ENHANCEMENT OF REVERBERANT SPEECH USING THE CELP POSTFILTER

Marco Jeub and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{jeub,vary}@ind.rwth-aachen.de

ABSTRACT

In this paper we investigate the application of adaptive postfiltering for the enhancement of reverberant speech. The considered method is commonly used in Code Excited Linear Prediction (CELP) speech coding to lower the impact of quantization noise in the excitation signal and the spectral envelope. We show that the underlying additive noise model is accurate enough to enhance speech which is recorded in an enclosed space where the resulting early reflections are usually modeled as a convolutive distortion. By means of adaptive filtering, the amplitudes of the unwanted peaks in the excitation signal are attenuated and the signal components at the harmonic peaks are emphasized. Both, single- and multi-channel dereverberation algorithms are proposed having a moderate computational complexity. Experiments have shown that this approach is capable of reducing early reverberation and attenuate the 'distance-effect' arising from room reflections.

Index Terms— Dereverberation, speech enhancement, CELP, adaptive postfilter, linear prediction

1. INTRODUCTION

Over the last years many algorithms for speech dereverberation based on a model of speech production have been proposed, cf. [1, 2, 3, 4]. They are based on a simplified model of speech production consisting of an excitation source and a vocal tract filter and use the linear prediction (LP) technique to estimate such parameters.

It is well-known that reflections in an enclosed space mainly affect the excitation signal in terms of spurious peaks, cf. [5]. These can, especially in voiced speech where the excitation signal is a periodic pulse-train, degrade the speech quality significantly. The reverberant linear prediction coefficients, which represent the spectral envelope of the speech signal, are approximately equivalent to those of anechoic speech in a spatially averaged sense [6]. Therefore, it has been found beneficial to manipulate mainly the LP residual or excitation signal.

Early studies in [2] apply an adaptive weight function to the LP residual. This should emphasize regions with a high signal-to-reverberation ratio (SRR) and attenuate low SRR regions. A different approach in [3] exploits the kurtosis of the LP residual, which is an indicator for the peakedness. While it has a more Gaussian distribution (smaller kurtosis) in reverberant environments, the kurtosis becomes larger with decreasing reverberation times. An adaptive filter is designed to maximize the kurtosis and hence, to minimize the effect of reverberation. A further method to reduce these peaks is to average the residual signal between consecutive cycles of opening and closing of the glottis (larynx cycle) while excluding the segments around the glottal closure instances (GCI) [4]. A GCI estimation is performed and the LP residual is multiplied with a cosine window having the length of one larynx cycle. Afterwards an averaging over the nearest neighboring cycles is carried out.

In this paper we follow the idea of a source-model based dereverberation from the perspective of speech coding. Commonly used CELP codecs analyze speech frames and extract parameters for the spectral envelope and the excitation signal. These parameters are vector-quantized and used in the decoder for speech reconstruction. Especially at low bitrates, the effect of additive quantization noise can degrade the speech intelligibility. The general idea of a postfilter is to reduce these effects after the decoding process, cf. [7, 8]. We take advantage of this concept and apply it for the enhancement of reverberant speech. However, this model is quite coarse, because reverberation is usually assumed as a convolutive distortion for early reflections and only as additive noise for late reflections, when the sound becomes diffuse.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the concept of adaptive postfiltering. Section 3 discusses the proposed single- and multichannel algorithms and in Section 4 the experiments and results are presented. Finally, in Section 5 we draw conclusions.

2. ADAPTIVE POSTFILTERING

The main objective of adaptive postfiltering in speech coding is the reduction of effects due to quantization noise. This is usually done by means of a cascade of separate filters for the spectral envelope and the spectral fine structure. The overall

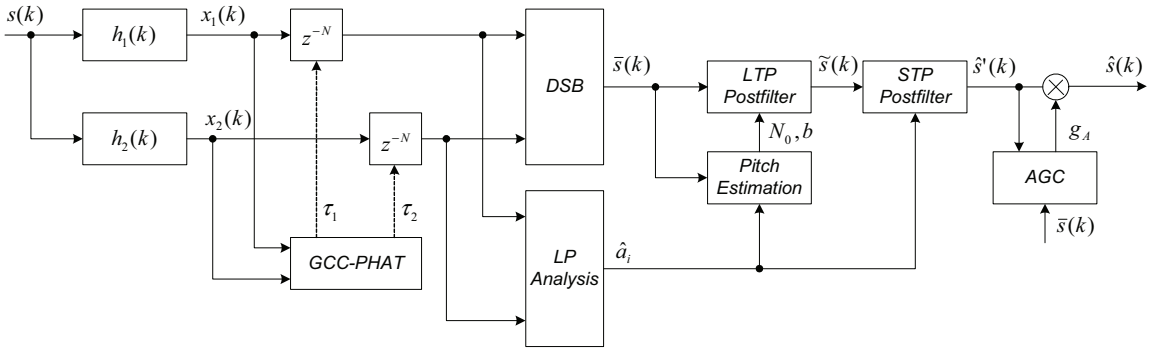

Fig. 1. Block diagram of the proposed algorithm for the dual-channel case ($M = 2$)

form introduced in [8] is given by

$$H(z) = g_P \cdot H_{LTP}(z) \cdot H_{STP}(z) \cdot H_T(z) \tag{1}$$

with a gain factor g_P, the long-term and short-term postfilters $H_{LTP}(z)$ and $H_{STP}(z)$ and a tilt correction filter $H_T(z)$.

The aim of the long-term postfilter (LTP) is the amplification of peaks which are associated with the fundamental frequency and its harmonics. Therefore this filter should only be active in voiced speech having a pulse-like excitation signal. The filter uses an estimation of the pitch period N_0, the pitch gain b ($|b| \leq 1$) and two constants, λ_1 and λ_2 as follows

$$H_{LTP}(z) = \frac{1 + \epsilon \cdot z^{-N_0}}{1 - \eta \cdot z^{-N_0}} \tag{2}$$

where $\epsilon = \lambda_1 \cdot b$ and $\eta = \lambda_2 \cdot b$ for voiced and $\epsilon = \eta = 0$ for unvoiced speech. In order to ensure power equality after the filtering, the following scaling factor is utilized

$$g_P = \frac{1 - \eta/b}{1 + \epsilon/b}. \tag{3}$$

In contrast to the LTP postfilter, the short-term postfilter (STP) has an effect on the spectral envelope where it emphasizes the formant frequencies. The transfer function is given by

$$H_{STP}(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{4}$$

with the prediction analysis filter

$$A(z/\gamma_m) = 1 + \sum_{i=1}^{n} a_i z^{-i} \gamma_m^i, \tag{5}$$

constants γ_m ($m = 1, 2$) and the prediction order n. The drawback of this approach is an unwanted high-pass effect which can be controlled using the additional tilt correction filter

$$H_T(z) = 1 - \mu \cdot z^{-1}. \tag{6}$$

An adaptive gain control (AGC) is used to compensate for gain differences between the unfiltered and the post-filtered signal. The final enhanced speech signal $\hat{s}(k)$ is obtained by multiplication with

$$g_A = \sqrt{\frac{\sum_{k=1}^{N} \bar{s}^2(k)}{\sum_{k=1}^{N} \hat{s}'^2(k)}}, \tag{7}$$

where $\bar{s}(k)$ is the unfiltered speech, $\hat{s}'(k)$ the filtered (and unscaled) speech and N the frame length. The parameters λ_m and γ_m are usually fixed and have to be chosen appropriately. μ can be constant or dependent on the first reflection coefficient of the truncated transfer function $H_{STP}(z)$.

3. PROPOSED ALGORITHM

3.1. Multi-channel speech enhancement

As the most simple multi-channel system we consider the dual-channel case ($M = 2$) as depicted in Figure 1 in the following. The derivation of the special case $M = 1$ or the general case $M > 2$ is straightforward. We assume the anechoic speech signal $s(k)$ to be convolved with the room impulse responses (RIR) $h_j(k)$, where M stands for the total number of microphones and $j = 1, .., M$. The reverberant signals are

$$x_j(k) = s(k) * h_j(k). \tag{8}$$

First it is necessary to perform a temporal alignment of the M channels (Delay elements). The time delay of arrival (TDOA) is calculated with the well-established generalized cross-correlation with phase transform (GCC-PHAT) algorithm, cf. [9, Chapter 6]. Afterwards, the autocorrelation approach is used for a joint LP analysis. We compute the short-term autocorrelation vectors $\varphi_{xx}^{(j)}$ and the corresponding correlation matrices $R_{xx}^{(j)} = \varphi_{xx}^{(j)} \cdot (\varphi_{xx}^{(j)})^T$ individually for each channel, take the spatial average over M microphones and use the Levinson-Durbin algorithm to solve the following equation

$$\hat{a}_i = \hat{R}_{xx}^{-1} \hat{\varphi}_{xx} \tag{9}$$

	Room 1 ($T_{60} \approx$ 440ms, $d =$ 1m)			Room 2 ($T_{60} \approx$ 640ms, $d =$ 3m)		
	ΔSRR_{seg} [dB]	$\Delta PESQ$	$\Delta Kurt_{seg}$	ΔSRR_{seg} [dB]	$\Delta PESQ$	$\Delta Kurt_{seg}$
Single-channel with postfilter	+1.70	+0.08	+1.61	+1.18	+0.05	+1.35
DSB without postfilter	+1.87	+0.22	+1.96	+1.63	+0.26	+2.47
DSB with postfilter (a)	+2.17	+0.27	+3.65	+1.72	+0.30	+4.08
DSB with postfilter (b)	+2.21	+0.26	+3.08	+1.76	+0.27	+3.37

Table 1. Evaluation of the proposed dereverberation algorithms for one ($M = 1$) and eight microphones ($M = 8$). The table entries give the difference to the reverberant speech (plus indicates improvement). The postfilter parameters for the multi-channel case are derived (a) from the DSB signal and (b) from M microphones.

with the spatial averages

$$\hat{R}_{xx} = \frac{1}{M} \sum_{j=1}^{M} R_{xx}^{(j)}, \tag{10}$$

$$\hat{\varphi}_{xx} = \frac{1}{M} \sum_{j=1}^{M} \hat{\varphi}_{xx}^{(j)} \tag{11}$$

and $i = 1, .., n$ where n represents the prediction order. The next step comprises a delay-and-sum beamforming (DSB) of the input samples $x_j(k)$ taking the time delays τ_j into account

$$\overline{s}(k) = \frac{1}{M} \sum_{j=1}^{M} x_j(k - \tau_j). \tag{12}$$

Out of the averaged LP coefficients and the DSB signal an estimation of the pitch period N_0 and the pitch gain b is estimated by the correlation method proposed in [10], which has been found to be robust in reverberant environments. This comprises the computation of the LP residual signal followed by autocorrelation estimation. Based on the pitch and the speech signal, a prediction gain can be computed (see [10] for details). The estimated pitch gain b is used for a voiced/unvoiced classification for each frame as follows

$$voiced = \begin{cases} 1 & \text{for} \quad b > \lambda_3 \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

with the threshold λ_3. Finally, the signal $\overline{s}(k)$ is passed through the postfilter given by Eq.1 and the adaptive gain control. As mentioned above, the LTP filter influences only the residual signal for voiced speech frames. Therefore the parameters ϵ and η are set to zero for $voiced \neq 1$. The overall estimation of $s(k)$ is given by $\hat{s}(k)$.

3.2. Single-channel speech enhancement

The single-channel case is a reduction of the above presented dereverberation algorithm to $M = 1$ channels. Regarding the diagram in Figure 1, the time-alignment as well as the beamforming block can be omitted because $\overline{s}(k) = x_1(k)$. Furthermore the LP analysis block simplifies to a single-channel computation without any spatial averaging.

4. EXPERIMENTS AND RESULTS

For the experiments we employ simulations using speech files from the TIMIT database convolved with two different RIRs from the MARDY database [11]. The first RIR was recorded in an enclosed space having a reverberation time of $T_{60} \approx$ 440ms at a source-microphone distance of $d =$ 1m and the second one is recorded in the same room at $d =$ 3m resulting in $T_{60} \approx$ 640ms. The RIRs for the multi-channel algorithms represent an horizontal circular array with eight microphones at adjacent distances of $d_M = 0.05$m. The constants for the postfilter are heuristically determined and stated in Table 2.

Parameter	γ_1	γ_2	μ	λ_1	λ_2	λ_3
Value	0.8	0.5	0.6	0.5	0.0	0.4

Table 2. Chosen postfilter constants

Some further parameters are: 20ms frames, 50% overlap, Hann window, prediction order $n = 16$ and $f_s =$ 16kHz. Dereverberation is performed in four different ways. The single-channel algorithm uses the described postfilter and one microphone. For the multi-channel case we apply spatial averaging of the time-aligned multiple channels and perform the postfilter. The parameters are derived (a) from the DSB signal $\overline{s}(k)$ and (b) from the time-aligned multiple channels $x_j(k - \tau_j)$. The performance is evaluated with three different measurements and informal listening tests. The segmental signal-to-reverberation ratio indicates the reduction of reverberation compared to the direct path $s_d(k)$ of the anechoic signal ($s_d(k) = s(k) * h_d(k)$, where $h_d(k)$ is the direct part of the RIR). The SRR_{seg} is calculated for every frame l with speech activity and length N by

$$\frac{\overline{SRR}_{seg}(l)}{dB} = 10 \cdot lg \left(\frac{\sum_{k=1}^{N} s_d(k)^2}{\sum_{k=1}^{N} (s_d(k) - \hat{s}(k))^2} \right) \tag{14}$$

and averaging over K considered frames to obtain

$$\frac{SRR_{seg}}{dB} = \frac{1}{K} \sum_{l=1}^{K} \overline{SRR}_{seg}(l). \tag{15}$$

For the reduction of the coloration has, so far, has no adequate measurement been established. The commonly used spectral

distance measure is not appropriate because it is mainly affected by variations in the spectral envelope and cannot identify improvements made in the residual. In order to evaluate the attenuation of the unwanted peaks we employ the segmental kurtosis defined as having a value of 0 for the normal distribution:

$$Kurt_{seg} = \frac{1}{K}\sum_{l=1}^{K}\frac{E\left\{d_l(k)^4\right\}}{E\left\{d_l(k)^2\right\}^2} - 3, \qquad (16)$$

where $d_l(k)$ indicates the residual signal of the lth frame and $E\{\cdot\}$ the expectation operator. Both segmental measures are performed on frames with 20ms and 50% overlap. Speech activity is determined in the anechoic speech with the voice activity detector (VAD) of the AMR-WB speech codec [12]. Additionally, we employ the perceptual evaluation of speech quality (PESQ) score [13] which is widely used for the evaluation of speech codecs and shows a good performance for the assessment of postfilters as well, cf. [14]. The results of the enhancement over 6300 reverberated speech files ($t \approx 5.4$h) are shown in Table 1. It can be seen that all algorithms are capable of reducing the effect of reverberation. In case of the multi-channel methods, the postfilter derived from the DSB signal gives similar results to the one derived from multiple channels in terms of SRR_{seg}. The kurtosis and PESQ values are slightly better if the postfilter parameters are derived from the DSB signal. Due to a performed informal listening test we conclude that the effect of early reverberation has been reduced and the speech sounds more 'near' without audible distortions. A first comparison to existing source-model algorithms showed that the speech distortions produced by the postfilter are less audible compared to [2] and [4].

5. CONCLUSIONS

In this paper we propose and evaluate the application of adaptive postfilter known from speech coding for the purpose of speech dereverberation. Although in speech coding the quantization noise is additive, these methods are adequate to reduce the effect which room reverberation has on the residual signal and the spectral envelope. Taking advantage of this concept, we propose a single- and multi-channel algorithm having a moderate computational complexity which may be combined with a speech encoder. The experiments show that this approach is capable of enhancing reverberant speech while avoiding disturbing artifacts. This approach is suitable for the reduction of early reverberation and make the speech sound more 'near'.

6. REFERENCES

[1] S. Griebel and M. Brandstein, "Wavelet transform extrema clustering for multi-channel speech dereverberation," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Pocono Manor, USA, 1999.

[2] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 3, pp. 267–281, May 2000.

[3] B. W. Gillespie, H. S. Malvar, and D. A. F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Acoustics, Speech, and Signal Proc., IEEE Int. Conf. on*, Salt Lake City, USA, 2001, vol. 6, pp. 3701–3704.

[4] M.R. P. Thomas, N.D. Gaubitch, J. Gudnason, and P.A. Naylor, "A practical multichannel dereverberation algorithm using multichannel dypsa and spatiotemporal averaging," in *Proc. IEEE Workshop on App. of Signal Proc. to Audio and Acoustics*, 2007, pp. 50–53.

[5] N.D. Gaubitch, P.A. Naylor, and D.B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 2003.

[6] N.D. Gaubitch, B. Ward Darren, and P.A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. Am.*, vol. 120(6), 2006.

[7] P. Vary and R. Martin, *Digital Speech Transmission. Enhancement, Coding and Error Concealment*, Wiley&Sons, Chichester, 2006.

[8] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 59–71, 1995.

[9] R. Martin, U. Heute, and C. Antweiler, Eds., *Advances in Digital Speech Transmission*, Wiley&Sons, Chichester, 2008.

[10] P. Ojala, P. Haavisto, A. Lakaniemi, and J. Vainio, "A novel pitch-lag search method using adaptive weighting and median filtering," in *Proc. Speech Coding, IEEE Workshop on*, Porvoo, Finland, 1999, pp. 114–116.

[11] J.Y.C. Wen, N.D. Gaubitch, E.A.P. Habets, T. Myatt, and P.A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, 2006.

[12] TS 26.194, *Adaptive Multi-Rate - Wideband speech codec, Voice Activity Detector*, V6.0.0, 3GPP, 2004.

[13] ITU-T Rec. P.862, *Perceptual evaluation of speech quality (PESQ)*, ITU, Geneva, 2001.

[14] V. Grancharov, J. H. Plasberg, J. Samuelsson, and B. W. Kleijn, "Generalized postfilter for speech quality enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 57–64, 2008.