Proceedings of the Institute of Acoustics

DEREVERBERATION OF SPEECH BY POWER ENVELOPE INVERSE FILTERING AND PITCH EMPHASIS PROCESSING

M. Yamazaki(1), S. Hirobayashi(1), H. Kimura(1), M. Tohyama(2)

(1)Faculty of Engineering, Kanazawa University, Kanazawa-shi, Ishikawa, Japan
(2)Kogakuin University, Hachioji, Tokyo, Japan

ABSTRACT

Source waveform recovery under reverberant conditions is important for speech recognition and teleconference systems. We describe dereverberation of speech based on blind deconvolution by approximating reverberation time and statistical characteristics of speech without measuring a room transfer function. The main characteristic of reverberant speech is that the envelope rounds widely and fine spectral structure is distorted. We proposed that (a) power envelope inverse filtering is a global recovery technique, and (b) pitch emphasis processing can examine the harmonics of speech as a local technique. We recovered the source speech from reverberant speech using a sub-band power envelope inverse filter to approximate room transfer function with an exponential function. Furthermore, we reduced spectral distortion, and improved speech intelligibility by emphasising the harmonic structure of speech. As a result, the envelope distortion of speech recovered from reverberant speech was improved on bands of nearly 90% by power envelope processing without inverse filtering of the measured room transfer function. Furthermore, we restored the harmonic structure and improved the clearness by altering and emphasizing the pitch to recover the nature of speech.

1 INTRODUCTION

In speech recognition and teleconference systems, reflection of sound waves produces distortion on a source signal. Therefore, some inverse filtering methods that recover a source signal from an observed reverberant signal in an indoor space have been proposed. S. T. Neely and J. B. Allen have proposed the minimum phase inverse filter method[1] which removed only the minimum phase component of a room transfer function from the received signal with a single microphone and recovered the source signal. If the room transfer function is in the minimum phase, it can suppress the reverberant influence effectively. However, it is difficult to remove the influence of reverberation from an observed signal and to restore a source speech accurately with the minimum phase inverse filter because the influence of all-pass phase components on the transfer system increases along with the increment of reverberation time. In other words, the longer the reverberation time, the lower the recovery precision of a source speech because all-pass phase components can not be remove. On the other hand, M. Miyoshi and Y. Kaneda proposed inverse filtering of room acoustics that used a plural microphone[2]. Using microphones with 1 or more of source signals, even if the transfer system is not the minimum phase characteristic, a source signal can be restored accurately if the seros of the transfer characteristics that reach to each microphone from each source signal do not overlap. In addition, H. Wang and F. Itakura proposed and demonstrated a sub-band inverse filter theory of room transfer function[3]. For each transfer system between the sound signal and plural microphones, it obtains optimal inverse filters that minimize the error of the source signal and the recovery signal in every band and recovers the source signal from the reverberant signal using the filter.

These reverberation recovery techniques must measure the reverberant transfer function beforehand when the inverse filter is estimated. Furthermore, recovery precision decreases if the transfer function is not measured at each environmental variation and if inverse filtering of the transfer function is not performed accurately because the transfer function always varies even if the reverberant impulse response is measured. Accordingly, it is difficult to practically to perform inverse filtering using the measured transfer characteristic as the transfer system varies.

The possibility that a source signal could be recovered from a reverberant signal without accurate inverse filtering of a reverberant transfer function was suggested[4]. Reverberation influences on the envelope of the waveform and the spectral distortion. We proposed a technique to recover the source signal using a source signal and transfer system based on the theory of MTF (Modulation Transfer Function)[5] and examined the power envelope of the signal. Furthermore, we expanded this technique to a sub-band process and examined a method that reduces the influence of reverberation without measuring the transfer function.

The envelope information of the time axis direction was reduced in the inverse filtering by the power envelope, but the spectral distortion of the frequency axis direction was not removed. Therefore, we examined the pitch with information such as language, individuality and nature of a speech. We recovered the speech by altering and emphasizing the pitch after power envelope inverse filtering.

2 INFLUENCE ON SOURCE SPEECH BY REVERBERATION

Fig. 1 schematically shows the spectrogram of reverberant speech. Reverberation causes the envelope of source speech to decrease with time. The reverberant speech is indistinct. Reverberation is removed with power envelope inverse filtering. However, because we are approximating power envelope of reverberant impulse response with an exponential function in power envelope inverse filtering, the process is inadequate. Therefor, we considered that the frequency component other than harmonic structure in voiced parts (vowel) influenced reverberation. We altered the pitch, emphasized it and recovered the speech.

3 SOURCE SIGNAL AND OBSERVATION SIGNAL IN REVERBERANT CONDITION

As shown in Fig. 2, a reverberant signal (an observed signal) is described,

$$y(t) \;=\; \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \qquad (1)$$

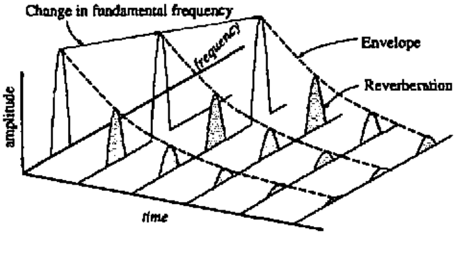

Fig. 1. Influence of a reverberation product on source speech.

$$= \;\; x(t) * h(t). \qquad (2)$$

where $x(t)$ is a source signal, $h(t)$ is an room impulse response and $*$ is the convolution product. Equation (1) is expressed in the frequency domain with

$$Y(z) = X(z)H(z). \qquad (3)$$

where $z = e^{j\omega T_s}$ it is (T_s is Sampling cycle). Accordingly, the source signal is expressed with

$$X(z) = Y(z)H^{-1}(z) = \frac{Y(z)}{H(z)}. \qquad (4)$$

In a linear system, estimation of the transfer system is necessary to obtain a source signal from an observation signal. However, source signal restoration by inverse filtering is difficult practically because the transfer function varies and requires adaptable estimation.

Therefore, we considered that the envelope of a reverberant impulse response is attenuated about the exponential function.

3.1 POWER ENVELOPE INVERSE FILTERING

In this paper, the envelope of room impulse response was modeled with an exponential function. We restored a source signal from an observed reverberant signal by envelope restoration processing without measuring the transfer function. The predicted theory of speech legibility by MTF [5] evaluates the distortion of a power envelope of speech.

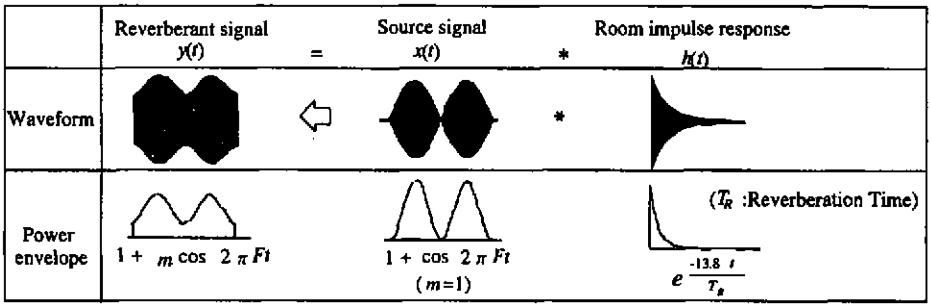

Fig. 3. Source and reverberant signals for a modulated noise signal of the modulation index m.

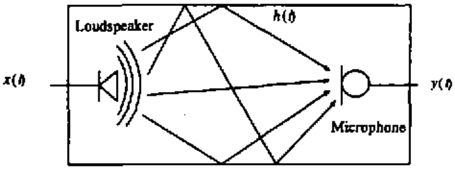

Fig. 2. Loudspeaker and microphone locations in an indoor space.

If distortion of the power envelope depends on reverberation, we can obtain MTF from the room impulse response[6].

Under reverberant conditions, the modulation factor (m=1) of the signal that modulates noise 100% with a sine wave, decreases in an observation point due to the influence of reverberation as shown in Fig. 3. The index m is called MTF, and it is the function of the center frequency and modulation frequency F of the source signal. It has been confirmed that the index represents important physical properties for predicting speech intelligibility scores[7].

Based on the MTF, speech signals can be modeled by

$$x(t) = e_a(t)n_1(t) \tag{5}$$

where $n_1(t)$ is white noise, $e_a(t)$ is an envelope. Similarly, room impulse response is represented as follows:

$$\begin{aligned} h(t) &= e_h(t)n_2(t) \\ e_h(t) &= ae^{\frac{-6.9t}{T_R}} \end{aligned} \tag{6}$$

where a is a constant, $T_R(s)$ is reverberation time, and $n_2(t)$ is white noise. Signals $n_1(t)$ and $n_2(t)$ demonstrate the following relationship:

$$< n_k(t)n_k(t+\tau) >= \delta(\tau). \tag{7}$$

$< * >$ indicates the average for a given set ($k = 1, 2$). Here, we obtain the power envelope from reverberant speech which is shown with the convolution of equations (5) and (6) from equation (1). From equation (1), (5) ~ (7), square set average of a reverberant speech is

$$\begin{aligned} < y(t)^2 > &= < \{\int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau\}^2 > \\ &= \int_{-\infty}^{\infty} d\tau_1 \int_{-\infty}^{\infty} d\tau_2 e_a(\tau_1)e_a(\tau_2) \\ &\quad \times e_h(t-\tau_1)e_h(t-\tau_2) \\ &\quad \times < n_1(\tau_1)n_1(\tau_2) > \\ &\quad \times < n_2(t-\tau_1)n_2(t-\tau_2) > \\ &= \int_{-\infty}^{\infty} e_a(\tau)^2 e_h(t-\tau)^2 d\tau \\ &= e_a(t)^2 * e_h(t)^2 \end{aligned} \tag{8}$$

In other words, the power envelope of a reverberant signal $y(t)$ is shown with the convolution of each power envelope of a source signal $x(t)$ and room impulse response $h(t)$. Accordingly, we can recover a power envelope of speech using an inverse characteristic of a power envelope of room impulse response. From equation (6),

$$e_h(t)^2 = a^2 e^{\frac{-13.8t}{T_R}}. \qquad (9)$$

If a power envelope of an room impulse response is approximated with an exponential function, we can obtain a power envelope inverse filter that has the inverse characteristics of reverberation time (T_R). If the sampling cycle is T_s(s),a power envelope transfer function $P_h(z)$ for equation (9) on $t > 0$ is

$$\begin{aligned} P_h(z) &= a^2 + a^2\alpha z^{-1} + a^2\alpha^2 z^{-2} + a^2\alpha^3 z^{-3} + \cdots \\ &= \frac{a^2}{1-\alpha z^{-1}} \qquad \text{Yet, } \alpha = e^{\frac{-13.8T_s}{T_R}} \qquad (10) \end{aligned}$$

Accordingly, the power envelope characteristic of speech $P_s(z)$ is

$$P_s(z) = \frac{P_y(z)}{P_h(z)} = \frac{1}{a^2}(1-\alpha z^{-1})P_y(z) \qquad (11)$$

It is expressed with the product of the power envelope characteristic of a reverberant speech $P_y(z)$ and an inverse characteristic of the power envelope transfer function $P_h(z)$. $P_s(z)$ and $P_y(z)$ are z conversions of $e_s(t)^2$ and $e_y(t)^2$. Using the positive square root (in other words, amplitude envelope) of $e_a(t)^2$ and $e_y(t)^2$, recovery speech $\hat{x}(t)$ is

$$\hat{x}(t) = e_a(t) \times \frac{y(t)}{e_y(t)} \qquad (12)$$

[8, 9]. The right side is the recovery envelope obtained from the inverse z conversion of equation (11). In addition, $y(t)/e_y(t)$ is the fine structure of reverberant speech, it is the flattened amplitude envelope of the output signal. Equation (12) shows that the estimated amplitude envelope is added to the output signal that flattened an amplitude envelope. Furthermore, the explanation reflects continuous time system, but the experiments were performed in a separation time system. The possibility of reverberant speech recovery was explained using envelope information of reverberant speech without considering the local influence of spectral distortion. We called equation (12) power envelope inverse filtering.

3.2 SUB-BAND POWER ENVELOPE INVERSE FILTERING

In section 3.1, we recovered envelope information using power envelope inverse filtering by approximating the power envelope of reverberant speech with 1 exponential function. However, the envelopes of several bands of the source speech were different. Therefore, the spectral distortion which could not be recovered by the process in section 3.1 may be recovered by recovering several power envelopes of several bands. A block diagram of sub-band power envelope inverse filtering is shown in Fig. 4. However, we used CQF to satisfied the reconstruction filter bank[10].

4 RECOVERY SPEECH BY PITCH EMPHASIS PROCESSING

We approximated the envelope of a reverberant impulse response with an exponential function in the power envelope inverse filtering. Furthermore, we assumed that the fine structure signal of source speech was not correlated and applied it to reverberant speech. We consider that recovery by power envelope inverse filtering is possible. However compared to a theoretical noise model, it has little effect. The main cause was the spectral distortion due to the influence of reverberation. As shown in Fig. 5, it was not completely removed by power envelope inverse filtering. We reconstructed the harmonic structure of speech to remove the spectral distortion.

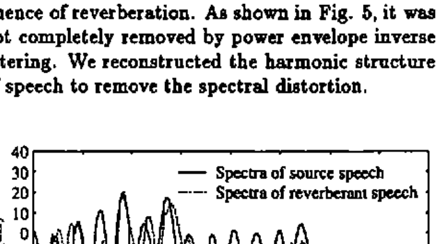

Fig. 5. Spectral distortions of voiced source and reverberant speeches.

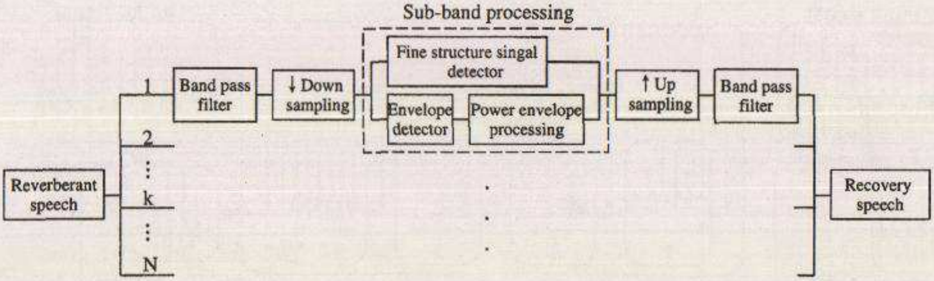

Fig. 4. Block diagram of sub-band power envelope inverse filtering (The number of separated bands is N).

Fig. 6 shows the procedure for the pitch emphasis processing. In reconstructing the harmonic structure of speech, we cut out a frame waveform from speech using hamming windows of 32 (ms), altered the pitch, and estimated fundamental frequency. We defined evaluation function $E(f)$ for the estimated fundamental frequency f and its harmonic components.

$$E(f) = \sum_{k=1}^{N} \log(\hat{X}(kf)) \quad (\text{Yet}, N = 5) \qquad (13)$$

We considered 5 harmonic components of estimated fundamental frequency. We detected the fundamental frequency f for the maximum value of E, changed it and the all harmonic components from the frequency f to a sampling frequency of the speech, and synthesized the speech using hamming windows. The processing is referred to as pitch emphasis processing in this paper.

5 EXPERIMENT

The waveform and spectrogram of source speech are shown in Fig. 7, and those of reverberant speech are shown in Fig. 8. Its reverberation time is 1 (s). Pitch of the source signal was evident. In reverberant speech, spectra of the past frame overlap the future frame and the pitch is buried in the reverberation. Furthermore, the change in fundamental frequency was not be confirmed.

We removed the reverberation by sub-band power envelope inverse filtering in section 5.1, and emphasized the pitch by pitch emphasis processing in section 5.2.

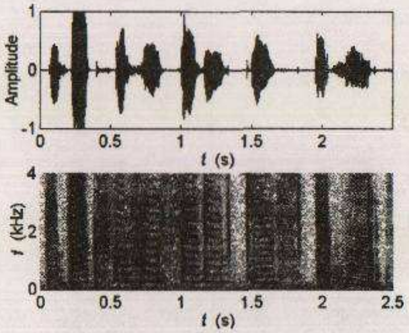

Fig. 7. Wave form and spectrogram of source signal.

5.1 EXPERIMENT OF SUB-BAND POWER ENVELOPE INVERSE FILTERING

Narrow band separation of the speech may recover to a fine structure signal. On the other hand, a power envelope of reverberation is very different from an

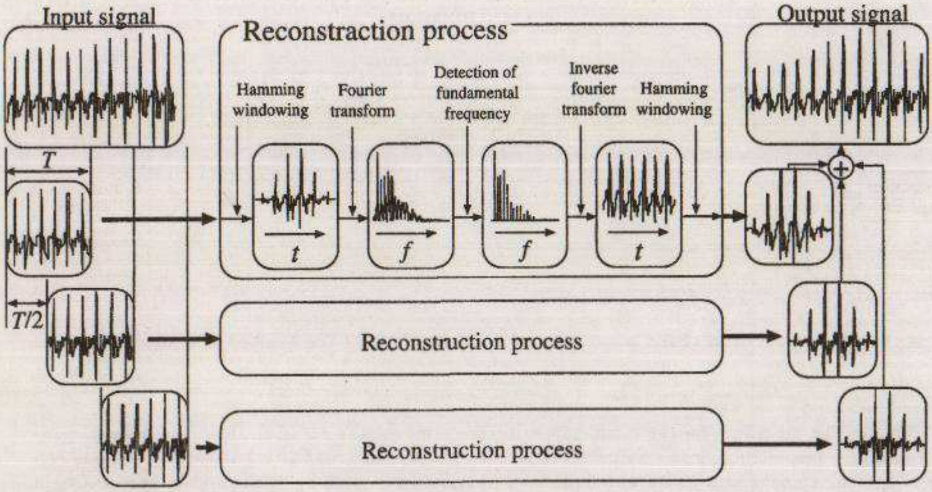

Fig. 6. Procedure of pitch emphasis processing (T=32 (ms)).

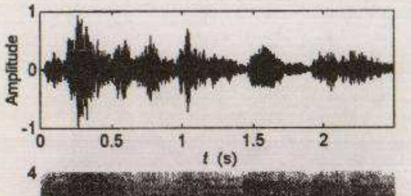

Fig. 8. Wave form and spectrogram of reverberant signal.

exponential function. Therefore, we recovered the speech by sub-band power envelope inverse filtering for many separated bands. We defined the improvement index of power envelope distortion and evaluated the recovery degree.

$$I_p = 10\log_{10}\frac{\int_0^T \{e_x(t)^2 - e_y(t)^2\}^2 dt}{\int_0^T \{e_x(t)^2 - \hat{e}_x(t)^2\}^2 dt}(\mathrm{dB}) \qquad (14)$$

where $\hat{e}_x(t)^2$ is the power envelope of the recovery speech, T is analysis time (2 (s)). The index value increased as the envelope of a reverberant signal recovered. If there was no improvement effect it was zero. At this time, we recovered the reverberant speech ($T_R = 1(s)$) for some separation numbers and calculated improvement indexes of power envelope distortion. The average is shown in Fig. 9. The speech improved most in 32 separation.

We recovered the speech by sub-band power envelope inverse filtering. The improvement index of power envelope distortion is shown in Fig. 10. The index was negative in some bands, but the envelope improved on bands of nearly 90% and the mean in-

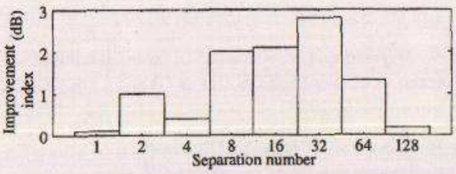

Fig. 9. Average of improvement index of power envelope distortion for some numbers of separated band.

dex became 2.8 (dB). Reverberation was effectively reduced.

The waveform and spectrogram of recovery speech are shown in Fig. 11. The influence of the reverberation from the past frame to the future frame was confirmed. However, the pitch was not periodic and did not recovered to a fine structure signal when the harmonic structure to the frequency axis direction was considered.

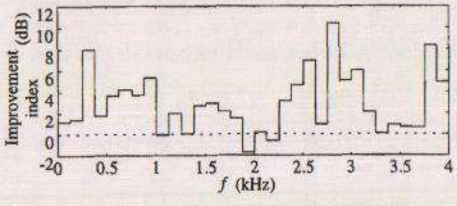

Fig. 10. Improvement index of power envelope distortion for each band.

5.2 EXPERIMENT OF PITCH EMPHASIS PROCESSING

Waveform and spectrogram of recovery speech by the pitch emphasis processing are shown in Fig. 12. The pitch of the recovery speech became marked, and recovered to a fine structure signal.

We also compared the change in spectra before and after the process regarding the frame as shown in

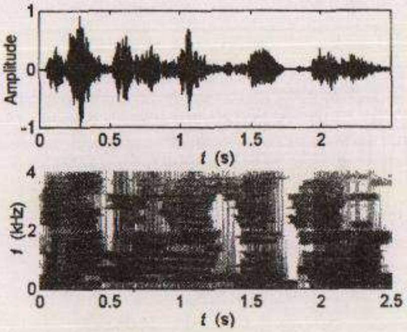

Fig. 11. Wave form and spectrogram of recovery speech by the sub-band power envelope inverse filtering.

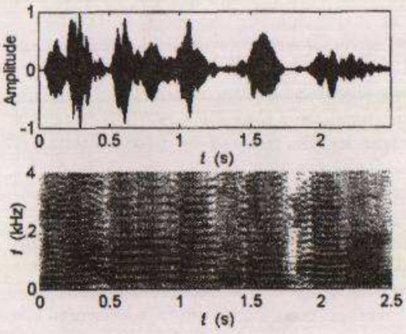

Fig. 12. Wave form and spectrogram of recovery speech by pitch emphasis processing.

Fig. 13. Before the process, spectral distortion occurred during the pitch. However after the process, it was removed and the pitch became marked.

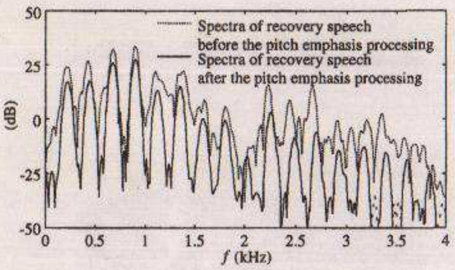

Fig. 13. Spectral distortions of voiced part of recovery speech before and after pitch emphasis processing.

6 CONCLUSION

We examined power envelope inverse filtering that used a reverberation time only without measuring the room transfer function and the pitch emphasis processing that emphasized pitch and included information such as language, individuality and nature of a speech. We showed that the power envelope of an observed reverberant signal is expressed by the convolution of each power envelope of a source signal and a room transfer function. Furthermore, we showed that the power envelope of a source signal can be restored by inverse filtering of an exponential function as a power envelope of a room transfer function. In addition, we demonstrated that pitch emphasis processing emphasized pitch and included information concerning language, individuality and nature of a speech. In experiments to recover source speech, we recovered speech by reducing reverberation. Furthermore, we restored the harmonic structure and recovered fine structure signals.

7 REFERENCES

[1] S. T. Neely and J. B. Allen,"Invertibility of a room impulse response," J. Acoust. Soc. Am, vol. 66, no. 1, pp. 165-169, 1979

[2] M. Miyoshi and Y. Kaneda,"Inverse filtering of room acoustics," IEEE Trans. ASSP. vol. 36, pp. 145-152, 1988

[3] H. Wang and F. Itakura,"Realization of acoustic inverse filtering through multi-microphone sub-band processing," IEICE Jpn, Trans. vol. E75A, no. 11, pp. 1474-1483, 1992

[4] S. Hirobayashi, T. Koike, M. Tohyama and R. Lyon,"Source waveform recovery from reverberant minimum phase component and power envelope," Technical Report of IEICE. EA93-81, Dec. 1993

[5] T. Houtgast, H. J. M. Steeneken and R. Plomp, "Predicting speech intelligibility in room acoustics," ACUSTICA, vol. 46, pp. 60-72, 1980

[6] M. R. Schroeder,"Modulation transfer functions: definition and measurement," ACUSTICA, vol. 49, pp. 179-182, 1981

[7] H. Nomura, H. Miyata, and T. Houtgast,"Speech Intelligibility and Modulation Transfer Function in Non-exponential Decay Fields," ACUSTICA, vol. 69, pp. 151-155, 1989

[8] R. Drullman, J. M. Festen, and R. Plomp,"Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am, vol. 95, no. 5. pp. 2670-2680, May, 1994

[9] R. Drullman,"Temporal envelope and fine structure cues for speech intelligibity," J. Acoust. Soc. Am, vol. 97, no. 1. pp. 585-592, Jan, 1995

[10] Boaz Porat,"A course in digital signal processing," John Wiley & Sons. Inc, pp. 489-492, 1997

[11] H. -L. Nguyen Thi and C. Jutten, "Blind source separation for convolutive mixtures," Signal Processing, vol. 45, pp. 209-229, 1995