Minimum mean squared error based warped complex cepstrum analysis for statistical parametric speech synthesis

Ranniery Maia†, Mark J. F. Gales†, Yannis Stylianou†, Masami Akamine‡

†Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK
‡Toshiba Corporation, Corporate Research and Development, Kawasaki, Japan
ranniery.maia@crl.toshiba.co.uk

Abstract

This paper presents an approach for complex cepstrum analysis based on the minimum mean squared error criterion, and describes its application to statistical parametric speech synthesis. The proposed method alleviates some of the issues associated with conventional complex cepstrum analysis, such as choice of the window, phase unwrapping, and the need for accurate pitch marks. Given initial estimates of warped complex cepstra and respective analysis instants, the method iteratively optimizes the complex cepstrum on a warped quefrency domain by minimizing the mean squared error between the natural and the reconstructed speech waveforms. When applied to statistical parametric speech synthesis, the optimized complex cepstrum results in better performance in terms of synthesized speech quality, specially for emotional databases, when compared with the complex cepstrum calculated through conventional methods.

Index Terms: Speech synthesis, statistical parametric speech synthesis, complex cepstrum, cepstral analysis

1. Introduction

In statistical parametric speech synthesis [1], the use of a source-filter model of speech production that incorporates the effects of lip radiation and glottal excitation, aside from the usual vocal tract effect, can result in a significant improvement on the quality of the synthetic speech. In order to address this problem, the use of the complex cepstrum to incorporate glottal pulse information into statistical parametric speech synthesis systems has been proposed [2, 3]. From the perspective of the speech production mechanism in source-filter modeling, the use of the complex cepstrum has an advantage over the commonly used cepstrum of minimum-phase cepstrum because it better represents the mixed-phase characteristics of speech signals [4]. However, complex cepstrum analysis has certain drawbacks. The speech signal must be windowed at the glottal closure instants (GCI). The accuracy of the detection of the GCIs, as well as the type of window used for analysis, have a direct impact on the estimation of the complex cepstrum [5, 6, 7]. In addition, a phase unwrapping procedure is usually performed to obtain the phase spectrum of the speech segment as a continuous function of the frequency. Furthermore, statistical parametric synthesizers usually require the observation vectors to be extracted at a fixed frame rate rather than at the GCIs. In our previous work [2, 3] these issues were mostly solved through the use of accurate pitch marks, high-order Discrete Fourier Transforms (DFT), linear interpolation of amplitude and phase, and use of the anti-causal cepstrum as glottal excitation parameters.

To alleviate the dependency on accurate pitch marks, choice of the window and phase unwrapping during the complex cepstrum analysis, we have proposed a complex cepstrum analysis method based on the minimum mean squared error (MSE), where the complex cepstrum is calculated in a two-step optimization process [8]. In the first step, given initial estimates of complex cepstra, their corresponding analysis instants are updated. The analysis instants are represented as a multi-pulse excitation signal. After that, complex cepstra are recalculated given the modified excitation signal. Both procedures are conducted in a way that the MSE between natural and reconstructed speech is minimized. Because the method is based on time-varying filtering of the excitation signal, no windowing is performed. Furthermore, because the optimization is conducted in the cepstral domain, phase unwrapping is not necessary. Finally, the optimization is conducted in a frame basis, resulting in frame-based complex cepstra, which is suitable for statistical parametric speech synthesis. This paper extends the work presented in [8] by implementing frequency warping in the complex cepstrum optimization process, and by adjusting the joint excitation-cepstrum optimization procedure so as to suit statistical parametric synthesizers. Subjective tests on an emotional corpus are also presented to evaluate the performance of the proposed complex cepstrum analysis under statistical parametric synthesizers.

This paper is organized as follows: Section 2 gives an overview of speech modeling using warped complex cepstra; Section 3 describes the framework of warped complex cepstrum analysis based on the MSE; Section 4 describes its application to statistical parametric speech synthesis; Section 5 shows some experiments, and the conclusions are in Section 6.

2. Speech analysis and synthesis using warped complex cepstrum

We assume that the speech signal, \( s(n) \), is produced by

\[
s(n) = h(n) * e(n),
\]

where \( h(n) \) is a slowly varying non-causal impulse response representing the effects of the glottal flow, vocal tract, and lip radiation. The excitation signal, \( e(n) \), is assumed to be composed of delta pulses (amplitude one) or white noise for voiced and unvoiced portions of the speech signal, respectively [4].

The synthesis filter impulse response, \( h(n) \), can be derived from the speech signal, \( s(n) \), by warped cepstral analysis

\[
\hat{S} \left( e^{j\beta z^{-1} \omega} \right) = \sum_{n=-\infty}^{\infty} s(n)e^{-j\beta z^{-1} \omega} n;
\]

\[
\hat{s}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \hat{S} \left( e^{j\beta z^{-1} \omega} \right) e^{jwn} d\omega,
\]

\( Z \)
where $\tilde{S}\left(e^{j\theta^{-1}}(\omega)\right)$ is the warped complex spectrum of $s(n)$, and
\[ \beta_\alpha(\omega) = \frac{(1 - \alpha^2) \sin \omega}{(1 + \alpha^2) \cos \omega - 2\alpha}, \] (4)
is a bilinear transform to warp the angular frequency $\omega$ [9]. The parameter $\alpha$ controls the degree of warping. Usually, at 16 kHz, $\alpha = 0.42$ and $\alpha = 0.55$ give good approximations to the mel and bark scales, respectively [10]. $\hat{s}(n)$ is by definition an infinite and non-causal sequence. If pitch synchronous analysis with an appropriate window to select two pitch periods is performed, then samples of $\hat{s}(n)$ tend to zero as $n \to \infty$. In this case, if the signal $e(n)$ is a delta pulse or white noise sequence then a cepstral representation of $\hat{h}(n)$ can be given by lifting the cepstral sequence $\hat{s}(n)$. This can be obtained by making $\hat{h}(n) = \hat{s}(n)$, so that $|n| \leq C$, where $C$ is the cepstral order.

To synthesize speech, the warped complex cepstrum of $s(n)$, $\hat{h}(n)$, must be converted into the non-causal impulse response $\tilde{h}(n)$
\[ H(e^{j\omega}) = \exp \sum_{n = -C}^{C} \hat{h}(n)e^{-j\beta_\alpha(\omega)n}, \] (5)
\[ \hat{h}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln H(e^{j\omega}) e^{j\omega n} d\omega, \] (6)
where $H(e^{j\omega})$ is the complex spectrum of $h(n)$. Finally, speech can be reconstructed through (1).

Other possibilities to obtain a warped complex cepstrum is to apply the bilinear transform directly on the non-causal cepstral sequence $\{\hat{h}(-C), \ldots, \hat{h}(C)\}$ [9]. However, the form presented here, by taking a inverse DFT of the warped spectrum, is the formulation that will be used in Section 3. When compared to the linear case, it reduces the dimensionality of the cepstral vector utilized in the optimization. Samples of the warped spectrum, $\{S\left(e^{j\theta^{-1}}(\omega_0)\right), \ldots, S\left(e^{j\theta^{-1}}(\omega_L)\right)\}$, where $\{\omega_0, \ldots, \omega_L\}$ are $L + 1$ frequencies equally (linearly) sampled between $\omega = 0$ and $\omega = \pi$, can be calculated through a non-uniform DFT [11] or a warped DFT [12]. Henceforth in this paper the term warped will be dropped and unless stated differently complex cepstrum means warped complex cepstrum.

3. Warped complex cepstrum analysis based on the MSE

To compute the complex cepstrum based on the MSE, the analysis-by-synthesis scheme of Figure 1 is used [8]. The idea is that initial estimates of the complex cepstrum are optimized so that the error between natural and reconstructed speech is minimized. Here we consider only the voiced portions of the speech signal, therefore the excitation signal, $e(n)$, is composed solely of pulses located at the glottal closure instants.

Complex cepstrum analysis is performed in two steps. In the first one, the locations of the pulses of the excitation signal, $e(n)$, representing the glottal closure instants, are optimized given the complex cepstrum, $\hat{h}(n)$. In the second step, the complex cepstrum $\hat{h}(n)$ at each frame of the speech signal is estimated given the updated excitation signal, $e(n)$. Both procedures are conducted in a way that the MSE between natural, $s(n)$, and reconstructed speech, $\hat{s}(n)$, is minimized. Stopping criterion can be based on the segmental signal-to-noise ratio (SNRseg) between natural and reconstructed speech or simply the minimum MSE achieved.

Excitation optimization is conducted by keeping $\hat{h}(n)$ fixed while updating the amplitudes, $a = \{a_0, \ldots, a_{Z-1}\}$, and locations, $p = \{p_0, \ldots, p_{Z-1}\}$, of the non-zero samples of $e(n)$, where $Z$ is the number of non-zero samples or pulses. During the process the mean power of the error signal, $E\{w^2(n)\}$, is minimized in a fashion that resembles the multi-pulse excited speech coding algorithm [13].

By considering matrix notation, the error of the system of Figure 1, $w(n)$, can be written as
\[ w = s - \hat{s} = s - He, \] (7)
where
\[ s = \begin{bmatrix} 0 & \cdots & 0 & s(0) & \cdots & s(N - 1) & 0 & \cdots & 0 \end{bmatrix}^T, \] (8)
\[ e = \begin{bmatrix} e(0) & \cdots & e(N - 1) \end{bmatrix}^T, \] (9)
with $s$ being a $N + M$-size vector whose elements are samples of the speech signal, $s(n)$, padded with zero. $e$ contains samples of the excitation signal, $e(n)$, $M$ is the order of $\hat{h}(n)$, and $N$ is the number of samples of $s(n)$. The $(M + N) \times N$ synthesis filter impulse response matrix $H$ has the following shape
\[ H = \begin{bmatrix} g_0 & \cdots & g_{N-1} \end{bmatrix}, \] (10)
\[ g_n = \begin{bmatrix} 0 & \cdots & 0 & h_n & 0 & \cdots & 0 \end{bmatrix}^T, \] (11)
\[ h_n = \begin{bmatrix} h_n(-\frac{M}{2}) & \cdots & h_n(\frac{M}{2}) \end{bmatrix}^T, \] (12)
where $h_n$ contains the impulse response of $\hat{H}(z)$ at the $n$-th sample position. Considering that the vector $e$ has only $Z$ non-zero samples, then $\hat{s}$ can be written as
\[ \hat{s} = He = \sum_{z=0}^{Z-1} a_z g_{p_z}, \] (13)
where $p_z$ and $a_z$ are the positions and amplitudes of the $Z$ non-zero samples of $e(n)$. The mean squared error as a function of the excitation $e(n)$ then can be approximated as
\[ \varepsilon(p, a) = \frac{1}{N} \left[ s - \sum_{z=0}^{Z-1} a_z g_{p_z} \right]^T \left[ s - \sum_{z=0}^{Z-1} a_z g_{p_z} \right]. \] (14)

The $z$-th pulse amplitude $a_z$, which minimizes (14) can be found by making $\frac{\partial \varepsilon(p, a)}{\partial a_z} = 0$, which results in
\[ a_z = g_{p_z} \left[ s - \sum_{z=0}^{Z-1} a_z g_{p_z} \right] g_{p_z}^T \] (15)
and the respective best position given by
\[
\hat{p}_z = \arg \max_{p_z = p_1 - \Delta p, \ldots, p_z + \Delta p} \left\{ \sum_{t=0}^{T-1} g_{p_z} \left[ s - \sum_{t=0}^{T-1} A_t g_{p_z} \right] \right\}^2
\]
with \(\Delta p\) being the range of samples in which the search for the best position in the neighborhood of \(p_z\) is conducted.

### 3.2. Complex cepstrum estimation

Because the impulse response \(h(t)\) is associated with each frame \(t\) of the speech signal, the reconstructed speech vector \(\hat{s}\) can also be written in matrix form as
\[
\hat{s} = \sum_{t=0}^{T-1} A_t h_t,
\]
where \(T\) is the number of frames in the sentence, and \(h_t = [h_t(-\Delta), \ldots, h_t(\Delta)]^T\) contains the synthesis filter impulse response at the \(t\)-th frame of the speech signal. The \((K + M) \times (M + 1)\) excitation matrix \(A_t\) is given by
\[
A_t = \begin{bmatrix} u_{-\frac{M}{2}} & \cdots & u_{\frac{M}{2}} \\
0 & \cdots & 0 & \cdots & 0 \\
\end{bmatrix},
\]
(18)
where \(e_t\) is the excitation vector where only samples belonging to the \(t\)-th frame are non-zero, and \(K\) is the number of samples per frame. By considering (17), the MSE can be written as
\[
\varepsilon = \frac{1}{N} \left[ s - \sum_{t=0}^{T-1} A_t h_t \right]^T \left[ s - \sum_{t=0}^{T-1} A_t h_t \right].
\]
(21)

The optimization must be performed in the warped cepstral domain. The relationship between the impulse response vector, \(h_t\), and its corresponding complex cepstrum vector, \(h_t = [\hat{h}_t(-C), \ldots, \hat{h}_t(C)]^T\), can be written as
\[
h_t = f(\hat{h}_t) = \frac{1}{2L+1} D_2 \exp(D_1 \hat{h}_t),
\]
(22)
where \(\exp(\cdot)\) means a matrix formed by taking the exponential of each element of the matrix argument, and \(L\) is the number of one-sided sampled frequencies in the spectral domain. The elements of the \((2L+1) \times (2C+1)\) matrix \(D_1\), and the \((M+1) \times (2L+1)\) matrix \(D_2\) are given respectively by
\[
D_1(i,j) = e^{-j\omega_j \rho_1}, \quad -\frac{M}{2} \leq i \leq \frac{M}{2}, -C \leq j \leq C,
\]
(23)
\[
D_2(i,j) = e^{j\omega_j \rho_1}, \quad -L \leq i \leq L, -L \leq j \leq L,
\]
(24)
where \(\rho_0 = 0, \omega_{-\frac{1}{2}} = \pi, \omega_{-1} = -\omega_0\). By substituting (22) into (21), which is the cost function to be minimized during the complex cepstrum estimation
\[
\varepsilon(\hat{h}_t) = \frac{1}{N} \left[ r_t^T r_t - 2r_t A_t f(\hat{h}_t) + f(\hat{h}_t)^T \Phi_t f(\hat{h}_t) \right],
\]
(25)
where \(\Phi_t = A_t^T A_t\) and
\[
r_t = s - \sum_{t=0, j \neq t}^{T-1} A_t f(\hat{h}_t).
\]
Since the relationship between cepstrum and impulse response, \(h_t = f(\hat{h}_t)\), is non-linear, a gradient method can be utilized to optimize the complex cepstrum. The gradient of \(\varepsilon(\hat{h}_t)\) with respect to \(\hat{h}_t\) is
\[
\nabla_{\hat{h}_t} \varepsilon(\hat{h}_t) = -\frac{2}{N(2L+1)} D_1^T \text{diag}(\exp(D_1 \hat{h}_t)) D_2^T A_t^T [r_t - A_t f(\hat{h}_t)].
\]
(27)

### 4. Application to statistical parametric speech synthesis

#### 4.1. Modification of the excitation optimization process

During the waveform generation stage of most statistical parametric synthesizers the construction of the voiced excitation signal must be done from the generated fundamental frequency \(F_0\) solely. To overcome this problem, pulse amplitudes \(\{a_0, \ldots, a_{L-1}\}\) are all forced to one, i.e. \(a_n = 1, \forall n\), prior to the complex cepstrum estimation process described in Section 3.2. Such procedure forces the gain term of the complex cepstrum to encapsulates the power of the speech signal. However, this is effective only when the initial cepstrum is computed through the use of unnormalized windows.

Aside from that, pulse elimination to avoid half-pitch period problems is conducted based on the amplitudes \(a_n\) [8].

#### 4.2. Acoustic modeling of complex cepstra

Statistical modeling of complex cepstrum follows the proposal described in [2]. The final frame-based complex cepstrum is split into minimum-phase cepstrum and some phase parameters. The minimum-phase cepstrum is obtained by making
\[
\hat{h}_m(n) = \begin{cases} \hat{h}(0), & n = 0, \\
\hat{h}(n) + \hat{h}(-n), & 1 \leq n \leq C, \\
\end{cases}
\]
(28)
while the phase parameters are defined as
\[
\phi(n-1) = -\hat{h}(n-1), \quad 1 \leq n \leq C_n,
\]
(29)
where \(C_n\) is the number of phase parameters. The phase parameters \(\phi(n)\) are related to the glottal flow excitation [3] and are clustered independently during the training of the synthesizer.

#### 4.3. Synthesis

Synthesis is also performed according to [2], \(\phi(n)\) is used to derive the phase response, \(\theta_n(n)\), of a glottal filter
\[
\theta_n(\omega_l) = -2 \sum_{n=1}^{C_n} \phi(n-1) \sin(\beta_n(\omega_l) n),
\]
(30)
for \(l = 0, \ldots, L\), whose impulse response is given by
\[
h_n(n) = \frac{1}{2L+1} \left\{ 1 + 2 \sum_{l=1}^{L} \cos(\omega_l n + \theta_n(\omega_l)) \right\},
\]
(31)
for \(n = -P_n, \ldots, P_n\), where \(P_n\) is the glottal filter order.
5. Experiments

5.1. Experimental conditions

The following emotional corpora, with audio sampled at 16 kHz, were utilized throughout this experimental section:

- F1-N: female American English in neutral style;
- F2-A, F2-H, F2-N, F2-S: female British English in angry, happy, neutral and sad styles, respectively;
- F3-A, F3-H, F3-N, F3-S: female American English in angry, happy, neutral and sad styles, respectively;
- M1-N: male American English in neutral style.

Pitch marks and F0 were both extracted using a proprietary tool for the speakers F1, F3, M1, and M2, and ESPS (Entropic Signal Processing Software) Epoch [14] for F2.

Cepstral order utilized was \( C = 39 \). Initial cepstra were calculated from warped amplitude and phase spectra, obtained through a 8192-size non-uniform DFT [11], with warping factor \( \alpha = 0.42 \) (see Section 2) to approximate the mel scale [10]. Phase unwrapping was performed as described in [2, 3].

For complex cepstrum analysis based on the MSE, during the excitation optimization process described in Section 3.1, each pulse was searched in the neighborhood of 20 samples, i.e. \( \Delta p = 20 \). For the complex cepstrum estimation process shown in Section 3.2, the synthesis filter order and the number of frames, and number of voiced frames.

5.2. Speech analysis and reconstruction

To verify the performance of the proposed MSE-based complex cepstrum analysis method in terms of speech modeling, fifty test utterances were randomly selected from each corpus and analyzed according to described in Section 5.1. After that, the utterances were synthesized by making \( s(n) = h(n) \ast e(n) \), where \( h(n) \) is a non-causal impulse response derived from the complex cepstrum according to Section 2, and \( e(n) \) is an excitation signal with unit pulses in the voiced regions, whose locations correspond to the locations of the pitch marks, and white noise in the unvoiced regions, defined according to the \( F_0 \).

Figure 2 shows the results in terms of segmented signal-to-noise ratio of voiced segments (SNRseg-v), given by

\[
d = \frac{10}{T_\nu} \sum_{t=0}^{T-1} w_t \log_{10} \left( \frac{\sum_{n=0}^{K-1} s^2 (tK + n)}{\sum_{n=0}^{N-1} [\hat{s} (tK + n) - \bar{s} (tK + n)]^2} \right),
\]

where \( s(n) \) and \( \hat{s}(n) \) are respectively the natural and reconstructed speech signals, and \( T \) and \( T_\nu \) are respectively the number of frames, and number of voiced frames. \( w_t \) is a weight so that \( w_t = 0 \) if the frame is unvoiced, and \( w_t = 1 \) if the frame is voiced. From the results shown in Figure 2 it can be seen that the SNRseg-v has a logarithmic behavior across joint iterations of excitation optimization and complex cepstrum estimation.

5.3. Experiments with synthesized speech

The F2 emotional corpus (British English) was utilized to train emotion-dependent statistical parametric speech systems. The amount of training was: 758 angry, 759 happy, 3140 neutral, and 753 sad utterances. For each corpus, two synthesizers were trained: one with features derived from the initial cepstrum and another with features derived from the final optimized cepstrum, according to Section 4.2. Each observation vector was composed of six streams: (1) 40 minimum-phase cepstra derived from complex cepstra, plus delta (\( \Delta \)) and delta-delta (\( \Delta^2 \)); (2) \( \ln F_0 \); (3) \( \Delta \ln F_0 \); (4) \( \Delta^2 \ln F_0 \); (5) 22 Bark band-aperiodicity coefficients [16], plus \( \Delta \) and \( \Delta^2 \); (6) 19 phase parameters derived from complex cepstra, plus \( \Delta \) and \( \Delta^2 \).

For each corpus, 50 test sentences were synthesized by each one of the two versions of the systems, according to Section 4.3. The test utterances were submitted to Amazon Mechanical Turk [17] for carrying out preference tests. Some mechanisms to detect cheating, such as minimum number of tested samples and listener’s disagreement index were used. The average listener’s preference are shown in the graphics of Figure 3. It can be noticed that the proposed MSE-based complex cepstrum method performances well for the angry and happy emotional data. This indicates the strength of the proposed method on data with high \( F_0 \) fluctuations. The poor performance on sad and indistinguishable quality on neutral may infer that the adjustments on the excitation optimization described in Section 4.1, such as elimination of pulses with negative amplitude, or the fixed range of samples for pulse search, \( \Delta p = 20 \), were not adequate for databases with low \( F_0 \) fluctuations.

6. Conclusions

This paper presented a warped complex cepstrum analysis method based on the minimum mean squared error criterion, and described its application to statistical parametric speech synthesis. The approach alleviates crucial issues related to conventional complex cepstrum analysis, such as choice of the analysis window, phase unwrapping, and consistent derivation of frame-based complex cepstra for statistical modeling. Experimental results shows that the resulting complex cepstrum results in parametric synthesizers with good performance in databases where there is high \( F_0 \) excursions, as in speech uttered with intense emotions such as anger and happiness. Future work includes robustness on data with low \( F_0 \) fluctuations.
7. References


