Pseudo Cepstral Analysis of Czech Vowels

Robert Vích

Institute of Radio Engineering and Electronics,
Academy of Sciences of the Czech Republic,
Chaberská 57 CZ-182 52 Prague 8, Czech Republic
vich@ure.cas.cz

Abstract. Real generalized cepstral analysis is introduced and applied to speech deconvolution. Real pseudo cepstrum of the vocal tract model impulse response is defined and applied to the analysis of Czech vowels. The energy concentration measure of the real pseudo cepstrum of the vocal tract model impulse response is introduced and evaluated for Czech vowels pronounced by male and female speakers. The goal of this investigation is to find a robust and more reliable method of vocal tract modeling also for voices with high fundamental frequency, i.e. for female and child voices. From the investigation follows that vowel and speaker dependent generalized cepstral analysis can be found which is more robust in speech modeling than cepstral and LPC analysis.

1 Introduction

In the papers [1-3] a parametric speech modeling approach based on homomorphic signal processing [4] using spectral analysis was presented and applied to speech synthesis.

In 1979 Lim [5] suggested a new nonlinear signal transformation, which converts the convolution of two signals, one of which is a train of pulses, into another convolution, in which the transformed impulse response is shorter than the original one and better separated from the other signal. Using this transformation it is possible to deconvolve a speech signal, i.e. to extract the transformed impulse response from the transformed speech signal by applying a suitable window and after inverse transformation to obtain the impulse response of the vocal tract model with greater accuracy than in classical homomorphic deconvolution. Let us call this nonlinear signal transformation generalized homomorphic approach.

In papers [6-8] this principle of generalized homomorphic signal analysis was applied to speech deconvolution and to vocal tract modeling. A comparison of the computational complexity of cepstral IIR and FIR vocal tract models may be found in [9]. In this contribution generalized homomorphic signal analysis is used for the analysis of Czech vowels uttered by male and female speakers and compared with results obtained by homomorphic signal analysis.

2 Generalized Cepstral Analysis

At first the procedure proposed by Lim is briefly summarized and applied to speech analysis. The voiced speech signal $s(n)$ may be described by the convolution

$$s(n) = p(n) * h(n) \ . \tag{1}$$

$p(n)$ is the sequence of the vocal tract excitation impulses with the fundamental frequency period L and $h(n)$ is the impulse response of the vocal tract model. The parameter L is the fundamental frequency period expressed by the number of speech samples.

Fourier transform of the convolution (1) leads to

$$S(\omega) = P(\omega) \cdot H(\omega) \ . \tag{2}$$

$S(\omega)$ represents the speech signal spectrum, $P(\omega)$ is the spectrum of the excitation signal $p(n)$ and $H(\omega)$ is the frequency response of the vocal tract model. For spectrum calculation we use fast Fourier transform with the dimension N_F **.**

A new transformation is searched for, which converts the convolution (1) into another convolution

$$\breve{s}(n) = \breve{p}(n) * \breve{h}(n) \tag{3}$$

with shorter "impulse response" $\breve{h}(n)$, which is better recognizable in the transformed speech signal $\breve{s}(n)$. The sequences $\breve{s}(n)$, $\breve{p}(n)$ and $\breve{h}(n)$ are the transformed sequences of the corresponding signals $s(n)$, $p(n)$ and $h(n)$ in (1).

This generally nonlinear transformation is performed in the frequency domain followed by inverse Fourier transform. The aim is to find a suitable function $f(S(\omega))$ for speech spectrum transformation.

In homomorphic analysis the function $f(S(\omega)) = \ln S(\omega)$ is applied in the definition of the complex cepstrum. In real cepstrum computation $f(S(\omega)) = \ln|S(\omega)|$ is used and for estimation of the autocorrelation sequence in the time domain we use $f(S(\omega)) = |S(\omega)|^2$. In [9] a unifying view on cepstral and correlation analysis was presented. Lim proposed for the spectrum transformation the function

$$f(S(\omega)) = (S(\omega))^{\gamma} \ , \ -1 \le \gamma \le 1 \ . \tag{4}$$

In this contribution we shall not use the complex spectrum $S(\omega)$ as the argument of the function $f(\cdot)$, like in the definition of the complex speech cepstrum, but the magnitude speech spectrum $|S(\omega)|$ as in the computation of the real cepstrum. The transformation in the frequency domain is therefore defined as

$$f(S(\omega)) = |S(\omega)|^{\gamma} \ . \tag{5}$$

The application of this transformation to (2) results in

$$\breve{S}_{\gamma}(\omega) = |S(\omega)|^{\gamma} = |P(\omega)|^{\gamma} \cdot |H(\omega)|^{\gamma} = \breve{P}_{\gamma}(\omega) \cdot \breve{H}_{\gamma}(\omega) \ . \tag{6}$$

The symbols $\breve{S}_{\gamma}(\omega)$, $\breve{P}_{\gamma}(\omega)$ and $\breve{H}_{\gamma}(\omega)$ are introduced for the Fourier transforms of the magnitude spectra transformed with the parameter γ.

By inverse Fourier transform the convolution in the form of (3) is obtained, but with new sequences $\breve{s}_{\gamma}(n)$, $\breve{p}_{\gamma}(n)$ and $\breve{h}_{\gamma}(n)$, i.e.

$$\breve{s}_{\gamma}(n) = \breve{p}_{\gamma}(n) * \breve{h}_{\gamma}(n) \ . \tag{7}$$

We shall call the transformed sequences *real pseudo cepstra* corresponding to the signals $s(n)$, $p(n)$ and $h(n)$, respectively they could be called *pseudo correlation sequences*. Lim calls them *spectral root cepstra***.**

The real pseudo cepstra are *two sided*, they have a *causal* and an *anticipative* part. Since the Fourier transforms in (6) of the real pseudo cepstra are real, for the sequences in (7) hold

$$\breve{s}_{\gamma}(n) = \breve{s}_{\gamma}(-n), \ \breve{p}_{\gamma}(n) = \breve{p}_{\gamma}(-n), \ \breve{h}_{\gamma}(n) = \breve{h}_{\gamma}(-n) \ .$$

The pseudo cepstrum $\breve{p}_{\gamma}(n)$ of the periodic excitation contains a quasi periodical component with the fundamental period *L* of the voiced excitation. In the following $\breve{p}_{\gamma}(n)$ will not be examined.

As already mentioned in Chapter 1, the aim of the pseudo cepstral approach is the robust extraction of the pseudo cepstrum $\breve{h}_{\gamma}(n)$ by windowing the speech pseudo cepstrum $\breve{s}_{\gamma}(n)$. Several approaches for inverse pseudo cepstral transformation of $\breve{h}_{\gamma}(n)$, i.e. for approximate estimation of $h(n)$, are summarized in [6-8].

3 Concentration Measure of the Transformed Impulse Response

For evaluation of the effective duration of the transformed impulse response $\breve{h}_{\gamma}(n)$ of the vocal tract model we define, according to Lim, the *energy concentration measure* $d_M(\gamma)$ for $M = Lp$, where *L* is the fundamental frequency period of the voiced excitation and p is a chosen constant, $1 \geq p \geq 0$. In our experiments we use $p = 0.95$, i.e. we apply for windowing of the transformed impulse response $\breve{h}_{\gamma}(n)$ a rectangular window of length $M = 0.95L + 1$.

The concentration measure is given as

$$d_M(\gamma) = \frac{\sum_{n=1}^{M} \breve{h}_\gamma^2(n)}{\sum_{n=1}^{N_F/2} \breve{h}_\gamma^2(n)} \ . \tag{9}$$

The low summation index is set $n = 1$, since $\breve{h}_\gamma(0)$ corresponds to the mean value of the transformed spectral function $\breve{S}_\gamma(\omega) = |S(\omega)|^\gamma$ and it is not convenient to consider it in the concentration measure $d_M(\gamma)$.

4 Concentration Measure for Male and Female Voices

In the following experiment we evaluate $d_M(\gamma)$ for Czech vowels uttered by a male and a female for several values of γ in the interval $-1 \le \gamma \le 1$. We use the stationary parts of the sounds *a, e, i, o, u* sampled with the sampling frequency $F_s = 16$ kHz. The dimension of the applied FFT in speech spectral analysis is $N_F = 1024$. The fundamental frequency period of the male speaker is approximately $L = 186$ ($F_0 = 86$ Hz), for the female speakerc $L = 91$ ($F_0 = 176$ Hz). The coefficient $p = 0.95$ and the frame length for both voices was set $N = N_F$. For spectrum analysis Hamming windowing was applied.

In Fig. 1 and 2 we see the energy concentration measure $d_M(\gamma)$ for the male and female voice respectively, as function of the parameter γ. The concentration measure for the real logarithmic cepstrum $d_M(0)$ is shown with an asterisk on all curves. It corresponds to the value $\gamma = 0$.

It can be seen that the curves $d_M(\gamma)$ have a maximum in the neighborhood of $\gamma = 0$. The positions of the maxima depend on the ratio of the numbers of poles and zeros of the corresponding speech models, which was already stated in the paper by Lim using an experimental signal model. For a system with only zeros in its transfer function, i.e. for a finite impulse response system (FIR), the maximum of $d_M(\gamma)$ is located in the neighborhood of $\gamma = 1$. In the case of an all pole transfer function, the maximum lies at $\gamma = -1$. For an infinite impulse response system (IIR) with equal number of poles and zeros the maximum of $d_M(\gamma)$ is positioned at $\gamma = 0$. This statement is not peremptory, the maximizing value of γ depends also on the mutual position of the formants and on the fundamental frequency period L, i.e. it is speaker dependent.

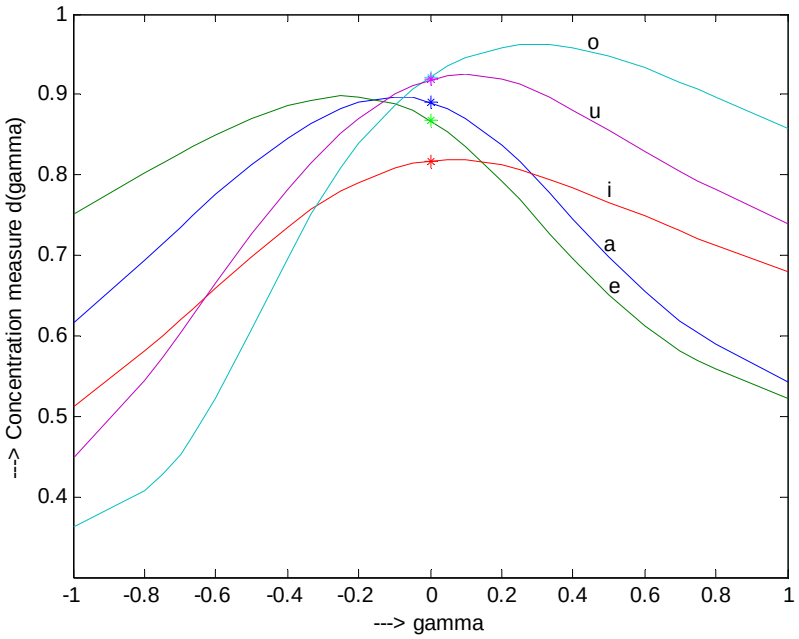

Fig. 1. Energy concentration measure $d_M(\gamma)$ for the male voice.

When comparing Fig. 1 and Fig. 2 it seems that the speech models for the male and female voice have different number of poles and zeros for the same vowel. Further the maximum values $d_M(\gamma)$ for the female voice are mostly smaller than for the male voice, which is given by the shorter fundamental frequency period L of the female voice in relation to the effective length of its transformed speech model impulse response $\breve{h}_\gamma(n)$.

5 Conclusion

The aim of the pseudo cepstral approach to speech analysis is to achieve greater accuracy of the vocal tract model in comparison to the accuracy obtained using cepstral speech modeling. The optimum value of the parameter γ is in relation to the number of formants and antiformants and to the fundamental frequency of voiced sounds. The pseudo cepstral speech model for $\gamma = 0$ corresponds to the cepstral speech model. The approximation error of the pseudo cepstral speech model depends on the parameter γ and on the length and type of the window used for pseudo cepstrum weighting.

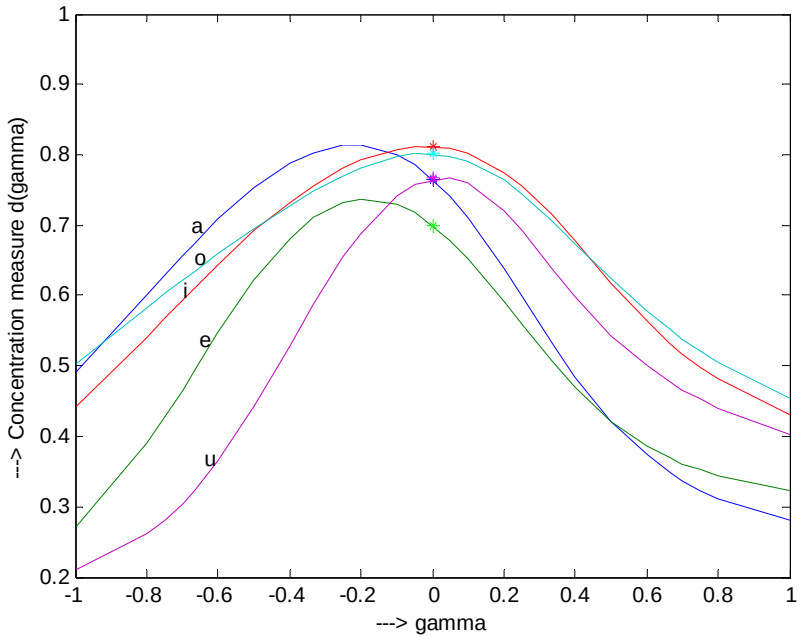

Fig. 2. Energy concentration measure $d_M(\gamma)$ for the female voice.

The second goal of this investigation is to find a robust and more reliable method of vocal tract modeling also for voices with high fundamental frequency, i.e. for female and child voices. The signal transformation described in this paper has been tested using a synthetic signal with three formants and a variable fundamental frequency [10]. It has been verified that the accuracy of the pseudo cepstral speech model is greater than that of the cepstral one, but the difference is not so convincing.

Pseudo cepstral speech analysis has been applied in the last years also in recognition of noisy speech, e.g. in papers by Alexandre and Lockwood [11], Zühlke [12] and Chilton and Marvi [14]. In these contributions an improvement was registered, but for different values of the parameter γ. Another application of generalized homomorphic speech analysis is its use for fundamental frequency estimation, which is summarized in [15].

References

1. Vích, R.: Cepstrales Sprachmodell, Kettenbrüche und Anregungsanpassung in der Sprachsynthese. Wissenschaftliche Zeitschrift der Technischen Universtät Dresden, Vol. 49 No 4/5 (2000) 119-121

2. Vích R.: Cepstral Speech Model, Padé Approximation, Excitation and Gain Matching in Cepstral Speech Synthesis. In: Proc. of the 15th Biennial International EURASIP Conference BIOSIGNAL Brno (2000) 77-82
3. Vích, R., Smékal, Z.: Speech Signals and their Models. In: Horová, I. (ed.): Summer School DATASTAT'01, Proceedings, Folia Facultatis Scientiarum Naturalium Universitatis Masarykianae Brunensis, Mathematica 11 (2002) 275-289
4. Oppenheim, A.V., Schafer, R.W.: Digital Signal Processing. Prentice-Hall, N. Jersey (1989)
5. Lim, J. S.: Spectral Root Homomorphic Deconvolution System. IEEE Transactions on Acoustics Speech, and Signal Processing, Vol. ASSP-27 No. 3 (1979) 223-233
6. Vích, R.: Experimente mit der Anwendung der Pseudokorrelation bei der Vokaltraktmodellierung. In: R. Hoffmann (Ed.): Tagungsband der 13. Konferenz Elektronische Sprachsignalverarbeitung, Dresden, Studientexte zur Sprachsignalverabeitung Vol. 24 (2002) 253-260
7. Vích, R.: Pseudocepstrale Sprachanalyse und Konstruktion eines Vokaltraktmodells mit endlicher Impulsantwort. In: D. Wolf (Ed.) Signaltheorie und Signalverarbeitung, Akustik und Sprachakustik, Informationstechnik. Arild Lacroix zum 60. Geburtststag, Studientexte zur Sprachkommunikation Vol. 29 (2003) 126-132
8. Vích, R.: Pseudo Cepstral Speech Analysis. In: Horová, I. (ed.): Summer School DATASTAT'03, Proceedings, Folia Facultatis Scientiarum Naturalium Universitatis Masarykianae Brunensis, Mathematica 15, (2004) 277-291
9. Vích, R., Horák, P., Schwarzenberg, M.: Korrelation von Sprachsignalen im Zeit-und Frequenzbereich. In: Hoffmann, R., Ose, R. (Eds.): Tagungsband der 6. Konferenz Elektronische Sprachsignalverarbeitung (1995) 10-13
10. Vondra, M., Vích, R.: Design of FIR Vocal Tract Models with Linear and Nonlinear Phase. In: R. Vích (Ed.): Proc. of the 12th Czech-German Workshop on Speech Processing Prague (2002) 28-32
11. Vích, R., Plšek, M.: New Methods for Speech Spectrum Smoothing and Formant Estimation. In: Tagungsband des 48. Internationalen Wissenschaftlichen Kolloquiums der TU Ilmenau Reihe 3.3 Sprachverarbeitung (2003) CD ROM
12. Alexandre, P., Lockwood, P.: Root Cepstral Analysis: A Unified View. Application to Speech Processing in Car Noise Environments. Speech Communication Vol. 12 (1993) 277-288
13. Zühlke, W.: Vergleich der Pseudokorrelationsbereiche mit dem Cepstralbereich. In: Konvens 2000 Sprachkommunikation Ilmenau (2000) 141-144
14. Chilton, E., Marvi, H.: Two-Dimensional Root Cepstrum as Feature Extraction Method for Speech Recognition. Electronics Letters, Vol. 39 No.10 (2003) 815-816
15. Hess, W.: Pitch Determination of Speech Signals. Algorithms and Devices. Springer-Verlag, Berlin Heidelberg New York Tokyo (1983)

Acknowledgements:

This paper was prepared within the framework of the research project AVOZ 20670512 and has been supported by the Ministry of Education, Youth and Sports of the Czech Republic OC 277.001 "Transformation of Segmental and Suprasegmental Speech Models".