MULTI-CHANNEL SPEECH DEREVERBERATION BASED ON A STATISTICAL MODEL OF LATE REVERBERATION

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

Technische Universiteit Eindhoven, Department of Electrical Engineering,
Signal Processing Systems Group, EH 3.27, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

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 a multi–channel speech dereverberation algorithm is presented which reduces spectral coloration and late reverberation. A spatially averaged amplitude spectrum is used to estimate the instantaneous amplitude spectrum of the clean speech signal, which is then further enhanced using an estimate of the power spectrum of the late reverberant signal. The power spectrum of the late reverberant signal is constructed from multiple microphone signals and a statistical model of late reverberation. The algorithm is tested using synthetic reverberated signals. The performances for different room impulse responses with reverberation times ranging from approximately 150 to 350 ms show significant reverberation reduction with little signal distortion.

1. INTRODUCTION

In general, 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. Reverberation and spectral coloration cause users of hearing aids to complain of being unable to distinguish one voice from another in a crowded room. We have investigated the application of signal processing techniques to improve the quality of speech or music distorted in an acoustic environment.

Early room echoes mainly contribute to coloration, or spectral distortion, while late echoes, or long term reverberation, contribute noise-like perceptions or tails to speech signals [1]. Reverberation reduction processes may generally be divided into single or multiple microphone methods and into those primarily affecting coloration or those affecting reverberant tails.

One important effect of reverberation on speech is overlap–masking, i.e. phonemes are smeared over time, thereby overlapping following phonemes. Lebart et.al. [2] introduced a single–channel speech dereverberation method based on Spectral Subtraction to reduce this effect. The described method estimates the power spectrum of the reverberation based on a statistical model of late reverberation. In this paper we show how this estimate can be improved using multiple microphone signals. Additionally, the fine-structure of the speech signal is partially restored due to spatial averaging of the received amplitude spectra.

Thanks to the Dutch Technology Foundation STW (project EEL 4921) for funding.

The outline of this paper is as follows. In Section 2 we introduce the statistical model for late reverberation. Section 3 describes the reverberant signal model. The Short Time Spectral Modification is described in Section 4. We discuss the complete algorithm and related implementation aspects in Section 5. The performances for different reverberation times are discussed in Section 6, and finally we discuss our conclusions in the last section.

2. ROOM IMPULSE RESPONSE MODEL

Polack [3] developed a time–domain model in which a Room Impulse Response (RIR) is described as one realization of a non-stationary stochastic process. A simplified version of this model can be expressed as

$$h(t) = \begin{cases} b(t)e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases}, \tag{1}$$

where $b(t)$ is a white zero–mean Gaussian stationary noise and α is linked to the reverberation time T_r through

$$\alpha \triangleq \frac{3\ln(10)}{T_r}.$$

The energy envelope of the RIR can be expressed as

$$E_h\{h^2(t)\} = \sigma^2 e^{-2\alpha t}, \tag{2}$$

where σ^2 denotes the variance of $b(t)$ and $E_h\{\cdot\}$ denotes ensemble averaging over h, i.e. over different realizations of the stochastic process in (1).

It can be shown that different realizations of this stochastic process are obtained by varying the position of the receiver with a fixed source position or by varying the position of the source with a fixed receiver position (or of course by varying both positions). We note that the same stochastic process will be observed, irrespective of position, provided that the time origin be defined with reference to the signal emitted by the source and not w.r.t. the arrival time of the direct sound at the receiver. This implies that we can assume ergodicity and evaluate the ensemble average in (2) by spatial averaging.

The RIR can be split into two components, $h_d(t)$ and $h_r(t)$ so that

$$h(t) = \begin{cases} 0 & t < 0 \\ h_d(t) & 0 \leq t < T \\ h_r(t) & t \geq T \end{cases}$$

The value T is chosen such that $h_d(t)$ consists of the direct signal and a few early echoes and $h_r(t)$ consists of all later echoes, i.e. late reverberation. T usually ranges from 40 to 80 ms.

3. REVERBERATION SIGNAL MODEL

The reverberant signal results from the convolution of the anechoic speech signal $s(t)$ and the causal RIR $h(t)$:

$$x(t) = \int_{-\infty}^{t} s(\theta)h(t-\theta)\ d\theta.$$

The auto-correlation $r_{xx}(t,t+\tau) = E_x\{x(t)x(t+\tau)\}$ of the reverberant signal x at time t and lag τ for a fixed source-receiver configuration is

$$r_{xx}(t,t+\tau;h) = \int_{-\infty}^{t}\int_{-\infty}^{t+\tau} E_s\{s(\theta)s(\theta')\}h(t-\theta)h(t+\tau-\theta')\ d\theta d\theta'.$$

The spatially averaged auto-correlation results in

$$r_{xx}(t,t+\tau) = E_h\{r_{xx}(t,t+\tau;h)\} =$$

$$\int_{-\infty}^{t}\int_{-\infty}^{t+\tau} E_s\{s(\theta)s(\theta')\}E_h\{h(t-\theta)h(t+\tau-\theta')\}\ d\theta d\theta'. \quad (3)$$

Using the theory described in Section 2 it follows that

$$E_h\{h(t-\theta)h(t+\tau-\theta')\} = e^{-2\alpha t}\sigma^2 e^{\alpha(\theta+\theta'-\tau)}\delta(\theta-\theta'+\tau),$$

where $\delta(\cdot)$ denotes the Dirac function. Equation (3) leads to

$$\begin{aligned} r_{xx}(t,t+\tau) &= e^{-2\alpha t}\int_{-\infty}^{t} E_s\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\alpha\theta}\ d\theta \\ &= e^{-2\alpha t}\int_{t-T}^{t} E_s\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\alpha\theta}\ d\theta \\ &\quad + e^{-2\alpha t}\int_{-\infty}^{t-T} E_s\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\alpha\theta}\ d\theta. \end{aligned}$$

The auto-correlation at time t can be divided into two terms. The first term depends on the direct signal between time $t-T$ and t, whereas the second depends on the late reverberant signal and is responsible for overlap–masking. Let us consider the spatially averaged auto-correlation at time $t-T$

$$r_{xx}(t-T,t-T+\tau) = \quad (4)$$

$$e^{-2\alpha(t-T)}\int_{-\infty}^{t-T} E_s\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\alpha\theta}\ d\theta.$$

We can now see that the auto-correlation at time t can be expressed as

$$r_{xx}(t,t+\tau) = r_{x_d x_d}(t,t+\tau) + r_{x_r x_r}(t,t+\tau),$$

with

$$r_{x_d x_d}(t,t+\tau) = e^{-2\alpha t}\int_{t-T}^{t} E_s\{s(\theta)s(\theta+\tau)\}\sigma^2 e^{2\alpha\theta}\ d\theta,$$

$$r_{x_r x_r}(t,t+\tau) = e^{-2\alpha T} r_{xx}(t-T,t-T+\tau). \quad (5)$$

In practice the signals can be considered as stationary over periods of time that are short compared to the reverberation time T_r. This is justified by the fact that the exponential decay is very slow, and that speech is quasi-stationary. Let T_s be the time span over which the speech signal can be considered stationary, which is usually around 20-40 ms. We consider that $T_s \leq T \ll T_r$. Under these assumptions, the counterparts of (4) and (5) in terms of the short-term PSDs are approximately:

$$\begin{aligned} \gamma_{xx}(t,f) &= \gamma_{x_d x_d}(t,f) + \gamma_{x_r x_r}(t,f), \\ \gamma_{x_r x_r}(t,f) &= e^{-2\alpha T}\gamma_{xx}(t-T,f). \end{aligned}$$

Therefore, we can estimate the PSD of the direct signal by spectral subtraction of the late reverberant PSD.

4. SHORT TIME SPECTRAL MODIFICATION

Numerous techniques for the enhancement of noisy speech degraded with uncorrelated additive noise have been proposed in literature. Among them the 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 noise reduction process can be related to the estimation of a Short-Time Spectral Attenuation factor. Since the spectral components are assumed to be statistical independent, this factor is adjusted individually as a function of the relative local *A Posteriori Signal to Noise Ratio* on each frequency. The A Posteriori SNR is defined as

$$SNR_{post}(t,f) \triangleq \frac{|X(t,f)|^2}{\gamma_{x_r x_r}(t,f)}. \quad (6)$$

Using informal listening tests we concluded, similar to the findings in [2], that Magnitude Subtraction gives very good performance. The gain function related to the Magnitude Subtraction is given by

$$G(t,f) = 1 - \frac{1}{\sqrt{SNR_{post}(t,f)}}. \quad (7)$$

The estimate of the amplitude spectrum of the signal is given by

$$|\hat{S}(t,f)| = G(t,f)|X(t,f)|. \quad (8)$$

In all frames it is however possible that for some frequencies the estimated amplitude of the noise spectrum is larger than the instantaneous amplitude of the noisy speech spectrum $|X(t,f)|$. Since this could lead to negative estimates for the amplitude of the clean speech spectrum $|\hat{S}(t,f)|$, for these frequencies the gain function $G(t,f)$ is usually put to zero (i.e. half-wave rectification) or equal to a small noise floor value as proposed in [4]. Applying above modification to the gain function in (7) results in the following gain function

$$G(t,f) = \begin{cases} 1 - \frac{1}{\sqrt{SNR_{post}(t,f)}} & \text{if } |\hat{S}(t,f)| \geq \lambda|X(t,f)| \\ \lambda & \text{otherwise} \end{cases} \quad (9)$$

where λ denotes the threshold value.

For single-channel noise reduction additional effort has to be made to reduce residual noise which is mainly caused by the random variations due to the reverberation in $|X(t,f)|$. Under the assumption that the speech signals are time aligned it can be shown that in the multi-channel case this variance can be reduced by replacing the amplitude spectrum $|X(t,f)|$ in (6) by a spatially averaged value, i.e.

$$\overline{|X(t,f)|} = \frac{1}{N}\sum_{n=0}^{N-1}|X_n(t,f)|,$$

where N denotes the number of microphones. Finally we can also use this term in (8) resulting in a partial reconstruction of the fine structure of the speech signal.

5. IMPLEMENTATION

The signals are digitized with a sampling rate of 8 kHz. In the following, the discrete time and frame indices will be denoted by n and m, respectively, and the discrete frequency index by k. An overview of the complete algorithm is presented in Figure 2.

The different stages of the algorithm can be described as follows:

Time Frequency Analysis and Synthesis. The Time Frequency (TF) analysis can be performed in many ways. As an example we used the Short Time Fourier Transform. Although this analysis results in a constant time-frequency bandwidth product, it performs well and has a low computational complexity. The analysis window is a 128 point hamming window, and the overlap between two successive windows is set to 75%. Each frame is zero padded to 256 points in order to avoid wrap around errors. The estimated dereverberated signal $\hat{s}(n)$ is then reconstructed through the overlap-add technique [5] from the estimated amplitude spectrum $|\hat{S}(m,k)|$ and the phase of a Delay & Sum beamformer output. The output of a Delay & Sum beamformer in time-frequency domain can be expressed as:

$$X_{ds}(m,k) = \frac{1}{N}\sum_{n=0}^{N-1} X_n(m,k).$$

Blind Estimation of $\mathbf{T_r}$ and $\gamma_{\mathbf{x_r x_r}}(\mathbf{m},\mathbf{k})$. In order to estimate the reverberant PSD we need to estimate the reverberation time of the room. Partially blind and blind methods have been developed in recent years. For evaluation purposes we have used the average reverberation time measured directly from the synthetic RIRs using Schroeder's method. The short-term PSD $\gamma_{x_r x_r}(m,k)$ is estimated by

$$\begin{aligned}\hat{\gamma}_{xx}(m,k) &= \beta\hat{\gamma}_{xx}(m-1,k) + \frac{1-\beta}{N}\sum_{n=0}^{N-1}|X_n(m,k)|^2,\\ \hat{\gamma}_{x_r x_r}(m,k) &= e^{-2\alpha T}\hat{\gamma}_{xx}(m-T',k),\end{aligned}$$

with $\beta = 0.9$ and $T' = \lfloor \frac{Tf_s}{32} \rfloor$.

Magnitude Spectral Subtraction. The discrete versions of (6), (9) and (8) are used to obtain the estimate $|\hat{S}(m,k)|$. Since the RIR model does not incorporate the direct delay caused by the source-receiver distance we will assume that this delay is fixed, thereby fixing the distance between the source and receiver. This will ensure that the received signals are time-aligned w.r.t. the direct speech signal. The threshold λ in (9) was set to 0.1, corresponding to a maximum attenuation of 20 dB.

6. EVALUATION

The reverberant microphone signals were obtained by convolution of an anechoic female voice of 12 seconds by different RIRs. The experimental setup is depicted in Figure 1. An odd number of microphones were uniformly spaced on an arc, with source-receiver distance $r_d = 3$ m and $\theta = 30^\circ$. The dimensions of the room are 5 m x 6 m x 4 m (l x w x h). The RIRs were constructed using a modified image method [6].

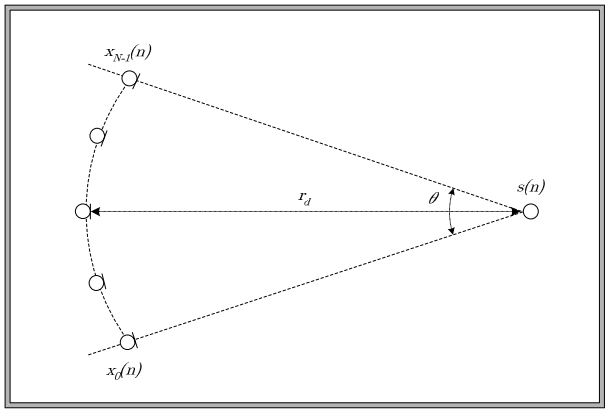

Fig. 1. Experimental setup.

6.1. Objective Measurements

Objective measurements for both the reverberation reduction and the speech distortion were used for this evaluation. The reverberant signal of the center microphone was decomposed into the sum of a direct signal $d_{in}(n)$ and a reverberant part $r_{in}(n)$, obtained by convolving the anechoic signal with the first 6 ms (w.r.t. the direct sound) of the RIR, and with the the RIR minus this part. While the complete reverberant signal was being processed, the time-varying, signal dependent gain function was recorded. The recorded gain was then applied separately to the reverberant part, giving $r_{out}(n)$.

Reverberation Reduction When no speech was present in the anechoic signal the global reverberation reduction was calculated using

$$RR = 10\log_{10}\left(\frac{\sum_{n\in\Omega_{\text{Silence}}} r_{in}^2(n)}{\sum_{n\in\Omega_{\text{Silence}}} r_{out}^2(n)}\right).$$

The separation between speech and silence zones was made trough manual segmentation.

Speech Distortion The cepstral distance between the direct signal $d_{in}(n)$ and the dereverberated signal $\hat{s}(n)$ was used as a measure of distortion. The cepstral distance in frame m is defined by the Euclidian distance between the first eight cepstral coefficients of the direct signal and the dereverberated signal.

$$CD(m) = 2\sum_{k=1}^{8}\left(c_{d_{in}}(m,k) - c_{\hat{s}}(m,k)\right)^2.$$

The cepstral coefficients $c(m,k)$ can be derived directly from the LPC coefficients of the m^{th} frame [7]. Finally, the mean cepstral distance over the periods of speech was calculated.

6.2. Results

The global reverberation reduction and mean cepstral distance are shown in Figure 3 and 4, respectively, for N = {0,1,3,5,7}. The results for $N = 0$ denote the results w.r.t. the reverberant signal obtained from the center microphone. The solid lines are the results for $T = 40$ ms and the dashed lines for $T = 60$ ms. For $T = 40$ ms the reverberation reduction was clearly increased, however in the single-channel case it also resulted in an unacceptable amount of distortion.

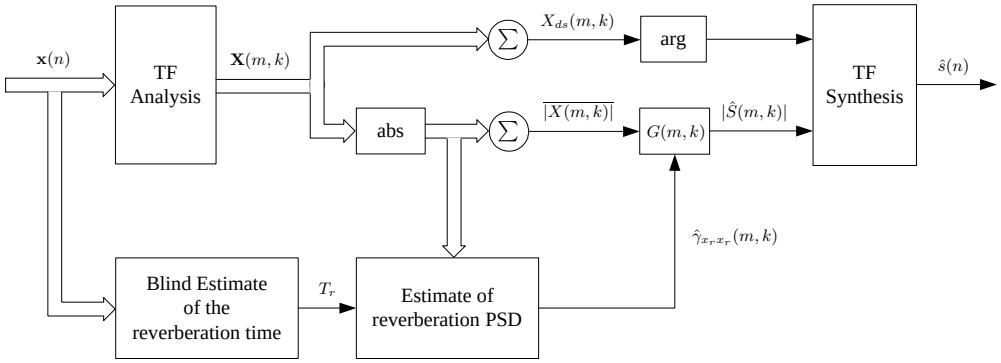

Fig. 2. Overview of the algorithm.

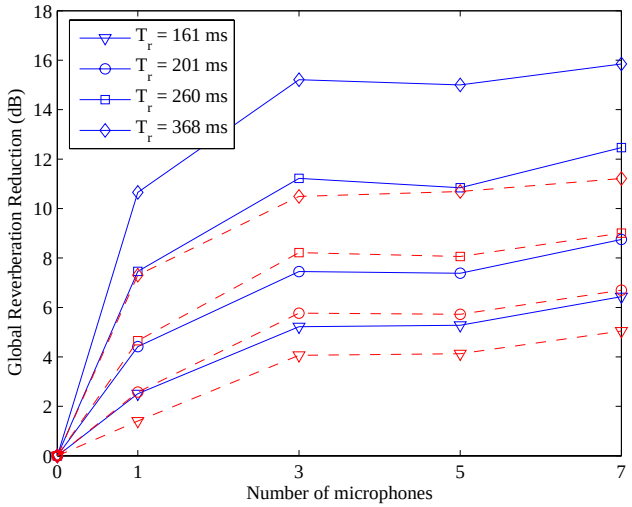

Fig. 3. Reverberation Reduction.

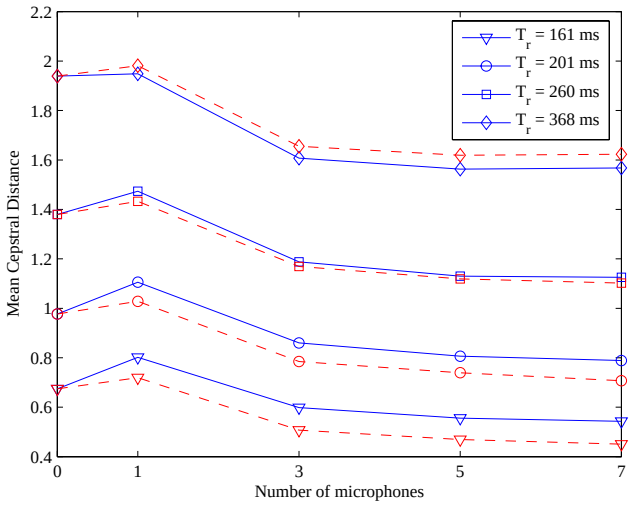

Fig. 4. Speech Distortion.

The results are available for listening on the following web page: *http://www.sps.ele.tue.nl/members/e.a.p.habets/icassp05/icassp05.html*

7. CONCLUSIONS

This paper presents a new multi-channel speech dereverberation algorithm based on a statistical model of late reverberation. We have shown how multiple microphone signals can be used to obtain an accurate estimate of the power spectrum of the late reverberant signal. Experimental results show a decrease in reverberation and distortion when using more microphones. Additionally, the fine structure of the speech signal is partially restored due to spatial averaging. Future work will focus on more accurate modeling of the RIR, loosening the assumptions w.r.t. the geometry of the microphone array and application in a real acoustic environment, rather than a simulated one.

8. ACKNOWLEDGEMENTS

The author expresses his thanks to STW and would like to thank Ir. J. v.d. Laar and his supervisor Dr. Ir. P.C.W. Sommen very much for carefully proofreading the manuscript.

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

[3] J.D. Polack, *La transmission de l'énergie sonore dans les salles*, Thèse de doctorat d'etat, Université du Maine, La mans, 1988.

[4] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE ICASSP'79*, vol. 4, pp. 208–211, 1979.

[5] J. Allen and L. Radiner, "A unified approach to short-time fourier analysis and synthesis," *IEEE Proceedings*, vol. 65, 1977.

[6] 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.

[7] A.H. Gray and J.D. Markel, "Distance measures for speech processing," *IEEE Transaction on acoustic, speech and signal processing*, vol. 24, no. 5, pp. 380–391, October 1976.