An Algorithm for Phase Manipulation in a Speech Signal

Darko Pekar[1], Siniša Suzić[2], Robert Mak[2], Meir Friedlander[3], and Milan Sečujski[2](✉)

[1] AlfaNum – Speech Technologies, Novi Sad, Serbia
darko.pekar@alfanum.co.rs
[2] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
{sinisa.suzic, robert.mak, secujski}@uns.ac.rs
[3] Speech Morphing, Inc., Campbell, CA, USA
meir@speechmorphing.com

Abstract. While human auditory system is predominantly sensitive to the amplitude spectrum of an incoming sound, a number of sound perception studies have shown that the phase spectrum is also perceptually relevant. In case of speech, its relevance can be established through experiments with speech vocoding or parametric speech synthesis, where particular ways of manipulating the phase of voiced excitation (i.e. setting it to zero or random values) can be shown to affect voice quality. In such experiments the phase should be manipulated with as little distortion of the amplitude spectrum as possible, lest the degradation in voice quality perceived through listening tests, caused by the distortion of amplitude spectrum, be incorrectly attributed to the influence of phase. The paper presents an algorithm for phase manipulation of a speech signal, based on inverse filtering, which introduces negligible distortion into the amplitude spectrum, and demonstrates its accuracy on a number of examples.

Keywords: Phase perception · Parametric speech synthesis · Zero phase · Random phase · Inverse filtering

1 Introduction

Due to the early studies of sound perception, it has been assumed for a long time that human auditory system recovers all information from the amplitude spectrum of the incoming signal and that it does not rely on phase spectrum at all [1, 2][1]. However, more recent studies have shown that, on the contrary, our sound perception is sensitive to phase information to a certain degree [4–6]. In practice this dependence is still often ignored. For example, most contemporary automatic speech recognition systems still rely on features extracted from amplitude spectra only [7], and in speech enhancement it is common practice to modify the magnitude spectrum and keep the corrupt phase spectrum [8, 9]. Nevertheless, a number of recent studies have shown that pitch information has particular relevance in case of speech signals [10], and that, e.g. the

[1] A review of most important early studies in phase perception can be found e.g. in [3].

A. Ronzhin et al. (Eds.): SPECOM 2016, LNAI 9811, pp. 92–99, 2016.
DOI: 10.1007/978-3-319-43958-7_10

accuracy of both human and automatic speech recognition can be improved if phase information is taken into account in some way [7, 11, 12].

Most practical studies related to the influence of phase to speech perception are based on listening tests in which listeners are presented with speech samples containing a number of different versions of the same utterance, which are assumed to have identical amplitude spectra and different phase spectra. A direct consequence of this approach is that any difference in quality that is perceived between sound samples will inevitably be attributed to the influence of the phase information. However, in reality it is not possible to modify the phase spectrum of an utterance in an arbitrary way without affecting the amplitude spectrum as well, which implies that the interpretation of the results of such studies has to take into account the specific algorithm used to modify the phase spectrum. For instance, some representative phase perception studies, including most notably [13], are based on the manipulation of either excitation signal or speech itself by superimposing overlapping frames previously shifted in time, which is known to affect the amplitude spectrum as well. This paper presents an alternative approach, based on the decomposition of the speech signal according to the source-filter model and phase modification by using an all-pass filter whose phase response varies from sample to sample of the input signal.

The remainder of the paper is organized as follows. Section 2 presents the source-filter separation algorithm which enables subsequent manipulation of the phase spectrum. Section 3 presents the proposed algorithm for phase manipulation in detail, Sect. 4 presents the results of a simple experiment carried out on recordings of four speakers of Serbian and English, and Sect. 4 concludes the paper, giving an outline of the future work as well.

2 Source-Filter Separation Algorithm

The source-filter separation algorithm used in this research is based on inverse filtering with filters obtained through the estimation of MFCC coefficients of the spectrum, as outlined in Fig. 1. Since this study is concerned with phase information, the discussion will be restricted to the case of voiced speech segments, with well-defined fundamental frequency (f_0). The speech signal under analysis is firstly separated into individual frames using overlapping Hamming windows, positioned pitch-synchronously, which requires the knowledge of f_0 as well as glottal closure instants. A slightly modified

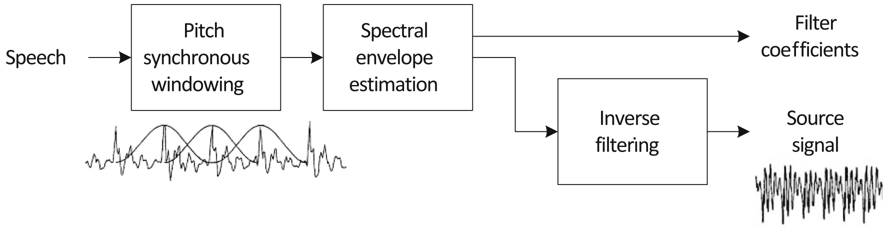

Fig. 1. Outline of the source-filter separation algorithm for voiced speech.

approach can be used even if these are not available, but with inferior results. The spectral envelope for each frame is estimated in the following steps:

(1) FFT of size N is calculated for each frame in a sufficient number of points.
(2) Obtained values are used to calculate the estimated value of the spectrum at integer multiples of f_0, using linear interpolation.
(3) Based on the values of the spectrum at integer multiples of f_0, FFT coefficients are re-interpolated.
(4) Thus obtained FFT coefficients are passed through a filter bank of K overlapping mel-spaced triangular filters, where K is one of the analysis parameters.

The estimates of filter bank outputs at equidistant reference time instants (the distance d between them being another analysis parameter) are obtained by linear interpolation between filter bank outputs previously obtained at glottal closure instants (i.e. pitch-synchronously). Thus obtained values are used to interpolate the full spectrum envelope in L points (where L is also specified as input parameter) for each reference time instant. Linear interpolation is used, since some other techniques, which were also investigated, did not lead to an improvement. In the i-th reference time instant the spectrum envelope is normalized by:

$$norm_i = \sqrt{\frac{1}{WL}\sum\nolimits_{k=0}^{L-1} X^2(k)}, \tag{1}$$

where $X(k)$ is the interpolated spectrum and W is the length of the analysis window.

Thus obtained spectrum envelope can be converted into the impulse response of the corresponding filter using IFFT. The actual frequency response of the inverse filter at each sample between reference time instants $n - 1$ and n is given by:

$$I(f) = \frac{1}{H(f)} = \frac{1}{c_1 H_{n-1}(f) + c_2 H_n(f)}, \tag{2}$$

where $H_{n-1}(f)$ and $H_n(f)$ are the frequency responses at reference time instants $n - 1$ and n, and c_1 and c_2 represent the relative distance of the sample under consideration from time instants $n - 1$ and n respectively. It is important to note that the inverse filter obtained in this way varies with each sample of the speech signal. The excitation signal $e(n)$ can now be easily obtained by inverse filtering.

It should be noted that the described procedure is computationally too demanding to be performed at real time due to the necessity of calculating the inverse filter coefficients for each sample. However, having in mind its purpose, this is only a minor disadvantage, particularly having in mind that the resynthesis process does not suffer from the same drawback. Namely, for resynthesis purposes it is possible to keep only the filters related to reference time instants and to perform linear interpolation in time domain.

3 Phase Manipulation

In this approach phase manipulation is performed on the excitation signal, which is processed within symmetrical rectangular windows of length equal to the current T_0. The windows are centered around glottal closure instants (GCIs), which are also used for calculating the current T_0 and as reference points during the analysis. Assuming that the aim is to obtain zero phase spectrum, the phase spectrum for each voiced frame is manipulated in the following steps:

(1) Firstly, the T_0 of each fundamental period is calculated by subtracting two adjacent GCIs, and this value is assigned to the centre of that period (i.e. it will be assumed that at that point T_0 has this value). Then, each sample is assigned its current T_0, calculated by interpolating the values of T_0 of the two adjacent fundamental periods between whose centres the sample in question is located.

(2) A DFT of size T_0 (in samples) is calculated for each sample, using a symmetrical rectangular window, without zero padding. The obtained phase spectrum is unwrapped in order to avoid phase discontinuities, using an appropriate function from *Matlab*.

(3) For each sample, the corresponding phase spectrum is calculated from the previously obtained unwrapped phase spectrum by adding the term $2k\pi\delta$, where k is the index of a particular spectral component and δ is the parameter that provides the relative distance of the relevant sample of the signal from the previous GCI ($0 \leq \delta < 1$). In this way a phase spectrum that varies with each sample of the excitation signal is obtained.

(4) For each sample it is now possible to obtain an all-pass filter with a phase spectrum exactly the inverse of the phase spectrum obtained in the previous step. By filtering the excitation signal $e(n)$ using the time-variant all-pass filter, a modified excitation signal $e'(n)$ with zero-phase spectrum is obtained. It is important to note that all-pass filtering is carried out on a sample-by-sample basis, i.e. although the all-pass filter obtained for sample i of the input signal operates on a number of samples in the neighbourhood of i, it produces only one sample of the output signal. For each new sample of the input signal, a new filter has to be obtained.

The algorithm is summarized by the pseudocode given in Fig. 2.

It should be noted that it would be possible to shape the phase spectrum in any other way by applying an obvious modification to the step 3. However, to obtain a random-phase excitation signal it is not necessary to perform a sample-by-sample modification proposed above, i.e. it is sufficient to manipulate the excitation signal all at once, by using a time-invariant all-pass filter. Figure 3 illustrates a segment of the excitation signal for the vowel /a/ before and after phase manipulation. The modified speech segment belongs to the corpus of studio recordings of a female speaker of Serbian (referred to as D in the following section), used for expressive speech synthesis [14]. It should be noted that the excitation signal modified in this way has the most of its energy concentrated around GCIs, which makes it more robust to windowing. This, in turn, is beneficial for a number of speech processing techniques that include overlap-add modifications of excitation signals. Figure 4 illustrates the difference between the original phase spectrum, the phase spectrum set to zero using the proposed algorithm, and the phase spectrum of the generic HTS (Hidden Markov model based TTS) excitation [15], shown here as a reference.

```
algorithm phase_mod is
  input: Excitation signal $e$
  output: Modified excitation signal $e'$
  for each sample $i$ in $e$ do
    if voiced($i$) then
      find 3 closest GCIs to $i$ and calculate $T_{01}$ and $T_{02}$
      let $d_1$ and $d_2$ be distances to centres of periods $T_{01}$ and $T_{02}$
      $T_0 \leftarrow (d_2 * T_{01} + d_1 * T_{02})/(d_1 + d_2)$
      let $f$ be the array of samples from $i - T_0/2$ to $i + T_0/2$
      $F \leftarrow$ DFT($f$)
      phase($F$) $\leftarrow$ unwrap(phase($F$))
      let $i_1$ and $i_2$ be positions of previous and next GCI in $e$
      $delta \leftarrow (i - i_1)/(i_2 - i_1)$
      for each spectral sample $F(k)$ in $F$ do
         phase($F(k)$) $\leftarrow$ phase($F(k)$) + 2 * $k$ * $delta$ * pi
      let $AP$ be the all-pass filter with phase -phase($F(k)$)
      $e'(i) \leftarrow AP(e(i))$
    else
      $e'(i) \leftarrow e(i)$
    end if
  return $e'$
```

Fig. 2. Pseudocode of the phase modification algorithm.

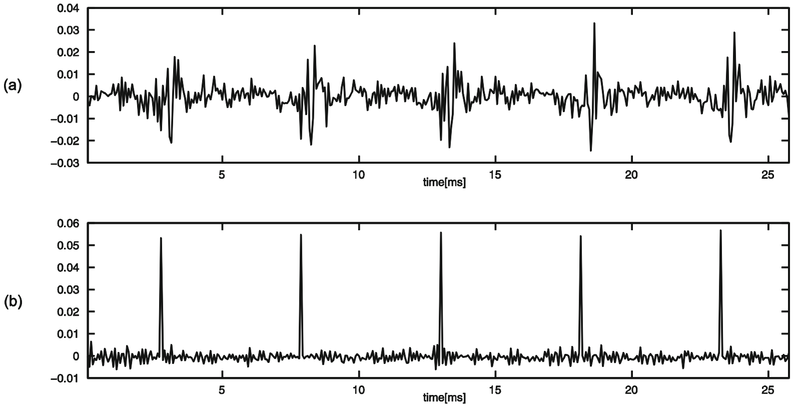

Fig. 3. A segment of voiced speech excitation (vowel /a/) (a) with original phase distribution (b) with phase spectrum set to zero using the proposed method.

4 Experiment Results

Preliminary experiments aimed at the verification of the proposed methods for source-filter separation and phase manipulation were carried out on speech recordings of four speakers – one female and one male speaker of Serbian, as well as one female and one male speaker of English. Five recordings from each speaker were analyzed, selected at random from large speech corpora recorded for the purposes of expressive speech synthesis. All of the corpora were made in a professional studio using high quality equipment. The selected recordings were downsampled to 16 kHz and the original bit depth of 16 bits per sample was kept.

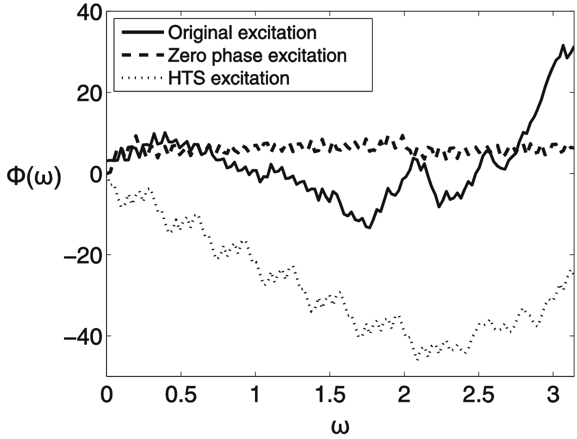

Fig. 4. Phase spectra of the original excitation, the excitation with phase spectrum set to zero, and the generic HTS excitation (pulse train with noise).

Firstly, in order to verify the proposed method for source-filter separation, the selected recordings were separated into their source and filter components and then resynthesized, with no phase manipulation (Experiment 1). The amplitude spectra of the original recordings were then compared frame by frame to the amplitude spectra of their resynthesized versions, and the results are shown in Table 1. The comparison was carried out on the full frequency range by calculating the amplitude of the FFT at 30 ms long successive overlapping frames with a shift of 5 ms, and then averaging the result. A very small distortion of the amplitude spectrum occurs, which was con-firmed by informal listening tests, where no perceptual difference between the original and the resynthesized sample was detected by any of the listeners.

Secondly, in order to verify the proposed phase manipulation algorithm and to examine the extent of the distortion of the amplitude spectrum that it introduces, the source component obtained in Experiment 1 was re-created with zero phase spectrum at voiced segments, while on unvoiced segments it was not modified in any way. It has been confirmed by listening tests that the different treatment of frames due to the difference in their voicing does not lead to audible discontinuities in resulting speech. The phase modification was limited to the frequencies below 5 kHz. The signal was then resynthesized from the source component modified in this way, using the filter component which was not altered (Experiment 2). The difference in the amplitude spectra was then calculated in the same way as in Experiment 1, and the results are shown in Table 2. The distortion that occurs is somewhat greater than in the first case, but can still be considered very small, while the informal listening tests reveal only a slight perceptual difference between the original and the resynthesized signal.

Table 1. Results of Experiment 1: Original vs. resynthesized recordings.

	Serbian		English		Average
	D (female)	J (male)	K (female)	D (male)	
Mean [%]	1.38	0.71	1.42	0.56	1.02
Variance [%]	0.36	0.14	0.23	0.06	0.20

Table 2. Results of Experiment 2: Original vs. zero-phase resynthesized recordings.

	Serbian		English		Average
	D (female)	J (male)	K (female)	D (male)	
Mean [%]	5.21	9.65	5.24	5.86	6.49
Variance [%]	0.63	0.98	0.66	1.07	0.83

5 Conclusion

In this paper we have proposed a novel technique for manipulating the phase spectrum of a speech signal by firstly decomposing the signal into its source and filter components by inverse filtering and then modifying the phase of the excitation on a sample-by-sample basis. The proposed technique thus avoids the need for overlap-add, which is often used for phase manipulation regardless of the fact that it can affect the amplitude spectrum significantly. The results of the preliminary experiments, including the resynthesis of the original speech using the original excitation signal and the excitation signal with phase spectra set to zero, confirm that the distortion of the amplitude spectrum caused by phase manipulation is indeed relatively small. This makes the proposed technique useful in a range of speech processing scenarios which require the modification of excitation phase spectrum, including the construction of excitation signals for parametric speech synthesis or speech vocoding. Our future work will include the re-implementation of most widely used phase modification algorithms from their descriptions in the literature, and their direct comparison with the proposed algorithm, both through objective measurements as well as more extensive listening tests.

Acknowledgments. The presented study was supported in part by the Ministry of Education and Science of the Republic of Serbia (grant TR32035), in part by the project "SP2: SCOPES Project for Speech Prosody" (No. CRSII2-147611/1), financed by the Swiss National Science Foundation, and in part by the company Speech Morphing, Inc. from Campbell, CA, USA, which also provided some of the speech corpora used in the experiments.

References

1. Ohm, G.S.: Über die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. Annalen der Physik und Chemie **135**(8), 513–565 (1843)
2. von Helmholtz, H.L.F.: Über die Klangfarbe der Vocale. Annalen der Physik und Chemie **18**, 280–290 (1859)
3. Plomp, R., Steeneken, H.J.M.: Effect of phase on the timbre of complex tones. J. Acoust. Soc. Am. **46**(2B), 409–421 (1969)
4. Schroeder, M.R.: Models of hearing. Proc. of the IEEE **63**, 1332–1350 (1975)
5. Oppenheim, A.V., Lim, J.S.: The importance of phase in signals. Proc. IEEE **69**, 529–541 (1981)
6. Patterson, R.D.: A pulse ribbon model of monaural phase perception. J. Acoust. Soc. Am. **82**(5), 1560–1586 (1987)
7. Paliwal, K.K., Alsteris, L.D.: On the usefulness of STFT phase spectrum in human listening tests. Speech Commun. **45**(2), 153–170 (2005)
8. Lim, J.S., Oppenheim, A.V.: Enhancement and bandwidth compression of noisy speech. Proc. IEEE **67**, 1586–1604 (1979)
9. Wang, D.L., Lim, J.S.: The unimportance of phase in speech enhancement. IEEE Trans. Speech Signal Process. **30**(4), 679–681 (1982)
10. Pobloth, H., Kleijn, W.B: On phase perception in speech. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 29–32 (1999)
11. Shi, G., Shanechi, M.M., Aarabi, P.: On the importance of phase in human speech recognition. IEEE Trans. Audio Speech Lang. Process. **14**(5), 1867–1874 (2006)
12. Schluter, R., Ney, H.: Using phase spectrum information for improved speech recognition performance. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 133–136 (2001)
13. Raitio, T., Juvela, L., Suni, A., Vainio, M., Alku, P.: Phase perception of the glottal excitation and its relevance in statistical parametric speech synthesis. Speech Communication (in press, 2016)
14. Sečujski, M., Ostrogonac, S., Suzić, S., Pekar, D.: Speech database production and tagset design aimed at expressive text-to-speech in Serbian. In: Proceedings of Digital Signal and Image Processing (DOGS), Novi Sad, Serbia, pp. 51–54 (2014)
15. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: The HMM-based speech synthesis system version 2.0. In: Proceedings of ISCA Speech Synthesis Workshop (2007)