Journal of ELECTRICAL ENGINEERING, VOL. 53, NO. 1-2, 2002, 46–49

COMMUNICATIONS

AUTOREGRESSIVE AND CEPSTRAL PARAMETRIZATION IN HARMONIC SPEECH MODELLING

Anna Madlová *

Two methods for coding the amplitude and minimum-phase spectra of the speech signal are compared. Autoregressive (AR) and cepstral parametrizations of the same spectral envelope use the same number of parameters. Computational complexity of the analysis is compared for the two methods. Using both AR and cepstral parametrization the harmonic synthesis is compared for concatenation of pitch-synchronous frames and overlap-and-add (OLA) of consecutive pairs of the same pitch-synchronous frames. Quality of synthesis is compared using the RMS log spectral measure. The mean value of the spectral measure is rather similar for all the four combinations of analysis and synthesis methods. However, the standard deviation of the spectral measure is lower for the OLA synthesis.

K e y w o r d s: harmonic speech model, autoregressive and cepstral parametrization

1 INTRODUCTION

The harmonic speech model [1], [2] is performed as a sum of harmonically related sine waves with frequencies given by pitch harmonics, and amplitudes and phases given by sampling the transfer function of the vocal tract model at these frequencies. For voiced speech, using a minimum-phase assumption of the vocal tract as well as the glottal pulse contribution, the logarithm of the magnitude frequency response and the phase frequency response form a Hilbert transform pair. The phases are randomized for unvoiced speech and for voiced speech above the voicing transition frequency. The harmonic model has already been used with autoregressive (AR) parametrization [3], and with cepstral parametrization [4]. Comparison of AR and cepstral parametrization together with concatenation and OLA synthesis is presented here.

2 HARMONIC MODEL

The speech signal synthesized by the harmonic model during one pitch period is given by

$$s(l) = \sum_{m=1}^{M} A_m \cos(\omega_m l + \varphi_m)\,, \tag{1}$$

where frequencies ω_m are given by pitch harmonics, and amplitudes A_m and phases φ_m are given by sampling the transfer function of the vocal tract model at these frequencies. For voiced speech, the voicing transition frequency is computed from the magnitude spectrum comparing the frequency distances between the pitch harmonics and the spectral local maxima. The phases are randomized for unvoiced speech and for voiced speech above the voicing transition frequency. The harmonic speech model is drawn in Fig. 1. Input parameters for the vocal tract transfer function are represented by pitch and AR or cepstral parameters.

3 SPEECH SPECTRAL ENVELOPE

Before describing AR and cepstral parametrization let us introduce a speech spectral envelope which is used to compute AR and cepstral parameters. First, each speech frame is weighted by the normalized Hamming window. Then, the staircase log spectral envelope is determined using steps of a pitch-frequency width. In each of the intervals of a pitch width the local maxima are found by detection of the slope change from positive to negative. The mean value of their amplitudes is chosen as the amplitude of the step. If no local maximum is found in the interval using this algorithm, the mean value of the interval boundary amplitudes is chosen as the amplitude of the step. The resulting staircase envelope is smoothed using the weighted moving average having the shape of the normalized Blackman window. Block diagram of the method can be seen in Fig. 2.

4 AUTOREGRESSIVE PARAMETRIZATION

Using N parameters of the AR model, its magnitude frequency response is written as

$$\left|P_A\left(e^{j\omega}\right)\right| = \frac{G}{\left|1 + \sum_{n=1}^{N-1} a_n \exp(-jn\omega)\right|}\,, \tag{2}$$

where the coefficients a_n and the gain G are computed using the standard autocorrelation method, however, applied to the time-domain signal corresponding to

* Department of Radioelectronics, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia, e-mail: madlova@elf.stuba.sk

ISSN 1335-3632

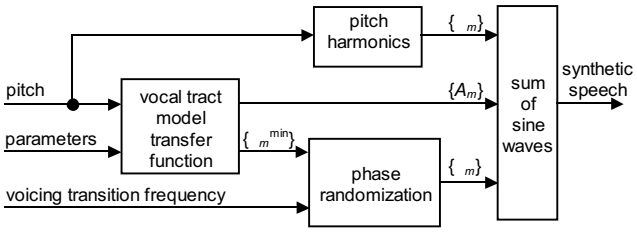

Fig. 1. Block diagram of the harmonic speech model.

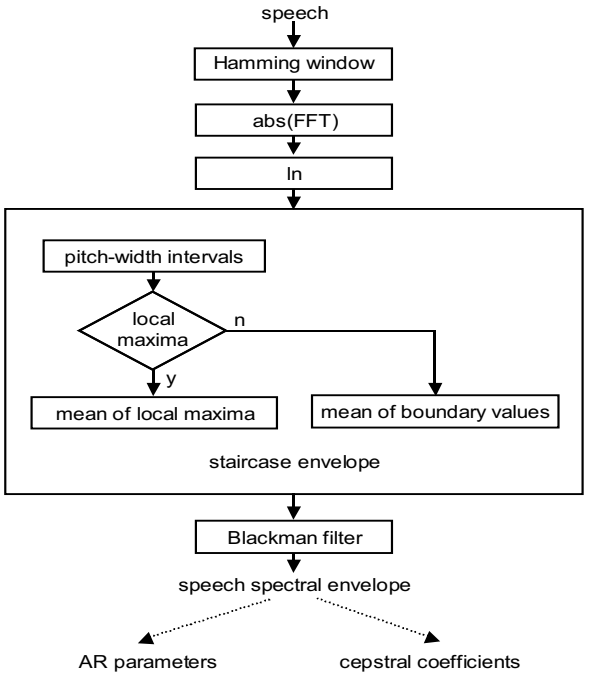

Fig. 2. Block diagram of the speech spectral envelope determination.

the speech spectral envelope, described in Section 3, instead of the original speech signal. Sampling the function $\left|P_A\left(e^{j\omega}\right)\right|$ and the Hilbert transform of its logarithm at frequencies ω_m, gives the amplitudes A_m and phases $\varphi_m^{\min}$ necessary for the harmonic model implementation using (1) and Fig. 1.

5 CEPSTRAL PARAMETRIZATION

Using N cepstral coefficients, the magnitude frequency response of the vocal tract model can be written in the following way

$$\left|P_C\left(e^{j\omega}\right)\right| = \exp\Big(c_0 + 2\sum_{n=1}^{N-1} c_n \cos n\omega\Big). \qquad (3)$$

Here, the cepstrum c_n is given by the inverse Fourier transform of the logarithm of the spectral envelope described in Section 3. Sampling the function $\left|P_C(e^{j\omega})\right|$ and the Hilbert transform of its logarithm at the frequencies ω_m, gives the amplitudes A_m and phases $\varphi_m^{\min}$ necessary for the harmonic model implementation using (1) and Fig. 1. However, instead of sampling $\left|P_C\left(e^{j\omega}\right)\right|$, explicit analytic relations may be used

$$A_m = \exp\Big(c_0 + 2\sum_{n=1}^{N-1} c_n \cos n\omega_m\Big), \qquad (4)$$

$$\varphi_m^{\min} = -2\sum_{n=1}^{N-1} c_n \sin n\omega_m . \qquad (5)$$

6 EXPERIMENTAL RESULTS

It had been found out by simulation that the minimum number of 26 cepstral coefficients is necessary for sufficient log spectrum approximation at 8-kHz sampling [5], and 51 cepstral coefficients are necessary for sufficient log spectrum approximation at 16-kHz sampling [6]. For that reason 26 parameters were used for autoregressive as well as cepstral parametrization using the sampling frequency of 8 kHz.

Comparison of the speech spectra obtained from the AR and cepstral parameters is shown in Figs. 3 and 4. Figure 3 shows the vocal tract model magnitude frequency response using the same number of 26 parameters computed from the same speech spectral envelope determined by the method shown in Fig. 2. It can be seen that sampling the AR model magnitude frequency response approximates the original spectral peaks more properly than sampling the cepstral model magnitude frequency response. However, increasing the number of model parameters the frequency responses of the AR and cepstral vocal tract models are rather similar and they both converge to the real spectral envelope. It can be seen in Fig. 4 for 42 AR as well as cepstral parameters.

Apart from two methods of analysis (AR and cepstral parametrization), two methods of synthesis were performed: concatenation of pitch-synchronous frames and overlap-and-adding (OLA) of consecutive pairs of the same pitch-synchronous frames. Prior to OLA each pair of frames was weighted by an asymmetric Hanning window. For every pair of consecutive pitch-synchronous frames the pitch period of the first frame is used for determination of the pitch harmonics. The AR or cepstral parameters of the first and the second frame of the pair are averaged. The harmonic parameters are determined according to Section 4 or 5. Speech is synthesized as a sum of sine waves during two consecutive pitch-synchronous frames. Then, this pair of frames is weighted by an asymmetric Hanning window with its left and right parts corresponding to the pitch periods of the first and the second frame so that the left part of the current asymmetric window has the same length as the right part of the previous window, and the right part of the current window has the

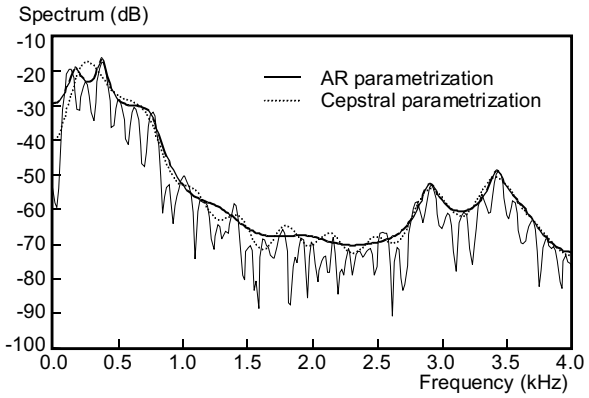

Fig. 3. Comparison of the spectra obtained from 26 parameters of the AR and cepstral model with prior spectral envelope for a 24-ms frame of a vowel "U" spoken by the male voice.

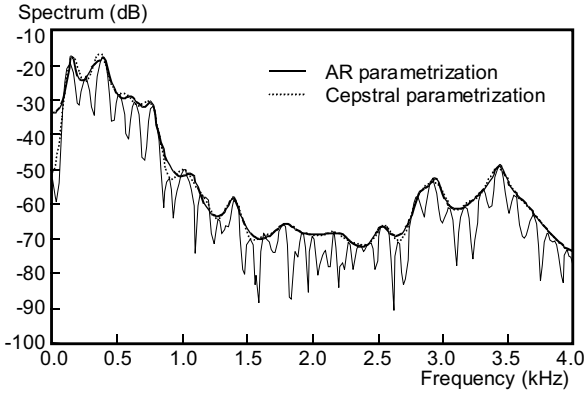

Fig. 4. Comparison of the spectra obtained from 42 parameters of the AR and cepstral model with prior spectral envelope for a 24-ms frame of a vowel "U" spoken by the male voice.

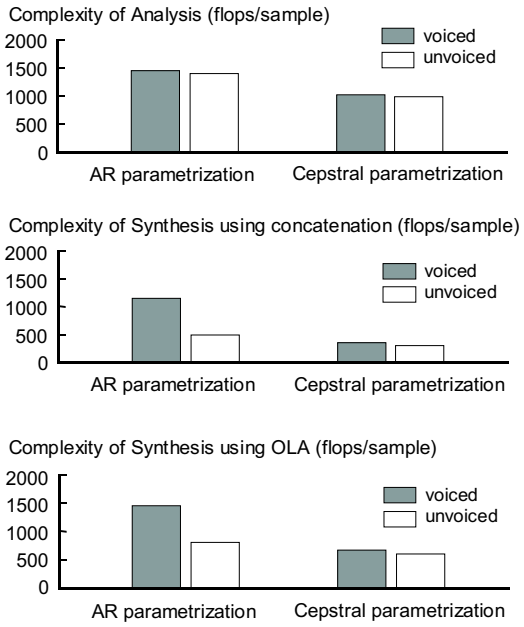

Fig. 5. Computational complexity for the harmonic model with AR and cepstral parametrizations of 26 parameters using concatenated and OLA synthesis.

same length as the left part of the next window, and the overlapped asymmetric windows are complementary. For the final synthesis the weighted overlapped consecutive pairs of pitch-synchronous frames are added to avoid discontinuities at the frame boundaries.

The computational complexity of the methods was measured in the number of floating-point operations per sample. The computational complexity of analysis using AR as well as cepstral parametrization and synthesis using concatenation as well as OLA is shown in Fig. 5. In each case voiced and unvoiced frames are evaluated separately. It is evident that cepstral parametrization is computationally less expensive than AR parametrization. Computational complexity of analysis using cepstral parametrization is lower because it needs only simple IFFT for cepstral parameters determination while AR parametrization needs transform of the logarithmic envelope into a linear spectral scale and application of autocorrelation method apart from IFFT. During synthesis with AR parametrization the frequency response must be computed and sampled and thus its computational complexity is higher than that of the cepstral parametrization where explicit analytic relations are used instead of sampling the frequency response. The highest computational complexity is required for voiced speech with AR parametrization because of computation of logarithm and its Hilbert transform. OLA is rather computationally expensive when compared with simple concatenation because an asymmetric window used for weighting pairs of frames must be computed for every frame pair, and multiplication of the window and pair of frames must be performed. The highest total computational complexity is necessary for AR parametrization with OLA, and the lowest total computational complexity is necessary for cepstral parametrization with concatenation.

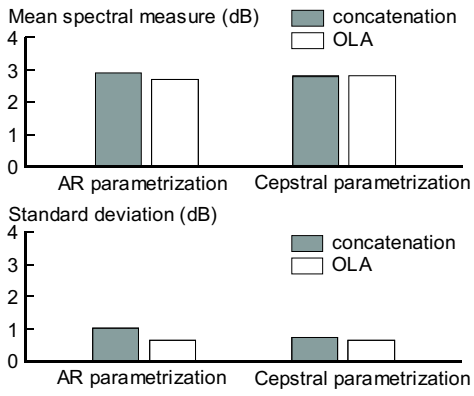

Fig. 6. RMS log spectral measure between the original and synthetic speech (concatenated and OLA) for the harmonic model with AR and cepstral parametrizations using 26 parameters.

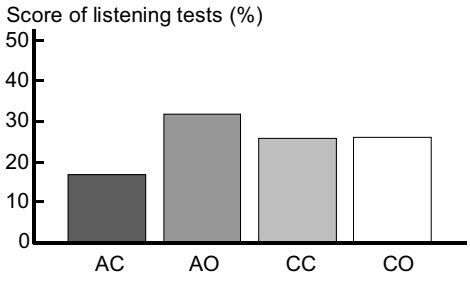

Fig. 7. Listening tests comparing 20 words synthesized using pitch-synchronous concatenation and OLA with AR and cepstral parametrization.
AC = AR parametrization with concatenative synthesis
AO = AR parametrization with OLA synthesis
CC = cepstral parametrization with concatenative synthesis
CO = cepstral parametrization with OLA synthesis

The RMS log spectral measure [7] was used to compare the smoothed spectra of original and resynthesized speech. The speech material consisted of about 450 stationary parts of 5 vowels and 2 nasals. The spectral measure was computed for the spectra of the speech frames weighed by a 24-ms Hamming window zero padded to 2048-point FFT. In Figure 6 we can see that the highest mean RMS log spectral measure as well as the highest standard deviation is given by AR parametrization with concatenation. The lowest mean value is observed for AR parametrization with OLA, while the standard deviation is almost the same for AR and cepstral parametrization with OLA.

Listening tests were performed in order to compare four combinations of the methods: AR parametrization with concatenative synthesis, AR parametrization with OLA synthesis, cepstral parametrization with concatenative synthesis, and cepstral parametrization with OLA synthesis. Twenty words synthesized by these methods were grouped into pairs of the same word synthesized by two methods. The words of the pairs were grouped in a random order and listeners had to choose the word with better resemblance to the original. Eight independent listening tests were performed. Scores given by the listeners are summarized in Fig. 7. We can see that the listening tests results are corresponding to the results of the RMS log spectral measure in Fig. 6, *ie* the highest score of listening tests corresponds to the lowest RMS log spectral measure and vice versa. It means that the use of the RMS log spectral measure is justified for determining perceptual resemblance of two speech signals.

7 CONCLUSION

Experiments have shown that the harmonic model with AR parametrization and OLA using an asymmetric Hanning window outperforms the cepstral parametrization and simple concatenative pitch-synchronous synthesis. However, it is achieved at the expense of increasing the computational complexity. In real situation a compromise must be found between speech quality and computational complexity.

References

[1] McAULAY, R. J.—QUATIERI, T. F.: Low-Rate Speech Coding Based on the Sinusoidal Model, Advances in Speech Signal Processing (Furui, S., Sondhi, M.M., eds.), Marcel Dekker, New York, 1992, pp. 165–208.

[2] McAULAY, R. J.—QUATIERI, T. F.: Sinusoidal Coding, Speech Coding and Synthesis (Kleijn, W.B., Paliwal, K.K., eds.), Elsevier Science, Amsterdam, 1995, pp. 121–173.

[3] MADLOVÁ, A.: An Experiment with Childish Voice Analysis and Synthesis., Radioelektronika '2000, Bratislava, Slovak Republic, pp. III–112-115, September 2000.

[4] MADLOVÁ, A: Harmonic Speech Model with Cepstral Parametrization, Speech Processing, 10th Czech-German Workshop (Vích, R., ed.), pp. 56–58,, Prague, Czech Republic, September 2000.

[5] PŘIBIL, J.: Comparison of Speech Spectral Features Using LPC Parameters and Cepstral Coefficients, 32nd Czech Acoustic Conference, pp. 63–66, Prague, Czech Republic, September 1995.

[6] PŘIBIL, J.: Comparison of Quality and Computational Complexity of Cepstral Speech Synthesis for Sampling Frequencies of 8 and 16 kHz, Applied Electronics, pp. 140–144, Plzeň, Czech Republic, September 1999. (in Czech)

[7] GRAY, A.—MARKEL, J. D.: Distance Measures for Speech Processing, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24, no. 5, pp. 380–391, October 1976.

Received 19 July 2001

Anna Madlová (Ing) was born in Hlohovec, Czechoslovakia in 1962. She received her Ing (MSc) degree in radioelectronics (medical electronics) from the Slovak University of Technology in Bratislava in 1985. For six years she had been with Chirana Research Centre for Medical Equipment as a research assistant. Since 1992 she has been working as a university teacher at the Department of Radioelectronics, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava