COMPLEX CEPSTRUM ANALYSIS BASED ON THE MINIMUM MEAN SQUARED ERROR

Ranniery Maia[†], *Masami Akamine*[‡], *M. J. F. Gales*[†]

[†]Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK
[‡]Toshiba Corporation, Corporate Research and Development Center, Kawasaki, Japan

ABSTRACT

This paper introduces a novel approach for complex cepstrum analysis. Given initial estimates of complex cepstra and respective instants of glottal closure, the method iteratively optimizes the complex cepstrum and instants of glottal closure so that the mean squared error between natural and reconstructed speech waveforms is minimized. The proposed approach results in a more accurate speech representation based on the complex cepstrum, with no need of windowing or phase unwrapping. Experimental results show that the proposed method produces reconstructed speech with higher segmental signal-to-noise ratio scores when compared with conventional methods of complex cepstrum analysis. Because this approach can derive the complex cepstrum at fixed periods, it can be applied to statistical modeling in parametric speech synthesizers.

***Index Terms*—** Complex cepstrum, cepstral analysis, speech analysis, speech synthesis

1. INTRODUCTION

The speech production models that are usually utilized in most of speech applications have mostly relied on a simplified parametric model of speech production where a minimum-phase filter is excited by a signal which consists of a mixture of pulses and noise. The use of a minimum-phase filter, as an approximation for the effects of the vocal tract and lip radiation of the human speech production model, has been a legacy of the speech coding area where causality is essential [1]. However, in speech coders the limitations of this simplification can be compensated for by the excitation signals that are usually derived through a frame-based analysis-by-synthesis procedure in the encoding part [2]. In other speech applications, such as statistical parametric speech synthesis [3], causality is usually not a requirement. In this case, the use of a more accurate speech model could have a significant impact 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 [4]. 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. The complex cepstrum representation of the speech signal allows a non-causal modeling of short-time speech segments, which is actually observed in natural speech [1]. However, though theoretically advantageous, 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]. 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. A high-order Fast Fourier Transform (FFT) often improves the performance of this phase unwrapping process as well as avoiding aliasing, at a cost of an increase of computational complexity [7].

This paper introduces a novel approach for complex cepstrum analysis. The proposed method calculates the complex cepstrum in a two-step optimization process. In the first one, given initial complex cepstra the GCIs are updated. After that, the complex cepstra are recalculated given the modified GCIs, the excitation signal, and natural speech. Both procedures are conducted in a way that the mean squared error between natural and reconstructed speech is minimized. Because the proposed 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 more suitable for most of speech applications.

This paper is organized as follows: Section 2 gives an overview of speech modeling using the complex cepstrum; Section 3 describes the proposed complex cepstrum analysis method; Section 4 shows some experiments, and the conclusions are in Section 5.

2. COMPLEX CEPSTRUM-BASED SPEECH MODELING

We assume a digital model in which speech is produced by

$$s(n) = h(n) * e(n), \tag{1}$$

where $h(n)$ is a slowly varying impulse response representing the effects of the glottal flow, vocal tract, and lip radiation [1]. The excitation signal, $e(n)$, is composed of delta pulses (amplitude one) or white noise for voiced and unvoiced portions of the speech signal, respectively.

The synthesis filter impulse response, $h(n)$, can be derived from the speech signal, $s(n)$, through cepstral analysis. The cepstrum of $s(n)$ is given by [7]

$$\hat{s}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \ln \left| S\left(e^{\jmath\omega}\right) \right| + \jmath\theta(\omega) \right\} e^{\jmath\omega n} d\omega, \tag{2}$$

$$S\left(e^{\jmath\omega}\right) = \sum_{n=-\infty}^{\infty} s(n) e^{-\jmath\omega n} = \left| S\left(e^{\jmath\omega}\right) \right| e^{\jmath\theta(\omega)}, \tag{3}$$

where $\left| S\left(e^{\jmath\omega}\right) \right|$ and $\theta(\omega)$ are respectively the amplitude and phase spectrum of $s(n)$. $\hat{s}(n)$ 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 $h(n)$, here defined as the *complex cepstrum of* $s(n)$, can be given by $\hat{h}(n) = \hat{s}(n)$, so that $|n| \leq C$, where C is the *cepstral order*.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

To synthesize speech, the complex cepstrum of $s(n)$, $\hat{h}(n)$, must be converted into the impulse response $h(n)$

$$H\left(e^{\jmath\omega}\right) = \exp \sum_{n=-C}^{C} \hat{h}(n) e^{-\jmath\omega n}, \tag{4}$$

$$h(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H\left(e^{\jmath\omega}\right) e^{\jmath\omega n} d\omega, \tag{5}$$

where $H\left(e^{\jmath\omega}\right)$ is the complex spectrum of $h(n)$. Finally, speech can be reconstructed through (1).

Theoretically, the use of the complex cepstrum results in a more accurate model of the speech signal when compared to the minimum-phase synthesis filter approach, which discards the glottal flow information contained in $\hat{h}(n)$ [5]. However, complex cepstrum analysis is very sensitive to the location of the analysis and shape of window utilized [6, 8], as well as to the the performance of the phase unwrapping algorithm used to estimate the continuous phase response $\theta(\omega)$[9, 10].

3. PROPOSED COMPLEX CEPSTRUM ANALYSIS

To overcome the complex cepstrum analysis issues commented in Section 2, the analysis-by-synthesis scheme of Fig. 1 is proposed. The idea is that initial estimates for 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.

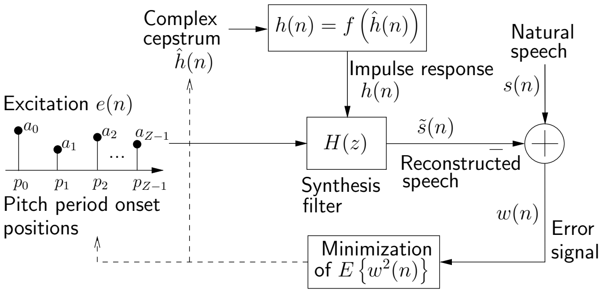

Fig. 1. Illustration of the proposed process for complex cepstrum analysis based on the minimum mean squared error.

The proposed complex cepstrum analysis approach is performed in two steps. In the first one, the locations of the pulses of the excitation signal, $e(n)$, representing the GCIs, 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 excitation signal, $e(n)$, with updated pulse positions. Both procedures are conducted in a way that the mean squared error (MSE) between natural, $s(n)$, and reconstructed speech, $\tilde{s}(n)$, is minimized. In the following sections these two procedures are described.

3.1. Pitch period onset position optimization

Pitch period onset position optimization is conducted by keeping $h(n)$ fixed while updating the amplitudes, $\{a_0, \ldots, a_{Z-1}\}$, and locations, $\{p_0, \ldots, p_{Z-1}\}$, of the pulses of $e(n)$, where Z is the number of GCIs. During the process the mean power of the error signal, $E\left\{w^2(n)\right\}$, is minimized in a fashion that resembles the multi-pulse excited speech coding algorithm [2].

By considering matrix notation, $w(n)$ can be written as

$$\boldsymbol{w} = \boldsymbol{s} - \tilde{\boldsymbol{s}} = \boldsymbol{s} - \boldsymbol{He}, \tag{6}$$

where

$$\boldsymbol{s} = \Bigg[\underbrace{0 \ \cdots \ 0}_{\frac{M}{2}} \quad s\left(0\right) \quad \cdots \quad s\left(N-1\right) \quad \underbrace{0 \ \cdots \ 0}_{\frac{M}{2}}\Bigg]^{\top}, \tag{7}$$

$$\boldsymbol{e} = \left[e\left(0\right) \quad \cdots \quad e\left(N-1\right)\right]^{\top}, \tag{8}$$

with $\boldsymbol{s}$ being a $N+M$-size vector whose elements are mostly samples of the speech signal, $s(n)$, $\boldsymbol{e}$ contains samples of the excitation signal $e(n)$, M is the order of $h(n)$, and is N the number of samples of $s(n)$. The $(M+N) \times N$ matrix $\boldsymbol{H}$ has the following shape

$$\boldsymbol{H} = \left[\boldsymbol{g}_0 \quad \cdots \quad \boldsymbol{g}_{N-1}\right], \tag{9}$$

$$\boldsymbol{g}_n = \Bigg[\underbrace{0 \ \cdots \ 0}_{n} \quad \boldsymbol{h}_n^{\top} \quad \underbrace{0 \ \cdots \ 0}_{N-n-1}\Bigg]^{\top}, \tag{10}$$

$$\boldsymbol{h}_n = \left[h_n\left(-\tfrac{M}{2}\right) \quad \cdots \quad h_n(\tfrac{M}{2})\right]^{\top}, \tag{11}$$

where $\boldsymbol{h}_n$ contains the impulse response of $H(z)$ at the n-th sample position. Considering that the vector $\boldsymbol{e}$ has only Z non-zero samples (voiced excitation), then $\tilde{\boldsymbol{s}}$ can be written as

$$\tilde{\boldsymbol{s}} = \boldsymbol{He} = \sum_{z=0}^{Z-1} a_z \boldsymbol{g}_z, \tag{12}$$

where $\{a_0, \ldots, a_{Z-1}\}$ are the amplitudes of the Z non-zero samples of $e(n)$. The mean squared error then becomes

$$\varepsilon = \frac{1}{N}\boldsymbol{w}^{\top}\boldsymbol{w} = \frac{1}{N}\left(\boldsymbol{s} - \sum_{z=0}^{Z-1} a_z \boldsymbol{g}_z\right)^{\top}\left(\boldsymbol{s} - \sum_{z=0}^{Z-1} a_z \boldsymbol{g}_z\right). \tag{13}$$

The z-th pulse amplitude $\hat{a}_z$ which minimizes (13) can be found by making $\frac{\partial\varepsilon}{\partial a_z} = 0$, which results in

$$\hat{a}_z = \frac{\boldsymbol{g}_z^{\top}\left(\boldsymbol{s} - \displaystyle\sum_{\substack{i=0\\i\neq z}}^{Z-1} a_i \boldsymbol{g}_i\right)}{\boldsymbol{g}_z^{\top}\boldsymbol{g}_z}. \tag{14}$$

By substituting (14) into (13), and considering the terms which depend only on the z-th pulse, the following expression for the best position $\hat{p}_z$ is obtained

$$\hat{p}_z = \underset{p_z = p_z - \frac{\Delta p}{2}, \ldots, p_z + \frac{\Delta p}{2}}{\arg\max} \frac{\left[\boldsymbol{g}_z{\top}\left(\boldsymbol{s} - \sum_{\substack{i=0\\i\neq z}}^{Z-1} a_i \boldsymbol{g}_i\right)\right]^2}{\boldsymbol{g}_z^{\top}\boldsymbol{g}_z}. \tag{15}$$

The term Δp is the range of samples in which the search for the best position in the neighborhood of p_z is conducted.

3.2. Complex cepstrum re-estimation

Because the impulse response $h(n)$ is associated with each frame t of the speech signal, the reconstructed speech vector $\tilde{\boldsymbol{s}}$ can also be written in matrix form as

$$\tilde{\boldsymbol{s}} = \sum_{t=0}^{T-1} \boldsymbol{A}_t \boldsymbol{h}_t, \tag{16}$$

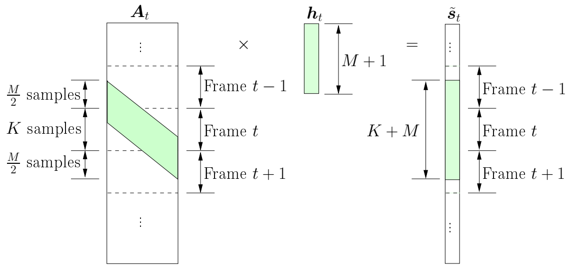

Fig. 2. Illustration of the product $\boldsymbol{A}_t\boldsymbol{h}_t$. Shaded parts indicate possible non-zero samples.

where T is the number of frames in the sentence, and $\boldsymbol{h}_t = \left[h_t\left(-\frac{M}{2}\right) \quad \cdots \quad h_t\left(\frac{M}{2}\right)\right]^\top$ contains the synthesis filter impulse response at the t-th frame of $s(n)$. The $(K+M)\times(M+1)$ matrix $\boldsymbol{A}_t$ is given by

$$\boldsymbol{A}_t = \left[\boldsymbol{u}_{-\frac{M}{2}} \quad \cdots \quad \boldsymbol{u}_{\frac{M}{2}}\right], \tag{17}$$

$$\boldsymbol{u}_m = \left[\underbrace{0 \ \cdots \ 0}_{\frac{M}{2}+m} \quad \boldsymbol{e}_t^\top \quad \underbrace{0 \ \cdots \ 0}_{\frac{M}{2}-m}\right]^\top, \tag{18}$$

$$\boldsymbol{e}_t = \left[\underbrace{0 \ \cdots \ 0}_{tK} \ e\left(tK\right) \ \cdots \ e\left((t+1)K-1\right) \ \underbrace{0 \ \cdots \ 0}_{N-(t+1)K}\right]^\top, \tag{19}$$

where $\boldsymbol{e}_t$ 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. Fig. 2 gives and illustration of the matrix product $\boldsymbol{A}_t\boldsymbol{h}_t$. By considering (16), the MSE can be written as

$$\varepsilon = \frac{1}{N}\left(\boldsymbol{s} - \sum_{t=0}^{T-1}\boldsymbol{A}_t\boldsymbol{h}_t\right)^\top\left(\boldsymbol{s} - \sum_{t=0}^{T-1}\boldsymbol{A}_t\boldsymbol{h}_t\right). \tag{20}$$

The optimization must be performed in the cepstral domain [11]. The relationship between the impulse response vector, $\boldsymbol{h}_t$, and its corresponding complex cepstrum vector, $\hat{\boldsymbol{h}}_t = \left[\hat{h}_t(-C) \quad \cdots \quad \hat{h}_t(C)\right]^\top$, can be written as

$$\boldsymbol{h}_t = f\left(\hat{\boldsymbol{h}}_t\right) = \frac{1}{2L+1}\boldsymbol{D}_2\exp\left(\boldsymbol{D}_1\hat{\boldsymbol{h}}_t\right), \tag{21}$$

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 $\boldsymbol{D}_1$, and the $(M+1)\times(2L+1)$ matrix $\boldsymbol{D}_2$ are given respectively by

$$D_1(i,j) = e^{-\jmath\omega_i j}, \qquad -L \le i \le L, -C \le j \le C \tag{22}$$

$$D_2(i,j) = e^{\jmath\omega_j i}, \qquad -\frac{M}{2} \le i \le \frac{M}{2}, -L \le j \le L \tag{23}$$

where $\{\omega_{-L}, \ldots, \omega_L\}$ are the sampled frequencies in the spectrum domain, with $\omega_0 = 0$, $\omega_L = \pi$, and $\omega_{-l} = -\omega_l$. By substituting (21) into (20) a cost function relating the MSE with $\hat{\boldsymbol{h}}_t$ can be obtained

$$\varepsilon\left(\hat{\boldsymbol{h}}_t\right) = \frac{1}{N}\left[\boldsymbol{r}_t^\top\boldsymbol{r}_t - 2\boldsymbol{r}_t\boldsymbol{A}_t f\left(\hat{\boldsymbol{h}}_t\right) + f\left(\hat{\boldsymbol{h}}_t^\top\right)\boldsymbol{A}_t^\top\boldsymbol{A}_t f\left(\hat{\boldsymbol{h}}_t\right)\right], \tag{24}$$

where

$$\boldsymbol{r}_t = \boldsymbol{s} - \sum_{j=0, j\neq t}^{T-1}\boldsymbol{A}_j f\left(\hat{\boldsymbol{h}}_j\right). \tag{25}$$

Table 1. Algorithm for complex cepstrum analysis based on the minimum mean squared error.

```
Initialization
1) Initialize $\{p_0, \ldots, p_{Z-1}\}$ as the instants used for initial cepstrum calculation
2) Make $a_z = 1, \quad 0 \le z \le Z-1$
3) Get initial estimates of the complex cepstrum for each frame: $\left\{\hat{\boldsymbol{h}}_0^{(0)}, \ldots, \hat{\boldsymbol{h}}_{T-1}^{(0)}\right\}$
Recursion
1) For each pulse position $\{p_0, \ldots, p_{Z-1}\}$
        1.1) Determine the best position $\hat{p}_z$ using (15)
        1.2) Update the optimal amplitude $\hat{a}_z$ using (14)
2) For each pulse amplitude $\{a_0, \ldots, a_{Z-1}\}$
        2.1) Make $a_z = 0$ if $a_z < 0$, or $a_z = 1$ if $a_z > 0$
3) For each frame $\{t = 0, \ldots, T-1\}$
        3.1) For $i = 1, 2, 3, \ldots$
                3.1.1) Estimate $\hat{\boldsymbol{h}}_t^{(i+1)}$ according to (26)
                3.1.2) Stop if $10\log_{10}\left(\frac{\varepsilon\left(\hat{\boldsymbol{h}}_t^{(i+1)}\right)}{\varepsilon\left(\hat{\boldsymbol{h}}_t^{(i)}\right)}\right) \ge 0$ dB
4) Stop of the SNRseg between $s(n)$ and $\tilde{s}(n)$ is satisfying
```

Since the relationship between cepstrum and impulse response, $\boldsymbol{h}_t = f\left(\hat{\boldsymbol{h}}_t\right)$, is non-linear, a gradient method [12] is utilized to optimize the complex cepstrum. Accordingly, a new estimation for the complex cepstrum can be obtained through

$$\hat{\boldsymbol{h}}_t^{(i+1)} = \hat{\boldsymbol{h}}_t^{(i)} - \gamma\bar{\boldsymbol{\nabla}}_{\hat{\boldsymbol{h}}_t}\varepsilon\left(\hat{\boldsymbol{h}}_t\right), \tag{26}$$

where $\bar{\boldsymbol{\nabla}}_{\hat{\boldsymbol{h}}_t}\varepsilon\left(\hat{\boldsymbol{h}}_t\right) = \frac{\boldsymbol{\nabla}_{\hat{\boldsymbol{h}}_t}\varepsilon\left(\hat{\boldsymbol{h}}_t\right)}{\left\|\boldsymbol{\nabla}_{\hat{\boldsymbol{h}}_t}\varepsilon\left(\hat{\boldsymbol{h}}_t\right)\right\|}$ is the normalized gradient of $\varepsilon\left(\hat{\boldsymbol{h}}_t\right)$ with respect to $\hat{\boldsymbol{h}}_t$, γ is a convergence factor, and i is an iteration index. The gradient vector is given by

$$\boldsymbol{\nabla}_{\hat{\boldsymbol{h}}_t}\varepsilon\left(\hat{\boldsymbol{h}}_t\right) = -\frac{2}{N(2L+1)}\boldsymbol{D}_1^\top \operatorname{diag}\left(\exp\left(\boldsymbol{D}_1\hat{\boldsymbol{h}}_t\right)\right)\boldsymbol{D}_2^\top \boldsymbol{A}_t^\top\left[\boldsymbol{r}_t - \boldsymbol{A}_t f\left(\hat{\boldsymbol{h}}_t\right)\right], \tag{27}$$

where $\operatorname{diag}(\cdot)$ means a diagonal matrix formed with the elements of the argument vector.

3.3. Iterative algorithm

Table 1 shows an algorithm that implements the proposed complex cepstrum analysis. Initial estimates of the complex cepstrum can be derived by conventional complex cepstrum analysis. The respective analysis instants can be used to represent the pulse positions $\{p_0, \ldots, p_{Z-1}\}$. Estimates of initial frame-based complex cepstra can be taken in several ways. One form is to consider each $\hat{\boldsymbol{h}}_t$ vector equal to the complex cepstrum vector obtained in the GCI immediately before frame t. Other possible ways are interpolation of pitch-synchronous cepstra over the frame, or interpolation of amplitude and phase spectra [4].

During the pulse optimization process, negative amplitudes $a_z < 0$ are strong indicators of false GCI detection. To solve this first problem, amplitudes are set to zero $a_z = 0$ whenever the algorithm finds negative amplitudes (recursive Step 2). Naturally, for this solution to make sense, it is assumed that there is no polarity reversal in the initial complex cepstra estimates [7]. In the same step, positive amplitudes are set to one. This is done to force the gain information to be captured by the cepstrum rather than the excitation signal, and it is useful in applications where the excitation signal has to be constructed basically from F_0 information, as in [13].

Table 2. Results of the optimization process for different corpora. APDPM stands for *average percentage of deleted pitch marks. SNRseg values are in dB and represent the average of the sentences.* A, H, N and S mean respectively angry, happy, neutral and sad.

	FZL-A	FZL-H	FZL-N	FZL-S	FLH
Initial SNRseg	2.0	2.9	2.4	2.6	4.5
APDPM	3.0	3.6	2.8	8.5	3.4
Final SNRseg	7.2	7.6	6.3	6.6	10.3
	FSP-A	FSP-H	FSP-N	FSP-S	MGT
Initial SNRseg	1.7	2.6	0.9	3.3	1.3
APDPM	2.1	1.7	26.7	19.7	9.3
Final SNRseg	6.2	7.6	4.8	7.7	7.5

4. EXPERIMENTS

4.1. Speech modeling properties

Experiments of speech analysis and reconstruction using the proposed complex cepstrum analysis method were conducted on 500 utterances, divided as follows: 200 utterances from a female British English speaker (FZL), with 50 utterances spoken in each one the following emotions: angry, happy, neutral, sad; 200 utterances from a female American English speaker (FSP), also with 50 utterances per emotion: angry, happy, neutral, sad; 50 utterances from another female American English speaker (FLH) in neutral style; and 50 sentences from a male American English speaker (MGT) in neutral style. For FZL and FSP, pitch marks were extracted using the Entropic Signal Processing Software [14]. For FLH and MGT pitch marks were extracted by a proprietary tool. Fundamental frequencies were also extracted using the Entropic Signal Processing Software.

Initial cepstra were obtained as in [4]. Cepstrum and filter orders were $C = 512$ and $M = 512$, respectively. The algorithm shown in Table 1 was utilized to derive the final complex cepstra. The maximum number of iterations and convergence factor for complex cepstrum optimization (Step 3) were $\gamma = 0.1$ and 50, respectively. Two recursive iterations were conducted for each sentence. Speech was reconstructed by converting the cepstra into synthesis filter impulse responses according to (21). The corresponding GCIs were used to determine the positions of the unit pulses of the excitation signal, $e(n)$. For the frames where $F_0 = 0$, samples of $e(n)$ were taken from a white noise generator. Speech was reconstructed through (1).

Table 2 shows the average SNRseg (average of the SNRseg from all the sentences of a given corpus) and average percentage of GCIs eliminated per sentence in the process. Speakers FZL and FSP had the lowest gains in terms of SNRseg. This could be due to inaccurate initial pitch marks for FSP (the number of deleted pitch marks is high when compared with the other data), and high F_0 excursions for FZL (from 80 to 500 Hz). Fig 3 shows a high pitched portion of natural speech of FZL-neutral, and its synthesized versions using initial and final cepstra. It can be seen that the proposed method corrects visible distortions on the synthesized speech due to the high F_0.

4.2. Experiments with statistical parametric speech synthesis

To test the method in statistical parametric synthesis [3], 759 FZL-happy utterances, with a rather intense emotional style, and 1071 FZL-neutral utterances were selected. For these corpora two different synthesizers were constructed: one with the initial complex cepstrum, and another with the final complex cepstrum. For statistical modeling, each complex cepstrum set was decomposed into its minimum-phase and all-pass components, and the all-pass component was transformed into phase features, as shown in [4].

The resulting 512-order minimum-phase component and 512-dimension phase parameters were warped onto 39 mel cepstral coefficients and 19 mel phase parameters, respectively [15]. Each observation vector had the following streams: (1) mel cepstrum plus delta and delta-delta; (2,3,4) logarithm of the fundamental frequency, delta, and delta-delta, respectively; (5) 22 band-aperiodicity parameters, calculated as described in [4], plus delta and delta-delta; (6) phase parameters, plus delta and delta-delta. The final system observation vectors were used to train five-state no-skip left-to-right HSMMs [3]. Stream weight was set to zero for the band-aperiodicity and phase parameter streams. At synthesis time, generated mel cepstrum, fundamental frequency, band-aperiodicity and mel phase parameters were used to synthesize speech according to [4].

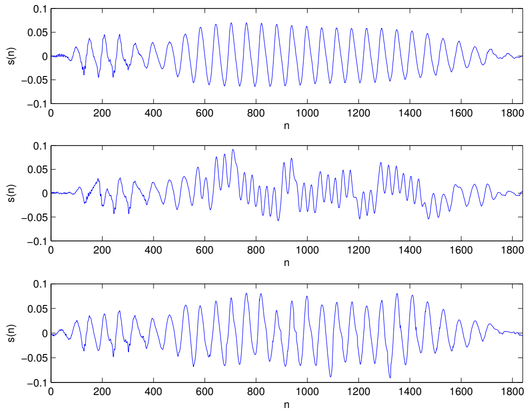

Fig. 3. Examples of waveforms. Top: natural speech; middle: speech reconstructed with the initial cepstra; bottom: speech reconstructed with the final cepstra.

Table 3. Subjects' average preferences. Initial cepstra are extracted as in [4]. *No phase* means system *Ini. Cep.* without phase features.

Corpus	No phase	Ini. cep.	Fin. cep.	No pref.	p-value
FZL-happy	37.7	48.3		14.0	0.004
FZL-happy		39.3	51.9	8.7	0.091
FZL-neutral	43.7	47.5		8.8	0.177
FZL-neutral		44.3	47.3	8.4	0.229

Fifty test sentences were synthesized. For the synthesizer with initial cepstra, the sentences were also synthesized without the phase parameter to simulate the system described in [13]. The test samples were submitted to the Amazon Mechanical Turk. On average, the opinions of 75 listeners were used to calculate the average preferences in each test shown in Table 3. Methods to detect cheating were used during the analysis of the results. Note that the addition of the phase parameter already results in improvement for the baseline system with initial cepstra. The proposed system with optimized cepstra achieved higher preference rates. Important to note the impact of the optimized cepstrum on FZL-happy, which is usually difficult to be processed due to its large F_0 excursions.

5. CONCLUSION

A method for complex cepstrum analysis based on the minimum mean squared error between natural and reconstructed speech has been proposed. The approach searches for the best analysis instants given initial estimates of the complex cepstrum, followed by complex cepstrum re-estimation given the updated analysis locations. This method produces closer to natural reconstructed speech when compared to conventional complex cepstrum analysis methods. Experiments with statistical parametric synthesis also show that the resulting complex cepstrum produces better quality.

6. REFERENCES

[1] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proaks, *Discrete-Time Processing of Speech Signals*, IEEE Press Classic Reissue, New York, 2000.

[2] W. Chu, *Speech Coding Algorithms*, Wiley-Interscience, USA, 2003.

[3] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[4] R. Maia, M. Akamine, and M. F. J. Gales, "Complex cepstrum as phase information for statistical parametric speech synthesis," in *Proc. of ICASSP*, 2012, pp. 4581–4584.

[5] T. F. Quatieri, Jr., "Minimum and mixed phase speech analysis-synthesis by adaptive homomorphic deconvolution," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 4, pp. 328–335, Aug. 1979.

[6] W. Verhelst and O. Steenhaut, "A new model for the short-time complex cepstrum of voiced speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 1, pp. 43–51, Feb. 1986.

[7] A. W. Oppenheim, *Discrete-time signal processing*, Pearson, 2010.

[8] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, pp. 855–866, 2011.

[9] J. M. Tribolet, "A new phase unwrapping algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-25, no. 2, pp. 170–177, Apr. 1977.

[10] B. Bhanu and J. H. McClellan, "On the computation of the complex cepstrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, , no. 5, pp. 583–585, Oct. 1980.

[11] R. Maia, H. Zen, and M.J.F. Gales, "Statistical parametric speech synthesis with the joint estimation of acoustic and excitation parameters," in *Proc. of SSW7*, 2010, pp. 88–93.

[12] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.

[13] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis for Blizzard Challenge 2005," *IEICE Trans. on Inf. and Systems*, vol. E90-D, no. 1, pp. 325–333, 2007.

[14] "ESPS," Entropic Signal Processing Software.

[15] Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai, "Mel-generalized cepstral analysis —a unified approach to speech spectral estimation," in *Proc. of ICSLP*, 1994, pp. 1043–1046.