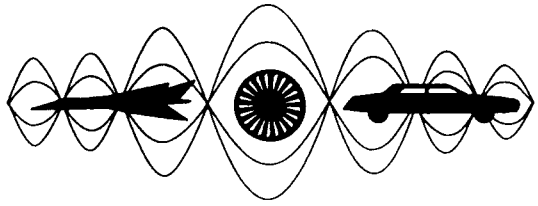

FIFTH INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
DECEMBER 15-18, 1997
ADELAIDE, SOUTH AUSTRALIA

BLIND DEREVERBERATION USING SHORT TIME CEPSTRUM FRAME SUBTRACTION

J.S. van Eeghem (1), T.Koike (2) and M.Tohyama (1)

(1) Kogakuin University, Institute of Computer Acoustics and Hearing, Tokyo, Japan
(2) NTT Advanced Technology, Tokyo, Japan

ABSTRACT

The authors present a blind dereverberation method that can be used to separate reverberant speech into an impulse response and a dry speech contribution. This method is based on the assumption that the contribution of an impulse response to reverberant speech varies slowly compared to that of the dry speech. Processing the reverberant signal for short-time frames and using the special properties of the cepstum domain, allows a recursive scheme to remove the equal impulse response contributions. Although short-time frames are processed, the effects of an amply longer impulse response can be separated from the dry speech. The assumption of a constant impulse response in only two frames and the iterative processing make the method inherently robust to changes in the impulse response. A drawback of our method is that the noise, related to the ambiguous linear phase addition when calculating the inverse cepstrum, requires additional processing.

INTRODUCTION

When the cepstrum domain contributions of a dry-speech signal and an unknown impuls response are divided between a low time and a high-time region, blind speech dereverberation was shown to be achievable [1]. In general the contributions to the magnitude cepstrum will

be spread over the entire cepstrum region. In this paper we will present a blind dereverberation method that is applicable to this more general type of impulse response.

BASIC IDEA

The transition from dry speech to reverberant speech can be represented by Fig. 1

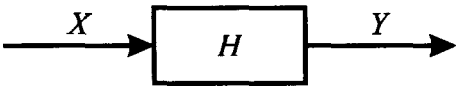

Figure 1. General system representation

where X, Y and H refer to the dry speech, the reverberant speech and a reverberant impulse response respectively. The relations between the in- and output of the system from Fig 1. for the time, frequency and cepstrum domain are given by

$$y(t) = h(t) * x(t) = \int h(T)x(t-T)dT \tag{1}$$

$$Y(\omega) = H(\omega) \cdot X(\omega) \tag{2}$$

$$C_y(\tau) = C_h(\tau) + C_x(\tau) \tag{3}$$

respectively. Where C_y represents the cepstrum of signal $y(t)$

$$C_y(\tau) = \Im^{-1} \ln \Im \; y(t) \tag{4}$$

and τ refers to the spatial variable of the cepstrum domain, the quefrency [s] (see also [3]), $\Im$ denotes taking the Fourier transform and *ln* refers to the natural logarithm. Before taking the natural logarithm of the frequency data

$$Y(\omega) = |Y(\omega)| \cdot e^{j \arg[Y(\omega)]} \tag{5}$$

the phase $\arg[Y(\omega)]$ is unwrapped [4] and its linear part removed [1]. This linear phase value is unambiguously connected to the signal $y(t)$, and is therefore required when transferring back to the time-domain

In general the contributions of the speech to the reverberant signal vary faster than those of the impulse response. Where a speech signal varies significantly in intervals longer than about 20 miliseconds, the impulse response varies much slower or is even constant. Therefor it is possible to subtract the nearly equal impulse response contributions of two consecutive time frames, and with pair wise processing, perform an on-line dereverberation. For the following derivation we use the following assumptions:

- the speech signal can be considered stationary within one frame.
- the TF contribution to the cepstra of two adjacent frames is nearly equal.

Using these assumptions together with Eq. 3 the effect of the reverberant impulse response can be partly canceled by simply subtracting the magnitude cepstra of adjacent frames.

$$\begin{aligned} C_{y(t-1)} &= C_{x(t-1)} + C_{h(t-1)} \\ C_{y(t)} &= C_{x(t)} + C_{h(t)} \quad ; \quad C_{h(t)} \approx C_{h(t-1)} \\ \Delta C_{y(t)} &= C_{x(t)} - C_{x(t-1)} \end{aligned} \tag{6}$$

The dry speech can now be obtained frame-wise from:

$$C_{x(t)} = \Delta C_{y(t)} + C_{x(i-1)} \tag{7}$$

Eq. 7 shows an recursive formula that uses only the previous frame cepstrum data.

The dry speech cepstrum $C_{x(0)}$ is unavailable as the initial condition. Instead an estimation for $C_{x(0)}$ is used being the first frame reverberant speech cepstrum

$$\hat{C}_{x(0} = C_{y(0)} \tag{8}$$

Now that $\hat{C}_{x(t)}$ is estimated, its convergence towards $C_{x(t)}$ is enabled by introducing a weighting for $\Delta C_{y(t)}$, α ($\alpha > 1$). This results in the following iterative estimation

$$\hat{C}_{x(t)} = \alpha \, \Delta C_{y(i)} + \hat{C}_{x(i-1)} \tag{9}$$

The recovered speech signal $\hat{x}(t)$ is obtained when transferring $\hat{C}_{x(t)}$ back to the time domain using Eq. 10.

$$\hat{x}(t) = \Im^{-1} \exp \Im \ \hat{C}(\tau) \tag{10}$$

Here the linear phase removed when calculating the cepstrum is added. However the linear phase matching the dry-speech signal is unavailable, and instead the linear phase removed from the reverberant signal is used. This unambiguous phase manipulation, will result in a phase error, causing the frames to not connect at the borders.

EXAMPLE

Now an example of the dereverberation of speech using the presented method is given.The reverberant speech results from the convolution between a sample of dry speech and an impulse response. The dry speech is performed by a Japanese woman and is sampled at 8000 Hz.

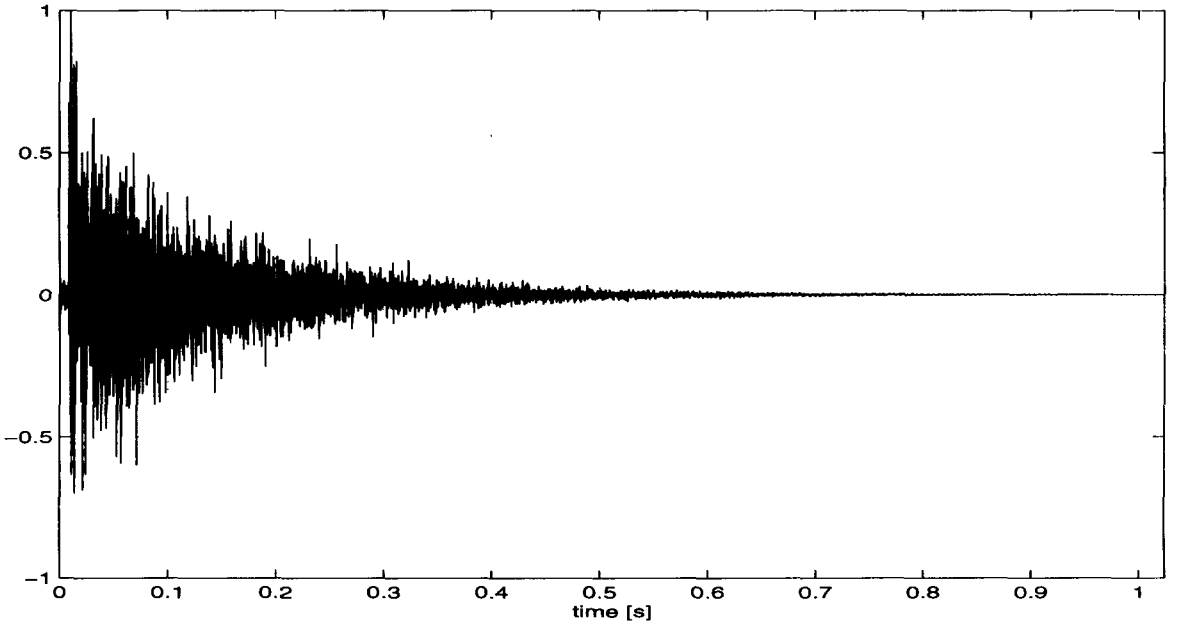

Figure 2. Impulse response data

The impulse response, printed in Figure 2, originates from a measurement inside a Japanese

concert hall and has an approximated reverberation time of 1 second. The impulse response is 8192 samples of length (about 1 second; sampled at 8000 Hz). Besides for building the reverberant speech, the impulse response data is not used for the dereverberation algorithm. The reverberant speech is processed in frames of 128 samples (0.016 second). No overlap or frame windowing is used. For this example the ratio between the frame length T_f and the impulse response length T_{ir} amounts

$$T_f / T_{ir} = 8192/128 = 64$$

For the weighting of the cepstrum (Eq. 9) difference $\alpha = 1.75$ is used. In Figure 2 respectively, the dry speech, the reverberant and the recovered speech are printed.

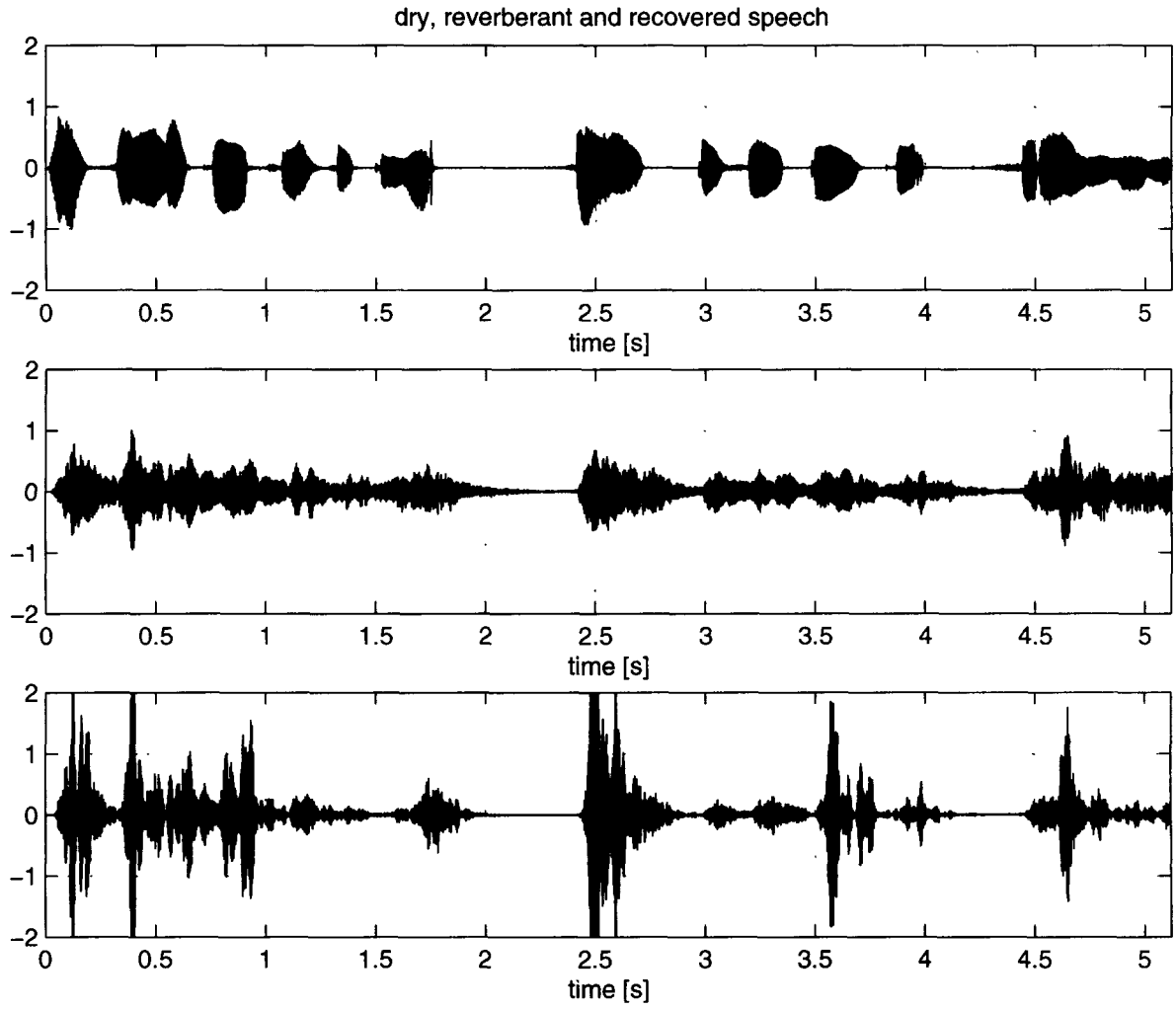

Figure 3.

First of all the recovered speech can be seen to have a time delay relative to the dry speech. As this delay represents the traveling time between source and receiver and one frame-length of sampling time, we can not expect to account for this delay without any preview knowledge. As discussed in the previous section, noise due to the unambiguous phase manipulation was to be expected. Next to the phase error, estimation errors in $\hat{C}_{x(i)}$ result in time-domain errors with a rather high peak levels. Remarkably the estimation of $\hat{C}_{x(i)}$ is robust for errors in the phase cepstrum as nearly the same dereverberation performance is achieved when estimating only the magnitude cepstrum of the dry-speech. Even if the phase cepstrum is disregarded, the speech is still audible. In spite of the errors the recovered speech is still clearly audible and can be heared to have a decreased reverberation. The dereverberation result can best be observed in the parts resulting from the silent parts of the original speech, where the reverberant signal has a decay similar to the impulse response.

CONCLUSION

Although short-time frames are processed, the presented method can successfully separate the effects of amply longer impulse responses from the reverberant speech. The presented method can be used to recover dry speech online is inherently robust for changes in the impulse response as it is only assumed constant in two adjacent frames. However additional processing is required, due to noise occurrence related to cepstrum estimation errors and unambiguous phase manipulation. For further research the investigation into the origins of the observed noise is recommended. Also the extension to processing longer time frames semi-online, to identify a the large size TF should provide performance enhancement.

REFERENCES

[1] A.V. Oppenheim . Digital Signal Processing, Prentice-Hall, 1975

[2] M. Tohyama. The Nature and Technology of Acoustic Space, Academic Press, 1995

[3] R.B. Randall. Frequency Analysis, Bruel and Kjaer, 1987

[4] J.M. Tribolet. A new phase unwrapping algorithm, IEEE, 1977