Speech enhancement based on magnitude estimation using the Gamma prior

Tran Huy Dat, Weifeng Li, Kazuya Takeda and Fumitada Itakura

CIAIR, Department of IE., Nagoya University, Japan

Abstract

In this paper, we propose a speech enhancement method based on spectral magnitude estimation. We modify the noise estimation from the minimum statistics method and combine with a maximum a posterior (MAP) decomposition, using the Rice-conditional probability and a non-Gaussian statistic model of the speech. We derive two versions of magnitude decomposition and magnitude-phase decomposition and compare to spectral subtraction and other MAP methods based on the Gaussian statistic (MMSE, LSA). The experiments show the advantage of the proposed method in the improvement of both SNR (up to 12 dB) and recognition accuracy rate (up to 21 % to base line).

1. Introduction

Reducing additive noise on the magnitude or power domain is an important problem of speech enhancement and speech recognition. In general, the speech enhancement systems can be divided into single and multi-microphones processing. In this paper we focus on one-microphone speech enhancement systems. Among this group, spectral subtraction and maximum a posterior (MAP) decomposition methods are commonly used.

The advantages of the spectral subtraction methods are their low computation cost and relative efficiency. Recently, a new approach, minimum statistic-spectral subtraction method, is proposed by R.Martin [3] and has shown robustness to cases of both stationary and non-stationary noises.

However, in the spectral subtractions methods, the cross-correlation between noise and speech in total power estimation is simplified via an over subtraction coefficient. Moreover, the statistical characteristics of speech and noise, greatly affecting to the robustness of the denoising method, are not considered.

MAP decomposition methods, based on statistical modeling of speech and noise, show better results in literatures [1], [2]. Up to the present, the MMSE version of MAP on the magnitude domain, proposed by Ephraim & Malah [1], is considered the optimal spectral amplitude estimation. Later, the simpler MMSE decomposition on the power domain proposed by Wolf [5] has shown the same results as in [1].


Although MMSE and LSA are popular used, there are some drawbacks. In both cases a Gaussian statistic model of speech is assumed. But the speech signals on the frequency domain is specified by the formant, and the AM-FM structure in general can not be Gaussian.

In this work we propose a MAP decomposition method on the magnitude domain with a Gamma prior model of speech magnitude. For noise estimation we modified the minimum statistic method and combined it with our proposed MAP decomposition. The effectiveness of our proposed method, compared to popular MMSE and LSA versions assuming the Gaussian prior, is demonstrated in both SNR improvement and speech recognition accuracy.

The organization of this paper is as follows: In section 2 we derive the proposed MAP decomposition method on the magnitude domain, in section 3 we describe the noise estimation used in this work. In section 4 we report the results of the comparison experiments in SNR improvement and speech recognition accuracy. Section 5 gives a discussion and section 6 summarizes this paper.

2. MAP decomposition using Gamma prior

We begin with the knowledge of that the speech magnitude always has few large values while others are small. Theoretically three families of distributions can be used for modeling: (1) Gamma distribution, (2) Positive super Gaussian distribution; and (3) Positive Gaussian mixture distribution. In this work we perform MAP decomposition for the Gamma prior because it gives a closed form solution. For MAP decomposition the Gamma distribution of the spectral magnitude is given by following:

$$f(|S|) = \frac{1}{\sigma_s} \, b^a \, G(a) \left( \frac{|S|}{\sigma_s} \right)^{a-1} \exp \left[ \frac{-b |S|}{\sigma_s} \right]$$

where $\sigma_s^2$ is variance of the random variable $|S|$ at each frame and frequency index. We note that, the distribution (1) is given over all the frequency bins and analysis frames. Therefore the normalized variable is not strictly unit-variance distributed.

2.1 Background model

We consider the model of noisy speech in the short time DFT (STDFT) domain:
\[ X(l, k) = \begin{cases} S(l, k) + N(l, k) & \text{speech + noise} \\ N(l, k) & \text{noise} \end{cases} \]  
(2)

Where \( S(l, k) \), \( N(l, k) \) and \( X(l, k) \) are the complex STDF components of speech, noise and noisy speech at each frame index \( l \) and frequency bin \( k \). Each complex component in (2) is presented in the form of real and imaginary parts, magnitude and phase.

\[ X = |X| e^{j\phi} \]
(3)

We consider the independent zero-mean Gaussian model of the real and imaginary parts of noise [6].

The joint conditional probability of the DFT components (first proposed by Rice) of the noisy speech, given the speech, is

\[ f(X_s, X_n | S_l, N_l, X_l) = \frac{1}{2\pi \sigma_s^2 \sigma_n^2} \exp\left( -\frac{(X_s - E(X_s))^2 + (X_n - E(X_n))^2}{2\sigma_s^2} \right) \]
(4)

where \( \sigma_s^2 \) is variance of the noise real and imaginary parts. The conditional joint probability for the magnitude and phase is given by the transform through the Jacobian [6],

\[ f([X_s, X_n], [s, \phi]) = \frac{1}{2\pi \sigma_s^2 \sigma_n^2} \exp\left( -\frac{(X_s - E(X_s))^2 + (X_n - E(X_n))^2}{2\sigma_s^2} \right) \]
(5)

where the Jacobian is

\[ J = \frac{\partial (X_s, X_n)}{\partial (X_s, X_n)} = \begin{vmatrix} \cos \phi & -|X| \sin \phi \\ \sin \phi & |X| \cos \phi \end{vmatrix} = |X| \]
(6)

Substituting the (6) into (5) gives

\[ f(X_s, X_n | [s, \phi]) = \frac{1}{2\pi \sigma_s^2 \sigma_n^2} \exp\left( -\frac{|X|^2 - 2|X||s| \cos(\phi - \phi) + |s|^2}{2\sigma_s^2} \right) \]
(7)

### 2.2 Rice-Gamma decomposition version 1

The joint MAP decomposition for both magnitude and phase is derived in this section

\[ \hat{S}_l, \hat{\phi} = \arg \max_{S_l, \phi} \log J([X_s, X_n], [S_l, \phi]) \]
(8)

Where the joint distribution

\[ J([X_s, X_n], [S_l, \phi], [s, \phi]) = f([X_s, X_n], [S_l, \phi], [s, \phi]) f(S_l f(\phi)) \]
(9)

The MAP decomposition equations lead to system of equations:

\[ \frac{\partial \log J}{\partial [s]} = 0 \quad \frac{\partial \log J}{\partial \phi} = 0 \]
(10)

Substituting (1) and (7) into (10) assuming a uniform distribution of phase, gives an estimation of phase of speech:

\[ \hat{\phi} = \phi \]
(11)

Then the estimation equation for the magnitude gives:

\[ \frac{|\hat{S}_l|}{\sigma_s} = \left( \frac{|X_s| \cos(\phi_s - \hat{\phi}) - \frac{b}{\sigma_s}}{\sigma_s^2} \right) + \frac{(a-1)}{S} \]
(12)

Substituting (11) into (12) gives the solutions of (12) as follows:

\[ |\hat{S}_l| = \frac{1}{2} \left( \left| X_s \right| - \frac{h}{\sigma_s^2} \right) \pm \sqrt{\left( \left| X_s \right| - \frac{h}{\sigma_s^2} \right)^2 + 4(a-1)\sigma_s^2} \]
(13)

The global maximum of posterior can be found strictly:

\[ |\hat{S}_l| < 0 \Rightarrow |\hat{S}_l| = |S| \]
\[ |\hat{S}_l| > 0 \Rightarrow |\hat{S}_l| = 0 \]
(14)

### 2.3 Rice-Gamma decomposition version 2

Another version of MAP decomposition can be derived using conditional probability just for magnitude components.

Assuming a uniform distribution of the noisy phase and taking the integral over the noisy phase gives the Rician conditional probability for the magnitude which is not phase-dependent.

\[ f([X_s]|s) = \frac{|X_s|}{2\pi \sigma_s^2} \exp\left( -\frac{|X_s|^2 + |s|^2}{\sigma_s^2} \right) I_0\left( \frac{|X_s||s|}{\sigma_s} \right) \]
(15)

Where \( I_0(x) \) is the Bessel function defined by

\[ I_0(x) = \int_0^\infty \exp[-x \cos(\varphi - \phi)] d\varphi \]
(16)

Making an approximation of the Bessel function

\[ I_0(x) = \frac{1}{\sqrt{2\pi x}} e^x \]
(17)

By the same way as in the 2.2 gives the solution:

\[ |\hat{S}_l| = \frac{1}{2} \left( \left| X_s \right| - \frac{h}{\sigma_s^2} \right) \pm \sqrt{\left( \left| X_s \right| - \frac{h}{\sigma_s^2} \right)^2 + 4(a-1)\sigma_s^2} \]
(18)

The global maximum is given by (14).

The estimation (13) or (18) is applied to the cases where speech is present. In general the estimation should take into account the speech detector evidence

\[ |\hat{S}_l| = \frac{\text{MAP}}{\mu \sigma_s^2} p = 1 \]
(19)

Where \( \text{MAP} \) is a decomposition following (13) or (18) and \( \mu \) is a small smoothing coefficient. Thus the obtained decompositions depend on the statistical parameters of speech and noise as well as the speech detector. In the next section we describe the estimation of these parameters.

### 3. The speech and noise variances estimation

Assuming a zero-mean imaginary and real parts of speech and noise in the (2), the variances \( \sigma_s^2 \) and \( \sigma_n^2 \) coincide with a full and half of speech and noise power.
spectrum (the expectation of square of magnitude). Therefore we can apply the same idea of noise estimation as used in spectral subtraction methods here.

\[ \text{UPDATE SPEECH VARIANCE ESTIMATE: Eq. (22)} \]

\[ \text{UPDATE NOISE VARIANCE ESTIMATE: Eq. (24)} \]

\[ \text{SPEECH-NON-SPEECH DETECTOR: Eq. (20)} \]

\[ X, \hat{S}, 2\sigma_s^2, n\sigma_n^2, 2n\mu, \sigma_n^2, \mu, \mu_0, \mu_1, \mu_2 \]

are the smoothing coefficients. Thus the noisy power to a noise threshold is estimated using the MS method

\[ \beta \sigma_n^2 \geq \gamma \sigma_s^2 \]

\[ \beta \sigma_n^2 \leq \gamma \sigma_s^2 \]

Thus the minimum statistic (MS) method [3] does not require the Gaussian statistic model, in this work we modify it to use in our enhancement systems. The diagram of noise and speech variance estimation is shown in the Fig.1. This modification differs from the MS version as below. Firstly, we take one more smoothing procedure to reduce the correlation on the frequency direction. The speech detection is done by comparing the smoothness of the MAP estimation in final step variance estimation. This procedure provides more robustness to estimation of speech power on each frame (we used a frame length of 4 ms with shift of 2 ms). Secondly we take feedback from final estimation of speech and noise (MS version as below. Firstly, we take one more smoothing procedure to reduce the correlation on the frequency direction. The speech detection is done by comparing the smoothness of the MAP estimation in final step variance estimation. This procedure provides more robustness to estimation of speech power on each frame (we used a frame length of 4 ms with shift of 2 ms).

Here \( \gamma_s, \gamma_n, \gamma_s \) are the smoothing coefficients. Thus with magnitude completely estimated, the wave form can be synthesized by the IDFT for each frame (using the noisy phases) and overlapping added.

4. Experiments

We first perform an experiment to define \((a, b)\). 100 utterances from 10 speakers are taken from the CIAIR in-car speech database [7]. Their speech variances are estimated following the procedure in section 3. Then the normalized variables of speech magnitudes are obtained and the maximum likelihood estimation performed to estimate \((a, b)\). In this experiment we obtained the values \(a = 2, b = 4\) used in the following experiments.

The performance evaluation of the proposed method consists of three parts. First we test the MAP decomposition methods with the modified noise and signal variance estimation using a simulation experiment. Second we test the methods using real noisy signals. Third we perform phoneme recognition experiments for both MAP decompositions. In both cases we compare our results with the original minimum statistic version of spectral subtraction.

White noise is used to for noise simulation. Clean speech is taken from the closed-talking microphone in the CIAIR database and consists of 10 utterances from 10 male and female speakers.

The segmental SNR defined by

\[ \text{snr}_{seg} = \frac{10}{N} \sum_{n} \log \frac{E_s[n]}{E_{s+n}[n]} . \] (25)

where \( n \) and \( E \) denote respectively the frame index and total energy on each frame. In this experiment the frame length of 25 ms and the shift of 10 ms are used for evaluation. The parameters used in the speech and noise power estimation are as follows: \( T = 5, \mu = 0.01 \)

\( \gamma_0 = 0.9, \gamma_n = 0.9, \gamma_s = 0.8, L = 60 \) and \( \beta = 5 \).

The MAP decompositions (MMSE [1], LSA [2], our RG1 and RG2 (Rice-gamma decomposition version 1 in section 2.2 and version 2 in section 2.3)) are performed and compared. The original minimum statistic spectral subtraction (MS-SS) is also implemented in this work.

Fig 2 shows the results of average SNR improvements according to the different level of the input SNR. For the real noisy signals, since the exact clean signal and noise are not available in SNR definition (25), we estimated the SNR_GMM of the log10 of energy on each frame (we used a frame length of 4 ms with shift of 2 ms to estimate the SNR_GMM)

\[ \text{snr}_{gmm} = 10 \left[ \mu - \mu_2 \right] . \] (25)

where: \( \mu_1, \mu_2 \) are the estimated means of the mixtures.

In next experiment we perform the experiment analyzing the SNR_GMM improvements for the signals from the CIAIR in-car data base. The noisy speech is
taken from a distant microphone at visor position. The 12 signals are collected from 3 male and 3 female speakers. Fig 3 compares the average SNR of both MAPs and MS-SS. Notes that the last column is the original SNR_GMM of the noisy speech.

Fig 2 Simulated white noise condition –segmental SNR improvement comparison

Fig 3 Real car noise condition –SNR_GMM comparison

Fig 4 Summary of recognition results

In the third experiment we perform phoneme recognition comparing the effectiveness of the denoising methods for speech recognition. The data is taken from the CIAIR in-car speech database--300 phonetically balanced sentences uttered by 15 male and 15 female speakers for training data. The test data includes another 100 phonetically balanced sentences uttered by 5 male and 5 female speakers. The training data was collected from the close-talking microphone using a headset when the car is stopped with the engine running and the speaker seated in the driver’s position. The test data was collected from a distant microphone (at visor position). Each frame (25ms of length and 10ms of shift) is comprised of 12-dimensional features vector including 12 estimated MFCC. There are a total of 15,244 examples in the training data and 4937 patterns in the test data. We use a Gaussian kernel support vector machine for training. The denoising methods are applied to both training and testing. Summary results are shown in the Fig4. Note that the baseline is performed without any enhancement technique.

5. Discussion

The SNR in two definitions are different but in both cases it shows the advantage in SNR improvements of the MAP (especially the RG methods) to the MS-SS. The SNR improvements are found in both the low and high input SNR cases. For the RG methods SNR is improved by 12dB. The very high values of SNR_GMM after doing the RG decompositions can be explained by good noise reduction. Of course the SNR improvement is not always coincident with sound quality but in most of cases the sound quality of RG is shown to be accepted. The improvements in recognition rate using Rice-gamma methods are 21% compared to base line, 3% compared to LSA, 5% compared to MMSE and more then 6% compared to MS-SS. One open question is that, although the choice of $a = 2, b = 4$ seems to be appropriate, further optimization of the values should improve more sound quality and SNR even further.

6. Conclusion

We proposed a posterior decomposition on the DFT magnitude domain using Rice-conditional probability and Gamma prior of speech. This method shows an advantage in both the SNR improvement and recognition accuracy rate, compared to spectral subtraction and other MAP methods based on Gaussian statistic assumption of speech.

7. References