COMPLEX CEPSTRUM FACTORIZATION FOR STATISTICAL PARAMETRIC SYNTHESIS

Ranniery Maia, Yannis Stylianou

Toshiba Research Europe Limited
Cambridge Research Laboratory, Cambridge, UK

ABSTRACT

This paper presents a study on complex cepstrum-based speech factorization for acoustic modeling in statistical parametric synthesizers. The factorization is conducted assuming that both vocal tract resonance and glottal flow effect are fully represented by the complex cepstrum. We investigated four different forms to represent the complex cepstrum in the acoustic models and compared their performances in terms of objective measures between reconstructed and natural waveforms and final quality of the synthesized speech. According to experimental results, the all-pass/minimum-phase and real cepstrum/phase cepstrum decompositions are the best ones in terms of preserving the complex cepstrum information after the parameter generation process.

Index Terms— Speech synthesis, statistical parametric speech synthesis, speech production models, complex cepstrum

1. INTRODUCTION

Homomorphic deconvolution is a method of processing signals that are assumed to be convolved [1]. In case of speech, cepstral analysis has been largely used to separate the effects of the vocal tract from the glottal flow [2, 3]. Different from the usual real cepstrum of speech, the complex cepstrum is a full representation of the speech signal because it contains not only the amplitude but also its phase spectrum [1, 2, 4]. However, there are two main drawbacks to the use of the complex cepstrum: (1) the speech signal usually must be segmented at the glottal closure instants (GCI); (2) because the phase response usually corresponds the principal value of the phase, a phase unwrapping mechanism must be conducted. In the last decades many authors have proposed methods to overcome or alleviate these issues, e.g. [5, 6, 7], and [8, 9].

We have applied the complex cepstrum to statistical parametric speech synthesis and obtained good results in terms of synthesized speech quality [10, 11]. However, despite the good performance achieved, analysis issues related to the speech segmentation and phase unwrapping still existed. To alleviate these problems, and also as a way to derive frame-based complex cepstrum as required by statistical models, we proposed an iterative method of complex cepstrum analysis based on the minimum mean squared error (MSE) [12]. The proposed approach showed to be robust to inaccurate glottal closure instant (GCI) indications. Furthermore, no phase unwrapping mechanism was necessary to take place during the optimization procedure. This technique resulted in good synthesized speech quality, specially for data with high F_0 fluctuations, such as expressive speech. Later, more improvements in terms of speech quality and robustness to expressive data were achieved by doing complex cepstrum optimization on a warped scale, and by adjusting the pulse optimization procedure so as to match the generation process of statistical parametric synthesizers [13].

In this paper we investigate on possible ways to model the complex cepstrum information in the hidden semi-Markov models so as to result in better speech representation at the synthesis stage. We test four different ways to model the complex cepstrum, with three of them based on factorizations. The decompositions were considered according to speech signal components, such as amplitude and phase, and also based on factors that are connected to the way in which the speech signal is produced, such as vocal tract and glottal flow parameters.

The organization of this paper is as follows. Section 2 outlines speech modeling using the complex cepstrum; Section 3 presents the considered factorizations and respective connections with the speech signal, and speech production mechanism; Section 4 shows some experiments; and Section 5 presents the conclusions.

2. COMPLEX CEPSTRUM-BASED SPEECH REPRESENTATION

Assuming that $s(n)$ is a two-pitch segment of speech, selected through an appropriate window with center at the GCI, the warped complex cepstrum of $s(n)$ can be given by [13]

$$\hat{h}(n) = \int_{-\pi}^{\pi} \left[\ln \left| S\left(e^{\jmath\beta_\alpha^{-1}(\omega)}\right) \right| + \jmath\theta\left(\beta_\alpha^{-1}(\omega)\right) \right] e^{\jmath\omega n} \frac{d\omega}{2\pi}, \quad (1)$$

for $-C \leq n \leq C$, where C is the cepstral order, $\left| S\left(e^{\jmath\beta_\alpha^{-1}(\omega)}\right) \right|$ and $\theta\left(\beta_\alpha^{-1}(\omega)\right)$ are the frequency warped amplitude and phase spectrum of $s(n)$, respectively, and $\beta_\alpha(\omega)$ is a bilinear function,

$$\beta_\alpha(\omega) = \frac{\left(1-\alpha^2\right)\sin\omega}{\left(1+\alpha^2\right)\cos\omega - 2\alpha}, \quad (2)$$

to warp the angular frequency ω, with α being a factor that indicates the degree of warping [14]. The cepstral coefficients, $\hat{h}(n)$, usually encapsulate the effects of the vocal tract resonance and glottal flow.

At synthesis time, $\hat{h}(n)$ can be converted to an impulse response, $h(n)$, through the following operations

$$H\left(e^{\jmath\omega}\right) = \exp \sum_{n=-C}^{C} \hat{h}(n) e^{-\jmath\beta_\alpha(\omega)n}, \quad (3)$$

$$h(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H\left(e^{\jmath\omega}\right) e^{\jmath\omega n} d\omega, \quad (4)$$

for $-M/2 \leq n \leq M/2$, where M is the impulse response order. To reconstruct the speech signal, synthesis can be done by performing overlap-and-add of the impulse responses, $h(n)$, at the corresponding analysis instants. Fig. 1 shows examples of natural and reconstructed waveforms. Note that the similarity between the two signals indicates that the phase information of the speech signal is being represented by $\hat{h}(n)$.

For simplicity, $\beta_\alpha(\omega)$ and $\beta_\alpha^{-1}(\omega)$ will be dropped henceforth.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

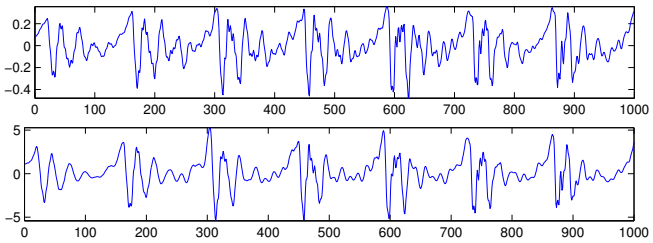

Fig. 1. Example of natural (top) and reconstructed (bottom) voiced speech segments. The reconstructed segment was obtained through overlap-and-add of the impulse responses derived from the complex cepstrum, $h(n)$, at the respective analysis instants.

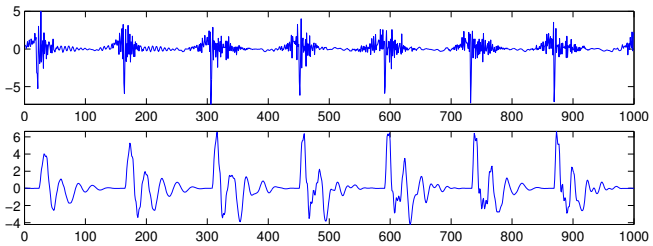

Fig. 2. All-pass (top) and minimum-phase (bottom) components of the natural segment shown in Fig. 1. The waveforms were obtained through the overlap-and-add of $h_a(n)$ and $h_m(n)$ at the corresponding analysis instants.

3. COMPLEX CEPSTRUM-BASED SPEECH FACTORIZATION

For statistical parametric synthesis it is interesting to factorize the complex cepstrum for acoustic modeling due to two main reasons. The first one is that depending on the size of the database, high-dimensional features like $\hat{h}(n)$ can be poorly estimated through high-dimensional Gaussian distributions [15]. The second reason is that as mentioned in Section 2, the complex cepstrum is assumed to be a full representation of the speech signal and consequently includes both effects of the vocal tract and glottal excitation. Because the glottal flow excitation and the vocal tract resonance play different roles in the speech production mechanism, it is natural to consider that they should have separate decision trees. In the following we present intuitive ways in which the complex cepstrum can be decomposed for statistical modeling.

3.1. All-pass/minimum-phase decomposition

Any stable system with impulse response $h(n)$ can be represented as a convolution of a minimum-phase, $h_m(n)$, and an all-pass, $h_a(n)$, impulse responses [1], i.e. $h(n) = h_m(n) * h_a(n)$. In the cepstral domain this convolution becomes a summation: $\hat{h}(n) = \hat{h}_m(n) + \hat{h}_a(n)$. Assuming that the complex cepstrum of an equivalent sequence with the same amplitude spectrum and minimum phase response is given by [1]

$$\hat{h}_m(n) = \begin{cases} 0, & n < 0, \\ \hat{h}(0), & n = 0, \\ \hat{h}(n) + \hat{h}(-n), & 1 \leq n \leq C, \end{cases} \tag{5}$$

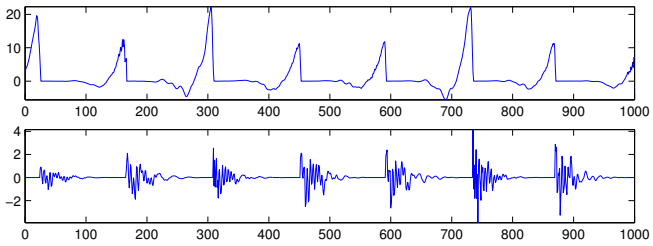

Fig. 3. Anti-causal (top) and causal (bottom) components of the natural speech segment shown in Fig. 1. The waveforms were obtained through the overlap-and-add of $h_{ac}(n)$ and $h_{ca}(n)$ at the corresponding analysis instants.

then the so-defined *all pass cepstrum* becomes

$$\hat{h}_a(n) = \hat{h}(n) - \hat{h}_m(n) = \begin{cases} \hat{h}(n), & n < 0, \\ 0, & n = 0, \\ -\hat{h}(-n), & n > 0, \end{cases} \tag{6}$$

where it can be noticed that $\hat{h}_a(n)$ contains solely samples of the anti-causal cepstrum, i.e. $\hat{h}(n), n < 0$. Therefore, the all-pass impulse response $h_a(n)$ represents the additional phase information that when added to the minimum-phase, results in the phase of the speech signal. This decomposition has advantages in terms of compatibility with systems based on the minimum-phase synthesis filter.

Fig. 2 shows the all-pass and minimum-phase components of the segment of speech shown in the top part of Fig. 1. The waveforms were produced through the overlap-and-add of the impulse responses $h_a(n)$ and $h_m(n)$ at the respective analysis instants.

3.2. Anti-causal/causal decomposition

Another way to factorize the complex cepstrum is to simply split it into its anti-causal and causal components, i.e.

$$\hat{h}_{ac}(n) = \begin{cases} \hat{h}(n), & -C \leq n < 0, \\ 0, & 0 \leq n \leq C, \end{cases} \tag{7}$$

$$\hat{h}_{ca}(n) = \begin{cases} 0, & -C \leq n < 0, \\ \hat{h}(n), & 0 \leq n \leq C. \end{cases} \tag{8}$$

In this case, studies have shown (e.g. [8]) that the impulse response derived from the anti-causal cepstrum, $\hat{h}_{ac}(n)$, is related to the glottal flow. Therefore, assuming that the complex cepstrum fully represents the speech signal, the causal cepstrum $\hat{h}_{ca}(n)$ is then related mostly to the vocal tract information. Fig. 3 shows the anti-causal and causal components of the speech segment shown in Fig. 1. The waveforms were created through the overlap-and-add of $h_{ac}(n)$ and $h_{ca}(n)$ at the analysis instants. It can be noticed that $h_{ac}(n)$ produces a signal that resembles a sequence of glottal pulses.

This form of decomposition has advantages in terms of explicitly separating the effects of the glottal flow from the vocal tract. However, as investigated in [8], the separation of the anti-causal and causal speech components using the complex cepstrum is not a trivial task. Although the waveforms shown in Fig. 3 indicate that the separation has more or less succeeded, in many cases the components produced by the inaccurate anti-causal and causal cepstra do not make much sense in terms of speech production mechanism. This decomposition is highly sensitive to the estimation of the complex cepstrum.

Table 1. Amplitude and phase responses of the components, in terms of $\hat{h}(n)$, assuming the decompositions shown in Section 3.

Minimum phase		All pass	
$\lvert H_m\left(e^{\jmath\omega}\right)\rvert = \exp\sum_{n=0}^{C}\hat{h}(n)\cos\omega n$ $\cdot\exp\sum_{n=1}^{C}\hat{h}(-n)\cos\omega n$	$\theta_m\left(\omega\right) = -\sum_{n=1}^{C}\hat{h}(n)\sin\omega n$ $-\sum_{n=1}^{C}\hat{h}(-n)\sin\omega n$	$\lvert H_a\left(e^{\jmath\omega}\right)\rvert = 1, \forall\omega$	$\theta_a\left(\omega\right) = 2\sum_{n=1}^{C}\hat{h}(-n)\sin\omega n$
Causal		**Anti-causal**	
$\lvert H_{ca}\left(e^{\jmath\omega}\right)\rvert = \exp\sum_{n=0}^{C}\hat{h}(n)\cos\omega n$	$\theta_{ca}\left(\omega\right) = -\sum_{n=1}^{C}\hat{h}(n)\sin\omega n$	$\lvert H_{ac}\left(e^{\jmath\omega}\right)\rvert =$ $\exp\sum_{n=1}^{C}\hat{h}(-n)\cos\omega n$	$\theta_{ac}\left(\omega\right) = \sum_{n=1}^{C}\hat{h}(-n)\sin\omega n$
Real cepstrum		**Phase cepstrum**	
$\lvert H_r\left(e^{\jmath\omega}\right)\rvert = \exp\sum_{n=0}^{C}\hat{h}(n)\cos\omega n$ $\cdot\exp\sum_{n=1}^{C}\hat{h}(-n)\cos\omega n$	$\theta_r\left(\omega\right) = 0, \forall\omega$	$\lvert H_p\left(e^{\jmath\omega}\right)\rvert = 1, \forall\omega$	$\theta_p\left(\omega\right) = -\sum_{n=1}^{C}\hat{h}(n)\sin\omega n$ $+\sum_{n=1}^{C}\hat{h}(-n)\sin\omega n$

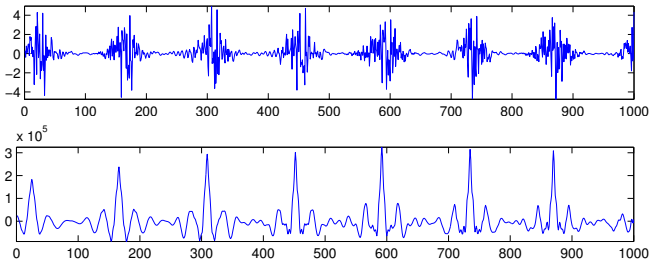

Fig. 4. Only-phase (top) and zero-phase components of the natural speech segment shown in Fig. 1. The waveforms were constructed through the overlap-and-add of $h_r(n)$ and $h_r(n)$ at the corresponding analysis instants.

3.3. Phase cepstrum/real cepstrum decomposition

The complex cepstrum can also be decomposed into the real cepstrum, which carries solely information on the amplitude spectrum of speech, and the so-defined *phase cepstrum*, which carries information solely on the phase spectrum of speech. The real cepstrum corresponds to the even component of the complex cepstrum [1]

$$\hat{h}_r(n) = \frac{\hat{h}(n) + \hat{h}(-n)}{2}, \quad -C \leq n \leq C, \tag{9}$$

and consequently the phase cepstrum becomes

$$\hat{h}_p(n) = \hat{h}(n) - \hat{h}_r(n) = \frac{\hat{h}(n) - \hat{h}(-n)}{2}, \tag{10}$$

for $-C \leq n \leq C$. It can be demonstrated that the real and phase cepstrum can also be derived by taking the inverse Fourier transforms of the log magnitude and phase spectra, respectively. Fig. 4 shows two waveforms. The one at the top contains the phase information of the speech segment of Fig. 1, while the waveform at the bottom is its zero-phase version. The waveforms were produced through the overlap-and-add of $h_r(n)$ and $h_p(n)$ at the analysis instants. Note the similarity between Figures 2 and 4. The only difference between the two forms of decomposition regards the minimum-phase response.

3.4. Summary of the decompositions

Table 1 shows the amplitude and phase responses of each component of the decompositions considered in Section 3. Assuming that the anti-causal cepstrum represents the glottal flow information, and the causal cepstrum the vocal tract[1], the following observations can be made: (1) the real cepstrum merges the amplitude responses of the vocal tract and glottal flow; (2) the minimum-phase cepstrum corresponds the vocal tract impulse response convolved with a time flipped, causal version of the glottal flow impulse response. Table 2 summarizes the decompositions, showing which sort of information each component embeds in terms of speech signal properties and speech production mechanism.

[1]Note that in this work we are disregarding that the lip radiation effect.

Table 2. Summary of the decompositions considered in terms of what each component represents.

Factorization	Comp.	Information embedded
All pass (AP) / Minimum phase (MP)	MP	Vocal tract amplitude response Glottal flow amplitude response Vocal tract phase response Glottal flow phase to some extent
	AP	Glottal flow phase response
Anti-causal (AC) / Causal (CA)	CA	Vocal tract amplitude response Vocal tract phase response
	AC	Glottal flow amplitude response Glottal flow phase response
Phase cepstrum (PC) / Real cepstrum (RC)	RC	Vocal tract amplitude response Glottal flow amplitude response
	PC	Vocal tract phase response Glottal flow phase response

4. EXPERIMENT

Two databases were used to check which factorizations presented in Section 3 result in better performance in terms of speech synthesis quality under the statistical parametric method. The first database consisted of 3180 utterances from a female English speaker sampled at 16 kHz. The second database corresponded to 2551 utterances from a male English speaker, also sampled at 16 kHz.

The databases were processed by the MSE-based complex cepstrum analysis method presented in [13], with cepstral order $C = 39$, warping factor $\alpha = 0.42$, impulse response order $M = 1024$, and the number of sampled frequency for the warped spectrum, was set to $L + 1 = 513$. A total of five iterations per utterance was conducted, where each iteration included a GCI marking optimization and a complex cepstrum re-estimation step. F_0 and initial GCI markings were extracted using the Entropic Signal Processing Tools.

The complex cepstra were decomposed in three different ways, according to the factorizations described in Section 3. Four different systems were trained. The systems differed in the way the complex cepstrum was modeled in the statistical parametric synthesis framework. Table 3 shows how stream organization and weight, γ, were adjusted. The complex cepstrum streams were modeled using continuous Gaussian probabilities while the $\ln F_0$ streams were modeled using multi-space Gaussian distributions [16]. Note that stream weight $\gamma = 1$ was set to all the streams in System 3, since in our un-

Table 3. Trained systems. The feature vectors are $\hat{\boldsymbol{h}}_m^\top = [\hat{h}_m(0) \cdots \hat{h}_m(C)]$, $\hat{\boldsymbol{h}}_a^\top = [\hat{h}_a(1) \cdots \hat{h}_a(C)]$, $\hat{\boldsymbol{h}}_{ca}^\top = [\hat{h}_{ca}(0) \cdots \hat{h}_{ca}(C)]$, $\hat{\boldsymbol{h}}_{ac}^\top = [\hat{h}_{ac}(-1) \cdots \hat{h}_{ac}(-C)]$, $\hat{\boldsymbol{h}}_r^\top = [\hat{h}_r(0) \cdots \hat{h}_r(C)]$, $\hat{\boldsymbol{h}}_p^\top = [\hat{h}_p(1) \cdots \hat{h}_p(C)]$, and $\hat{\boldsymbol{h}}^\top = [\hat{h}(-C) \cdots \hat{h}(C)]$, and γ is the stream weight. Other streams considered were $\{\boldsymbol{o}_2, \boldsymbol{o}_3, \boldsymbol{o}_4\} = \{\ln F_0, \Delta \ln F_0, \Delta^2 \ln F_0\}$.

System	Factorization	Stream organization	γ
1	All-pass/ min. phase	$\boldsymbol{o}_1^\top = [\hat{\boldsymbol{h}}_m^\top \ \Delta\hat{\boldsymbol{h}}_m^\top \ \Delta^2\hat{\boldsymbol{h}}_m^\top]$ $\boldsymbol{o}_5 = [\boldsymbol{h}_a^\top \ \Delta\boldsymbol{h}_a^\top \ \Delta^2\boldsymbol{h}_a^\top]$	1 0
2	Anti-causal/ causal	$\boldsymbol{o}_1^\top = [\hat{\boldsymbol{h}}_{ca}^\top \ \Delta\hat{\boldsymbol{h}}_{ca}^\top \ \Delta^2\hat{\boldsymbol{h}}_{ca}^\top]$ $\boldsymbol{o}_5^\top = [\hat{\boldsymbol{h}}_{ac}^\top \ \Delta\hat{\boldsymbol{h}}_{ac}^\top \ \Delta^2\hat{\boldsymbol{h}}_{ac}^\top]$	1 0
3	Phase cep./ real cep.	$\boldsymbol{o}_1^\top = [\hat{\boldsymbol{h}}_r^\top \ \Delta\hat{\boldsymbol{h}}_r^\top \ \Delta^2\hat{\boldsymbol{h}}_r^\top]$ $\boldsymbol{o}_5^\top = [\hat{\boldsymbol{h}}_p^\top \ \Delta\hat{\boldsymbol{h}}_p^\top \ \Delta^2\hat{\boldsymbol{h}}_p^\top]$	1 1
4	None	$\boldsymbol{o}_1^\top = [\hat{\boldsymbol{h}}^\top \ \Delta\hat{\boldsymbol{h}}^\top \ \Delta^2\hat{\boldsymbol{h}}^\top]$	1

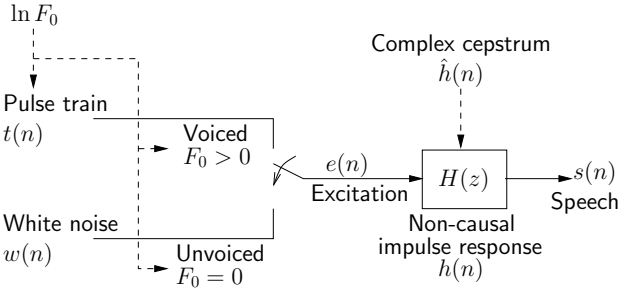

Fig. 5. Synthesis time. $\ln F_0$ is used to generate a simple pulse train/white noise excitation signal while the non-causal impulse response of the synthesis filter is derived from the complex cepstrum.

derstanding the *full* phase information should also have an influence on the state alignment. Decision tree-clustering for the systems listed in Table 3 was performed in three steps. In the first two steps, the minimum description length (MDL) criterion was used [17]. Each step consisted of untying the models and decision tree-based context clustering, followed by 5 iterations of embedded re-estimation. In the last step, cross validation with hierarchical priors as presented in [18] was used to derive the final decision trees. This procedure was conducted to avoid over smoothing problems due to the empirical adjustment of the MDL factor for tree growth.

At synthesis time, in systems 1,2 and 3, the generated components of the complex cepstrum were combined to result in frame-based complex cepstrum sequences. In System 4, frame-based complex cepstra was directly obtained from the parameter generation stage. Once sequences of complex cepstrum and $\ln F_0$ were available, speech synthesis was performed as shown in Fig. 5. The generated the complex cepstrum was used to derive the synthesis filter non-causal impulse response according to (3) and (4).

4.1. Objective evaluation

A set of 50 sentences from each speaker used in the training were used for an objective evaluation. Each of the selected sentences were re-synthesized using natural $\ln F_0$, GCI markings obtained in the last iteration of the MSE-based complex cepstrum analysis, phone durations taken from the training labels, and generated complex cepstrum. The idea was to verify the impact of the acoustic modeling using the different decompositions. The distance between natural and reconstructed speech was measured in terms of segmented signal to noise ratio of the voiced regions (SNRseg-v). Table 4 shows the results of this test for both male and female speakers. The sentences were randomly selected from the training material.

Table 4. Segmented signal-to-noise ratio for the voiced regions (SNRseg-v) in dB, between natural and reconstructed speech.

	Natural cepstrum	Cepstrum generated from system			
		1	2	3	4
Female	17.25	4.12	3.59	4.12	4.20
Male	13.42	0.18	0.20	0.06	0.32

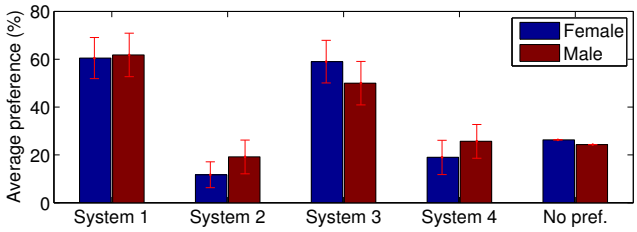

Fig. 6. Results of the subjective test in terms of average listener's preference, with the corresponding 95% confidence intervals.

It can be noticed that the degradation caused by the acoustic modeling is evident, with losses in average of 13 dB. From Table 4 it can also be seen that systems 1, 2 and 4 are similar but superior in quality to System 3. This shows that the instability issues related to the anti-causal/causal decomposition were not smoothed out.

4.2. Subjective test

A subjective listening preference test was conducted to compare the four systems in terms of quality. In total, 20 open sentences were used. Thirteen subjects, including eight speech synthesis specialists, took part in the test. The subjects were instructed to listen carefully to the samples as many time as possible in order to choose the one with the best quality. Each subject listened to 20 comparison pairs that were randomly selected from the 120 possible ones. Table 6 shows the results of the test in terms of average preference. According to these results, the minimum-phase/all-pass and real cepstrum/phase cepstrum decompositions are the ones that mostly preserves complex cepstrum information.

One can note that although the full complex cepstrum modeling performed as one of the best in the objective evaluation with closed sentences, they performed poorly in the listening test. Because the sentences used in the listening test were not part of the training set, this indicates that the high-dimensional complex cepstrum vector is not being properly represented by a single Gaussian distribution. By observing the results of System 2 it can be concluded that the unsuccessful decomposition instances between causal and anti-causal components directly affects that quality of the synthetic speech. Finally, given these conditions, two good choices to model complex cepstrum information for speech synthesis are through the all-pass/minimum-phase and real/phase decompositions.

5. CONCLUSIONS

We investigated different forms to model complex cepstrum information for statistical parametric speech synthesis. In order to do that, different ways to factorize the complex cepstrum in terms of relationship with the speech production mechanism, and in terms of phase and amplitude characteristics, were taken into account. According to the results, among the tested modeling forms, the ones that used all-pass/minimum-phase and the real cepstrum/phase cepstrum decompositions were the ones that mostly preserved the complex cepstrum information, and consequently speech quality.

6. REFERENCES

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

[2] T. F. Quatieri, *Speech signal processing*, Prentice Hall Signal Processing Series, 2002.

[3] 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.

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

[5] J. B. Bednar and T. L. Watt, "Calculating the complex cepstrum without phase unwrapping or integration," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 4, pp. 1014–1017, Aug. 1985.

[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] 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.

[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 *Proc. of Interspeech*, 2010, pp. 657–660.

[10] 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.

[11] R. Maia, M. Akamine, and M.J.F. Gales, "Complex cepstrum for statistical parametric speech synthesis," *Speech Communication*, vol. 5, no. 55, pp. 606–618, June 2013.

[12] 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.

[13] R. Maia, M.J.F. Gales, Y. Stylianou, and M. Akamine, "Minimum mean squared error based warped complex cepstrum analysis for statistical parametric speech synthesis," in *Proc. of Interspeech*, 2013, pp. 2336–2340.

[14] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," *Proceedings of IEEE*, vol. 60, no. 6, pp. 681–691, June 1972.

[15] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[16] K. Tokuda, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information & Systems*, vol. E85-D, no. 3, pp. 455–464, Mar. 2002.

[17] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. of Eurospeech*, 1997, pp. 99–102.

[18] H. Zen and M. J. F. Gales, "Decision tree-based context clustering based on cross validation and hierarchical priors," in *Proc. of ICASSP*, 2011, pp. 4560–4563.