SPEECH CONVERSION USING A MIXED-PHASE CEPSTRAL VOCODER

Martin Vondra and Robert Vich

Institute of Photonics and Electronics,
Academy of Sciences of the Czech Republic
vondra@ufe.cz, vich@ufe.cz

Abstract: In the study a simple conversion of the voice character using a modification of the glottal pulse is shortly described. The glottal signal is estimated by homomorphic speech deconvolution of the speech signal into the maximum- and minimum-phase parts. The maximum-phase part is an approximation of the glottal signal. For speech reconstruction the parametric mixed phase speech generation model based on the complex cepstrum is used, which takes into account not only the magnitude spectrum of the modeled speech frame, but also the phase spectrum.

1 Introduction

The parametric speech generation model based on the complex cepstrum was described in the papers [1, 2]. This modeling approach takes into account not only the magnitude speech spectrum, but also the phase properties of the speech signal. For that reason the speech signal is approximated with higher accuracy.

The conventional parametric speech model (Fig. 1) of the source/filter type contains the excitation model, a periodic pulse generator for voiced speech and a white noise source for unvoiced speech and a digital filter, which models the vocal tract. This filter can be constructed using linear prediction or based on the real cepstrum and the Padé approximation [3, 4, 5]. Both approaches lead to a minimum-phase filter. The initial part of the impulse response of a minimum-phase filter contains the maximum energy of the impulse response which is the cause of the so called buzz effect of the resulting synthetic speech.

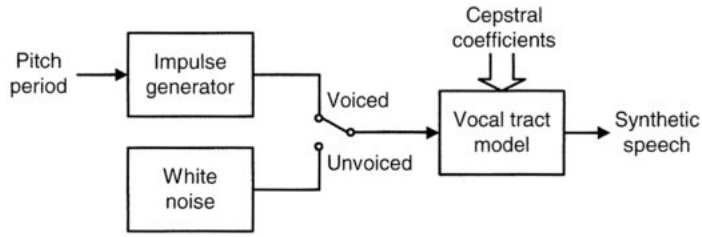

Figure 1 – Parametric cepstral speech model.

A vocal tract model based on the complex cepstrum is a mixed phase system [1, 2]. Such a speech generation model does not suffer from the buzz effect and the speech reconstructed using this system is more natural than in the case of a minimum-phase model.

The complex cepstrum also allows the decompositions of the speech signal into the minimum- and maximum-phase parts [6]. In the paper [7] it is stated that the maximum-phase part of the speech signal is first of all given by the open phase of the glottis. Shortly, the maximum-phase part of the speech signal can be considered as the excitation, i.e. as the glottal signal.

Considering that the speech model based on the complex cepstrum leads to a very natural speech synthesis and enables the decomposition of the speech signal into excitation and the vocal tract impulse response, we decided to experiment with the maximum-phase speech

component, i.e. with the excitation signal, with the aim to change the voice quality, e.g. normal, tense, loose, etc. An inspiration for this was the master thesis [8] about the analysis of the glottal signal.

2 Complex Cepstral Speech Analysis

For complex cepstral speech analysis we use pitch synchronous segmentation, which has to be preceded by pitch pulses localization (estimation of glottal closure instants GCI). One analyzed speech frame consists of two pitch periods, where the pitch pulse is in the middle of the frame. Frame shift is set to one pitch period. Then a weighting window is applied on the frame. The type of the weighting window has an impact on the spectrum and also on the cepstrum of the speech frame. The widely used Hamming window is in this case not suitable, because it does not suppress the periodicity of the frame completely and for this reason the spectrum of the frame is partially harmonic and the cepstrum contains also small periodic lobes. A better choice is the Hann or Blackman window. Some information about appropriate time windows for cepstrum analysis can be found in [6].

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) with the dimension M. 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.

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

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. The linear component in the phase unwrapping is an important parameter for the synthesis part of the vocoder. 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 $0 < n$ *anticipative cepstrum*.

3 Speech Reconstruction from the Complex Cepstrum

Speech is synthesized by a parametric model. Voiced speech is generated by excitation of the vocal tract model by the Dirac impulse. Unvoiced speech is generated by excitation of the vocal tract by a white noise generator. Because the convolution of the Dirac impulse and the impulse response of the vocal tract model is senseless, the convolution for voiced speech can be omitted. After the convolution we must use the overlap and add (OLA) algorithm to fold up the resulting speech, because the segmentation in the analysis is performed also with overlapping.

In the inverse cepstral transformation it is important to use the linear component from the phase unwrapping, which can be interpreted as a signal delay. If we omit this term, the successive impulse responses can have different time shifts, which results in a rough quality of the synthetic speech.

The vocal tract model is designed as a *finite impulse response* (FIR) filter and performs the convolution of the excitation and the impulse response, which is computed as the inverse cepstral transformation of the windowed complex cepstrum into the time domain. The used cepstral window is rectangular and selects only the relevant cepstral coefficients near $n = 0$. These cepstral coefficients correspond to slow changes in the spectrum magnitude, the fast

changes can be caused by the fundamental frequency or by other disturbances. The length of the cepstral window must be chosen with regard to the accuracy of the vocal tract model, the frame length and the sampling frequency. For $F_S = 8$ kHz we use a cepstral window of the length $N = 50$, which is centered on $n = 0$.

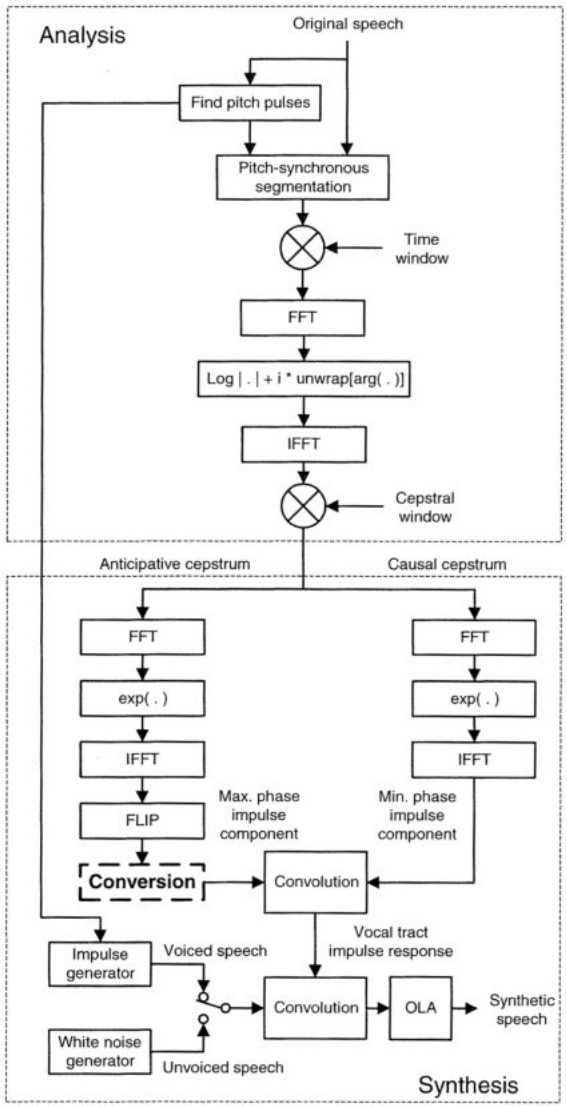

Figure 2 - Complex cepstrum speech analysis and synthesis with conversion of the maximum phase component obtained from the anticipative complex cepstrum.

4 Conversion of the Maximum-Phase Part of the Speech Signal

Many authors agree in the fact that the character of the emotional speech is given by the excitation produced by the glottis. As already mentioned the complex cepstrum is composed by the causal and the anticipative parts. After separate inverse cepstral transformation of the

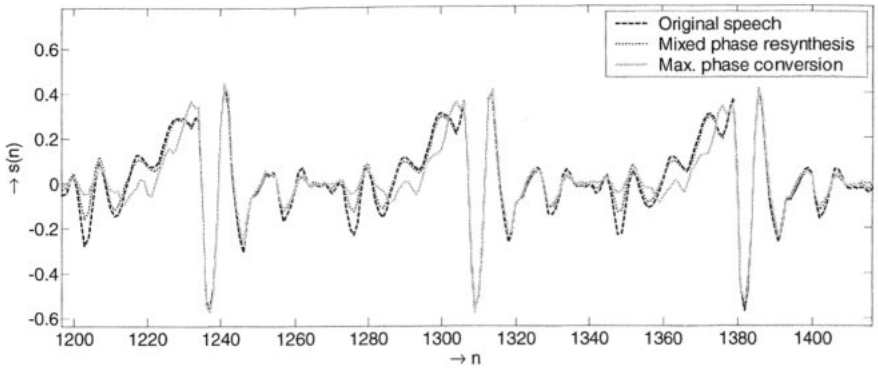

Figure 4 - Comparison of the original speech, mixed phase resynthesis and mixed phase resynthesis with the converted maximum-phase component (OQ = 0.54 → 0.36)

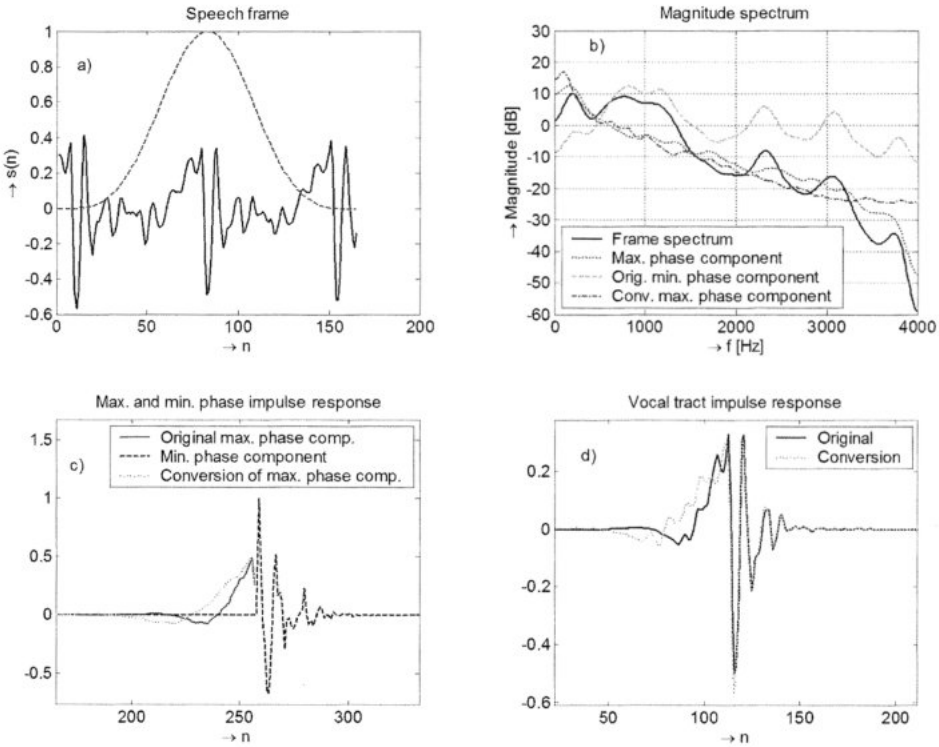

Figure 5 - Complex cepstrum speech decomposition in maximum- and minimum-phase components and the extension of the maximum-phase part.

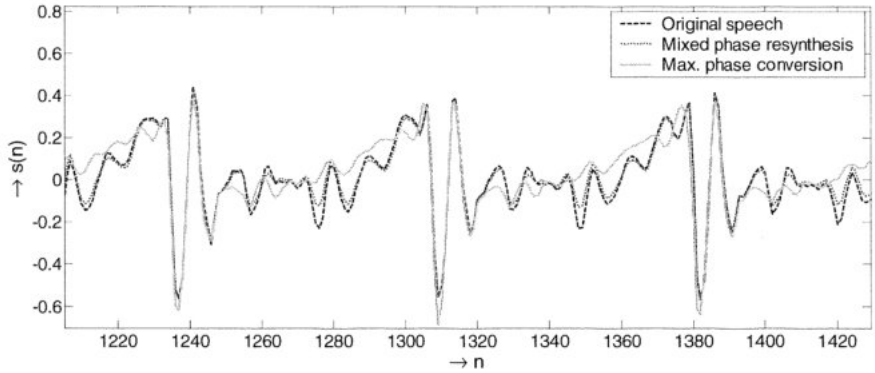

Figure 6 - Comparison of the original speech, mixed phase resynthesis and mixed phase resyntesis with the converted maximum-phase component (OQ = 0.54 → 0.90).

5 Conclusion

The aim of the described speech conversion is only a modification of the glottal pulse. This results in a change of the voice character and does not alter the voice identity. Together with a change of the prosody, this conversion could be exploited for the synthesis of emotional speech.

Speech converted in our study sounds according to the assumption a little bit tense for the shortening of the maximum-phase speech component and a little bit loose for the extension of the maximum-phase component. When trying to change the style of emotional speech we have to change also in addition to this speech modification the prosody of the speech (i.e. F0, timing and intensity). As discussed further we had some problems in some speech parts with the maximum- and minimum-phase speech decomposition using the complex cepstrum. Therefore we performed these experiments only with steady parts of vowels.

The decomposition of the speech signal based on the complex cepstrum into the minimum- and maximum-phase parts requires precise localization of the pitch pulses, i.e. the GCI for pitch synchronous segmentation and is very sensitive to the applied window for the weighting of the modeled speech frame [6, 7]. The best results can be achieved using the Hann-Blackman or the Blackman window. These windows are on the other hand not optimal from the point of view of the speech reconstruction, because after the application of the OLA algorithm in speech synthesis they do not ensure a constant (not rippled) signal envelope. From this point of view the best choice is the Hann window, which fulfils the constant response in adding the synthesized frames by OLA.

The estimation of the complex cepstrum requires also a reliable algorithm for phase unwrapping. This can be sometimes achieved using a sufficient high dimension of the FFT. Nevertheless in the cepstral analysis of natural sounds it can occur that the estimation of the maximum-phase part does not lead to a reasonable glottal pulse (the spectrum of the maximum-phase part is not falling). Also in this case the speech reconstruction is possible, but the described conversion of the glottal pulse results in a total destruction of the synthesized signal. For that reason we shall try to apply the zeros of the Z transform of the windowed speech frame for the calculation of the complex cepstrum [7]. This approach eliminates phase unwrapping and may help to understand the reasons of sometimes unreliable maximum-phase components.

Acknowledgment

This paper has been supported within the framework of COST2102 by the Ministry of Education, Youth and Sport of the Czech Republic, project number OC08010 and by the research project 102/09/0989 by the Grant Agency of the Czech Republic.

References

[1] Vích, R.: Nichtkausales cepstrales Sprachmodell. In R. Hoffmann (Ed): Elektronische Sprachsignalverarbeitung 2009, Dresden, September 21.-23. 2009, Vol.1, Studientexte zur Sprachkommunikation: 53, TUDpress Dresden, 2009, pp.107-114.

[2] Vích, R., Vondra, M.: Complex Cepstrum in Speech Synthesis. In: J. Jan, (Ed.), Proceedings of Biosignal 2010: Analysis of Biomedical Signals and Images, Vol. 20, June 27-29, Brno University of Technology, 2010, pp. 37-42.

[3] Vích, R.: Pitch Synchronous Linear Predictive Czech and Slovak Text-to-Speech Synthesis. In: Proc. of the 15th International Congress on Acoustics ICA 95, Trondheim, Norway, 1995, Vol.III, pp. 181-184.

[4] Vích, R.: Cepstrales Sprachmodell, Kettenbrüche und Anregungsanpassung in der Sprachsynthese. Wissenschaftliche Zeitschrift der Technischen Universitaet Dresden, Vol. 49, No. 4/5, 2000, pp. 116-121.

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

[6] Drugman, T., Bozkurt, B., Dutoit, T.: Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation. Interspeech 2009, Brighton, U.K., 2009, pp. 116-119.

[7] Bozkurt, B.: Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signal. PhD. Thesis, Faculté Polytechnique de Mons, Belgium 2005.

[8] Pulakka, H.: Analysis of Human Voice Production Using Inverse Filtering, High-Speed Imaging, and Electroglottography. Master Thesis, Helsinki University of Technology, 2005.

[9] Bozkurt, B., Couvreur, L., Dutoit, T.: Chirp group delay analysis of speech signals. Speech Communication, Vol. 49, 2007, pp. 159-176.