COMPLEX CEPSTRUM AS PHASE INFORMATION IN STATISTICAL PARAMETRIC SPEECH SYNTHESIS

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

Statistical parametric synthesizers usually rely on a simplified model of speech production where a minimum-phase filter is driven by a zero or random phase excitation signal. However, this procedure does not take into account the natural mixed-phase characteristics of the speech signal. This paper addresses this issue by proposing the use of the complex cepstrum for modeling phase information in statistical parametric speech synthesizers. Here a frame-based complex cepstrum is calculated through the interpolation of pitch-synchronous magnitude and unwrapped phase spectra. The non-causal part of the frame-based complex cepstrum is then modeled as phase features in the statistical parametric synthesizer. At synthesis time, the generated phase parameters are used to derive coefficients of a glottal filter. Experimental results show that the proposed approach effectively embeds phase information in the synthetic speech, resulting in close-to-natural waveforms and better speech quality.

Index Terms— Speech synthesis, statistical parametric speech synthesis, spectral analysis, cepstral analysis, complex cepstrum.

1. INTRODUCTION

Various studies, e.g. [1], have reported the importance of phase information for speech processing systems. For systems that assume the source-filter model of speech production, investigations into the usefulness of phase information have mostly concentrated on glottal pulse models, such as the Liljencrants-Fant (LF) model [2]. The idea is that the glottal pulses models can introduce phase information which is *lost* due to the minimum-phase vocal-tract filter assumption in simplified source-filter models [3]. For statistical parametric speech synthesis [4], glottal pulse models based on the estimation of the LF model parameters, and glottal inverse filtering have been used, e.g. [5, 6]. In terms of using explicit phase parameters, in [7] the use of specific *phase features* was reported to be successful for unit-concatenation-based systems. However, there was no investigation of their use with statistical parametric synthesis.

This paper proposes the use of complex cepstrum for incorporating phase information in statistical parametric speech synthesis. The use of complex cepstrum for glottal source estimation has been recently discussed in the area of speech analysis [8]. It has been reported to be advantageous in terms of computational cost, for instance when compared with the zeros of the z-transform. From the speech production mechanism viewpoint, the use of complex cepstrum theoretically has a clear advantage against the commonly used complex cepstrum of minimum phase sequences[1] since it represents the natural mixed-phase characteristics of the speech signal. Despite this advantage, the use of complex cepstrum for speech processing has some drawbacks due to analysis issues such as the need for pitch-synchronous analysis and phase unwrapping. This paper addresses these problems by proposing a frame-based calculation of the complex cepstrum. One approach to derive a frame-based complex cepstrum is based on chirp analysis [9]. Alternatively, in this work the interpolation of pitch-synchronous spectra was chosen as this has less computational complexity, and still yields an accurate representation of the speech spectral envelope by removing the F_0 effect [10]. Frame-based complex cepstra are then decomposed into all-pass and minimum-phase components. The all-pass cepstra, which are directly related to the non-causal part of the complex cepstrum, are represented as *phase parameters* and modelled by hidden semi-Markov models (HSMM) in an additional observation vector stream. At synthesis time, the generated phase parameters are used to implement an all-pass glottal filter.

Section 2 outlines speech analysis/synthesis using complex cepstrum and Section 3 describes the incorporation of complex cepstrum into HMM-based speech synthesis. Experiments are shown in Section 4, and the conclusions are in Section 5.

[1]Henceforth regarded as the *minimum-phase cepstrum*.

2. COMPLEX CEPSTRUM ANALYSIS/SYNTHESIS

Cepstral analysis is an outcome of the field of homomorphic deconvolution [11]. The complex cepstrum $\hat{s}(n)$ of a $N+1$-sample windowed speech segment $s(n)$ is given by

$$S\left(e^{\jmath\omega}\right) = \sum_{n=-N/2}^{N/2} s(n) e^{-\jmath\omega n}, \tag{1}$$

$$\ln S\left(e^{\jmath\omega}\right) = \ln\left|S\left(e^{\jmath\omega}\right)\right| + \jmath\theta(\omega), \tag{2}$$

$$\hat{s}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln S\left(e^{\jmath\omega}\right) e^{\jmath\omega n} d\omega, \tag{3}$$

where $S\left(e^{\jmath\omega}\right)$ is the Discrete-Time Fourier Transform (DTFT) of $s(n)$ and $\theta(\omega)$ is the unwrapped phase spectrum [11]. The complex cepstrum is an infinite, non-causal and non-symmetric sequence. The impulse response related to the spectral envelope of $s(n)$, $x(n)$, can be obtained from a truncated version of $\hat{s}(n)$ through

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

$$X\left(e^{\jmath\omega}\right) = \exp\left\{\mathrm{Re}\left[\hat{X}\left(e^{\jmath\omega}\right)\right] + \jmath\,\mathrm{Im}\left[\hat{X}\left(e^{\jmath\omega}\right)\right]\right\}, \tag{5}$$

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X\left(e^{\jmath\omega}\right) e^{\jmath\omega n} d\omega, \tag{6}$$

where C is the cepstral order, and $\hat{x}(n)$ is a truncated version of $\hat{s}(n)$ so that $\hat{x}(n) = 0$ for $|n| > C$.

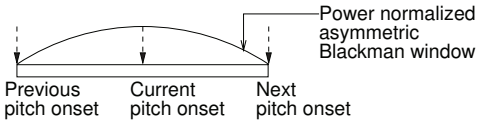

Fig. 1. Windowing for pitch-synchronous spectral analysis.

For statistical parametric speech synthesis the minimum-phase cepstrum is normally used. However, the speech signal is naturally a mixed-phase sequence. This minimum-phase assumption for the synthesis filter comes from the theory of linear prediction analysis, where the inverse synthesis filter must be causal and stable [3] so as to minimize the energy of the so-called *residual signal* during the calculation of the linear prediction coefficients. The minimum-phase cepstrum is a causal sequence, i.e. $\hat{x}_m(n) = 0, n < 0$, and can be obtained from the complex cepstrum $\hat{x}(n)$ through a simple operation in the quefrency domain (see Section 3.2).

3. COMPLEX CEPSTRUM AS PHASE FEATURES FOR STATISTICAL PARAMETRIC SYNTHESIS

To use phase information from complex cepstrum in statistical parametric synthesis, the following issues must be addressed: (1) frame-based complex cepstrum analysis; (2) acoustic modeling of phase parameters; (3) and synthesis with phase features.

3.1. Frame-based complex cepstrum

In this work the frame-based cepstrum are obtained through the interpolation of pitch-synchronous amplitude and phase responses.

From the literature, the location of the window for complex cepstrum analysis should be at the pitch onsets in order to yield an appropriate phase representation [11, 8]. To enable this, windowing is performed with asymmetric Blackman windows as illustrated in Fig. 1. Pitch-synchronous spectra $S\left(e^{\jmath\omega}\right)$ can then be obtained by taking the DTFT of the windowed speech segments.

After extracting the spectra phase unwrapping must be performed. By noting that $\{\theta(\omega_0), \ldots, \theta(\omega_L)\}$ are $L+1$ samples of $\theta(\omega)$ between $\omega_0 = 0$ and $\omega_L = \pi$, phase unwrapping involves

$$\theta(\omega_l) = \begin{cases} \Theta(\omega_l), & l = 0, \\ \Theta(\omega_l) - \sum_{k=0}^{l-1} \psi(k), & l = 1, \ldots, L, \end{cases} \tag{7}$$

where $\{\Theta(\omega_0), \ldots, \Theta(\omega_L)\}$ are samples of the phase function modulo 2π, and

$$\psi(k) = \left\lfloor \frac{\Theta(\omega_{k+1}) - \Theta(\omega_k) + \pi}{2\pi} \right\rfloor 2\pi, \quad k = 0, \ldots, L-1. \tag{8}$$

The final step is to remove the linear phase component using

$$\theta(\omega_l) = \theta(\omega_l) - \frac{l}{L}\theta(\pi), \quad l = 1, \ldots, L. \tag{9}$$

Once pitch-synchronous spectra have been calculated, frame-based complex cepstrum can be obtained through linear interpolation. In this work the interpolation of features in three different domains were investigated:

1. complex spectrum: $S\left(e^{\jmath\omega}\right) = \mathrm{Re}\left[S\left(e^{\jmath\omega}\right)\right] + \jmath\,\mathrm{Im}\left[S\left(e^{\jmath\omega}\right)\right]$;
2. magnitude and unwrapped phase spectrum: $\left|S\left(e^{\jmath\omega}\right)\right|, \theta(\omega)$;
3. complex cepstrum: $\hat{x}(n)$.

Table 1. SNRseg and LSDs (dB) results for the interpolation of the three different features described in Section 3.1.

Interpolated of feature	1	2	3
SNRseg	1.54	1.60	1.64
LSD	4.14	3.95	4.01

To evaluate the interpolation schemes, 100 sentences uttered by an American English speaker sampled at 16 kHz were used. The speech signals were analyzed, and re-synthesized using frame-based cepstrum based on (4), (5), and (6) to obtain a non-causal synthesis filter impulse response, with $C = 39$. The filter was driven by a simple excitation signal constructed from the extracted F_0. Two objective measures were used to evaluate the distortion between natural and re-synthesized speech: (1) segmental signal-to-noise ratio (SNRseg) [3]; and (2) log spectral distance (LSD) given by

$$d_{LS} = \sqrt{\frac{100}{2L+1}\left\{\left[\log\frac{\left|S\left(e^{\jmath\omega_0}\right)\right|}{\left|\tilde{S}\left(e^{\jmath\omega_0}\right)\right|}\right]^2 + 2\sum_{l=1}^{L}\left[\log\frac{\left|S\left(e^{\jmath\omega_l}\right)\right|}{\left|\tilde{S}\left(e^{\jmath\omega_l}\right)\right|}\right]^2\right\}}, \tag{10}$$

where $\left\{S\left(e^{\jmath\omega_0}\right), \ldots, S\left(e^{\jmath\omega_L}\right)\right\}$ and $\left\{\tilde{S}\left(e^{\jmath\omega_0}\right), \ldots, \tilde{S}\left(e^{\jmath\omega_L}\right)\right\}$ are samples between $\omega_0 = 0$ and $\omega_L = \pi$ of the DTFT of natural and re-synthesized speech, respectively, and $\log$ means the common logarithm. The LSD by (10) is calculated for each frame and its mean value used as the final measure. Table 1 shows the results for $L = 512$. It can be seen that there is basically no difference between interpolation of pitch-synchronous complex cepstrum and amplitude and phase response. As it is more flexible, the latter was chosen to derive frame-based complex cepstra in this work.

3.2. Acoustic modeling of phase information

A given sequence $x(n)$, for which the complex cepstrum $\hat{x}(n)$ exists, can be decomposed into its minimum-phase, $x_m(n)$, and all-pass, $x_a(n)$, components [11]. Thus

$$x(n) = x_m(n) * x_a(n). \tag{11}$$

The minimum-phase cepstrum, $\hat{x}_m(n)$, is a causal sequence and can be obtained from the complex cepstrum, $\hat{x}(n)$, as follows

$$\hat{x}_m(n) = \begin{cases} 0, & n = -C, \ldots, -1, \\ \hat{x}(n), & n = 0, \\ \hat{x}(n) + \hat{x}(-n), & n = 1, \ldots, C, \end{cases} \tag{12}$$

where C is the cepstral order. The all-pass cepstrum $\hat{x}_a(n)$ can then be simply retrieved from the complex and minimum-phase cepstrum as

$$\hat{x}_a(n) = \hat{x}(n) - \hat{x}_m(n), \qquad n = -C, \ldots, C. \tag{13}$$

By substituting (12) into (13) it can be noticed that the all-pass cepstrum $\hat{x}_a(n)$ is non-causal and anti-symmetric, and only depends on the non-causal part of $\hat{x}(n)$

$$\hat{x}_a(n) = \begin{cases} \hat{x}(n), & n = -C, \ldots, -1, \\ 0, & n = 0, \\ -\hat{x}(-n), & n = 1, \ldots, C, \end{cases} \tag{14}$$

Therefore, $\{\hat{x}(-C), \ldots, \hat{x}(-1)\}$ carries the extra phase information which is not usually taken into account in conventional source-filter models based on minimum-phase filter impulse responses.

Finally, for use in acoustic modeling *phase parameters* are derived, defined as the non-causal part of $\hat{x}(n)$,

$$\phi(n) = -\hat{x}(-n-1) = \hat{x}_a(n+1), \quad n = 0, \ldots, C_a, \tag{15}$$

where $C_a < C$ is the order of the phase parameters.

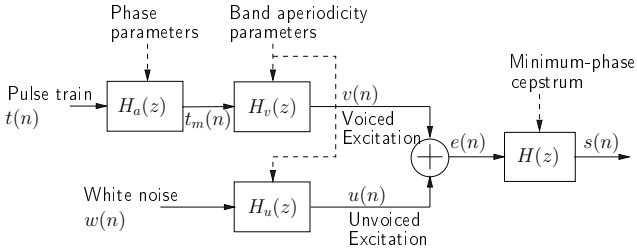

Fig. 2. Synthesis time. The new phase parameters are used to include phase information in the pulse train through the all-pass filter $H_a(z)$.

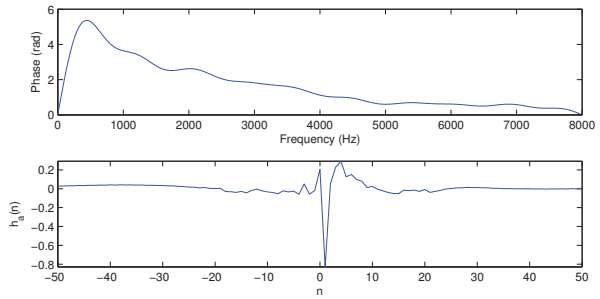

Fig. 3. Phase (top) and impulse (bottom) responses of the all-pass filter $H_a(z)$, obtained from the generated phase parameters $\hat{\phi}(n)$ according to (16) and (17), respectively, for a given speech frame.

3.3. Synthesis under band-aperiodicity-based mixed excitation

To incorporate the complex cepstrum into the waveform synthesis stage, the time domain process shown in Fig. 2 is used. First, F_0, phase, and band-aperiodicity parameters are generated as described in [10]. These are used to derive the pulse train $t(n)$, all-pass filter $H_a(z)$, and voiced and unvoiced filters $H_v(z)$ and $H_u(z)$ respectively. After that, $t(n)$ is passed through $H_a(n)$ to result in the filtered pulse train $t_m(n)$. The mixed excitation signal $e(n)$ is formed by passing $t_m(n)$ and $w(n)$ through $H_v(z)$ and $H_u(z)$, respectively. Finally, the speech signal $s(n)$ is synthesized by passing $e(n)$ through $H(z)$. This filter can be implemented directly by using the minimum-phase cepstral coefficients $\hat{x}_m(n)$ with the Mel log approximation filter [12].

The impulse response of the non-causal all-pass filter, $h_a(n)$, is obtained from the generated phase parameters $\tilde{\phi}(n)$. By using the relationship of (15), and taking the inverse complex cepstrum operation from equations (4), (5), and (6), the phase and impulse responses of the all-pass filter $H_a(z)$ become, respectively

$$\theta_a(\omega_l) = -2\sum_{n=0}^{C_a-1}\tilde{\phi}(n)\sin[\omega_l(n+1)], \quad (16)$$

$$h_a(n) = \frac{1}{2L+1}\left\{1+2\sum_{l=1}^{L}\cos[\omega_l n+\theta_a(\omega_l)]\right\}, \quad (17)$$

for $n=-P_a,...,P_a$, where P_a is the order of the all-pass filter impulse response, and $\{\omega_0,\ldots,\omega_L\}$ are the $L+1$ frequencies in which $S(e^{\jmath\omega})$ is sampled, which $\omega_0=0$ and $\omega_L=\pi$. Fig. 3 shows examples of $\theta_a(\omega)$ and $h_a(n)$ obtained from generated phase parameters. Fig. 4 shows the effect of the resulting filter, where single pulses, represented by $t(n)$, are transformed into the glottal pulses $t_m(n)$.

The band-aperiodicity parameters are interpolated to result in $L+1$ aperiodicity coefficients $\{\alpha_0,\ldots,\alpha_L\}$. The aperiodicity parameters are used to derive the voiced and unvoiced filter impulse

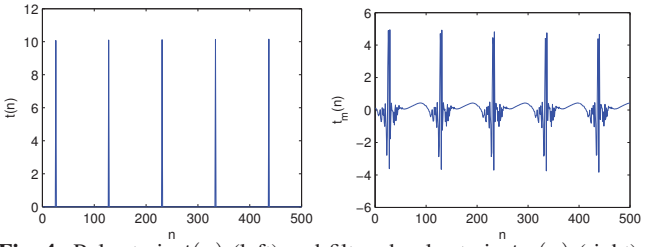

Fig. 4. Pulse train $t(n)$ (left) and filtered pulse train $t_m(n)$ (right) for a given segment.

responses, $h_v(n)$ and $h_u(n)$, respectively. By considering that these filters have zero-phase response, i.e., $H_v(e^{\jmath\omega}) = H_v(e^{-\jmath\omega}) = 1-\alpha_l$, and $H_u(e^{\jmath\omega}) = H_u(e^{-\jmath\omega}) = \alpha_l$, and taking the inverse DTFT, the impulse responses can be obtained from $\{\alpha_0,\ldots,\alpha_L\}$ as

$$h_v(n) = \frac{1}{2L+1}\left[1-\alpha_0+2\sum_{l=1}^{L}(1-\alpha_l)\cos(\omega_l n)\right], \quad (18)$$

$$h_u(n) = \frac{1}{2L+1}\left[\alpha_0+2\sum_{l=1}^{L}\alpha_l\cos(\omega_l n)\right]. \quad (19)$$

4. EXPERIMENTS

A database comprising 4898 sentences, spoken by an American English female speaker and sampled at 16 kHz, was utilized to train two statistical parametric synthesizers: one without phase information, *the baseline system*; and another with phase information, the *proposed system*. The complex cepstrum was extracted through interpolation of pitch-synchronous amplitude and phase responses, as shown in Section 3.1. It was then decomposed into its minimum and all-pass components as described in Section 3.2. The minimum-phase cepstra was warped to yield 40 Mel-cepstral coefficients ($C=39$) [12]. The all-pass cepstra was converted into 19 phase parameters ($C_a=19$). The aperiodicity coefficients $\{\alpha_0,\ldots,\alpha_L\}$ were calculated as the ratio between the voiced and unvoiced speech component amplitude spectra [13], and converted into band-aperiodicity according to the procedure described in [10]. For the proposed system, each observation vectors was composed of:

- stream 1: 40 Mel-cepstral coefficients, delta and delta-delta;
- streams 2, 3, 4: $\ln F_0$, $\Delta \ln F_0$ and $\Delta\Delta \ln F_0$, respectively;
- stream 5: 5 band-aperiodicity, delta and delta-delta;
- stream 6: 19 phase parameters, delta and delta-delta.

The baseline system did not have the stream of phase parameters. For both systems, the observation vectors were used to train HSMMs with 5 states and a left-to-right no-skip topology. All the streams were independently clustered using the decision tree method. At synthesis time, parameter generation with global variance [4] was used for minimum-phase Mel-cepstrum and $\ln F_0$.

To investigate the impact of the phase features on the subjective quality of the synthetic speech, a preference test was conducted with 33 listeners and 25 sentences using a web-based recruitment system. For the test, each of the 25 sentences were synthesized using the two systems. The baseline system emulated the excitation method described in [10] through the use of the same engine as the proposed method, Fig 2, but constraining $H_a(z)=1$, i.e., by removing the all-pass filter. For the proposed system the all-pass filter was implemented with the phase parameters. Table 2 shows the results for this preference test. There is a statistically significant preference for the

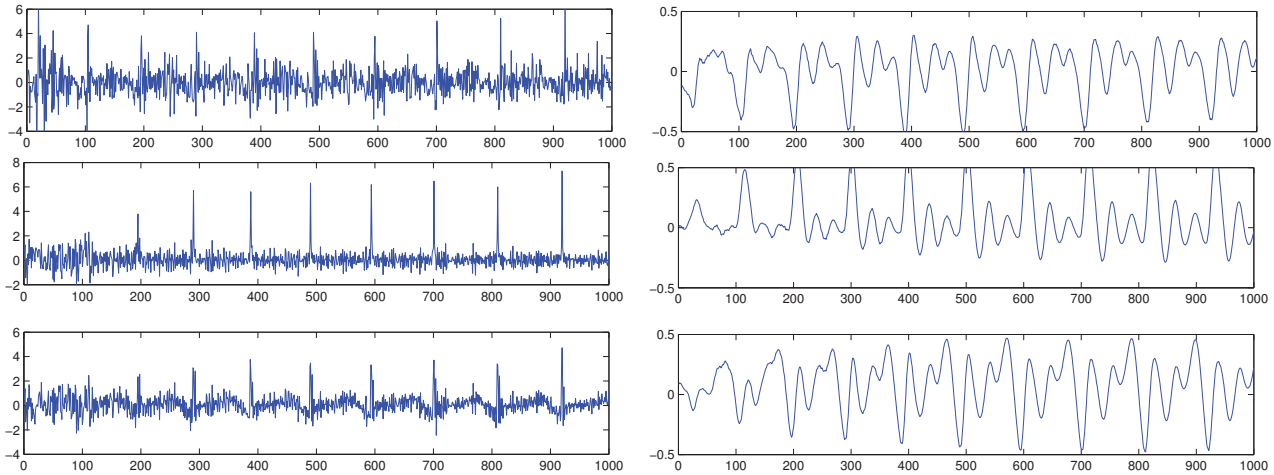

Fig. 5. From top to bottom: residual signal (left) and natural speech (right), mixed excitation without phase information (left) and corresponding synthetic speech (right), and mixed excitation with phase feature (left) and corresponding synthetic speech.

Table 2. Results of the preference test (in percentage) with 36 subjects and 25 sentences.

Baseline	Proposed	No preference	p-value
35.5	43.3	21.2	0.043

Table 3. SNRseg and LSDs (dB) for the sentences used in the listening test for each system.

	Baseline	Proposed		Baseline	Proposed
SNRseg	-4	-0.77	LSD	5.78	5.65

proposed system ($p < 0.05$), which implies that the use of the phase parameters can produce synthetic speech with improved quality.

Objective measures such as SNRseg and LSD were also calculated for the test sentences. To avoid duration and pitch onset mismatches, the original phonetic durations were employed for parameter generation, and pitch marks extracted from natural speech used for waveform generation. The results are shown in Table 3. It can be seen that the proposed system produces waveforms that are closer to the natural speech, especially when measured by SNRseg. Fig. 5 shows some examples of excitation signals and the resulting synthetic waveforms generated with the natural spectrum, duration and pitch marks. The excitation signal with phase information results in a speech waveform that is closer to the natural one.

5. CONCLUSIONS

An approach to model phase information in HMM-based speech synthesizers through the use of complex cepstrum has been presented. At the parameter extraction stage, the interpolation of pitch-synchronous magnitude and phase spectra has been shown to be effective for obtaining frame-based complex cepstrum. For acoustic modeling, the minimum-phase/all-pass decomposition of complex cepstra was used. The all-pass component was derived from the non-causal part of the complex cepstrum and converted into phase parameters. At synthesis time these phase parameters were used to implement a glottal filter. Experimental results under the framework of band-aperiodicity-based mixed excitation show that the inclusion of phase information increases synthetic speech quality. However, there is additional computational complexity for the proposed system due to the phase unwrapping procedure at analysis time and glottal filter implementation at synthesis time. Future work will investigate the use of frame-based complex cepstrum analysis for improving expressive speech synthesis.

6. REFERENCES

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

[2] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of the glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 001–013, 1985.

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

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

[5] J. P. Cabral, J. Yamagishi, and K. Richmond, "HMM-based speech synthesiser using the LF-model of the glottal model," in *ICASSP*, 2011, pp. 4704–4709.

[6] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system using glottal inverse filtering," in *Interspeech*, 2008, pp. 1881–1884.

[7] M. Tamura, T. Kagoshima, and M. Akamine, "Sub-band basis spectrum model for pitch-synchronous log-spectrum and phase based on approximation of sparse coding," in *Interspeech*, 2010, pp. 2406–2409.

[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] T. Drugman and T. Dutoit, "Chirp complex cepstrum-based decomposition for asynchonous glottal analysis," in *Interspeech*, 2010, pp. 657–660.

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

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

[12] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *ICASSP*, 1992, pp. 137–140.

[13] P. J. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, Oct. 2001.