Preemphasis Influence on Harmonic Speech Model with Autoregressive Parameterization

Anna PŘIBILOVÁ

Department of Radio Electronics, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia

pribilova@kre.elf.stuba.sk

Abstract. *Autoregressive speech parameterization with and without preemphasis is discussed for the source-filter model and the harmonic model. Quality of synthetic speech is compared for the harmonic speech model using autoregressive parameterization without preemphasis, with constant and adaptive preemphasis. Experimental results are evaluated by the RMS log spectral measure between the smoothed spectra of original and synthesized male, female, and childish speech sampled at 8 kHz and 16 kHz. Although the harmonic model is used, the benefit of the adaptive preemphasis could be valid for the source-filter model, as well.*

Keywords

Autoregressive parameterization, harmonic model, source-filter model, preemphasis, spectral measure.

1. Introduction

An autoregressive (AR) model is well known in speech processing as a linear predictive coding (LPC) model being an all-pole model of a vocal tract. For the LPC model, preemphasis should be performed prior to the analysis and postemphasis should be performed as the last step of the synthesis [1]. Preemphasis is a simple and effective way of accenting the higher formants, thus allowing more accurate formant tracking results [2]. Almost invariably the first order preemphasis is used. Apart from the source-filter speech model, the all-pole model is also used for sine-wave amplitude coding, known as the minimum phase harmonic sine-wave speech model [3], [4]. However, no preemphasis is used here, although the authors describe a postfilter resembling the postemphasis [2] with the filter coefficient computed from the synthetic speech instead of the original speech, however, without any prefilter resembling the preemphasis. Neither other authors of harmonic speech modelling use any preemphasis [5] - [9]. On the other hand, adaptive preemphasis is used for example in source-filter based speech coding [10].

The presented paper investigates use of constant and adaptive preemphases for different voices synthesized by the harmonic speech model with AR parameterization. The RMS log spectral measure [11] is used as a comparison criterion, as in [12] it has been shown that it corresponds to the listening tests results.

2. Source-Filter Speech Model

The principle of the source-filter speech model is shown in Fig. 1. The model parameters determine the vocal tract transfer function $P(z)$. For voiced speech the excitation is formed by the impulse train. For unvoiced speech it is formed by the random noise. The output of the filter is a synthesized speech signal.

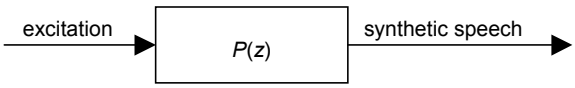

Fig. 1. The principle of the source-filter speech model.

Depending on the parameters describing the vocal tract transfer function, the source-filter model is called either autoregressive or cepstral. The AR model will be of concern here. Its transfer function $P(z)$ is given by the gain G and coefficients $\{a_n\}$ in the form

$$P(z) = \frac{G}{A(z)} = \frac{G}{1 + \sum_{n=1}^{N_A} a_n z^{-n}} \quad , \tag{1}$$

where N_A is the order of the AR model.

3. Harmonic Speech Model

The principle of the harmonic speech model is shown in Fig. 2.

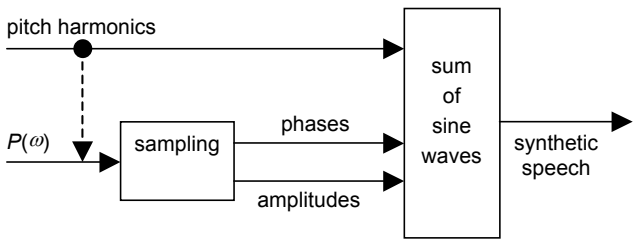

Fig. 2. The principle of the harmonic speech model.

It is performed as a sum of harmonically related sine waves with frequencies given by pitch harmonics, and amplitudes and phases given by sampling the frequency response of the vocal tract model at these frequencies. More details can be found in [13], [14].

AR parameterization of the harmonic model uses relations similar to that of the AR source-filter model. The magnitude frequency response of the AR model is given by

$$\left|P\left(e^{j\omega}\right)\right| = \frac{G}{\left|1+\sum_{n=1}^{N_A} a_n \exp\left(-jn\omega\right)\right|} . \qquad (2)$$

4. Preemphasis in AR Source-Filter Speech Modeling

The first-order preemphasis is performed as a single-zero filter $H_p(z) = 1 - \mu z^{-1}$ shown in a dashed block in Fig. 3. The coefficient μ is chosen between 0.9 and 0.95. In [2] it had been stated that the optimal preemphasis filter is the one which maximizes the output spectral-flatness measure, and it will have $\mu = r(1) / r(0)$, where $\{r(n)\}$ represents the autocorrelation sequence for the input speech data sequence. Thus, the adaptive preemphasis is computed in each speech frame from the present speech data. Apart from the parameters G and $\{a_n\}$, the pitch period L is computed from the non-preemphasized speech signal by some of the pitch determination algorithms [15].

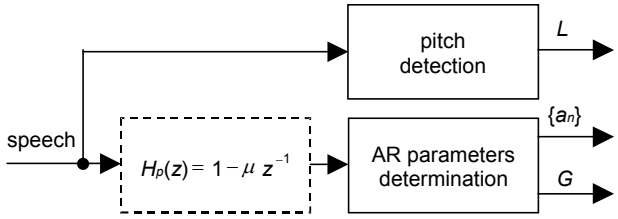

Fig. 3. Preemphasis in the AR speech analysis.

During synthesis the output of the all-pole vocal tract model, given by parameters G and $\{a_n\}$, is passed through a single-pole filter $1 / H_p(z)$ reciprocal of the preemphasis (Fig. 4). The coefficient μ is either constant or changes from frame to frame as determined during analysis.

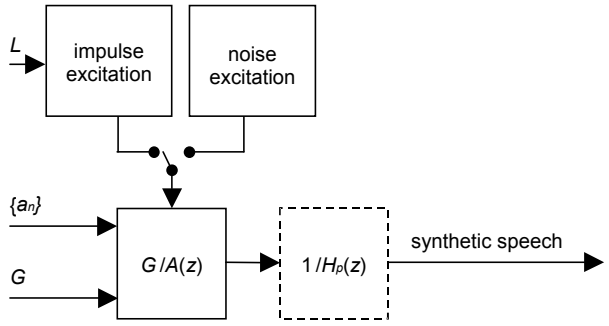

Fig. 4. Postemphasis in the AR source-filter speech synthesis.

For preemphasized speech one formant occurs approximately every 1 kHz and one pole pair is necessary to model each formant, so the AR order should be at least twice the bandwidth of the signal in kilohertz [16]. If no preemphasis and postemphasis were used, the vocal tract model should comprise the dashed block of Fig. 4 and the resulting all-pole model should have the order higher by the order of the postemphasis. In such a way, the AR order for speech without preemphasis must be odd instead of even AR order for speech with the first order preemphasis [14].

5. Preemphasis in Harmonic Speech Modeling with AR Parameterization

For the harmonic speech model with AR parameterization, the same speech analysis is performed as shown in Fig. 3. For AR parameters determination, the staircase log spectral envelope is smoothed and inverse Fourier transformed to get the time-domain signal corresponding to the spectral envelope of the original speech signal [13], [14]. Then the Levinson-Durbin autocorrelation algorithm is used to compute the parameters $\{a_n\}$ and G. Detection of the pitch period L is performed by the Rabiner autocorrelation method with three-level clipping of the signal [15] using different length of analysis frames for male, female, and childish voices.

The synthesis of the AR harmonic model (Fig. 5) uses the same input parameters as the synthesis of the AR source-filter model (Fig. 4). These parameters (L , $\{a_n\}$, G) are used to compute the parameters of the harmonic model ($\{\omega_m\}$, $\{\varphi_m\}$, $\{A_m\}$). Pitch harmonics $\{\omega_m\}$ are derived from the pitch period L. Amplitudes $\{A_m\}$ are given by sampling the function (2) at frequencies $\{\omega_m\}$. Phases $\{\varphi_m\}$ of voiced speech given by sampling the Hilbert transform of the logarithm of (2) correspond to the impulse excitation of the minimum-phase model. Randomized phases of unvoiced speech correspond to the noise excitation. The synthetic speech during one pitch period is given by

$$s(l) = \sum_{m=1}^{M} A_m \cos\left(\omega_m l + \varphi_m\right), \qquad 0 \le l \le L. \qquad (3)$$

After overlap-adding (OLA) pairs of frames the synthesized signal is postemphasized to get resulting synthetic speech.

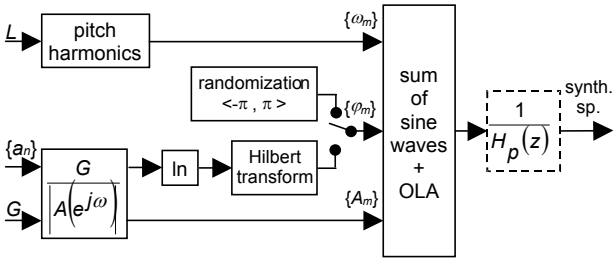

Fig. 5. Postemphasis in the AR harmonic speech synthesis.

6. Experimental Results

Speech material consisted of about 500 speech frames of vowels and nasals for each of three voices: male, female, and childish, sampled at 8 kHz and 16 kHz. Three AR orders for each method were examined: the minimum AR order, twice the minimum, and three times the minimum when preemphasis was used, the same orders increased by one when no preemphasis was used. The RMS log spectral measure [11] determined the error or difference between the smoothed spectra of original and resynthesis. Mean and standard deviation of this measure were evaluated.

For the male voice with the mean pitch frequency of about 110 Hz the analysis speech frame had the duration of 24 ms.

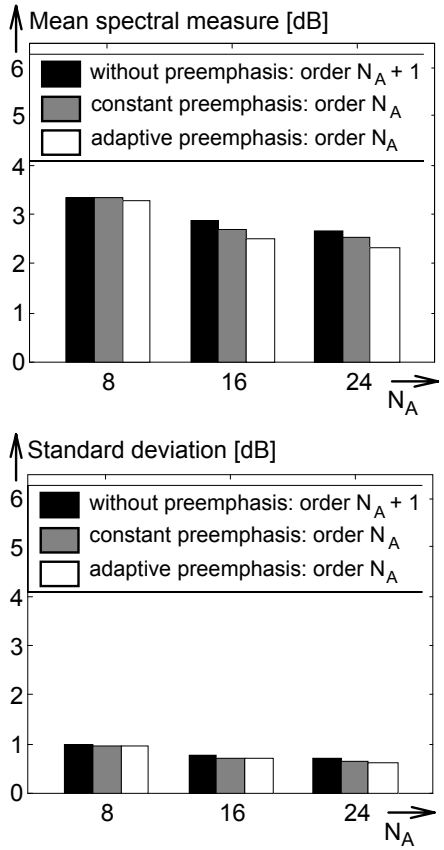

Fig. 6. RMS log spectral measure for the male voice sampled at 8 kHz.

Results for 8-kHz sampling are shown in Fig. 6. For each AR order, the lowest mean spectral measure as well as its standard deviation correspond to adaptive preemphasis. The greatest difference of 0.41 dB is between the mean spectral measure of the 17th AR order without preemphasis and the 16th AR order with adaptive preemphasis.

Results for 16-kHz sampling shown in Fig. 7 give the similar trend. However, the greatest difference is only 0.19 dB between the mean spectral measure of the 49th AR order without preemphasis and the 48th AR order with adaptive preemphasis.

For the female voice with the mean pitch frequency of about 200 Hz the duration of the analysis frame was 16 ms.

As shown in Fig. 8, for 8-kHz sampling, the model with constant preemphasis gives better frequency properties than the higher order model without preemphasis, and the model with adaptive preemphasis gives better frequency properties than the model with constant preemphasis. The greatest difference of 0.46 dB is between the mean spectral measure of the 9th order without preemphasis and the 8th AR order with adaptive pre-emphasis.

For the female voice sampled at 16 kHz (Fig. 9) the results are not so unambiguous as the results described so far (Figs. 6-8). The same trend of decreasing the mean spectral measure can be observed only for the 17th order without preemphasis, and the 16th order with preemphasis, with the greatest difference of 0.28 dB.

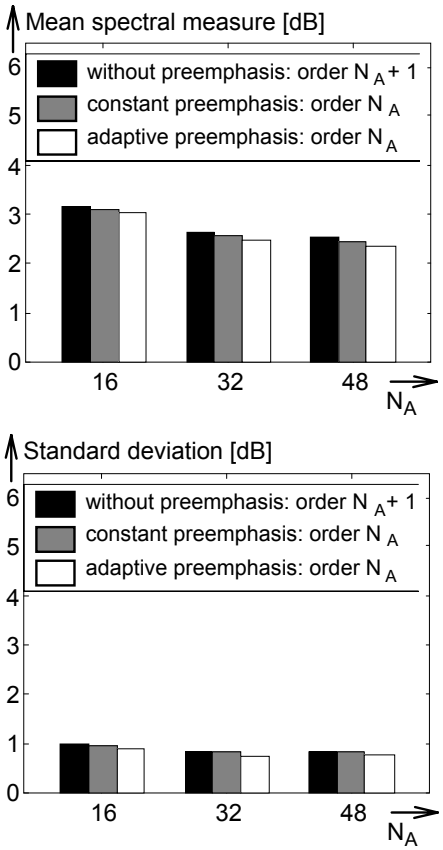

Fig. 7. RMS log spectral measure for the male voice sampled at 16 kHz.

For the childish voice with the mean pitch frequency of about 300 Hz the analysis speech frame of 10 ms was used.

For 8-kHz sampling of the childish voice, the model with adaptive preemphasis gives better results than the model without preemphasis for the lowest and the highest examined orders. For the 16th AR order the adaptive preemphasis is worse than no preemphasis. The mean spectral measure for the 16th AR order with adaptive preemphasis is higher by 0.16 dB than that for the 17th order without preemphasis.

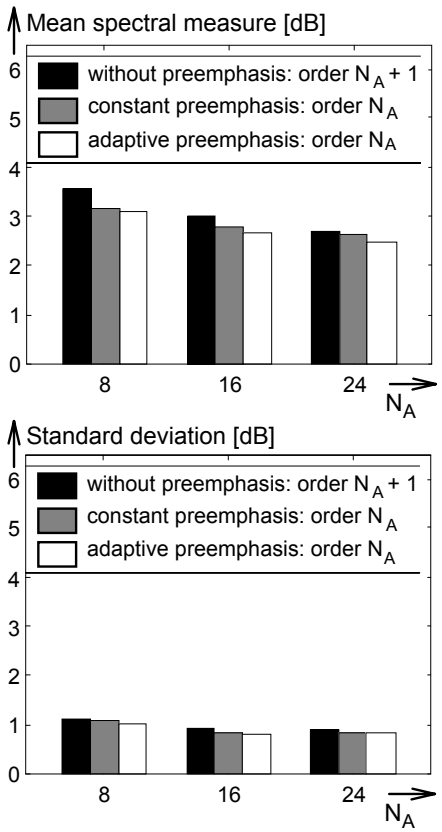

Fig. 8. RMS log spectral measure for the female voice sampled at 8 kHz.

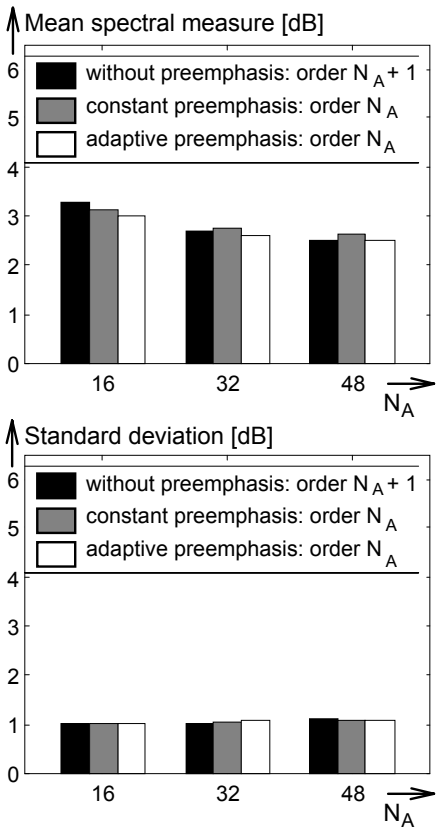

Fig. 9. RMS log spectral measure for the female voice sampled at 16 kHz.

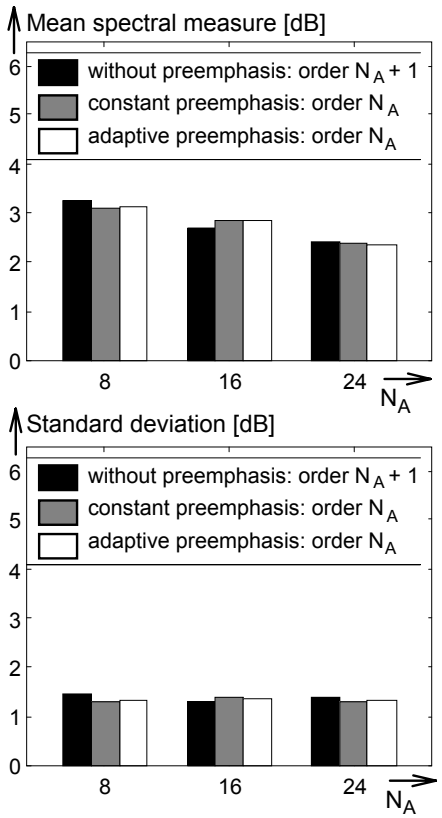

Fig. 10. RMS log spectral measure for the childish voice sampled at 8 kHz.

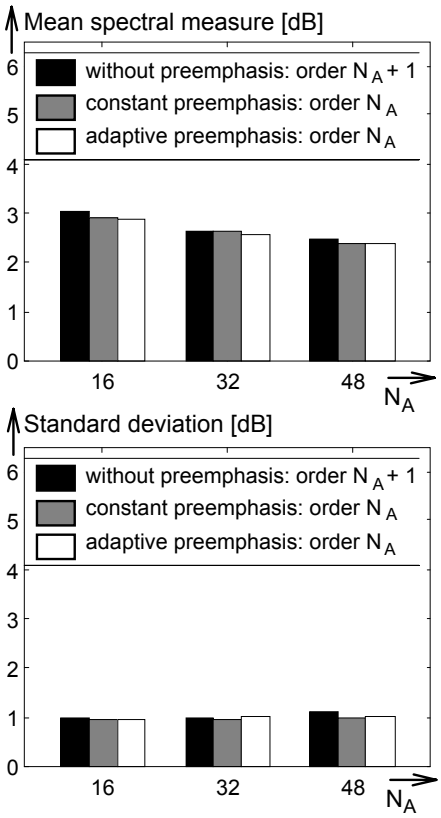

Fig. 11. RMS log spectral measure for the childish voice sampled at 16 kHz.

The mean spectral measure for the childish voice sampled at 16 kHz (Fig. 11) exhibits the similar trend as that for the male voice sampled at 16 kHz (Fig. 7). However, differences are slightly lower, although the greatest difference is 0.18 dB (between the 17th order without preemphasis and the 16th order with adaptive preemphasis).

7. Conclusion

Parametric modeling (either source-filter or sinusoidal) of speech signals finds its use in speech analysis and synthesis, speech coding, speech recognition, and speaker verification and identification.

The aim of the paper was to investigate whether use of preemphasis is justified in the harmonic speech model. It tries to fill in the gap in the area of sinusoidal and harmonic speech modeling where no preemphasis has been used so far. From the theoretical point of view, use of preemphasis and postemphasis was compared for the source-filter model and the harmonic model, both with AR parametrization. It has been shown that the AR model without preemphasis should have the order higher by the order of the preemphasis so that comparison of preemphasized and non-preemphasized speech processing would be relevant. Constant and adaptive preemphasis has also been compared.

Experiments have shown that in majority cases the preemphasis improves speech synthesis quality, and the adaptive preemphasis improves it even more. The results are mostly marked for the male voice processing. For female and childish voices, the constant preemphasis gives sometimes worse results than a higher-order model without preemphasis. However, the adaptive preemphasis is almost always the best solution in female and childish voice processing, too.

Although the harmonic speech model was used to examine influence of preemphasis, the results could also be valid for the source-filter model. Better representation of the male speech spectrum is due to higher number of composite sine waves of the harmonic model. Although the synthetic speech production is different in the source-filter model, convolution of the impulse train of lower frequency with the vocal tract transfer function results in higher number of spectral peaks corresponding to the harmonics of the pitch frequency. Thereby, the source-filter model should represent the speech spectrum with similar quality as the harmonic model. From these facts it can be concluded that the adaptive preemphasis should be preferred in the harmonic speech modeling, and in the source-filter speech modeling, as well.

Acknowledgement

The work was supported by the Ministry of Education of the Slovak Republic under the Grant N^{o} 1/0144/03 and 102/VTP/2000.

References

[1] MARKEL, J. D., GRAY, A. H. A linear prediction vocoder simulation based upon the autocorrelation method. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1974, vol. ASSP-22, no. 2, p. 124 – 134.

[2] GRAY, A. H., MARKEL, J. D. A spectral–flatness measure for studying the autocorrelation method of linear prediction of speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1974, vol. ASSP-22, no. 3, p. 207 – 217.

[3] McAULAY, R. J., QUATIERI, T. F. Sinusoidal coding. In *Speech Coding and Synthesis* (Kleijn, W. B., Paliwal, K. K., Eds.). Elsevier, 1995, p. 121 – 173.

[4] QUATIERI, T., F., McAULAY, R., J. Audio signal processing based on sinusoidal analysis/synthesis. In *Applications of Digital Signal Processing to Audio and Acoustics* (Kahrs, M., Brandenburg, K., Eds.). Kluwer Academic Publishers, 2001, p. 343 - 416.

[5] AHMADI, S., SPANIAS, A., S. A new phase model for sinusoidal transform coding of speech. *IEEE Transactions on Speech and Audio Processing*, 1998, vol. 6, no. 5, p. 495 – 501.

[6] STYLIANOU, Y. Concatenative speech synthesis using a harmonic plus noise model. *Third ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, B. Mountains (Australia), 1998, p. 261 – 266.

[7] CRESPO, M., Á., R., et al. On the use of a sinusoidal model for speech synthesis in text-to-speech. In: *Progress in Speech Synthesis* (van Santen, J., P., H. et al., Eds.). New York: Springer-Verlag, 1997, p. 57 - 70.

[8] LI, C. *Analysis-by-synthesis multimode harmonic speech coding at low bit rate*. PhD Thesis. Santa Barbara (USA): University of California, 2000.

[9] AHMADI, S., SPANIAS, A. S. Low bit-rate speech coding based on an improved sinusoidal model. *Speech Communication*, 2001, vol. 34, no. 4, p. 369 - 390.

[10] SKOGLUND, J. Analysis and quantization of glottal pulse shapes. *Speech Communication*. 1998, vol. 24, no. 2, p. 133 - 152.

[11] GRAY, A. H., MARKEL, J. D. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1976, vol. ASSP-24, no. 5, p. 380 - 391.

[12] MADLOVÁ, A. Comparison of spectral measure and listening tests results. *Measurement of Speech and Audio Quality in Networks*. International On-line Workshop. 2002, p. 37 – 40. (http://wireless.feld.cvut.cz/mesaqin2002/full07.pdf).

[13] MADLOVÁ, A. Autoregressive and cepstral parametrization in harmonic speech modelling. *Journal of Electrical Engineering*. 2002, vol. 53, no. 1-2, p. 46 – 49.

[14] MADLOVÁ, A. *Some parametric methods of speech processing*. PhD Thesis. Bratislava (Slovakia): Slovak University of Technology, 2001.

[15] HESS, W. *Pitch determination of speech signals. Algorithms and devices*. Berlin: Springer-Verlag, 1983.

[16] GARDNER, W. R., RAO, B. D. Noncausal all-pole modeling of voiced speech. *IEEE Transactions on Speech and Audio Processing*. 1997, vol. 5, no. 1, p. 1 – 10.

About Authors...

Anna PŘIBILOVÁ (**MADLOVÁ**) was born in Hlohovec in 1962. She received her Ing (MSc) degree in radio (medical) electronics in 1985 and her PhD degree in electronics in 2002 from the Slovak University of Technology (SUT). 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 Dept. of Radioelectronics, SUT in Bratislava.