SPATIOTEMPORAL AVERAGING METHOD FOR ENHANCEMENT OF REVERBERANT SPEECH

Nikolay D. Gaubitch, Patrick A. Naylor

Dept. of Electrical and Electronic Engineering, Imperial College London

ABSTRACT

We present a reverberant speech enhancement algorithm which, operating on the linear prediction residual of spatially averaged multi-microphone observations, utilizes temporal averaging of neighbouring larynx cycles. The enhanced larynx cycles are used to design an equalization filter, which is applied in order to dereverberate both voiced and unvoiced speech. The DYPSA algorithm is employed and evaluated for larynx cycle segmentation of reverberant speech. Simulation results show reverberation reduction with up to 5 dB in terms of segmental signal-to-reverberant ratio and 0.34 in terms of normalized Bark spectral distortion score.

***Index Terms*—** Speech dereverberation, GCI identification

1. INTRODUCTION

Hands-free speech acquisition is often desired in modern telecommunications applications operating inside typical office environments. When a speech signal, $s(n)$, is produced in an enclosed space at some distance from an array of M microphones, the observed signals are affected by noise, $\nu_m(n)$, and reverberation. The latter arises from multipath propagation of the acoustic signal due to reflections off walls and surrounding objects, and is characterized by the L-tap room impulse response (RIR), $\mathbf{h}_m = [h_{m,0}\ h_{m,1}\ \ldots\ h_{m,L-1}]^T$. The observed signals are given by

$$x_m(n) = {\mathbf{h}_m}^T \mathbf{s}(n) + \nu_m(n), \quad m = 1, 2, \ldots, M, \tag{1}$$

where $\mathbf{s}(n) = [s(n)\ s(n-1)\ \ldots\ s(n-L+1)]^T$.

Reverberation reduces the perceived quality of speech and alters the characteristics of the speech signal which can be problematic in applications including speech recognition and source localization. The deleterious effects are magnified as the distance between talker and microphones is increased. Thus, dereverberation is important for the future of hands-free applications. The aim of dereverberation is to process the observations, $x_m(n)$, so as to form $\hat{s}(n)$, an estimate of $s(n)$. This is a *blind problem* since, in most practical cases, neither the signal $s(n)$ nor the RIRs $\mathbf{h}_m$ are available.

Dereverberation algorithms can be classified into [1]: (i) *beamforming*, where the microphone signals are delayed, weighted and summed [2]; (ii) *speech enhancement*, where the speech signals are modified to better fit some *a priori* model of the speech signal [2, 3, 4, 5, 6]; (iii) *blind deconvolution*, where the RIRs are identified blindly and equalized [7]. In theory, blind deconvolution provides a means for exact dereverberation. However, existing methods suffer many practical problems such as RIR order estimation, noise robustness and high computational load. Speech enhancement methods, on the other hand, are computationally efficient; they do not require explicit RIR identification and, although they only achieve limited dereverberation, form practically applicable algorithms.

In this paper, we present the **S**patiotemporal averaging **M**ethod for **E**nhancement of **R**everberant **S**peec**h** (SMERSH): a speech enhancement algorithm based on processing of the linear prediction (LP) residual. The speech signals are first spatially averaged followed by temporal larynx cycle averaging of voiced speech LP residual. The effect of the intercycle averaging in the LP residual domain is embodied into an equalization filter which is subsequently applied to both voiced and unvoiced LP residual. Finally, a speech signal with reduced reverberation is synthesized with the enhanced LP residual. In this way, SMERSH provides a more delicate and linear processing compared to previous LP residual processing methods [2, 3, 5]. In such previous methods the aim has been to attenuate the signal between glottal closure instants (GCIs), without specific consideration of the actual shape of the signal. This often results in distortions and reduced naturalness of the processed speech. An early version of the ideas behind SMERSH was presented in [8]. The new developments and results presented here include: multichannel calculation of the LP coefficients, DYPSA for larynx cycle segmentation, a new weighting function for the inter-cycle averaging, and the design and use of the equalization filter.

The remainder of the paper is organized as follows. LP residual processing for dereverberation is reviewed in Section 2. In Section 3, the building blocks of the spatiotemporal averaging method are described. Simulation results are provided in Section 4 and conclusions are drawn in Section 5.

2. REVERBERANT LP RESIDUAL PROCESSING

A speech signal, $s(n)$, can be written in terms of pth order linear prediction as [9]

$$s(n) = -\mathbf{a}^T \mathbf{s}(n-1) + e(n), \tag{2}$$

where $\mathbf{a} = [a_1\ a_2\ \ldots\ a_p]^T$ are the LP coefficients, $e(n)$ is the prediction residual and $\mathbf{s}(n-1) = [s(n-1)\ s(n-2)\ \ldots\ s(n-p)]^T$. Similarly, the mth reverberant observation, $x_m(n)$ can be written

$$x_m(n) = -\mathbf{b}_m^T \mathbf{x}_m(n-1) + e_m(n), \tag{3}$$

with $\mathbf{b}_m = [b_{m,1}\ b_{m,2}\ \ldots\ b_{m,p}]^T$ and $x_m(n-1) = [x_m(n-1)\ x_m(n-2)\ \ldots\ x_m(n-p)]^T$. The LP coefficients can be found by minimizing $e(n)$ or, in the multichannel case, $e_m(n)$. Alternatively, in our approach the LP coefficients, $\mathbf{b}$, are obtained by jointly minimizing the M-channel cost function

$$J_M = \tfrac{1}{M} \textstyle\sum_{m=1}^{M} \sum_{n=-\infty}^{\infty} e_m^2(n). \tag{4}$$

It was shown in [10] that this provides a close estimate of the LP coefficients calculated from clean speech, $\mathbf{a}$. Consequently, the effects

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC).

1-4244-0882-2/07/$20.00 ©2007 IEEE

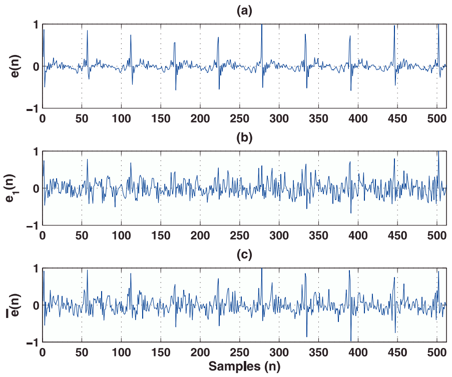

Fig. 1. LP residuals obtained from (a) clean, (b) reverberant, and (c) spatially averaged speech.

of reverberation mainly reside in the LP residual, and a dereverberated speech signal, $\hat{s}(n)$, can be obtained by processing $e_m(n)$ to find an estimate of the true LP residual, $\hat{e}(n) \approx e(n)$.

3. THE SPATIOTEMPORAL AVERAGING METHOD

The observed speech signals are spatially averaged using a delay-and-sum beamformer (DSB) [2]

$$\bar{x}(n) = \tfrac{1}{M} \textstyle\sum_{m=1}^{M} x_m(n - \tau_m), \tag{5}$$

where τ_m is a delay to compensate for the different propagation times between the source and the mth microphone, and is assumed here to be known. We next introduce the inter-cycle averaging.

Consider Fig. 1 showing a portion of the LP residual obtained from (a) clean speech, (b) reverberant speech, and (c) the output of a DSB. The effect of reverberation on the LP residual can be clearly seen, constituting many random peaks of similar strength to the periodic peaks representing the GCIs in clean speech. The following specific observations can be made from the LP residuals in this and in other examples:

(i) The LP residual from the DSB speech differs from that in clean speech by seemingly random peaks that are left unattenuated after the spatial averaging; these appear uncorrelated among consecutive larynx-cycles.

(ii) The main features between consecutive larynx-cycles in the clean speech LP residual change slowly and show high inter-cycle correlation.

(iii) Strong periodic peaks in the LP residual from the DSB speech appear to represent the GCIs seen in the clean speech.

Property (i) arises from the quasi-periodic nature of voiced excitation. Property (ii) is well-known in speech processing and has been applied in, for example, TD-PSOLA [9]. Motivated by these observations, it is proposed that applying a moving average operation on neighbouring larynx cycles in voiced speech will suppress the uncorrelated features and, hence, enhance the LP residual. There are two issues to consider: first, 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 GCIs are important to speech quality [11] and should remain unchanged. Thus, they should be excluded from the averaging process.

DYPSA performs automatic GCI identification in speech [12]. GCI candidates are generated based on the positive zero-crossings of the phase-slope function; additional candidates are obtained by phase-slope projection when a local minimum is followed by a local maximum without crossing a zero. Next, characteristics of voiced speech are used to form a cost function, which is minimized using dynamic programming so to select a subset of the GCI candidates which are most likely to correspond to the true ones. Thus, at the output of DYPSA we obtain the estimated time, n_ℓ, of the ℓth GCI. The dynamic programming makes DYPSA robust to spurious peaks in the prediction residual. This is attractive for GCI identification in reverberant (or spatially averaged) speech and can be expected to discriminate many of the erroneous candidates due to reverberation. Experimental results confirming this are given in Section 4.

In order to leave the glottal pulse undisturbed, a weight function is applied on each larynx frame prior to the averaging. The weight function should, ideally, exclude only the true glottal pulse. However, in practice, GCIs are identified to an uncertainty in the order of 1 ms [12] and the glottal pulse is not a true impulse but is spread in time [9]. A weight function which was found suitable, with a reasonable trade-off between the issues described above, is the time-domain Tukey window defined as [13]

$$w_u = \begin{cases} 0.5 + 0.5\cos\left(\frac{2\pi u}{\beta(\mathcal{L}-1)} - \pi\right), & u < \frac{\beta\mathcal{L}}{2} \\ 0.5 + 0.5\cos\left(\frac{2\pi}{\beta} - \frac{2\pi u}{\beta(\mathcal{L}-1)} - \pi\right), & u > \mathcal{L} - \frac{\beta\mathcal{L}}{2} - 1 \\ 1.0, & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathcal{L}$ is the length of one larynx-cycle (in samples) and $0 \leq \beta \leq 1$ is the taper ratio of the window. An example of the weighting function with $\beta = 0.3$ is shown in Fig. 2. The taper ratio offers a tunable parameter with the beneficial ability to control the amount of the larynx cycle to be included in the averaging process and can be adjusted, for example, in some proportion to the estimation error variance of the GCI identification algorithm. Following the averaging procedure, the inverse weight function with weights , $1 - w_u$, is applied to the larynx frame under consideration to restore the original glottal pulse shape.

Thus, each enhanced larynx cycle in a voiced speech segment is obtained by averaging the current weighted larynx cycle frame under consideration with $\mathcal{I}$ 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

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

where $\bar{\mathbf{e}}_\ell = [\bar{e}(n_\ell)\ \bar{e}(n_\ell + 1)\ \ldots\ \bar{e}(n_\ell + \mathcal{L} - 1)]^T$ is the ℓth larynx-cycle at the output of the DSB with its GCI at time n_ℓ, $\hat{\mathbf{e}}_\ell = [\hat{e}(n_\ell)\ \hat{e}(n_\ell + 1)\ \ldots\ \hat{e}(n_\ell + \mathcal{L} - 1)]^T$ is the ℓth larynx cycle of the enhanced residual, $\mathbf{I}$ is the identity matrix and $\mathbf{W} = \text{diag}\{w_0\, w_1\, \ldots\, w_{\mathcal{L}-1}\}$ is a diagonal weighting matrix. Since the larynx-cycles are not strictly periodic but may vary within a few samples, $\mathcal{L}$ is set to equal the length of the larynx cycle being processed. Other larynx cycles used in the averaging that have less than $\mathcal{L}$ samples are padded with zeros while those with more than $\mathcal{L}$ samples are truncated.

The choice of $\mathcal{I}$ is important: if too many cycles are included, the averaging will remove uncorrelated portions from the original excitation; if too few cycles are considered, erroneous peaks due to

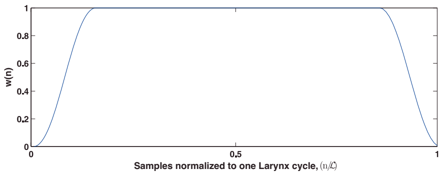

Fig. 2. Larynx weighting function defined in (6) with $\beta = 0.3$.

reverberation will remain. For the results presented here, the number of cycles for averaging was set to $\mathcal{I} = 4$. Generally, it was found through several experiments that $\mathcal{I} > 4$ provides less accurate results.

Although the spatiotemporal averaging attenuates the reverberant components in the prediction residual, two unresolved issues remain: (i) only voiced speech segments are processed, leaving reverberant effects clearly audible in the unvoiced speech and silence, and (ii) the algorithm does not take advantage of past correct larynx-cycle frames in case of erroneous larynx-cycle segmentation due to inaccuracies in the GCI identification. These issues are addressed here by introducing an $L_{\rm i}$-tap FIR filter with coefficients $\mathbf{g}_\ell = [g_{\ell,0}\ g_{\ell,1}\ \dots\ g_{\ell,L_{\rm i}-1}]$, which performs the equivalent operation of the inter-cycle averaging. A least squares estimate of $\mathbf{g}_\ell$ is found from $\hat{\mathbf{g}}_\ell = \min_{\mathbf{g}_\ell} \|\mathbf{g}_\ell^T \bar{\mathbf{e}}_\ell - \hat{e}(n_\ell)\|^2$ and is used to update a slowly varying filter

$$\hat{\mathbf{g}}(n_\ell) = \gamma \hat{\mathbf{g}}(n_{\ell-1}) + (1-\gamma)\hat{\mathbf{g}}_\ell, \qquad (8)$$

where $0 \leqslant \gamma \leqslant 1$ is a forgetting factor with typical values in the range $\{0.1 - 0.3\}$. The filter is initialized to $\hat{\mathbf{g}}(0) = [1\ 0\ \dots\ 0]^T$ with the update performed only during voiced speech segments; in unvoiced speech or silence it is applied at its last update.

4. SIMULATIONS AND RESULTS

Simulation results are provided to demonstrate the performance of the DYPSA and SMERSH algorithms. The APLAWD database [14] 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. Each recording includes a Laryngograph signal, accommodating accurate GCI identification with the HQTx algorithm [15, 12]. Autocorrelation LPC with 30 ms frames overlapping by 50% and prediction order $p = 13$ was used in all experiments. Reverberation was simulated by convolution of RIRs, generated with the source-image method [16], and the anechoic speech samples. The simulated room dimensions were arbitrarily set to $5 \times 4 \times 3$ m with an eight element linear microphone array with 0.05 m uniform element separation. The talker was positioned at a distance of 2.5 m from the centre of the microphone array. The reverberation time, T_{60}, was varied between 0.1 s and 0.5 s.

We begin by evaluating the robustness of the DYPSA algorithm's GCI identification in the presence of reverberation. Following the approach in [12], the GCIs obtained with HQTx were used as the reference. Two metrics were employed for performance evaluation of the DYPSA algorithm on voiced speech: *detection rate*, defined as the percentage of reference GCIs for which exactly one GCI is detected with DYPSA, and *identification accuracy* defined as the standard deviation of the distance between the reference and the identified GCIs. The GCIs from the HQTx algorithm were used as a voiced speech detector. Only those GCIs obtained with DYPSA that are within a voiced region were kept. However, within this voiced region, the GCIs were left as found by DYPSA, i.e. including possible false or inaccurate estimates.

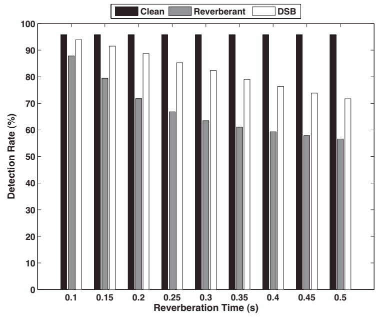

Fig. 3. Detection rate of DYPSA for clean, reverberant, and DSB pre-processed speech.

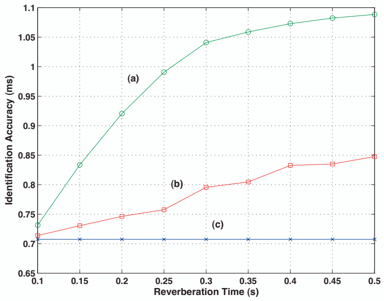

Fig. 4. Identification accuracy of DYPSA for (a) reverberant, (b) DSB pre-processed, and (c) clean speech.

Figure 3 shows a plot of the detection rate versus reverberation time, for clean, reverberant, DSB pre-processed speech, and the corresponding identification accuracy is shown in Fig. 4. These results are the average over utterances in APLAWD. The clean speech results are the same as in [12], showing a detection rate of 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. Remarkably, DYPSA with the DSB as a pre-processor, provides $71.8 - 93.9\%$ in detection rate with accuracy in the range $0.71 - 0.85$ ms, which, although worse than for anechoic speech, is comparable to other algorithms operating on clean speech as seen from the results presented in [12].

The sentence 'George made the girl measure a good blue vase' was used as an illustrative example for the dereverberation experiment. Segmental signal-to-reverberation ratio (SRR) and Bark spectral distortion (BSD) [1] were employed as evaluation metrics.

The results in terms of segmental SRR, averaged over all ten talkers in APLAWD, are shown in Fig. 5 for (a) reverberant speech at

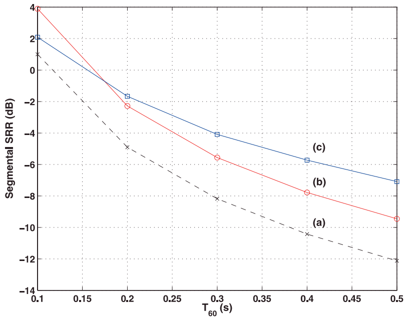

Fig. 5. Segmental SRR vs. reverberation time for (a) reverberant, (b) DSB processed, and (c) SMERSH processed speech.

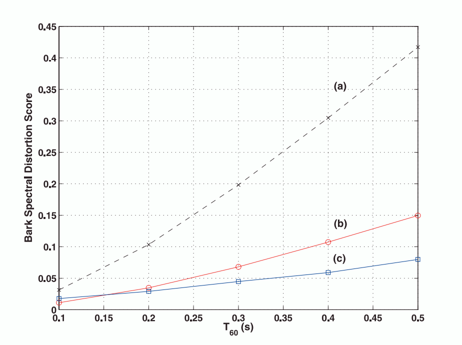

Fig. 6. BSD vs. reverberation time for (a) reverberant, (b) DSB processed, and (c) SMERSH processed speech.

the microphone closest to the talker, (b) DSB speech and (c) speech processed with SMERSH; the corresponding results in terms of BSD are shown in Fig. 6. Reverberation reduction of up to 5 dB in segmental SRR and 0.34 in BSD score is observed at $T_{60} = 0.5$ s, which corresponds to an improvement of 2.4 dB and 0.07 over the DSB. The following can be noted regarding perceptual quality of the processed speech: (i) the reverberant effects are reduced and (ii) the talker appears to be closer to the microphone.

5. CONCLUSIONS

We have presented a multimicrophone method for enhancement of reverberant speech using spatial averaging of the speech signals and temporal inter-cycle averaging in the LP residual. Since the latter relies on accurate GCI identification, we have demonstrated that the DYPSA algorithm can successfully identify GCIs in reverberant speech pre-processed with a delay-and-sum beamformer. Moreover, we introduced the use of an equalization filter, calculated from the enhanced larynx cycles, in order to tackle processing of both voiced and unvoiced speech. Example simulation results confirm the improvement achieved by the proposed algorithm.

6. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, “Speech dereverberation,” in *Proc. Int. Workshop Acoust. Echo Noise Control*, Eindhoven, The Netherlands, Sept. 2005.

[2] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, 1 edition, 2001.

[3] B. Yegnanarayana and P. Satyanarayana, “Enhancement of reverberant speech using LP residual signal,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[4] E. A. P. Habets, “Multi-channel speech dereverberation based on a statistical model of late reverberation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, Mar. 2005, vol. 4, pp. iv/173 – iv/176.

[5] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, “Speech dereverberation via maximum-kurtosis subband adaptive filtering,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, vol. 6, pp. 3701–3704.

[6] M. Delcroix, T. Hikichi, and M. Miyoshi, “Precise dereverberation using multichannel linear prediction,” *IEEE Trans. Audio, Speech, Language Porcessing*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[7] Y. Huang, J. Benesty, and J. Chen, “A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 882–895, Sept. 2005.

[8] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, “Mulit-microphone speech dereverberation using spatio-temporal averaging,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Vienna, Austria, Sept. 2004, pp. 809–812.

[9] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, Macmillan, 1993.

[10] N. D. Gaubitch, D. B. Ward, and P. A. Naylor, “Statistical analysis of the autoregressive modeling of reverberant speech,” *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4031–4039, Dec. 2006.

[11] B. Yegnanarayana, J. M. Naik, and D. G. Childers, “Voice simulation: factors affecting quality and naturalness,” in *Proc. Conf. of the Association for Computational Linguists*, Stanford, California, USA, July 1984, pp. 530–533.

[12] 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.

[13] F. J. Harris, “On the use of windows for harmonic analysis with the Discrete Fourier Transform,” *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[14] G. Lindsey, A. Breen, and S. Nevard, “SPAR’s archivable actual-word databases,” Tech. Rep., University College London, June 1987.

[15] M. Huckvale, “Speech filing system: Tools for speech research,” [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/, July 2003.

[16] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.