ON THE USE OF LINEAR PREDICTION FOR DEREVERBERATION OF SPEECH

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

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

ABSTRACT

In recent years, several multichannel speech dereverberation algorithms have been proposed based on the enhancement of the Linear Prediction (LP) residual signal. In common, they rely on the observation that in reverberant conditions the LP residual contains the original excitation impulses followed by several other peaks due to reverberation. Moreover, they rely on the important assumption that the calculated coefficients of the all-pole filter are unaffected by the multi-path effects of the room.

In this paper, we suggest that this latter assumption holds only in a spatially averaged sense, and that it can not be guaranteed at a single point in space for a given room. Consequently, we present experimental results to demonstrate that an average of the predictor coefficients obtained from spatially distributed microphones can greatly improve the performance of dereverberation algorithms based on LP residual processing compared to those using coefficients from a single channel.

1. INTRODUCTION

The quality of speech recorded in enclosed spaces is degraded by reverberation due to sound wave reflections from surrounding walls and objects. Moreover, the severity of the quality degradation is magnified as the distance between speaker and microphone increases. Therefore, dereverberation of recorded speech is vital for the enhancement of perceived speech quality and for tasks such as speech recognition and speaker verification in "hands-free" telephony applications.

Recently, several dereverberation algorithms based on the source-filter speech production model have been proposed by various authors [1, 2, 3]. The source-filter model describes speech production in terms of an excitation sequence exciting a time-varying all-pole filter. The excitation sequence consists of random noise for unvoiced speech and quasi-periodic pulses for voiced speech, while the filter models the human vocal tract. The all-pole filter coefficients can be estimated through Linear Predictive (LP) analysis of the recorded speech and subsequently, the excitation sequence, or the LP residual, can be obtained by inverse filtering the speech waveform [4, 5].

The motivation for the proposed methods is the observation that in reverberant environments, the LP residual contains the original impulses followed by several other peaks due to multi-path reflections. Furthermore, an important assumption is made that the predictor coefficients obtained from the LP analysis are unaffected by reverberation. 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 and the time-varying all-pole filter with coefficients calculated from the reverberant speech.

To the best of our knowledge, the effects of reverberation on the LP coefficients have not been studied explicitly. However, the validity of the assumption of the pole equivalence for all-pole filters obtained from LP analysis is vital in the dereverberation algorithms based on LP residual enhancement. Therefore, we believe there is a need for a comparative study of the LP coefficients obtained from clean and reverberated speech.

The remainder of this paper is organized as follows. Section 2 provides a theoretical problem formulation for a single and for multiple microphones. Section 3 describes the set of simulation experiments conducted and presents the results. Finally, in Section 4 conclusions are made about the use of LP residual processing for dereverberation of speech based on our current results.

2. PROBLEM FORMULATION

2.1. LP residual processing - single microphone

We consider a clean speech signal, $s(n)$, produced in a reverberant room. The signal received by a microphone is $x(n) = h(n) * s(n)$ where $h(n)$ is the room impulse response relative to the source and microphone position, and $*$ denotes convolution.

Furthermore, applying LP analysis, a speech signal can

be expressed as a linear combination of its p past sample values. The clean and the reverberant speech then become, respectively,

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + e_s(n), \quad (1)$$

$$x(n) = -\sum_{k=1}^{p} b_k x(n-k) + e_x(n), \quad (2)$$

where a_k and b_k are the corresponding LP coefficients and $e_s(n)$ and $e_x(n)$ are, respectively, the clean and the reverberant prediction error signals or LP residuals. In general, the LP coefficients are obtained by minimizing the total sum of the squared prediction error with respect to each of the coefficients and form the analysis and synthesis filters [5]

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}, \quad (3)$$

$$\Lambda(z) = \frac{1}{A(z)}. \quad (4)$$

The source-filter model links the linear prediction to speech signals. The LP residual represents the excitation sequence and the filter $\Lambda(z)$ accounts for the vocal tract effects [4]. Inversely, the LP residual is found by inverse filtering the speech signal. As it was mentioned earlier, the LP residual from reverberant speech contains the original excitation impulses and other significant peaks due to reverberation.

Several speech dereverberation methods have been proposed recently, which aim to obtain an enhanced LP residual, $\hat{e}(n)$ so that $\hat{e}(n) \cong e_s(n)$. Moreover, these algorithms assume that the LP coefficients are unaffected by multi-path reflections, such that $a_k = b_k$, and can thus obtain a clean speech estimate, $\hat{s}(n)$, from the microphone signal as

$$\hat{s}(n) = -\sum_{k=1}^{p} b_k x(n-k) + \hat{e}(n) \quad (5)$$

Griebel and Brandstein [2] use a multichannel approach to obtain a rough estimate of the room impulse response for each channel. They further apply a matched filter type operation to provide a weighting function for the reverberant LP residuals of each channel. Finally the enhanced speech signals are used in a beamforming procedure to provide the final estimate of the enhanced speech signal. Yegnanarayana et. al. [3] use Hilbert envelopes to represent the strength of the peaks in the LP residuals. The Hilbert envelopes from the individual channels are then time-aligned and added. The resulting weight vector is applied to the LP residuals of one of the channels giving the enhanced excitation sequence. This is finally used to synthesize the dereverberated speech. A different approach is proposed by Gillespie et. al. [1], where the kurtosis of the LP residual is shown to be a valid reverberation metric. Consequently, they apply an adaptive filter maximizing the kurtosis of the excitation sequence. This filter is duplicated and applied to the speech signal directly, thus avoiding reliance on the LP coefficient equivalence assumption. This method is also extended to multiple channels where each channel is treated individually.

2.2. LP residual processing - multiple microphones

Applying statistical room acoustic (SRA) theory [6] we are able to show that the spatially expected values of the predictor coefficients obtained from reverberant speech are equivalent to those obtained from clean speech [7], i.e.

$$E\{b_k\} = a_k, \quad (6)$$

where $E\{\cdot\}$ denotes the expectation taken over space. Significantly, this equivalence is only true if one takes expectation over a spatial region, it is not true at a single point in space. The discussion in Section 2.1 can be extended into multiple channels with M microphones so that the speech signal captured at the m^{th} microphone is $x_m(n) = h_m(n) * s(n)$, where $h_m(n)$ is the room impulse response relative to the m^{th} microphone.

Consequently, an enhanced speech signal can be obtained from the m^{th} sensor with an enhanced LP residual in terms of its LP coefficients as

$$\hat{s}(n) = -\sum_{k=1}^{p} E\{b_k\} x_m(n-k) + \hat{e}_m(n). \quad (7)$$

This result suggests that an averaged value of the coefficients obtained from M spatially distributed microphones will provide a value similar to that of the coefficients obtained from clean speech. However, it also implies that, in accordance with SRA, this equivalence can not be guaranteed at a single point in space of a reverberant room.

3. EXPERIMENTS AND RESULTS

We present simulation results to demonstrate the deviation of the LP coefficients obtained from reverberant speech from those obtained form cleans speech and how this deviation is reduced via spatial averaging. Moreover, the advantages of using the averaged coefficients are demonstrated in the context of a speech dereverberation application. For the purpose of the simulation experiments, we assume a room of dimensions 4x4x3m. A speaker is situated at coordinates $(1, 3, 1.5)$ and an array of $M = 15$ equidistant microphones is positioned along one of the room walls at $([x_1, ..., x_M], 1, 1.5)$m. The distance between successive sensors is $|x_m - x_{m-1}| = 0.1875m$. The reverberation

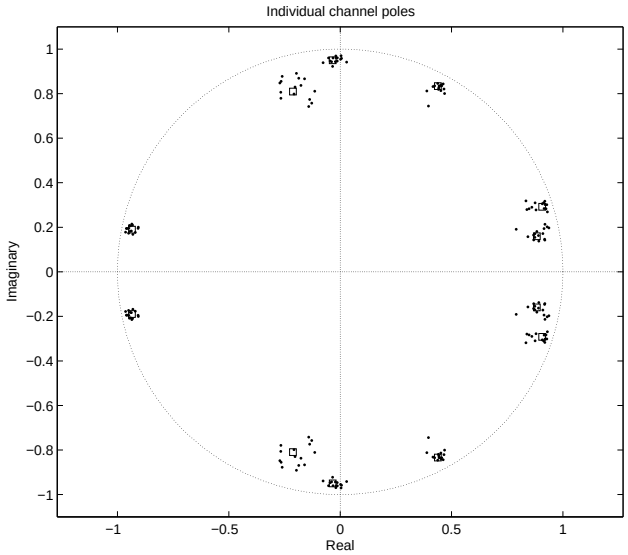

Fig. 1. Poles from 15 channels(dots), and the clean speech poles(squares).

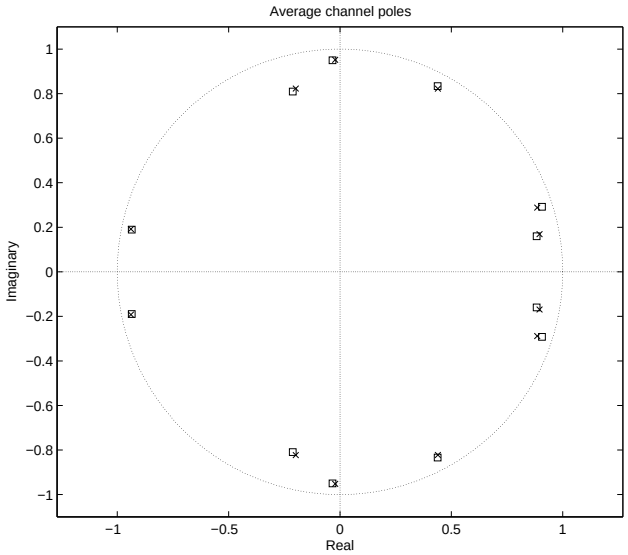

Fig. 2. Averaged poles from 15 channels(crosses) and the clean speech poles(squares)

time is set to $T_{60} = 100$ms. We simulate this environment using finite impulse responses obtained from the Allen and Berkley source-image model [8]. Synthetic speech sampled at 8kHz, with known pitch period and pole order is used for all simulations. We compute the LP coefficients for each microphone as well as for the clean speech using 12^{th} order LP analysis with 30ms, 50% overlapping frames. These values coincide with the exact order and pitch period of the speech signal used.

Furthermore, we apply a simple averaging scheme for the poles obtained from the various microphones as follows. Each channel's poles are arranged according to their angular position on the unit circle and corresponding values are averaged over all channels. If real speech is to be used this may not be directly applicable since there may be many spurious poles due to inaccurate filter order estimation or e.g. nasal sounds for which the zeros introduced in the system are approximated by poles. Therefore, in real applications a more sophisticated averaging algorithm would be required. Consequently, in this paper we have chosen to present results based on synthetic speech only since it provides an unbiased view of the situation.

Figure 1 shows the clean speech poles plotted on the z-plane together with the poles from the 15 channels for a single analysis frame. It can be seen that the poles calculated from the reverberant speech cluster around those obtained from clean speech. However, there are cases in most pole positions where some of the poles deviate from the clean speech value and thus, it can not be guaranteed that the poles from a single sensor are a good approximation to the poles of the clean speech. The result of the averaged poles for the same frame is shown in Figure 2. It is apparent that the averaged poles provide a reasonably good approximation of the clean speech poles. This is again confirmed in Figure 3 where the spectra of a single channel, the clean speech and the averaged channels are shown. Together, these results validate the result stated in (6).

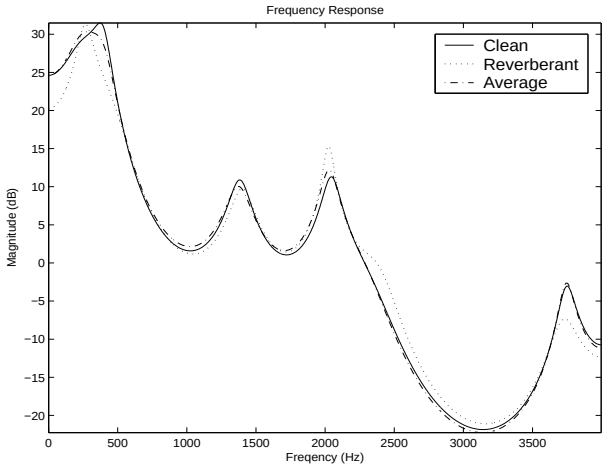

Fig. 3. Spectra for clean speech poles, single microphone reverberant speech poles and average poles for reverberant speech over $M = 15$ microphones

Subsequently, we apply the method proposed by Yegnanarayana et. al. [3] with the reverberated speech and synthesize the enhanced LP residual with the LP coefficients from a single channel as well as with coefficients from poles averaged over all channels. We chose this algorithm since it

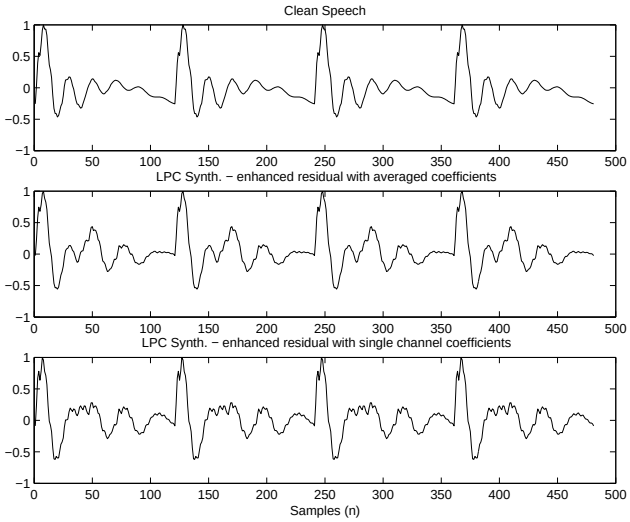

Fig. 4. Synthesized speech with LP residual enhanced with the method in [3] and poles from the reverberant speech

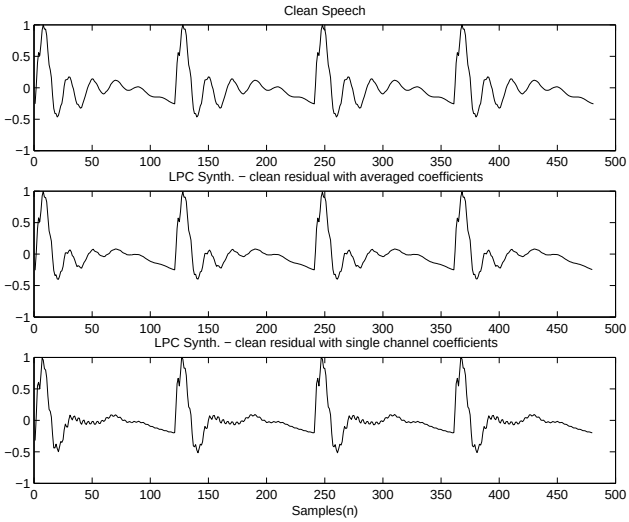

Fig. 5. Synthesized speech with clean speech residual and poles from reverberant speech

explicitly uses poles from a single channel in its dereverberation procedure. The result is shown in Figure 4. The small distortion in the case of the processed residual is due to the fact that in the enhanced residual the peaks due to reverberation have been attenuated and not entirely removed. A similar experiment is conducted where the LP residual from clean speech is used, again, with the LP coefficients from a single channel and with the averaged coefficients. This scenario represents an ideal situation, that is when the residual is processed to contain only the original excitation sequence. Figure 5 shows the results. In both cases a clear improvement can be observed when using the averaged coefficients rather than those from a single channel.

4. CONCLUSIONS

We have demonstrated through experiments that the expected value of the predictor coefficients calculated from LP analysis of reverberant speech are approximately equal to those from clean speech. Significantly our results show that this can not be guaranteed at a single microphone. Therefore, in order to obtain an accurate estimate of the predictor coefficients, an average over several spatially distributed microphones is required. Our results have shown that these averaged coefficients can provide potentially better results when used in conjunction with dereverberation algorithms that rely on LP residual enhancement. It can finally be concluded that in order to use LP residual enhancement techniques for speech dereverberation, unless a method similar to that proposed by [1] is considered, where the dependence on the pole equivalence is eliminated, a spatial average of the predictor coefficients would be preferable.

5. REFERENCES

[1] B.W. Gillespie, H.S. Malvar and D.A.F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in ICASSP 2001, vol. 6, pp. 3701-3704.

[2] S.M. Griebel and M.S. Brandstein, "Microphone array speech dereverberation using coarse channel modeling," in ICASSP 2001, vol.1, pp. 201-204

[3] B. Yegnanarayana, S.R. Mahadeva Prasanna and K. Sreenivasa Rao, "Speech enhancement using excitation source information," in ICASSP 2002, vol.1, pp. 541-544

[4] B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol. 50, no. 2, pp. 637-655, 1971

[5] J. Makhoul, "Linear Prediction: A tutorial review," IEEE Proc., vol. 63, no. 4, pp. 561-580, 1975

[6] A.D. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*, McGraw-Hill, New York, 1981

[7] N.D. Gaubitch, P.A. Naylor, D.B. Ward, "AR modelling of reverberant speech," in preparation

[8] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am., vol. 65, no.4, pp. 943-950, 1979