ITERATIVE ESTIMATION OF PHASE USING COMPLEX CEPSTRUM REPRESENTATION

Ranniery Maia, Yannis Stylianou

Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK

ABSTRACT

This paper presents a method to iteratively estimate phase information from speech in the cepstrum domain. It assumes that correct markings of pitch periods, which may not correspond to glottal closure instants (GCI), are available and can be used to extract the smooth spectral envelope of speech. By using this information, the minimum-phase cepstrum is derived and used as prior information in a modified version of a previously proposed scheme of complex cepstrum analysis based on the mean squared error. Experiments with an emotional database show that the proposed method achieves better performance in terms of continuous phase spectrum estimation, when compared with approaches that rely on accurate GCI markings and high-resolution phase unwrapping mechanisms. In addition, similar results to the full optimization of the complex cepstrum vector are reached, at a lower computational complexity.

Index Terms— Speech representation, speech analysis, phase estimation

1. INTRODUCTION

Phase-aware speech processing methods have been subject of study recently, e.g. [1, 2]. More than a decade ago, some researchers investigated on the usefulness of phase information for speech recognition, e.g. [3]. In the speech coding area, the best performances have been obtained by the family of analysis-by-synthesis coders. Basically, these vocoders usually build a sophisticate excitation signal, which together with the minimum-phase synthesis filter, attempts to reproduce the mixed-phase nature of speech [4]. In speech synthesis, more specifically in statistical parametric speech synthesis [5], although in the past it has been implied that the use of phase does not result in better synthesized speech quality due to the nature of statistical machines based on hidden Markov models (HMMs), recent advances in deep learning for text-to-speech (TTS) [6, 7] have create new frameworks in which the inclusion of phase information can lead to improvement. In fact, even for HMMs, if phase is properly estimated it can increase naturalness of synthesized speech, as shown in our previous work [8]. In many ways, vocoding methods that attempt to mimic the glottal flow or residual information are indeed implicitly moving beyond the minimum-phase assumption and consequently modeling phase information.

This paper presents an approach to phase estimation through a modified version of our previous work on complex cepstrum analysis based on mean squared error (MSE) [9]. The complex cepstrum is a set of parameters that theoretically contains the full information of the speech signal: amplitude and phase. However, in practical terms, complex cepstrum estimation is difficult to achieve due to inaccuracies in speech segmentation, due to the need of the detection of glottal closure instants (GCI), and phase unwrapping. These problems have been alleviated by the MSE-based complex cepstrum analysis. Although the referred scheme had advantages against conventional complex cepstrum analysis such as no need for windowing, no need for phase unwrapping, use of soft glottal closure instant (GCI) marks, and estimation of frame-based cepstra, the method suffered from high computational complexity. The phase estimation method proposed here is a simplification of the MSE complex cepstrum analysis scheme. By assuming that the initial pitch marks are good enough, a good estimation of the spectral envelope of speech can be obtained, and consequently the minimum-phase cepstrum. Since the minimum-cepstrum is given, only the all-pass component of the complex cepstrum, which is related to the residual (or dispersion) phase, need to be estimated. The proposed method has the advantages of the MSE cepstrum analysis when it comes to representing phase information, at a lower computational complexity.

This paper is organized as follows. Section 2 outlines the method of complex cepstrum analysis based on the MSE. Section 3 describes the proposed method for phase estimation. Experiments are shown in Section 4, and the conclusions in Section 5.

2. MSE-BASED COMPLEX CEPSTRUM ANALYSIS

2.1. Complex cepstrum-based speech modeling

In our approach of complex cepstrum-based speech modeling, the speech signal, $s(n)$, is assumed to be produced by the following convolution: $s(n) = h(n) * e(n)$, where $h(n)$ is a slowly varying impulse response representing the effects of the glottal flow, vocal tract, and lip radiation. The excitation signal, $e(n)$, is composed of pulses located at the pitch period onset times. The cepstrum of $s(n)$, $\hat{h}(n)$, is given by

$$\hat{h}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\Big\{\ln|S(e^{\jmath\omega})| + \jmath\theta(\omega)\Big\}e^{\jmath\omega n}d\omega, \qquad (1)$$

for $-C \leq n \leq C$, where C is the cepstral order and $|S(e^{\jmath\omega})|$ and $\theta(\omega)$ are respectively the amplitude and phase spectrum of $s(n)$. To synthesize speech, $\hat{h}(n)$ must be converted into the impulse response

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

for $-P \leq n \leq P$, P is the impulse response order and $H(e^{\jmath\omega})$ is the complex spectrum of $h(n)$. Finally, speech is reconstructed by making $s(n) = h(n) * e(n)$.

2.2. Iterative estimation of the complex cepstrum

The MSE-based approach overcomes two main issues of conventional complex cepstrum analysis [9]: (1) no need for accurate GCI markings; (2) no need for phase unwrapping. MSE complex cepstrum analysis (MSE-CCEP) is performed in a two-step procedure: (1) estimation of the analysis instants; (2) cepstrum optimization at the frame level using a gradient method. Fig. 1 illustrates the MSE-CCEP process.

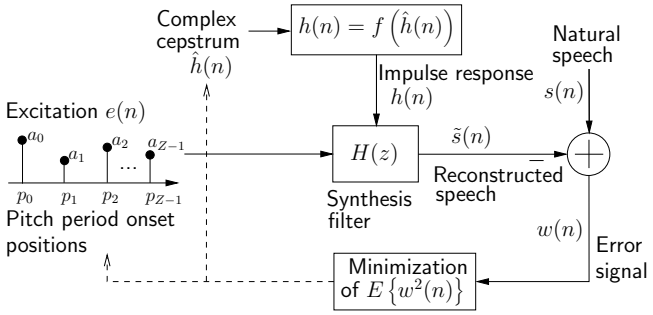

Fig. 1. MSE-based complex cepstrum analysis.

2.2.1. *Pulse optimization*

Pitch period onset position estimation is done by keeping $\hat{h}(n)$ 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 pulses. By using matrix notation, the MSE between $s(n)$ and $\tilde{s}(n)$ (Fig. 1), is

$$\varepsilon\left(a_z, p_z\right) = \frac{1}{N}\left[\boldsymbol{r}_z^\top \boldsymbol{r}_z - 2a_z \boldsymbol{g}_{p_z}^\top \boldsymbol{r}_z + a_z^2 \boldsymbol{g}_{p_z}^\top \boldsymbol{g}_{p_z}\right], \quad (3)$$

with N being the number of samples of $s(n)$, $\boldsymbol{r}_z = \boldsymbol{s} - \sum_{j=0, j\neq z}^{Z-1} a_j \boldsymbol{g}_{p_j}$ the target signal for pulse z, and

$$\boldsymbol{s} = \left[\underbrace{0 \ \cdots \ 0}_{P} \quad s\left(0\right) \quad \cdots \quad s\left(N-1\right) \quad \underbrace{0 \ \cdots \ 0}_{P}\right]^\top, \quad (4)$$

$$\boldsymbol{g}_n = \left[\underbrace{0 \ \cdots \ 0}_{n} \quad \boldsymbol{h}_n^\top \quad \underbrace{0 \ \cdots \ 0}_{N-n-1}\right]^\top, \quad (5)$$

$$\boldsymbol{h}_n = \left[h_n\left(-P\right) \quad \cdots \quad h_n(P)\right]^\top, \quad (6)$$

where $\boldsymbol{h}_n$ contains the impulse response $h(n)$ at the n-th sample position. The z-th pulse position which minimizes (3) can be found by making $\frac{\partial \varepsilon(a_z, p_z)}{\partial a_z} = 0$,

$$\hat{p}_z = \underset{p_z = p_z - \Delta p, \ldots, p_z + \Delta p}{\arg\max} \frac{\left(\boldsymbol{g}_{p_z}^\top \boldsymbol{r}_z\right)^2}{\boldsymbol{g}_{p_z}^\top \boldsymbol{g}_{p_z}}, \quad (7)$$

where Δp is the range of samples, and $\hat{a}_z = \frac{\boldsymbol{g}_{p_z}^\top \boldsymbol{r}_z}{\boldsymbol{g}_{p_z}^\top \boldsymbol{g}_{p_z}}$ is used to update all the amplitudes.

2.2.2. *Complex cepstrum optimization*

For complex cepstrum estimation, the MSE as function of $\hat{\boldsymbol{h}}_t$, where t is frame index, becomes

$$\varepsilon\left(\hat{\boldsymbol{h}}_t\right) = \frac{1}{N}\left[\boldsymbol{r}_t^\top \boldsymbol{r}_t - 2\boldsymbol{r}_t \boldsymbol{A}_t f\left(\hat{\boldsymbol{h}}_t\right) + f\left(\hat{\boldsymbol{h}}_t^\top\right) \boldsymbol{A}_t^\top \boldsymbol{A}_t f\left(\hat{\boldsymbol{h}}_t\right)\right], \quad (8)$$

where $\boldsymbol{r}_t = \boldsymbol{s} - \sum_{j=0, j\neq t}^{T-1} \boldsymbol{A}_j f\left(\hat{\boldsymbol{h}}_j\right)$ is the target vector at frame t, T is the number of frames in the sentence, and $\hat{\boldsymbol{h}}_t$ in the t-th complex cepstrum vector. The $(K+M)\times(M+1)$ matrix $\boldsymbol{A}_t$ is given by

$$\boldsymbol{A}_t = \left[\boldsymbol{u}_{-P} \quad \cdots \quad \boldsymbol{u}_P\right], \quad (9)$$

$$\boldsymbol{u}_m = \left[\underbrace{0 \ \cdots \ 0}_{P+m} \quad \boldsymbol{e}_t^\top \quad \underbrace{0 \ \cdots \ 0}_{P-m}\right]^\top, \quad (10)$$

$$\boldsymbol{e}_t = \left[\underbrace{0 \ \cdots \ 0}_{tK} \ e\left(tK\right) \ \cdots \ e\left((t+1)K-1\right) \ \underbrace{0 \ \cdots \ 0}_{N-(t+1)K}\right]^\top, \quad (11)$$

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. The relationship between $\boldsymbol{h}_t = \left[h_t(-P) \quad \cdots \quad h_t(P)\right]^\top$ and $\hat{\boldsymbol{h}}_t = \left[\hat{h}_t(-C) \quad \cdots \quad \hat{h}_t(C)\right]^\top$ can be written as

$$\boldsymbol{h}_t = f\left(\hat{\boldsymbol{h}}_t\right) = \frac{1}{2L}\boldsymbol{D}_2 \exp\left(\boldsymbol{D}_1 \hat{\boldsymbol{h}}_t\right), \quad (12)$$

where the elements of $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are given respectively by

$$D_1(i,j) = e^{-\jmath\omega_i j}, \qquad -L+1 \leq i \leq L, -C \leq j \leq C \quad (13)$$

$$D_2(i,j) = e^{\jmath\omega_j i}, \qquad -P \leq i \leq P, -L+1 \leq j \leq L \quad (14)$$

with $\{\omega_{-L+1}, \ldots, \omega_L\}$ being the sampled frequencies in the spectrum domain, with $\omega_0 = 0$, $\omega_L = \pi$, and $\omega_{-l} = -\omega_l$. A new estimation for the complex cepstrum can be given by

$$\hat{\boldsymbol{h}}_t^{(k+1)} = \hat{\boldsymbol{h}}_t^{(k)} - \gamma \bar{\boldsymbol{\nabla}}_{\hat{\boldsymbol{h}}_t} \varepsilon\left(\hat{\boldsymbol{h}}_t\right), \quad (15)$$

where $\bar{\boldsymbol{\nabla}}_{\hat{\boldsymbol{h}}_t} \varepsilon\left(\hat{\boldsymbol{h}}_t\right) = \frac{\boldsymbol{\nabla}_{\hat{\boldsymbol{h}}_t} \varepsilon(\hat{\boldsymbol{h}}_t)}{\left\|\boldsymbol{\nabla}_{\hat{\boldsymbol{h}}_t} \varepsilon(\hat{\boldsymbol{h}}_t)\right\|}$ 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 k is an iteration index. The gradient vector is given by

$$\boldsymbol{\nabla}_{\hat{\boldsymbol{h}}_t} \varepsilon\left(\hat{\boldsymbol{h}}_t\right) = -\frac{1}{NL}\boldsymbol{D}_1^\top \operatorname{diag}\left[\exp\left(\boldsymbol{D}_m \hat{\boldsymbol{h}}_t\right)\right] \boldsymbol{D}_2^\top \boldsymbol{A}_t^\top \left[\boldsymbol{r}_t - \boldsymbol{A}_t f\left(\hat{\boldsymbol{h}}_t\right)\right], \quad (16)$$

and $\operatorname{diag}(\cdot)$ means diagonal matrix made with argument vector.

3. ITERATIVE ESTIMATION OF PHASE USING MSE COMPLEX CEPSTRUM ANALYSIS

3.1. The idea

In MSE-CCEP, the pulse positions $\{p_0, \ldots, p_{Z-1}\}$ are optimized so that they: (1) are pitch-synchronous; (2) indicate somewhere nearby the GCIs. In (1), the goal is to remove the F_0 effect, so that $\hat{h}(n)$ may represent the smooth spectral envelope. In (2), the goal is the removal of the linear phase component of the phase response.

In many situations it is straightforward to obtain *pitch marks*, i.e. indications of pitch periods, rather than exact moments where phase information can be theoretically retrieved, such as GCIs [10, 11, 12] or related instants where phase can be estimated [13]. Therefore, by assuming initial pitch marks, a simplified version of MSE-CCEP can be used for phase estimation solely. This is done by removing the contribution of the minimum-phase cepstrum, which is derived from the amplitude spectrum of speech. For this, minimum-phase/all-pass factorization of speech in the cepstral domain is used.

For the proposed approach to work, the following conditions regarding the excitation signal $e(n)$ in Fig 1 should ideally be met: (1) there are no missing pulses (no pitch period left *unmarked*); (2) positions $\{p_0, \ldots, p_{Z-1}\}$ accurately mark pitch periods, although they may not indicate the GCI. By assuming that, the real cepstrum, $\hat{h}_r(n)$, can be obtained from the amplitude spectrum of speech as follows

$$\hat{h}_r(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|S\left(e^{\jmath\omega}\right)\right| e^{\jmath\omega n} d\omega, \quad (17)$$

for $-C \leq n \leq C$, and the minimum-phase cepstrum by making

$$\hat{h}_m(n) = \begin{cases} \hat{h}_r(n), & n = 0, \\ 2 * \hat{h}_r(n), & 0 < n \leq C, \end{cases} \quad (18)$$

In this case, it is necessary to estimate only the so-defined *all-pass cepstrum*, which is the difference between complex and minimum-phase cepstra

$$\hat{h}_a(n) = \hat{h}(n) - \hat{h}_m(n) = \begin{cases} \hat{h}(n), & -C \leq n < 0, \\ 0, & n = 0, \\ -\hat{h}(-n), & 0 < n \leq C. \end{cases} \quad (19)$$

since $\hat{h}_m(n) = \hat{h}(n) + \hat{h}(-n)$, for $1 \leq n \leq C$. The all-pass cepstrum, $\{h_a(-C), \ldots, h_a(C)\}$, represents the *residual phase* (or dispersion phase) of speech. The residual phase contains the phase response of the glottal flow and represents an important factor in order to achieve high-quality speech parameterization [14].

3.2. Iterative estimation of phase

Fig. 2 shows the proposed method for phase estimation, derived as a simplification of the MSE-CCEP framework. Basically, the differences between this method and the one outlined in Section 2 are:

- the minimum-phase cepstrum is calculated directly from natural speech by using the positions $\{p_0, \ldots, p_{Z-1}\}$;
- optimization is done on the all-pass cepstrum, $\hat{h}_a(n)$;
- optimized all-pass cepstrum, $\hat{h}_a(n)$, and fixed minimum-phase cepstrum, $\hat{h}_m(n)$, are added together to compose the complex cepstrum, $\hat{h}(n)$.

Therefore, like in the MSE-CCEP method, phase estimation is performed in a two-step optimization procedure. The first one corresponds to pulse optimization, which is exactly the same scheme as described in Section 2.2.1, and for this reason will be omitted.

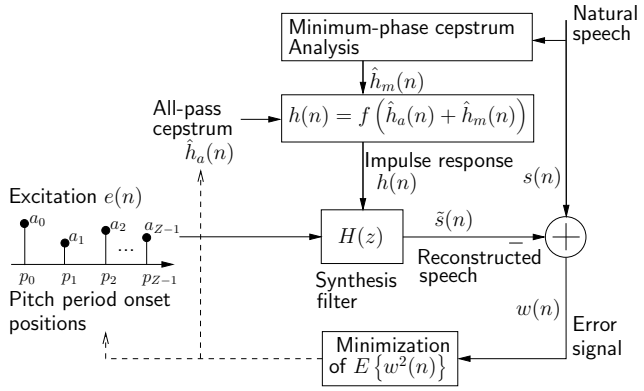

Fig. 2. Proposed method for iterative estimation of phase.

3.2.1. All-pass cepstrum optimization

Since $\hat{h}(n) = \hat{h}_m(n) + \hat{h}_a(n)$, from Fig. 2 the MSE between $s(n)$ ad $\tilde{s}(n)$, as function of $\hat{\boldsymbol{h}}_{a,t} = \left[\hat{h}_{a,t}(-C) \quad \cdots \quad \hat{h}_{a,t}(C)\right]^\top$, is

$$\varepsilon\left(\hat{\boldsymbol{h}}_{a,t}\right) = \frac{1}{N}\Bigg[\boldsymbol{r}_t^\top \boldsymbol{r}_t - 2\boldsymbol{r}_t \boldsymbol{A}_t f\left(\hat{\boldsymbol{h}}_{m,t} + \hat{\boldsymbol{h}}_{a,t}\right) + + f\left(\hat{\boldsymbol{h}}_{m,t} + \hat{\boldsymbol{h}}_{a,t}\right)^\top \boldsymbol{A}_t^\top \boldsymbol{A}_t f\left(\hat{\boldsymbol{h}}_{m,t} + \hat{\boldsymbol{h}}_{a,t}\right)\Bigg]. \quad (20)$$

By exploiting the fact that the all-pass cepstrum is anti-symmetric, i.e. $\hat{h}_a(n) = -\hat{h}_a(-n)$, with $\hat{h}_a(0) = 0$, a vector of *phase features* can be defined as: $\boldsymbol{\phi}_t = \left[\hat{h}_{a,t}(1) \quad \cdots \quad \hat{h}_{a,t}(C)\right]^\top$. By using the gradient descent approach in order to optimize $\boldsymbol{\phi}_t$, similarly to the complex cepstrum optimization case shown in Section 2.2.2, a new update for vector $\boldsymbol{\phi}_t$ can be given by

$$\boldsymbol{\phi}_t^{(k+1)} = \boldsymbol{\phi}_t^{(k)} - \gamma \bar{\boldsymbol{\nabla}}_{\boldsymbol{\phi}_t} \varepsilon\left(\boldsymbol{\phi}_t\right), \quad (21)$$

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

$$\boldsymbol{\nabla}_{\hat{\boldsymbol{h}}_t} \varepsilon\left(\boldsymbol{\phi}_t\right) = -\frac{1}{NL} \boldsymbol{D}_{a,1}^\top \operatorname{diag}\left[\exp\left(\boldsymbol{D}_{m,1} \hat{\boldsymbol{h}}_{m,t} + \boldsymbol{D}_{a,1} \boldsymbol{\phi}_t\right)\right] \boldsymbol{D}_2^\top \boldsymbol{A}_t^\top \left[\boldsymbol{r}_t - \boldsymbol{A}_t f\left(\hat{\boldsymbol{h}}_{m,t} + \hat{\boldsymbol{h}}_{a,t}\right)\right], \quad (22)$$

where matrix $\boldsymbol{D}_{m,1}$ is the same as $\boldsymbol{D}_1$, but for $0 \leq j \leq C$, and the elements of $\boldsymbol{D}_{a,1}$ are given by

$$D_{a,1}(i,j) = -2\jmath \sin \omega_i j, \quad -L+1 \leq i \leq L, 1 \leq j \leq C. \quad (23)$$

4. EXPERIMENTS

4.1. Conditions

Some studies suggest that the glottal flow derivative is connected to the speaker's voice style or emotion [15, 16]. Although this is true for the spectral tilt, i.e. the amplitude response of the glottal flow, here we utilize an emotional database to check the performance of the proposed algorithm in order to verify whether it works satisfactorily for different voice styles. The data comprises 50 sentences uttered in six different styles: anger, fear, happiness, neutral), sadness, and tenderness. The data was recorded in studio by a female British actress. The audio was originally sampled at 48 kHz and down-sampled to 22.05 kHz.

We compared the proposed approach for phase estimation with two methods: (1) pitch synchronous spectral analysis at the GCI, retrieved by the DYPSA algorithm [10], followed by high-resolution phase unwrapping using the simple algorithm shown in [8]; (2) MSE-CCEP approach [9] outlined in Section 2. For the first method, anti-symmetric Hann windows were centered at the GCIs, covering two pitch periods. Then a 4096-point fast Fourier transform (FFT) was taken to obtain a high-resolution wrapped phase spectrum, and increase the accuracy of the unwrapping process. For both MSE-CCEP and proposed algorithm, initialization was done with the GCIs provided by DYPSA and with complex cepstrum estimated as in [8]. P and L were set to $P = 128$ and $L = 512$ in both methods.

4.2. Evaluation criterion

The evaluation criterion is based on speech signal analysis and reconstruction using original amplitude and estimated phase spectra, as shown in Fig. 3. In the baseline approach, the phase spectrum is obtained by unwrapping the principal value of the phase, and then linear phase term removal. For the MSE-CCEP and proposed approach, since both methods yield a complex cepstrum at the end[1], the continuous phase response can be retrieved by

$$\theta(\omega) = -\sum_{n=-C}^{C} \hat{h}(n) \sin \omega n. \quad (24)$$

[1]Note that in the proposed method the minimum-phase cepstrum is calculated initially from natural speech and kept constant during the entire process.

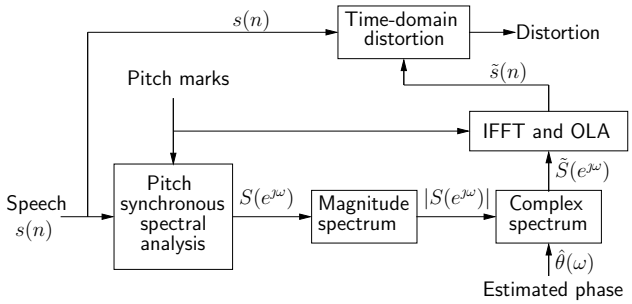

Fig. 3. Evaluation of phase estimation.

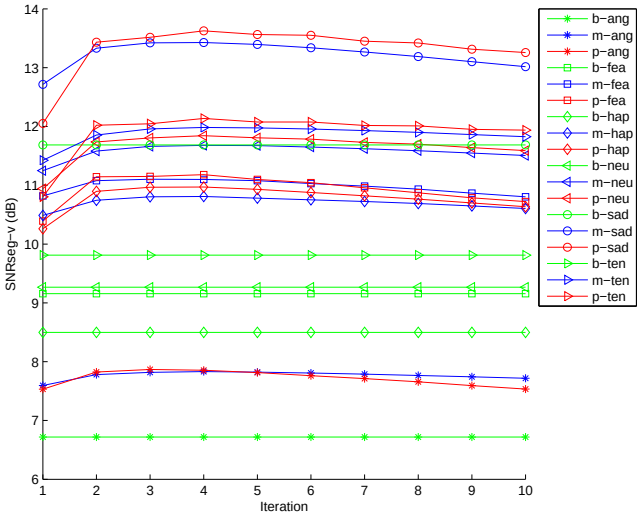

Fig. 4. SNRseg-v for the compared methods in all the speech styles. *b*, *m* and *p* mean respectively baseline, MSE-CCEP, and proposed.

4.3. Results

Fig. 4 shows the results of the compared methods for the six difference voice styles: anger (ang), fear (fea), happiness (hap), neutral (neu), tender (ten), sadness (sad). Note that 10 iterations were used for MSE-CCEP and proposed method. For sake of visualization, the values obtained by the baseline are depicted as constant lines across the iterations. As distortion measure, the segmented signal-to-noise ratio of voiced frames (SNRseg-v) was used. It can be seen that while better than the conventional approach for all the styles, the proposed method achieves performance similar to the MSE-CCEP, at a lower computational complexity. Interesting to note is that all the systems performed much better for sad than for ang. This was expected since phase estimation for highly expressive voices is challenging. In terms of CPU processing time, the proposed approach runs in average three times faster than MSE-CCEP. Informal listenings showed that there were no difference between the two methods.

To check if the proposed method was similar to MSE-CCEP in terms of vocoded speech as well, we used the final complex cepstra resulted from both methods to derive non-causal impulse responses, $h(n)$, and F_0 and optimized $\{p_0, \ldots, p_{Z-1}\}$ to create the excitation signals, $e(n)$. Then speech was reconstructed by making $\tilde{s}(n) = h(n) * e(n)$. Results in terms of SNRseg-v across 10 iterations are shown in Fig. 5. In this case it is clear that the MSE-CCEP approach, which optimizes the entire complex cepstrum vector, achieves better performance for all the speech styles. This is illustrated in Fig. 6, which shows amplitude and phase spectra derived from the cepstrum obtained at the end of the optimization process, in both MSE-CCEP and proposed method. It can be noticed that the difference comes mostly from the minimum-phase cepstrum, which is also optimized together with phase information in MSE-CCEP. This is shown by the differences in amplitude and minimum-phase spectra of Fig. 6.

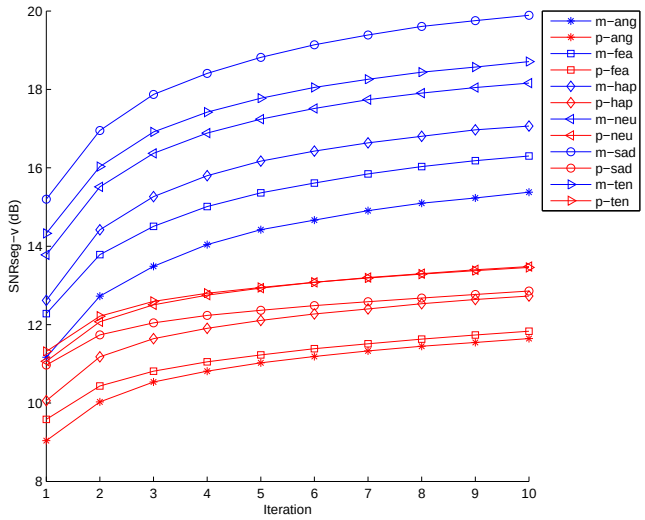

Fig. 5. SNRseg-v for MSE-CCEP and proposed methods in all speech styles, using optimized cepstra to re-synthesize speech. *m* and *p* mean respectively MSE-CCEP and proposed.

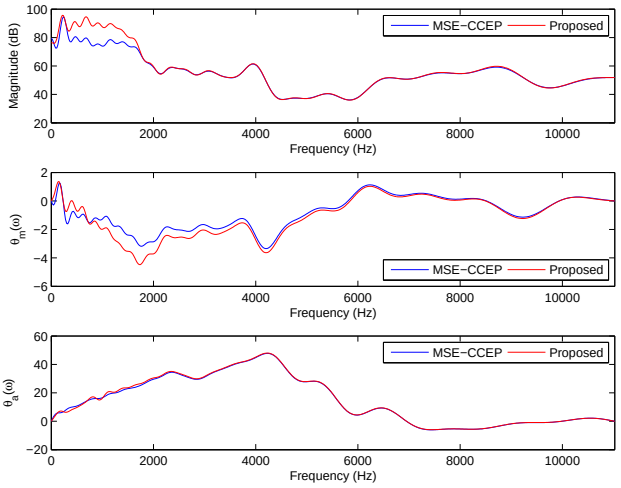

Fig. 6. Example of amplitude (top), minimum-phase (middle) and residual phase (bottom) derived from the complex cepstrum from the optimization process in MSE-CCEP and proposed approach.

5. CONCLUSIONS

This paper presented an approach for phase estimation based on a simplified version of our previously published work on MSE complex cepstrum analysis. In the proposed scheme, rather than the full complex cepstrum vector, only the all-pass cepstrum is optimized. The idea, however, assumes good enough initial pitch marks in the sense that they effectively mark pitch periods, with little missing problems. Experiments on an emotional database showed that the proposed method achieves much better performance than GCI marking followed by multi-resolution spectral analysis and phase unwrapping, while reaching similar performance to the full optimization of to the complex cepstrum with less computational load.

6. REFERENCES

[1] T. Gerkmann, M. Krawczyk, and Robert Rehr, "Phase estimation in speech enhancement - unimportant, important, or impossible?," in *Proc. of IEEE Convention of Electrical and Electronics Engineers in Israel*, 2012, pp. 1–5.

[2] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Phase importance in speech processing applications," in *Proc. of Interspeech*, 2015, pp. 1623–1627.

[3] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. of Eurospeech*, 2003, pp. 2117–2120.

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

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

[6] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. of Interspeech*, 2014, pp. 1964–1968.

[7] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. of ICASSP*, 2015, pp. 4470–4474.

[8] 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.

[9] R. Maia, M. Akamine, and M.J.F Gales, "Complex cepstrum analysis based on the minimum mean squared error," in *Proc. of ICASSP*, 2013, pp. 7972–7976.

[10] P.A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[11] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 16021613, Nov. 2008.

[12] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, Nov. 2011.

[13] R. Maia, Y. Stylianou, and M. Akamine, "A maximum likelihood approach to the detection of moments of maximum excitation and its application to high-quality speech parameterization," in *Proc. of Interspeech*, 2015, pp. 603–607.

[14] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer speech and language*, pp. 20–34, Apr. 2011.

[15] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Proc. of Interspeech*, 2007, pp. 1410–1413.

[16] É. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameteres," in *Proc. of Interspeech*, 2011, pp. 2409–2412.