Sinusoidal model parameterization for HMM-based TTS system

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

Abstract

A sinusoidal representation of speech is an alternative to the source-filter model. It is widely used in speech coding and unit-selection TTS, but is less common in statistical TTS frameworks. In this work we utilize Regularized Cepstral Coefficients (RCC) estimated in mel-frequency scale for amplitude spectrum envelope modeling within an HMM-based TTS platform. Improved subjective quality for mel-frequency RCC (MRCC) combined with the sinusoidal model based reconstruction is reported, compared to the state-of-the-art MGC-LSP parameters.

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

1. Introduction

An HMM-based speech synthesis becomes a leading technology for low footprint TTS systems. Although it provides high intelligibility the quality of the synthesized speech has some room for improvement. 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 HTS 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:}

\[ s_{a}(n) = \hat{s}_{a}(n) = w(n) \sum_{i=0}^{L} A_i \sin(\theta_i n + \phi_i), \]

where \( w(n) \) is a symmetric window, e.g. Hamming or Hann, \( \{A_i\} \) and \( \{\phi_i\} \) are harmonic amplitudes and phases correspondingly and \( \theta_i \) is the position of the highest local maximum found on the short time amplitude spectrum \( |S_a(\omega)| \) in a close vicinity of \( \omega_k \), i.e. the k-th multiple of the angular pitch frequency \( \theta_k \). Equation (1) is equivalent to the frequency domain equation (2)

\[ s_a(\omega) = \hat{s}_a(\omega) = \sum_{k=1}^{K} A_k e^{i\omega_k} W(\omega - \theta_k), \]

where \( s_a(\omega) \) is a short time spectrum and \( W(\omega) \) is the Fourier transform of \( w(n) \). The vector \( c \triangleq \{c_k\}_{k=0}^{K} \triangleq \{A_k e^{i\omega_k}\}_{k=0}^{K} \) 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:

\[ E = \sum_{n=0}^{N-1} \| S_a(\omega_n) - \hat{S}_a(\omega_n) \|^2, \]

where \( N \) 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 voice 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

$$\theta_k = \theta_k + \Delta \theta_k$$

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

In order to achieve smooth evolvement of the synthesized waveform in time it is crucial to perform phase alignment of the consecutive line spectra. To this end we calculate two linear phase terms to be applied to the current ($m$-th) line spectrum. The tangent of the first term, given by (5) maximizes the cross-correlation between the current and the previous line spectra.

$$\tau_1 = \arg \max ||\text{IFFT} (e^{\imath \theta} e^{\imath (m-1)})||$$

The tangent of the second linear term $\tau_2$ accounts for the constant time offset between the current and previous frames of the reconstructed signal:

$$\tau_2 = 0.5(\theta_2(m) + \theta_2(m-1))N,$$

where N is the frame offset in samples. Finally, the alignment phase of synthesized line spectrum is given by

$$\phi_k = (\tau_1 + \tau_2)k + \phi_0$$

Once the alignment phase (7) is incorporated in the synthesized line spectrum, the windowed waveforms are reconstructed in the frequency domain according to (2), 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 [2][4][9]. 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 $d_t$ represent a continuous amplitude spectrum $B(\theta)$:

$$a \triangleq \| \log |B(\theta)| \|^t = d_t + 2 \sum_{i=1}^{\infty} \cos (\theta_i) \triangleq \text{Md},$$

where $M$ is a real cepstrum transform matrix and $p$ is the order of the cepstrum

$$M = \begin{bmatrix}
1 & 2 \cos (\theta_1) & \ldots & 2 \cos (p \theta_1) \\
\vdots & \vdots & \ddots & \vdots \\
1 & 2 \cos (\theta_L) & \ldots & 2 \cos (p \theta_L)
\end{bmatrix}$$

The RCC parameters are obtained by the minimization of the composite criterion (10)

$$e = \sum_t \left( \log |B(\theta)_{k,t}|^t - d_t \right)^2 + \lambda \Re (B(\theta)),$$

where the first term is the squared error between the sinusoidal amplitudes and the modeled spectrum sampled at the pitch harmonics while the second regularization term given by

$$\Re (S(\theta)) = 2 \pi \int_{\theta}^{\theta+\Delta} \left( \frac{d \Theta}{d \theta} \right)^2 d \theta$$

accounts for the smoothness of the model spectrum. In (10) $\lambda$ is the regularization control parameter.

The minimization of (10) with respect to cepstral coefficients $d_t$ in (8) results in [9]:

$$d_t = (M^T M + \lambda R)^{-1} \text{Md},$$

where $R$ is a regularization diagonal matrix:

$$R = 8 \pi^2 \begin{bmatrix}
0 & 1 & \vdots & \vdots \\
1 & 2 & \vdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \vdots & p^2
\end{bmatrix}$$

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 $f(\theta)$

$$\theta = f(\theta), 0 \leq \theta \leq \pi$$

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:

$$f(\theta) = \arctan \left( \frac{(1 - \alpha) \sin \theta}{(1 + \alpha) \cos \theta - 2 \alpha} \right),$$

where $\alpha$ is a warping control parameter. Figure 1 shows warping curves corresponding to some values of $\alpha$ along with the ones corresponding to bark and mel scales.

In [9] it was proposed to apply bark-scale warping prior to the RCC solution (12), combined with $\lambda = 5 e^{-2}$, 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 λ 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.

\[ \lambda \in [1e^{-4}, 5e^{-4}] \] 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:

\[
\text{arg } B_k(i) \big|_{i=\alpha} = -2\sum_d \sin(\theta_d) \delta_{i0},
\]

where \( M \) is:

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

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 (4) and perform the high band phase extrapolation [5] (with 4kHz threshold frequency 4kHz 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 (8) and the corresponding minimal phase spectrum is reconstructed by (16).

Figure 3 presents the spectrograms of the speech signals synthesized using MRCC and MGC-LSP parameters. It is clearly seen that the MRCC synthesized spectrum conveys more details than the MGC-based spectrum, thus reducing the spectral smearing effect.

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.

Figure 2: Objective quality of RCC amplitude parameterization for a male speaker with various regularization constants and RCC order (p). The results are presented for different frequency warping setting: parametric (\( \alpha = 0.2,0.4,0.6 \)), ‘mel’ and ‘bark’.
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 4.

The evaluations have revealed that both system (B) and system (C) significantly \( p<0.000001 \) outperform the baseline HTS synthesis (A). The preference of the proposed MRCC based synthesis (C) compared to the MGC-LSP-based harmonic synthesis (B) 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 [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 improve 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