Phase perception of the glottal excitation of vocoded speech

Tuomo Raitio1, Lauri Juvela1, Antti Suni2, Martti Vainio2, Paavo Alku1

1Department of Signal Processing and Acoustics, Aalto University, Finland
2Institute of Behavioural Sciences, University of Helsinki, Finland
firstname.lastname@aalto.fi, firstname.lastname@helsinki.fi

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. Especially in speech synthesis, the phase information is often omitted for simplicity. This study investigates the impact of phase information of the excitation signal of voiced speech. The experiments in the study involve analysis-synthesis of speech using a vocoder that utilizes natural glottal flow pulses for reconstructing the voiced excitation. Firstly, the phase spectra of the glottal flow waveforms are converted to either zero-phase or random-phase. Secondly, the quality of vocoded speech using the two phase-modified pulses is compared in subjective listening tests to the corresponding signal excited with the natural-phase pulse. The results indicate that phase has a perceptually relevant effect in vocoded speech and the use of natural phase improves the synthesis quality.

Index Terms: statistical parametric speech synthesis, vocoding, phase perception, glottal flow excitation, voice quality

1. Introduction

In statistical parametric speech synthesis (SPSS), several vocoding techniques have been used in the past decade [1, 2]. 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 source of speech production, the glottal flow. While the characteristics of the amplitude spectrum of the voiced excitation have been studied widely both in natural [3, 4] and synthetic [5, 6] 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 [7]. In addition, previous studies show that the phase spectrum has a perceptually relevant role especially in speech signals [8] and that incorporating phase information is advantageous, for example, in feature extraction of speech recognition [9].

The common tradition of discarding the 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 the phase unwrapping, in processing the phase spectrum. In addition, previous studies indicate that the perception of phase has a complex dependency on the signal’s fundamental frequency (f0), intensity, and bandwidth [7, 10]. Despite 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., [11], the current investigation aims to gather new knowledge on the perceptual relevance of phase that is embedded in speech excitation that is used by the vocoder in SPSS. More specifically, this study explores how the perception of phase 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 to either zero-phase or random-phase. Secondly, the quality of vocoded speech in each case is compared in subjective listening tests to the corresponding signal excited with the original, natural phase. Experiments are conducted using material from various speakers with varying speaking styles: breathy, normal, and Lombard speech.

The paper is organized as follows. First, Section 2 shortly presents the properties of a periodic signal, discusses previous studies on phase perception and utilizing phase information in SPSS. Section 3 describes the general methodology of phase modification, and Section 4 details the experiments conducted and presents the consequent results. Section 5 finally 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

\[ s(t) = \sum_{n=1}^{\infty} a_n \sin(2\pi nf_0 t + \varphi_n) \]  

where \( a_n \) and \( \varphi_n \) are the amplitudes and the phases (in radians) of the \( n \)th 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 \( \varphi_n \), 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 the human hearing is not sensitive to phase due to the early studies on phase perception [12, 13]. More recent studies, however, show that the ear is not phase deaf [7, 14–18], 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, the phase information plays a perceptually relevant role in certain important signals, such as in speech [7, 8, 10, 11, 15, 17, 18].

The human hearing is less sensitive to phase in signals with high repetition rate (pitch) than in signals with low repetition rate [15]. Depending on the signal, the phase differences become inaudible at repetition rates from 400 Hz to 800 Hz [7, 10].

Copyright © 2015 ISCA
Thus, the perception of phase plays a more significant role in male speech than in female speech [19]. It has been shown that the phase changes are perceivable when the frequency separation between the spectral components is not too large [14]. This can be explained by the limited frequency resolution of the human hearing and the processing of sounds within critical bands [20]. The differences in phase spectrum can be classified into local and global phase changes [7] or to within-channel and between-channel phase changes [21]. Only the relative phase changes within the channels are considered perceptually important [7, 21]. Therefore, the higher the $f_0$, the less the number of harmonics in the same critical band, and thus the phase relationship becomes perceptually less important as the proportion of perceptually irrelevant (between-channel) phase components to the total number of harmonics becomes smaller.

In summary, previous studies indicate that there is a rather good consensus on the fundamentals of phase perception. It is, though, worth noting that these investigations have mainly used stationary synthetic signals, constructed from individual sinusoids, with simple magnitude spectra. There are, however, less studies on phase perception using complicated, real world signals in modern applications such as SPSS. As stated before, phase perception has a complex dependency on the properties of the signal [7, 10]. Thus it is not clear what the role of phase is in SPSS, where the excitation consists of the quasiperiodic glottal excitation mixed with (modulated) noise, whose magnitude is finally modified by the vocal tract resonances.

### 2.3. Previous studies on utilizing phase in speech synthesis

Already Rosenberg et al. [22] showed in their studies with formant synthesis that the glottal pulse shape has an effect on the perceived quality of synthetic vowels. Recently, SPSS [1, 2] has taken its place as one of the most important speech synthesis technologies along with unit selection synthesis. However, the conventional SPSS techniques employ an excitation signal that consists of a train of zero-phase impulses mixed with noise, which is different from the natural glottal flow excitation that has specific non-zero phase characteristics. Due to the inherent difficulties in modeling the phase spectrum, many studies have utilized parametric models of the glottal flow [23–26]. Other studies have utilized natural glottal flow or residual waveforms to create the excitation of synthetic speech to account for the natural phase characteristics. For example, individual natural glottal flow pulses [27–29] or pulse libraries [30–32], residual codebooks [33], and principal component analysis (PCA) based decomposition of residual/glottal flow signals [6, 34–36] have been used for the excitation modeling in SPSS. Also a deep neural network (DNN) based approach for modeling the glottal flow waveform in SPSS was proposed in [37–39]. In [40], it was shown that a natural residual is beneficial in a male voice but not with a female voice, being consistent with the earlier studies in which the phase information is observed to be more important with male than female voices. The study also concludes that a noise model is essential in improving the synthesis quality, but the perceptual impact of the noise time-envelope, i.e., the distribution of noise energy in time per pitch period, seems to be negligible.

It is noteworthy that the phase characteristics of the speech residual have been also modeled explicitly in [41] by using the complex cepstrum. Their studies show that the modeling of the phase characteristics results in synthesized waveforms that are closer to natural ones and they are also perceived as more natural. Also STRAIGHT [42, 43], the most widely used vocoder in SPSS today, uses a phase manipulation of the zero-phase impulse excitation in order mimic the temporal fine-structure of natural excitation [44]. However, there is no information on the perceptual effect of this processing. Despite the aforementioned studies and findings, it is not yet well known what is the relevance of the phase information in advanced voice source modeling methods in SPSS.

### 3. Phase modification

In order to investigate the effect of the excitation phase characteristics, a phase manipulation scheme was developed in which the phase of a glottal flow waveform is altered while keeping the magnitude spectrum unchanged.

Let $x(n)$ be a glottal flow derivative signal consisting of two pitch periods having altogether $2N_T$ samples, where $N_T$ is the number of samples per fundamental period. The signal is delimited by two glottal closure instants (GCIs) located at $n = 0$ and $n = 2N_T - 1$. A third GCI is located approximately in the center of the frame. The signal is windowed using the Hann window of length $2N_T$. An example of such a signal and its phase spectrum is shown in Fig. 1a. The phase manipulation process begins by computing the discrete Fourier transform (DFT) of the signal by

$$X(k) = \mathcal{F}\{x(n)\},$$

and then evaluating the phase spectrum by

$$\varphi(k) = \angle X(k) = \arctan\left(\frac{\mathcal{I}\{X(k)\}}{\mathcal{R}\{X(k)\}}\right).$$

The phase spectrum can be then modified for the purpose of the
study. A zero-phase signal can be easily created by setting
\[ \varphi_{\text{zero}}(k) = 0 \quad \forall k. \]  
(4)
This implies that all the cosine terms have the same instantaneous phase (of zero rad) at \( n = 0 \). The signal is symmetric along its center point, but the energy of the signal is concentrated at the beginning and at the end of the signal (see Fig. 1b).

In order to enable pitch-synchronous overlap-add (PSOLA) synthesis, a circular time shift is performed (see Fig. 1c), which equals to adding a linear component to the phase spectrum.

A random-phase signal can be generated by sampling the phase from the uniform distribution
\[ \hat{x}(n) = \mathbb{R}\{ \mathcal{F}^{-1}\{ r(k)e^{i\hat{\varphi}(k)} \} \}, \]  
(6)
where \( \hat{\varphi}(k) \) is the modified phase and \( r(k) \) is the radius
\[ r(k) = |X(k)| = \left( \Re\{X(k)\}^2 + \Im\{X(k)\}^2 \right)^{\frac{1}{2}}. \]  
(7)
In the case of random-phase modification, the energy of the resulting signal must be normalized to correspond to the energy of the original signal. It is important to note that none of the aforementioned phase modifications change the magnitude spectrum of the signal. The aforementioned phase manipulation techniques form the basis for the following experiment to investigate the perceptual effect of phase in vocoded speech.

4. Experiments

The aim of the present study was to investigate the perceptual effect of phase in vocoded speech. A vocoder that utilizes natural glottal flow pulses in reconstructing the voiced excitation was used. Natural and phase-modified glottal flow pulses were used in the analysis-synthesis of sentences from 3 male and 2 female speakers with different speaking styles, and the quality of the samples were evaluated in subjective listening tests.

4.1. Methodology

The GlottHMM vocoder \([28, 30]\) was used in the experiments for analysis-synthesis of speech. GlottHMM was chosen since it utilizes natural glottal flow pulses for reconstructing the voiced excitation. Thus it is possible to apply phase-modifications to the glottal flow pulses used in synthesis and thereby evaluate the effect of phase in vocoded speech.

The GlottHMM vocoder utilizes glottal inverse filtering (GIF) for decomposing voiced speech into the vocal tract filter contribution and the voice source signal. The iterative adaptive inverse filtering (IAIF) \([45]\) is used for GIF, inside which linear prediction (LP) is used for spectrum estimation. GlottHMM performs a frame-wise analysis of speech and parameterizes it into 3 types of speech features: 1) LP spectrum (order 20–30) converted to LSF, 2) LP voice source spectrum (order 6–10) converted to LSF, 3) \( f_0 \), 4) frame energy, and 5) harmonic-to-noise ratio (HNR) of 5 bands. GlottHMM also detects the GCIIs from the voice source signal and extracts individual two-pitch-period glottal flow derivative pulses that can be used in synthesis for reconstructing the voiced excitation. GlottHMM provides several synthesis methods for the voiced excitation. In this work, a single two-pitch-period glottal flow derivative waveform is used for synthesizing an utterance. In synthesis, the waveform is first scaled to have appropriate length and energy, after which noise is added according to the HNR in order to obtain a desired degree of voicing. The waveforms are then overlap-added to create the voiced excitation, which is then modified using a spectral matching filter to obtain an appropriate excitation spectrum. Finally, the excitation signal is filtered using the vocal tract filter to obtain speech.

For obtaining glottal flow pulses for synthesis, the following methodology was used. For each test utterance, all glottal flow derivative waveforms were extracted using GlottHMM. The pulses were interpolated to constant length and a mean waveform was computed. From the extracted glottal flow pulses, the closest one to the mean in terms of mean squared error was selected for synthesizing the utterance, thus preserving as much of the speaker and voice quality characteristics as possible using a single pulse per utterance.

Two types of phase manipulations were used for the glottal flow derivative pulses, thus resulting in three systems: 1) natural-phase pulse\(^2\), 2) zero-phase pulse, and 3) random-phase pulse. The zero-phase and random-phase pulses were achieved using the methodology described in Sec. 3. Since the random-phase pulse is constant throughout an utterance, a cyclostationary-random-phase excitation is created.

4.2. Speech material

Finnish speech corpora of 3 males and 2 females was utilized with 3 different speaking styles: breathy, normal, and Lombard speech. The Lombard speech was elicited by playing babble noise from the NOISEX-92 database with 80 dB SPL to the speaker’s ears through Sennheiser HD 250 linear II headphones. The speaker’s own feedback of speech was set to correspond to a situation of hearing the speaker’s own voice in a quiet room without headphones. The breathy speaking style was elicited by increasing the level of the speaker’s own feedback through headphones as well as instructing the subjects to keep their speech voiced. Speech was recorded in a soundproof studio, standing 10 cm away from an AKG 4000B large diaphragm condenser microphone with a sampling rate of 48 kHz with 24-bit resolution, and converted to 16 kHz for the experiments. More details on the recordings can be found in \([6]\).

Five sentences spoken by the 5 speakers with the 3 speaking styles were selected for the experiment, resulting in a total of 75 sentences. Fig. 2 shows the speech parameter statistics of the data extracted using the GlottHMM vocoder.

4.3. Subjective listening tests

All the 75 sentences were vocoded using GlottHMM with the three types of pulses described in the previous section: natural-

\(^2\) Although the glottal flow pulses used for synthesis are interpolated according to \( f_0 \), which slightly shifts the frequencies of the pulse and thus, e.g., the glottal formant position and bandwidth, the changes are considered only minor, and the overall phase structure can be assumed the remain approximately the same.
phase, zero-phase, and random-phase. In order to evaluate the perceptual differences between the vocoded utterances 3, a test methodology similar to the MUSHRA test [46] was used. Since MUSHRA is intended for the comparison of high quality reference sounds with several lower quality test sounds, this test type was deemed to be best suited for the task. In a MUSHRA test, the test subject listens to a number of test samples and rates the quality of them in comparison to a known reference sample on a mean opinion score (MOS) scale ranging from bad (0) to excellent (100). Included in the test samples, there is a hidden reference. Thus, the compared methods were the following:

1. Natural speech (reference and hidden reference)
2. Vocoded with natural-phase pulse
3. Vocoded with zero-phase pulse
4. Vocoded with random-phase pulse (cyclostationary-random-phase excitation)

The tests were performed in quiet listening booths using Sennheiser HD650 headphones. The test interface was based on the MUSHRAM software [47] and run in Matlab. The test took approximately one hour per listener. A total of 15 Finnish listeners with no reported hearing problems between the age of 23 and 37 participated in the test.

4.4. Results

The ratings for each method averaged across each speaker, gender, speaking style, and all speakers are shown in Fig. 3. The average score for the hidden reference is close to 100, confirming the validity of the test and the listeners (not shown in the figure for clarity). The scores for the natural-phase, zero-phase, and random-phase vocoded speech vary across speakers. For males on average the natural-phase signal is rated higher than the zero-phase signal, which is further ranked higher than the random-phase signal. For female speech, the natural-phase signal is rated higher than the zero-phase signal, but random-phase signal shows no statistically significant difference in comparison to the two other systems. For every speaking style, the order obeys the same order except that the random-phase signal is not significantly worse than the two others with Lombard speech. Overall, the natural-phase signals are rated best among vocoded speech. The zero-phase and random-phase signals are rated clearly worse, but the zero-phase signals are rated better than the random-phase signals. The natural-phase signals are rated higher with males than females, and higher for low-pitch breathy and normal than high-pitch Lombard speech, confirming the previous studies indicating that the phase characteristics are perceptually more relevant with low-pitched speech.

5. Conclusions

The results unambiguously show that using the natural phase information of the glottal excitation in vocoding improves the speech quality for all speaking styles. The benefit is especially notable with low-f0 male speech. However, even with high-pitched female speech the natural phase information seems to give benefit, although these observations are not statistically significant.

The natural phase spectrum of the excitation signal, in addition to correct amplitude spectrum, is probably one of the reasons for the recent success of glottal-flow-excited vocoders (e.g., [28, 30, 35]). Future work will include evaluations also with synthetic speech generated with SPSS.

6. Acknowledgements

This work has been supported by the Academy of Finland (256961, 284671).

7. References
