Advances in Spectral Parameterization for Statistical (HMM-Based) TTS

Slava Shechtman, Alex Sorin
IBM Research, Haifa Research Lab, Israel
slava@il.ibm.com, sorin@il.ibm.com

Abstract

HMM-based parametric speech synthesis has recently become an alternative to the concatenative TTS approach, especially when low footprint and general speech domain are required. A majority of speech parameterization models used in state-of-the-art HMM TTS systems employ source-filter waveform synthesis schemes. Sinusoidal representation and waveform generation of speech is an alternative to the source-filter model, which is successfully applied in speech coding, unit-selection TTS and voice conversion, but rarely used for HMM TTS systems. In this paper we utilize Regularized Cepstral Coefficients (RCC) estimated in mel-frequency scale for sinusoidal amplitude spectrum envelope modeling within an HMM-based TTS framework. Improved subjective quality for mel-frequency RCC (MRCC) combined with the sinusoidal model based reconstruction is reported and compared to the state-of-the-art MGC-LSP parameters.

Index Terms: speech synthesis, HMM-based TTS, sinusoidal model, speech parameterization

1. Introduction

Statistical (HMM-based) speech synthesis (STTS) contrasts with its more mature predecessor, the concatenative TTS (CTTS) in many key aspects. While CTTS aims to combine natural speech waveform units with as little modification as possible, STTS generates the most likely speech frames using some pre-trained statistical model in some parametric domain. As such, STTS systems rely heavily on signal manipulation; speech waveforms are decomposed into speech parameters, modeled and reconstructed. In terms of speech quality, concatenative systems produce variable quality speech, which at its best is highly natural. STTS systems have more consistent speech quality which is still far from natural speech quality. However, the generated speech is smooth and stable and the systems show good predictable behavior with respect to unseen contexts.

In Figure 1 a block diagram of the HMM-based speech synthesis system is presented [1][8]. The training part performs training of context-dependent multi-stream (e.g. spectrum, F0) HMM models, where both linguistic and prosodic context is taken into account. Each of the parametric streams is clustered individually by using context decision trees. Context-dependent Gaussian distributions for HMM state-durations are typically estimated at the final stages of the training, and also clustered individually by using context decision trees.

During the synthesis, a given word sequence is converted into a sequence of linguistic and acoustic context labels, and then the utterance HMM \( \tilde{\lambda} = \{ \tilde{\lambda}_1, ..., \tilde{\lambda}_C \} \) is constructed by concatenating the context-dependent HMMs according to the label sequence. The HMM state durations are then estimated from the corresponding context dependent state-duration distributions, thus defining the state-sequence of HMM states for the whole utterance \( \tilde{\mathbf{q}} = \{ q_1, ..., q_T \} \). In the typical case of single multi-variate Gaussian state-output distributions, \( N(\mathbf{o}_t; \mu, \Sigma) \), the ML parameter generation is given by:

\[
\hat{o} = \arg \max \left\{ N(\mathbf{o}; \hat{\mu}_q, \Sigma_q) \right\}
\]

To facilitate the smooth parameter generation, the dynamic features are introduced as a part of a state-output vector \( \mathbf{o}_t \):

\[
\mathbf{o}_t = [c^t, \Delta c^t, \Delta^2 c^t] \triangleq \mathbf{Wc}_t,
\]

where \( \mathbf{W} \) is a matrix that constructs augmented feature vectors \( \mathbf{c}_t \) from static features \( \mathbf{c}_j \). In that case the minimization (1) can be written as:

\[
\hat{c} = \arg \max \left\{ N(\mathbf{Wc}; \hat{\mu}_q, \Sigma_q) \right\}
\]

and determined by a solution of the following set of linear equations:

\[
W^T \Sigma_q^{-1} \mathbf{W} \hat{c} = W^T \Sigma_q^{-1} \mu_q
\]

Finally, a speech waveform is synthesized from the generated spectral and excitation parameters.

Figure 1: Block diagram of HMM-based speech synthesis system [1]

A majority of speech parameterization models used in state-of-the art HMM TTS systems imply source-filter waveform synthesis schemes [1]. Sinusoidal representation and waveform generation of speech [2] is an alternative for the source-filter model and is successfully applied in speech coding [2], unit-selection TTS [3][4] and voice morphing [5]. Few works are related to its usage for HMM TTS [6][7]. In [6] a pitch synchronous approach is taken. To fit this approach to the constant frame modeling of the HMM based TTS system, an additional time interpolation of HMM TTS output parameters is required. Banos et al. propose in [7] a constant
frame harmonic/stochastic modeling. However, the harmonic amplitudes at synthesis time are obtained by sampling of MGC-LSP spectrum [8], which is not optimized for sinusoidal waveform synthesis.

In this work we utilize an alternative amplitude parameterization [9] of the constant frame sinusoidal model [5], featuring a constant number of frame parameters and suitable for HMM-based TTS. We parameterize a continuous amplitude spectrum (a.k.a. spectral envelope) by Regularized Cepstral Coefficients (RCC) in perceptual frequency scale [9], with jointly optimized selection of appropriate frequency warping and regularization constant, determined experimentally. The performance of the proposed spectral parameters (Mel Regularized Cepstral Coefficients, or MRCC) is compared to the state-of-the-art MGC-LSP parameterization within HMM-based TTS parameter generation [8], followed by the sinusoidal waveform generation [7].

The paper is organized as follows. First, we review a constant frame sinusoidal representation, capable of high quality speech reconstruction and pitch modification [5]. Second, we review the Regularized Cepstral Coefficient representation for sinusoidal amplitudes [9] and determine experimentally the appropriate frequency warping and the regularization constant for the best reconstruction quality. Finally, the full HMM-based TTS system is described, and experimental results are presented and discussed.

2. Sinusoidal model parameterization

2.1. Sinusoidal model analysis

Many speech production models represent voiced speech by a sum of quasi periodic (i.e. harmonic) and noise-like signals [3][10]. Stationary sinusoidal modeling is widely used to describe the harmonic part of voiced speech, due to its simplicity and accuracy [2][3]. With slight extensions (e.g. frequency jittering [5]), it is capable of high quality synthesis of semi voiced speech signal as well.

The sinusoidal model analysis may be either pitch-synchronous [3][6] or pitch-asynchronous [5][7], i.e. having constant frame update rate. The former requires thorough determination of pitch coherent analysis window centers [3], while the latter requires precise frame alignment procedure during the synthesis [5]. In the current work we adopt a constant frame sinusoidal model, for which the model is updated at a fixed rate (e.g., 200Hz), since this approach is more robust and can be easily incorporated into the HMM-based TTS framework. This model is capable of high quality speech reconstruction (MOS = 4.2) and modification [5]. Here we outline the system that was successfully used for wideband speech reconstruction and voice transformations [5].

The sinusoidal model representation of an analysis window extracted from a speech waveform is given by:

\[ s_w(n) = W(n) \sum_{k=0}^{K} A_k \sin(\theta_k n + \phi_k), \]

where \( W(n) \) is a symmetric window, e.g. Hamming or Hann, \( \{A\} \) and \( \{\phi\} \) are harmonic amplitudes and phases correspondingly and \( \theta_k \) is the position of the highest local maximum found on the short time amplitude spectrum \( S_w(\theta) \) in a close vicinity of \( \theta_k, \) i.e. the k-th multiple of the angular pitch frequency \( \theta_p \). Equation (5) is equivalent to the frequency domain equation (6)

\[ S_\omega(\theta) = S_w(\theta) = \sum_{k=-L}^{L} A_k e^{j\theta_k} W(\theta - \theta_k), \]

where \( S_\omega \) is a short time spectrum and \( W(\theta) \) is the Fourier transform of \( w(n) \). The vector \( \{e^{j\theta_k}\} \) is referred to as line spectrum.

Once the harmonic frequencies are determined, the line spectrum estimate can be obtained by minimizing a spectrum approximation error:

\[ E = \sum_{m=0}^{M-1} |\hat{S}_\omega(\theta_m) - \hat{S}_w(\theta_m)|^2, \]

where \( M \) is half of the FFT length. The minimization is accomplished by solving an over-determined set of linear equations [5].

Besides a pitch detection, the detection of a maximal voicing frequency [5][10] is also desirable during the synthesis, to avoid buzziness within voiced regions.

Reasonable approximation of the line spectra evolving in time is obtained with a sequence of constant length analysis windows. A more accurate approximation is achieved by: a) using analysis window of pitch-dependent length (e.g. constant number of pitch periods in voiced regions and predefined length in unvoiced regions); and b) by centering the voiced analysis windows at pitch marks. We estimate the pitch marks as the positions of high peaks of the speech waveform located at approximately one pitch period distance from each other.

The estimation procedure described above is carried out for fully voiced and semi voiced frames only. For pure unvoiced frames the short-time Fourier transform (STFT) is used as the line spectrum estimate.

Even though the explicit phase values are important for achieving high quality speech representation [3], in the framework of the current HMM-based TTS application the original phase is not transmitted, but rather estimated from the amplitude spectrum, as described in section 2.3.

2.2. Sinusoidal model synthesis

During the synthesis of voiced frames, the harmonic frequencies are estimated as

\[ \theta_k = \begin{cases} \theta_k, k \leq k_v \\ \theta_k + \Delta \theta, k > k_v \end{cases} \]

where \( \Delta \theta \) is a random frequency jitter [5] and \( k_v \) is defined by a maximal voicing frequency, either constant [6] or estimated per frame [5][10].

In order to achieve smooth evolution of the synthesized waveform in time it is crucial to perform phase alignment of the consecutive line spectra [5]. After the phase alignment stage the windowed waveforms are reconstructed in the frequency domain according to (6), and overlap-and-added in the time domain.

2.3. Line spectrum parameterization (MRCC)

The line spectrum can be interpreted as a result of sampling of a continuous complex spectral envelope at the harmonic frequencies. Once the line spectrum estimate is obtained, it is possible to calculate the complex spectral envelope at any frequency using an appropriate interpolative model. However, each frame has a large and variable number of line spectrum parameters. For the purposes of speech modeling and pitch modification, it is desirable to have a robust representation of the continuous spectrum with a constant (and reduced) number of parameters.

Several techniques of spectral envelope parameterization, based on sinusoidal amplitudes, were developed in the past.
They are based on various parameters, but share in common a prerequisite stage of harmonic amplitude extraction and a utilization of error criteria with respect to the amplitudes approximation.

The Regularized Cepstral Coefficients (RCC) [9] parameterization was successfully applied for quantization of the sinusoidal model amplitudes [11]. The RCC coefficients \( d_i \) represent a continuous amplitude spectrum \( B(\theta) \):

\[
a \triangleq \left\{ \log |B(\theta)| \right\}_{i=0}^L = d + 2\sum_{i=0}^L d_i \cos(\theta_i) \triangleq \text{Md},
\]

where \( \text{M} \) is a real cepstrum transform matrix [9] and \( p \) is the order of the cepstrum. The RCC parameters are obtained by the minimization of the composite criterion (10):

\[
\epsilon = \sum_{i=0}^L \log A_i - \log |B(\theta_i)| + \lambda \Re(B(\theta)),
\]

where the first term is the squared error between the sinusoidal amplitudes and the modeled spectrum sampled at the pitch harmonics while the second is a log-spectral smoothness regularization term [9] and \( \lambda \) is the control parameter. The minimization of (10) with respect to cepstral coefficients \( d_i \) in (9) results in:

\[
d = \left[ M' + \lambda R \right]^{-1} M'a.
\]

where \( R \) is a regularization diagonal matrix [9].

Frequency warping is a common practice to improve the perceptual quality of speech envelope parameters [12]. When performing parameterization in a warped scale, the lower band is approximated at a higher frequency resolution than the higher band in accordance with the human perception of audio signals. A normalized frequency warping is defined by a concave bijective function \( f(\theta) \)

\[
\hat{\theta} = f(\theta), 0 \leq \theta, \hat{\theta} \leq \pi
\]

Several warping functions are widely used in speech processing, e.g. mel-scale or bark-scale [12] transformations. Alternatively, a bilinear parametric frequency warping [12] may be used:

\[
f(\theta) = \arctan \left( \frac{1 - \alpha^2}{1 + \alpha^2} \sin \theta - 2\alpha \right),
\]

where \( \alpha \) is a warping control parameter.

In [9] it was proposed to apply bark-scale warping prior to the RCC solution (11), combined with \( \lambda = 5e^{-4} \), however, alternative warping functions were not examined. The selection of warping function implies appropriate selection of the regularization constant \( \lambda \). The stronger the warping is, the higher \( \lambda \) should be, in order to prevent unstable and diverged behavior of the continuous spectrum at the low frequency region. On the other hand, increasing the \( \lambda \) value increases the original line spectrum reconstruction error, given by the first term in (10), and may gradually reduce the synthesized speech quality. That is why the warping function and the corresponding regularization constant have to be jointly selected.

To select the regularization constant and the frequency warping settings for the sinusoidal modeling scheme, we conducted a short objective quality evaluation experiment. 25 male and 25 female sentences were analyzed and represented by the sinusoidal model. Then the line spectrum amplitudes were transformed to the RCC parameters using several selections of \( \lambda \) values and warping functions. The RCC-reconstructed signals were compared to the signals reconstructed directly from the sinusoidal amplitudes and phases, using PESQ standard [13] for wideband (16kHz) signals. The averaged results for the male voice are presented in Figure 2 for different frequency warping settings (linear, parametric / \( \alpha = 0.2 \), \( \alpha = 0.4 / \), mel-scale, bark-scale). The results for the female voice, which are not presented here, exhibit similar trends and have slightly higher PESQ values.

One can notice that the mel-scale warping with \( \lambda \in [1e^{-4}, 5e^{-4}] \) provides the best PESQ results (the same conclusion is derived from the female voice results). It is clearly observed that both no warping (linear scale) and too strong warping (bark scale) reduce the subjective quality of the reconstructed speech. Hence we choose to use Mel Regularized Cepstral Coefficients (MRCC) with \( \lambda = 2e^{-4} \) for the HMM-based TTS modeling, described later.

In the current implementation the phase of the line spectrum is estimated from the corresponding MRCC amplitude parameters as a minimal phase spectrum sampled at harmonic frequencies:

\[
\{ \arg B(\theta_i) \}_{i=0}^L = -2\sum_{i=0}^L d_i \sin(\theta_i) \triangleq \text{Md},
\]

where \( \text{M} \) is:

\[
\text{M} = \begin{bmatrix}
0 & -2\sin(\theta_0) & -2\sin(2\theta_0) & \ldots & -2\sin(p\theta_0) \\
-2\sin(\theta_1) & -2\sin(2\theta_1) & \ldots & -2\sin(p\theta_1) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
-2\sin(\theta_L) & -2\sin(2\theta_L) & \ldots & -2\sin(p\theta_L)
\end{bmatrix}
\]

Figure 2: Objective quality of RCC amplitude parameterization for a male speaker with various regularization constants and RCC order (p).

3. Statistical Parametric System

3.1. System Configuration

The HMM based Speech Synthesis system (HTS,[8]) was used for simultaneous multi-stream modeling of spectral (either MGC-LSP or MRCC) and excitation (logF0) parameters, considering the Global Variance (GV). A standard build of 16 kHz US English female voice (slt) [14] was performed twice, with 34 MGC-LSP spectral parameters and with 34 MRCC parameters (p=33). The 20 first recordings were excluded from the training and used as testing sentences.

3.2. Combination with sinusoidal model

After the generation of spectral and prosodic feature vectors at the HMM-based TTS synthesis stage, they are converted to line spectra appropriate for the constant frame sinusoidal reconstruction. During the reconstruction we apply a constant frequency jittering (8) and perform the high band phase extrapolation [5] (with 4kHz threshold frequency for the both operations).
The two types of synthesized spectral feature vectors (MGC-LSP and MRCC) are converted to the complex line spectra. For MGC-LSP, the all-pole spectral envelope is sampled at the pitch frequency multiples to obtain sinusoidal amplitudes and phases, as proposed by Banos et al [7]. For MRCC, the continuous amplitude spectrum is reconstructed according to [9] and the corresponding minimal phase spectrum is reconstructed by (14).

3.3. Experiments and Results

In order to ease the subjective evaluation of the amplitude spectrum solely, we have synthesized the test sentences imposing the original phone durations.

Three systems have been compared for the slt voice: A) the baseline HTS synthesis with a MLSA filter [8]; B) the MGC-LSP-based sinusoidal synthesis, sharing the same features for training as in A, according to [7]; and C) the MRCC-based system with MRCC parameters used both for training and sinusoidal synthesis. In our subjective pair comparison evaluations each of 10 subjects (7 non-experts and 3 experts in speech science) listened to 33 sentence pairs containing samples randomly selected from the outputs of the three systems. The subjects were instructed to choose between 5 options: no preference, strong preference to either side or weak preference to either side. The results are presented at Figure 3.

The evaluations have revealed that both system (B) and system (C) significantly ($p<0.001$) outperform the baseline HTS synthesis (A). The preference of the proposed MRCC (MRCC) as spectral parameters for the HMM-based TTS system (B) compared to the MGC-LSP-based HTS synthesis (C) is also statistically significant ($p<0.001$).

4. Summary

In the current work we proposed to use RCC parameterization [9] with mel-scale frequency warping (i.e. MRCC) as spectral parameters for the HMM-based TTS system, followed by constant frame sinusoidal waveform reconstruction [5].

Subjective evaluations give a clear indication that the MRCC parameter generation followed by a sinusoidal waveform reconstruction provides a preferable alternative to the MGC-LSP parameter generation within the HMM-based TTS system [8] followed by either a simple MLSA source-filter waveform generation or a sinusoidal model waveform generation.

Also the evaluations confirmed previously reported findings [7] that sinusoidal waveform synthesis (even being fed with the same synthesized MGC-LSP parameters) is preferable compared to a very simple source-filter synthesis. However, advanced source-filter schemes (e.g. STRAIGHT [15]) were not evaluated by us yet.

In the current work we did not strive to enhance the excitation/phase modeling, which seems to be essential for further improvement of the synthesized speech quality.

5. Acknowledgements

Authors would like to thank the HTS development group for making HTS tools for HMM-based TTS modeling and synthesis publicly available.

6. References


