DUAL-MICROPHONE SPEECH DEREVERBERATION USING A REFERENCE SIGNAL

E.A.P. Habets, Student member, IEEE

Department of Electrical Engineering
Technische Universiteit Eindhoven
Eindhoven, The Netherlands
Email: *e.habets@ieee.org*

S. Gannot, Senior Member, IEEE

School of Engineering
Bar-Ilan University
Ramat-Gan, Israel
Email: *gannot@eng.biu.ac.il*

ABSTRACT

Speech signals recorded with a distant microphone usually contain reverberation, which degrades the fidelity and intelligibility of speech, and the recognition performance of automatic speech recognition systems. In this paper we propose a speech dereverberation system which uses two microphones. A Generalized Sidelobe Canceller (GSC) type of structure is used to enhance the desired speech signal. The GSC structure is used to create two signals. The first signal is the output of a standard delay and sum beamformer, and the second signal is a reference signal which is constructed such that the direct speech signal is blocked. We propose to utilize the reverberation which is present in the reference signal to enhance the output of the delay and sum beamformer. The power envelope of the reference signal and the power envelope of the output of the delay and sum beamformer are used to estimate the residual reverberation in the output of the delay and sum beamformer. The output of the delay and sum beamformer is then enhanced using a spectral enhancement technique. The proposed method only requires an estimate of the direction of arrival of the desired speech source. Experiments using simulated room impulse responses are presented and show significant reverberation reduction while keeping the speech distortion low.

Index Terms— Speech dereverberation, speech enhancement.

1. INTRODUCTION

Acoustic signals radiated within a room are linearly distorted by reflections from walls and other objects. These distortions degrade the fidelity and intelligibility of speech, and the recognition performance of automatic speech recognition systems. Early reflections mainly contribute to coloration, or spectral distortion, while late reflections, or late reverberation, contribute noise-like perceptions or tails to speech signals [1]. Spectral coloration and late reverberation cause users of hearing aids to complain of being unable to distinguish one voice from another in a crowded room. One of the reasons why reverberation degrades speech intelligibility is the effect of overlap-masking, in which segments of an acoustic signal are affected by reverberation components of previous segments. In this paper we have investigated the application of signal processing techniques to improve the quality of speech recorded in an acoustic environment.

Dereverberation algorithms can be divided into two classes. The classification depends on whether the Room Impulse Responses (RIRs) need to be known or estimated beforehand. Until now blind estimation of the RIRs, in a practical scenario, remains an unsolved but challenging problem [2]. Even if the RIRs could be estimated, the inversion and tracking would be very difficult. While these techniques try to recover the anechoic speech signal we like to suppress the tail of the RIR by means of spectral enhancement. In the last decade many speech enhancement solutions have been proposed which do not require an estimate of the RIR. For example algorithms based on processing of the linear prediction (LP) residual signal [3, 4]. Other algorithms are based on spectral enhancement techniques and utilize a statistical reverberation model [5, 6, 7]. The later algorithms do not require detailed knowledge on the RIR structure, but require some *a priori* information about room characteristics, for example the reverberation time.

In this paper we propose a dual-microphone speech dereverberation system. A Generalized Sidelobe Canceller (GSC) [8] type of structure is used to enhance the desired speech signal. The GSC structure is used to create two signals. The first signal is the output of a standard delay and sum beamformer, and the second signal is a reference signal which is constructed such that the direct speech signal is blocked. We propose a novel method which utilizes the reverberation present in the reference signal to enhance the output of the delay and sum beamformer. The power envelope of the reference signal and the power envelope of the output of the delay and sum beamformer are used to estimate the residual reverberation in the output of the delay and sum beamformer. The signal is then enhanced using a spectral enhancement technique. An advantage of the proposed method is that it requires a minimum amount of *a priori* knowledge, since we only require an estimate of the Direction of Arrival (DOA) of the desired speech source.

The outline of this paper is as follows, in Section 2 the problem is described. In Section 3 we describe the proposed dereverberation algorithm. Evaluation using simulated RIRs are presented in Section 4. Discussion and conclusions can be found in Section 5.

2. PROBLEM STATEMENT

The m^{th} microphone signal ($m \in \{1, 2\}$) is denoted by $z_m(n)$ and results from the convolution of the anechoic speech signal $s(n)$ and the RIR between the source and the corresponding microphone. The RIR between the source and the m^{th} microphone, at time n, is modelled as a finite impulse response of length L, and is denoted by $\mathbf{a}_m(n) = [a_{m,0}(n), \ldots, a_{m,L-1}(n)]^T$. The RIR is divided into two parts such that

$$a_{m,j}(n) = a^{\text{d}}_{m,j}(n) + a^{\text{r}}_{m,j}(n) \tag{1}$$

where j is the coefficient index, $\mathbf{a}^{\text{d}}_m(n)$ consists of the direct path, and $\mathbf{a}^{\text{r}}_m(n)$ consists of all echoes. In the sequel we assume that the

Thanks to the Dutch Technology Foundation STW (project EEL 4921), applied science division of NWO, for funding.

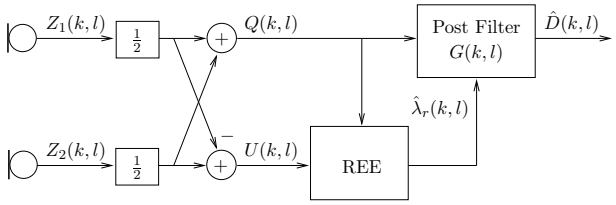

Fig. 1. Dual Microphone Speech Dereverberation System (REE: Reverberant Energy Estimator).

microphone array is steered towards the desired source using an estimate of the DOA of the direct signal, i.e., the direct speech signals in $z_1(n)$ and $z_2(n)$ are time-aligned. The m^{th} microphone signal is given by

$$z_m(n) = \underbrace{\left(\mathbf{a}_m^{\text{d}}(n)\right)^T \mathbf{s}(n)}_{d_m(n)} + \underbrace{\left(\mathbf{a}_m^{\text{r}}(n)\right)^T \mathbf{s}(n)}_{r_m(n)}, \tag{2}$$

where $\mathbf{s}(n) = [s(n), \ldots, s(n-L+1)]^T$, $d_m(n)$ is the desired (direct) speech component, and $r_m(n)$ denotes the reverberant component which contains all reflections. Using the Short-Time Fourier Transform (STFT), we have in the time-frequency domain

$$Z_m(k,l) = D_m(k,l) + R_m(k,l) \quad \forall m \in \{1,2\}, \tag{3}$$

where k represents the frequency bin index, and l the frame index.

Figure 1 shows the proposed dual-microphone speech dereverberation system. The time-frequency signal $Q(k,l)$ is the output of a delay and sum beamformer (in this case with zero delay), i.e.,

$$\begin{aligned} Q(k,l) &= \tfrac{1}{2}\left(Z_1(k,l) + Z_2(k,l)\right), \\ &= D(k,l) + R_q(k,l), \end{aligned}$$

where $D(k,l)$ denotes the direct speech, and $R_q(k,l) = \frac{1}{2}(R_1(k,l) + R_2(k,l))$ denotes the residual reverberation of the speech in $Q(k,l)$. The reference signal $U(k,l)$ is constructed using the difference between the two microphone signals, i.e.,

$$U(k,l) = \tfrac{1}{2}\left(Z_1(k,l) - Z_2(k,l)\right). \tag{4}$$

In case there are no steering errors the direct signal is perfectly blocked, i.e., $D_1(k,l) - D_2(k,l) = 0$, such that

$$U(k,l) = \tfrac{1}{2}\left(R_1(k,l) - R_2(k,l)\right). \tag{5}$$

We can now see that $U(k,l)$ contains the (spatially filtered) reverberation.

Note that the exact relation between $U(k,l)$ and $R_q(k,l)$ is very complex due to the spatial filtering, e.g., for low frequencies the delay and sum beamformer is omnidirectional, while the null beamformer, which is used to create the reference signal, will not only suppress the direct signal but also some reflections. However, using the statistical reverberation model used in [7] it can be shown that for frequencies above the Schroeder frequency $\mathcal{E}\{|U(k,l)|^2\} \approx \mathcal{E}\{|R_q(k,l)|^2\}$, where $\mathcal{E}\{\cdot\}$ denotes the mathematical expectation.

The spectral speech component $\hat{D}(k,l)$ is obtained by applying a frame and frequency dependent spectral gain function $G(k,l)$ (see Section 3) to the spectral component $Q(k,l)$, i.e.,

$$\hat{D}(k,l) = G(k,l)\, Q(k,l). \tag{6}$$

The dereverberated speech signal $\hat{d}(n)$ can be obtained using the inverse STFT and the weighted overlap-add method.

3. PROPOSED METHOD

In this section we show how the residual reverberant energy can be estimated using the reference signal. Additionally, we design a post filter which uses this estimate to enhance the speech signal.

3.1. Reverberant Energy Estimator

First the power envelopes of the output of the delay and sum beamformer $Q(k,l)$ and the reference signal $U(k,l)$ are recursively estimated, using

$$\lambda_q(k,l) = \beta\lambda_q(k,l-1) + (1-\beta)|Q(k,l)|^2, \tag{7}$$

and

$$\lambda_u(k,l) = \beta\lambda_u(k,l-1) + (1-\beta)|U(k,l)|^2, \tag{8}$$

respectively, where β $(0 \leq \beta < 1)$ is the forgetting factor.

Let us assume that the estimated residual reverberant energy in frequency bin k at frame l can be estimated using

$$\hat{\lambda}_r(k,l) = W(k)\lambda_u(k,l-\Delta), \tag{9}$$

where $W(k)$ is a frequency dependent constant. The parameter Δ can be used to control the end point of the uncompensated part of the residual reverberation, e.g., by increasing Δ one can reduce only late reflections while leaving the early reflections intact. The end point is measured with respect to the arrival time of the direct speech signal. Note that Δ is a positive integer value. The time related to Δ is given by $\frac{\Delta F}{f_s}$, where f_s denotes the sampling frequency and F denotes the frame rate in samples of the STFT. The frame rate F depends on the window length and the overlap of the STFT.

We now define an error signal $\lambda_e(k,l)$ as,

$$\lambda_e(k,l) = \lambda_q(k,l) - \hat{\lambda}_r(k,l). \tag{10}$$

An adaptive algorithm is used to minimize the following quadratic cost function

$$J = \left(\lambda_e(k,l)\right)^2, \tag{11}$$

such that

$$\hat{W}(k,l+1) = \hat{W}(k,l) - \frac{\mu}{2}\nabla J_W, \tag{12}$$

where μ denotes the step-size parameter, and ∇J_W denotes the gradient with respect to $W(k,l)$, which is given by

$$\nabla J_W = -2\lambda_e(k,l)\lambda_u(k,l-\Delta). \tag{13}$$

Note that $\lambda_e(k,l)$ and $\lambda_u(k,l)$ are real and positive values for all k and l.

3.2. Post Filter

Many spectral enhancement techniques are described in the literature. Spectral subtraction methods are the most widely used due to the simplicity of implementation and the low computational load, which makes them the primary choice for real-time applications. A common feature of this technique is that the interference reduction process can be related to the estimation of a short-time spectral attenuation factor [9]. Since the spectral components are assumed to be statistically independent, this factor is adjusted individually as a function of the relative local *a posteriori* Signal to Interference Ratio (SIR) on each frequency. The *a posteriori* SIR is defined as

$$\gamma(k,l) \triangleq \frac{|Q(k,l)|^2}{\hat{\lambda}_r(k,l)}. \tag{14}$$

Using informal listening tests we concluded that magnitude subtraction gives very good performance. The gain function related to the magnitude subtraction is given by [9]

$$G(k,l) = \max\left\{1 - \frac{1}{\sqrt{\gamma(k,l)}}, G_{\text{min}}\right\}, \qquad (15)$$

where G_{min} is a lower-bound constraint for the spectral gain function which allows us to control the maximum amount of reverberation that is reduced. In the following experiments G_{min} was set to 0.1, which corresponds to maximum attenuation of 20 dB.

4. EVALUATION

In this section we present evaluation results that were obtained using synthetically reverberated signals. One speech fragment which consists of a female voice of 20 seconds and a male voice of 20 seconds, sampled at 8 kHz, was used in all experiments. The synthetic RIRs were generated using the image method [10], and the reflection coefficients were set such that the reverberation time, denoted by T_{60} was equal to approximately 200, 400 and 600 ms. Experiments were conducted using different distances between the source and the center of the array, denoted by d, ranging from 1 to 2 m. The distance between the two microphones was set to 10 cm.

The analysis window of the STFT was a 256 point Hamming window, and the overlap between two successive frames is set to 75%. Each frame is zero padded with 256 points to avoid wrap around errors. The forgetting factor β in (7) and (8) was set to 0.9, and the step-size μ in (12) was set to 0.2.

We used the segmental Signal to Interference Ratio (SIR_{seg}), Bark Spectral Distortion (BSD), and a recently proposed evaluation measure developed by Wen and Naylor called the Reverberation Decay Tail (R_{DT}) [11] to evaluate the proposed algorithm. The R_{DT} jointly characterizes the relative energy in the tail of the RIR and the rate of decay. In [12] the R_{DT} measure was tested using three dereverberation methods, the results were compared to the subjective amount of reverberation indicated by 26 normal hearing subjects. The results showed a strong correlation between the R_{DT} values and the amount of reverberation perceived by the subjects. Note that higher R_{DT} values correspond to a higher amount of relative energy in the tail and/or a slower decay rate. The (properly delayed) anechoic speech signal was used as a reference signal for these speech quality measures. As a reference dereverberation method we show the quality measures calculated from the output of the delay and sum beamformer (DSB). In Table 1 the results are shown for $d = 1$ m and $d = 2$, and $\Delta = 0$ [1]. The quality measures are calculated using 40 seconds of speech data after the filter coefficients have converged. We can see that the SIR_{seg} is increased in almost all scenarios. The BSD measure indicates that the average Bark spectral distance is slightly increased. The R_{DT} values are very consistent and indicate a clear improvement in all cases.

In Figure 2 the spectrogram of the anechoic signal, the microphone signal $z_1(n)$ and the output of the proposed algorithm for $\Delta = 0$ and $\Delta = 16$ are depicted ($d = 2$ m and $T_{60} = 400$ ms). Note that the effect of overlap-masking is reduced and that the first reflections can be preserved by increasing Δ. In Figure 3 the microphone signal $z_1(n)$ and the output of the proposed algorithm, using $d = 2$ m, $T_{60} = 400$ ms and $\Delta = 8$, are depicted. In both figures is can be seen that the smearing caused by reverberation, is reduced.

[1] The results are available for listening on the following web page: *http://www.sps.ele.tue.nl/members/e.a.p.habets/icassp07*.

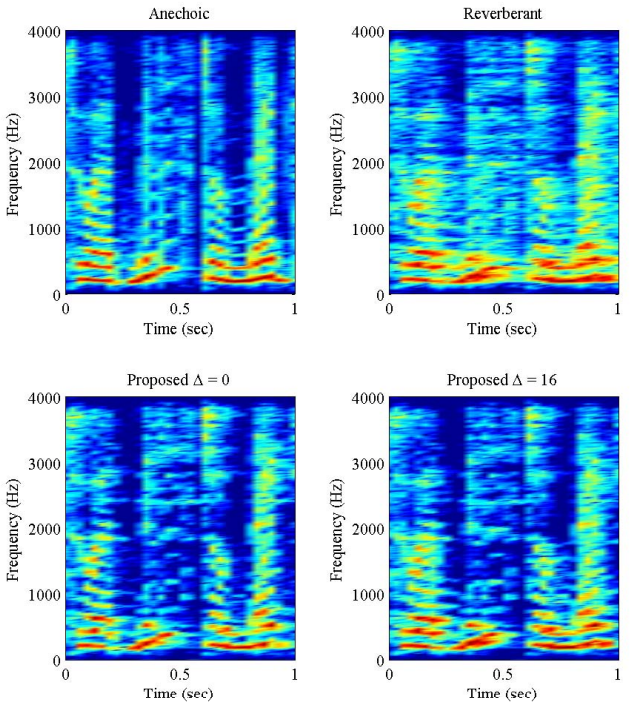

Fig. 2. Spectrograms of the anechoic, reverberant and proposed signals using $\Delta = 0$ and $\Delta = 16$ ($d = 2$ m and $T_{60} = 400$ ms).

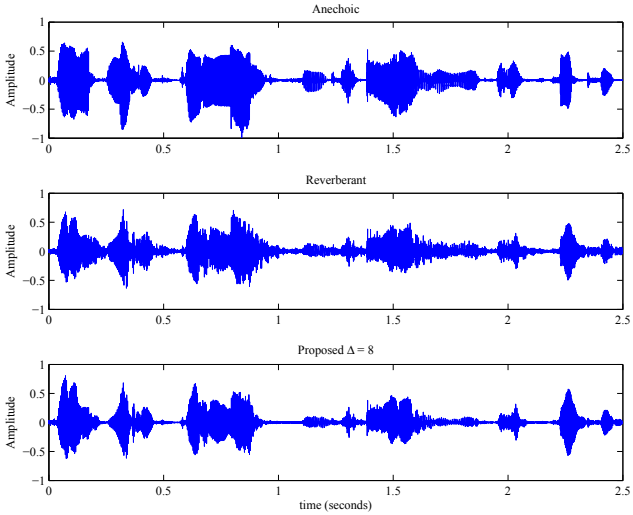

Fig. 3. Anechoic, reverberant and proposed ($\Delta = 8$, $d = 2$ m, and $T_{60} = 400$ ms) signals.

In case the DOA estimation is not perfect the direct speech signal will leak into the reference signal. To study the effects of steering errors due to errors in the DOA estimate we introduced a steering error of 5 degrees. The spectrogram of the processed signals, with and without steering error, are depicted in Figure 4. We can see that the proposed algorithm is still able to suppress a significant amount of reverberation. However, it can also be seen that some additional distortion was introduced by the proposed dereverberation algorithm.

Table 1. Experimental results in terms of segmental Signal to Interference Ratio (SIR_{seg}), Bark Spectral Distortion (BSD) and Reverberation Decay Tail (R_{DT}) for $\Delta = 0$.

d	Method	$T_{60} = 200$ ms			$T_{60} = 400$ ms			$T_{60} = 600$ ms		
		SIR_{seg}	BSD	R_{DT}	SIR_{seg}	BSD	R_{DT}	SIR_{seg}	BSD	R_{DT}
1 m	Unprocessed	8.40 dB	0.05 dB	53	-0.13 dB	0.13 dB	250	-4.31 dB	0.20 dB	568
	DSB	9.03 dB	0.04 dB	42	0.37 dB	0.10 dB	175	-3.95 dB	0.17 dB	463
	Proposed	6.83 dB	0.06 dB	23	2.30 dB	0.13 dB	126	-0.26 dB	0.18 dB	180
2 m	Unprocessed	3.52 dB	0.15 dB	89	-4.17 dB	0.31 dB	454	-8.15 dB	0.41 dB	939
	DSB	4.41 dB	0.12 dB	74	-3.35 dB	0.23 dB	337	-7.43 dB	0.33 dB	766
	Proposed	4.05 dB	0.18 dB	66	-0.12 dB	0.34 dB	200	-2.83 dB	0.45 dB	296

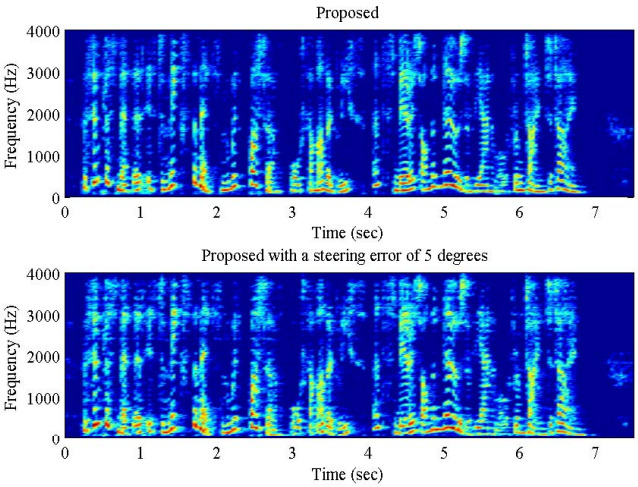

Fig. 4. Spectrograms of the processed signal with and without a steering error of 5 degrees ($\Delta = 0$, $d = 2$ m, and $T_{60} = 400$ ms).

5. DISCUSSION AND CONCLUSIONS

In this paper we proposed a dual-microphone speech dereverberation algorithm. A GSC type of structure was used to enhance the desired speech signal. We proposed to use a reference signal to enhance the output of the delay and sum beamfomer. The advantage of the proposed solution is that we only require an estimate of the DOA. Although no additional interferences have been taken into account, i.e., coherent or non-coherent noise sources, we would like to point out that the power envelope of the reverberant component could also be estimated in a noisy environment (see for example [7]). Experimental results have shown that the proposed solution can be used to reduce the reverberation while keeping speech distortion low. Future research will focus on the extension to multi-microphones, which allows better estimation of the residual reverberant energy, and to more realistic situations where additional interferences are present.

6. ACKNOWLEDGEMENT

The authors express there thanks to Jimi Wen from the Imperial College London, United Kingdom, for making the code for the R_{DT} measure available.

7. REFERENCES

[1] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[2] Y. Huang, J. Benesty, and J. Chen, "Identification of acoustic MIMO systems: Challenges and opportunities," *Signal Processing*, vol. 6, no. 86, pp. 1278–1295, 2006.

[3] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, 2000.

[4] N. D. Gaubitch, P. A. Naylor, and D. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. of the International Workshop on Acoutsic Echo and Noise Control (IWAENC'03)*, Kyoto, Japan, 2003, pp. 99–102.

[5] K. Lebart and J.M. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[6] E.A.P. Habets, "Multi-Channel Speech Dereverberation based on a Statistical Model of Late Reverberation," in *Proc. of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Philadelphia, USA*, March 2005, pp. 173–176.

[7] E.A.P. Habets, S. Gannot, and I. Cohen, "Dual-Microphone Speech Dereverberation in a Noisy Environment," in *Proc. of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2006), Vancouver, Canada*, August 2006, pp. 651–655.

[8] L.J. Griffiths and C.W. Jim, "An Alternate Approach to Linearly Constrained Adaptive Beamforming," *IEEE Transaction on Antennas and Propagation*, vol. 1, no. 30, pp. 27–34, 1982.

[9] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 112–120, April 1979.

[10] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[11] J.Y.C. Wen and P. Naylor, "An evaluation measure for reverberant speech using tail decay modelling," in *Proc. of the European Signal Processing Conference (EUSIPCO 2006)*, Florence, Italy, 2006, pp. 1–4.

[12] J.Y.C. Wen, N.D. Gaubitch, E.A.P. Habets, T. Myatt, and P.A. Naylor, "Evaluation of Speech Dereverberation Algorithms using the MARDY Database," in *Proc. of the 10th International Workshop of Acoutsic Echo and Noise Control (IWAENC 2006), Paris, France*, September 2006, pp. 1–4.