PHASE RANDOMIZATION – A NEW PARADIGM FOR SINGLE-CHANNEL SIGNAL ENHANCEMENT

Akihiko Sugiyama and Ryoji Miyahara†

Information and Media Processing Laboratories
NEC Corporation
† Internet Terminal Division, NEC Engineering
1753 Shimonumabe, Nakahara-ku, Kawasaki 211-8666, Japan
aks@ak.jp.nec.com

ABSTRACT

This paper proposes a new paradigm with phase randomization for single-channel signal enhancement. In contrast to literatures which pursue better target signal quality, the new method tries to minimize artifacts in the residual noise. Applications of signal enhancement are revisited to highlight today's examples where environmental signal is often considered as a part of target and SNR may take a negative value. A signal example demonstrates that conventional signal enhancement with magnitude-only modification is insufficient from both objective and subjective points of view. A new framework with phase randomization as well as a specific algorithm is developed. Enhanced signals show that phase randomization is an integral component for sufficient enhancement. A subjective evaluation result demonstrates that the new paradigm with phase randomization is superior to the magnitude-only enhancement with statistically significant differences.

Index Terms— Speech enhancement, Noise suppressor, Digital still camera, Residual noise, Random phase

1. INTRODUCTION

Speech enhancement [1, 2, 3, 4] is an indispensable technology in these days to make a target speech contaminated by other signals much easier to listen to. Traditionally, speech enhancement is performed mostly in the frequency-domain and consists of four blocks, namely, forward and inverse transforms, noise estimation, and magnitude modification. Magnitude of the noisy (or degraded) speech is modified while the enhanced-speech phase is copied from the noisy speech. This structure has been the standard since Lim stated that the short-time spectral amplitude rather than phase is principally important for speech intelligibility [1].

In Lim's days, the most important application of speech enhancement was speech communication [1]. However, there are more applications today, such as audio-visual (AV) recording by digital still cameras (DSCs) and camcorders, which is as important as speech communication. Good intelligibility is no longer sufficient and high fidelity is essential in AV recording, especially, with today's dissemination of HD (high-definition) pictures. In addition to speech, the environmental signal is an important part of the target signal.

Mechanical noise such as zooming and auto-focusing noise during movie recording is recently recognized as a serious problem [5, 6]. It is often stronger than environmental signal and speech is sometimes absent. As a result, a negative signal-to-noise (or target-to-noise) ratio (SNR or TNR), which is not assumed in speech communication, is often encountered in AV recording. With a negative SNR, phase information of the mechanical noise is dominant in the noisy speech phase. It is no longer justified to use the noisy speech phase as the enhanced speech phase.

Phase has been paid much less attention than magnitude. Wang et al. showed that phase is practically more useful only in low SNR's where it is harder to estimate and for long windows where longer delays are inevitable [7]. This result agrees with that in [8]. Experiments by Paliwal et al. [9] and Shannon et al.[10] revealed that magnitude is much more important for a window size of 20 to 30 ms that is most widely used. Vary theoretically derived [11] that there is no perceivable speech degradation by keeping the noisy phase when the local SNR is greater than 6 dB. Loweimi et al. evaluated the clean phase combined with a noisy magnitude and the clean magnitude with a noisy phase to show that 1.1 and 2.2 PESQ [12] improvement, respectively, are achievable [13]. This result indicates that phase is important but magnitude is much more important. It was reported by Wójcicki et al. that phase modification is sometimes more effective than magnitude modification when a noisy magnitude and a modified phase are combined [14]. They proposed to add a real-valued constant which is anti-symmetric around zero to all frequency components to obtain a modified phase. SNR dependency of the constant was improved by weighting by an estimated noise [15].

Recently, some new techniques to estimate phase spectrum have been proposed. Fardkhaleghi et al. proposed three cost functions for minimizing the difference between the enhanced signal and the zero-phase Wiener-filtered signal [16]. Phase prediction from neighboring time-frequency tiles was proposed by Rad to show that there is some correlation in phase [17]. Mehmetcik et al. and Krawczyk et al. independently proposed a phase reconstruction method for harmonics in voiced sections [18, 19]. Phase in a frame can be calculated from that in the previous frame once the initial phase is given. The initial phase can be approximated by noisy phase. In the context of signal separation, a general version of MMSE (minimum mean square error) estimate that includes an estimated phase based on an estimated magnitude was proposed by Moulaee et al. [20, 21].

All of these phase estimation techniques are for better quality of the target signal component. However, the serious problem in AV recording is the quality of the residual noise. Therefore, a new technique is needed to make the artifacts in the residual noise less audible. In addition, it should also be effective for making boundaries of a noise and a non-noise sections less noticeable.

This paper proposes a new paradigm with phase randomization for single-channel signal enhancement. The following section anal-

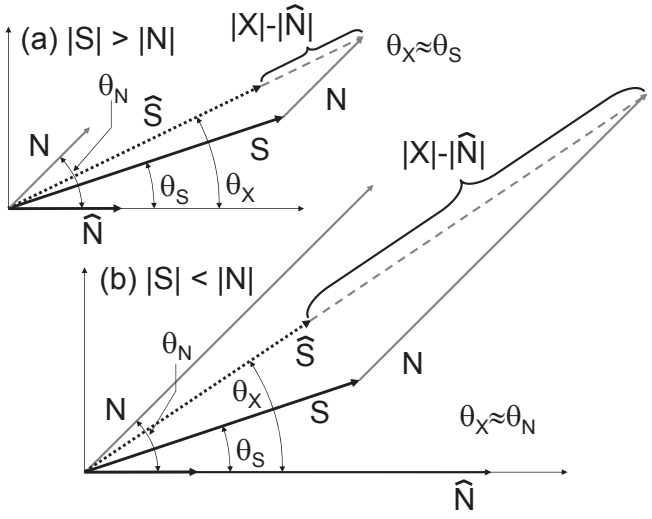

Fig. 1. Noisy signal vector with a positive and a negative SNR.

yses the phase relationship of the noisy and the enhanced signals. Section 3 presents a new signal enhancement algorithm with phase randomization. Finally, in Section 4, enhanced signals are demonstrated and subjective evaluation results are provided to confirm the effect of phase randomization.

2. SIGNIFICANCE OF PHASE

2.1. Vector Representation of Signal Enhancement

Figure 1 illustrates a vector representation of signal enhancement. Let us assume, for simplicity, that the target signal is speech and "noise" is to be suppressed. (a) and (b) exhibit a positive and a negative SNR case, respectively. The noisy signal (degraded signal) X is the resultant vector of a speech vector S and a noise vector N as

$$\begin{aligned} X &= S + N, \\ &= |S|\exp\{j\theta_S\} + |N|\exp\{j\theta_N\}, \\ &= |X|\exp\{j\theta_X\}, \end{aligned} \tag{1}$$

$$\theta_X = \tan^{-1}\frac{|S|\sin\theta_S + |N|\sin\theta_N}{|S|\cos\theta_S + |N|\cos\theta_N}, \tag{2}$$

where $|\cdot|$ is an absolute value operator and $j = \sqrt{-1}$. θ_S, θ_N, and θ_X are the speech, the noise, and the noisy speech phases. Assuming Spectral Subtraction [2], an enhanced speech vector $\hat{S}$ is expressed as in (3) where $\hat{N} = |\hat{N}|\exp\{j0\}$ is an estimated noise.

$$\hat{S} = (|X| - |\hat{N}|)\exp\{j\theta_X\}. \tag{3}$$

Equation (3) indicates that the magnitude of the enhanced signal is the difference of magnitude in the noisy speech $|X|$ and the estimated noise $|\hat{N}|$. The phase is copied from X. As the SNR becomes lower, $|N|$ becomes much bigger than $|S|$. Therefore, the phase θ_X is more dominated by θ_N as is understood from Fig. 1. It is also indicated by modifying (2) as

$$\theta_X = \tan^{-1}\frac{|S|/|N|\cdot\sin\theta_S + \sin\theta_N}{|S|/|N|\cdot\cos\theta_S + \cos\theta_N} \approx \theta_N. \tag{4}$$

On the contrary, when $|S|$ is much bigger than $|N|$, (2) is approximated as

$$\theta_X = \tan^{-1}\frac{\sin\theta_S + |N|/|S|\cdot\sin\theta_N}{\cos\theta_S + |N|/|S|\cdot\cos\theta_N} \approx \theta_S. \tag{5}$$

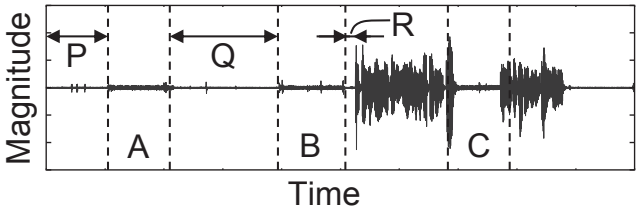

Fig. 2. Typical zooming noise with speech.

It means that θ_X is more heavily dependent on the phase of either the speech or the noise, whichever has a larger magnitude. It causes a more serious problem in a negative SNR case ($|S| < |N|$) in Fig. 1 (b) than the other in (a). The enhanced signal phase θ_X is more like the noise phase, *i.e.* θ_N. Thus, the enhanced signal should be more contaminated by the noise characteristics represented by its phase, leading to its insufficient suppression. In case of zooming noise, this situation happens in non-speech sections, where the zooming noise is larger in magnitude than the environmental signal to be preserved.

2.2. Negative SNR Example

Figure 2 shows an example of a zooming noise mixed with speech. Zooming noise is intermittent and has clear noise sections like A, B, and C where zooming noise is to be suppressed. Please note that the environmental signal should be considered as a part of target (speech), because it is to be preserved for fidelity. This fact makes the SNRs (or TNRs) in sections A and B negative. Sections P, Q, and R are environmental-noise sections with no speech nor zooming noise. In A and B, the enhanced signal level, *i.e.* the residual noise level, is adjusted to the environmental signal level for continuity. If there is any phase characteristics originating from the zooming noise in the enhanced signal, it is easily noticeable at the beginning and the ending points of zooming-noise sections. This is because sections A and B have a negative SNR and phase of the zooming noise is dominant in the enhanced-signal phase. In order to make such an artifact inaudible even at boundaries of zooming-noise sections, phase randomization is effective.

2.3. Phase Effect in Zooming Noise

In order to confirm this problem, Fig. 3 compares artificial zooming-noise suppression with and without phase randomization. In Fig. 3 (a), zooming noise in white is shown over the noisy signal in gray, which contains target speech, environmental signal, and zooming noise. A, B, and C are zooming-noise sections where the zooming noise is to be suppressed. (b) is the ideal enhanced signal that consists of target speech and environmental signal and no zooming noise at all. The result of magnitude-only modification from (a) is depicted in (c). The magnitude in (c) contains that of speech and environmental signal and no information about zooming noise. On the contrary, its phase is that of the noisy signal that contains speech, environmental signal, and zooming noise. It means that (c) is the ideal zooming noise suppression in the conventional framework based on magnitude-only modification. Although there is no clear visible difference between (b) and (c), the difference is audible. In order to highlight the invisible difference, the signals in (b) and (c) are bandlimited to a frequency range from 6 to 15 kHz and shown in (d). This frequency range was selected so that the difference between (b) and (c) is not masked by other signal components such as speech and environmental signal. The black exuding areas in (d) represent

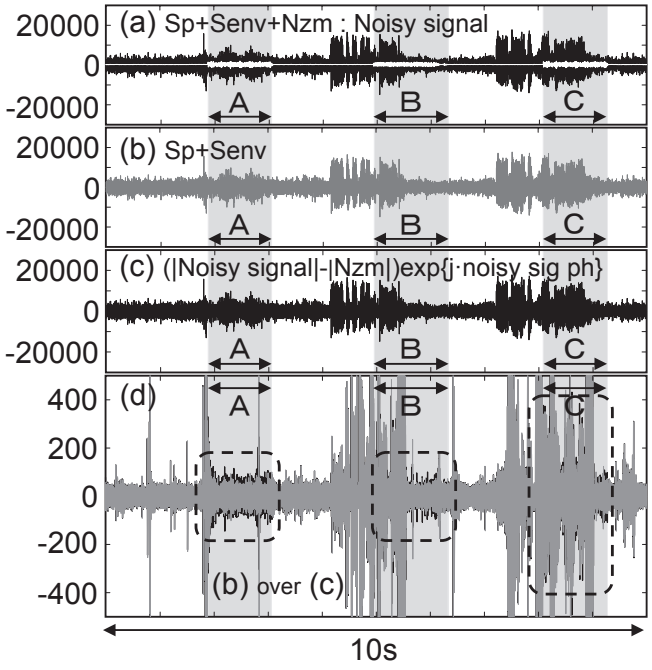

Fig. 3. Phase effect in zooming noise. (a) Zooming noise (white) over speech+env. signal+zooming noise (gray), (b) speech+env. signal (c) (a) after magnitude-only modification (subtraction of true zooming-noise magnitude), (d) (b) over (c) after bandlimitation to a frequency range from 6 to 15 kHz.

the residual zooming-noise that is audible. Because masking of the residual noise by speech and environmental signal in a frequency range from 0 to 6 kHz does not extend higher frequencies, the residual noise in black in (d) is audible.

3. PHASE RANDOMIZATION – A NEW PARADIGM

3.1. Concept and Implementation of Phase Modification

Figure 3 clearly suggests that some phase modification is needed to make the residual noise inaudible. However, it is still unclear what kind of modification to be applied. The residual zooming (or any) noise is audible because it carries significant character from the noise in its phase. Although it is not easy to find out exactly which character makes the residual noise audible, the least thing one can say is that it is caused by some phase correlations along the time and the frequency axes. It is a natural consequence that phase modification is implemented as phase randomization for decorrelation.

It should be noted that phase randomization is applied to frequency components that are perceived, after magnitude modification, as residual noise. This is because residual noise is audible only when its magnitude exceeds a level determined by the principle of psychoacoustics. When the magnitude takes a near-zero value, phase does not matter. In addition, large-magnitude signals should be accompanied by its original phase because they should have a high SNR. Thus, small magnitude signals above the psychoacoustic masking threshold should have a randomized phase.

3.2. Detailed Design of Enhancement Algorithm

Based on these considerations, a more specific algorithm of a zooming noise suppression was developed as shown in Fig. 4. Noise

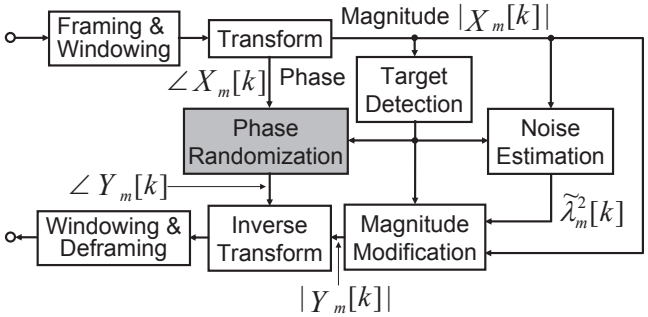

Fig. 4. Blockdiagram of zooming noise suppressor.

suppression is performed only in noise sections that is signaled by the DSC. Significant target signal components are detected as peaks in the frequency domain and their magnitude is not modified. The noisy signal phase is copied for the target signal components. Magnitudes of non-target components, which are small, are replaced with an estimated environmental signal level and their phases are randomized.

Peaks are detected in the way described in [22] and given a peak flag $p_n[k] = 1$. Otherwise, $p_n[k] = 0$. In the process of peak detection, hangovers are considered. Hangover is determined when there is any peak in a past period to fill gaps in a speech section. A hangover index $h_m[k]$ is set as

$$h_m[k] = \begin{cases} 1 & \sum_{i=m-Q+1}^{m} p_i[k] > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where integers m and k are the frame and the frequency index and an integer Q is a hangover period.

An estimate of the environmental signal $\tilde{\lambda}_m[k]^2$ is updated based on a first-order leaky integration (recursive filter) with a leaky factor γ in non-peak frequency bins.

For a simple description, a suppression flag $f_m[k]$ that indicates detailed suppression is introduced. For peak bins and non-peak-non-hangover bins, $f_m[k]$ is defined by

$$f_m[k] = \begin{cases} 0 & p_m[k] = 1 \\ 2 & p_m[k] + h_m[k] = 0 \end{cases}. \tag{7}$$

For non-peak-hangover bins,

$$f_m[k] = \begin{cases} 2 & |X_m[k]|^2 \geq |X_{m-1}[k]|^2 + \delta\text{dB} \\ 0 & |X_m[k]|^2 < |X_{m-1}[k]|^2 \\ 1 & \text{otherwise} \end{cases}, \tag{8}$$

Based on the suppression flag $f_m[k]$, amplitude of the noise suppressed signal $|Y_m[k]|^2$ is obtained by

$$|Y_m[k]|^2 = \begin{cases} |X_m[k]|^2 & f_m[k] = 0 \\ |X_{m-1}[k]|^2 & f_m[k] = 1 \\ \tilde{\lambda}_m^2[k] & f_m[k] = 2 \end{cases}. \tag{9}$$

For $f_m[k] = 2$, a randomization index $r_m[k]$ is set to 1 and the phase is randomized. Otherwise, $r_m[k]$ is set to 0 to preserve the noisy-signal phase.

The input noisy signal phase $\theta_{X_m[k]}$ is randomized based on $r_m[k]$ in Phase Randomization to obtain the enhanced signal phase $\theta_{Y_m[k]}$ as

$$\theta_{Y_m[k]} = \theta_{X_m[k]} + r_m[k] \cdot \phi_m[k], \tag{10}$$

where $\phi_m[k]$ is a random value between $\pm\pi$. The enhanced signal at the output is reconstructed from $|Y_m[k]|^2$ and $\theta_{Y_m[k]}$.

Table 1. Speech and noise used for evaluations.

Speech	Male and female speech
Noise 1	Street noise with crow caws
Noise 2	Office noise with telephone rings
SNR[1]	$-0.62 \leq SNR \leq +5.7$

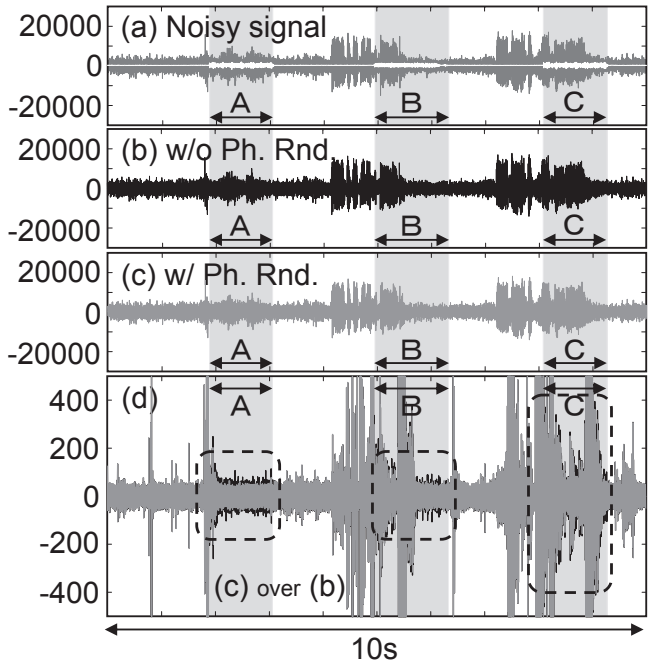

Fig. 5. Signal enhancement for zooming noise. (a) Noisy signal (speech+env. signal+zoom. noise), (b) Enhanced signal w/o phase randomization, (c) Enhanced signal w/ phase randomization, (d) (c) over (b) after bandlimitation to $6 - 15$ kHz.

4. EVALUATIONS

Evaluations were performed using zooming noise of a digital still camera and environmental signal on a street and in an office. Male and female speech signals were also prepared. All signals were sampled at 44.1 kHz. More detailed information about these signals is shown in Tab. 1. Q, δ, and γ were set to 16, 3dB, and 0.98, respectively.

4.1. Evaluation by signals

Figure 5 shows signal enhancement for a zooming noise. The noisy signal contained speech, environmental signal (Noise 1), and zooming noise as depicted in Fig. 5 (a). The white trajectory at the center of the ordinate is the zooming noise. (c) and (b) compare the enhanced signal with and without phase randomization. Although there is difference in zooming sections A, B, and C, it may not be easy to see it clearly. For better comparison, the signals in (b) and (c) were bandlimited to a frequency range from 6 to 15 kHz. The bandlimited version of (c) (gray) is overlaid on that of (b) (black) in Fig. 5 (d). The gray curve with phase randomization in zooming sections achieves a comparable signal level to that of the neighboring regions. On the contrary, the black curve without phase randomization has a higher enhanced signal level in zooming sections. This is a sign of insufficient suppression (significant residual noise) as will be confirmed by a subjective evaluation.

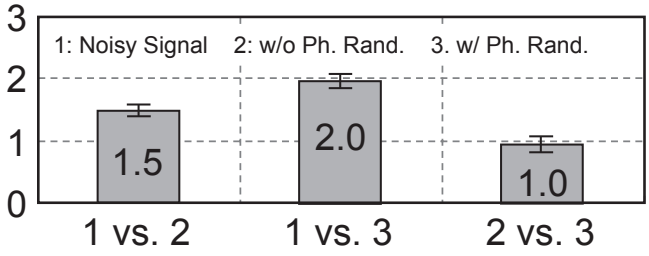

Fig. 6. Subjective evaluation result with modified CCR.

4.2. Evaluation by subjective assessment

Male and female speech with three different levels including the zero level were mixed with the zooming noise and the street or the office noise. A total of 14 subjects were asked to give an integer score between ± 3 following a 7-grade modified CCR (Comparison Category Rating)[2][23].

Figure 6 depicts the result in a bar chart with a 95% confidence interval. "1," "3," and "2," in the figure represents the noisy signal, the enhanced signal with and without phase randomization. A positive score means that "β" of "α vs. β" is superior to "α." The leftmost bar compares the subjective quality of the noisy signal and that of the enhanced signal without phase randomization. This corresponds to the subjective quality of the conventional magnitude-only modification. The score is 1.5 and it has a statistically significant difference because the lower limit of the 95 % confidence interval lies in the positive region. The center bar exhibits the subjective quality of the enhanced signal with magnitude modification and phase randomization. It outperforms the magnitude-only modification with a score of 2.0 and a statistically significant difference. Finally, the rightmost bar represents the direct comparison of subjective qualities by enhanced signals with and without phase randomization. It is clearly demonstrated that phase randomization in zooming noise suppression brings improved subjective quality of almost 1.0 with a statistically significant difference.

5. CONCLUSION

A new paradigm for single-channel signal enhancement with phase randomization has been proposed for the purpose of minimum artifacts in the residual noise. A vector representation of signal enhancement has revealed the significance of phase in negative SNR cases. It has been demonstrated with an example that magnitude-only modification is not sufficient for high fidelity of the enhanced signal. A zooming noise suppressor in the new paradigm has been designed and evaluated by signals as well as subjective assessment. A subjective evaluation result has shown that phase randomization successfully improved a modified CCR score by 1.0 with a statistically significant difference over its counterpart with no phase modification.

[1]This SNR applies to only subjective evaluation of zooming noise suppression. Speech plus environmental signal and the AF noise are represented by S and N, respectively. Mixtures without speech are also included.

[2]The modified CCR method uses processed reference samples but without noise suppression whereas the standard CCR method uses unprocessed reference samples.

6. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of speech," Proc. of IEEE, Vol. 67, No. 12, Dec. 1979.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. ASSP, vol. 27, no. 2, pp.113–120, Apr. 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. ASSP, vol. 32, no. 6, pp.1109–1121, Dec. 1984.

[4] J. Benesty, S. Makino, and J. Chen, Eds., "Speech enhancement," Springer, Berlin, Mar. 2005.

[5] A. Sugiyama, T. Maeda, and K. Park, "A mechanical noise suppressor based on *a priori* information for digital still cameras and camcorders," Proc. of ICCE2011, pp. 426–427, Jan. 2011.

[6] A. Sugiyama and R. Miyahara, "An auto-focusing-noise suppressor for cellphone movies based on multiple noise references," Proc. of ICCE2012, pp. 45–46, Jan. 2012.

[7] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," IEEE Trans. ASSP, vol. 30, no. 4, pp.679–681, Aug. 1982.

[8] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," Proc. IEEE, Vol. 69, No. 5, pp.529–541, May 1981.

[9] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," Speech Commun., Vol. 45, pp.153–170, 2005.

[10] B. J. Shannon and K. K. Paliwal, "Role of phase estimation in speech enhancement," Proc. of INTERSPEECH2006-ICSLP, pp.1423–1426, Sep. 2006.

[11] P. Vary, "Noise suppression by spectral magnitude estimation, – Mechanism and theoretical limits –," Signal Processing, pp.387–400, vol. 8, no. 4, July 1985.

[12] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-endspeech quality assessment of narrowband telephone networks and speech codecs," 2000.

[13] E. Loweimi, S. M. Ahadi, S. Loveymi, "On the importance of phase and magnitude spectra in speech enhancement," Proc. ICEE2011, CD-ROM, May 2011.

[14] K. K. Wójcicki, M. Milacic, A. Stark, J. G. Lyons and K. K. Paliwal, "Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement", IEEE Signal Processing Letters, Vol. 15, pp. 461–464, Dec. 2008.

[15] A. P. Stark, K. K. Wójcicki, J. G. Lyons and K. K. Paliwal, "Noise driven short time phase spectrum compensation procedure for speech enhancement", Proc. INTERSPEECH 2008, pp. 549–552, Sep. 2008.

[16] P. Fardkhaleghi and M. H. Savoji, "New approaches to speech enhancement using phase correction in Wiener filtering," Proc. IST2010, pp.895–899, Dec. 2010.

[17] A. B. Rad and T. Virtanen, "Phase spectrum prediction of audio signals," Proc. ISCCSP2012, CD-ROM, May 2012.

[18] E. Mehmetcik and T. Ciloglu, "Speech enhancement by maintaining phase continuity," Proc. SIU2012, CD-ROM, Apr. 2012.

[19] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," Proc. IWAENC2012, CD-ROM, Sep. 2012.

[20] P. Moulaee and R. Martin, "On phase importance in parameter estimation for single-channel source separation," Proc. IWAENC2012, CD-ROM, Sep. 2012.

[21] —— "On phase importance in single-channel source separation," Proc. AUDIS2012, CD-ROM, Sep. 2012.

[22] ISO/IEC 11172-3:1993, "Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3 : Audio," Aug. 1993.

[23] 3GPP TS 06.77 V8.1.1, "Minimum performance requirements for noise suppresser application to the AMR speech encoder," Apr. 2001.