, the training part of the statistical parametric approach can be viewed as a two-step optimization and sub-optimal: extract vocoder parameters by fitting a generative model of speech signals then model trajectories of the extracted vocoder parameters by a separate generative model for time series (Tokuda, 2011). There have been attempts to integrate these two steps into a single one (Toda & Tokuda, 2008; Wu & Tokuda, 2008; Maia et al., 2010; Nakamura et al., 2014; Muthukumar & Black, 2014; Tokuda & Zen, 2015; 2016; Takaki & Yamagishi, 2016). For example, Tokuda & Zen (2016) integrated non-stationary, nonzero-mean Gaussian process generative model of speech signals and LSTM-RNN-based sequence generative model to a single one and jointly optimized them by back-propagation. Although they showed that this model could approximate natural speech signals, its segmental naturalness was significantly worse than the non-integrated model due to over-generalization and over-estimation of noise components in speech signals.

The conventional generative models of raw audio signals have a number of assumptions which are inspired from the speech production, such as

- Use of fixed-length analysis window; They are typically based on a stationary stochastic process (Itakura & Saito, 1970; Imai & Furuichi, 1988; Poritz, 1982; Juang & Rabiner, 1985; Kameoka et al., 2010). To model time-varying speech signals by a stationary stochastic process, parameters of these generative models are estimated within a fixed-length, overlapping and shifting analysis window (typically its length is 20 to 30 milliseconds, and shift is 5 to 10 milliseconds). However, some phones such as stops are time-limited by less than 20 milliseconds (Rabiner & Juang, 1993). Therefore, using such fixed-size analysis window has limitations.
- Linear filter; These generative models are typically realized as a linear time-invariant filter (Itakura & Saito, 1970; Imai & Furuichi, 1988; Poritz, 1982; Juang & Rabiner, 1985; Kameoka et al., 2010) within a windowed frame. However, the relationship between successive audio samples can be highly non-linear.