Forensic Speech Enhancement Based On Two-Dimensional Fractional Fourier Transform Domain

[1]Wei Zhong, [2]Xiangwei Kong, [3]Xingang You, [4]Bo Wang
*[1,2,3, 4,] School of Information and Communication Engineering, Dalian University of Technology,
E-mail: zww110221@163.com , kongxw@dlut.edu.cn, youxg@dlut.edu.cn, bowang@dlut.edu.cn

Abstract

Forensic speech usually suffers from noise, distortion, interfering sounds, and other signal processing challenges that can impede proper analysis. The most common enhancement issue for forensic speech only could remove amplitude spectrum of background noise from the record. This paper proposes a new two-dimensional speech enhancement method which adopts parallel Wiener filter array in fractional Fourier transform domain to remove amplitude spectrum and phase spectrum of background noise. Theoretical analysis and experiments are conducted and the experiment shows that it has significantly superior Signal-to-Noise ratio over wiener filtering under the colored noise and white noise. It has wide application prospects in forensic speech recognition and speaker recognition.

Keywords: *Digital Forensics, Speech Enhancement, Fractional Fourier Transform*

1. Introduction

Over the past 40 years, the field of speech forensics [1] involves many topics such as speech recognition, speaker identification, and speech enhancement. R. C. Maher discussed the progress of the audio forensics technology in 2009 [2], In the same year, J. P. Campbell and others analysis and discussed the forensic speaker recognition [3].The two papers summarize audio forensics of the past forty years which also promote the further development of forensic speech research and point out the forensic speech enhancement should not only remove background noise but also preserve underlying phoneme. However, current technology of speech enhancement [4] only could reduce the amplitude spectrum of background noise but the phase spectrum of noise was reserved usually[5].

Speech enhancement in forensic audio recordings remains a challenging task. Forensic audio signals typically suffer from noise, distortion, even artificial manipulation, all of which impede effective proper analysis. Poor placement of the recording device and wind interference are typical examples of interference. Most forensic speech enhancement only aimed at amplitude spectrum of the noise. Y.Lu, S.D.Kamath and P.C.Loizou proposed the enhancement techniques of spectral subtraction [6][7]. Dynamic time warping and wavelet theory have been used in the forensic speech enhancement [8] to improve the processing results.

V.Namias calculated the Fractional Fourier Transform (FRFT) by using the Hermite polynomial in 1980, which put forward a definition of FRFT for the first time [9]. L.B.Almeida, H.M.Qzatkas [10] etc. found out that the FRFT of signal of α order corresponds to $\alpha\pi/2$ angle in its time-frequency surface. By selecting the appropriate fractional factor, we can achieve the best separation of speech and noise in the fractional Fourier transform which correspond to reducing the amplitude and phase spectrum of noise, Therefore we could accumulated enhanced speech in the fractional Fourier transform of multiple α to obtain the purpose of reducing the amplitude and phase spectrum of noise coinstantaneously for best α was different in every speech frame.

The paper will first discuss the feasibility of reducing the phase noise in the fractional Fourier transform domain. Then it'll discuss the theoretical limit in noise reduction in fractional Fourier transform domain. Finally, it proposes a two-dimensional fractional Fourier transform scheme for speech enhancement.

2. Two-dimensional filtering algorithm in Fractional Fourier transform domain

Due to the fact the recording of speech evidence is often obtained in a poor recording environment and often the speakers are in a movement position while the recording is taking place. So Interference and fading are usually time-varying and non-stationary which could not be eliminated directly from the frequency domain. The introduction of fractional Fourier transform theory provides a new and flexible way to achieve the best voice and interference noise separation, and to further reduce the fading effects.

2.1. Analysis of phase noise in Fractional Fourier transform domain

Assume the recorded signal $x(t)$ composed of source signals $s(t)$ and additive background noise $n_b(t)$:

$$x(t) = s(t) * h(t) + n_b(t) \tag{1}$$

Currently, most noise reduction algorithms transform x(t) into other domain, such as Fourier transform domain. And it is assumed that the noise is stationary noise to obtain the estimation of the amplitude spectrum of voice multiplied by the phase spectrum of noise and voice, and then inverse transform them to time domain to obtain the enhanced speech.

Let $\theta_x(\omega) = \theta_s(\omega) + \Delta\theta(\omega)$,

$$\hat{S}(\omega) = |\hat{S}(\omega)| \exp(-j\theta_x(\omega)) = |\hat{S}(\omega)| \exp(-j\theta_s(\omega)) \exp(-j\Delta\theta(\omega)) \tag{2}$$

For the sake of simplicity, assume that the amplitude spectrum noise reduction algorithm can produce the estimate consistent with the original clean speech. That is $|\hat{S}(\omega)| = |S(\omega)|$, so that:

$\hat{S}(\omega) = S(\omega)\exp(-j\Delta\theta(\omega))$. $\theta_x(\omega)$, $\theta_s(\omega)$ is part of the noisy speech phase spectrum of clean speech. $\Delta\theta(\omega)$ is the interference and noise caused by phase error. Applying inverse Fourier transform:

$$\Rightarrow \hat{s}(t) * IDFT(\exp(-j\Delta\theta(\omega))) \Rightarrow \hat{s}(t - \tau(t)) \tag{3}$$

$\tau(t)$ is the delay, equation (1), (3)shows that due to channel characteristics and noise , not only does the amplitude spectrum of speech signal estimated exist errors, but also the noisy phase spectrum introduces random delay in the time-domain waveform. Therefore, It should reduce the amplitude spectrum and phase spectrum of noise.

The fractional Fourier transform algorithm defined by Namias [7] is represented by equation:

$$S_a(u) = F^a(s(t)) = \begin{cases} \sqrt{\dfrac{1 - j\cot a}{2\pi}} \displaystyle\int_{-\infty}^{+\infty} \exp\left(j\frac{u^2 + t^2}{2}\cot a - \frac{jut}{\sin a} \right) s(t)dt & a \neq n\pi \\ s(t) & a = 2n\pi \\ s(-t) & a = (2n \pm 1)\pi \end{cases} \tag{4}$$

where $a = p\pi/2, 0 < |p| < 2$.

To obtain the fractional Fourier transform of $\hat{S}(\omega)$:

$$\hat{S}_{a+\frac{\pi}{2}}(u) = \sqrt{\frac{1 - j\cot a}{2\pi}} \int_{-\infty}^{+\infty} \exp\left(j\frac{(u^2 + \omega^2)\cos a - 2u\omega}{2\sin a} \right) \hat{S}(\omega)d\omega \tag{5}$$

Substitute (3) into (5) and denote $\Delta\theta(\omega) = \delta\omega$, δ is the random variable with a small mean value：

$$\hat{S}_{a+\frac{\pi}{2}}(u) = \sqrt{\frac{1 - j\cot a}{2\pi}} \int_{-\infty}^{+\infty} \exp\left(j\frac{(u^2 + \omega^2)\cos a - 2u\omega}{2\sin a} \right) S(\omega)\exp(-j\delta\omega)d\omega \tag{6}$$

Let $u' = u + \frac{\delta \sin a}{2}$, therefore :

$$\hat{S}_{a+\frac{\pi}{2}}(u) = \exp\left(j\left(\frac{\delta^2 \sin a \cos a}{8} - \frac{u'\delta \cos a}{2} \right) \right) \sqrt{\frac{1 - j\cot a}{2\pi}} \int_{-\infty}^{+\infty} \exp\left(j \frac{(u'^2 + \omega^2)\cos a - 2u'\omega}{2\sin a} \right) S(\omega) d\omega \tag{7}$$

According to the superposition property of the fractional Fourier transform,

$$\hat{S}_{a+\frac{\pi}{2}}(u) = \exp\left(j\left(-\frac{\delta^2 \sin 2a}{16} - \frac{u\delta \cos a}{2} \right) \right) S_{a+\frac{\pi}{2}}\left(u + \frac{\delta \sin a}{2}\right) \tag{8}$$

Comparing with equation (2) (3), we can conclude: the frequency domain phase noise $\Delta\theta(\omega)$ in a fractional Fourier domain is transferred into amplitude spectrum. This indicates phase spectrum and amplitude spectrum in fractional Fourier transform domain are interchangeable. So we could reduce phase spectrum of noise by reducing the amplitude spectrum in fractional Fourier transform.

2.2. Speech enhancement in two-dimensional Fractional Fourier transform domain

It has been shown in the last section that amplitude spectrum and phase spectrum of a recorded signal in the fractional Fourier transform domain are interchanged. Therefore phase noise elimination can be achieved in the time domain by the reduction of amplitude spectrum in fractional Fourier domain.

Consider time-domain stationary random noise $n(t)$ with mean δ_0 .

Applying the fractional Fourier transform, there are:

$$E[N(u)] = E\left[\sqrt{\frac{1 - j\cot a}{2\pi}} \int_{-\infty}^{+\infty} \exp\left(j \frac{(u^2 + t^2)\cos a - 2ut}{2\sin a} \right) n(t) dt \right] \tag{9}$$

Substituting $E[n(t)] = \delta_0$ into the above equation, it follows

$$E[N(u)] = \delta_0 \sqrt{\frac{1 - j\cot a}{2\pi}} \int_{-\infty}^{+\infty} \exp\left(j \frac{(u^2 + t^2)\cos a - 2ut}{2\sin a} \right) dt$$

$$\Rightarrow | E[N(u)] | = \delta_0 \tag{10}$$

Equation (9), (10) shows a stationary time-domain signal $n(t)$ is no longer a stationary random process after the fractional Fourier transform, but its amplitude spectrum of the signal remains stationary. It is feasible that complicated algorithm may be applied to the forensic speech enhancement, because that processing is not necessarily real-time.

With the foregoing analysis, speech enhancement can be modeled as follows:

The source signal $s(t)$ was assumed as speech sequence, $a_1, a_2, ..., a_M$ is the multiple α order of fractional Fourier transform. According to the cycle characteristic of the fractional Fourier transform, $-a_1, -a_2, ..., -a_M$ order of fractional Fourier transform is the inverse transformation of $a_1, a_2, ..., a_M$ order one.

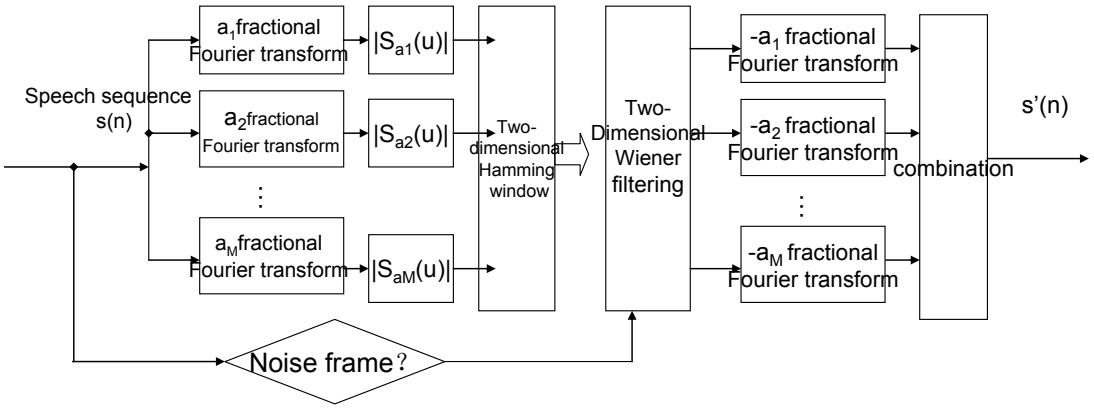

Figure 1. Two-dimensional fractional Fourier transform domain model for speech enhancement

Figure 1 shows the general model of speech enhancement based on fractional Fourier transform. Input speech sequence are framed with size N=256, 50% frame overlapping, After going through parallel number $M(M=16)$ of different fractional Fourier transform order, two-dimensional Hamming windowing operation is applied. The order factors for the fractional Fourier transform are $a_i \in (0,2\pi)$, $a_i(i=1,2,...,M)$. For simplicity, analysis can be made:

$$a_i = a_0 + i\Delta a \qquad i=1,2,...,M, \tag{11}$$

In equation (8) a_0 is the initiation factors, factors Δa is the fractional interval.

Two-dimensional Hamming window is defined as:

$$h(i,j)=\left(0.54-0.46\cos\left(\frac{2\pi i}{255}\right)\right)\left(0.54-0.46\cos\left(\frac{2\pi j}{15}\right)\right) \tag{12}$$

In which $i \in [0,255]$, $j \in [0,15]$

The outcome from the algorithm is shown below:

$$\hat{s}(t)=\frac{1}{M}\sum_{i=1}^{M}F^{-a_i}\left(\hat{S}_i(u)\right) \tag{13}$$

In the equation, $\hat{S}_i(u)$ is estimated speech of a_i -order fractional Fourier transform. Following the assumptions:

$$\begin{aligned}\hat{S}_i(u) &= |S_i(u)|\exp(-j\varphi_x(i,u)) \\ &= |S_i(u)|\exp(-j\varphi_s(i,u))\exp(-j\Delta\theta_i(u)) = S_i(u)\exp(-j\Delta\theta_i(u))\end{aligned} \tag{14}$$

In the equation, $\varphi_x(i,u)$、$\varphi_s(i,u)$ stand for a_i order noisy speech, clean speech phase spectrum in fractional Fourier transform, $\Delta\theta_i(u)$ is phase error,

Two-dimensional fractional Fourier transform domain filtering is derived as follows:

$$|\hat{S}(i,j)|^2 = |X(i,j)|^2 - |\hat{N}(i,j)|^2 = |X(i,j)|^2\left(1-\frac{|\hat{N}(i,j)|^2}{|X(i,j)|^2}\right) \tag{15}$$

Two-dimensional Wiener filter can be obtained as follows:

$$\begin{aligned}\hat{S}(i,j) &= X(i,j)H(u,v) \\ &= X(i,j)\sqrt{\max\left(1-\frac{1}{SNR(u,v)},\beta\right)}\end{aligned} \tag{16}$$

$\beta \geq 0$ to ensure $H(u,v)>0$.

$$SNR(u,v) = \frac{|\overline{X}(u,v)|^2}{|\overline{N}(u,v)|^2} \tag{17}$$

On which:

$$|\overline{X}(u,v)| = \sum_{i=v-1}^{v+1}\sum_{j=u-1}^{j=u+1}|X(i,j)| \quad |\overline{N}(u,v)| = \sum_{i=v-1}^{v+1}\sum_{j=u-1}^{j=u+1}|N(i,j)|$$

(18)

Equation (16)(17)(18) are the two-dimensional Wiener filtering process in the fractional Fourier transform domain. The noise spectrum $|\overline{N(u,v)}|$ was obtained by spectrum statistics of forward no-speech frame and the noised speech spectrum $|\overline{X(u,v)}|$ was obtained by spectrum statistics of current and forward speech frame.

3. Experiment and Analysis

In order to evaluate the performance of the two-dimensional fractional Fourier transform domain model for speech enhancement. Different noise levels on the voice recording are compared to each other.

All voices used in the experiments were recorded from young men and women. The sampling frequency is 8kHz while the quantization precision is 8-bit. Superimposed on the original clean speech is Gaussian white noise and non-stationary noise supplied by Voice Research Centre RSRE of the Netherlands belongs.

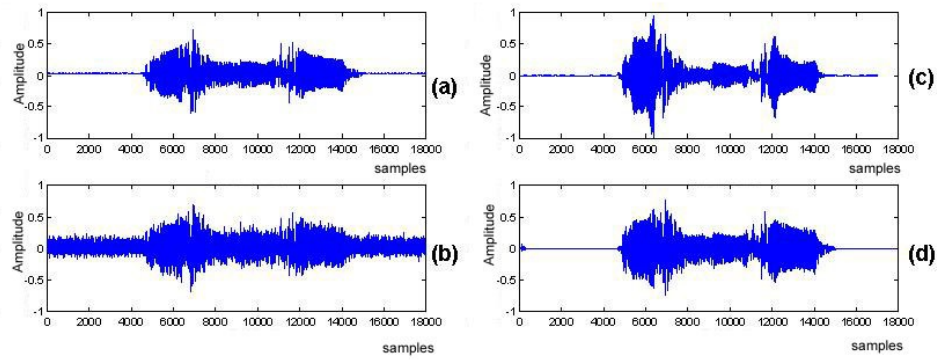

Figure 2. Wiener filter in two-dimension fractional domain compared with the normal filter
(a) pure speech; (b) 5dB noise speech(AGWN); (c) Ordinary Wiener [11];
(d) Wiener filter in two-dimension fractional domain

As it can be seen from the figure, the performance that fractional Fourier transform domain for two-dimension Wiener filtering to Gaussian white noise filtering is superior to ordinary Wiener filter [9] (3.19dB improvement when input SNR is 0dB;2.73dB improvement when input SNR is 5dB).

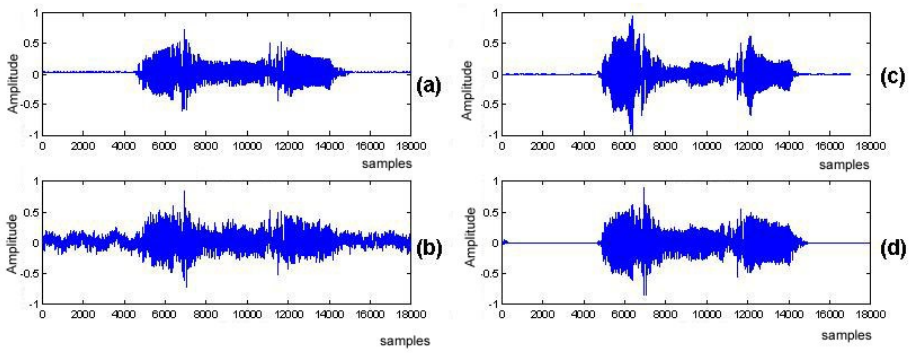

Figure 3. Two-dimensional wiener filter in the fractional Fourier transform compared with the normal filter
(a) pure speech; (b) 5dB noise speech(pink noise); (c) Ordinary Wiener [11];
(d) two dimension fractional Fourier transform domain filtering

As can be seen from Figure 3, the general Wiener filtering method has some distortion in the voice segment while the performance of two-dimensional wiener filter in the fractional Fourier transform domain to pink noise is superior to ordinary Wiener filter [11].

4. Conclusions

In this paper, we have proposed the new two-dimensional methods for forensic speech enhancement which adopts parallel Wiener filter array in fractional Fourier transform domain. Our proposed methods can generate clean speech from the given noisy speech including the pink noise, and the zero-mean white Gaussian noise. The proposed technique also has potential application in robust speech recognition tasks.

5. Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 60971095, and also the Fundamental Research Funds for the Central Universities.

6. References

[1] Advisory Panel on White House Tapes.The Executive Office Building Tape of June 20, 1972: Report on a technical investigation. United States District Court for the District of Columbia, May 31,1974. Available:http://www.aes.org/aeshc/docs/forensic audio/watergate.tapes. report. Pdf
[2] Robert C. Maher, Audio Forensic Examination (Authenticity, enhancement, and interpretation), IEEE Signal Processing Magazine, vol. 26, no.2, pp.84-94, 2009.
[3] Joseph P.Campbell, Wade Shen, William M.Campbell, Reva Schwartz, Jean-Francois Bonastre, Driss Matrouf, Forensic Speaker Recognition, IEEE Signal Processing Magazine, vol. 26, no.2, pp.95-103, 2009.
[4] Lv Gang, Zhao Heming, Tracking formant trajectory of continuous Chinese whispered speech with hidden dynamic model based on dynamic target orientation, Journal of Convergence Information Technology, vol. 5, no.9, pp.222-230, 2010.
[5] Kirill Sakhnov, Ekaterina Verteletskaya, Boris Simak, Echo delay estimation using algorithms based on cross-correlation, Journal of Convergence Information Technology, vol. 6, no.4, pp.1-11, 2011.
[6] Yang Lu and Philipos C.Loizou, A geometric approach to spectral subtraction, Speech Communication, vol. 50, no.6, pp.453-466, 2008.

[7] Sunil D.Kamath, Philipos C. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, In Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-02, vol. 4, pp.IV-4164, 2002.
[8] L.Singh, S.Sridharan, Speech enhancement for forensic applications using dynamic time warping and wavelet packet analysis, In Proceedings of IEEE Region 10 Annual International Conference on Speech and Image Technologies for Computing and Telecommunications, vol. 2, pp.475-478, 1997.
[9] V.Namias, The fractional order Fourier transform and its application to quantum mechanics, IMA Journal of Applied Mathematics, vol. 25, no.3, pp.241-265, 1980.
[10] Ozaktas Haldun M., Zeev Zalevsky, M. Alper Kutay, The Fractional Fourier Transform, New York:wiley,2001.
[11] Wang, Dongxia, Fan, Zhenwei, Li Bo, An adaptive beamforming method based on post-multistage Wiener filter for the speech enhancement, In Proceedings of the 2010 2nd International Conference on Signal Processing Systems.2, pp.V2-360-V2-362, 2010.