Speech enhancement on smartphone voice recording

Bagus Tris Atmaja, Mifta Nur Farid, Dhany Arifianto

Dept. of Engineering Physics, Faculty of Industrial Technology, Institut Teknologi Sepuluh Nopember, Kampus ITS Sukolilo, Surabaya 60111, Indonesia

E-mail: bagus@ep.its.ac.id

Abstract. Speech enhancement is challenging task in audio signal processing to enhance the quality of targeted speech signal while suppress other noises. In the beginning, the speech enhancement algorithm growth rapidly from spectral subtraction, Wiener filtering, spectral amplitude MMSE estimator to Non-negative Matrix Factorization (NMF). Smartphone as revolutionary device now is being used in all aspect of life including journalism; personally and professionally. Although many smartphones have two microphones (main and rear) the only main microphone is widely used for voice recording. This is why the NMF algorithm widely used for this purpose of speech enhancement. This paper evaluate speech enhancement on smartphone voice recording by using some algorithms mentioned previously. We also extend the NMF algorithm to Kulback-Leibler NMF with supervised separation. The last algorithm shows improved result compared to others by spectrogram and PESQ score evaluation.

1. Introduction

Signal enhancement can be viewed from many perspectives, one of the currently widely used is from the view of source separation perspective. Source separation is the process of decoupling of two or more sources with no or little prior information. From this definition, speech enhancement can be categorized as natural application for source separation because the goal is to enhance target speech from noises.

The use of smartphone as multimedia device has increased exponentially in current era. As we can seen on the media, most of journalist use smartphone devices instead of traditional recorder to record speech sound. In the other speech events, traditional recorder was almost gone replaced by smartphone voice recording feature to record human voice. The voice recording feature on the smartphone also can be used for voice biometrics, voice tapping, video recording, voice commands and voice calling. The use of smartphone as voice recording tools should be evaluated for measurement and it can be enhanced by signal processing technique addressing background noises.

This paper evaluates signal enhancement on smart phone voice recording from traditional method to the advanced one. From the traditional view, spectral subtraction method by Boll [1] was used with three different derivation: power spectral subtraction, magnitude spectral subtraction and over spectral subtraction [2]. The second approach to enhance speech on smartphone voice recording is by wiener filtering to estimate minimum mean square error (MMSE) as proposed by Lim and Oppenheim [3]. On the third approach, estimation of MMSE was done by short-time spectral amplitude (STSA) as suggested by Ephraim and Malah [4].

Moving to current speech enhancement algorithm, Non-Negative Matrix Factorization (NMF) was used as fourth approach based on [5], and finally on the fifth approach, NMF was modified by using supervised separation [6]. All of five approaches was evaluated by spectrogram and PESQ score to evaluate the performance of speech enhancement over smartphone device.

2. Speech Enhancement Algorithms

In the following section we briefly describe some algorithms used on this research: spectral subtraction, Wiener filtering, MMSE STSA and NMF.

2.1. Spectral subtraction

Speech signal recorded by smartphone device can be modelled by addition of clean signal with by noises. This phenomena can be write in frequency domain as follows,

$$Y(\omega) = X(\omega) + D(\omega) \tag{1}$$

where $Y(\omega)$ is recorded noisy signal, $X(\omega)$ is clean target signal and $D(\omega)$ is noise. The goal is to obtain target signal from noisy signal. One of the simple solution is by subtracting $Y(\omega)$ with estimation of $D(\omega)$ to obtain $X(\omega)$,

$$|\hat{X}(\omega)| = |Y(\omega)| - E[|D(\omega)|] \tag{2}$$

By taking real part of that equation, it can be obtained the magnitude of spectral subtraction as follows.

$$|\hat{X}(\omega)| = max\left\{|Y(\omega)| - E[|D(\omega)|], 0\right\} \tag{3}$$

Instead of calculating magnitude of spectral subtraction, the power of each components can be calculated to achieve power of spectral subtraction.

$$|\hat{X}(\omega)|^{\alpha} = max\left\{|Y(\omega)|^{\alpha} - E[|D(\omega)|^{\alpha}], 0\right\} \tag{4}$$

in which $\alpha = 2$. From the power spectral subtraction 4, the gain can be expressed as follows

$$H(\omega) = \frac{|\hat{X}(\omega)|}{|Y(\omega)|} = \sqrt{\frac{|Y(\omega)| - E[|D(\omega)|]}{|\hat{X}(\omega)|^2}} = \sqrt{\frac{\gamma(\omega) - 1}{\gamma(\omega)}} \tag{5}$$

where $\gamma(\omega)$ is called the a-posteriori SNR $\gamma(\omega) = \frac{|Y(\omega)|^2}{E[|D(\omega)|^2]}$. Different gain can be obtained for various α.

$$H(\omega) = \frac{|\hat{X}(\omega)|}{|Y(\omega)|} = \left(\frac{\gamma(\omega)^{\alpha/2} - 1}{\gamma(\omega)^{\alpha/2}}\right)^{1/\alpha} \tag{6}$$

The final method on spectral subtraction is overspectral subtraction which is proposed by Berouti et al. [2] by adding weighting coefficient α and β to improve separation result.

$$|\hat{X}(\omega)|^2 = max\left\{|Y(\omega)|^2 - \alpha E[|D(\omega)|^2], \beta E[|D(\omega)|^2\right\} \tag{7}$$

The weighting coefficient α and β are used for reduce noise peaks which in α should be dependent on the frame segmental SNR (γ), less attenuation (small α) for high SNR and more attenuation (large α) for low SNR.

2.2. Wiener Filtering

Wiener filter is designed to find optimal linear filter that outputs the desired signal i.e. clean signal. It can be obtained by find $h*$ that minimizes $E[e^2(n)]$ by solving $\frac{\partial E[e^2(n)]}{\partial h} = 0$. The gain is defined by,

$$H(\omega) = \frac{\hat{X}(\omega)}{Y(\omega)} \tag{8}$$

Implementing Wiener filter,

$$e(n) = x(n) - \hat{x}(n) \tag{9}$$

In frequency domain, by assuming a non-causal IIR filter and by using the convolution theorem,

$$E(\omega) = X(\omega) - H(\omega)Y(\omega) \tag{10}$$

By minimizing $E[|E(\omega)|^2]$ with respect to $H(\omega)$ we have Wiener filter as follows,

$$H(\omega) = \frac{E[|X(\omega)|^2]}{E[|Y(\omega)|^2]} = \frac{E[|X(\omega)|^2]}{E[|X(\omega)|^2] + E[|D(\omega)^2]} \tag{11}$$

Or more generally,

$$H(\omega) = \left(\frac{E[|X(\omega)|^2]}{E[|X(\omega)|^2] + \alpha E[|D(\omega)|^2]} \right)^{\beta} \tag{12}$$

Since, $E[|X(\omega)|^2]$ is unknown and if $\alpha = 1$, $\beta = 1/2$ then $E[|X(\omega)|^2 = |\hat{X}(\omega)|^2$, with the result

$$|\hat{X}(\omega)| = H(\omega)|Y(\omega)| = \sqrt{\frac{|\hat{X}(\omega)|^2}{|\hat{X}(\omega)|^2 + E[|D(\omega)|^2]}}|Y(\omega)| \tag{13}$$

$$|\hat{X}(\omega)|^2(|\hat{X}(\omega)|^2 + E[|D(\omega)|^2]) = |\hat{X}(\omega)|^2|Y(\omega)|^2 \tag{14}$$

gives two solutions $|\hat{X}(\omega)|^2 = |\hat{Y}(\omega)|^2 - E[|D(\omega)|^2]$ or $|\hat{X}(\omega)|^2$ which is essentially the power spectral subtraction algorithm.

If $E[|X(\omega)|^2]$ is replaced by $|\hat{Y}(\omega)|^2 - E[|D(\omega)|^2]$, then Wiener filter is

$$H(\omega) = \frac{E[|X(\omega)|^2]}{E[|X(\omega)|^2] + E[|D(\omega)|^2]} = \frac{\gamma(\omega) - 1}{\gamma(\omega)} \tag{15}$$

where $\gamma(\omega) = \dfrac{|\hat{Y}(\omega)|^2}{E[|D(\omega)|^2]}$.

2.3. MMSE-STSA

Minimum mean square error based on short-time spectral amplitude (MMSE STSA) estimator is estimator that minimizes the mean square error of the spectral magnitude. It minimizes,

$$min\, E[(X_k - \hat{X}_k)^2] \tag{16}$$

Defined $\gamma(\omega)$ is called the a-posteriori SNR,

$$\gamma(\omega) = \frac{|Y(\omega)|^2}{E[|D(\omega)|^2]} \tag{17}$$

and

$$\xi_k = \frac{E[|X(\omega_k)|^2]}{E[|D(\omega_k)|^2]} \tag{18}$$

MMSE STSA estimator is function of the gain containing $\xi(\omega)$ and $\gamma(\omega)$. Moreover, the full derivation can be obtained from [4].

$$\hat{X}(\omega) = G(\xi(\omega), \gamma(\omega))Y(\omega) \tag{19}$$

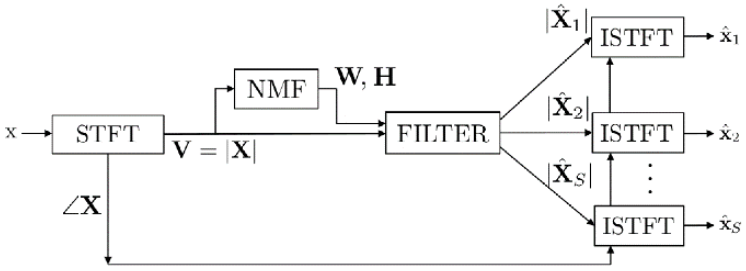

Figure 1. Block diagram of NMF signal enhancement by source separation [8]

2.4. NMF

NMF is matrix Factorization where everything is non-negative. It can be used for signal enhancement by source separation method. Source separation is the core of this work after modeling sound mixture. Separation principle consist of the following steps:

(i) STFT
(ii) NMF (Non-negative Matrix Factorization)
(iii) FILTER (Masking)
(iv) ISTFT

Those steps of signal enhancement by source separation can be organized by the block diagram in figure 1. As seen in the diagram, the source signal can be decomposed into its components, for example $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$. The target signal is the only desired signal while the others can be neglected or assumes as noises.

(i) Unsupervised NMF
In NMF, the matrix output magnitude $|X|$ from STFT process is decomposed into two matrices there are basis vectors W and weights H. Then a subset of basis vectors W_s and activation H_s is chosen to reconstruct source s by estimate the source s magnitude.

$$|\hat{X}_s| = W_s H_s = \sum_{i \in s} (w_i h_i^T) \tag{20}$$

To solve W and H for given a known V, a optimization problem frame is applied as

$$minimize\ D(V||WH) \tag{21}$$

which $W, H \geq 0$ and D is a measure of "divergence" that obtained by Kullback-Leibler, then to minimize

$$D(V||WH) = \sum_{i}, j \left(v_{ij} log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + \hat{(}WH)_{ij} \right) \tag{22}$$

Now we just iterate between:
(a) Updating W
(b) Updating H
(c) Checking $D(V||WH)$. If the change since the last iteration is small, then declare the convergence.

(ii) Supervised NMF
In unsupervised NMF, to control which basis vector explain which source is remain difficult, therefore supervised NMF is used. The general idea of supervised NMF is the use isolated

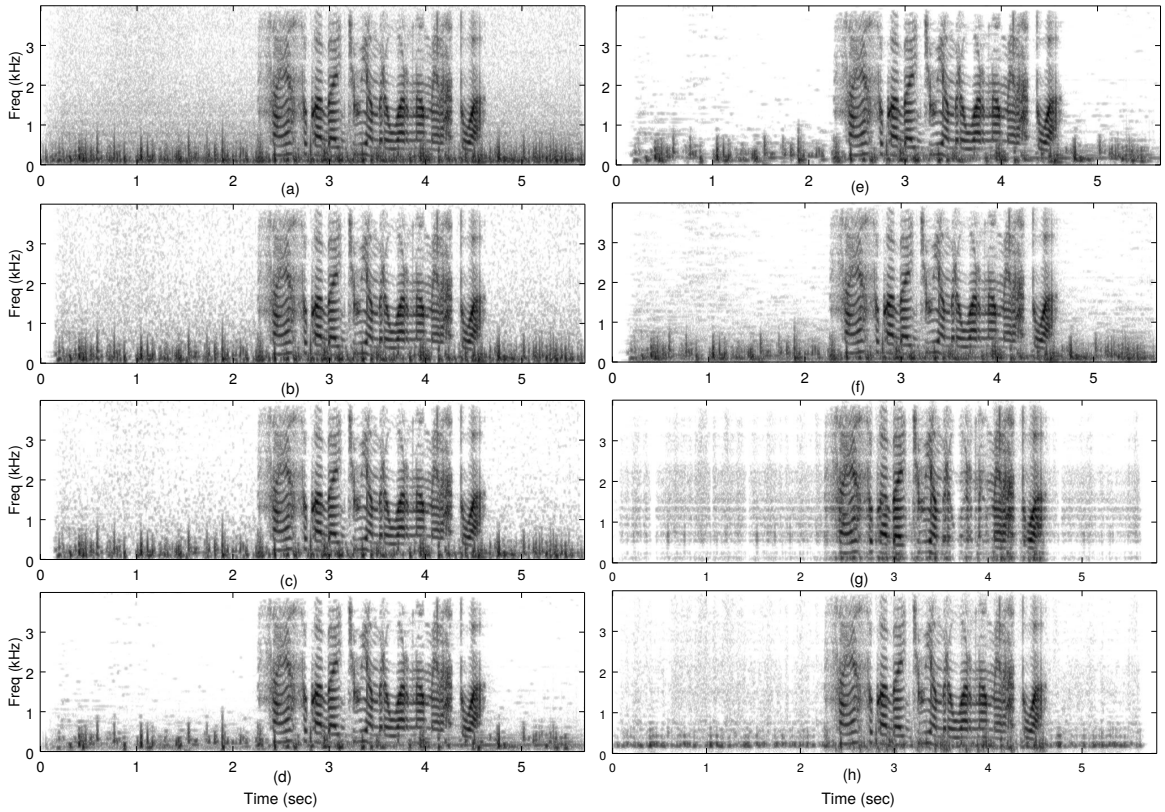

Figure 2. Spectrogram of speech enhancement result: (a) Noisy signal (b) spectral subtraction power (c) spectral subtraction magnitude (d) spectral subtraction over (e) Wiener filtering (f) mmse stsa (g) NMF (h) supervised NMF

training data of each source within a mixture to pre-learn individual models of each source. The process of supervised NMF obtained from [8] consist of the following steps,

(a) Use isolated training data to learn a Factorization ($W_s H_s$) for each source
(b) Throw away activation H_s for each source s
(c) Concatenate basis vectors of each source (W_1, W_2) for complete dictionary W
(d) Hold W fixed, and factorise unknown mixture of source V (only estimate H)
(e) Once complete, use W and H as before to filter and separate each source

3. Experiment

The experiment is conducted by using smartphone Motorola Razr IS12M to record speech target and fan noise with the voice recording software, Easy Voice Recorder$^{\text{TM}}$. The sampling rate was used at 8000 Hz and recorded in semi anechoic chamber. Sound source used is laptop speaker with Indonesian speech database and background noise from electric fan. This fan noise is made to imitate wind noises when making a call in the road on vehicle. The experiment set-up can be seen in Figure 3. The sound data, Matlab/Octave codes along with .tex files of this paper are openly available at: `http://bitbucket.org/bagustris/icopia2016` in the spirit of open science.

4. Result and Discussion

The recorded sound explained from the previous section analyzed offline by PC computation. Each .wav files were evaluated with seven algorithms explained in section 2 and enhanced

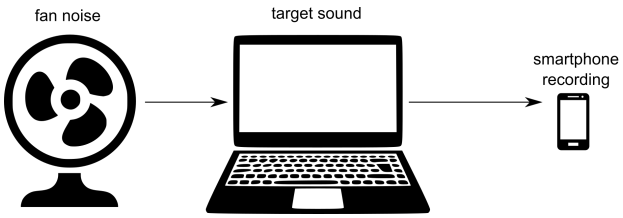

Figure 3. Experiment set-up

Algorithms	PESQ
spec sub pow	2.3982
spec sub mag	2.4493
spec sub over	2.4994
wiener filtering	2.5224
mmse-stsa	2.4922
nmf	1.9156
supervised nmf	2.5843

Table 1. PESQ score

speeches were evaluated by means of spectrogram and PESQ score [7]. Figure 2 shows spectrogram of noisy signal (a) with results from seven different speech enhancement algorithms (b to h).

From the spectrogram, it is shown that NMF algorithms both unsupervised and supervised can suppress noise much more other speech enhancement algorithms. Unsupervised NMF even shows almost no noise in the lower part of spectrogram. However, from the PESQ score, it is clearly shown that supervised has the highest speech intelligibility compared to other algorithms.

The PESQ score also shows that NMF has lowest score although its spectrogram shows it can suppress noise more than other algorithms. When listened, the sound resulted from this unsupervised NMF has degraded voice quality that impact its intelligibility score. However, the spectrogram of enhanced speech from Unsupervised NMF showed filtered noises on lower frequencies more than other methods. We choose PESQ score for objective evaluation because it is the current standard in telecommunication. The listened sound also shows consistency with the PESQ score.

5. Conclusion

This paper review some speech enhancement algorithms from classical to modern method by evaluating its directly in smartphone voice recording. From objective evaluation by using spectrogram and PESQ score, it can be concluded that the modern method, supervised NMF has highest PESQ score and spectrogram similarity compared to the original clean signal. On future research, computation time should be studied and enhanced for real time implementation.

References

[1] Steven Boll 1979 Suppression of acoustic noise in speech using spectral subtraction *Acoustics, Speech and Signal Processing* IEEE Transactions on 27 no. 2, pp. 113—120

[2] M. Berouti, M. Schwartz, and J. Makhoul, 1979 Enhancement of speech corrupted by acoustic noise, *IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP âĂŹ79)*, pp. 208—211

[3] Jae S Lim and Alan V Oppenheim 1979 Enhancement and bandwidth compression of noisy speech *Proceedings of the IEEE 67* no. 12, pp. 1586—1604

[4] Y. Ephraim and D. Malah 1984 Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator *IEEE Transactions on Acoustics, Speech and Signal Processing* pp. 1109—1121

[5] P. Smaragdis and J.C. Brown 2003 Non-negative matrix factorization for polyphonic music transcription *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* pp. 177—180

[6] P Smaragdis, B Raj, M Shashanka 2007 Supervised and semi-supervised separation of sounds from single-channel mixtures, *International Conference on Independent Component Analysis and Signal Separation* Springer—Verlag pp. 414—421

[7] ITU 2000 Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. *ITU—T Recommendation* P. 862

[8] M. M. Nandakumar and K. E. Bijoy 2014 Performance evaluation of single channel speech separation using non-negative matrix factorization *2014 National Conference on Communication, Signal Processing and Networking (NCCSN)* pp. 1—4