DEREVERBERATION OF SPEECH SIGNALS BASED ON THE DISCRETE MODEL OF SPEECH PRODUCTION

Marco Jeub and Peter Vary

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

Abstract: Over the last years several algorithms for the dereverberation of speech signals based on the discrete model of speech production have been proposed. They have in common that they rely on a model consisting of an excitation source and a time-varying vocal tract filter. In this paper we investigate the application of the postfilter algorithm used in Code Excited Linear Prediction (CELP) speech coding for the purpose of speech dereverberation. We show that in case of a reverberant signal, the amplitudes of the unwanted peaks in the excitation signal are attenuated and that this approach is capable of reducing early reverberation. Furthermore, we compare the new approach with state-of-the-art speech source-model dereverberation algorithms and evaluate the proposed method in a single-channel approach.

1 Introduction

In speech communication systems with a hands-free operation mode the signal is often degraded by reverberation caused by the room characteristics. This applies for telephone systems as well as for hearing aids. Over the last years, several algorithms for speech dereverberation based on a discrete model of speech production have been proposed. They are based on a simplified model consisting of an excitation source and a time-varying vocal tract filter. The corresponding model parameters are estimated by means of Linear Prediction (LP) techniques. In order to reduce the effect of room reverberation, the spectral envelope as well as the excitation signal can be modified.

It is well-known that reflections in an enclosure mainly affect the excitation signal in terms of spurious peaks due to the multipath reflections [1]. These can degrade the speech quality significantly, especially in voiced speech segments where the model excitation signal is a periodic pulse-train. Therefore state-of-the-art dereverberation algorithms, which are based on the source-filter model, mainly modify the LP residual or excitation signal.

In the following, we assume the anechoic speech signal $s(k)$ to be convolved with a room impulse response (RIR) $h(k)$

$$x(k) = s(k) * h(k), \tag{1}$$

with the reverberant signal $x(k)$. The general block diagram of a source-model dereverberation algorithm is depicted in Figure 1.

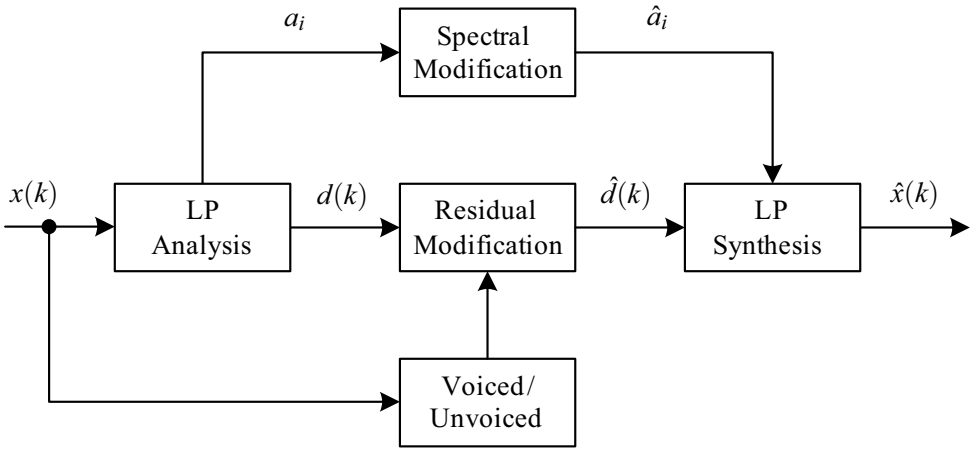

Figure 1 - General block diagram of speech source-model based dereverberation algorithms.

From the reverberant signal $x(k)$, the vocal tract filter $A(z)$ is computed through linear prediction analysis

$$A(z) = \sum_{i=1}^{n} a_i \cdot z^{-i}, \tag{2}$$

where a_i represent the LP coefficients and n the prediction order. The excitation or residual signal $d(k)$ is calculated by filtering the speech signal as follows

$$d(k) = x(k) - \sum_{i=1}^{n} a_i \cdot x(k-i). \tag{3}$$

The residual signal $d(k)$ can then be modified to obtain an estimate $\hat{d}(k)$ of the anechoic residual signal. What follows is the synthesis with the inverse analysis filter to obtain an enhanced signal $\hat{x}(k)$. Besides the modification of the excitation signal, the parameters of the spectral envelope a_i can be modified as well. Additionally, the algorithm requires a classification whether the current speech frame is either voiced or unvoiced.

In this paper we give an overview of source-model based dereverberation algorithms and present our algorithm which is based on the adaptive CELP postfilter [2]. Then, we provide a comparison to state-of-the-art single-channel dereverberation algorithms based on the above mentioned concept. The evaluation is performed in reverberant environments with different acoustical properties using real-measured room impulse responses.

The remainder of this paper is organized as follows. In Section 2, we give an overview of source-model dereverberation algorithms. Section 3 follows [2] and presents a dereverberation method based on the CELP postfilter. In Section 4 the experiments and results are presented. Finally, in Section 5 we draw conclusions.

2 Source-Model based Dereverberation

In this section we give a brief overview into dereverberation methods based on the discrete model of speech production. They all have in common that they mainly modify the LP residual signal in order to enhance the degraded speech signal. Early studies in [3] apply an adaptive time domain weighting 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 [4] 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 unwanted peaks in the LP residual 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) [5]. Estimation of GCI is performed and the LP residual is multiplied in the time domain with a cosine window having the length of one larynx cycle. Afterwards, averaging over the nearest neighboring cycles is carried out. This technique is performed on voiced speech only where a pulse-like excitation is assumed. Since it leaves unvoiced speech unaffected, a further improvement was presented e.g. in [6]. An equalization filter is applied to perform the equivalent operation of temporal averaging on unvoiced and silence speech.

In [7] a wavelet clustering algorithm is applied to the LP residual. The basic idea is to cluster the multiple channel signals according to their wavelet extrema to obtain a single residual signal. This is resynthesized to obtain a dereverberated signal. The same authors suggest in [8] to perform a rough estimate of the room impulse response for each channel in a multi-channel approach. A weight function is computed for each channel by applying a matched filter type operation. Each weighted residual signal is then aligned and added. Investigations of the influence of room reverberation on the spectral envelope, especially for multi-channel algorithms, can be found e.g. in [1].

3 CELP Postfilter

In this chapter we follow the idea of a source-model based dereverberation from the perspective of speech coding [2]. 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 quality. The general idea of a postfilter is to reduce these effects after the decoding process. We take advantage of this concept and apply it for the enhancement of reverberant speech.

The main objective of adaptive postfiltering in speech coding is the reduction of subjective 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 form introduced in [9] is given by

$$H(z) = g_{\mathrm{P}} \cdot H_{\mathrm{LTPF}}(z) \cdot H_{\mathrm{STPF}}(z) \cdot H_{\mathrm{T}}(z) \tag{4}$$

with a gain factor g_{P}, the long-term and short-term postfilters $H_{\mathrm{LTPF}}(z)$ and $H_{\mathrm{STPF}}(z)$ and a tilt correction filter $H_{\mathrm{T}}(z)$.

The aim of the long-term postfilter (LTPF) 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

$$H_{\mathrm{LTPF}}(z) = \frac{1 + \varepsilon \cdot z^{-N_0}}{1 - \eta \cdot z^{-N_0}} \tag{5}$$

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

$$g_{\mathrm{P}} = \frac{1 - \eta/b}{1 + \varepsilon/b}. \tag{6}$$

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

$$H_{\mathrm{STPF}}(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \tag{7}$$

with the prediction analysis filter

$$A(z/\gamma_m) = 1 + \sum_{i=1}^{n} a_i z^{-i} \gamma_m^i, \tag{8}$$

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 by using the additional tilt correction filter

$$H_{\mathrm{T}}(z) = 1 - \mu \cdot z^{-1}. \tag{9}$$

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 $\hat{x}(k)$ is obtained by multiplication with

$$g_{\mathrm{A}} = \sqrt{\frac{\sum_{k=1}^{N} \bar{x}^2(k)}{\sum_{k=1}^{N} \tilde{x}^2(k)}}, \tag{10}$$

where $\bar{x}(k)$ is the unfiltered speech, $\tilde{x}(k)$ 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 $H_{\mathrm{STPF}}(z)$. In this paper we take only single-channel algorithms into consideration and therefore set the number of microphones to $M = 1$. The presented algorithm can be extended to multiple channels as described in [2].

4 Experiments and Results

We employ simulations using speech files from the TIMIT database convolved with different RIRs from the Aachen Impulse Response (AIR) database [10]. The RIRs are measured at different source-microphone distances in the presence of a dummy head. During the measurements by means of Maximum Length Sequences (MLS), the capturing microphone was placed 1cm next to the pinna. This ensures a realistic environment for an application in digital hearing aids. Room types, the corresponding reverberation times RT_{60} and source-microphone distances d_{LM} are as follows

- Low-reverberant room: $\mathrm{RT}_{60} = 0.11\,\mathrm{s}$, $d_{\mathrm{LM}} = 1\,\mathrm{m}$,
- Meeting room: $\mathrm{RT}_{60} = 0.21\,\mathrm{s}$, $d_{\mathrm{LM}} = 1.45\,\mathrm{m}$,
- Office room: $\mathrm{RT}_{60} = 0.37\,\mathrm{s}$, $d_{\mathrm{LM}} = 1\,\mathrm{m}$,
- Lecture room: $\mathrm{RT}_{60} = 0.70\,\mathrm{s}$, $d_{\mathrm{LM}} = 4\,\mathrm{m}$.

For the evaluation we compare the algorithm described in Section 3 (CELP) with the methods by Yegnanarayana (YEG) [3], Gaubitch (GAUB) [6] and Gillespie (GIL) [4]. The simulations are performed on $f_{\mathrm{s}} = 8\,\mathrm{kHz}$ sampled signals. The constants for the postfilter are heuristically determined as in [2]. The selected pitch estimation and voiced/unvoiced classification is based on a weighted autocorrelation approach. All algorithms are performed on 20 ms frames with 50 % overlap and a prediction order $n = 10$. For the GAUB algorithm we perform a weighted

averaging of 5 neighboring residuals. The performance is evaluated with two different objective measurements and informal listening tests. A non-intrusive measurement based on the Speech to Reverberation Modulation energy Ratio (SRMR) [11] is used. It is calculated by means of a gammatone filterbank analysis of temporal envelopes of the speech signal and shows a good correlation with subjective ratings of the overall speech quality and intelligibility. In order to evaluate the attenuation of the unwanted peaks in the LP residual signal, we employ the segmental kurtosis defined as having a value of 0 for the normal distribution

$$\text{SegKurt} = \frac{1}{K}\sum_{l=1}^{K}\frac{\hat{E}\left\{d_l(k)^4\right\}}{\hat{E}\left\{d_l(k)^2\right\}^2} - 3, \tag{11}$$

where $d_l(k)$ indicates the residual signal of the lth frame, $\hat{E}\{\cdot\}$ the short-term expectation operator and K the number of considered frames. Silence periods have been removed before evaluation using the voice activity detector (VAD) of the AMR-WB speech codec. The signal levels are normalized to $-26\,\text{dBov}$ using the ITU-T Rec. P.56 speech voltmeter [12]. The results of the enhancement over 100 reverberated speech files from the TIMIT database are shown in Figure 2. It can be seen that all algorithms are capable of reducing the effect of room

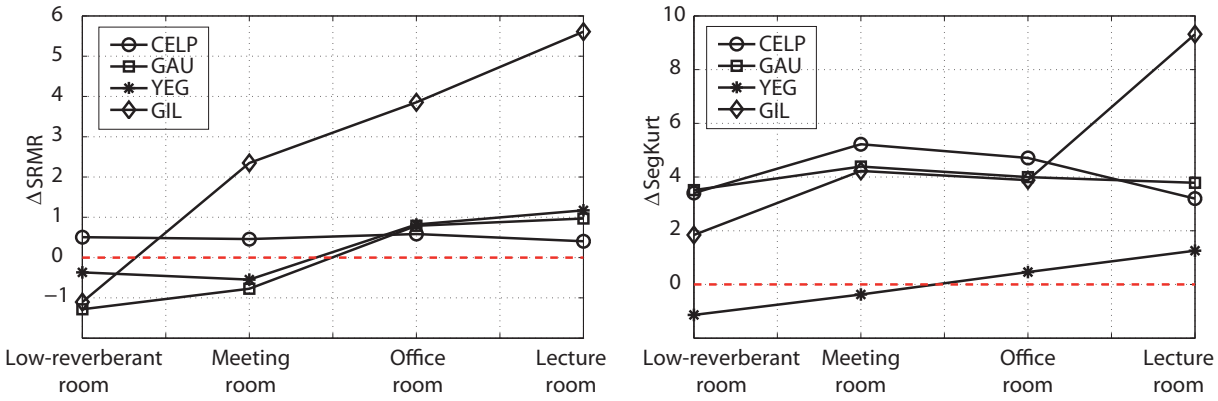

Figure 2 - Evaluation of the proposed dereverberation algorithms using speech files from the TIMIT and room impulse responses from the AIR database. The values give the difference to the reverberant speech (positive values indicate improvement).

reverberation in terms of the objective measurements. However, while some of them show a similar performance in all four tested scenarios, some perform better in rooms with a higher reverberation time. The CELP postfilter has a similar dereverberation performance independent of the reverberation time. The GIL algorithm is greatly dependent on the reverberation time and performs better for highly reverberant environments. In the low-reverberant room, the SRMR value becomes even negative, which is also an indication of speech distortions. A similar dependency can be seen for the YEG and GAU algorithms. In terms of the SRMR measure, the algorithms show improvements only for reverberation times $> 0.3\,\text{s}$. Due to a performed informal listening test we conclude that the effect of reverberation has been reduced and that the speech sounds more near for all four algorithms. The CELP algorithm has a similar performance in all measured rooms without audible distortions. YEG and GAUB also reduce the reverberation effect on the expense of a high amount of speech distortions, especially for lower reverberation times. Among the four tested algorithms, the YEG algorithm gives lowest degree of improvement. The adaptive filter (GIL) gives the best listening impression but also causes a disturbing lowpass-filter effect.

5 Conclusions

In this paper we propose and evaluate an adaptive postfilter as 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 that room reverberation has on the residual signal and the spectral envelope. Taking advantage of this concept, we propose an algorithm having a moderate computational complexity. The experiments show that this approach is capable of enhancing reverberant speech while avoiding disturbing artifacts for different acoustical environments. In comparison to other state-of-the-art speech source-model algorithms, the dereverberation performance is fairly constant in all measured rooms. It outperforms other algorithms, especially for enclosures having a reverberation time of less than 0.7 s. All other algorithms show a very high dependency on the acoustical environment and can, in rooms with moderate reverberation, cause a high amount of speech distortions.

References

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

[2] M. Jeub and P. Vary, "Enhancement of reverberant speech using the CELP postfilter," in *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 3993–3996.

[3] 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.

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

[5] N.D. Gaubitch, P.A. Naylor, and D.B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *Proc. of the European Signal Processing Conference (EUSIPCO)*, Vienna, Austria, 2005, p. 809–812.

[6] 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 (WASPAA)*, 2007, pp. 50–53.

[7] 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.

[8] S. Griebel and M. Brandstein, "Microphone array speech dereverberation using coarse channel modeling," in *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, USA, 2001, vol. 1, pp. 201–204.

[9] 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.

[10] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. Int. Conference on Digital Signal Processing (DSP)*, Santorini, Greece, 2009.

[11] T.H. Falk and W.-Y. Chan, "A non-intrusive quality measure of dereverberated speech," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, USA, 2008.

[12] ITU-T Rec. P.56, *Objective measurement of active speech level*, ITU, 1993.