Comparison of Speech Enhancement Algorithms for Forensic Applications

Ahmed Kamil Hasan, David Dean, Bouchra Senadji and Vinod Chandran

School of Electrical Engineering and Computer Science , Queensland University of Technology

ahmedkamilhasan.alali@hdr.qut.edu.au, ddean@ieee.org, {b.senadji, v.chandran}@qut.edu.au

Abstract

Speech enhancement algorithms play an essential role in forensic applications, and enhanced speech signals can be used in court as evidence in criminal cases. This paper compares the performance of single channel (spectral subtraction and level dependent wavelet threshold techniques) and multiple channel (independent component analysis or ICA) speech enhancement algorithms to remove real environmental noise from noisy audio recording signals. Experimental results demonstrate that ICA achieves a significant improvement in average signal to noise ratio (SNR) enhancement compared to single channel speech enhancement algorithms, when 100 sentences from a forensic voice comparison database were corrupted with a car, street and factory noise at input SNR (-10 to 10 dB).

Index Terms: speech enhancement, independent component analysis, spectral subtraction, wavelet threshold technique

1. Introduction

Speech recordings obtained in the context of law enforcement agencies are often degraded by various types of real environmental noise. The police agencies often use hidden microphones to record the speech from the criminal in public places. Such forensic audio recordings may be far away from the hidden microphones and these recordings are often corrupted by car, street or machinery (factory) noise. It is difficult to directly use these recordings in court as a part of evidence in criminal cases, because their intelligibility is poor. Therefore, speech enhancement algorithms in real-life casework may be more complicated than those in theoretical research. Choosing the most reliable method for speech enhancement algorithm under these conditions play an important role in forensic applications. The enhanced speech signal can be used to eventually establish or confirm the identity of the criminal [1].

Speech enhancement algorithms can be divided into single channel and multiple channel algorithms depending on the number of the microphones that are used for collecting the noisy speech signal. Various algorithms for single channel speech enhancement, such as spectral subtraction [2] and wavelet threshold techniques [3] have been proposed in the last few decades, but these methods do not achieve great improvements in speech quality when the speech signal is corrupted with real environmental noise.

The spectral subtraction algorithm [2] is based on subtracting the estimated spectrum of the noise signal from the spectrum of the noisy speech signal. Since the spectrum of real environmental noise and speech signal are not uniformly distributed over the whole frequency bands, the musical noise will appear in the enhanced speech signal. This noise will lead to reduction in the quality of the denoised speech signal [4]. Wavelet denoising techniques are widely used to suppress noise from noisy speech signals [3] [4]. Noise is removed by applying an appropriate threshold to the wavelet cofficients for high frequency bands (detailed coefficients). This is based on the assumption that detail coefficients below significant energy levels arises from background noise rather than speech [3]. Wavelet threshold techniques fail to suppress noise in high SNR cases [5]. Colored noise is a non-stationary signal and the distribution of colored noise is spread non-uniformly over different frequency sub bands [4]. Such noise can have significant energy in the wavelet coefficients for low frequency band (approximation) or detail wavelet coefficients. If the power spectral density of the colored noise is concentrated at low frequency sub bands, a threshold applied to high frequency components of the noisy speech signals will not eliminate the low frequency components of the noise and will lead to a poor signal to noise ratio at the output.

Multichannel speech enhancement algorithms can be used to suppress and improve the quality of the speech signal under noisy conditions [6] . Independent component analysis (ICA) is widely used in multi channel speech enhancement and it is used to separate the speech from the noise by transforming the noisy speech signals into components which are statistically independent [7]. The principle of estimating independent components is based on maximizing the non-Gaussian distribution of one independent component [8]. The difference between a Gaussian distribution and the distribution of the independent component is measured using a contrast parameter, such as kurtosis, which is maximized by the ICA algorithm [8].

Single and multiple speech enhancement algorithms were also used to suppress the noise from noisy forensic recording signal in the existing literature review. Single channel speech enhancement was used with dynamic time warping and wavelet packet threshold techniques to suppress co-talker interference noise from forensic audio recordings in [9]. Multichannel speech enhancement was used to remove co-talker noise from the noisy forensic recording by using the delay and sum beamforming algorithm in [6]. Spectral subtraction was used to remove colored noise from mixed speech signals and convolutive ICA was used to separate one speaker from another in [1] to improve the performance of speaker identification. The original contribution of this research is to investigate the performance of ICA to suppress real environmental noise from short utterance of noisy forensic recordings, and to compare this performance with single speech enhancement algorithms under such conditions.

2. Model of ICA

Let the speech and noise signals emitted from N sources be represented as $s(t) = \{s_1(t), s_2(t), ..., s_N(t)\}$. The noisy speech signals can be recorded instantaneously by using M microphones in a street for forensic applications and be expressed as $x(t) = \{x_1(t), x_2(t), \cdots, x_M(t)\}$. Instantaneous ICA can be defined as a linear transformation of noisy speech signals

into components which are statistically independent, and can be represented as [8]

$$x = As \tag{1}$$

where A is an unknown mixing matrix

The goal of ICA is to estimate the original sources from the mixed signals. The estimates of speech and noise signals $(\hat{s})$ can be represented by the following equation:

$$\hat{s} = Wx \tag{2}$$

where *W* is the unmixing matrix which equals the inverse of the mixing matrix A when the matrix is square.

In this paper, we use two sources (speech and noise) and two microphones to record the noisy speech signals$(M = N = 2)$. Therefore, the mixing and unmixing matrices are square and they have a size of 2×2.

2.1. Fast ICA Algorithm

The procedure for a fast ICA algorithm for one unit can be illustrated by the following steps [8] :

1. Remove the mean value from the noisy signal and center its distribution.
2. Whiten the noisy speech signal (x) to get (x_w) by using eigenvalue decomposition of the covariance of the noisy speech signal.

$$x_w = VD^{-1/2}V^T x \tag{3}$$

   where V is the eigenvector matrix of the covariance of the noisy speech signal, and $D^{-1/2}$ is the inverse square root diagonal of the eigenvalue matrix.

3. Choose an initial vector of unmixing matrix *W*.
4. Estimate a row vector of unmixing matrix

$$w^+ = E\{x_w\, g(w^T\, x_w)\} - E\{g'(w^T\, x_w)\}w \tag{4}$$

   where w^+ is the new value of the row vector of the unmixing matrix, E is the sample mean, g and g' are the first and the second derivatives of the contrast function respectively.

5. Normalize the row vector of w^+

$$w^* = \frac{w^+}{\|w^+\|} \tag{5}$$

   where w^* is the normalization of the new row vector of the unmixing matrix.

6. Insert $w = w^*$ in step 4 and repeat the procedure until there is convergence.

The criterion of convergence is that the direction of previous and new values of w must be in the same direction, i.e. the dot product of these *w* points is almost equal to one.

This algorithm is based on separating one non-Gaussian component each time under the assumption that the sum of the others has a Gaussian distribution. It is necessary to prevent different row vectors of w from converging to the same maxima and this can be performed by using a deflation decorrelation of the output $w_1^T x, w_2^T x, \cdots, w_p^T x$ after every iteration.

3. Denoising by Wavelet Thresholding

Removing noise components by thresholding the wavelet coefficients is based on the assumption that in a noisy speech signal, the energy of the speech signal is mainly concentrated in a small number of wavelet dimensions [3]. The energy of these coefficients has larger values compared with other coefficients (especially noise) that have their energy spread over a large number of wavelet coefficients [3]. Thresholding the smaller wavelet coefficients to zero may reduce the noise components from a noisy speech signals [3].

Level dependent wavelet threshold techniques are used widely to suppress the noise from the noisy speech signal and improve the intelligibility of the speech [3]. This method is used in this paper because the forensic audio recording is corrupted with different types of colored noise and these noises have different distributions in different frequency subbands. Thresholding the wavelet coefficients for high frequencies(detail) of the noisy speech signal at each level may reduce the effect of the colored noise at high levels of noise. Level dependent threshold (λ) can be represented by [3]:

$$\lambda = \sigma_j(\sqrt{2\log\, N_j}) \tag{6}$$

$$\sigma_j = \frac{\mathrm{MAD}(D_j)}{0.6745} \tag{7}$$

where MAD is the median absolute deviation of the detailed coefficients for each level (D_j) and N_j is the length of the noisy speech signal for each level.

The procedure of level dependent wavelet threshold techniques can be illustrated by the following steps.

- Frame the noisy speech signal into several segments by using a Hamming window. The frame duration used in this paper is 25 msec.
- Compute the wavelet coefficients of the noisy speech signal by using discrete wavelet transform (DWT).
- Threshold the detailed coefficients of the noisy speech signal by using a hard or a soft level dependent threshold. Hard (T_{hard}) and soft (T_{soft}) thresholds can be expressed by the following equations:

$$T_{hard}(D_j) = \begin{cases} D_j, |D_j| > \lambda \\ 0, |D_j| \leqslant \lambda \end{cases} \tag{8}$$

$$T_{soft}(D_j) = \begin{cases} sign(D_j) * (|D_j| - \lambda), |D_j| > \lambda \\ 0, |D_j| \leqslant \lambda \end{cases} \tag{9}$$

- Reconstruct the enhanced speech signal by applying the inverse discrete wavelet transform to the thresholded wavelet coefficients.

4. Spectral Subtraction

This method is based on subtracting the estimated power spectrum of the noise from the power spectrum of the noisy speech signal, without prior knowledge of the power spectral density of the clean speech and noise signals. Spectral subtraction can be used to suppress background noise by assuming the noise is stationary or changing slowly during the non-speech and speech activity periods [2].

The procedure of spectral subtraction can be summarized by the following steps. Firstly, the noisy speech signal is framed into several overlapping segments by using a Hamming window. The duration of the frame used in this paper is 25 msec and

the duration of the overlap between two successive windows is 12.5 msec [10]. Secondly, the noise is estimated by computing the average power spectrum of noise from several silence frames (noise only). Spectral distance voice activty detector is used to determine the noise frames. Then, Fourier transform has been applied to the windowed frames of the noisy speech signal . Spectral subtraction can be computed as [10]:

$$|\hat{S}(k)|^2 = \begin{cases} |X(k)|^2 - \delta|\hat{D}(k)|^2, |X(k)|^2 - \delta|\hat{D}(k)|^2 > \beta|\hat{D}(k)|^2 \\ \beta|\hat{D}(k)|^2, \text{Otherwise} \end{cases} \tag{10}$$

where $X(k)$, $\hat{S}(k)$ and $\hat{D}(k)$ are the magnitude power spectrum of the segment of corrupted speech, estimated speech and estimated noise respectively, δ is the over subtraction factor and it depends on a posteriori segmental SNR, and β is the spectral factor with values between 0 and 1. For a large value of β, the spectral floor is high and the remaining noise is audible, while choosing a small value of β, the noise is significantly reduced, but the remnant noise becomes annoying. Hence,the optimum value of β used in this paper is 0.03 [10]. Finally, the enhanced speech signal $\hat{s}(t)$ can be obtained by applying an inverse Fourier transform to the phase function of discrete Fourier transform of the input speech signal and the estimated spectrum of the speech $|\hat{S}(k)|$.

5. Simulation Results

In this section, we present the simulation results of the independent component analysis, as well as a comparison with spectral subtraction and wavelet denoising techniques for the speech enhancement algorithms. For this paper, 4 levels and Daubechies 8 of the wavelet family were used, respectively. One hundred sentences from forensic voice comparison databases were used for simulation. The forensic voice comparison databases Australian English: 500+ speakers [11] consisted of 532 Australian speakers. Each speaker was recorded in three speaking styles (informal telephone conversation, information exchange task over the telephone and pseudo police style) which are popular speaking styles in forensic applications. The speech was sampled at 44.1 kHz and 16 bit/sample resolution in this database.

Various types of real environmental noise were used in this test from NOISEX-92 [12] and QUT-NOISE databases [13]. The NOISEX-92 database consists of various types of real environmental noise, recorded at 19.98 kHz sample rate with 16 bit resolution [12]. The QUT-NOISE database was created by collecting 10 hours of background noise in 5 common scenarios (cafe, home, street, car and reverberation). Each type of noise was recorded in two locations and the noise signal was sampled at 48 kHz sample rate with 16 bit resolution [13].

In these simulated results, the first and second microphones (x_1 and x_2) have the same distance to the clean speech source from forensic voice comparison database, but the noise(car, street noise from QUT-NOISE database and factory noise from NOISEX-92 database) has different distance to the second microphone resulting in the value of the mixing coefficient of the noise (α). These noises were used in this paper because these types of real environmental noise are more likely to occur in real covert forensic recordings. The mixed speech signal in an ICA algorithm can be represented by :

$$x_1 = s(n) + e(n) \tag{11}$$

$$x_2 = s(n) + \alpha e(n) \tag{12}$$

where $s(n)$ is the original speech signal and $e(n)$ is the noise.

Two down sampling frequencies were used in this paper. Firstly, the car and street noises were down sampled to 44.1 kHz before mixing with clean speech signal. Secondly, the speech signal was also down sampled to 19.98 kHz when factory noise was corrupted with clean speech signal. The down sampled is necessary to match the sampling frequencies of the clean speech and noise signals.

The mixed speech signals are separated using the fast ICA algorithm and the contrast function used in fast ICA has a Gaussian function [8]

There is an arbitrariness in the sign upon inversion. The problem of the sign change of the samples of estimated speech compared with samples of original speech in an ICA algorithm is solved by multiplying all samples of the estimated speech signal by -1 if the maximum correlation coefficient between estimated and original speech has a negative sign.

To evaluate the performance of speech enhancement algorithms in removing the noise from the speech signal, we use the reconstruction SNR or SNR output. It is defined as follows [3]:

$$\mathrm{SNR_o} = \frac{\sum_n s^2(n)}{\sum_n |s(n) - \hat{s}(n)|^2} \tag{13}$$

where $s(n)$ is the original speech signal, and $\hat{s}(n)$ is the estimated original speech signal. The SNR enhancement($\mathrm{SNR_e}$) in (dB) can be defined by:

$$\mathrm{SNR_e} = \mathrm{SNR_o} - \mathrm{SNR_i} \tag{14}$$

where $\mathrm{SNR_i}$ is the input SNR and it can be computed by the ratio of the sum squared of the clean speech to that of the noise from the first microphone (x_1).

To evaluate the effect of the changing mixing coefficient (α) on the performance of ICA to separate the noise from the noisy speech signal, we chose different values of α, ranging from 0.4 to 2.0. Experimental results demonstrated that increasing the value of α decreased the average SNR enhancement when car, street and factory noise were added to 100 sentences from forensic voice comparison database.

Figures (1-3) show comparisons of the average and standard deviation of SNR enhancement for different speech enhancement algorithms when 100 sentences from the forensic voice comparison database were corrupted with street noise, factory noise and car noise. Standard deviations from the Monte Carlo simulation are given on the bars. The value of (α) used in the simulation results of Figures (1-3) was 2 to compare the performance of multiple speech enhancement algorithm (ICA) under worst case conditions with single channel algorithms.

From Figures (1 to 3) we conclude the following:

- Independent component analysis achieves significant improvement in average SNR, compared with spectral subtraction and wavelet threshold techniques, when the speech signals were corrupted with street, car and factory noise for input SNR ranging from -10 to 10 dB.
- Level dependent wavelet denoising techniques achieve higher average SNR enhancement compared with the spectral subtraction algorithm for the same conditions, because real environmental noise are not uniformly distributed over the whole frequencies. Thresholding the detail coefficients in each high frequency sub band by using level dependent wavelet threshold will lead to improved average SNR enhancement at high levels of noise.

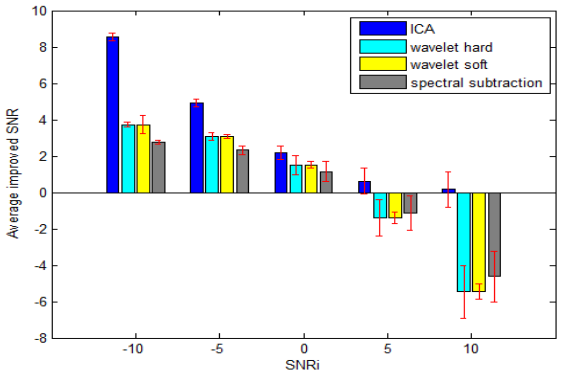

Figure 1: *Comparison of average SNR enhancement when street noise is added to the forensic database*

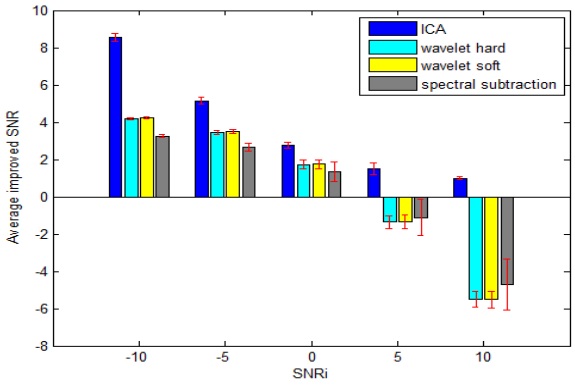

Figure 2: *Comparison of average SNR enhancement when factory noise is added to the forensic database*

- Level dependent threshold and spectral subtraction fails to suppress real environmental noise for input SNR in the range 5 to 10 dB, because power spectral densities of real environmental noise are concentrated at certain frequencies. Using a fixed oversubtracting parameter in spectral subtraction or thresholding all high frequency sub bands of the noisy speech signal at low levels of noise will lead to a distortion of the enhanced speech signal.

6. Conclusions

This paper compares the performance of ICA with specral subtraction and wavelet level dependent threshold techniques to suppress real environmental noise from noisy forensic recordings. Computer simulation results show that ICA achieves higher average SNR improvement than single speech enhancement algorithms for SNR levels in the range -10 dB to 10 dB. Further work is required to investigate the effect of channel delay duration on the performance of the convolutive ICA to suppress the noise from noisy speech signal and compare this result with other single speech enhancement algorithm (Wiener filter) and multiple speech enhancement (beamforming algorithm) for forensic applications.

7. References

[1] Denk, F., da Costa, J. P. C. L. and Silveira, M. A., "Enhanced forensic multiple speaker recognition in the presence of coloured noise", 8th IEEE Int. Conf. Signal Process. Commun. Syst., 2014, pp. 1-7.

[2] Berouti , M., Schwartz, R. and Makhoul, J., "Enhancement of speech corrupted by acoustic noise", IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 4, 1979, pp. 208-211.

[3] Ghanbari , Y. and Karami-Mollaei, M.R., "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets", Speech Commun., vol. 48, no. 8, pp. 927-940, 2006.

[4] Ruwei , L., Changchun, B., Bingyin, X. and Maoshen, J., "Speech enhancement using the combination of adaptive wavelet threshold and spectral subtraction based on wavelet packet decomposition", 11 th IEEE Int Conf. Signal Process., 2012, pp. 481-484.

[5] Donho, D.L and Johnston, I.M., "Ideal spatial adapation by wavelet shrinkage", Biometrika J., vol. 81, pp. 425-455,1994.

[6] Cao, Y., Sridharan, S. and Moody, M.P., "Post-microphone-array speech enhancement with adaptive filters for forensic application", Int. Symp. Speech, Image Process. Neural Netw., 1994, pp. 253-255.

[7] Zou , X., Jancovic, P., Liu, J. and Kokuer, M., "Speech Signal Enhancement Based on MAP Algorithm in the ICA Space", IEEE Trans. Signal Process., vol. 56, no. 5, pp. 1812-1820, 2008.

[8] Hyvarinen, A. and Oja, E., "Independent component analysis: algorithms and applications", Neural Netw., vol. 13, no. 4, pp. 411-430, 2000.

[9] Singh, L. and Sridharan, S., "Speech enhancement for forensic applications using dynamic time warping and wavelet packet analysis", 10 Annual Conf. IEEE Region Speech Image Technologies Computing Telecommun., vol. 2, 1997, pp. 475-478.

[10] Upadhyay, N. and Karmakar, A., "Speech Enhancement using Spectral Subtraction-type Algorithms: A Comparison and Simulation Study", Procedia Comput. Sci., vol. 54, pp. 574-584, 2015.

[11] Morrison G.S., Zhang C., Enzinger E., Ochoa F., Bleach D., Johnson M., Folkes B.K., De Souza S., Cummins N. and Chow D., "Forensic database of voice recordings of 500+ Australian English speakers", 2015. Available: http://databases.forensic-voice-comparison.net/.

[12] Varga, A. and Steeneken, H.J., "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems", Speech Commun., vol. 12, no. 3, pp. 247-251, 1993.

[13] Dean, D.B., Sridharan, S., Vogt, R.J. and Mason, M.W., "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms", Proc. Interspeech, Makuharia, Japan,2010, pp. 26-30.

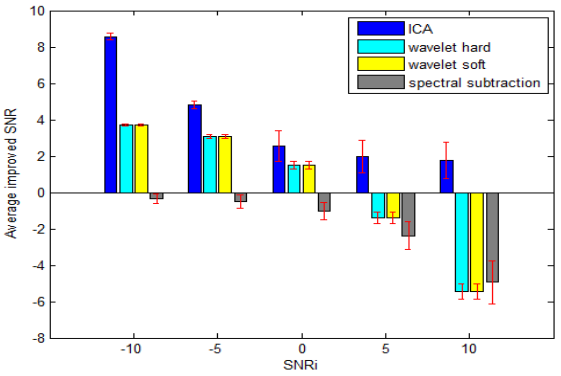

Figure 3: *Comparison of average SNR enhancement when car noise is added to the forensic database*