

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

Emanuël Habets

e.a.p.habets@tue.nl

Technische Universiteit Eindhoven, Dept. of Electrical Engineering,
P.O.Box 512, 5600 MB Eindhoven, The Netherlands

ABSTRACT

Speech signals recorded with a distant microphone usually contain reverberation and noise, which degrade the fidelity and intelligibility of speech, and the recognition performance of automatic speech recognition systems. In [1] we have presented a multi-channel speech dereverberation algorithm which reduces spectral coloration and late reverberation. In this paper we present experimental results obtained using microphone signals measured in a real acoustic environment. Additionally, we present a modification to reduce additive sensor noise which is present in the recorded microphone signals.

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 voices in a crowded room. We have investigated the application of signal processing techniques to improve the quality of speech 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. 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 of the reasons that 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 [1] we have introduced a multichannel speech dereverberation method based on Spectral Subtraction to reduce this effect. The described method estimates the Power Spectrum Density (PSD) of the reverberation based on Polack's statistical model of late reverberation. We have shown how this estimate can be produced using multiple microphone signals and that the fine-structure of the speech signal is partially restored due to spatial averaging of the received power spectra.

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

2. ROOM IMPULSE RESPONSE MODEL

Polack [2] 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

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

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

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

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, 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 $h_r(t)$:

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

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. SIGNAL MODEL

A noise-free reverberant signal $x(t)$ results from the convolution of an anechoic speech signal $s(t)$ and the causal RIR $h(t)$:

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

The n^{th} microphone signal, denoted by $z_n(t)$, contains uncorrelated additive sensor noise, i.e.

$$z_n(t) = x_n(t) + \nu_n(t).$$

Using the theory described in Section 2, we have shown in [1] that the spatially averaged auto-correlations of the noise-free signals, i.e. $r_{xx}(t, t+\tau) = 1/N \sum_{n=0}^{N-1} r_{x_n x_n}(t, t+\tau)$, 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. The spatially averaged auto-correlation at time t can be expressed as

$$r_{xx}(t, t+\tau) = r_{x_d x_d}(t, t+\tau) + r_{x_r x_r}(t, t+\tau), \quad (3)$$

with

$$r_{x_d x_d}(t, t+\tau) = e^{-2\alpha t} \int_{t-T}^t E_s \{s(\theta)s(\theta+\tau)\} \sigma^2 e^{2\alpha\theta} d\theta, \\ r_{x_r x_r}(t, t+\tau) = e^{-2\alpha T} r_{xx}(t-T, t-T+\tau). \quad (4)$$

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 (3) and (4) in terms of the short-term PSD's at time t and frequency f are approximately:

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

We can estimate the PSD of the direct signal by spectral subtraction of the late reverberant PSD which can be estimated from the sensor PSD's only. The spatially averaged short-term PSD's of the sensor signals, i.e. $\gamma_{zz}(t, f) = 1/N \sum_{n=0}^{N-1} \gamma_{z_n z_n}(t, f)$, can be expressed as

$$\gamma_{zz}(t, f) = \gamma_{xx}(t, f) + \gamma_{\nu\nu}(t, f).$$

In this paper we will use the delayed and attenuated version of $\gamma_{zz}(t, f)$ to estimate the short-term PSD of the late reverberant signal, i.e.

$$\gamma_{z_r z_r}(t, f) = e^{-2\alpha T} \gamma_{zz}(t-T, f) \\ = e^{-2\alpha T} (\gamma_{xx}(t-T, f) + \gamma_{\nu\nu}(t-T, f)). \quad (5)$$

Without additional processing the short-term PSD of the late reverberant signal will include a fraction of the noise equal to $e^{-2\alpha T} \gamma_{\nu\nu}(t-T, f)$. In case we would apply the dereverberation algorithm proposed in [1] the noise would be shaped in such a way that the residual noise becomes very annoying. Due to the additional noise reduction step this problem has been solved.

4. SHORT TIME SPECTRAL MODIFICATION

Numerous techniques for the enhancement of noisy speech degraded by uncorrelated additive noise have been proposed in the 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. Both the noise and reverberation reduction processes are based on a *Magnitude Subtraction* technique. These are expressed in terms of time and frequency dependent gain functions $G_\nu(t, f)$ and $G_r(t, f)$, respectively, and will be explained in Section 4.1 and 4.2, respectively. The influence of the noise term in (5) will be discussed in Section 4.3.

To reduce random variations in the estimated amplitude spectrum $A(t, f)$ and to ensure a proper relation with the estimated short-term PSD of the late reverberant signal we use the squared spatially averaged power spectra of the received microphone signals, i.e.

$$A(t, f) = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} |Z_n(t, f)|^2},$$

where N denotes the number of microphones. In case the direct source signals are properly aligned this will also result in a partial reconstruction of the fine structure of the speech signal. The phase spectrum is defined as

$$\phi_{ds}(t, f) = \arg \left\{ \frac{1}{N} \sum_{n=0}^{N-1} Z_n(t, f) \right\},$$

which is equal to the phase spectrum of a delay and sum beamformer with zero delay. The estimated clean speech spectrum is given according to the Spectral Subtraction method:

$$\hat{S}(t, f) = Q(t, f) e^{j\phi_{ds}(t, f)},$$

with

$$Q(t, f) \triangleq G_r(t, f) G_\nu(t, f) A(t, f). \quad (6)$$

4.1. Noise Reduction

Since we know a priori that $e^{-2\alpha T} \gamma_{\nu\nu}(t, f)$ can be reduced by the dereverberation process, to avoid over subtraction we may only suppress $(1 - e^{-2\alpha T}) \gamma_{\nu\nu}(t, f)$ using the noise reduction procedure. The partially de-noised amplitude spectrum is given by

$$Y(t, f) = G_\nu(t, f) A(t, f). \quad (7)$$

The gain function is defined as

$$G_\nu(t, f) = 1 - \frac{1}{\sqrt{SNR_{inst}(t, f) + 1}}, \quad (8)$$

where SNR_{inst} denotes the *instantaneous Signal to Noise Ratio* which is defined as

$$SNR_{inst}(t, f) \triangleq \frac{A^2(t, f)}{(1 - e^{-2\alpha T}) \gamma_{\nu\nu}(t, f)} - 1.$$

However, in all frames it is possible that for some frequencies the estimated amplitude of the noise spectrum is larger than the instantaneous amplitude of the noisy speech spectrum $A(t, f)$. Since this could lead to negative estimates for the amplitude of the partially de-noised speech spectrum $Y(t, f)$, the gain function $G_\nu(t, f)$ is usually put to zero or equal to a small noise floor value. However, because of the non-stationary character of the speech signal, this non-linear rectification leads to a specific kind of residual noise, called musical noise, which consists of short tones with randomly distributed frequencies. The residual noise problem is usually alleviated using two standard modifications. The first modification reduces the random variations due to the contribution of noise in $A(t, f)$ by averaging SNR_{inst} which results in an estimate of the *A Priori SNR* (SNR_{prio}). In practice we can estimate SNR_{prio} using a recursive average of SNR_{inst} :

$$SNR_{prio}(t, f) = \beta_n SNR_{prio}(t-1, f) \\ + (1 - \beta_n) \mathbf{P}[SNR_{inst}(t, f)],$$

where \mathbf{P} denotes half-wave rectification and β_n ($0 \leq \beta_n \leq 1$) denotes the forgetting factor. The second modification consists of using a threshold. Instead of setting the negative values of

$Y(t, f)$ to zero, the values of $Y(t, f)$ less than $\lambda_n A(t, f)$ are set to this value. Applying above modifications to the standard gain function in (8) results in the following gain function

$$G_\nu(t, f) = \begin{cases} 1 - \frac{1}{\sqrt{SNR_{prio}(t, f)+1}} & \text{if } Y(t, f) \geq \lambda_n A(t, f) \\ \lambda_n & \text{otherwise.} \end{cases}$$

4.2. Reverberation Reduction

The gain function used for the dereverberation process is given by

$$G_r(t, f) = \begin{cases} 1 - \frac{1}{\sqrt{SRR(t, f)}} & \text{if } Q(t, f) \geq \lambda_r Y(t, f) \\ \lambda_r & \text{otherwise,} \end{cases}$$

where λ_r denotes the threshold value defining the maximum amount of late reverberation and/or noise that is reduced. The *Signal to Reverberation Ratio* (SRR) is defined as

$$SRR(t, f) \triangleq \frac{Y^2(t, f)}{\gamma_{z_r z_r}(t, f)}.$$

4.3. Discussion

The amplitude spectrum which is subtracted by applying $G_r(t, f)$ in (6) is equal to

$$\sqrt{\hat{\gamma}_{z_r z_r}(t, f)} \approx \sqrt{\gamma_{x_r x_r}(t, f) + e^{-2\alpha T} \gamma_{\nu\nu}(t - T, f)}.$$

Because $\hat{\gamma}_{z_r z_r}(t, f)$ contains noise and late reverberation it will influence the reverberation reduction. In case $\hat{\gamma}_{z_r z_r}(t, f)$ contains more late reverberation than noise the subtracted amplitude spectrum will strongly depend on the amount of late reverberation, i.e. the amount of noise reduced by $G_r(t, f)$ is limited. However, in case $\hat{\gamma}_{z_r z_r}(t, f)$ contains more noise than late reverberation the subtracted amplitude will strongly depend on the amount of noise. The received microphone signal can be divided into 4 different time regions:

1. *Noise: The noise is reduced by G_ν and G_r .*
2. *Direct Speech + Noise: The noise is reduced by G_ν and G_r , note that $\gamma_{x_r x_r}$ is still zero.*
3. *Direct Speech + Late Reverberation + Noise: Part of the noise will be reduced by G_ν while G_r reduces late reverberation. We assume that the amount of additive sensor noise in this region is small compared to the late reverberation. Noise that is not canceled will therefore be partially masked by the speech signal.*
4. *Late Reverberation + Noise: Part of the noise is reduced by G_ν while the amount of reverberation and/or noise reduced by G_r now depends on their contribution to $\hat{\gamma}_{z_r z_r}$.*

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 1.

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 [3] from the estimated amplitude spectrum $Q(m, k)$ and the phase spectrum $\phi_{ds}(m, k)$.

Estimation of T_r and $\gamma_{z_r z_r}(m, k)$. In order to estimate the late reverberant PSD we need to estimate the reverberation time of the room. Partially blind and blind methods have been developed recently, c.f. [4]. For evaluation purposes we have used the average reverberation time measured directly from the RIR's using Schroeder's method. The RIR's were estimated using a Maximum Length Sequence measurement. The short-term PSD $\gamma_{z_r z_r}(m, k)$ is estimated by

$$\hat{\gamma}_{zz}(m, k) = \beta_r \hat{\gamma}_{zz}(m - 1, k) + \frac{1 - \beta_r}{N} \sum_{n=0}^{N-1} |Z_n(m, k)|^2,$$

$$\hat{\gamma}_{z_r z_r}(m, k) = e^{-2\alpha T} \hat{\gamma}_{zz}(m - T', k),$$

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

Estimation of $\gamma_{\nu\nu}(m, k)$. The PSD estimate of the noise can be obtained using a Voice Activity Detector in combination with a PSD estimation procedure. An alternative would be to use a noise PSD estimation based on optimal smoothing and minimum statistics procedure as proposed by R. Martin [5]. For evaluation purposes we have measured the average noise PSD using a standard periodogram averaging method.

Magnitude Spectral Subtraction. The discrete versions of the equations presented in Section 4 are used to obtain the estimate $Q(m, k)$. The thresholds λ_r and λ_n are set to 0.1 and the forgetting factor β_n to 0.9.

6. EVALUATION

The source signal consisted of an anechoic female voice of 8.5 seconds, which was played through a loudspeaker. The reverberant microphone signals were recorded using 5 omnidirectional microphones. The microphones formed an uniform linear array with an inner microphone distance of 10 cm, the source-receiver distance was 3 m. The dimensions of the room are 4.8 m x 4.6 m x 2.8 m (l x w x h). The reverberation time (T_r) was 298 ms.

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 (measured w.r.t. the arrival time of the direct sound) of the RIR, and with the RIR minus this part. While the complete noise-free reverberant signals were being processed, the time-varying, signal dependent gain function $G_r(t, f)$ was recorded. The recorded gain was then applied separately to the reverberant parts $r_{in}(n)$, giving $r_{out}(n)$. Because the reverberation reduction is calculated using the noise-free reverberant signals we will calculate an upper bound of the reverberation reduction.

Reverberation Reduction When no speech was present in the anechoic signal the Global Reverberation Reduction (GRR) was calculated using

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

