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

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