MULTI-MICROPHONE SPEECH DEREVERBERATION USING SPATIO-TEMPORAL AVERAGING

Nikolay D. Gaubitch, Patrick A. Naylor, Darren B. Ward

Department of Electrical and Electronic Engineering
Imperial College London
SW7 2AZ, UK
E-mail: {nikolay.gaubitch, p.naylor, d.ward}@imperial.ac.uk

ABSTRACT

The use of the source-filter speech production model in methods for enhancement of reverberant speech has received considerable attention over the last few years. Furthermore, it has most recently been shown that *spatial averaging of the linear prediction (LP) coefficients* is required to improve accuracy in implementation of these types of algorithms. In this paper, we suggest and demonstrate experimentally that *LP coefficients obtained from spatially averaged multi-channel speech signals* achieve an equally satisfactory result. Consequently, we propose a novel multi-channel speech dereverberation approach operating on the LP residual, utilizing a combination of spatial averaging and a new approach based on inter-cycle temporal averaging. Simulation results and informal listening tests indicate an improvement in terms of direct-to-reverberant sound ratio and in perceived quality of the enhanced speech.

1. INTRODUCTION

Speech signals obtained by microphones placed at a distance from the speaker in an enclosed space are degraded in quality due to multiple reflections from the surrounding walls and other objects. The deleterious effect is further magnified as the distance between speaker and receiver increases. Consequently, listeners' perceptual experience and the intelligibility of the captured speech are significantly reduced. This is an important problem, often encountered in "hands-free" applications such as teleconferencing or speech-recognition aimed for use, for example, in offices.

Several dereverberation algorithms based on the source-filter speech production model [1] have been proposed by various authors [2, 3, 4]. The motivation for these methods is the observation that in reverberant environments, the linear prediction (LP) residual contains the original impulses followed by several other peaks due to multi-path reflections. In addition, it is assumed that the LP coefficients calculated from clean speech and those obtained from reverberant speech are equivalent. Consequently, dereverberation is achieved by attenuating the peaks in the excitation sequence due to multi-path reflections and synthesizing the enhanced speech waveform using the modified LP residual. It was recently suggested in [5] that the LP coefficients from reverberant speech should be spatially averaged for the assumed equivalence with the clean speech coefficients to hold.

The main advantage of the source-filter speech production model algorithms is that they can achieve dereverberation without specific knowledge of the room transfer function, which is known to be difficult to estimate. Furthermore, they are more robust to speaker movements. This makes them flexible to use in rooms of different acoustic characteristics.

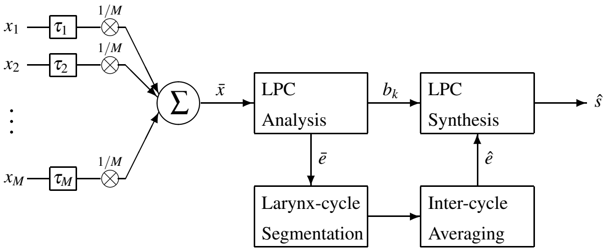

Figure 1: Proposed speech dereverberation algorithm.

Various methods for enhancing the LP residual exist. Griebel and Brandstein [2] use coarse estimates of the room impulse response for each channel and apply a matched filter type operation to obtain weighting functions for the reverberant LP residuals. Yegnanarayana et al. [3] use Hilbert envelopes to represent the strength of the peaks in the LP residuals. The time-aligned Hilbert envelopes from the individual channels are summed and used as a weight vector which is applied to the LP residual of one of the channels. Gillespie et al. [4] demonstrate the kurtosis of the LP residual to be a valid reverberation metric and apply an adaptive filter that maximizes the kurtosis of the excitation sequence. Although these algorithms perform reasonably well in their task of attenuating peaks due to reverberation, they do not consider the original structure of the excitation sequence. Consequently, the enhanced residual can differ significantly from the original excitation signal, resulting in less natural-sounding speech.

In this contribution, we propose a novel approach to use of the source-filter model in speech dereverberation. Further to [5], we show experimentally that an improvement in accuracy of the LP coefficients can be obtained from spatially averaged, time-aligned observations of speech signals produced in a reverberating room. Consequently, using the fact that the waveform between adjacent larynx-cycles varies slowly, each such cycle is replaced by an average of itself and its nearest neighboring cycles to provide the final enhancement of the LP residual. We also introduce the concept that the LP residual from the averaged sequence clearly shows the original instances of glottal closure, such that they can be identified by a suitable algorithm, e.g. DYPSA [6] or HQTx [7]. The major advantage of this approach is that the enhancement process preserves the structure of the clean speech residual. Thus, the enhanced speech is natural-sounding and perceptually close to the clean speech utterance. A diagrammatic summary of the algorithm is presented in Fig. 1.

The remainder of this paper is organized as follows. Section 2 introduces the details of the proposed method. In Section 3, the experimental environment is outlined and results from simulation experiments are presented and discussed. Finally, Section 4 provides concluding remarks about the proposed algorithm based on the current results.

2. SPEECH DEREVERBERATION ALGORITHM

2.1 Linear Prediction of spatially averaged observations

We consider a clean speech signal, $s(n)$, produced in a reverberating room and observed by an array of M microphones. Let the speech signal received at the mth microphone be $x_m(n) = h_m(n) * s(n)$, where $h_m(n)$ is the room impulse response relative to the source and the mth microphone position, and $*$ denotes convolution. The spatially averaged input, $\bar{x}(n)$, is obtained with

$$\bar{x}(n) = \frac{1}{M}\sum_{m=1}^{M} x_m(n-\tau_m). \qquad (1)$$

This is essentially a delay-and-sum (DS) beamformer where appropriate delays, τ_m, which are assumed to be available, are applied on the individual channels in order to time-align the inputs before these are averaged.

Applying LP analysis, we can express the clean speech and the spatially averaged observation signals as a linear combination of their p past samples, which for a single analysis frame becomes

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + e(n), \qquad (2)$$

$$\bar{x}(n) = -\sum_{k=1}^{p} b_k \bar{x}(n-k) + \bar{e}(n), \qquad (3)$$

where a_k and b_k are the corresponding LP coefficients and $e(n)$ and $\bar{e}(n)$ are, respectively, the clean and the DS beamformer output LP residuals. The LP residual is found by inverse filtering the speech signal [1].

Subsequently, we use a fixed microphone array and speaker geometry in the simulation environment specified in Section 3 with a reverberation time $T_{60} = 0.8s$ and using an example vowel /i/. We rotate and translate the entire source-receiver configuration to various randomly selected positions in the room and calculate the LP coefficients of the DS beamformer output for each case. Figure 2 shows a z-plane plot of the poles for a single frame resulting from 5 such cases, $\cdot$, as well as their mean, $\circ$. This suggests that, on average, the vocal tract filters from spatially averaged reverberant speech are close to those obtained from clean speech, $\square$, i.e. the pole positions arising from a_k are approximately equal to the pole positions arising in the spatially averaged case from b_k.

Finally, comparing the LP residuals obtained from the output of the DS beamformer as shown in Fig. 4c to the clean speech residual in Fig. 4a, we see that the peaks due to closure of the glottis become apparent. However, the two residuals are still dissimilar.

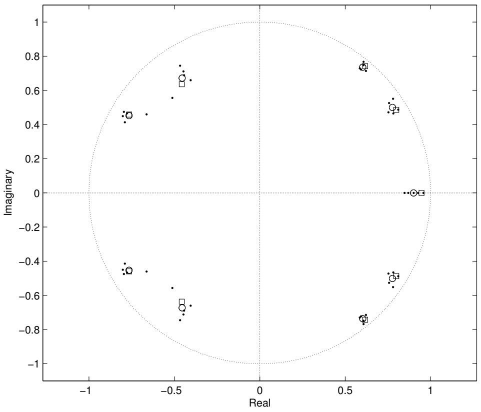

Figure 2: The poles from clean speech ($\square$), DS beamformer output from individual source-microphone positions ($\cdot$) and average over all the beamformer outputs at the different source-microphone positions ($\circ$).

2.2 Averaging across Larynx-cycles

A more detailed investigation of the excitation sequences shown in Fig. 4a and 4c obtained from (2) and (3) leads to the following observations:

- The LP residual obtained from the output of the DS beamformer differs from that in clean speech by seemingly random peaks which are left after the averaging. These peaks, due to the room effects, appear uncorrelated among consecutive larynx-cycles.
- In the case of the clean speech, the prediction residual between consecutive larynx-cycles changes slowly and shows high inter-cycle correlation. This property has also been applied in the context of TD-PSOLA [8].
- The impulses due to glottal closure appear to represent the original, clean speech excitation.

It is therefore proposed that applying a moving average type operation between successive larynx-cycles will enhance the prediction residual to closer resemble the original excitation sequence. There are two major parts in this averaging procedure. First it is necessary to correctly identify the instances of glottal closure so as to segment the larynx-cycles. Second, the true glottal pulse should remain unchanged and thus, should be excluded from the averaging process. Since the peaks due to glottal closure are clearly identifiable in the residual from the DS beamformer, we can apply an algorithm such as HQTx [7] or DYPSA [6] for finding their positions correctly. One of the features that is particularly desirable for the task in question and that DYPSA possesses, is robustness to spurious peaks.

We would like to perform the averaging on the LP residual between the successive larynx-cycles only and leave the glottal pulse undisturbed. Therefore, we apply a weighting function on each frame prior to the averaging, which ideally should exclude only the true glottal pulse. In practice, the position of the glottal pulse is not identified exactly but within a few samples and the glottal pulse is not an impulse but is spread in time. Consequently, a weighting function is needed to take these variations into consideration and the weights have to be chosen such that, as much as possible of

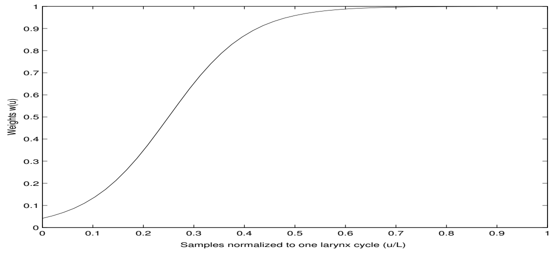

Figure 3: The weighting function $w(u)$ for one larynx-cycle.

the larynx-cycle is included in the averaging process.

The weighting function, $w(u)$, we found suitable to meet the set requirements with a reasonable trade-off between the issues described above is given by

$$w(u) = \frac{\tanh(-\frac{\pi}{2} + \frac{2\pi u}{L}) + 1}{2}, \qquad \text{for } u = 0, 1, \ldots, L-1 \quad (4)$$

where L is the number of samples in one larynx-cycle. The weighting function is depicted in Fig. 3. Moreover, we need to reverse the effect of the applied weights in order to restore the original glottal impulse at the end of the averaging process. This is done utilizing an inverse of (4), $1 - w(u)$.

Each enhanced larynx-cycle in a voiced speech segment is obtained by averaging it with D of its neighboring weighted cycles on each side. The result is then added to the original cycle weighted with the inverse weighting function according to

$$\hat{\mathbf{e}}(l) = \bar{\mathbf{e}}(l) \odot (1 - \mathbf{w}) + \frac{1}{2D+1} \sum_{d=-D}^{D} \bar{\mathbf{e}}(l+d) \odot \mathbf{w} \qquad (5)$$

where $\hat{\mathbf{e}}(l) = [\hat{e}(lL)\ \hat{e}(lL+1)\ \ldots\ \hat{e}(lL+L-1)]^T$ is the lth larynx-cycle of the enhanced residual, $\bar{\mathbf{e}}(n) = [\bar{e}(lL)\ \bar{e}(lL+1)\ \ldots\ \bar{e}(lL+L-1)]^T$ is the lth larynx-cycle from the DS beamformer output residual, $\odot$ is the Hadamard (element-by-element) product and $\mathbf{w} = [w(0)\ w(1)\ \ldots\ w(L-1)]^T$ is the weight vector. Since the larynx-cycles are not strictly periodic but may vary within a few samples, L is set to equal the longest cycle of the $2D+1$ considered, while those of less samples are padded with zeros.

The choice of D is also an important factor in the enhancement of the LP residual. If too many cycles are included the averaging will cancel uncorrelated portions from the original excitation. If too few cycles are considered, then peaks due to reverberation will still remain. We use $D = 2$ in all experiments and have found that $D > 3$ provides less accurate results.

Finally, we obtain an estimate of the clean speech signal, $\hat{s}(n)$, by synthesizing the speech signal using the enhanced residual $\hat{e}(n)$ and the time-varying LP coefficients, b_k, calculated with (3). The dereverberated speech signal is then given by

$$\hat{s}(n) = -\sum_{k=1}^{p} b_k \hat{s}(n-k) + \hat{e}(n). \qquad (6)$$

3. SIMULATIONS AND RESULTS

We present simulation results to demonstrate the performance of the proposed algorithm in terms of LP residual enhancement and dereverberated speech improvement. For the purpose of the experiments we use simulated, finite room impulse responses obtained with the image method [9] assuming a room with dimensions $6 \times 5 \times 4m$. An array of $M = 15$ microphones is positioned along a circular arc in front of the source, such that the source-microphone distance is exactly $2.15m$. The distance between two successive microphones is $0.2m$. Furthermore, we use speech samples taken from the APLAWD database [10], which also includes a Laryngograph (EGG) signal for each sample. In all experiments, the instances of glottal closure are found using the HQTx algorithm [7]. For the LP analysis/synthesis, we use an order of $p = 13$ and $30ms$, 50% overlapping Hamming windowed frames.

To measure improvement of the processed speech, we use segmental signal-to-noise ratio (SNR_{Seg}) defined as [11]

$$SNR_{Seg} = \frac{1}{K} \sum_{k=0}^{K-1} 10\log_{10} \left\{ \frac{\sum_{n=kN}^{kN+N-1} s^2(n)}{\sum_{n=kN}^{kN+N-1} (s(n) - \hat{s}(n))^2} \right\}, \quad (7)$$

where N is the frame length and K is the total number of frames considered. In terms of reverberation, the noise component is due to multi-path effects and thus the measure can be interpreted as a segmental direct-to-reverberant signal ratio.

For the experiments we use the vowel /i/ uttered by a male speaker as an example. Figure 4 shows a segment of the LP residual signals from clean speech, (a), reverberant speech, (b), processed speech by spatial averaging only, (c), and LP residual processed by the proposed algorithm, (d). It is clear from this result that the excitation sequence processed with the proposed method is significantly closer to the original, clean speech residual than that of the DS beamformer.

Figure 5 shows a plot of the SNR_{Seg} at different reverberation times of the reverberant speech for one of the channels, (a), speech enhanced with the spatial averaging only, (b) and speech enhanced with the proposed method, (c). This result indicates that the proposed method provides close to 1.5dB improvement over the DS beamformer at reverberation time $T_{60} = 0.8s$. Furthermore, our approach appears to be more stable to increased reverberation times.

Finally, the algorithm has been applied to a sentence from the APLAWD database and informal listening tests have been performed. The perceptual results of the processed speech can be summarized in the following four points: 1) the reverberant effects due to the room are significantly reduced, 2) the speaker appears to be closer to the microphone, 3) no deleterious artifacts are introduced by the processing and 4) at the end of voiced utterances, the reverberant tail effect is still apparent, which is explained below.

In the processing of these sentences only the voiced speech portions have been enhanced. Unvoiced speech portions have not been altered in the tests mentioned above.

4. CONCLUSIONS

We have proposed a novel multi-microphone speech dereverberation algorithm which utilizes both spatial and temporal averaging. The method is based on linear prediction of spatially averaged microphone array inputs and larynx synchronous temporal averaging of the LP residuals. The algorithm addresses only voiced speech segments.

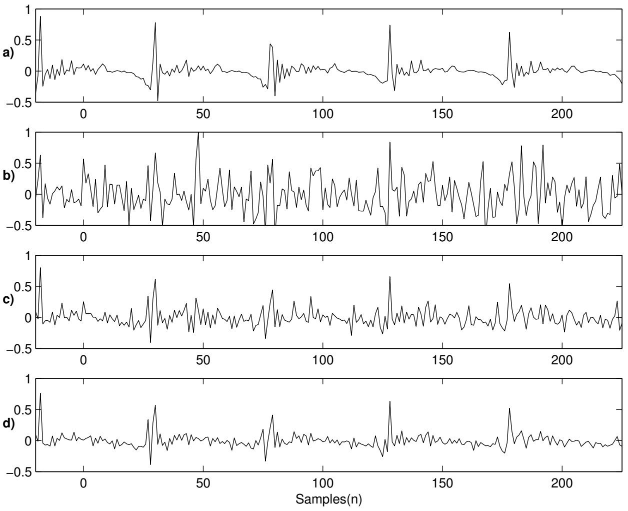

Figure 4: a) Clean speech residual, b) Reverberant speech residual c) DS beamformer output residual d) Residual after processing with proposed method.

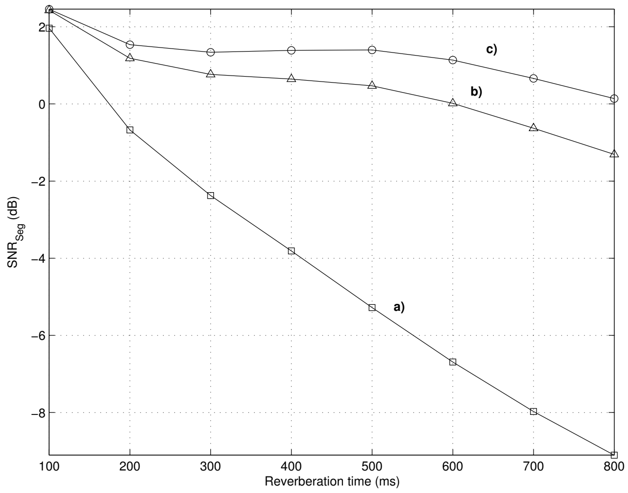

Figure 5: Segmental SNR vs. reverberation time for a) Reverberant speech b) Speech after DS beamforming and c) Speech processed with proposed algorithm.

We have demonstrated experimentally that, on average, the pole positions obtained by LP analysis from of the DS beamformer output show close correspondence to those obtained from clean speech and that this represents an alternative approach to averaging the pole positions [5]. The new approach of inter-cycle averaging has been shown to give further significant enhancement of the LP residual. A prediction residual close to that obtained from clean speech can therefore be found and used for subsequent LP synthesis to provide dereverberated speech. The performance of the algorithm is dependent on the accuracy of the larynx-cycle segmentation, the choice of weighting function and the number of larynx-cycles included in the averaging process. This paper has presented practical examples of these factors and work is undergoing to further examine these parameters.

We have provided simulation results to demonstrate the performance of the proposed algorithm, comparing the achieved performance to the DS beamformer. The results are presented in terms of how well the processed LP residual matches that of clean speech and also in terms of direct-to-reverberant ratio, using a segmental SNR measure for the latter. These show that the proposed method outperforms the DS beamformer, particularly when the reverberation times are high. Based on our current results the proposed algorithm provides up to 9dB improvement in SNR_{Seg} over a single channel of reverberant speech, which is 1.5dB better than the DS beamformer.

Finally, informal listening tests have shown that reverberant effects are reduced considerably, without introducing artifacts. One of the clearly audible features in the processed speech is that the ”distant” effect is reduced. A set of samples of clean, reverberant and processed speech can be found at *http://www.commsp.ee.ic.ac.uk/∼ndg/samples.htm*.

REFERENCES

[1] J. Makhoul, “Linear Prediction: A tutorial review,” *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

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

[3] B. Yegnanarayana, S. M. Prasanna, and K. S. Rao, “Speech enhancement using excitation source information,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2002, pp. 541–544.

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

[5] N. Gaubitch, P. Naylor, and D. Ward, “On the use of linear prediction for dereverberation of speech,” in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC-03)*, Kyoto, Japan, Sept. 2003.

[6] A. Kounoudes, P. Naylor, and M. Brookes, “The DYPSA algorithm for estimation of glottal closure instants in voiced speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, May 2002, pp. I–349 – I–352.

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

[8] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Commun.*, vol. 9, no. 5-6, pp. 453–467, Dec. 1990.

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

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

[11] A. Spanias, “Speech coding: A tutorial review,” *Proc. IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.