EXPLOITING THE BASEBAND PHASE STRUCTURE OF THE VOICED SPEECH FOR SPEECH ENHANCEMENT

Sanjay P. Patil, John N. Gowdy

Department of Electrical and Computer Engineering, Clemson University
Clemson SC 29634, USA
sanjayp@g.clemson.edu, jgowdy@clemson.edu

ABSTRACT

Performance of traditional speech enhancement techniques like spectral subtraction and log-Minimum Mean Squared Error Short Time Spectral Amplitude (log-MMSE STSA) estimation degrades in presence of highly non-stationary noises like babble noise. This is mainly due to inaccurate noise estimation during the voiced segment of the speech signal. In this paper, we propose to exploit the fine structure of the phase spectra of the voiced speech in the baseband STFT domain. This phase structure is used to detect the noise dominant frequency bins in the voiced frames. This information is used to achieve better non-stationary noise Power Spectral Density (PSD) estimation. Using this estimation, performance of spectral subtraction and log-MMSE STSA is improved overall by 0.3 and 0.2, respectively, in terms of Perceptual Evaluation of Speech Quality (PESQ) measure over the original algorithms when noisy speech is used for pitch estimation. We also present the combination of these two algorithms (spectral subtraction and log-MMSE STSA) to achieve the overall PESQ improvement of 0.5 over standard log-MMSE STSA when accurate pitch estimation is available.

Index Terms— Phase estimation, speech enhancement, PESQ

1. INTRODUCTION

Speech signals are often degraded due to presence of background noises like babble noise, train noise, machine noise, etc. Such speech signals make listening difficult and are highly undesirable for automatic speech processing tasks such as speech recognition, speech coders, speaker identification, hearing aids, etc. This has motivated several researchers over the past decades to develop robust speech enhancement systems. Currently, several speech enhancement algorithms exist which estimate the magnitude spectrum of the underlying clean speech signal. These algorithms segment the speech signal in windowed frames of 20-40 msec in length and then apply DFT analysis to estimate the clean speech. The most challenging and important step in this overall process is noise estimation [1, 2]. For stationary noises like white noise, noise estimation can be carried out using a basic voice activity detector. However, performance is not satisfactory for non-stationary noises. Several methods exist to address the issue of non-stationary noise estimation such as quantile based noise estimation [3], MCRA [4] and MS [5]. In these methods noise is estimated from PSD of noisy speech.

Additive noise corrupts both magnitude and phase spectrum of clean speech. Though phase spectrum is usually considered to be insignificant for human perception as compared to magnitude spectrum, this is true only for high SNR(>5 dB). For lower SNRs phase degradation leads to audible speech distortion [6]. This has motivated the enhancement of the phase for the noisy speech along with the magnitude. It has been shown that fine phase structure exists in the voiced frames of speech if the overlap between the successive speech frames is small (around 4 msec) [7, 8]. Using this fact speech quality was improved in terms of PESQ in [7, 8]. In [7], spectral subtraction was carried out separately on the real and imaginary spectrum of the noisy speech in the modulation domain. Then the enhanced real and imaginary spectra were used to estimate the phase spectrum of the underlying clean speech. This phase spectrum was used to reconstruct the original speech instead of using the noisy speech phase. This approach resulted in quality improvement in the voiced segment of the speech. In [8], the harmonic model for voiced speech was exploited to estimate the phase spectrum in the baseband STFT domain. The estimated phase spectrum was then used to reconstruct the original speech. This resulted in noise reduction in the voiced frames. But this approach introduces undesirable artifacts in the processed speech due to inexact harmonic modeling for voiced speech. Moreover, enhancement is achieved only in the voiced frames of the speech [8].

In this paper, we use the approach suggested in [8] to estimate the phase for the voiced speech. But instead of using this as a estimate to reconstruct the original clean speech, we use it to detect the noise dominant frequency bins in the voiced frames of the speech. Those frequency bins are used to further refine the VAD based noise estimation. This noise estimation is then combined with traditional spectral subtraction and log-MMSE STSA to demonstrate the significance of the approach. Similarly, the spectral sparsity detected by estimated phase values is used to redefine the spectral subtraction rule to avoid over-attenuation of low energy speech.

2. BACKGROUND: ESTIMATION OF CLEAN PHASE

Let *s(n)* denote the clean speech signal. Background noise *w(n)* degrades the clean speech producing the noisy signal as:

$$y(n) = s(n) + w(n). \tag{1}$$

To transform the noisy speech into the frequency domain, speech is windowed into overlapping frames of length N with a overlap of L samples, using the Hamming window. If the overlap is small, then the phase and magnitude are shown to be highly correlated [6]. Each frame is 32 msec long with a shift of 4 msec. Frames are then transformed into the frequency domain using DFT of length *N* represented by:

$$Y(\lambda, \mu) = S(\lambda, \mu) + W(\lambda, \mu) \tag{2}$$

where $S(\lambda, \mu)$ and $W(\lambda, \mu)$ represent spectral coefficients of speech and noise at frame λ and DFT bin μ. Baseband STFT for the noisy speech is given as:

$$Y_B(\lambda, \mu) = Y(\lambda, \mu) e^{-j\Omega_\mu \lambda L} \tag{3}$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

where $\Omega_\mu = \frac{2\pi\mu}{N}$ is the normalized angular frequency. Assuming a sinusoidal model for voiced speech, the clean phase for the voiced speech is estimated in [8] as:

$$\phi_{\tilde{S}_B}(\lambda,\mu) = \phi_{\tilde{S}_B}(\lambda-1,\mu) + (\Omega_h^\mu - \Omega_\mu)L \tag{4}$$

and

$$\phi_{\tilde{S}_B}(\lambda,\mu+i) = \phi_{\tilde{S}_B}(\lambda,\mu) + i(\pi - \frac{2\pi\lambda L}{N}) \tag{5}$$

where $\phi_{\tilde{S}_B}(\lambda,\mu)$ is the phase for voiced speech baseband Fourier coefficient at index μ and frame λ, L is the window shift in number of samples, $i \in [\lceil \frac{-f_0/2}{f_s}N\rceil, ..., \lceil \frac{f_0/2}{f_s}N\rceil]$, f_0 is the fundamental frequency and f_s is the sampling frequency. Ω_h^μ, the angular frequency of the harmonic closest to current DFT bin μ, can be expressed as:

$$\Omega_h^\mu = \underset{\Omega_h}{\operatorname{argmin}}(|\Omega_\mu - \Omega_h|) \tag{6}$$

where Ω_μ is angular frequency corresponding to current DFT bin μ and Ω_h is a harmonic frequency. Eq.(4) is used recursively to estimate the phase values at the frequency coefficient containing the harmonic component and (5) is used to estimate the phase between the harmonics in the frame. This algorithm uses the YIN [9] algorithm to estimate the pitch frequency.

3. PROPOSED METHOD

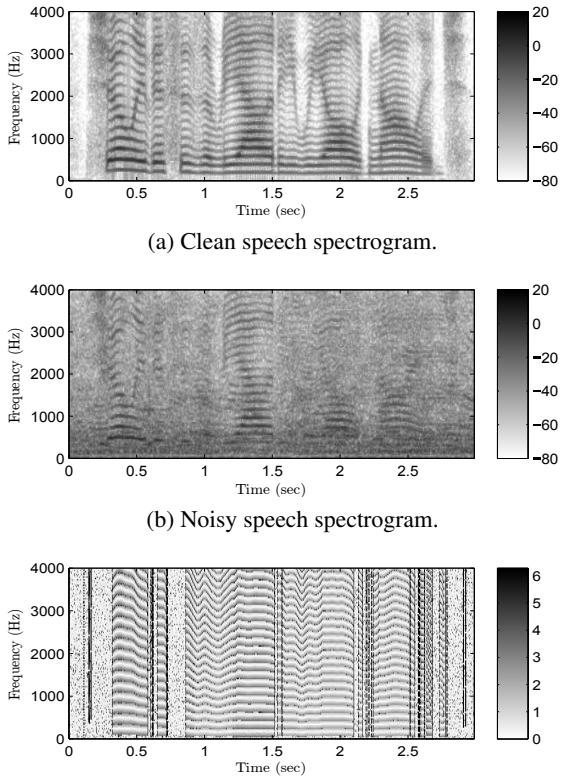
3.1. Determination of Noise Dominant Frequencies

In [8], the estimated clean speech phase given by (4) and (5) is used to reconstruct the speech, and the reconstructed speech is shown to be enhanced in the voiced segments. We used this phase estimation method to identify the noise dominant frequency bins in the voiced frames. These values are then used to further refine the final noise estimation. We compute the frame to frame phase difference from the above estimated clean phase as $\Delta\phi_{\tilde{S}_B}(\lambda,\mu) = \phi_{\tilde{S}_B}(\lambda,\mu) - \phi_{\tilde{S}_B}(\lambda-1,\mu)$. This phase difference is highly correlated with the magnitude of the underlying clean speech in the voiced frames as shown in Fig.1a and Fig.1c. Clean speech is corrupted by adding babble noise at 0dB global SNR (See Fig.1b). Estimated frame to frame phase difference for clean speech, i.e., $\Delta\phi_{\tilde{S}_B}(\lambda,\mu) = \phi_{\tilde{S}_B}(\lambda,\mu) - \phi_{\tilde{S}_B}(\lambda-1,\mu)$, is represented in Fig.1c. Here, we have plotted the absolute value of the phase difference in the range from 0 to 2π rad. From Fig.1c, it can be noted that phase difference can be used to determine the frequencies dominated by the harmonics and the frequencies containing high amount of noise in the voiced frames. Those noise dominant frequencies correspond to the gaps between the harmonics.

From (4) and (5), it can be noted that in the voiced frames the phase difference is close to zero for frequencies associated with the harmonics, and this phase difference deviates from zero for other frequencies. Thus, we use the following threshold(ϕ_T) based test to separate such frequencies as described below:

Let H be the total number of harmonics in a voiced frame, let F_h be the set of frequencies dominated by harmonic *h*, and let F_{nh} be the set of frequencies considered to be valid noise candidates in the neighboring of harmonic *h*. If μ_h is the DFT bin corresponding to harmonic *h* then we apply the following bin selecting rule in the range of frequencies $\mu_h + i$, where $i \in [\lceil \frac{-f_0/2}{f_s}N\rceil, ..., \lceil \frac{f_0/2}{f_s}N\rceil]$, for each harmonic:

$$\mu \in \begin{cases} F_{nh}, \text{if } \Delta\phi_{\tilde{S}_B}(\lambda,\mu) > \phi_T. \\ F_h, \text{otherwise.} \end{cases} \tag{7}$$



(a) Clean speech spectrogram.

(b) Noisy speech spectrogram.

(c) Estimated clean speech phase difference.

Fig. 1: Correlation between clean speech magnitude and baseband phase difference when pitch is estimated on clean speech by the YIN algorithm.

3.2. Computation of Noise PSD

For each band of the frequencies F_h and F_{nh}, the noise power is assumed to be constant and is given as the average of spectral magnitudes over F_{nh}. The noise estimate is calculated as:

$$N_{Fnh}(\lambda) = \sum_{j=1}^{|F_{nh}|} \frac{|Y(\lambda, F_{nh}(j))|^2}{|F_{nh}|} \text{.....for } \mu \in \mu_h + i. \tag{8}$$

where $|F_{nh}|$ denotes the cardinality of the set F_{nh}. This is repeated for each harmonic in a voiced frame, λ. Final noise PSD is obtained by combining the individual noise estimates and can be represented as:

$$|\hat{W}_\phi(\lambda)|^2 = \{N_{Fn1}(\lambda), N_{Fn2}(\lambda), N_{Fn3}(\lambda)....., N_{FnH}(\lambda)\}. \tag{9}$$

This noise estimation is valid only for voiced frames. In the unvoiced frames, noise estimation is carried out using standard VAD based noise estimation [10, 11]. When a voiced frame is detected, noise estimate is updated with the proposed noise PSD as (In our experiment we obtained the best results with the weighting factors 0.8 and 0.2.):

$$|\hat{W}(\lambda,\mu)|^2 = 0.8|\hat{W}(\lambda-1,\mu)|^2 + 0.2|\hat{W}_\phi(\lambda,\mu)|^2. \tag{10}$$

3.3. Use of Estimated Noise PSD for Speech Enhancement

We have verified the effectiveness of the proposed noise PSD estimation by using it in the standard spectral subtraction and log-MMSE

STSA algorithms. The spectral subtraction over-attenuation factor is also adjusted to provide less attenuation in the harmonic dominant frequency bins. But for log-MMSE STSA we have incorporated the estimated noise PSD without making any change in the original noise reduction rule. Spectral subtraction effectively suppresses noise at the expense of speech distortion, while log-MMSE STSA causes less speech distortion leaving some residual noise. We combine those two approaches to further improve the quality of speech.

The standard spectral subtraction rule is given as:

$$|\hat{S}_{SS}(\lambda,\mu)|^2 = \begin{cases} \hat{S}_{sub}, & \text{if} \hat{S}_{sub} > \beta|\hat{W}(\lambda,\mu)|^2 \\ \beta|\hat{W}(\lambda,\mu)|^2, & \text{otherwise.} \end{cases} \quad (11)$$

where $\hat{S}_{sub} = |Y(\lambda,\mu)|^2 - \alpha|\hat{W}(\lambda,\mu)|^2$, $|\hat{S}_{SS}(\lambda,\mu)|$ is the estimated clean speech magnitude, $|Y(\lambda,\mu)|$ is noisy speech magnitude, and $|\hat{W}(\lambda,\mu)|$ is estimated noise magnitude using (10). As we also have the knowledge of spectral sparsity in the voiced frame, we set $\alpha = 2.7$ if $\Delta\phi_{\hat{S}_B}(\lambda,\mu) < \phi_T$, else $\alpha = 5$. This leads to over-suppression of noise in the noise dominant period and less attenuation in harmonic dominant bins. In the unvoiced frames α is calculated using Berouti's rule [12].

In order to use the proposed noise PSD for log-MMSE STSA enhancement, we use the following rule [2] with noise estimation in (10):

$$|\hat{S}_{LMMSE}(\lambda,\mu)|^2 = G_{LMMSE}(\zeta,v)|Y(\lambda,\mu)|^2 \quad (12)$$

where $G_{LMMSE}(\zeta,v)$ and $v(\lambda,\mu)$ are defined by

$$G_{LMMSE}(\zeta,v) = \frac{\zeta(\lambda,\mu)}{\zeta(\lambda,\mu)+1} exp[\frac{1}{2}\int_{v(\lambda,\mu)}^{\infty} \frac{e^{-t}}{t} dt] \quad (13)$$

$$v(\lambda,\mu) = \frac{\zeta(\lambda,\mu)}{\zeta(\lambda,\mu)+1}\gamma(\lambda,\mu). \quad (14)$$

The terms $\zeta(\lambda,\mu)$ and $\gamma(\lambda,\mu)$ are the a priori and a posteriori SNRs.

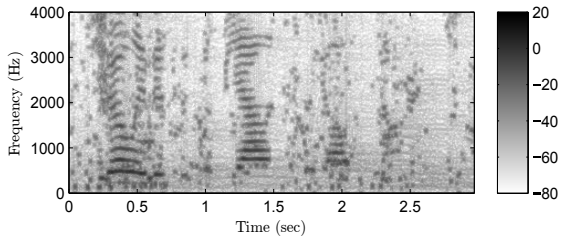
To achieve good noise suppression for noise dominant bins in voiced frames with overall less speech distortion we use the following combination:

$$|\hat{S}_{Fusion}(\lambda,\mu)|^2 = \begin{cases} |\hat{S}_{LMMSE}(\lambda,\mu)|^2 & \text{if } |Y(\lambda,\mu)| = U \\ & or \Delta\phi(\lambda,\mu) < \phi_T \\ \hat{S}_{Comb} & \text{otherwise.} \end{cases} \quad (15)$$

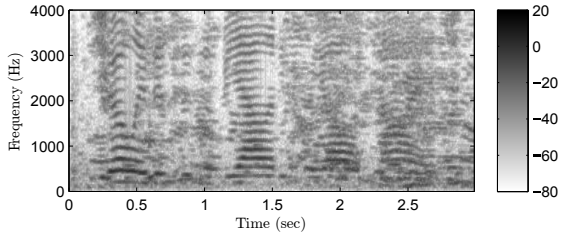
where $\hat{S}_{Comb} = 0.8 * |\hat{S}_{SS}(\lambda,\mu)|^2 + 0.2 * |\hat{S}_{LMMSE}(\lambda,\mu)|^2$. U and V denote the unvoiced and voiced frame detected by the YIN algorithm, respectively, and $|\hat{S}_{Fusion}(\lambda,\mu)|$ is the resultant magnitude of the combination.

4. EXPERIMENTAL SETUP

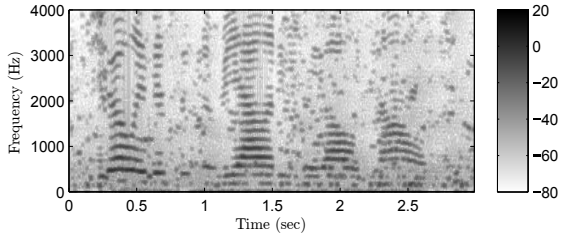
The performance of the proposed method was evaluated on the 100 sentences from the TIMIT database. The speech is corrupted by adding babble, restaurant and subway noise [13] with global SNRs from -5 dB to 10 dB. Frame length is set to 32 msec with 4 msec shift. This corresponds to 256 samples per frame with a shift of 32 samples. This large overlap is used so that speech magnitude and phase are correlated as stated earlier. Fundamental frequency estimation is carried out using the YIN algorithm [9]. Each speech frame is classified as voiced/unvoiced using the aperiodicity measure of the YIN algorithm. In this experiment it is set to 0.7. The bin selection threshold parameter ϕ_T is set to 0.5. The performance of the proposed method is very sensitive to accuracy of the pitch estimation. To see the effect of accurate pitch estimation on the performance, enhancement results are also presented when the YIN algorithm is run on the clean speech.



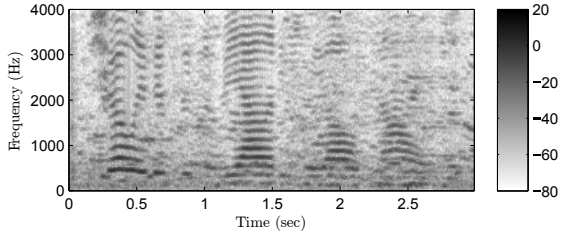
(a) Output of SS algorithm.



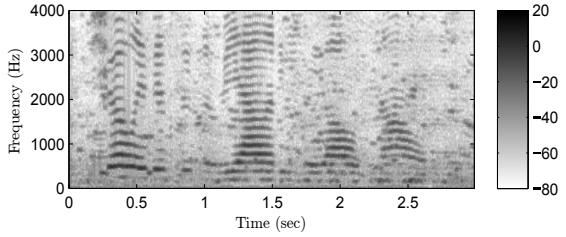
(b) Output of LMMSE algorithm.



(c) Output of SS-CPE algorithm.



(d) Output of LMMSE-CPE algorithm.



(e) Output of Fusion-CPE algorithm.

Fig. 2: Spectrograms of the speech processed by the discussed algorithms.

To quantify the comparison of the proposed method with the standard spectral subtraction and log-MMSE STSA, we used the Perceptual Evaluation of Speech Quality (PESQ) [14] to measure the speech quality. This choice of PESQ measure is motivated by

Table 1: PESQ evaluation of proposed algorithm against standard spectral subtraction and log-MMSE.

Noise	SNR(dB)	Noisy	SS	LMMSE	SS-NPE	SS-CPE	LMMSE-NPE	LMMSE-CPE	Fusion-NPE	Fusion-CPE
Babble	-5	1.32	1.21	1.41	1.47	1.76	1.56	1.69	1.50	1.98
	0	1.66	1.73	1.85	1.99	2.17	1.99	2.11	1.96	2.30
	5	2.02	2.17	2.26	2.38	2.47	2.34	2.41	2.32	2.58
	10	2.38	2.57	2.59	2.66	2.72	2.63	2.70	2.60	2.84
Restaurant	-5	1.35	1.12	1.42	1.45	1.78	1.55	1.68	1.49	1.97
	0	1.66	1.64	1.81	1.91	2.11	1.99	2.08	1.92	2.27
	5	2.00	2.07	2.16	2.31	2.42	2.34	2.42	2.24	2.53
	10	2.34	2.46	2.46	2.64	2.68	2.63	2.69	2.53	2.73
Subway	-5	1.22	1.13	1.36	1.58	1.73	1.64	1.76	1.66	1.93
	0	1.49	1.56	1.68	1.90	2.07	2.01	2.12	2.01	2.24
	5	1.81	2.01	2.05	2.29	2.37	2.37	2.46	2.36	2.51
	10	2.16	2.40	2.40	2.59	2.62	2.67	2.74	2.64	2.73

the fact that it is proven to be more reliable and correlated with Mean Opinion Score [15]. The implementations of the PESQ algorithm, spectral subtraction and log-MMSE STSA are taken from [10]. In future work we will evaluate the performance using subjective measures as well.

5. RESULTS AND DISCUSSIONS

Performance comparison of the proposed algorithm against standard spectral subtraction and log-MMSE is presented in Table 1. Spectrograms for the enhanced speech using standard spectral subtraction, log-MMSE and proposed methods are shown in Fig.2. Clean and noisy speech spectrograms are shown in Fig.1a and Fig.1b.

In Table 1, performance comparison using PESQ scores is shown for babble, restaurant and subway noise for SNR from -5dB to 10 dB. The spectral subtraction algorithm (SS) results in poor speech quality at low SNRs for non-stationary noises. When the proposed noise estimation algorithm is combined with spectral subtraction algorithm, significant improvement in the speech quality is observed. Performance of the standard log-MMSE algorithm (LMMSE) is also improved further when the proposed noise estimation is used. When we use clean speech to estimate the pitch and utilize the estimated noise PSD for noise reduction, the performance of resulting spectral subtraction (SS-CPE) and log-MMSE STSA (LMMSE-CPE) algorithms is superior to their counterpart spectral subtraction (SS-NPE) and log-MMSE STSA (LMMSE-NPE) when pitch is estimated from noisy speech. Accurate pitch estimation results in correct detection of noise dominant bins and hence gives better noise estimation. The combination of spectral subtraction and log-MMSE STSA results in even better speech quality when accurate pitch estimation is available. As shown in the Table 1, this combination with pitch estimated from clean speech is represented as Fusion-CPE and with pitch estimated from noisy speech is represented as Fusion-NPE. If the pitch estimate is not accurate, then harmonic dominant bins will be processed by the spectral subtraction algorithm instead of log-MMSE, and overall performance of Fusion-NPE will degrade.

The effectiveness of the proposed approach can also be confirmed by the spectrograms of the processed speech as shown in Fig.2. SS-CPE processed speech has less speech distortion as compared to SS processed speech as shown in Fig.2a and Fig.2c. Also, LMMSE-CPE results in better noise reduction than the standard LMMSE algorithm as shown in Fig.2b and Fig.2d. Fusion-CPE suppresses the noise present between the harmonics effectively as shown in Fig.2e. The combination of spectral subtraction and log-MMSE has an advantage of effective noise suppression due to the spectral subtraction rule and minimum musical noise because of log-MMSE as demonstrated in Fig.2e.

Identification of spectral sparsity in voiced frames results in better estimation of non-stationary noise PSD. As indicated in Table 1, frequent update of noise PSD results in better speech quality. The improvement is significant over standard spectral subtraction since its success depends totally on accurate noise PSD estimation. We also modified the selection of the over-attenuation factor to avoid the suppression of low energy speech. Though such modification is not carried out in log-MMSE STSA, speech quality is improved further by the proposed noise estimation algorithm. The combination of SS and LMMSE causes further reduction of noise in noise dominant periods in the voiced frames detected using phase difference. In this way gain of LMMSE is adjusted indirectly to cause effective attenuation of noise in voiced frames. In cases of stationary noises like white noise, no improvement is observed. In some cases speech quality is even worse than the standard noise reduction algorithm which uses the VAD based noise estimation technique. Although estimating noise in each voiced frames gives good noise estimation for non-stationary noise, for stationary noise this results in extra attenuation of low energy speech in the voiced frames. Thus, the improvement is limited to non-stationary noise where good estimation of pitch frequency is available.

6. CONCLUSIONS

In this study, we presented the usage of baseband phase difference as a means of detecting the noise-dominant frequencies in the voiced speech frames. We showed that baseband phase difference can be used to improve the performance of current speech enhancement algorithms. This approach is verified by combining the proposed algorithm with spectral subtraction and log-MMSE STSA. For both of these existing algorithms, incorporating the proposed noise estimation method improves the speech quality as measured by PESQ. Spectral subtraction and log-MMSE STSA are also combined to achieve better noise reduction in voiced frames as compared to pure log-MMSE STSA. The proposed noise estimation algorithm can be combined with any of the existing speech enhancement algorithms. The estimated baseband phase difference can also be used as prior information to provide robust estimation of other parameters such as Speech Presence Probability (SPP). The usage the baseband phase difference to further enhance the performance of existing advanced noise estimation algorithms should be explored in future work.

7. REFERENCES

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

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[3] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, June 2000, vol. 3, pp. 1875–1878.

[4] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters.*, vol. 9, no. 1, pp. 12–15, 2002.

[5] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing.*, vol. 9, no. 5, pp. 504–512, 2001.

[6] K. Paliwal and L.D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 52, no. 2, pp. 153–170, February 2005.

[7] Y.Zhang and Y.Zhao, "Spectral subtraction on real and imaginary modulation spectra," in *IEEE International Conference on Acoustics, Speech and Signal Processing.*, 2011, pp. 4744–4747.

[8] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," in *International Workshop on Acoustic Signal Enhancement*, September 2012, pp. 1–4.

[9] A. Cheveign and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, pp. 1917–1930, 2002.

[10] P.C. Philipos, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[11] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[12] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979, vol. 4, pp. 208–211.

[13] "Noise samples," `http://www.ee.columbia.edu/~dpwe/sounds/noise/`.

[14] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 2, pp. 749–752.

[15] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.