Complex Cepstrum in Speech Synthesis

Vích R., Vondra M.
Institute of Photonics and Electronics, Academy of Sciences of the Czech Republic
vich@ufe.cz

Abstract. In the contribution the use of the complex cepstrum in speech modeling and synthesis is shortly described. This approach leads to a mixed phase speech model in contrary to the conventionally used LPC, or Padé based real cepstrum minimum-phase speech modeling. The obtained mixed phase parametric speech model is of the finite impulse response type and contains also information about the phase properties of the modeled speech frame. The synthesized speech is more natural but the memory requirements and numerical complexity are much higher.

1 Introduction

Conventional parametric speech synthesis is based on the minimum-phase parametric speech production model with infinite impulse response (IIR), see Fig. 1.

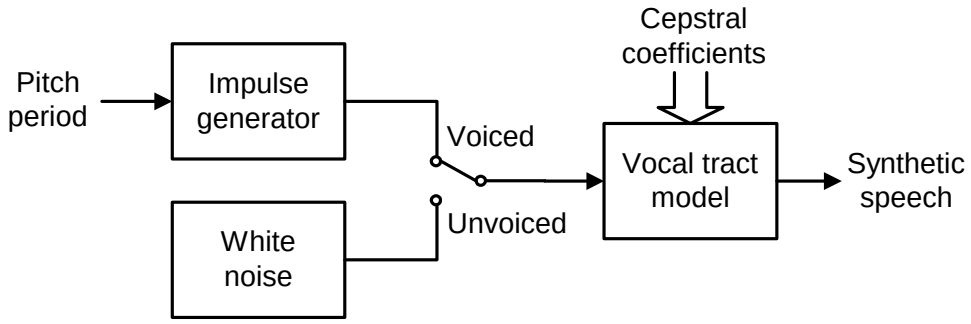

Fig 1. Parametric cepstral speech model.

The vocal tract model is a time varying digital filter based e.g. on linear prediction or on the real cepstrum. The minimum-phase speech model approximates only the magnitude spectrum of the speech frame. The pole-zero transfer function of the vocal tract model based on the real cepstrum using the Padé approximation is described in [1]. In this approach only the logarithmic magnitude frequency response of the corresponding speech frame is approximated and the stability of the model depends on the magnitude of the cepstral coefficients and the chosen order of the Padé approximation.

In this contribution the principle of cepstral speech modeling using the complex cepstrum is described [2,3,4,5,6]. The obtained mixed phase parametric speech model contains also the information about the phase properties of the modeled speech frame, is of the finite impulse response (FIR) type and therefore always stable. The mixed phase model approximates the speech signal with higher accuracy than the model based on the real cepstrum, but the numerical complexity and the memory requirements are at least twice greater.

2 Complex Cepstral Speech Analysis and Synthesis

Let $\{s_n\}$ be the windowed speech frame of the length *N*, sampled with the sampling frequency F_S. The corresponding complex cepstrum $\{\hat{s}_n\}$ is a *two sided real* , in general *asymmetric sequence*, which can be estimated using the fast Fourier transform (FFT). In this case we obtain a time aliased version of the complex cepstrum, but using a sufficient high

dimension of the FFT, $M > N$, the aliasing can be reduced.

$$\hat{s}_n = \frac{1}{M}\sum_{k=0}^{M-1}\hat{S}_k e^{\mathrm{j}2\pi kn/M} \quad , \qquad \hat{S}_k = \ln S_k = \ln|S_k| + \mathrm{j}\arg S_k \, , \qquad S_k = \sum_{n=0}^{M-1} s_n e^{-\mathrm{j}2\pi kn/M} \, . \tag{1}$$

The sequence $\{S_k\}$is the complex spectrum of the speech frame. The imaginary part of the logarithmic spectrum $\hat{S}_k$, i.e. $\arg S_k$, is the unwrapped phase sequence [2,3]. The part of the complex cepstrum $\{\hat{s}_n\}$ for $0 \le n$ will be called *causal cepstrum*, the part of $\{\hat{s}_n\}$ for $n < 0$ *anticipative cepstrum*.

The *minimum-phase* cepstral speech model is based on the *real cepstrum* $\{c_n\}$ defined by

$$c_n = \frac{1}{M}\sum_{k=0}^{M-1}\ln|S_k| e^{\mathrm{j}2\pi kn/M} \, . \tag{2}$$

The real cepstrum is also a *two sided*, but *symmetrical real sequence* $c_n = c_{-n}$, $0 < n$. It does not contain any information about the phase properties of the signal and has therefore only one half of the memory requirements of the complex cepstrum. It also holds

$$c_n = \frac{\hat{s}_n + \hat{s}_{-n}}{2} \, . \tag{3}$$

The complex cepstrum corresponding to a *minimum-phase signal* is *causal* and can be constructed by windowing the real cepstrum $\hat{s}_{\min n} = 2c_n$, for n > 0 and $\hat{s}_{\min 0} = c_0 = \hat{s}_0$.

As an example of the cepstral speech analysis and synthesis we use the stationary part of the vowel *a*. The sampling frequency is $F_S = 8\,\mathrm{kHz}$, the fundamental frequency $F_0 = 118\,\mathrm{Hz}$, the pitch synchronously windowed speech frame is equal to two fundamental periods in the case of a voiced signal with the length $N = 2\,fix(F_S / F_0) = 134$ and the dimension of the FFT $M = 512$. In the example the Blackman window centered on the glottal closure instant is used. The signal, the window, the magnitude and phase spectra and the complex cepstrum are shown in Fig 2. The signal delay resulting from the phase unwrapping in the example is $s = -65$.

In Fig. 3 the anticipative and causal cepstrum parts, the corresponding spectra and the anticipative and causal impulse responses are given. The sum of the logarithmic magnitudes of the anticipative and causal spectra is equal to the magnitude spectrum in Fig. 2. The convolution of the anticipative and causal impulse responses leads to the finite mixed phase impulse response $\{h_n\}$of the vocal tract model shown in Fig. 4. This impulse response can be directly obtained from the complex cepstrum using a symmetric centered rectangular cepstral window $\{w_n\}$ of the length $N_0 = fix(F_S / F_0) = N/2$. It is given by

$$h_n = \frac{1}{M}\sum_{k=0}^{M-1} H_k e^{\mathrm{j}2\pi kn/M} \quad , \qquad H_k = \exp(\hat{H}_k) \, , \qquad \hat{H}_k = \sum_{n=0}^{M-1}\hat{h}_n e^{-\mathrm{j}2\pi kn/M} \, , \tag{4}$$

where $\hat{h}_n = w_n \hat{s}_n$ is the windowed complex cepstrum.

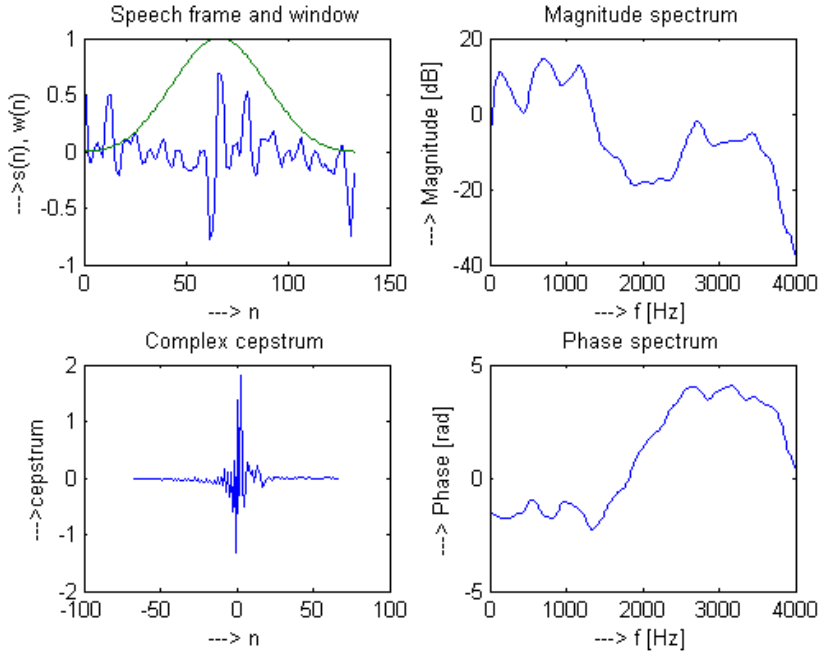

Fig 2. Signal, spectra and complex cepstrum of the stationary part of vowel *a*.

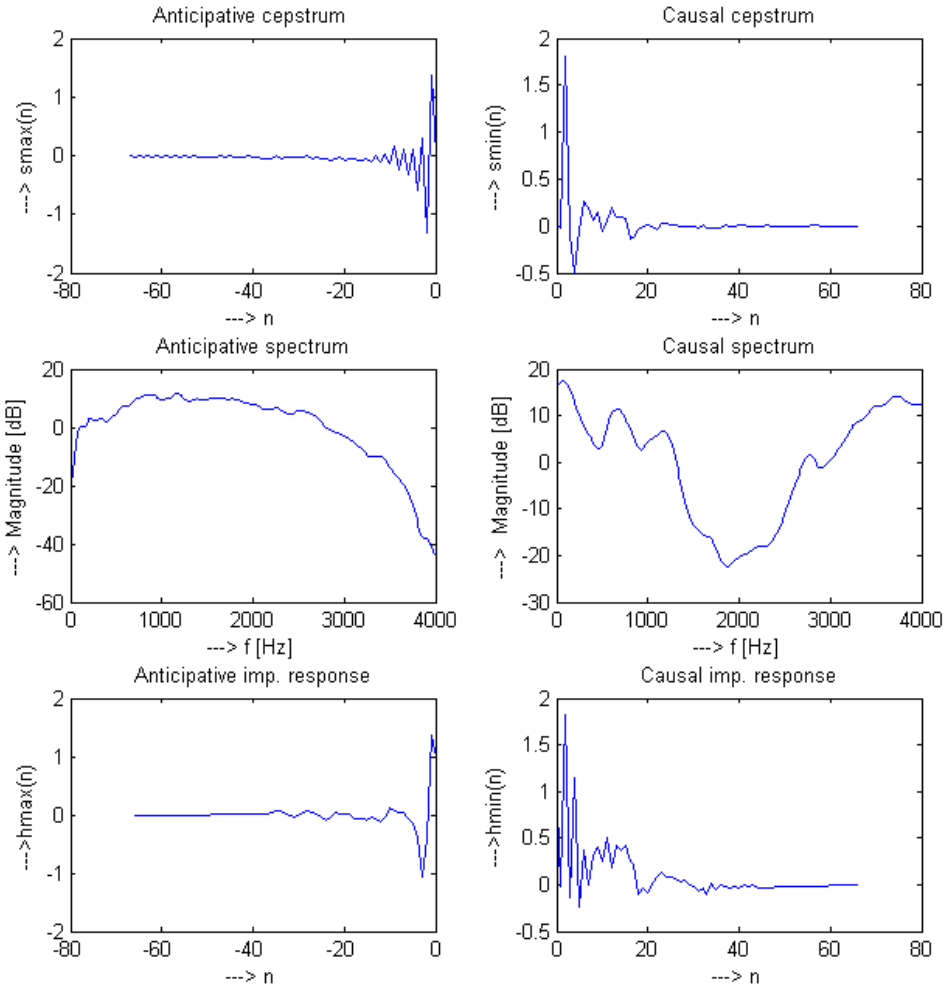

Fig 3. Anticipative and causal cepstra, the corresponding spectra and impulse responses.

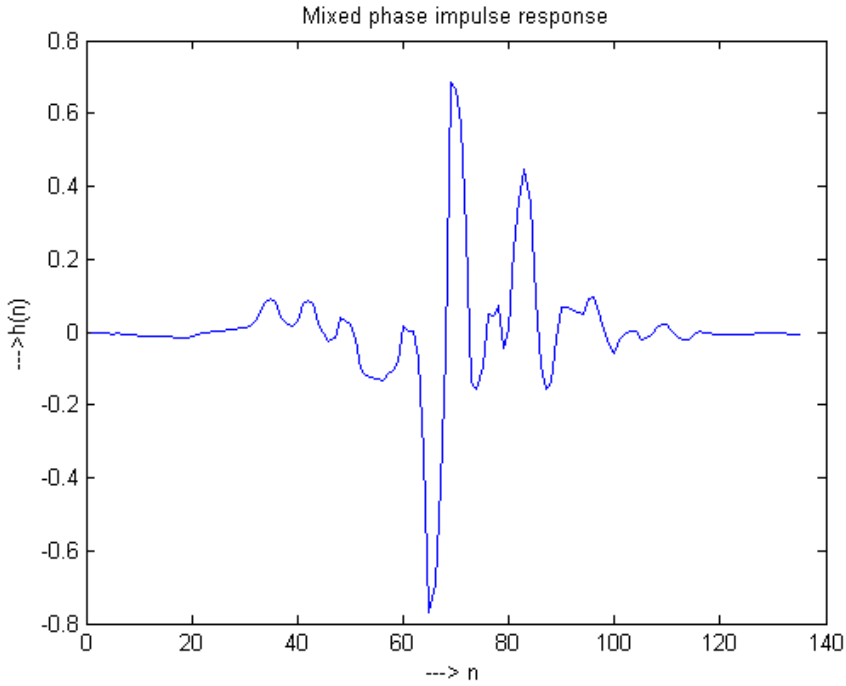

Fig 4. Impulse response of the mixed phase vocal tract model.

Synthetic speech is obtained using the speech production model in Fig.1 by convolving the mixed-phase impulse response $\{h_n\}$ and a periodic impulse train $\{p_n\}$, $p_n = \sum_{i=0}^{[N/67]} \delta_{n-i67}$, where $\{\delta_n\}$ is the unit sample sequence. This results in overlapped-and-added shifted impulse response. In the convolution the signal delay calculated in the phase unwrapping must by respected.

In Fig 5 the comparison of the original speech frame, the minimum-phase and the mixed phase synthesis of the vowel *a* are given. It can be seen that the mixed phase synthesis approximates the original speech signal with higher accuracy than the minimum-phase approach. The difference of this two synthesis approaches is audible.

Essential in this speech modeling using the complex cepstrum is the pitch synchronously windowed speech frame of the length equal to two pitch periods centred on the glottal closure instants with overlapping of one pitch period. The synthesis is also realized pitch synchronous with overlapping. Together with a convenient smooth window starting and ending with zero values the influence of the periodic excitation in the case of voiced sounds is destroyed. This results in a smoothed spectrum corresponding to the frequency response of the vocal tract. Moreover, the anticipative impulse response can be considered as an approximation of the glottal excitation.

In the case of unvoiced sounds the frame length in the analysis should be approximately equal to the mean frame length of the voiced sounds with overlapping of one half of the frame length.

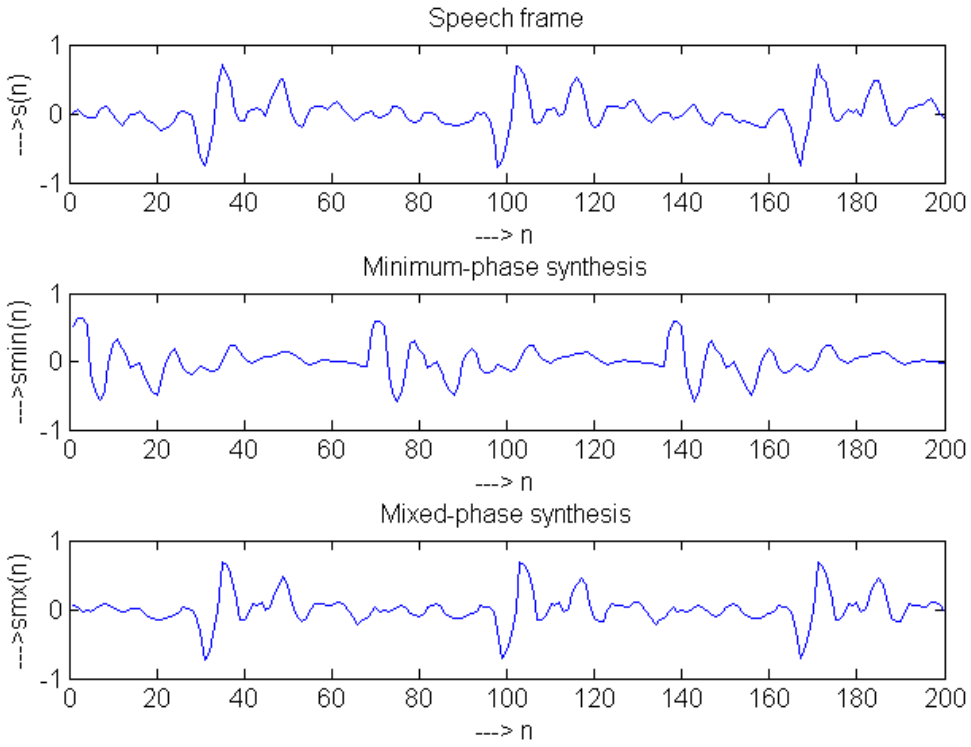

Fig 5.Original speech frame and minimum- and mixed phase synthesis.

3 Conclusions

The construction of the mixed phase FIR vocal tract model based on the complex cepstrum is straight forward and results in a more natural speech synthesis than the minimum-phase approach. It is given by the fact that also the information of the phase properties of the modeled speech frame is respected, or in other words that the true spectral properties of the glottal signal are incorporated into the modeling approach. The memory requirements and the numerical complexity are in consequence of the calculation and application of the complex cepstrum at least twice higher. Audio examples of male and female minimum-and mixed phase speech modeling and Czech triphone text-to-speech synthesis will be presented.

In paper [7] the importance of proper speech windowing for the estimation of the glottal signal is studied. A new type of window is proposed for which the Hann window and the Blackman windows are particular cases. By comparison of speech models obtained with Hann, Blackman, Blackman-Harris and that by Drugman et al. proposed windows using spectral distances and listening tests, no marked differences have been stated.

Acknowledgement

This paper has been supported within the framework of COST 2102 by the Ministry of Education, Youth and Sport of the Czech Republic, project number OC08010.

References

[1] Vích, R. Cepstral Speech Model, Padé Approximation, Excitation and Gain Matching in Cepstral Speech Synthesis. In: J. Jan, (Ed.) BIOSIGNAL 2000, Brno: VUTIUM, 2000:77-82.

[2] Oppenheim, A.V., Schafer, R.W. Discrete-Time Signal Processing. Prentice Hall, 1989:768-825.

[3] Vích, R. *Z*-transform Theory and Application. Dordrecht, D. Reidel Publ. Comp., 1987.
[4] Quatieri, T. F. Discrete-Time Speech Signal Processing. Prentice Hall, 2002:281-292.
[5] Vích, R. Komplexes Cepstrum in der Sprachsynthese. In: A. Lacroix (Ed.): Beiträge zur Signaltheorie, Signalverarbeitung, Sprachakustik und Elektroakustik, Studientexte zur Sprachkommunikation 52, Dresden: TUDpress 2009:216-223.
[6] Vích, R. Nichtkausales cepstrales Sprachmodell. In: R. Hoffmann (Ed): Elektronische Sprachsignalverarbeitung 2009, Studientexte zur Sprachkommunikation: 53, Dresden: TUDpress, 2009:107-114.
[7] Drugman T., Bozkurt, B. T., Dutoit T. Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation, Interspeech 2009, Brighton, U.K, 2009: On-line http://tcts.fpms.ac.be/~drugman/files/IS09-ComplexCepstrum.pdf.