TEMPORAL SELECTIVE DEREVERBERATION OF NOISY SPEECH USING ONE MICROPHONE

Emanuël A.P. Habets

Bar-Ilan University
School of Engineering
Ramat-Gan, Israel

Nikolay D. Gaubitch, Patrick A. Naylor

Imperial College London
Department of Electrical Engineering
London, UK

ABSTRACT

Reverberant speech can be described as sounding distant with noticeable coloration and echo. These detrimental perceptual effects are caused by early and late reflections, respectively, and reduces the fidelity and intelligibility of speech. It is well-known that the echo density of the reflections increases with time. Therefore, the temporal structure of early and late reflections differs. In this paper, we combine two different dereverberation techniques that were recently developed to suppress early and late reverberation separately. First, late reverberation is suppressed using a spectral processing technique that is based on a statistical reverberation model. Secondly, early reverberation and residual late reverberation are suppressed using a Linear Prediction (LP) residual processing technique. In addition, an objective measure based on the kurtosis of the LP residual is proposed to measure the coloration caused by early reflections. Experimental results demonstrate the beneficial use of the new single microphone system that reduces echo and coloration with little speech distortion.

Index Terms— Acoustic Signal Processing, Speech Dereverberation, Spectral Enhancement, LP Residual Enhancement.

1. INTRODUCTION

In typical speech communication systems, such as hands-free mobile telephones, voice-controlled systems, and hearing aids, the received microphone signal is degraded by room reverberation and background noise. This signal degradation can lead to reduced intelligibility of the speech and decreases the performance of automatic speech recognition systems.

The received microphone signal generally consists of a) a direct sound, b) reflections that arrive shortly after the direct sound (early reverberation), and c) reflections that arrive after the early reverberation (late reverberation). Reverberant speech can be described as sounding distant with noticeable colouration and echo [1]. These detrimental perceptual effects are caused by early and late reverberation, respectively, and generally increase with increasing distance between the source and microphone.

Dereverberation algorithms can be divided into two classes. The classification depends on whether the Room Impulse Responses (RIR) need to be known or estimated. Blind estimation of the RIRs, in a practical scenario, remains an unsolved and challenging problem [2]. Algorithms that do not require an estimate of the RIR are for example based on Linear Prediction (LP) residual processing [3, 4, 5] or spectral processing [6]. While most dereverberation algorithms exploit multiple microphones, many typical speech communication systems are equipped with a single microphone. Since, a smaller number of (practically feasible) single-microphone speech dereverberation algorithms have been proposed, the development of novel single-microphone dereverberation algorithms continues to be an important research topic.

Recently, a practically feasible single-microphone spectral processing technique has been developed that is able to suppress late reverberation and background noise [6]. The algorithm is based on a statistical reverberation model which is characterized by the reverberation time of the room and the Direct to Reverberation Ratio (DRR). In [5] a Spatiotemporal averaging Method for Enhancement of Reverberant Speech (SMERSH) has been developed which is based on processing of the residual signal following linear prediction. The speech signals are first spatially averaged followed by temporal averaging of the LP residual over adjacent larynx cycles of voiced speech. The effect of the inter-cycle averaging in the LP residual domain is embodied into an equalization filter which is subsequently applied to both voiced and unvoiced LP residual, and suppresses the effects of early and late reflections.

In this paper we develop a single-microphone system which consists of two-stages. In the first stage the spectral processing technique proposed in [6] is used to suppress late reverberation. In addition, we show how background noise can be suppressed in this stage. In the second stage, a single-microphone version of SMERSH algorithm is used to suppress early reverberation and residual late reverberation. The performance of the LP residual processing in the presence of background noise is analysed. Furthermore, we show that the kurtosis of the LP residual signal is highly correlated with the standard deviation of the log amplitude spectrum of the RIR, which is a channel-based objective measure for the colouration introduced by the RIR. Using the kurtosis measure we are able to show that the LP residual processing can be used to reduce the colouration caused by the early reflections. This also confirms the study in [3, 4] where an adaptive filter maximizing the kurtosis of the LP residual was used to suppress early reverberation.

The outline of the paper is as follows. In Section 2 the spectral and LP residual processing techniques are reviewed. In Section 3 the system containing the two dereverberation stages is discussed. Experimental results, for different source-microphone and reverberation times, are described in Section 4. Finally, conclusions are drawn in Section 5.

2. SPEECH DEREVERBERATION

In this section we will briefly review the spectral and LP residual processing techniques described in [6] and [5], respectively.

The speech signal $s(n)$ received at the microphone can be written

$$z(n) = s(n) * h(n) + v(n) = x(n) + v(n), \quad (1)$$

where $h(n)$ is the RIR and $v(n)$ denotes additive noise. The reverberant signal $x(n)$ consists of an early reverberation component, $x_e(n)$, and a late reverberation component, $x_r(n)$ such that $x(n) = x_e(n) + x_r(n)$. The goal of the dereverberation system is to obtain an estimate $\hat{s}(n)$ of the anechoic speech signal.

2.1. Spectral Processing

In [6] single- and multi-microphone speech dereverberation algorithms were developed based on a generalized statistical reverberation model. The model is characterized by two parameters. The first parameter is related to the reverberation time, T_{60}, of the room, and the seconds parameter, κ, is related to the DRR of the RIR. Methods for blind estimation of T_{60} and κ are described in [7], and [6], respectively. The parameter κ is important when the source-microphone distance is smaller than the critical distance, which is the distance at which the direct path energy is equal to the energy of all reflections [8].

The noisy and reverberant signal $z(n)$ is first transformed in the time-frequency domain by using the short time Fourier transform (STFT). The time frames are denoted by ℓ, and the discrete frequency bins are denoted by k. In the STFT domain the microphone signal can be written as $Z(\ell,k) = X_e(\ell,k) + X_r(\ell,k) + V(\ell,k)$. The spectral variance $\lambda_r(\ell,k) = \mathcal{E}\{|X_r(\ell,k)|^2\}$ of the late reverberant signal component $x_r(n)$ is then obtained using [6]

$$\lambda_r(\ell,k) = (1-\kappa(k))\,\mathrm{e}^{-2\delta(k)t_r}\lambda_r(\ell-1,k) + \kappa(k)\,\mathrm{e}^{-2\delta(k)t_r}\lambda_x(\ell-\frac{t_r f_s}{R},k), \quad (2)$$

where $\lambda_x(\ell,k) = \mathcal{E}\{|X(\ell,k)|^2\}$ denotes the spectral variance of the reverberant signal[1], R denotes the frame rate of the STFT, f_s denotes the sample frequency, and $\delta(k)$ is related to the (frequency dependent) reverberation time $T_{60}(k)$ through $\delta(k) = 3\ln(10)/T_{60}(k)$. The parameter t_r (in seconds) controls the time instance at which the late reverberation starts and is chosen such that $\frac{t_r f_s}{R}$ is an integer value. Its value usually ranges between 30 and 50 ms. The early spectral speech component $X_e(\ell,k)$ consists of the direct sound and early reverberation and is estimated by applying a time and frequency dependent gain function to $Z(\ell,k)$, i.e.,

$$\hat{X}_e(\ell,k) = G(\ell,k)Z(\ell,k). \quad (3)$$

In this case a modified magnitude subtraction approach is used. The corresponding gain function is given by

$$G(\ell,k) = \max\left\{1 - \frac{1}{\sqrt{\xi(\ell,k)+1}}, G_{\min}\right\}. \quad (4)$$

where $\xi(\ell,k)$ denotes the *a priori* Signal to Interference Ratio (SIR) which is given by

$$\xi(\ell,k) = \frac{|X_e(\ell,k)|^2}{\lambda_r(\ell,k)+\lambda_v(\ell,k)}, \quad (5)$$

and $\lambda_v(\ell,k)$ denotes the spectral variance of the background noise $v(n)$, which can be estimated using the Minima Controlled Recursive Average approach proposed by Cohen [9], and $G_{\min}$ denotes the gain floor. The *a priori* SIR can be estimated using the Decision-Directed approach proposed by Ephraim and Malah [10].

[1]The spectral variance of the reverberant signal can be estimated given an estimate of the spectral variance of the background noise. If the spectral variance of the noisy reverberant signal is used (2) will be biased.

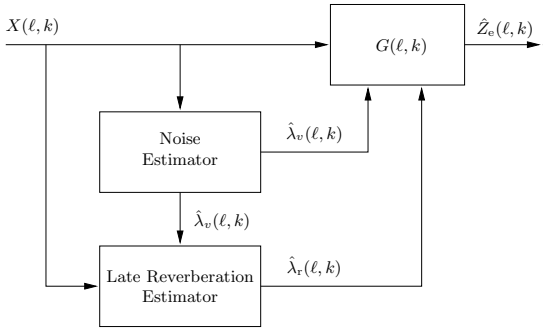

Fig. 1. First stage of the developed system: spectral processing.

The early speech component $x_e(n)$ can then be obtained using the inverse STFT and an overlap-save technique (see [6] and the references therein).

2.2. LP Residual Processing

The reverberant speech signal $x(n)$, can be written in terms of pth order linear predictor as

$$x(n) = -\mathbf{b}^T\mathbf{x}(n-1) + e(n), \quad (6)$$

where $\mathbf{b} = [b_1 b_2 \ldots b_p]^T$ are the LP coefficients, $e(n)$ is the prediction residual signal and $\mathbf{x}(n-1) = [x(n-1)x(n-2)\ldots x(n-p)]^T$. The LP coefficients can be found by minimizing $e(n)$.

When studying the effect of reverberation on the LP residual the following specific observations can be made. Firstly, the LP residual from the reverberant speech differs from that in clean speech by seemingly random peaks; these appear uncorrelated among consecutive larynx-cycles. Secondly, the main features between consecutive larynx-cycles in the clean speech LP residual change slowly and show high inter-cycle correlation. The first property arises from the quasi-periodic nature of voiced excitation and the effect of the RIR. The second property is well-known in speech processing. Motivated by these observations, it is proposed that applying a moving average operation across neighbouring larynx cycles in voiced speech will suppress the uncorrelated features and, hence, enhance the LP residual [5]. There are two issues to consider: Firstly, it is necessary to correctly identify the peaks that belong to the original excitation so as to segment the larynx cycles. Secondly, peaks attributed to Glottal Closure Instants (GCI) are important to the speech quality and should remain unchanged. Hence, they should be excluded from the averaging process.

DYPSA performs automatic GCI identification in speech [11]. At the output of DYPSA we obtain the estimated time, n_ℓ, of the ℓth GCI. The dynamic programming within DYPSA makes it robust to spurious peaks in the prediction residual. This is attractive for GCI identification in reverberant speech since it discriminates many of the erroneous candidates due to reverberation [5].

In order to leave the glottal pulse undisturbed, a weight function is applied on each larynx frame prior to the averaging. In practice, GCIs are identified to an uncertainty in the order of 1 ms [11] and the glottal pulse is not a true impulse but is spread in time [12]. In [5], a weight function was proposed with a reasonable trade-off between the issues described above.

Thus, each enhanced larynx cycle in a voiced speech segment is obtained by averaging the current weighted larynx cycle frame

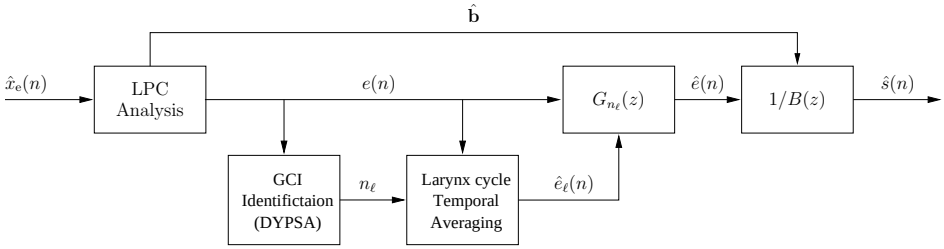

Fig. 2. Second stage of the developed system: LP residual processing.

under consideration with $2\mathcal{I}$ of its neighbouring weighted larynx cycles.The result is then added to the original larynx cycle weighted with the inverse weight function. The final expression for the ℓth enhanced larynx cycle becomes

$$\hat{\mathbf{e}}_\ell = (\mathbf{I} - \mathbf{W})\mathbf{e}_\ell + \frac{1}{2\mathcal{I}+1}\sum_{i=-\mathcal{I}}^{\mathcal{I}} \mathbf{W}\mathbf{e}_{\ell+i}, \tag{7}$$

where $\mathbf{e}_\ell = [e(n_\ell)\ \ e(n_\ell+1)\dots e(n_\ell+\mathcal{L}-1)]^T$ contains the samples of the ℓth larynx-cycle of length $\mathcal{L}$ with its GCI at time n_ℓ, $\hat{\mathbf{e}}_\ell = [\hat{e}(n_\ell)\ \ \hat{e}(n_\ell+1)\dots\hat{e}(n_\ell+\mathcal{L}-1)]^T$ is the ℓth larynx cycle of the enhanced residual, $\mathbf{I}$ is the identity matrix, and $\mathbf{W} = \text{diag}\{w_0\ w_1\dots w_{\mathcal{L}-1}\}$ is a diagonal weighting matrix with coefficients obtained from the time domain Tukey window [5].

Because the above procedure only affects the voiced speech segments and does not take advantage of the of past correct larynx-cycle frames, an L_i-tap FIR filter, denoted by $G_{n_\ell}(z)$, was proposed in [5] which performs the equivalent operation of the inter-cycle averaging. In this way interferences, such as reverberation and background noise, that are uncorrelated among consecutive frames, will be suppressed by the temporal averaging operation. Under the assumption that the time-span of the autocorrelation function of the background noise is smaller than the length of the larynx-cycle frames, the temporal averaging of the SMERSH algorithm will suppress the noise. The maximum amount of suppression is related to the number of frames that is averaged, i.e., $2\mathcal{I}+1$. Therefore, the choice of $\mathcal{I}$ is important: if too many cycles are included the averaging will remove uncorrelated portions for the original excitation; if too few cycles are considered, peaks due to reverberation will remain.

3. SYSTEM DESCRIPTION

In this section we describe the dereverberation system. To be able to extract the optimal LP coefficients (LPCs) and to identify the GCIs the SMERSH algorithm normally requires multiple microphone signals. In case only one noisy microphone signal is used the LPCs are biased due to early and late reverberation and background noise, and correct identification of the GCIs is difficult. By applying the spectral processing of Section 2.1 first we are able to i) reduce the overlap between subsequent phonemes, and ii) reduce part of the background noise. By reducing the overlap and the noise the obtained LPCs are good estimates of the LPCs of the anechoic speech signal. The block diagrams of the two stages are shown in Figs. 1 and 2, respectively.

4. EVALUATION

Simulation results are provided to demonstrate the performance of the developed system. The APLAWD database [13] was used for evaluation with the sampling frequency set to $f_s = 8$ kHz; it contains anechoic recordings comprising ten repetitions of five sentences uttered by five male and five female talkers. In all experiments the LPC analyses was performed using 30 ms frames overlapping by 50%, and the prediction order $p = 13$. The number of neighbouring weighted larynx cycles was $\mathcal{I} = 2$. Reverberation was simulated by convolution of the anechoic speech samples, and an RIR, measured in two conference rooms with a reverberation time of approximately 350 and 475 ms. The source was positioned at distances $d = 1.5$ m and $d = 3$ m from the microphone.

4.1. Objective Colouration Measure

The segmental Signal to Reverberation Ratio (SRR) and Bark spectral distortion (BSD) [1] are frequently employed evaluation metrics. In the context of reverberation their values depend on the DRR of the acoustic channel and are unable to measure the coloration caused by the early reflections independently. The perceptual coloration increases with increasing source-microphone distance until the source-microphone distance is equal to the critical distance, i.e., the DRR is smaller than 0 dB. The channel-based spectral deviation measure, which is defined as the standard deviation of the log spectral amplitude of the RIR [14], measures the 'flatness' of the amplitude spectrum of the RIR and is a known measure for the colouration. It was found that the kurtosis of the LP residual, which was also used in [3, 4], correlates well with the spectral deviation. In Fig. 3 the kurtosis of the LP residual and the spectral deviation are shown (averaged over all speech fragments in the APLAWD database) against the source-microphone distance (the reverberation time was 0.5 s and the critical distance was 0.8 m). The correlation coefficient between kurtosis and the spectral deviation is -0.989, which indicates a highly linear relation between the two values. Therefore, we choose to use the kurtosis of the LP residual as an objective measure for the colouration of reverberant speech.

4.2. Experimental Results

The performance of DYPSA for reverberant speech was evaluated in [5]. For the clean speech the GCIs detection rate was approximately 95.7% and identification accuracy of 0.71 ms. The detrimental effect of reverberation is apparent, with detection rate drop of up to 40% and accuracy in excess of 1 ms at $T_{60} = 0.5$ s. We found that the detection and accuracy are still sufficient for the SMERSH algorithm.

Segmental Signal to Interference Ratio (SIR), Bark spectral distortion (BSD), and the kurtosis of the LP residual (KLPR) were employed as evaluation metrics. The dereverberation system was tested using noisy speech signals that were generated by adding a white Gaussian noise to the reverberant speech signals, such that the Signal to Noise Ratio was equal to 15 and 25 dB. The results, averaged over all speech fragments in APLAWD, are shown in Table 1 for

Table 1. SegSIR, BSD and Kurtosis of the LP residual (KLPR) obtained using Spectral Processing (SP) and Spectral Processing and LP Residual Processing (SP+LPRP).

		SNR = 15 dB			SNR = 25 dB		
		SegSIR [dB]	BSD [dB]	KLPR	SegSIR [dB]	BSD [dB]	KLPR
Room 1 ($T_{60} \approx 475$ ms, $d = 3$ m)	Unprocessed	-6.50	0.24	4.51	-5.34	0.22	6.16
	SP	-4.03	0.18	5.00	-3.90	0.19	6.86
	SP + LPRP	-3.56	0.15	7.24	-3.42	0.16	8.64
Room 2 ($T_{60} \approx 350$ ms, $d = 1.5$ m)	Unprocessed	-6.87	0.19	1.43	-5.34	0.17	2.67
	SP	-3.79	0.14	1.81	-3.30	0.14	3.17
	SP + LPRP	-3.66	0.13	3.61	-3.28	0.14	4.56

(a) reverberant speech, (b) spectrally processed speech (SP), i.e., using the first stage of the system and (c) speech processed with both stages of the system (SP + LPRP). From these results and informal listening test[2] we can conclude that the reverberation is significantly reduced. Furthermore, it can be seen that the kurtosis of the LP residual was mainly increased by applying LP residual processing, which indicates that the coloration of the speech is reduced.

5. CONCLUSIONS

In this paper we have developed a single microphone speech dereverberation system that suppresses the effect of early and late reverberation. First, the late reverberant signal component is suppressed using spectral processing. The late reverberant spectral variance is estimated using a recently proposed generalized statistical reverberation model, and is suppressed using a spectral subtraction technique. The coloration caused by the early reflections is suppressed by LP residual processing in which adjacent larynx cycles are averaged in the LP residual domain. The larynx cycles can be identified using the spectrally processed reverberant signal. The effect of the temporal averaging in the LP residual domain is embodied into an equalization filter which is applied to both voiced and unvoiced LP residual. In addition, an objective measure is proposed to measure the coloration caused by early reflections. Experimental results demonstrate the beneficial use of the new single microphone system that reduces echo and coloration with little speech distortion, and of the objective measure which correlates well with the 'flatness' of the spectral amplitude of the RIR.

[2]Some results are available for listening on:
`http://home.tiscali.nl/ehabets/publications/icassp08`

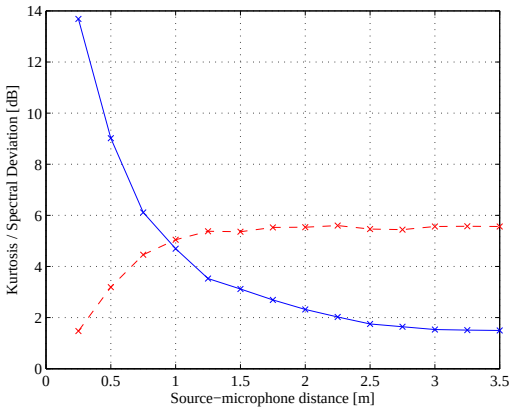

Fig. 3. Kurtosis of the LP residual signal (solid), Spectral Deviation of the room impulse response in dB (dashed).

6. REFERENCES

[1] P.A. Naylor and N.D. Gaubitch, "Speech dereverberation," in *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'05)*, 2005.

[2] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[3] B.W. Gillespie, H. Malvar, and D. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, 2001, vol. 6, pp. 3701–3074.

[4] M. Wu and D. Wang, "A two-stage algorithm for enhancement of reverberant speech," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 3, pp. 774–784, May 2006.

[5] N. Gaubitch and P.A. Naylor, "Spatiotemporal averagingmethod for enhancement of reverberant speech," in *Proc. of the 15th International Conference on Digital Signal Processing (DSP 2007)*, July 2007, pp. 607–610.

[6] E.A.P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.d. Thesis, Technische Universiteit Eindhoven, June 2007.

[7] R. Ratnam, D.L. Jones, B.C. Wheeler, W.D. O'Brien Jr., C.R. Lansing, and A.S. Feng, "Blind estimation of reverberation time," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[8] H. Kuttruff, *Room Acoustics*, Spon Press, London, fourth edition, 2000.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[11] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, Language Porcessing*, vol. 15, no. 1, pp. 34–43, 2007.

[12] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, New York: MacMillan, 1993.

[13] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Tech. Rep., University College London, June 1987.

[14] J.J. Jetzt, "Critical distance measurement of rooms from the sound energy spectral response," *Journal of the Acoustical Society of America*, vol. 65, no. 5, pp. 1204–1211, 1979.