Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 81 (2016) 138–147

www.elsevier.com/locate/specom

Phase distortion resulting in a just noticeable difference in the perceived quality of speech

Roger Chappel, Belinda Schwerin*, Kuldip Paliwal

Signal Processing Laboratory, School of Engineering, Griffith University, Nathan Campus, Brisbane, QLD 4111, Australia

Received 15 June 2015; received in revised form 28 March 2016; accepted 11 April 2016
Available online 22 April 2016

Abstract

Common speech enhancement methods based on the short-time Fourier analysis–modification–synthesis (AMS) framework, modify the magnitude spectrum while keeping the phase spectrum unchanged. This is justified by an assumption that the phase spectrum can be seen as unimportant to speech quality, and hence the noisy phase spectrum can be used as a reasonable estimate of the clean phase spectrum in signal reconstruction. In this work we show, by using an ideal magnitude estimator, that corruption in the phase spectrum can still affect the quality of the resulting speech in low SNR environments. Furthermore, we quantify the distortion in the phase spectrum which can be tolerated before it begins to affect speech quality. This is done through a series of experiments, using both subjective and objective tests, and statistical analysis to evaluate the results. The results show that the phase spectrum computed from noisy speech can be used as an estimate of the phase spectrum of the clean signal without noticeably affecting perceived speech quality, only if the segmental SNR of the noisy speech signal is greater than 7 dB.

Keywords: Just noticeable difference (JND); Phase spectrum; Speech enhancement; Short-time Fourier analysis; Analysis–modification–synthesis (AMS).

1. Introduction

The enhancement of speech has received much attention in recent years, both as a way of improving the human listening experience across various devices and environments, and to improve the performance of automatic speech recognition systems. As a result, there is an extensive number of speech enhancement methods in the literature. Some process the speech signal in the time domain, others in the frequency domain, some modifying either the magnitude or phase spectrum only, others process the complex spectrum. The choice of which method is best suited to an application is influenced by many factors, including the end purpose of the enhanced signal, any computational constraints, and most significantly, the type and level of noise present in the signal. Some of the most popular speech enhancement methods are Spectral subtraction (Boll, 1979; Lim and Oppenheim, 1979), MMSE magnitude estimation (Ephraim and Malah, 1984), Kalman filtering (Paliwal and Basu, 1987), Wiener filtering (Wiener, 1949) and Subspace methods (Ephraim and Van Trees, 1995). A detailed description of these methods can be found in (Loizou, 2007).

Many of the popular enhancement methods in the literature process speech signals in the frequency domain with a short-time Fourier analysis–modification–synthesis (AMS) framework (*e.g.*, Boll, 1979; Berouti et al., 1979; Ephraim and Malah, 1984; 1985), and modify only the short-time magnitude spectrum[1] (MS) of the noisy speech signal, in order to suppress noise and improve quality. Speech is then reconstructed by combining the short-time phase spectrum (PS) of the noisy signal with the processed MS. This use of the noisy PS in stimuli reconstruction is typically justified by the assumption that the PS carries little speech information when

* Corresponding author. Tel.: +61 755529296.
E-mail address: belsch71@gmail.com, b.schwerin@griffith.edu.au (B. Schwerin).

[1] In the remainder of this paper, when referencing the magnitude and phase spectra the STFT modifier will be implied.

http://dx.doi.org/10.1016/j.specom.2016.04.005
0167-6393/

processing stimuli using short window durations (Oppenheim et al., 1979; Shannon and Paliwal, 2006; Wang and Lim, 1982). Use of the noisy phase spectrum is also justified by the fact that it can be shown to be the minimum mean square error estimate of the clean phase spectrum (Ephraim and Malah, 1984).

More recent studies suggest that the PS can contribute useful information to speech intelligibility (Alsteris and Paliwal, 2004; 2006; Paliwal and Alsteris, 2003; Shi et al., 2006), as well as to quality (Paliwal et al., 2011; Vary, 1985), motivating investigations into the benefits of processing phase for speech enhancement (*e.g.*, Stark et al., 2008; Krawczyk and Gerkmann, 2014; Mowlaee and Kulmer, 2015). More specifically, Paliwal et al. (2011) have used noisy speech as input to an AMS system, replaced the noisy phase spectrum by the corresponding clean phase spectrum, and found the quality of the synthesised speech to be better than that of the noisy speech. Vary (1985), on the other hand, used clean speech as input to the AMS system, modified the phase spectrum by adding phase distortion to it, and found audible roughness in the synthesised speech provided the amount of additive distortion was greater than a certain threshold. When this distortion was less than this threshold, the synthesised speech sounded similar to the original clean speech. Through informal listening, he found this threshold to be $\pi/8$ to $\pi/4$. He related this threshold analytically to an instantaneous spectral signal-to-noise ratio (I-SNR) value equal to 6 dB (Vary, 1985). This threshold can be called the just noticeable difference (JND) in the phase spectrum.

In the present paper, our aim is to determine, through formal listening experiments, this JND in terms of additive phase distortion introduced to the phase spectrum. For this purpose we conduct four experiments, which are reported in the sections below. In the first experiment, we consider the approach of Vary (1985), and quantify the additive phase distortion which results in a JND in speech quality. In the second experiment, we quantify the JND with respect to I-SNR. The third experiment, then quantifies the JND with respect to a global segmental SNR, which can be applied to an entire speech utterance. The fourth experiment quantifies the JND with respect to I-SNR where the clean magnitude spectrum is estimated from the noisy spectrum using the log-MMSE (Ephraim and Malah, 1984) speech enhancement algorithm. Findings are then summarised in the last section.

2. Analysis–modification–synthesis framework

As mentioned in the introduction, many enhancement methods utilise a short-time Fourier analysis–modification–synthesis (AMS) framework, and modify just the magnitude spectrum, using the phase spectrum calculated from the noisy signal in stimuli reconstruction. In this study, we aim to quantify the effect of noise in the phase spectrum on the resulting speech quality. Therefore, like previous efforts to investigate the relative significance of the magnitude and phase spectral components, we make use of the short-time Fourier AMS framework. Using this framework, the speech signal is decomposed into its short-time magnitude and phase spectral components, which can be modified according to the associated treatment method (as described for each experiment). For reference, the AMS framework (as applied in this work) is described as follows.

In the analysis stage, short-time Fourier transform (STFT) analysis is applied to the discrete-time input signal to produce the complex frequency spectrum $X(n,\ k)$. For a discrete-time signal $x(n)$, the STFT is given by

$$X(n,k)=\sum_{l=-\infty}^{\infty}x(l)w(n-l)e^{-j2\pi kl/N}, \tag{1}$$

where n refers to the discrete-time index, k is the index of the discrete frequency, N is the frame duration (in samples), and $w(n)$ is the analysis window function. In speech processing, a frame duration of 20–40 ms is typically used, with a Hamming window used as the analysis window function (Huang et al., 2001; Paliwal and Wójcicki, 2008; Picone, 1993). In polar form, the STFT of the speech signal can be written as

$$X(n,k)=|X(n,k)|\mathrm{e}^{j\angle X(n,k)}, \tag{2}$$

where $|X(n,\ k)|$ denotes the short-time magnitude spectrum and $\angle X(n,\ k)$ denotes the short-time phase spectrum.

In the modification stage, the magnitude and/or phase spectrum can be modified according to the treatment being applied. In this work, our goal is to investigate the effects of adding noise to the phase spectrum only. Therefore, we only modify the phase spectrum while leaving the magnitude spectrum unchanged. The modified complex spectrum $\widehat{Y}(n,k)$ is therefore given by the combination of the clean magnitude spectrum $|X(n,\ k)|$ and the modified phase spectrum $\angle Y(n,\ k)$, that is

$$\widehat{Y}(n,k)=|X(n,k)|\mathrm{e}^{j\angle Y(n,k)}. \tag{3}$$

Finally, the synthesis stage reconstructs the modified speech, $y(n)$, by applying the inverse STFT to the modified spectrum, followed by least-squares overlap-add synthesis (Quatieri, 2002)):

$$y(n)=\sum_{l=-\infty}^{\infty}\left[\left(\frac{1}{N}\sum_{k=0}^{N-1}Y(l,k)\mathrm{e}^{j2\pi nk/N}\right)w_s(l-n)\right]. \tag{4}$$

Here, the modified Hann window (Griffin and Lim, 1984) was used as the synthesis window function $w_s(n)$.

A block diagram of the AMS framework used in this work is shown in Fig. 1. Throughout all experiments in this paper we have used a frame duration t_w of 32 ms with a 4 ms frame shift, and an FFT analysis length of $2N$ (where $N=t_wF_s$, and F_s is the sampling frequency of clean stimuli).

3. Background

How the noise, when added to the speech signal, corrupts the phase spectrum can be viewed in terms of complex vector analysis. Let us consider an additive noise model

$$y(n)=x(n)+d(n) \tag{5}$$

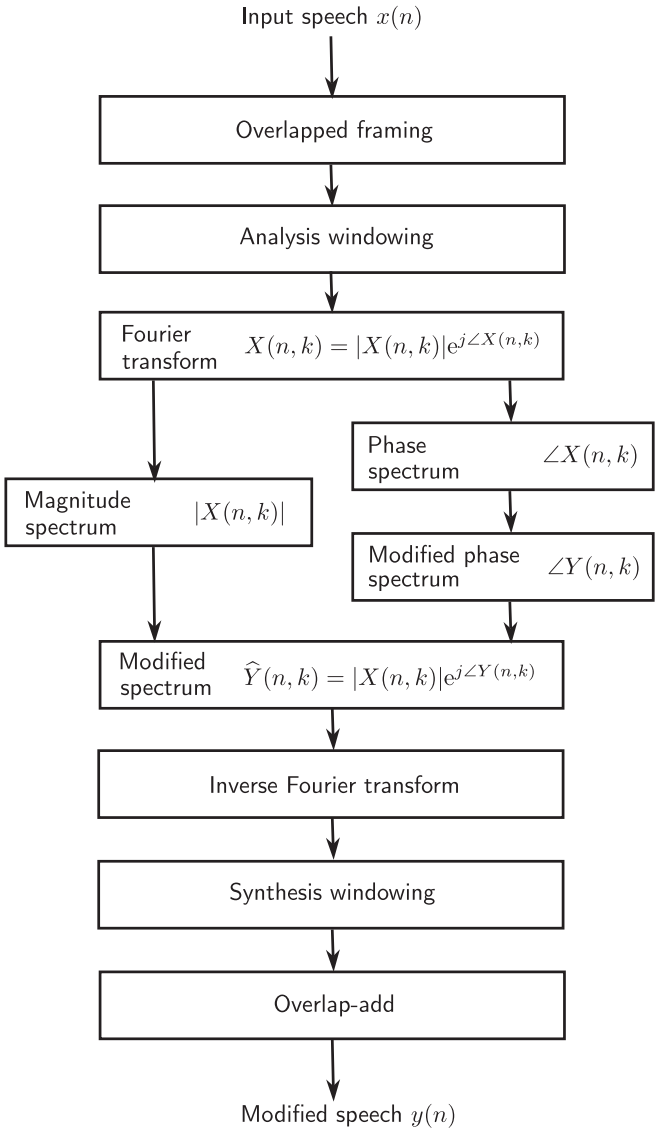

Fig. 1. Block diagram of the modified AMS framework used to the construct stimuli in this work.

where $y(n)$, $x(n)$, and $d(n)$ denote the discrete-time signals of noisy speech, clean speech and noise, respectively. Upon framing the signal and taking the STFT, Eq. (5) can be represented in polar form as

$$|Y(n,k)|e^{j\angle Y(n,k)} = |X(n,k)|e^{j\angle X(n,k)} + |D(n,k)|e^{j\angle D(n,k)}, \quad (6)$$

where $|Y(n,k)|$, $|X(n,k)|$ and $|D(n,k)|$ are the short-time magnitude spectra of the noisy speech, clean speech and additive noise signals, and $\angle Y(n,k)$, $\angle X(n,k)$ and $\angle D(n,k)$ are the short-time phase spectra of the noisy, clean and noise signals. In this equation, n is the time index (or beginning sample of the frame as shown in Eq. (1)), and k is the frequency index (or bin). In the rest of this section, we drop the indices n and k, for clarity of the presentation.

For the purpose of understanding the interactions between spectral terms, Eq. (6) can be represented geometrically, as illustrated in Fig. 2, to provide vector representations of the complex frequency spectrum for the clean, noisy and noise signals, $\vec{X}$, $\vec{Y}$ and $\vec{D}$, respectively.

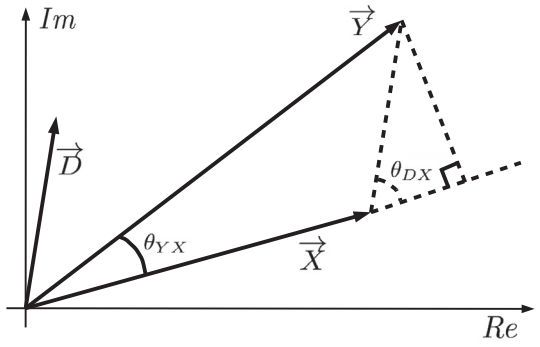

Fig. 2. Vector diagram representing the phase difference between the spectral vectors for clean speech, noisy speech and noise (Loizou, 2007) for a given frequency bin and frame index.

Applying trigonometry, the phase difference between the clean and noisy spectra, denoted by θ_{YX}, becomes

$$\tan(\theta_{YX}) = \frac{|D|\sin(\theta_{DX})}{|X| + |D|\cos(\theta_{DX})}. \quad (7)$$

As θ_{DX} becomes 90 deg (*i.e.*, $\theta_{DX}=\frac{\pi}{2}$), θ_{YX} reaches a maximum. Substituting $\theta_{DX} = \frac{\pi}{2}$ in Eq. (7) gives

$$\theta_{YX_{\max}} = \tan^{-1}\left(\frac{|D|}{|X|}\right). \quad (8)$$

The instantaneous spectral signal to noise ratio (I-SNR) ξ, can be defined for the given frequency bin and frame as the power of the clean vector divided by the power of the noise vector; that is $\xi = \frac{|X|^2}{|D|^2}$. Therefore, the maximum phase difference between the clean and noisy vectors can be expressed in terms of I-SNR, as

$$\theta_{YX_{\max}} = \tan^{-1}\left(\frac{1}{\sqrt{\xi}}\right). \quad (9)$$

In the remainder of this work, we will refer to $\theta_{YX_{\max}}$ as $\theta_{\max}$.

Fig. 3 illustrates the relationship between $\theta_{\max}$ (given by Eq. (9)) and the I-SNR, calculated as follows:

$$\text{I-SNR}_{dB} = 10\log_{10}\frac{|X|^2}{|D|^2} = 10\log_{10}\xi. \quad (10)$$

Here, $\theta_{\max}$ is shown to decrease as the I-SNR increases. When the I-SNR is low, $\theta_{\max}$ becomes large, indicating that phase changes in the noisy signal are likely to be large.

As mentioned in the introduction, Vary (1985) identified the JND for this phase difference $\theta_{\max}$ through informal listening as being between $\pi/8$ and $\pi/4$. From Fig. 3, this corresponds to an I-SNR of from 8 dB to 0 dB. Experiments presented in the remainder of this work aim to identify this JND more accurately by using formal human listening tests.

4. Speech corpus

In our experiments, we employ the Noizeus speech corpus (Hu and Loizou (2007)). This corpus is composed of

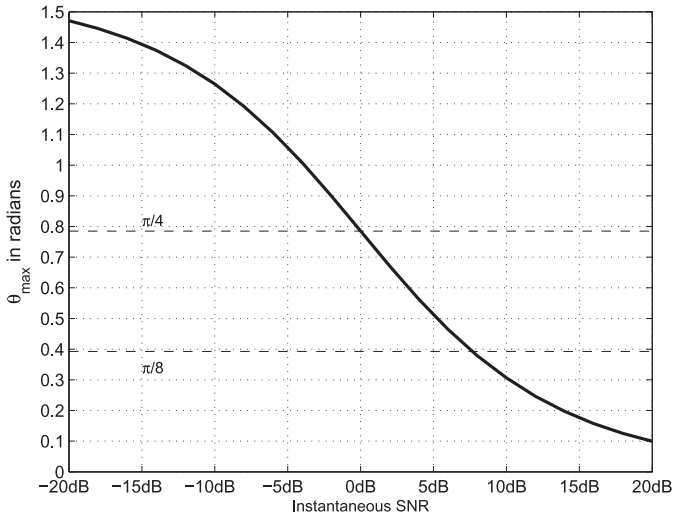

Fig. 3. Maximum phase difference between the clean and noisy speech as a function of I-SNR (Loizou, 2007).

30 phonetically-balanced sentences belonging to six speakers (three males and three females). Stimuli have been sampled at 8 kHz and filtered to simulate receiving frequency characteristics of telephone handsets. The corpus comes with non-stationary noises at different SNRs. However, for our experiments we keep the clean part of the corpus and generate noisy stimuli by degrading the clean stimuli as required for each experiment.

5. Experiment 1: JND in terms of phase distortion

In the first experiment, the goal was to quantify how much distortion can be introduced to the clean phase spectrum before the quality of the resulting (reconstructed) speech becomes audibly degraded. For this purpose, numbers uniformly distributed between $[-\theta_{max}, \theta_{max}]$ are added to the clean phase spectrum in the modification stage of the AMS system. The use of a uniform distribution for the distortion added to the clean phase spectrum ensured we maintained consistency with the earlier work of Vary (1985), and the applied range ensured that the maximum difference between the modified and clean phase angles is less than θ_{max}. By varying this θ_{max} and evaluating the quality of reconstructed stimuli using objective and subjective tests, we find the smallest maximum phase angle difference required to produce a JND in speech quality.

5.1. Stimuli generation

The stimuli used in the first experiment were generated by modifying the clean speech utterances of the Noizeus corpus (see Section 4). This modification involved processing speech utterances using the short-time AMS framework described in Section 2, and modifying the phase spectrum while leaving the magnitude spectrum unchanged. The rationale for this type of modification was twofold: it allowed us to directly observe the quality effects attributed to changes in the phase spectrum only; and it tells us when the noisy phase spectrum can be used as an estimate of the clean phase spectrum in a speech enhancement algorithm.

To modify the phase spectrum, a uniformly distributed random vector ϕ of length $2N$, with the first N elements uniformly distributed between $-\pi$ and π, and the remaining N anti-symmetric about the midpoint, was initially generated.[2] The modified phase spectrum was then constructed as

$$\angle Y(n,k) = \angle X(n,k) + \alpha\phi, \tag{11}$$

where $\angle Y(n, k)$ is the modified phase spectrum, $\angle X(n, k)$ is the phase spectrum of the clean signal, α is a multiplier, and $\alpha\pi = \theta_{max}$. By varying α between 0 and 1, we can investigate the effect of θ_{max} on the quality of reconstructed speech. In the present experiment, we have taken α in the range of 0.1–0.28 (in steps of 0.02). Applying $\angle Y(n, k)$ in Eq. (3), the modified signal was then reconstructed as described in Section 2.

5.2. Experiments

Objective and subjective experiments were conducted using the procedures described as follows.

5.2.1. Objective experiment

Objective experiments in this work make use of the perceptual evaluation of speech quality (PESQ) metric (Rix et al., 2001). PESQ provides a score between 0.5 and 4.5 which aims to predict the quality of a degraded speech signal.[3] A score of 0.5 indicates that the quality of the speech is low, and a score of 4.5 indicates that the speech is of high quality. Ten different modifications (corresponding to α values from 0.1 to 0.28 insteps of 0.02) of each sentence were constructed. The objective experiment then calculated the mean PESQ score across the 30 sentences of the corpus for each modification being investigated.

5.2.2. Subjective experiment

Subjective tests, which can provide a more accurate representation of stimuli quality, were also conducted. Ten English speaking listeners with normal hearing participated in each test. These tests were in the form of AB human listening experiments, in which listeners were asked to select a preferred stimulus for each stimuli pair. One clean stimulus was always paired with a modified (treated) stimulus. The playlist included each clean-modified stimulus-pair with both the clean stimulus played first (clean then modified), and the clean stimulus played second (modified then clean). This was done to compensate for any bias associated with listening order. In each test, stimulus-pairs were played back to the participants in randomised order.

Listeners were presented with three labelled options after listening to each stimulus-pair and asked to make a selection.

[2] Different randomly generated vectors were used for each frame and sentence.

[3] PESQ does not necessarily provide a reliable prediction of speech quality for phase-modified speech. However, it can indicate the effect of increasing distortion on speech quality.

The first and second options were used to indicate preference for the corresponding stimulus, while the third option was used to indicate that both stimuli sounded the same. Pairwise scoring was used, with a score of +1 awarded to the preferred version (either clean or modified utterance) and +0 to the other. For a pair of stimuli sounding the same, both (clean and modified) were awarded a score of +0.5. Listeners could replay stimuli if required.

Since it was unreasonable to use the entire corpus, two utterances (one from a male speaker, and one from a female speaker), modified as described in Section 5.1 for each value of α (*e.g.*, $\alpha = 0.1, 0.12, \ldots, 0.28$), were utilised for experiment 1. Thus, a total of 20 modified utterances were generated for the first subjective experiment, and since each stimulus-pair was also played in reverse order, each participant scored 40 stimulus-pairs in this listening test.

5.3. Results

Fig. 4 shows results of the objective experiment (see Section 5.2.1) which calculated the mean PESQ scores for signals constructed using each value of α (*i.e.*, $\alpha = 0.1, 0.12, \ldots, 0.28$). This graph indicates that as α increases, speech quality is reduced in an almost linear relationship. This pattern alone does not indicate the $\theta_{\max}$ corresponding to a JND in signal quality. However, it does indicate the relationship between degradation of the phase spectrum and perceived signal quality.

The results of the subjective listening tests are shown in Fig. 5. Mean scores were calculated, over the ten participating listeners, for modified stimuli using each α value. A score of 50% indicates that the listeners could not identify a difference between the clean and modified utterances, however, a score of 0% indicated that the listeners always found the modified utterance to be worse than the clean utterance in terms of speech quality. As illustrated in this figure, when α is small, little difference in quality is perceived. However, as α increases, differences in quality become audible.

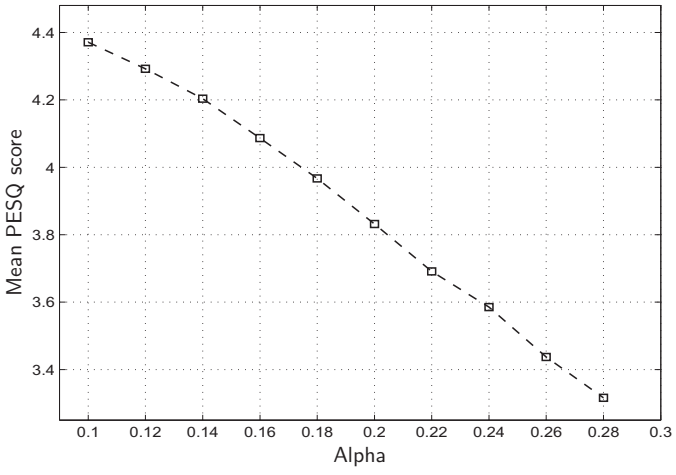

Fig. 4. Objective results in terms of mean PESQ scores for modified stimuli described in Section 5.1.

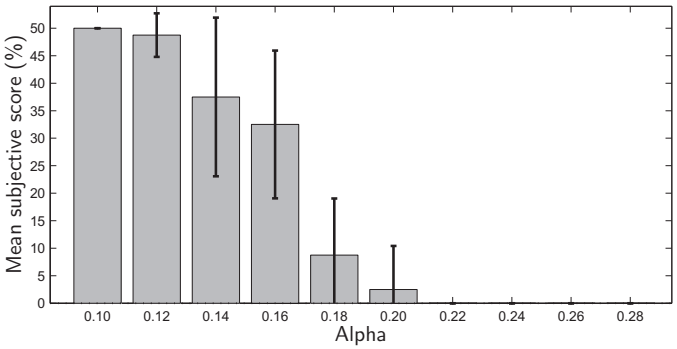

Fig. 5. Mean subjective scores (with standard error bars) for modified stimuli described in Section 5.1.

One-way analysis of variance (ANOVA) was used to test differences in subjective scores for ten α conditions. There was a significant effect observed for α ($F(9, 90) = 81.10$, $p < 0.0005$). *Posthoc* multiple comparison tests using Tukey's significant difference criterion (at 0.05 level of significance) were conducted to test statistical significance between the subjective scores for each α value. It was found that $\alpha = 0.1$ and 0.12 had a significantly higher score than the remaining eight α values ($0.14 - 0.28$). The smallest value of α which had a statistically significant difference in subjective score was $\alpha = 0.14$. Since $\theta_{\max} = \alpha\pi$, this means the JND for $\theta_{\max}$ is 0.14π. In other words, a difference in quality becomes noticeable for $\theta_{\max} \geq 0.14\pi = 0.44$.

5.4. Discussion

The above objective and subjective test results show there is a decrease in quality as α is increased. Small modifications of the phase spectrum are not noticed by the human listeners; however, as it is degraded further (by further increase in $\theta_{\max}$), the difference between the clean and noisy speech becomes noticeable.

Further, the subjective results showed that $\alpha = 0.14$ (corresponding to $\theta_{\max} = 0.44$ rad), is the largest angle of deviation from the clean phase spectrum that the modified phase spectrum can become before a difference in quality is perceived by the listener, suggesting the JND to be at $\theta_{\max} = 0.44$ rad. Relating this result to the relationship given in Eq. (9) (and shown in Fig. 3), a maximum phase difference of 0.44 corresponds to an I-SNR of 7 dB. This indicates, that the noisy phase spectrum can be used as an approximation of the clean phase spectrum without introducing audible degradation if the I-SNR is >7 dB, (or, the maximum phase angle deviation does not exceed 0.44 rad).

6. Experiment 2: JND in terms of I-SNR

In the previous experiment, we found the JND in terms of phase distortion $\theta_{\max}$, and related it to I-SNR using Eq. (9), to find the JND in terms of I-SNR in an indirect manner. We also used a uniform distribution for the distortion added to the phase spectrum, for consistency with the earlier work of

Vary (1985). However, we have observed that in practice the distribution of phase distortion is not uniform, but is more like a Gaussian distribution with a variance which increases with decreasing I-SNR. Therefore in this section, we perform a more direct and meaningful evaluation of the JND in terms of I-SNR. For this purpose, we conduct an experiment where I-SNR is set explicitly to a desired level, and the effect on quality is evaluated.

6.1. Stimuli generation

The clean stimuli from the Noizeus corpus (see Section 4) were processed using the short-time AMS framework outlined in Section 2. In the modification stage, the clean magnitude spectrum was again preserved, and the clean phase spectrum was replaced with a degraded phase spectrum.

To create the degraded phase spectrum, a noisy spectrum $Y(n,\ k)$ was generated such that the I-SNR across all STFT frequency bins had the required SNR. Given that the I-SNR, can be calculated in the frequency domain as

$$\text{I-SNR}(n,k) = 10\ \log_{10}\left(\frac{|X(n,k)|^2}{|D(n,k)|^2}\right), \tag{12}$$

then the magnitude of the noise spectrum required for $Y(n,\ k)$ to have an I-SNR of ξ is

$$|D(n,k)| = \sqrt{\frac{|X(n,k)|^2}{10^{\frac{\xi}{10}}}}. \tag{13}$$

Using random phase ϕ uniformly distributed between, $-\pi$ and π, the complex noise spectrum is constructed as

$$D(n,k) = |D(n,k)|e^{j\phi}. \tag{14}$$

Assuming noise is additive, the noisy spectrum can then be generated as

$$Y(n,k) = D(n,k) + X(n,k). \tag{15}$$

The phase of this noisy spectrum $Y(n,\ k)$ was then combined with the magnitude of the clean spectrum $|X(n,\ k)|$ using Eq. (3) to construct the modified spectrum $\widehat{Y}(n,k)$. This modified spectrum was then used for signal reconstruction as described in Section 2.

For the purpose of experimentation, each speech utterance of the Noizeus corpus was processed using the above method to generate stimuli with each of the following ten I-SNR levels: -3, 2, 4, 5, 6.5, 7, 8, 10, and 15 dB.

6.2. Results of experiments

The aim of this experiment was to quantitatively measure the effect of distortion in the phase spectrum at different I-SNRs on the quality of the resulting speech. The PESQ metric described in Section 5.2.1 was again used for the objective evaluation of speech quality. Results of this experiment, in terms of mean PESQ scores, are shown in Fig. 6. Unlike the objective test in experiment 1, the mean PESQ of the utterances in this experiment have a nonlinear decline for I-SNR values lower than 8 dB, with a rapid drop-off seen below around 4 dB. This reduction in quality is examined further by the following subjective test.

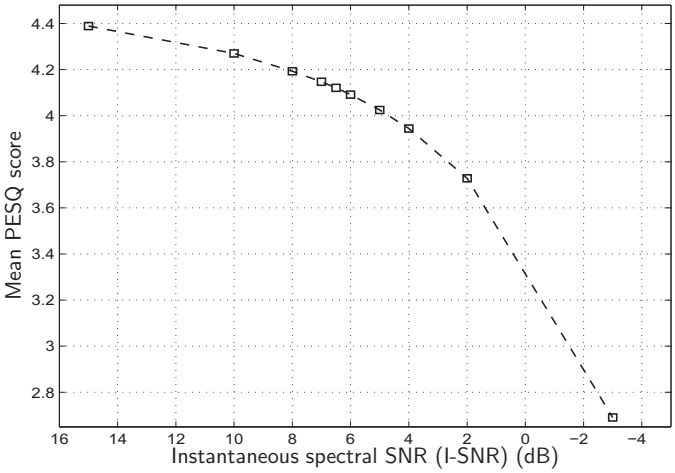

Fig. 6. Objective results in terms of mean PESQ scores for the modified stimuli described in Section 6.1.

The subjective evaluation of the quality of generated stimuli was in the form of a human listening test. The test was conducted in a single session under the same conditions as experiment 1, described in Section 5.2.2. Again, two sentences (one from a male and one from a female speaker), processed using the method described above for each I-SNR under investigation, were used for this experiment. Ten English speaking listeners with normal hearing participated in the test. The results of the subjective experiment are displayed with standard error bars in Fig. 7.

The subjective scores also illustrate a decrease in perceived quality as the I-SNR decreases. A distinct decrease in subjective scores can be seen between an I-SNR of 8 dB and 7 dB. This observation was supported by subjecting the listening test results to statistical analysis. There was a significant effect observed for changing I-SNR ($F(9, 90) = 55.79$, $p < 0.0005$). Tukey's significant difference criterion (at 0.05 level of significance) was again used for the *posthoc* multiple comparison test. It was found that I-SNRs of 15, 10 and 8 dB had a significantly higher score than the remaining seven SNR values (7 to -3 dB). The transitional value of I-SNR producing a

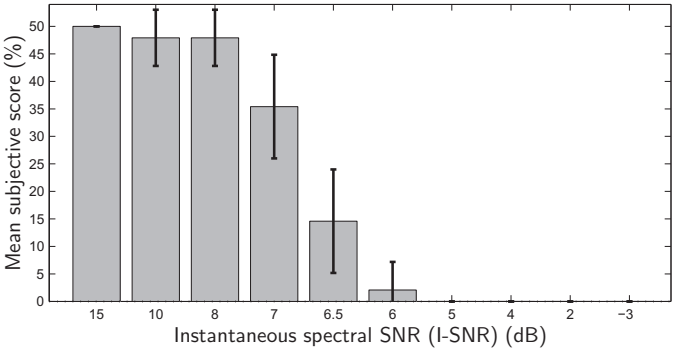

Fig. 7. Mean preference scores (with standard error bars) for modified stimuli described in Section 6.1.

significant change in subjective score occurred at I-SNR = 7 dB, indicating that 7 dB is the I-SNR value where a JND in quality was perceived.

6.3. Discussion

From the results of the objective and subjective experiments, we can see that as the I-SNR decreases the quality also decreases. In the subjective experiment, we showed that the JND for I-SNR is 7 dB. This means that an I-SNR of 7 dB is the highest value for which a change in quality is perceivable, which suggests that the noisy phase spectrum can be used as an approximation of the clean phase spectrum only if the I-SNR is >7 dB. This result is consistent with findings from experiment 1 (shown by Fig. 3).

It must be noted that we have used the clean magnitude spectrum in this experiment for constructing the stimuli. However, in real life, where a speech enhancement algorithm is deployed, we estimate the clean magnitude spectrum from the noisy spectrum. Thus, the estimated spectrum is different from the clean magnitude spectrum. That is, the estimated spectrum is a distorted version of the clean magnitude spectrum. Therefore, we have investigated the effect of phase distortion on speech quality as a function of I-SNR using the estimated magnitude spectrum. This is reported in Section 8 where we show that the JND in terms of I-SNR reduces from 7 dB for clean magnitude spectra, to 3.5 dB using an estimate of the clean magnitude spectra.

7. Experiment 3: JND in terms of segmental SNR

In the previous experiments, we have measured the JND in terms of I-SNR. However, since I-SNR is a measure of localised SNR, and for practical purposes, SNR will not be fixed across all the frames and frequency bins, in this section we evaluate the JND in signal quality due to degradation of the phase spectrum in terms of segmental SNR, which can be calculated across an entire utterance.

7.1. Stimuli generation

Similar to previous experiments, the clean stimuli from the Noizeus corpus (see Section 4), were processed using the short-time AMS framework (outlined in Section 2). In the modification stage, the clean magnitude spectrum was preserved, while the clean phase spectrum was replaced with a degraded phase spectrum from a noisy signal.

In order to calculate the SegSNR of a speech utterance we first calculate the SNR of each frame as

$$SNR_{frame}(m) = 10\log_{10}\frac{\sum_{n=0}^{N-1} x_m^2(n)}{\sum_{n=0}^{N-1}(x_m(n)-y_m(n))^2}, \tag{16}$$

where N is the frame length, n is the discrete-time index, m is the frame index, $x_m(n)$ is the n^{th} sample of the m^{th} frame of the clean signal, and $y_m(n)$ is the n^{th} sample of the m^{th} frame of the degraded signal. The segmental SNR (SegSNR) of a given utterance can then be found by averaging $SNR_{frame}(m)$ over all frames. However, using this approach, the segmental SNR of an utterance can be significantly affected by silent regions present in the speech utterance. During non-speech frames, $SNR_{frame}(m)$ values can be large and negative, which adversely affects the resulting SegSNR. To address this problem, we used a thresholded version of the segmental SNR measure. Here, non-speech frames were identified as those with $SNR_{frame}(m)$ below a given threshold $\gamma = (maxSegSNR - 45)$ dB [4], where $maxSegSNR = \max_{m=0}^{L-1} SNR_{frame}(m)$, and L is the number of frames in the utterance. Non-speech frames were excluded from the SegSNR calculation, and the segmental SNR of a speech utterance was then calculated from the remaining frames as

$$SegSNR = \frac{1}{M}\sum_{m=0}^{M-1} SNR_{frame}(m), \tag{17}$$

where M is the number of speech-present frames in the utterance.

To generate the noisy utterance (used to construct modified stimuli), the level of additive white Gaussian noise (AWGN) required to achieve the desired SegSNR was first calculated as

$$\varphi = \sqrt{\frac{10^{avClnPow/10}}{10^{SegSNR/10}}}, \tag{18}$$

and $avClnPow$ is the average power of non-silence regions of the utterance. The noise signal $d(n)$ was found as the product of φ and a randomly generated vector of numbers with Gaussian distribution from $[-1, 1]$ This noise was then added to the entire clean signal $x(n)$ (including silence regions) in the time domain to produce the noisy signal $y(n)$.

The noisy frames were then processed by taking the STFT of $y(n)$, to give the noisy magnitude spectrum $|Y(n,\ k)|$ and the noisy phase spectrum $\angle Y(n,\ k)$. The noisy phase spectrum $\angle Y(n,\ k)$ is then combined with the clean magnitude spectrum $|X(n,\ k)|$ to give the modified spectrum $\widehat{Y}(n,k)$ given by Eq. (3). Taking the inverse STFT of $\widehat{Y}(n,k)$, followed by overlap-add synthesis, yielded the modified stimuli $\widehat{y}(n)$.

This method was applied to the sentences of the Noizeus corpus, to generate stimuli with SegSNR values of -3, 2, 4, 5, 6.5, 7, 8, 10, 15 dB.

7.2. Results of experiments

Each of the stimuli of the Noizeus corpus, processed using the above method and SegSNRs of -3, 2, 4, 5, 6.5, 7, 8, 10, 15 dB, were used in the objective experiment. Mean PESQ scores were calculated for each SegSNR, and the results are shown in Fig. 8. We observe that the mean PESQ score

[4] With a $\gamma = (maxSegSNR - 45)$ dB threshold, about 5% of frames were identified as non-speech frames and not included in calculation of the SegSNR. By decreasing the threshold, the number of non-speech frames excluded will decrease, and vice versa.

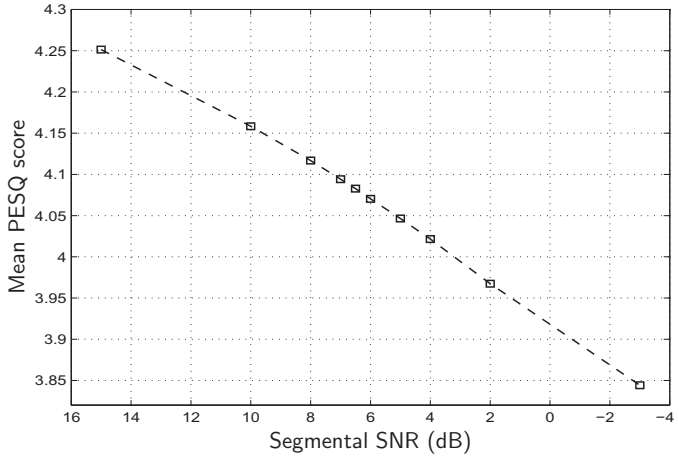

Fig. 8. Objective results in terms of mean PESQ score for the stimuli described in Section. 7.1.

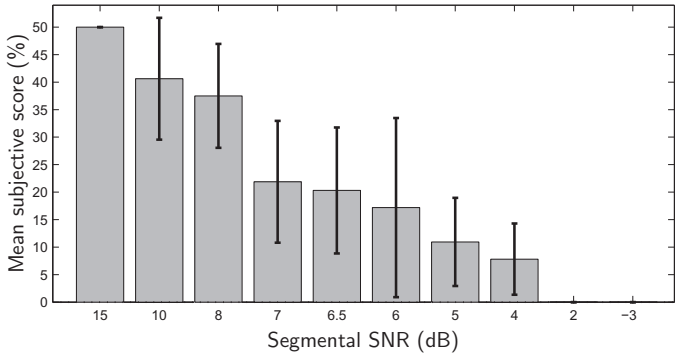

Fig. 9. Subjective quality results in terms of mean preference scores for stimuli described in Section 7.1.

decreases as the SegSNR decreases, suggesting that the quality of utterances decreases for decreasing SegSNR.

The subjective evaluation of the quality of the stimuli created in Section 7.1 was carried out again in the form of a human listening test, measuring subjective scores. For this purpose, two stimuli from the Noizeus corpus (one from a male and one from a female speaker), were utilised. This resulted in 20 stimuli (40 stimulus-pairs) being used in each subjective test. The test was conducted in a single session under the same conditions as described in Section 5.2.2. Ten English speaking listeners with normal hearing participated in the experiment.

The results of the subjective experiment are displayed with standard error bars in Fig. 9. These results illustrate a decrease in perceived quality as the SegSNR decreases. Relating the subjective scores to mean PESQ scores, there was found to be a high correlation between results ($r = 0.9354$). By observing the mean subjective scores, a pattern similar to that of Experiment 2 is seen, with a distinct transition in mean scores occurring between a SegSNR of 8 dB and 7 dB. One-way analysis of variance (ANOVA) was used to test differences in subjective scores for the ten SegSNR values. A significant effect was observed for SegSNR ($F(9, 90) = 30.61$, $p < 0.0005$). Again, Tukey's significant difference criterion (at 0.05 level of significance) was used for the *posthoc* multiple comparison test. It was found that SegSNRs of 15, 10 and 8 dB had a significantly higher score than the remaining seven values (7 dB to −3 dB). The transitional value of SegSNR, where a significant change in subjective score was indicated, occurred at 7 dB. Thus, 7 dB can be interpreted as the value of SegSNR for which the JND in speech quality is audible.

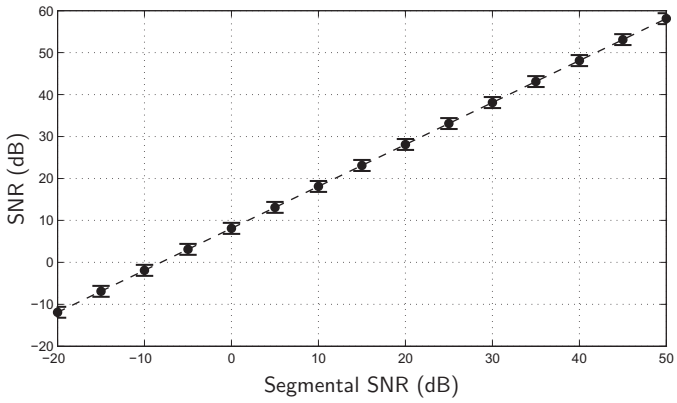

Fig. 10. Mean and standard deviation of SNR as a function of SegSNR.

In this experiment we have measured JND with respect to segmental SNR. However, overall SNR is another measure often used to characterise the noise present in the speech signal. For a given speech sentence (or utterance) it is calculated as

$$SNR = 10\log_{10}\frac{\sum_{\ell=1}^{L} x^2(\ell)}{\sum_{\ell=1}^{L}(x(\ell) - y(\ell))^2}, \tag{19}$$

where ℓ is the sample index, and L is the number of samples in the speech sentence. Unlike the previously described SegSNR, overall SNR is calculated over the whole speech utterance, and includes both silence and speech regions in its calculation. For the 30 utterances of Noizeus database, we have generated their noisy versions with a specified SegSNR. We computed the SNR of each of these noisy sentences using Eq. (19). The mean and the standard deviation of SNRs as a function of SegSNR are shown in Fig. 10. As can be seen, a SegSNR of 7 dB corresponds to an SNR of 14 dB. Thus, the JND in terms of SNR is 14 dB.

7.3. Discussion

In this experiment, we have used SegSNR to quantify the JND. It was found in both the objective and subjective experiments, that as the SegSNR decreased so did the quality of the noisy speech. Subjective listening tests demonstrated that the JND corresponds to a SegSNR of 7 dB. While these listening tests only used two utterances (from one male and one female speaker), which may introduce some bias to results reported, these utterances were randomly selected and findings are believed to reflect outcomes using other utterances.

As such the results reported are sufficient to be indicative of testing on a larger corpus. Thus, from this experiment we can conclude that the noisy phase spectrum can be used as an approximation of the clean phase spectrum only if the SegSNR is greater than 7 dB.

8. Experiment 4: JND for stimuli constructed using an estimate of the clean magnitude spectrum

In Experiment 2 Section 6, we investigated the effect of phase distortion in terms of instantaneous SNR (I-SNR). In those experiments, stimuli were reconstructed from the combination of the clean magnitude spectra and the phase spectra from noisy stimuli degraded with a known level of I-SNR. However when we employ speech enhancement algorithms in practice, we do not have access to the clean magnitude spectrum, and instead it is estimated from the noisy spectrum. This estimate is not ideal, and contains some distortions with respect to the clean magnitude spectrum. Therefore, in this section, we aim to evaluate the effect of reconstructing stimuli using an estimate of the clean magnitude spectrum, on the previously measured JND.

For this purpose, we make use of a popular speech enhancement method – the MMSE log-spectral magnitude estimator (Ephraim and Malah, 1985) (log-MMSE). Stimuli used in this experiment were generated using a procedure similar to the one described in Section 6.1. From Eq. (15), where $Y(n, k)$ is the noisy spectrum (generated to have an I-SNR ξ), log-MMSE was used to estimate the clean magnitude spectrum, denoted $|\hat{X}(n,k)|$. Stimuli pairs were then reconstructed as follows:

(a) The first stimuli of the pair was reconstructed from $Y_{CP}(n,k)$, which is found by combining the estimate of the clean magnitude spectrum with the original clean phase spectrum. That is,

$$Y_{CP}(n,k) = |\hat{X}(n,k)|e^{j\angle X(n,k)}. \quad (20)$$

(b) The second stimuli was reconstructed from $Y_{NP}(n,k)$, which is found by combining the same estimate of the clean magnitude spectrum with the noisy phase spectrum. That is,

$$Y_{NP}(n,k) = |\hat{X}(n,k)|e^{j\angle Y(n,k)}. \quad (21)$$

Comparison of these two treatment types, for a range of I-SNR values, allows us to study the effect of distortion in the phase spectrum (quantified in terms of I-SNR) on speech quality (in terms of the JND).

A subjective test similar to that conducted in Section 6 was carried out. Eleven listeners participated in the subjective test, and thirteen I-SNR values (8, 7, 6.5, 6, 5.5, 5, 4.5, 4, 3.5, 3, 2, 1, 0 dB) were evaluated. The result of the subjective experiment is shown in Fig. 11.

One-way analysis of variance (ANOVA) was used to test differences in subjective scores for varying I-SNR values. This statistical analysis confirmed there was a significant effect for changing I-SNR ($F(12, 130) = 19.35$, $p < 0.0005$). Turkey's significant difference criterion (at 0.05 level of significance) was again used for the *posthoc* multiple comparison test. This indicated a significant change in subjective score at I-SNR around 3.5 dB, indicating that an I-SNR of 3.5 dB corresponds to the JND in speech quality.

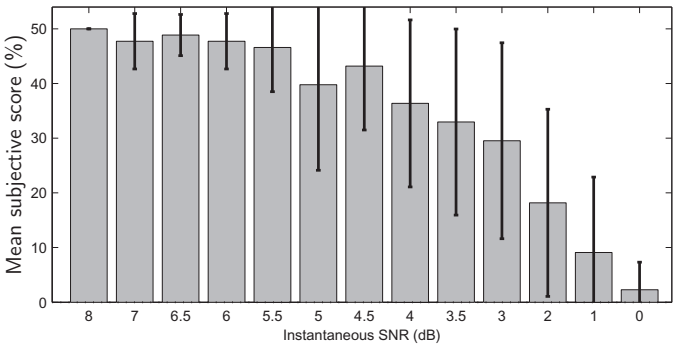

Fig. 11. Subjective quality results in terms of mean preference scores for stimuli described in Section 8.

Relating these findings to those of Experiment 2 (Section 6), we observe that the JND in terms of I-SNR reduces from 7 dB for clean magnitude spectra, to 3.5 dB using an estimate of the clean magnitude spectra. In other words, the presence of distortion in the magnitude spectrum makes listeners tolerate higher amounts of phase distortion.

9. Summary and conclusion

In this paper we have tried to quantify the level of distortion in the phase spectrum which can be tolerated before the quality of speech begins to be effected - that is, the just noticeable difference (JND) resulting from phase distortion. For this we conducted a series of experiments to investigate the effects that distortion in the phase spectrum has on the perceived quality of speech. We have measured the JND in terms of (1) phase difference between noisy phase and clean phase angles, (2) instantaneous spectral SNR (I-SNR), and (3) Segmental SNR (SegSNR). These JNDs are expected to be useful in speech enhancement applications where they can guide us when to use the noisy phase spectrum as an estimate of the clean phase spectrum. In terms of I-SNR and SegSNR, a JND of 7 dB can be recommended as a guideline for speech enhancement algorithms.

References

Alsteris, L., Paliwal, K., 2004. Importance of window shape for phase-only reconstruction of speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 1. Montreal, Quebec, Canada, pp. 573–576.

Alsteris, L., Paliwal, K., 2006. Further intelligibility results from human listening tests using the short-time phase spectrum. Speech Commun. 48 (6), 727–736.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 4. Washington, DC, USA, pp. 208–211.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), vol. 27, pp. 113–120.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 32, pp. 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 33, pp. 443–445.

Ephraim, Y., Van Trees, H., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3, 251–266.

Griffin, D., Lim, J., 1984. Signal estimation from modified short-time Fourier transform. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 32, pp. 236–243.

Hu, Y., Loizou, P.C., 2007. Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. 49 (7–8), 588–601.

Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, Upper Saddle River, New Jersey.

Krawczyk, M., Gerkmann, T., 2014. STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. IEEE Trans. Signal Process. 22 (12), 1931–1940.

Lim, J., Oppenheim, A., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. Taylor and Francis, Boca Raton, FL.

Mowlaee, P., Kulmer, J., 2015. Phase estimation in single channel speech enhancement: Limits-potential. IEEE Trans. Audio Speech Lang. Process. 23 (8), 1283–1294.

Oppenheim, A.V., Lim, J.S., Kopec, G., Pohlig, S.C., 1979. Phase in speech and pictures. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 4. Washington, DC, USA, pp. 632–637.

Paliwal, K., Alsteris, L., 2003. Usefulness of phase spectrum in human speech perception. In: Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). Geneva, Switzerland, pp. 2117–2120.

Paliwal, K., Basu, A., 1987. A speech enhancement method based on Kalman filtering. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 12, pp. 297–300.

Paliwal, K., Wójcicki, K., 2008. Effect of analysis window duration on speech intelligibility. IEEE Signal Process. Lett. 15, 785–788.

Paliwal, K., Wójcicki, K., Shannon, B., 2011. The importance of phase in speech enhancement. Speech Commun. 53, 465–494.

Picone, J., 1993. Signal modeling techniques in speech recognition. Proc. IEEE 81 (9), 1215–1247.

Quatieri, T., 2002. Discrete-Time Speech Signal Processing: Principles and Practice. Prentice Hall, Upper Saddle River, NJ.

Rix, A., Beerends, J., Hollier, M., Hekstra, A., 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P.862.

Shannon, B., Paliwal, K., 2006. Role of phase estimation in speech enhancement. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). Pittsburgh, PA, USA, pp. 1423–1426.

Shi, G., Shanechi, M., Aarabi, P., 2006. On the importance of phase in human speech recognition. IEEE Trans. Audio Speech Lang. Process. 14 (5), 1867–1874.

Stark, A., Wójcicki, K., Lyons, J., Paliwal, K., 2008. Noise Driven Short Time Phase Spectrum Compensation Procedure for Speech Enhancement. In: Proceedings of INTERSPEECH 2008. Brisbane, Australia, pp. 549–552.

Vary, P., 1985. Noise suppression by spectral magnitude estimation – mechanism and theoretical limits. Signal Process. 8 (4), 387–400.

Wang, D., Lim, J., 1982. The unimportance of phase in speech enhancement. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 30 (4), pp. 679–681.

Wiener, N., 1949. The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications. Wiley, New York.