SPEECH DEREVERBERATION

Patrick A. Naylor and Nikolay D. Gaubitch

Imperial College London

{p.naylor, nikolay.gaubitch}@imperial.ac.uk

ABSTRACT

The effect of reverberation on speech is to cause it to sound distant and spectrally distorted and can also reduce intelligibility. Dereverberation is therefore an important speech enhancement process for hands-free terminals. This is a blind problem and currently an unsolved problem. This paper reviews existing approaches and discuss current work on this topic in two categories - one based on processing of the LPC prediction residual and one based on a combination of blind channel estimation and channel inversion. The measurement of (de)reverberation is discussed.

1. INTRODUCTION

Reverberation is the process of multi-path propagation of an acoustic signal $s(n)$ from its source to one or more microphones. The observed signal at the m^{th} microphone can be written

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

where $\mathbf{h}_m = [h_{m,1}, h_{m,2}, \ldots, h_{m,L}]^T$ is the impulse response of the acoustic channel from the source to microphone m and $\nu_m(n)$ is observation noise. Reverberant speech can be described as sounding 'distant' with noticeable echo and colouration and these effects generally increase with increasing distance from source to microphone for a given reverberant room. Reverberation has a neglible effect in telephony applications with traditional handsets. However, in hands-free systems, reverberation affects the quality and intelligibility of speech and is a significant problem for both telecommunications and speech recognition applications.

The aim of dereverberation is to form $\hat{s}(n)$, an estimate of $s(n)$, from $x_m(n)$, $m = 1, \ldots, M$. This is a blind problems since neither the signal $s(n)$ nor the room impulse response $\mathbf{h}_m$ are available. Furthermore, typical room impulse responses are time-varying with several thousand coefficients, making the estimation problem extremely difficult.

Several dereverberation algorithms have been proposed and can be considered in two categories: (i) algorithms based on processing of the linear prediction (LP) residual and (ii) blind channel estimation/inversion algorithms. This paper aims to give a brief overview of these techniques and highlight some of the important limitations and open questions for research.

2. EVALUATION AND MEASUREMENT

Reliable quantitative measurement of the level of reverberation in a speech signal is particularly difficult and a unanimously accepted methodology has yet to emerge. In room acoustics, several studies of speech intelligibility in reverberant rooms have been presented. From the room impulse response, it is possible to measure the Direct-to-Reverberant Ratio (DRR) for an observed reverberant signal as the ratio of power due to the direct acoustic path to the power due to the non-direct paths. Several other possible variants of this measure exist [1]. The early reflections are generally considered to result in colouration whereas the later reflections cause a 'distant' and 'echo-ey' sound quality. Early reflections therefore have a less detrimental effect on intelligibility than the later reflections [1]. The Speech Transmission Index [1] is a measure of speech intelligibility in reverberant environments based on the reverberation time and the masking properties of the ear in different frequency bands. Objective measures based on the speech signals can also be adopted such as Segmental SNR or the Bark Spectral Distortion (BSD) [2][3]. In this paper we show results for simulated impulse responses [4] using BSD and the Segmental Signal-to-Reverberation Ratio (SRR) defined as

$$\text{SRR}_{\text{Seg}} = \frac{10}{K}\sum_{k=0}^{K-1}\log_{10}\left\{\frac{\sum_{n=kN}^{kN+N-1} s_d(n)^2}{\sum_{n=kN}^{kN+N-1}(s_d(n)-\hat{s}(n))^2}\right\}, \tag{2}$$

where K is the number of frames, $N = 512$ is the frame length in samples and $s_d(n) = s(n) * h_d(n)$ is the direct-path signal. The delay-and-sum beamformer (DSB) is used as a reference.

3. LP RESIDUAL PROCESSING

The residual signal following LP analysis has been observed to contain the effects of reverberation, comprising peaks corresponding to excitation events in voiced speech

together with additional peaks due to the reverberant channel [3][5]. Several LP residual processing techniques have been developed using established models of speech production. These aim to suppress the effects of reverberation without degrading the original characteristics of the residual such that dereverberated speech can be synthesised using the processed residual and the all-pole filter resulting from LP analysis on the reverberant speech. It is assumed in these techniques that the effect of reverberation on the LP coefficients is insignificant [3].

Griebel and Brandstein use wavelet extrema clustering in [3] to reconstruct an enhanced LP residual. In [6] the authors use coarse room impulse response estimates and applied a matched filter type operation to obtain weighting functions for the reverberant residuals. Yegnanarayana *et al* [7] use time-aligned Hilbert envelopes to represent the strength of the peaks in the LP residuals. The Hilbert envelopes are then summed and used as a weight vector which is applied to the LP residuals of one of the microphones. In [5] the authors derive a weighting function based on the direct-to-reverberant ratio in different regions of the LP residual. Gillespie *et al* [8] demonstrate the kurtosis of the residual to be a useful reverberation metric which they then maximize using an adaptive filter. Although these methods attenuate the impulses due to reverberation in the LP residual, they can also significantly reduce naturalness in the dereverberated speech.

An approach based on spatio-temporal averaging of the LP residual ameliorates this problem [9]. LP analysis is performed on the output of a beamformer resulting in improved identifiability of the voiced-speech excitation events in the LP residual. Other spurious impulses in the residual, which appear uncorrelated among consecutive larynx-cycles, are assumed to be due to reverberation and are suppressed by weighted averaging across $\mathcal{I}$ neighboring cycles on each side. Cycles are weighted according to their offset from the current cycle, $i = 0$, by a constant $1/(|i|+1)$. The result is then added to the original cycle weighted with the inverse weighting function. The ℓ^{th} enhanced cycle, $\tilde{\mathbf{e}}(\ell) = [\tilde{e}(\ell\mathcal{L})\ \tilde{e}(\ell\mathcal{L}+1)\ \ldots\ \tilde{e}(\ell\mathcal{L}+\mathcal{L}-1)]^T$, is thus obtained as:

$$\tilde{\mathbf{e}}(\ell) = \bar{\mathbf{e}}(\ell) \odot (1-\mathbf{w}) + \frac{\sum_{i=-\mathcal{I}}^{\mathcal{I}} \bar{\mathbf{e}}(\ell+i) \odot (\mathbf{w}/(|i|+1))}{\sum_{i=-\mathcal{I}}^{\mathcal{I}} \frac{1}{(|i|+1)}}$$

where $\bar{\mathbf{e}}(n) = [\bar{e}(\ell\mathcal{L})\ \bar{e}(\ell\mathcal{L}+1)\ \ldots\ \bar{e}(\ell\mathcal{L}+\mathcal{L}-1)]^T$ is the residual from the ℓ^{th} larynx-cycle of length $\mathcal{L}$ of a DSB output, $\odot$ is the Hadamard (element-by-element) product and $\mathbf{w} = [w(0)w(1)\ \ldots\ w(\mathcal{L}-1)]^T$ is a weight vector used to preserve the true excitation events. Dereverberation is then achieved using a smoothly time-varying least-squares inverse filter estimated from $\tilde{\mathbf{e}}(\ell)$ and $\bar{\mathbf{e}}(\ell)$.

Figures 1 and 2 show reverberation measurements for a sentence averaged over 5 male speakers and 10 realizations of the spatial setup with 11 microphones spaced 5 cm in a room with $T_{60} = 0.6$ s.

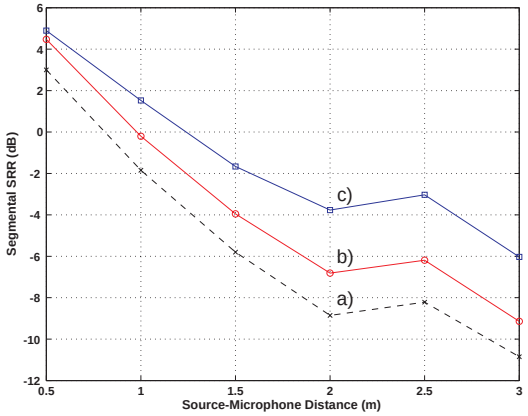

Figure 1: Segmental SRR for (a) reverberant speech, (b) delay-and-sum beamformer and (c) spatio-temporal averaging method.

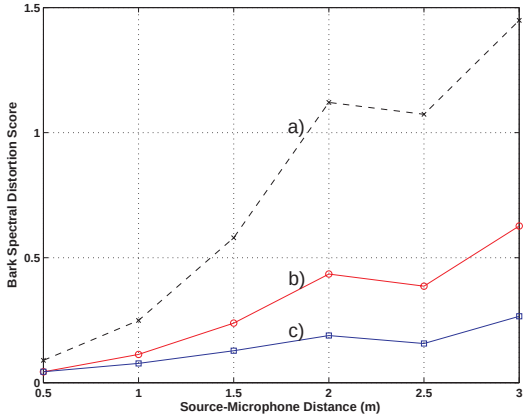

Figure 2: BSD for (a) reverberant speech, (b) delay-and-sum beamformer and (c) spatio-temporal averaging method.

4. CHANNEL ESTIMATION AND INVERSION

4.1. Acoustic Channel Estimation

Blind multi-channel system identification is often based on the cross-relation given for two channels by [10]:

$$x_1(n) * h_2(n) = s(n) * h_1(n) * h_2(n) = x_2(n) * h_1(n)$$

which leads to

$$\mathbf{R}\mathbf{h} = \mathbf{0}, \tag{3}$$

where in general for M channels $\mathbf{R}$ is a correlation-like matrix [11] and $\mathbf{h} = [\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_M^T]^T$ is the composite channel vector. Several closed form batch solutions for

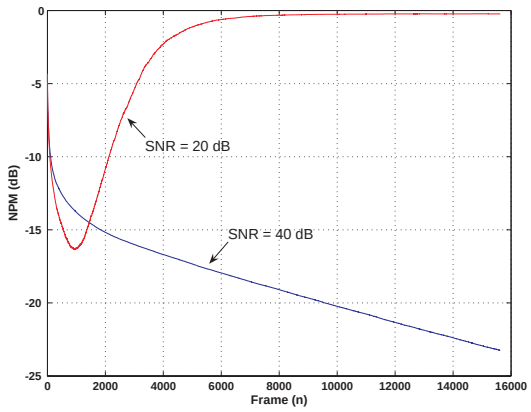

Figure 3: Blind channel estimation with 20 dB and 40 dB SNR using Normalized Multichannel Frequency Domain LMS. Input: white noise, $\mu = 0.5, \lambda = (1 - (1/(3 * L)))^L$, $M = 5$ microphones separated by 5 cm, source-mic distance 1 m.

$\mathbf{h}$ have been proposed and are reviewed in [12]. Gannot and Moonen [13] use subspace methods for dereverberation both in the fullband and in the subband domains. Recently, Huang and Benesty proposed the use of (3) as an error function for adaptive filters and used it to derive multichannel LMS and Newton adaptive filters both in the time domain [14] and in the frequency domain [11]. The Newton algorithms were shown to be able to identify channels of order of hundreds of taps, which is more realistic for acoustic room impulse responses.

Blind acoustic system identification of this type suffers from several limitations which are the subject of current research in the community. (i) Channels cannot be identified uniquely when they contain common zeros. Algorithm performance can be degraded significantly even if zeros are close but not exactly common [15]. Furthermore, the correlation matrix of the source signal $\mathrm{E}\left\{ \mathbf{s}(n)\mathbf{s}^T(n) \right\}$ must be full rank. (ii) Observation noise can cause the adaptive algorithms to diverge as in the example shown in Fig. 3 for Normalized Multichannel Frequency Domain LMS [11]. Several approaches have been developed to improve robustness [16], [17]. (iii) Many approaches assume knowledge of the order of the unknown system. This issue has been addressed, for example, in [13] and [18]. (iv) Solutions for $\mathbf{h}$ are nomally found only to within a scale factor.

4.2. Acoustic Channel Inversion

After performing an identification of the acoustic channels, $h_m(n)$, dereverberation can be achieved in principle by an inverse system with response $g_m(n)$ satisfying $h_m(n) * g_m(n) = \kappa\delta(n - k)$, where k and κ are an arbitrary scale factor and delay. Direct inversion of the acoustic channel is not normally feasible since: (i) it can be several thousand taps in length, (ii) have non-minimum phase [19] and (iii) may contain spectral nulls that after inversion give strong peaks in the spectrum causing narrow band noise amplication.

Several alternative approaches have been studied. Least squares (LS) inverse filters can be designed by minimizing the error $g_{opt,m}(n) = \min_{g_m} \|h_m(n) * g_m(n) - \kappa\delta(n - k)\|^2$ [20] which can also be applied in an adaptive framework [21]. Homomorphic inverse filtering has been investigated [4][20][22], where the impulse response is decomposed into a minimum phase component, $h_{mp}(n)$ and an all-pass component, $h_{ap}(n)$, such that $h(m) = h_{mp}(n) * h_{ap}(n)$. Consequently, magnitude and phase are equalized separately, where an exact inverse can be found for the magnitude, while the phase can be equalized e.g. using matched filtering [22]. It is important to note that magnitude compensation only results in audible distortions in the processed speech signal [19][22].

In the multi-channel case, an exact inverse can be achieved with the MINT method [23] and subband version [24]. If there are no common zeros between the two channel transfer functions, a pair of inverse filters, $g_1(n)$ and $g_2(n)$ can be found such that:

$$h_1(n) * g_1(n) + h_2(n) * g_2(n) = \delta(n). \qquad (4)$$

Thus, exact inverse filtering can be performed. However, undermodelled estimates of $h_m(n)$ are problematic. Furthermore, it has been observed that exact channel inverses are of limited value for practical dereverberation when the channel estimate contains even moderate estimation errors. Figure 4 shows an illustration for $L = 16$. The true impulse response $\mathbf{h}$ has been corrupted with noise to represent estimation error of 0 to -60 dB of Normalized Projection Misalignment (NPM). It can be seen in Fig. 4(a) that equalization using MINT inversion introduces significant spectral distortion for NPM levels greater than around -40 dB, which is unlikely to be achieved by current blind channel estimation techniques. As an alternative, least squares estimation seems more robust to channel estimation errors as shown in Fig. 4(b), although equalizers with very high order are typically required. A study of the effect of delay constraints in the context of acoustic channel inversion for dereverberation was presented in [25]. Its was shown that, for exact inversion, observation noise can be amplified (as in (iii) above) whereas LS solutions generally introduce significant delay which can be problematic in many communications applications.

5. CONCLUSIONS

We have briefly reviewed some of the current and previous work on speech dereverberation. Our aim has been

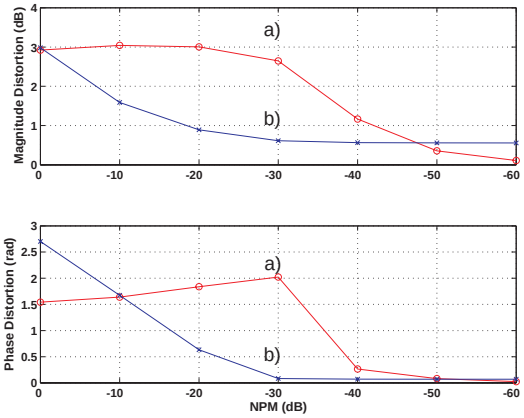

Figure 4: Equalizer magnitude and phase distortion for (a) MINT and (b) Least Squares estimation

to highlight the significant features of the dereverberation problem and show how some of these are being addressed. Dereverberation is a difficult but not impossible problem, presenting several new and exciting challenges to the speech enhancement research community.

6. REFERENCES

[1] H. Kuttruff, *Room Acoustics*, Taylor & Francis, 4 edition, Oct. 2000.

[2] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 819–829, 1992.

[3] M. S. Brandstein and S. M. Griebel, "Nonlinear, model-based microphone array speech enhancement," in *Acoustic Signal Processing For Telecommunication*, S. L. Gay and J. Benesty, Eds., pp. 261–279. Kluwer Academic Publishers, 2000.

[4] 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.

[5] 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.

[6] S. M. Griebel and M. S. Brandstein, "Microphone array speech dereverberation using coarse channel estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, vol. 1, pp. 201–204.

[7] B. Yegnanarayana, S. R. Mahadeva Prasanna, and K. Sreenivasa Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2002, vol. 1, pp. 541–544.

[8] 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.

[9] 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.

[10] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Processing*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[11] 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.

[12] L. Tong and S. Perreau, "Multichannel blind identification: from subspace to maximum likelihood methods," vol. 86, no. 10, pp. 1951–1968, Oct. 1998.

[13] S. Gannot and M. Moonen, "Subspace methods for multi-microphone speech dereverberation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1074–1090, Oct. 2003.

[14] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Process.*, vol. 82, no. 8, pp. 1127–1138, Aug. 2002.

[15] N. D. Gaubitch, J. Benesty, and P. A. Naylor, "Adaptive common root estimation and the common zeros problem in blind channel identification," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, Sept. 2005.

[16] Yiteng Huang, J. Benesty, and Jingdong Chen, "Optimal step size of the adaptive multichannel lms algorithm for blind simo identification," *IEEE Signal Processing Lett.*, vol. 12, no. 3, pp. 173–176, Mar. 2005.

[17] Md. K. Hasan, J. Benesty, P. A. Naylor, and D. B. Ward, "Improving robustness of blind adaptive multichannel identification algorithms using constraints," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, Sept. 2005.

[18] Ken'ichi Furuya and Yutaka Kaneda, "Two-channel blind deconvolution for non-minimum phase impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 1315–1318.

[19] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 165–169, July 1979.

[20] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1982, vol. 7, pp. 1858–1861.

[21] P. A. Nelson, F. Orduña-Brustamante, and H. Hamada, "Inverse filter design and equalization zones in multichannel sound reproduction," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 3, pp. 185–192, Nov. 1995.

[22] B. D. Radlović and R. A. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 6, pp. 728–737, Nov. 2000.

[23] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[24] K. Yamada, J. Wang, and F. Itakura, "Recovering of broad band reverberant speech signal by sub-band MINT method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 969–972.

[25] M. Hofbauer and H. Loeliger, "Limitations for FIR multi-microphone speech dereverberation in the low-delay case," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sept. 2003, pp. 103–106.