Comparison of speech signal reconstruction enhanced in the spectral domain by overlap-and-add and using the cepstral vocoder

Robert Vích, Martin Vondra

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

Introduction

Cepstral vocoder has been originally proposed for replacing the LPC vocoder for obtaining more natural text–to–speech synthesis [1]. For its high quality speech modeling it has been also successfully used for voice conversion [2]. In this paper the cepstral vocoder is further applied for speech reconstruction after speech enhancement in the spectral domain. In this case first, in the analysis step, an algorithm for noise suppression in the spectral domain is used and then the estimation of the cepstral model follows, which leads to more reliable speech parameters. Speech is synthesized by the cepstral vocoder, which works in the source-filter mode. For speech enhancement in the spectral domain the Minimum Mean-Square Error Log-Spectral Amplitude (MMSE LSA) algorithm developed by Ephraim and Malah [3] was chosen. Speech synthesized by the cepstral vocoder and by overlap-and-add (OLA) algorithm is compared.

Cepstral vocoder

The cepstral vocoder is composed of the analysis and synthesis parts, see Fig. 1. The speech parameters are estimated in the analysis part. For vocal tract modeling with 8 kHz sampling frequency the first 26 real cepstral coefficients are computed and the information about excitation given by the fundamental frequency is estimated.

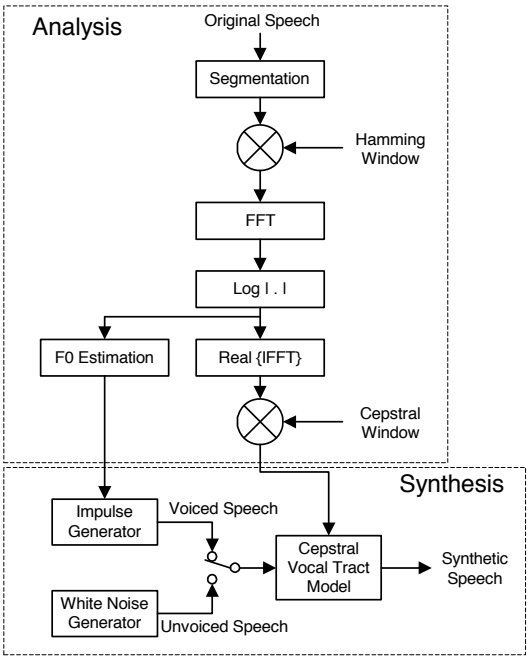

Figure 1: Block diagram of the cepstral vocoder.

The synthesis part consists of the excitation (for voiced speech an impulse sequence with period corresponding to the fundamental period of speech and for unvoiced speech white noise) and of the cepstral model of the vocal tract. The cepstral vocal tract model is modeled by a time varying digital filter with coefficients given by the minimum phase part of the real cepstrum.

If we would use the cepstral vocoder directly for noisy speech modeling, the output speech would be of decreased quality and also of low intelligibility. Noise would cause distortion of the vocal tract parameters and of the estimated fundamental frequency. For that reason we implement a classical speech enhancement algorithm in the analysis part of the cepstral vocoder. We use the MMSE LSA algorithm proposed by Ephraim and Malah [3]. This improves to a certain extent the quality of the output speech, but the enhanced magnitude spectrum can be in addition distorted by the enhancement, particular by not perfect estimation of the noise spectrum. This also influences the estimated parameters of the vocal tract. The distortion of the enhanced magnitude spectrum is smallest on its maxima. For this reason we use a sophisticated technique of magnitude spectrum smoothing that fits only on the maxima. This technique is called Non Linear Envelope Detection (NLED) and was adapted from [4]. The modified analysis part of the cepstral vocoder for application in speech enhancement is shown in Fig. 2.

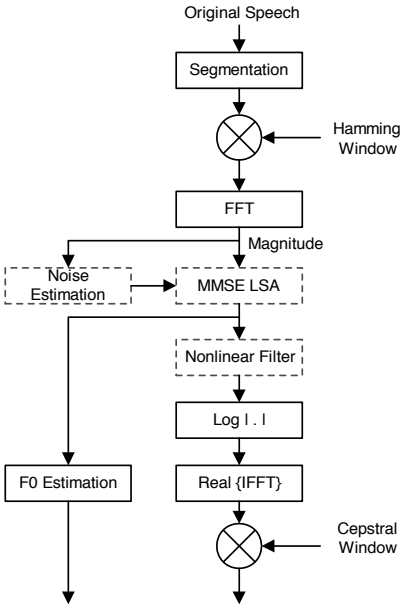

Figure 2: Analysis part of the cepstral vocoder for speech enhancement.

Nonlinear Envelope Detection

In the original cepstral vocoder, the vocal tract parameters are obtained by cepstral windowing. Cepstral windowing can be interpreted in the spectral domain like smoothing of the magnitude spectrum by a noncausal filter with impulse response given by the magnitude spectrum of the cepstral window (hidden homomorphic processing [5]). This smoothing is a linear convolution of the magnitude spectrum sequence with that impulse response of the noncausal filter. But classical linear convolution does not ignore spectral valleys which are distorted by spectral enhancement. For this reason we use nonlinear convolution that depends only on spectral peaks. Nonlinear convolution in the spectral domain is given by the following formula

$$Y(k) = \max_i \left[S(i) H(k-i) \right],$$

where $H(k)$ is the impulse response of the filter, $S(k)$ is the input magnitude spectrum and $Y(k)$ is the output smoothed spectrum. The effect of spectral envelope smoothing by classical cepstral windowing and by NLED on enhanced short-time magnitude spectrum can be seen in Fig. 3. The thin solid line is the enhanced speech magnitude spectrum. The dashed envelope is obtained by classical cepstral windowing. For 8 kHz sampling frequency we use a rectangular window of length $N_{\text{ceps}} = 26$. The dotted line is the output of the nonlinear convolution (NLED) of the enhanced magnitude spectrum. The impulse response $H(k)$ of the smoothing filter was set based on [5]. In this case the Hann window of length 25 was used, which corresponds for FFT length $N_{\text{F}} = 512$ to maximum fundamental frequency $F_{0\text{max}} = 200$ Hz. The thick solid line represents the resulting magnitude spectrum of the cepstral vocal tract model, which was obtained by cepstral windowing (rectangular window of length $N_{\text{ceps}} = 26$) of the cepstrum corresponding to the dotted NLED speech magnitude spectrum depicted in Fig. 3.

It can be seen that the thick solid envelope fits the spectral maxima. The peaks and valleys of this envelope have greater range and the area where speech is more present has higher energy than the envelope obtained by classical cepstral windowing.

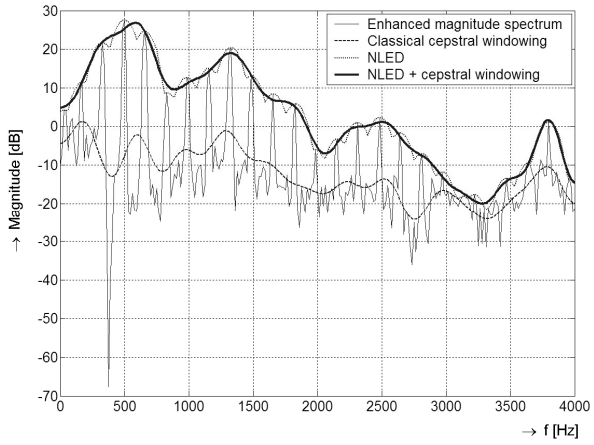

Figure 3: Comparison of magnitude spectra for various smoothing approaches.

Conclusion

In Fig. 4 the spectrograms of noisy speech, enhanced speech by MMSE LSA algorithm with reconstruction using OLA algorithm and the cepstral vocoder are shown. In the lower spectrogram it can be seen that the harmonic frequencies are more visible. This is given by the pure periodic excitation in the cepstral vocoder. The residual noise in the low frequency region is also better suppressed by the proposed approach using the cepstral vocoder. Subjective comparison is not such unambiguous. The output of the cepstral vocoder sounds a little buzzy for voiced speech, but according to our opinion the intelligibility of the enhanced speech reconstructed using the cepstral vocoder is somewhat higher than speech reconstructed by inverse Fourier transform and OLA algorithm.

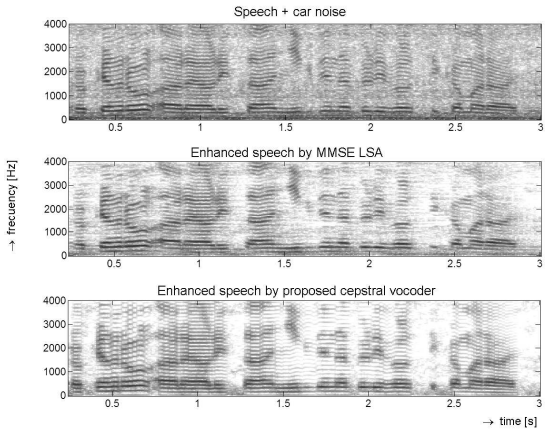

Figure 4: Comparison of spectrograms of noisy speech and enhanced speech with reconstruction by IFFT and OLA algorithm and using the proposed cepstral vocoder.

Acknowledgment

This paper has been supported by the National research program "Information Society" of the Academy of Sciences of the Czech Republic, project number 1ET301710509.

References

[1] Vích, R., Smékal, Z.: All-Pole and Zero-Pole Speech Modeling. In: Proceedings of the 14th Biennial International Conference Biosignal'98 June 23-25, 1998, Brno, Czech Republic, pp. 196-199

[2] Vondra, M.: Voice Transformation in Vocoders and TTS Systems. PhD. Dissertation (in Czech), Brno University of Technology, 2005

[3] Ephraim, Y., Malah, D.: Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator. IEEE Transactions on Acoustic, Speech, and Signal Processing, April 1985, ASSP-33, pp. 443–445

[4] Zhu, Q., Alwan, A.: Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise. Computer Speech and Language, Vol. 17, 2003, pp. 381-402

[5] Vích, R. Vondra, M.: Speech Spectrum Envelope Modeling. In: A. Esposito at al. (Eds.): Verbal and Nonverbal Commun. Behaviors, LNAI 4775, pp. 129-137. 2007