Statistical parametric speech synthesis

While concatenative synthesis can reach highly natural synthesized speech, the approach is inherently limited by properties of the speech corpus used for the unit selection process. Concatenative systems can only produce speech whose constituent segments (e.g., diphones) have been pre-recorded. In order to make the synthesis sound natural, large amounts of speech from a single speaker must therefore be available. This limits the flexibility of concatenative systems in producing different voices, speaking styles, emotional expressions, or other modifications to the sound that are common in everyday human communication.

As an alternative to the concatenative approach, statistical parametric speech synthesis (SPSS) is another TTS approach that has become highly popular in the speech technology field. This is because it addresses the main limitation of the concatenative systems — the lack of flexibility — by generating the speech using statistical models of speech instead of relying on pre-recorded segments. These statistical models are learned from speech corpora using machine learning techniques, and they encode information of how speech evolves as a function of time in the context of a given input text. In this respect, SPSS systems can be viewed as a mirror image of ASR systems: while an ASR system tries to convert speech from acoustic features to a string of words using machine learning models, an SPSS system tries to convert a string of words into acoustic features or directly to the acoustic waveform using machine learning models. Both ASR and SPSS systems are typically trained on a large amount of speech data with their transcriptions, resulting in a set of parameters that describe statistical characteristics of the speech data (hence "statistical parametric" speech synthesis).

A full SPSS system consists of text analysis, feature generation, and waveform generation modules. The classical approach to SPSS is based on a combination of a hidden-Markov model Gaussian mixture model (HMM-GMM) architecture for feature generation and a vocoder for waveform generation, and these will be discussed in more detail below. Recent advances in neural network-based SPSS are then reviewed at the end.

**Feature generation**

Given the linguistic description of the text-to-be-synthesized, the purpose of the feature generation is to transform the linguistic features into a corresponding description of the acoustic signal. Similarly to ASR, this component mediating the two levels is called an acoustic model. Technically speaking, the acoustic model converts the linguistic input into a series of acoustic features at a fixed frame-rate (e.g., one feature frame every 10 ms) using a probabilistic mapping between the two. The mapping is learned from a training speech corpus.

A standard approach for the probabilistic mapping has been to use a HMM-GMM as the statistical parametric model. Similarly to an HMM-GMM ASR system, the states of the HMM correspond to parts of subword units (e.g., parts of a phone, diphone, or triphone). Transition probabilities $P(s \mid s_{t-1})$ between the states describe how the speech evolves through each subword unit and from a unit to another. Acoustic characteristics associated with each state are modeled with a GMM, where the GMM describes a probability distribution $P(y \mid s)$ over the possible acoustic feature vectors in that state. Given a sequence of desired subword units (as instructed by the linguistic features), the model can be stochastically or deterministically sampled to produce a sequence of acoustic features. These are then fed to a waveform generation module to produce the actual speech signal. In the most basic form, self-transitions from an HMM state to itself account for the duration spent in that state (i.e., how many frames should the same acoustic content be repeated). However, separate more advanced duration models are often used to overcome the limitations of a first-order Markov chain in modeling the temporal dependencies and durational characteristics of speech.
A visual illustration of HMM-GMM-based speech feature generation. State sequence $s = \{s_1, s_2, ..., s_{10}\}$ required for word “cat” (/k æ t/) is shown on top, where each phoneme consists of three states: initial, center and final state (e.g., $k_1$, $k_2$, and $k_3$). Each state is associated with an $N$-dimensional Gaussian mixture model (GMM), where $N$ is the dimensionality of the speech features. At each time step, the GMM of the active state is sampled for a feature vector $y_t$. After this, a state transition can occur to a next state or back to the current state, controlling the durational aspects of the speech.

### State chain: $P(s_t | s_{t-1})$

$$P(y | s) = \sum_{k=1}^{K} \phi_{k,S} N(y | \mu_{k,S}, \Sigma_{k,S})$$

A typical high-quality speech waveform consists of "continuous" (e.g., 16-bit quantized) amplitude values sampled at 16 kHz. In addition, the shape of the waveform is affected by several factors that do not directly contribute to the naturalness or intelligibility of speech, such as signal gain or phase and amplitude characteristics of the recording and transmission chain. This means that mere 80 milliseconds of a raw waveform — a typical length of one vowel — would correspond to $0.08 \times 16 \times 1280 = 1280$-dimensional amplitude vector, and that this vector could take countless of shapes for perceptually highly similar sounds. Moreover, the values encoded in this vector would be highly correlated with each other (see LPC). Given the high dimensionality, variability, and redundant nature of the waveform signal representation, it is not an attractive target for statistical parametric modeling with classical machine learning techniques (but see also Neural SPSS below).

However, as we remember from speech feature extraction (see, e.g., SFFT), speech signal can be considered as quasi-stationary in short windows of approx. 10–30 ms in duration. Speech contents of the signal within these short windows can be described using a set of spectral and source features (such as MFCCs and F0) that are assumed to be fixed for that window. When extracting the features in a sliding window with short (e.g., 10 ms) window steps, the overall structure of the signal can be captured with a much lower dimensional and less variable representation than what the actual waveform would be. A vocoder, then, is an algorithm that can 1) parametrize a speech waveform into a more compact set of descriptive features as a function of time, but also to 2) synthesize the speech from the features with minimal loss in speech quality. In addition, many vocoders use features that are interpretable in terms of speech production or speech acoustics, enabling analysis and manipulation of the speech signal to observe or cause certain phenomena in the speech signal.

Compactness and invariance of the acoustic signal representation is also why vocoding is used in SPSS systems: instead of generating the speech waveform directly, the feature generation module first generates a lower-dimensional set of vocoder features that characterize the speech signal with its essential properties. A vocoder then takes these features as input and generates the corresponding waveform using a series of signal processing operations. These operations are essentially an inverse of the original feature extraction process, combined with some additional mechanisms for re-introducing (or inventing) information lost during the feature extraction process (such as signal phase that is discarded in standard spectral features).

For instance, when using the popular STRAIGHT vocoder (Kawahara et al., 1999), the HMM-GMM model first generates a sequence of feature vectors that encode spectral envelope, F0, and periodicity characteristics of the speech signal to-be-produced, as instructed by the text analysis module. These features are then fed to STRAIGHT that synthesizes the final speech waveform based on the features.
A schematic view of a vocoder and typical uses for vocoder features.

When used as a part of an SPSS system, vocoder features are generated by the parametric statistical model during the synthesis process.

**SPSS system training**

Training of an SPSS system refers to estimation of the parametric acoustic model (e.g., a HMM-GMM) that is responsible for mapping the linguistic features to the corresponding waveform generation (vocoder) features. This is achieved using a corpus of speech data, where each utterance comes with the corresponding text of what was said, and optionally with phonetic annotation describing the phonetic units and their temporal positions in the waveform. First, the text analysis module is used to create linguistic features of a training utterance while the vocoder is used to extract vocoder features from the corresponding speech waveform. Then the statistical model is trained to minimize prediction error of the given vocoder features when the linguistic features are used as inputs. Access to phonetic annotation allows more accurate temporal alignment between the linguistic features and the speech signal. Since the widely utilized HMM architecture for acoustic modeling is not ideal for modeling speech segment durations, a separate duration model is often trained to align the linguistic features (which are agnostic of speaking rate and rhythm in the actual speech data) with the phonetic units realized in the acoustic speech signal. In the figure below, both the acoustic model and the duration model are denoted jointly by the parametric statistical model block.

**Advantages and disadvantages of the HMM-GMM SPSS compared to concatenative synthesis**

Since the “instructions” for speech generation are encoded by parameters of the SPSS model, the model can easily be adapted to produce speech with different characteristics. For instance, the vocal tract characteristics of the training speaker are encoded by the means and variances of the Gaussian distributions in each HMM state whereas durational characteristics are encoded by the transition probability matrix of the HMM. Therefore, the system can be adapted to other speakers by simply adapting the pre-trained HMM-GMM using speech from a new talkers. In this case, standard techniques such as Maximum-a-posteriori (MAP) adaptation or Maximum-likelihood linear regression (MLLR) can be used to update the model parameters. In addition, since the parameters of the HMM-GMM are often interpretable in terms of speech spectral envelope or phonation characteristics, it is possible to either modify the models or to post-process the resulting acoustic features in order to achieve desired effects. For example, changing of the speech pitch can be done by simply adjusting the F0 parameter, whereas reduction of some synthesis artifacts such as muffled sound quality due to statistical averaging can be attempted by adjusting the GMM parameters with a chosen transformation.

The potential disadvantages of the statistical approach include sound quality issues (e.g., muffledness) due to statistical smoothing taking place in a stochastic generative model, sound quality of the used vocoders, and potential problems in robust statistical model estimation from finite data.

**Neural SPSS**
Recent advances in artificial neural networks (ANNs) have also led to new developments in speech synthesis (SPSS) beyond the classical HMM-GMM framework. In terms of vocoding, WaveNet (van den Oord et al., 2016) is a highly influential neural network waveform generator that can produce high-quality speech. It is based on an autoregressive convolutional neural network (CNN) architecture and it operates directly on the speech waveforms. Given a history of previous waveform samples and some “conditioning” information on what type of signal should be produced, the model predicts the most likely next speech waveform sample at each time step. For instance, WaveNet can be trained to produce speech from spectral features such as log-Mel energies and F0 information. Although WaveNet can reach near-human naturalness of the produced speech (with certain limitations), waveform-level autoregressive processing is also computationally extremely expensive. Development of computationally flexible high-quality neural vocoders is therefore still an active research area.

In addition to vocoding, neural networks have become commonplace replacements for the HMM-GMM in the feature generation stage. For instance, deep feed-forward networks or LSTMs can be utilized in the feature generation. Since LSTMs are especially good at modeling temporal dependencies, they can theoretically handle larger temporal ambiguity and variability between the input linguistic specifications and the target vocoder features.

Given sufficient training data, it is also possible to implement the entire chain from written text to the synthesized waveform using a neural network system. Tacotron 2 is an example of such a system, where the input text is processed by a sequence-to-sequence ANN model to directly create a log-Mel spectrogram corresponding to the input text (i.e., without a dedicated text analysis module). The spectrogram is then fed to the WaveNet module (see above) to produce the speech signal. As a result, Tacotron 2 and the WaveNet vocoder can together achieve highly impressive speech quality. The advantage of these type end-to-end approaches is that there are fewer assumptions regarding what kind of intermediate representations are good for the task at hand, reducing the risk that the pre-specified operations and representations cause a loss of relevant information in the pipeline. There is also no need for deep understanding of the linguistic structure underlying written and spoken language or access to pre-existing text analysis tools, making deployment of the systems possible for any language with sufficient training data (text and corresponding speech). Since all the components are based on differentiable neural network operations, it is also possible to jointly optimize the entire chain from waveform generation to text processing. In principle, neural SPSS systems are also highly flexible, as basically any type of side information can be injected to the system to adjust the characteristics of the produced speech.

Neural systems, however, also have their drawbacks. The amount of data and computational resources required to train these systems can be high. Runtime computational requirements of neural vocoders may also be problematic in some applications, although recent advances in vocoders and in parallelization of the computations has already led to significant advances in this respect. Another issue is the lack of interpretability and transparency of model parameters: while parameters of classical models such as HMMs and GMMs have relatively clear relationship with what are the inputs and outputs of the system, the same is not true for ANNs with multiple layers. This makes it much more difficult to understand the behavior of the model, especially when trying to overcome problems in model performance. Lack of transparency and interpretability also means that manual control of characteristics of the produced speech is more difficult. Finally, adaptation of the models to new data (e.g., a new speaker or speaking style) cannot make use of the well-understood mathematical solutions available to the classical models. In contrast, the design, training, and adaptation of ANNs are much more heuristics-driven, similarly to the use of ANNs in any other machine learning domain.

Further reading


