Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 81 (2016) 104–119

www.elsevier.com/locate/specom

Phase perception of the glottal excitation and its relevance in statistical parametric speech synthesis☆

Tuomo Raitio[a,*], Lauri Juvela[a], Antti Suni[b], Martti Vainio[b], Paavo Alku[a]

[a] Department of Signal Processing and Acoustics, Aalto University, Finland
[b] Institute of Behavioural Sciences, University of Helsinki, Finland

Received 14 June 2015; received in revised form 13 January 2016; accepted 30 January 2016
Available online 15 February 2016

Abstract

While the characteristics of the amplitude spectrum of the voiced excitation have been studied widely both in natural and synthetic speech, the role of the excitation phase has remained less explored. This contradicts findings observed in sound perception studies indicating that humans are not phase deaf. Especially in speech synthesis, phase information is often omitted for simplicity. This study investigates the impact of phase information of the excitation signal of voiced speech and its relevance in statistical parametric speech synthesis. The experiments in the study involve, firstly, converting the pitch-synchronously computed original phase spectra of the excitation waveforms (either glottal flow waveforms or residuals) to either zero phase, cyclostationary random phase, or random phase. Secondly, the quality of synthetic speech in each case is compared in subjective listening tests to the corresponding signal excited with the original, natural phase. Experiments are conducted with natural, vocoded, and synthetic speech using voice material from various speakers with varying speaking styles, such as breathy, normal, and Lombard speech. The results indicate that the phase spectrum of the voiced excitation has a perceptually relevant effect in natural, vocoded, and synthetic speech, and utilizing the phase information in speech synthesis leads to improved speech quality.

Keywords: Statistical parametric speech synthesis; Phase perception; Glottal flow excitation; Vocoding.

1. Introduction

In statistical parametric speech synthesis (SPSS), several vocoding techniques have been used in the past decade (Tokuda et al., 2013; Zen et al., 2009). The conventional vocoding approach employs excitation signals composed of impulses mixed with noise. The spectrum of this kind of simple excitation, both in terms of its amplitude and phase, is greatly different from the spectrum of the real voice source of speech, the glottal flow. While the characteristics of the amplitude spectrum of voice excitation have been studied widely both in natural (Childers and Lee, 1991; Gobl and Ní Chasaide, 1992) and synthetic (Klatt and Klatt, 1990; Raitio et al., 2014c) speech, the role of the excitation phase has remained less explored. This contradicts findings observed in sound perception studies indicating that humans are not phase deaf (Patterson, 1987). In addition, previous studies show that the phase spectrum has a perceptually relevant role especially in speech signals (Pobloth and Kleijn, 1999) and that incorporating phase information is advantageous, for example, in feature extraction of speech recognition (Alsteris and Paliwal, 2004; Paliwal, 2003; Zhu and Paliwal, 2004).

The common tradition of discarding phase information in speech processing stems from two issues. Firstly, the magnitude spectrum is perceptually more relevant than the phase spectrum. Secondly, there are inherent difficulties, such as phase unwrapping (Tribolet, 1977), in processing the phase spectrum. In addition, previous studies indicate that the perception of phase has a complex dependency on the signal's

☆ Audio files and additional figures can be found at http://research.spa.aalto.fi/publications/papers/specom-phase/.

* Corresponding author. Tel.: +1 4088589068.
E-mail addresses: tuomo.j.raitio@gmail.com, traitio@apple.com (T. Raitio), lauri.juvela@aalto.fi (L. Juvela), antti.suni@helsinki.fi (A. Suni), martti.vainio@helsinki.fi (M. Vainio), paavo.alku@aalto.fi (P. Alku).

http://dx.doi.org/10.1016/j.specom.2016.01.007
0167-6393/

fundamental frequency (f_0), intensity, and bandwidth (Laitinen et al., 2013; Patterson, 1987). Regardless of these factors, the present study was designed to investigate the impact of phase information in speech synthesis. Differently from the previous studies that utilize phase information that is extracted from speech pressure signals (e.g. Paliwal and Alsteris, 2005), the current investigation aims to gather new knowledge on the perceptual relevance of phase embedded in *speech excitation* that is used by the vocoder in SPSS. More specifically, this study explores how perception of phase information depends on factors related to speech material, such as gender, speaker, and speaking style. The experiments involve, firstly, converting the pitch-synchronously computed original phase spectra of the excitation waveforms (either glottal flow waveforms or residuals) to either zero phase, cyclostationary random phase, or random phase. Secondly, the quality of synthetic speech in each case is compared in subjective listening tests to the corresponding signal excited with the original, natural phase. Experiments are conducted with natural, vocoded, and synthetic speech using voice material from various speakers with varying speaking styles, such as breathy, normal, and Lombard speech.

The paper is organized as follows. Section 2 briefly presents the properties of a periodic signal and discusses previous studies on phase perception and its mechanisms. In addition, the relation of phase to speech production and voice quality is described, and previous studies on utilizing phase in SPSS are discussed. Section 3 first describes the methodology of phase modification, and then details three separate experiments conducted with natural, vocoded, and synthetic speech and presents the consequent results. Section 4 discusses the implications of the results, and finally Section 5 summarizes the findings and concludes the paper.

2. Background

2.1. Properties of a periodic signal

A steady-state periodic signal $s(t)$ can be represented by

$$s(t) = \sum_{n=1}^{\infty} a_n \sin(2\pi n f_0 t + \varphi_n) \quad (1)$$

where a_n and φ_n are the amplitudes and the phases (in radians) of the nth sinusoidal component, respectively, and f_0 is the fundamental frequency of the signal in Hertz (oscillations or cycles per second). According to Eq. (1), the waveform of the steady state periodic signal depends solely on a_n, which define the peak amplitude of each sinusoidal component, and φ_n which define the instantaneous phase of each sinusoid at $t = 0$.

2.2. Previous studies on phase perception

A common assumption has long been that human hearing is not sensitive to phase due to the early studies on phase perception[1] by Ohm (1843) and von Helmholtz (1859); 1863). More recent studies, however, show that the ear is not phase deaf (Bilsen, 1973; de Boer, 1961; Carlson et al., 1979; Goldstein, 1967; Licklider, 1957; Moore, 2002; Moore and Glasberg, 1989; Patterson, 1987; Plomp and Steeneken, 1969; Schroeder, 1959; Schroeder and Strube, 1986), that is, two harmonic signals with identical magnitude spectra but different phase spectra can be perceptually different from each other. Although the phase information is perceptually not as important as spectral amplitude information, phase information plays a perceptually relevant role in certain important signals, such as speech (Carlson et al., 1979; Laitinen et al., 2013; Paliwal and Alsteris, 2005; Patterson, 1987; Plomp and Steeneken, 1969; Pobloth and Kleijn, 1999; Schroeder and Strube, 1986).

Using a synthetic harmonic signal consisting of 10 harmonics with a slope of −6 dB/oct and f_0 of 146.2 Hz, Plomp and Steeneken (1969) reported that the processing of phase had a maximal perceptual effect[2] that was equal to the effect of changing the spectral slope of the signal by 2 dB/oct, or changing the overall sound pressure level by 2 dB. They also observed that the maximal effect of phase in speech signals was quantitatively comparable to the differences in timbre between the vowels [ø], [e], and [æ]. In Laitinen et al. (2013), 100 Hz flat-spectrum signals with a bandwidth of 24 kHz were used in a similar setting, and the difference between in-phase and random-phase signals was found to exceed the effect of random modulation of the harmonic amplitudes by 4 dB. Based on these findings, it can be argued that the effect of phase spectrum is perceptually less important than that of magnitude spectrum but is still perceptually relevant.[3]

Various studies conclude that human hearing is less sensitive to phase in signals with a high repetition rate (pitch) than in signals with a low repetition rate (Plomp and Steeneken, 1969). For example, Patterson (1987) suggests that humans are phase deaf for signals with a repetition rate above 400 Hz but not below 200 Hz, if the signal consists of 12 harmonics. In Laitinen et al. (2013), original-phase and random-phase periodic signals with repetition rates from 50 Hz to 1600 Hz were used as test stimuli, and the difference due to phase was found to gradually decrease with the repetition rate. Test subjects could not reliably tell the difference between the two signals based on the phase differences when the repetition rate was beyond 800 Hz. In line with these observations, many studies suggest that the perception of phase plays a more

[1] For a review on early studies on phase perception, see for example Plomp and Steeneken (1969), Goldstein (1967), and Patterson (1987).

[2] The maximal effect of phase on timbre was found to occur between a tone consisting of only sine/cosine terms and a tone consisting of alternate sine and cosine terms.

[3] It is generally believed that the magnitude spectrum plays a dominant role for small window durations (20–40 ms) while the phase spectrum is more important for large window durations (>1 s) (Oppenheim and Lim, 1981; Schroeder, 1975). However, Paliwal and Alsteris (2003, 2005) show that phase information can contribute as much to speech intelligibility as magnitude information, if analysis-synthesis parameters are appropriately set. It is, however, not clear if this is applicable in SPSS.