Perceptual Importance of the Phase Related Information in Speech

Ibon Saratxaga[1], *Inma Hernaez*[1], *Michael Pucher*[2], *Eva Navas*[1], *Iñaki Sainz*[1]

[1] Aholab Signal Processing Laboratory, University of the Basque Country UPV/EHU, Spain
[2] Telecommunication Research Center Vienna, Austria

{ibon,inma,eva,inaki}@aholab.ehu.es, pucher@ftw.at

Abstract

The importance of phase information in the perceptual quality of the speech signals is studied in this paper. Many speech synthesisers do not use the original phase information of the signals assuming their contribution is almost inaudible. The Relative Phase Shift (RPS) representation of the phase allows straightforward phase structure analysis, manipulation and resynthesis, and we use these features to do a comparative evaluation of some phase modifications usually found in speech models. The final intention of this study is to get an answer to the question of whether phases deserve elaborate models to get high quality synthetic speech, or their subtle effects justify overlooking them.

Index Terms: phase perception, RPS, speech synthesis.

1. Introduction

Phase information has been usually neglected in speech synthesis. Most coding algorithms use parameters related to the spectral module of the speech signal, i.e. related to the energy of the signal at every frequency. The phase of the signal is typically either derived from the amplitude envelope (minimum phase systems for instance) or imposed by a mathematical rule. Usually, the only concern is to avoid phase mismatches or discontinuities between synthesis frames, because they produce audible clicks in the synthetic signal.

Recently, we have proposed a representation for the phase information of the speech signal: the Relative Phase Shift (RPS) [1]. This representation shows in a clear way the true structure of the phase of the signal, thus allowing modelling and manipulation of the phases. We have successfully applied the RPS analysis for applications involving phase modelling, namely ASR [2], speaker verification and synthetic speech detection [3] among others. Due to the ease of phase control through the RPS representation, the manipulation of the phases in synthetic speech seems very feasible. However, the question to be answered in the first place is: does phase contribute to the quality of the synthetic speech?

Before going ahead, we should clarify that we are referring to the phase of the short-time spectra of the signal. The phase rules the temporal distribution of the signal so it is obvious that in the long-time spectrum it plays a crucial role in the intelligibility of the speech signal.

The question of the perceptual effect of the phase distortion has been a controversial one for a long time. In his pioneering work "Über die Definition des Tones", Ohm stated in his "phase law", that the quality of the sound depends only on its spectral power. Helmholtz, some years later, arrived to the same conclusion, and his experiments founded the idea of the hearing organ being similar to a spectral power analyzer, where phase information would not play any significant role.

Hemlholtz's conclusions were rebated very soon by a series of experiments using ever improving equipment. The research in this field can be arranged into three lines. First, the electro-acoustic line, which paid more and more attention to the perceptual importance of the phase distortion once audio technology had reached a high quality level, which allowed focusing on second order effects. Early works demonstrated that hearing was not deaf to the phase in the low-medium frequency band: phase shifts can be distinguished in synthetic tones, and even some signal's polarity inversion can be detected. Nevertheless, as Lipshitz concludes [4], the phase distortions, although audible for certain signals and conditions, are very subtle and can usually be disregarded.

The second line of research about the phase importance in audition came from hearing physiology studies. Summing up, findings in this area show that phase, in terms of waveform, is actually detected at least at the neural level of the audition process for the low-medium band of frequencies. In this range of frequencies the human ear would act as a linear half-wave rectifier. This ability disappears for higher frequencies [5].

The third line of research about phases is the speech processing field. In this area the importance of the phases in the intelligibility of the speech signal has been evaluated by several authors [6], [7]. The results showed that the phase was important when the analysis frames where long, and mostly negligible for short windows.

Phases have also been investigated from the vocoder point of view to determine how much of the phase information can be discarded with no or little perceptual effect in the transcoded signal [8], [9]. Finally, few authors have evaluated the effect of modifying the phases in the overall quality of the audio signal: the experiments usually imply heavy phase modifications (fixed or minimum phase) and synthetic signals instead of real speech (e.g. [10]).

In our case, we wanted to know how important it is the perceptual relevance of the use or manipulation of the phase information (via the RPS representation) in speech signals. The literature does not give a clear answer: we have not found extensive evaluations applied to speech signals, but extrapolating from other kind of signals (synthetic and music) we can suppose that phase manipulations should be audible but very subtle, especially for minor phase changes. So we decided to do our own evaluation taking advance of the broad possibilities the RPS offers to manipulate and resynthesise the speech signals. Namely, we wanted to check the possible impairment, if any, produced by some usual workarounds adopted by many speech coding algorithms to discard the real phase information, as random or minimum phase assumption. Should we care about real phases in speech synthesis?

This experiment is explained in the following sections. First the RPS representation is reviewed. Second the phase modifications to be evaluated are described. Then, the experiment design is outlined and finally the results are presented. Some conclusions close the paper.

2. The RPS representation

The Relative Phase Shift (RPS) is a representation for the harmonic phase information described in [1]. Harmonic analysis models each frame of a signal by means of a sum of

sinusoids harmonically related to the pitch or fundamental frequency.

$$h(t)=\sum_{k=1}^{N}A_k\cos\left(\varphi_k(t)\right)\qquad \varphi_k(t)=2\pi k f_o t+\theta_k \tag{1}$$

where N is the number of bands, A_k are the amplitudes, $\varphi_k(t)$ is the instantaneous phase, f_0 the pitch or fundamental frequency and θ_k is the initial phase shift of the k-th sinusoid.

The RPS representation consists in calculating the difference between the instantaneous phase of every harmonic and the instantaneous phase of the fundamental component, at a fixed point of the period of the signal, namely the point t_o where the instantaneous phase of the first harmonic is zero, $\varphi_1(t_o)=0$. At that point (t_o), the phase difference will keep constant while the waveform keeps constant.

Although the RPSs refer to a fixed point of the period, the analysis itself is pitch asynchronous, provided we assume local stationarity for the signal. In an arbitrary analysis time instant, t_a, the instantaneous phases of the fundamental component and the k-th harmonic are:

$$\varphi_1(t_a)=2\pi f_1 t_a+\theta_1 \qquad \varphi_k(t_a)=2\pi f_k t_a+\theta_k \tag{2}$$

The RPSs (ψ_k) is the phase difference in t_o:

$$\psi_k(t_a)\equiv\varphi_k(t_o)-\varphi_1(t_o)=\varphi_k(t_o) \tag{3}$$

If the signal is stationary we can extrapolate the phase of the k-th harmonic in t_o calculating the time difference between t_a and t_o from the phase of the fundamental frequency:

$$t_o=t_a-\frac{\varphi_1(t_a)}{2\pi f_1} \tag{4}$$

And substituting in (2) and (3) we get

$$\psi_k(t_a)=\varphi_k(t_o)=\varphi_k(t_a)-k\varphi_1(t_a) \tag{5}$$

This is the RPS transformation which allows computing the RPSs from the instantaneous phases at any point of the signal. The RPS values are wrapped to the [-π, π] interval.

The same expression (5) was obtained in an independent work in [11], where it was proposed to be used as a phase correction of the signal in order to avoid linear phase mismatches in concatenative speech synthesis.

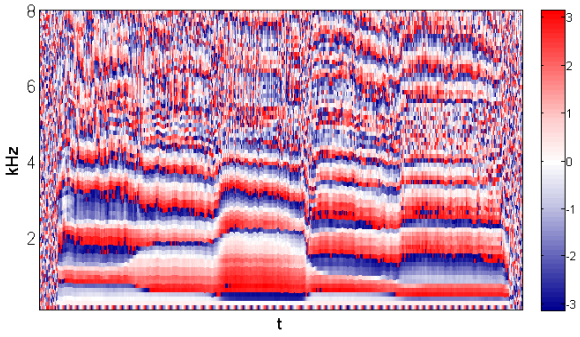

Figure 1: *RPS phasegram of a voiced segment /aeiou/.*

Among other interesting properties of the RPS a major feature is that it reveals a structured pattern in the phase information of the voiced segments. This can be noticed in Figure 1 which shows a "RPS phasegram" which, as its magnitude counterpart the spectrogram, shows the evolution along time of the RPS for each harmonic. The figure shows a voiced speech segment of five sustained vowels |*aeiou*|, where the stable pattern of every vowel can be clearly distinguished.

3. RPS manipulations

The RPS allows a very straightforward manipulation of the waveform shape. So we can evaluate very different transformations with different levels of preservation of the original information. Besides testing specific modelling properties of the RPS, we have chosen manipulations that reproduce the phase processing done by some vocoders. The transformations are explained next, and some examples are displayed in Figure 2, where RPS phasegram and a short segment of some phase manipulated test signals are shown.

3.1. Polarity inverted signals

Though literature from the acoustic research field has clearly stated that the polarity inversion of certain synthetic signals can be perceptible, it is a normal assumption in the speech technology area that polarity does not affect the speech signal. We wanted to check if this is really the case, so we have included this transformation in the test.

3.2. Time-constant RPSs

The most trivial transformation is to substitute the original RPS information by constant values, but it is a quite common workaround to generate the phases in several vocoders. Thus, we have that:

$$\psi_k(t_a)=\varphi_k(t_a)-k\varphi_1(t_a)=c_k \qquad \forall t_a \tag{6}$$

where c_k is a constant number. We will use two different values for c_k:

- Zero for every k. This transformation is called zero phase or cosine-phase in literature.
- A different random value in [-π, π] for each component.

3.3. Minimum phase

Due to its desirable properties, minimum phase response is a very common solution for imposing phase values in many vocoders, namely those based on LPC envelope modelling of the spectrum.

There is no direct way to convert RPS values to minimum phase RPS. Instead, prior to the RPS analysis, we convert the phases of the spectrum of every analysis frame to minimum phase using a cepstrum based non-parametric method [12]. Then, we have applied the RPS transformation to this minimum phase spectrum to obtain the minimum phase RPSs.

3.4. Phase models

The phase structure revealed by the RPS can be effectively modelled, as we have demonstrated applying it in an ASR application [2]. These models could also be useful for synthesis purposes, as they could allow the use of phase data in the coding information in a more efficient way than using the RPS values directly. We have evaluated two models:

- Linear model. This is the simplest possible model. For every frame, the RPS values are unwrapped along frequency, and the resulting curve is modelled by a line connecting the initial and final values. Admittedly, this is not a suitable model but we have chosen it as a limit case to check the sensitivity of hearing to such a rough phase change.
- The DCT-mel-RPS model. This is a variation of a real model which has produced good results for ASR tasks. It has been thoroughly explained in [2]. In this case we calculate the differences of the unwrapped RPS values,

filter them with a mel filterbank (32 filters) and apply a discrete cosine transform (DCT) to the resulting sequence. The DCT is truncated to 20 values.

The model parameters are used to recover the "original" RPS by reverting the calculations: an inverse DCT is first done, and the resulting differentiated RPS envelope is interpolated to obtain the values at the frequencies of the components. Finally, the discrete integral of those values is calculated, getting back the modelled RPS values. This model is not especially suitable for synthesis, because the interpolation and subsequent integration produces accumulative errors which distort the reconstructed values as frequency increases. Yet, the recovered phases are quite accurate for the low frequencies, where they should be more perceptible.

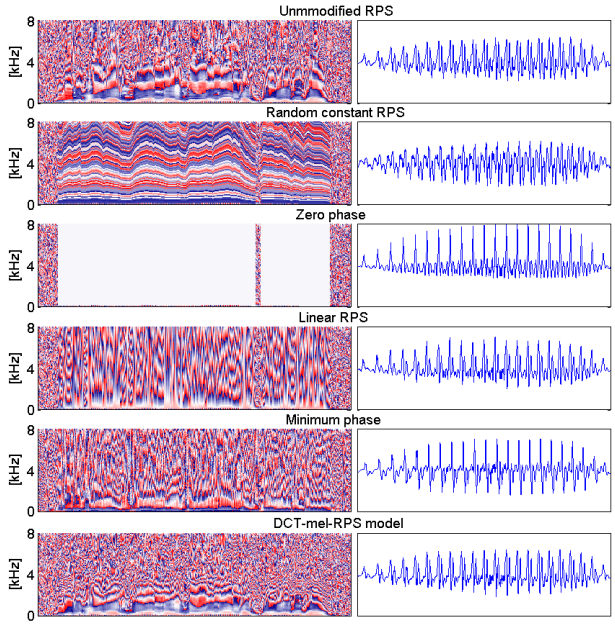

Figure 2: *RPS phasegram and signal sample for different phase manipulations.*

4. Design of the evaluation

The proposed phase modifications were expected to produce very subtle effects. Hence, the evaluation was designed trying to maximize the chances of hearing them.

This kind of evaluation presents some problems to be addressed. First, it has to be noted that the resynthesis technique itself can introduce degradations. To compensate this the reference signals with original phases are resynthesised. An additional distortion can appear because the harmonic analysis is not perfect and some energy of the stochastic component can be modelled as periodic. Subsequent phase manipulation which implies assigning deterministic phase to such components can produce tonal noises, which could be erroneously interpreted.

We have tried to overcome this effect by two means: on the one hand, we have used a voiced-only signal corpus for the evaluation, which minimises the stochastic energy and thus the leakage. On the other hand, we have applied a long analysis window (3 pitch periods, 10ms framerate) to produce a more robust estimation of the harmonic part. In any case, this problem is inherent to the harmonic assumption and usual phase manipulations worsen it. It is a factor to be taken into account when deciding which phase treatment will be chosen and consequently, it does not invalidate the evaluation results.

Other problems are related to the subtlety of the phase impairments and its perception. This applies to the choice of the speech signal corpus, the synthesis method, the evaluation procedure and even to the evaluators' profile. We explain all these aspects in the following subsections.

4.1. Evaluation corpus

Apart from the aforementioned stochastic leakage problem, the choice of a corpus with just voiced phonemes is also convenient to make it easier for the evaluators to concentrate on the timbre variations due to the phase, because the phase manipulated segments of the signal are longer and without interspersed unvoiced fragments.

Thus, we have recorded a speech database of voiced signals with several non professional speakers, both males and females in Spanish language (3 males and 3 females). 10 sentences were recorded by every speaker and 8 of them were selected to be used in the evaluation.

4.2. RPS manipulation and resynthesis

To generate the RPS manipulated versions of the signals we used our own implementation of a harmonic plus stochastic coding system, which supports RPS for the phases: the so-called Harmonic plus Stochastic with Iterative Multiband Analysis (HSM-iMBA) system. This algorithm models the speech signal by a sum of harmonic and stochastic parts.

The harmonic part is modelled by means of a sum of harmonic sinusoids weighted by different amplitudes and shifted some initial phases. The amplitudes are obtained by means of a modified iterative multiband excitation (MBE) analysis and the phases are derived from the spectrum and transformed into RPSs. The stochastic residual is calculated by spectral subtraction of the harmonic part and extends along the whole bandwidth. It is modelled by means of Gaussian white noise filtered by a LPC filter.

Once the analysis is done, the RPS data are manipulated, and the signals are resynthesised (sampling frecuency 16 kHz). In order to avoid any effect which could hinder the perception of the phase effects, the stochastic part has been suppressed in the resynthesised signals. The original signals, with unmodified RPS, were also resynthesised to be used as references in the test.

4.3. Evaluation procedure

We have used the double blind triple stimulus with hidden reference method to subjectively evaluate the small impairments between signals, as it has been proposed by the ITU [13]. The evaluators listen to three signals: the reference, A and B. One of A or B (randomly assigned) is the same signal as the reference, i.e. synthesised with the unmodified RPSs, and the other is the manipulated version. The evaluators have to rate the degradation of the two signals relative to the reference, using a 1 (very annoying) to 5 (imperceptible) continuous scale. They know that either A or B are the same as the reference so one of them should be rated 5.

Evaluators are also provided with some example signals to become familiar with the effects they will have to rate in the test and how to apply the grading scale. The evaluators are native speakers, and most of them are experts in speech technologies. They did the test via a webpage using their own audio equipment, always using headphones. In every test a sentence was randomly chosen from two male and two female speakers and the 6 transformations had to be evaluated for every sentence. 19 people took part in the evaluation.

5. Results

The first evident result is that the evaluation design has been quite successful in making the perceptual effects of the phase apparent, because scores are more categorical than we could expect from the literature. Detection rate, for instance, is shown in Figure 3a. The bar labelled 'OK' shows the percentage of phase-manipulated signals that were correctly detected as manipulated (75.8%), 'NOK' shows the mistaken detections, and 'Ind.' the indistinguishable ones.

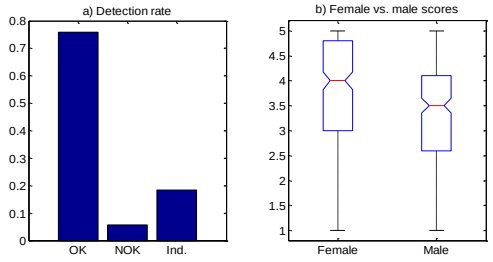

Figure 3: *Phase manipulated signals: a) Detection rate. b) Scores by speaker gender.*

Figure 3.b shows the grouped scores for phase manipulated female and male voices. Male voices seem to be more affected than female ones, obtaining worse scores (about 0.5 point less than the female counterpart). This is consistent with published results stating that the human ear is sensitive to low-medium frequency phase information: the male voice has more components in these low bands and so the phase manipulation is more manifest.

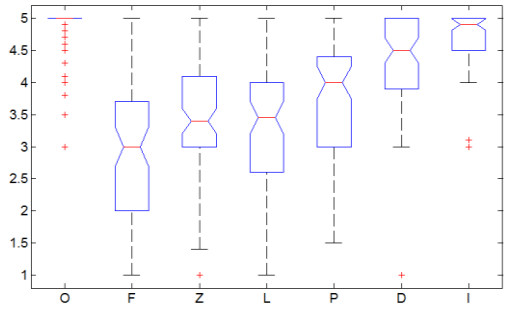

Figure 4: *Scores by manipulation type.*

It is also clear that phase, to some extent, matters. The results are quite consistent and scores are meaningfully related with the phase manipulation level, as it is shown in figure 4. It is worth noting that the variances in the scores are quite high, suggesting that the perception is highly dependent on the actual signal and/or the listener. We have tested the statistical significance of the scores of every manipulation using the Wilcoxon's matched pairs signed-ranks test, with a threshold of p=5%. According to this analysis, every manipulation produces worse scores than the original signal. There is a group composed by rough phase manipulations like constant phase assumption (zero-phase, 'Z' or random initial phase, 'F'), or linear models ('L') which have no significant differences among them, but they all show important degradation against the original unmodified signal ('O').

Minimum phase ('P') assumption stands in the middle of the impairment effects, producing a remarkable but mostly not annoying degradation in the speech quality. At the next level, the phases derived from our proposed DCT-mel-RPS ('D') model perform notably well, producing a slight and not annoying impairment. Finally, and quite unexpectedly for speech signals, polarity inversion ('I') can sometimes be detected and gives a very slight although statistically significant impairment against the original.

6. Conclusions

In this work we have evaluated the perceptual impairments produced in speech signals when their original phase information is disregarded and substituted by different approximations. We have used the RPS representation of the phases, which allows a complete control of the phase structure.

Our results show that these manipulations produce audible degradation of the speech signal, thus suggesting that signal quality can be improved using more elaborate phase models. We evaluate one of these models, the DCT-mel-RPS which performs well, producing very small impairments. In order to soundly appreciate the results, it has to be taken into account that the evaluation procedure was designed so as to maximize the perception of the phase effects, so results would probably be less marked with everyday speech signals. Nevertheless, it seems that phase information should be taken into account to produce the highest possible quality synthesis.

7. Acknowledgements

The authors want to thank all the evaluators for their effort. This work has been supported by the Spanish Gov. (TEC2009-14094-C04-02), the Basque Government (IE09-262, MV20090225) and the Austrian Science Fund (FWF): P23821-N23. The Competence Center Forschungszentrum Telekommunikation Wien GmbH (FTW) is funded within the program COMET by BMVIT, BMWA, and the City of Vienna. The COMET program is managed by the FFG.

8. References

[1] Saratxaga, I., Hernáez, I., Erro, D., Navas, E. and Sánchez, J., "Simple representation of signal phase for harmonic speech models", Electronics Letters 45: 381-383, 2009.

[2] Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., and Erro, D., "Using Harmonic Phase Information to Improve ASR Rate", Interspeech 2010, 1185-1188, 2010.

[3] Leon, P.D., Hernáez, I., Saratxaga, I., Pucher, M. and Yamagishi, J., "Detection of synthetic speech for the problem of imposture", ICASSP 2011, 2011.

[4] Lipshitz, S.P., Pocock, M. and Vanderkooy, J., "On the audibility of midrange phase distortion in audio systems," J. Audio Eng. Soc. 30: 580–595,1982.

[5] Pickles, J. O., "An Introduction to the Physiology of Hearing", Academic Press Inc, 2008.

[6] Liu, L., He, J., Palm, G., "Effects of phase on the perception of intervocalic stop consonants", Speech Com. 22, 403–417, 1997.

[7] Alsteris, L.D. and Paliwal, K.K., "Evaluation of the modified group delay feature for isolated word recognition," Procs. Signal Processing and its Applications 2. 715-718, 2005.

[8] Pobloth, H. and Kleijn, W. B. "On phase perception in speech," ICASSP 1999, 11-14, 1999.

[9] Kim, D. "On the perceptually irrelevant phase information in sinusoidal representation of speech," IEEE Trans. Speech and Audio Processing 9(8), 900-905, 2001.

[10] Banno, H., Lu, J., Nakamura, S., Shikano, K. and Kawahara, H. "Efficient representation of short-time phase based on group delay," ICASSP 1998, 861–864, 1998.

[11] Stylianou, Y., "Removing linear phase mismatches in concatenative speech synthesis," IEEE Trans. Speech and Audio Processing 9(3), 232-239, 2001.

[12] Smith, J.O., "Introduction to Digital Filters with Audio Applications," 2010.

[13] ITU-R, "ITU-R BS 1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunications Union, Geneva Switzerland, 1997:1-26, 1994.