Phase in the harmonic models of the speech signal: strategies for representation, processing and applications

Ibon Saratxaga, supervised by Inma Hernaez

Aholab Signal Processing Laboratory, University of the Basque Country UPV/EHU, Bilbao
ibon@aholab.ehu.es, inma@aholab.ehu.es

Abstract. This PhD dissertation was written by Ibon Saratxaga and supervised by Inma Hernáez. It was defended at the University of the Basque Country the 5th of March 2012. The committee members were Prof. José Manuel Pardo (UPM), Prof. Asunción Moreno (UPC), Dr..Eduardo Rodríguez Banga (UVigo), Dr. Eva Navas (UPV/EHU) and Dr. Daniel Erro (UPV/EHU).The dissertation was awarded an "apto cum laude" qualification.

Keywords: Phase, harmonic models, polarity, ASR, speaker recognition, speech perception.

1 Introduction

Phase spectrum information of the speech signal has traditionally been neglected by many applications in favour of the information derived from the spectral module. Phase information poses some difficulties from the signal processing viewpoint, due to its special characteristics like time and frequency dependency, phase-wrapping and representation difficulties. Moreover, phase spectrum seems to have little perceptual importance for the human hearing, its influence being of second order compared to the signal module spectrum.

These reasons and the fact that the spectral module is easy to understand and manipulate and it is directly related to perception, have favoured the exclusive use of the module spectrum both in synthesis applications (speech synthesis, voice conversion and transformation, for instance), and in analysis applications (speech and speaker recognition, among others). Let's review the role of phases in these two areas.

1.1 Phases and speech synthesis

In the speech synthesis area phases are often regarded as a problem instead of as a useful feature of the speech. In the case of sinusoidal [1][2] and harmonic models [3][4], for instance, the instantaneous phases of the sinusoidal components are calculated. Instantaneous phase depends on the time of analysis, so using it often implies further processing to assure that no phase mismatches between frames occur, especially when the signal is manipulated (pitch or time modifications) or concatenated.

To solve these problems, the linear phase term (which depends on the analysis instant) has to be removed from the instantaneous phase. Several techniques have been proposed to do this, e.g.: pitch synchronous analysis, calculation of constant points in every pitch period (onset times [5], gravity centres[6] or glottal closure instants).

Other models for speech synthesis, like those based on LPC, directly disregard the phase information and use artificial phase (usually minimum phase but also random or zero phase) to generate the synthetic speech.

These two approaches have direct consequences in the waveform shape of the resulting signal: changing the original phases leads to losing the original waveform shape. The actual importance of keeping the signal shape invariant is not definitely stated. Several authors [5],[7] have proposed shape invariant speech signal manipulation methods, but the effect of the phase manipulation (which leads to waveform shape modification) in the perceptual quality of the speech is not well studied.

The perceptual importance of the phases has been well studied in areas such as electro acoustics and hearing physiology, in general using artificial or musical sounds. In the area of speech technologies, phase contribution to the intelligibility has been evaluated by several authors [8], [9]. 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 [10], [11]. But very 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. [12]).

Summing up, the role of the phase in the speech synthesis domain can be described as uncertain: phases are tough to manipulate, are often discarded and it is not clear if they have a relevant paper in the resulting speech quality. This thesis tries to bring some light into these questions.

1.2 Phases and speech analysis

Short time spectral information plays a fundamental role in speech analysis, as it is in the base of most of the speech modelling techniques. However, in speech or speaker recognition applications, usually only spectral magnitude is used, while phase information is directly discarded. Phase lacks apparent structure or patterns and this is one reason not to use it for modelling purposes.

Nonetheless, there have been several attempts to include phase information in recognition applications. Group delay related information has been used both in automatic speech recognition (ASR) [13], [14] and in speaker recognition models [15], [16], [17]. In this last area several authors report significant benefits when phase information is combined with magnitude (MFCC) based parameters [18], [19], [20].

One area in speech analysis that is still an open problem is the detection of synthetic impostors, that is, synthetic speech which imitates a speaker's voice who is authorized in a speaker verification system. Current speaker verification systems have serious problems detecting high quality synthetic voice as synthetic and thus rejecting it as false claimant [21], [22]. Phase information could help to discriminate synthetic

speech in many cases, namely those where the original phases have been disregarded or severely altered.

2 Motivation and Objectives

In the framework of the harmonic models for the speech, we have proposed a new representation for the phase information whose features simplify its processing. The RPS (Relative Phase Shift) transformation converts the instantaneous phase of the spectrum to a measure of the relative phase shift between each harmonic component and the fundamental frequency.

This transformation eliminates the dependency of the phase with the frequency and the analysis instant. The RPSs are only dependent on the initial phase shifts between components, and thus, their interpretation and manipulation is much easier than the instantaneous phase's one. This representation is convenient for phase manipulation in synthesis using harmonic models, but more important, it makes the signal phase structure evident, as can be seen in the following figure.

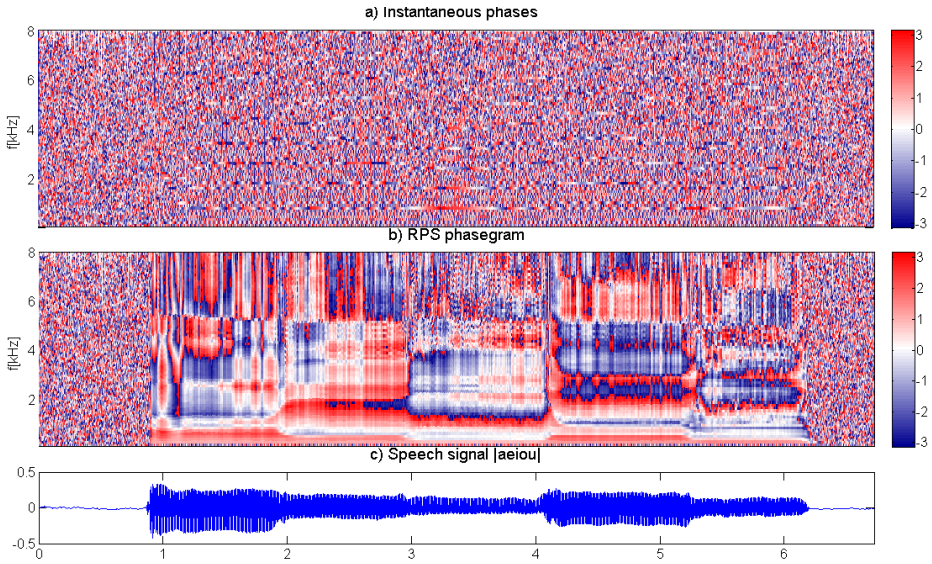

Fig. 1. Instantaneous phase (a) and RPS (b) phasegrams for the signal /aeiou/ (c)

Fig. 1 shows two phasegrams (similar to the spectrograms, but showing phase instead of module) of a speech signal with sustained vowels /aeiou/. The difference between (a) where the instantaneous phases do not show any structure and the clear phase patterns that arise in the RPS phasegram (b) is evident.

The main motivation of the thesis is to study the features and potential applications of this new phase representation, the RPS. Therefore, the main objectives of the thesis can be summarized as follows.

2.1 Objectives

The principal objective of the thesis has been to analyse and understand the potential of the phase information represented by means of the RPS, in different areas of the speech signal processing domain. This general objective was split into more specific ones:

- Phases in the voice signal characterization: Interpretation of the RPS and analysis of the relationship between their features and the voice characteristics like formants, polarity or source and vocal tract separation.
- Phase modelling: Finding a parameterization for the phase information suitable for statistical modelling.
- Phases in the speech signal analysis: Research on the influence of phoneme and speaker features on the phase information of the signal. Application in several tasks in the speech analysis area: speech recognition, speaker recognition and synthetic speech detection.
- Phases and perception: Evaluation of the perceptual influence of the phase manipulations, trying to answer to the question of the real importance of implementing phase handling strategies in speech synthesis.

3 Methodology

Due to the broadness of the objectives of the thesis, the research has been divided in several separate stages with different methodologies employed in each of them. Generally speaking, for each of the areas that have been covered in the thesis we have reviewed the relevant literature, focusing both in the state of the art of the area and in the reported applications of the phase information. Then, depending on the case, we analyse any relevant feature of the RPS, usually by means of some experiments or evaluations, comparing the performance against canonical systems when possible.

The specific methodology applied in each part of the dissertation can be summarized in:

- Design and implementation of a harmonic plus stochastic speech model including support of the RPS representation. For the analysis and interpretation of the RPS several tools have been developed: a specifically adapted pitch detection algorithm, CDP [23], polarity detection algorithms, inverse filtering tools, phasegram rendering tools, etc.
- Development of a parameterisation technique suitable for statistical modelling. The methodology here has been to evaluate different candidate parameter sets by means of an ASR application choosing those that maximise the accuracy of the system. The ASR used has been a canonical system (HTK) with a SpeechDat-like database in Basque language.
- In order to accomplish the objective of analysing how speech or speaker characteristics are reflected in the RPSs we have implemented and trained basic systems for speech recognition and speaker identification using the RPSs as parameters. Be-

sides evaluating the performance of the RPS based systems we have also compared it with the results of baseline MFCC systems, as well as combinations of both module and phase parameters. In all the cases the objective was not to evaluate the actual improvement RPS data can confer to state-of-the-art systems in these areas, but to test whether the RPSs carried information about these features of the speech signal.

- For the evaluation of the RPS application to the detection of synthetic impostors in speaker verification systems a module for synthetic speech detection was implemented, which works with the accepted signals of a standard GMM based speaker verification system. In this case we have compared the performance of the whole verification system with and without the synthetic detection module.
- Finally, for the evaluation of the perceptual importance of the phase manipulations in speech, we have designed a subjective quality evaluation aiming to highlight the subtle degradation that phase can introduce in speech signals. The evaluation procedure is based on the ITU-R BS 1116-1 recommendation for the subjective assessment of small impairments in audio [24], and was implemented in a web page. We have also designed and recorded a multispeaker and multilingual database to be used in this test: the utterances contain only voiced phonemes, and it consists of 10 sentences per speaker. There are speakers in 3 languages: 6 males and 6 females in Spanish and Basque languages, and 3 males and 3 females in German language.

4 Main Contributions of the Thesis

4.1 The Relative Phase Shift representation

The first contribution of the thesis is the Relative Phase Shift representation itself. From the instantaneous phases, calculated by means of the Fourier transform and the pitch, the proposed 'RPS transformation' allows calculating the RPS values in a pitch asynchronous way.

The new representation of the phases is useful when changing the duration and pitch of a speech signal by means of harmonic models. The RPSs allow calculating the phases for the modified signals in a straightforward way, similar to what it is done with the amplitudes. This way it is not necessary to do specific adjustments in the phases or limiting the analysis to specific pitch period points as other published methods require.

The RPS representation is more than a technique to ease phase manipulation. It shows many desirable properties, notably its direct relationship with the waveform of the signal and this allows direct visualization of the structure of the speech phases.

The relationship of the RPS with speech formants and source and vocal tract separation has also been analysed and described.

4.2 DCT-mel-RPS parameterization

The evident phase patterns that arise with the RPS phasegrams can be useful for humans but they are hardly usable by automatic applications. That is why we have developed a parameterization of the RPS values which is suitable for statistical modelling. This parameterization copes with some problematic characteristics of the RPS like wrapping, high dimensionality, variable number of values per frame at different frequencies, etc. The DCT-mel-RPS parameterization applies mel filtering and discrete cosine transform (DCT) to reduce the dimensionality of the original RPS values with the minimum loss of information. This parameterization has been tested with HMM based models in an ASR system and with GMM based models for speaker identification. In both experiments the parameterization has proven effective in this kind of statistical modelling.

In the experiments we have also proved that phase information is useful in these recognition systems either jointly with spectral module information (in the form of MFCC parameters) in the case of ASR, either just alone in the case of speaker identification. In both cases the inclusion of phase information can improve the results of traditional module only systems.

4.3 Polarity detector

One notable feature of the RPSs is the ability to distinguish the polarity of the signal. Inverting the polarity produces shifts of π radians in the RPSs values of the even components. Taking advantage of this fact we have proposed a polarity detection algorithm based on the comparison of the ripple of the RPS envelope along frequencies for every frame. This algorithm has been evaluated and compared with others reported in the literature and has obtained some excellent results even with short signals, outperforming the other algorithms.

4.4 Synthetic impostor detection

Phase information has been used to detect synthetic impostors in speaker verification systems with excellent results. This is the first time that the use of the signal phase has been proposed to discriminate the speech signal produced by synthesisers which imitate the voice of a specific person, and it opens a promising research line in an area where there were difficulties for this kind of impostors' detection with other methods.

The impostor detection system we have developed works as a separate stage after the verification system, trying to reject the synthetic impostors that have been wrongly accepted by the verification system. The synthetic speech detection (SSD) system models separately both the natural and the synthetic speech for every speaker, using the DCT-mel-RPS parameters, and classifies the claimant's speech accordingly.

4.5 Perceptual evaluation

One of the initial ideas for the thesis was to improve the treatment of the phases in the traditional speech synthesis systems where they are discarded or substantially modified. However, previously, we needed to know if caring about phases in synthesis was worth the effort, that is, if phase manipulations are perceived as impairments in the speech signal. To answer this question, we have evaluated speech signals with several phase transformations and compared them with the original one. This kind of experiments have been done before for other kind of audio signals (music, synthetic tones, etc.) but to our knowledge, there were no studies about the specific impact of the phase on the overall speech signal quality.

The results allow us to assure that human hearing is sensitive to the phase of the speech signal, and that major or less natural phase modifications are perceived as impairments. Nevertheless, concerning the convenience of adding the original phase information to the synthesis systems, the results are not categorical. Although the degradation produced by certain phase modifications is undeniable according to the results of the evaluation, it has to be noted that the test conditions were carefully designed to highlight the perception of these impairments, so their effect in a real system will probably be less perceptible than what our evaluation suggests.

In this sense it is worth noting the good results that the DCT-mel-RPS parameterization has obtained, with almost unnoticeable impairment in speech signals with phases calculated from this parameters.

5 Conclusions and opened research lines

As a result of this thesis, contributions were made in the speech phase representation and its application in different areas of speech processing, as has been detailed in the preceding section. Summing up, the thesis tries to highlight the usefulness of this new representation for the phases to be applied in various speech processing areas. The RPS representation removes the time dependency of the phase information, linking it to the waveform shape of the signal. These features make the RPS suitable to be used in applications where only spectral module information was used till now. We think that the experiments detailed in the thesis, though preliminary, can open the door for future and more extensive research to apply the RPS in these and other areas.

In this sense, one of the outcomes of this thesis is the opening of new research lines, which can be summarized in three blocks:

5.1 Improved RPS analysis

The calculation of the RPS can be improved to make it more robust to noisy signals, pitch errors or previous processing of the signal. This last aspect is crucial in order to apply the RPS analysis with telephone and coded speech signals.

5.2 Evaluation of the performance of the RPS in state of the art recognition systems

The experiments of the use of RPS information in ASR and speaker recognition systems detailed in the thesis were devised to check if the phase information was relevant for these tasks, and thus, classical recognition systems were employed. Since the results of these tests have proved the utility of the RPS, more research should be done to determine how the use of RPS data could improve the recognition rate in state of the art recognition systems, also using standard evaluation databases.

5.3 New application areas for the RPS

There are more areas of the speech processing domain where phase information can be applied. The results of the perception test clearly suggest the use of the DCT-mel-RPS model in a high quality statistic speech synthesis system. Also the low perceptual impact of the phases could be used to add information in the speech signal either for watermarking, or data hiding purposes.

6 Publications

The main results of this thesis have been published both in journals and conferences. The RPS representation has been described in a journal paper [25], and so has been the synthetic impostor detection [26]. Other results have been published in several important conferences: the polarity detection algorithm was described in [27] the application of the RPS in ASR systems has been reported in [28], speaker recognition results have been reported in [29], additional impostor detection experiments were also published in [30] and the evaluation of the perceptual importance of phases in speech signals has been published in [31].

Other research collaterally related with the thesis, namely the works around the pitch detection algorithm (CDP), have been published in [23] and [32]. This last paper received an award to one of the best student papers in the 5th JTH conference.

References

1. McAulay, R.J., Quatieri, T.F.: Speech analysis/synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech and Signal Processing. 34, 744–754 (1986).
2. George, E.B., Smith, M.J.T.: Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. IEEE Transactions on Speech and Audio Processing. 5, 389–406 (1997).
3. Stylianou, Y.: Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification, PhD Thesis, (1996).
4. Erro, D.: Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models, PhD Thesis, (2008).
5. Quatieri, T.F., Mcaulay, R.J.: Shape invariant time-scale and pitch modification of speech. IEEE Transactions on Signal Processing. 40, 497–510 (1992).

6. Stylianou, Y.: Removing linear phase mismatches in concatenative speech synthesis. IEEE Transactions on Speech and Audio Processing. 9, 232–239 (2001).
7. O'Brien, D., Monaghan, A.: Shape Invariant Pitch and Time-Scale Modification of Speech Based on a Harmonic Model. Improvements in Speech Synthesis. pp. 64–75. Wiley Online Library (2002).
8. Liu, L., He, J., Palm, G.: Effects of phase on the perception of intervocalic stop consonants. Speech Communication. 22, 403–417 (1997).
9. Alsteris, L.D., Paliwal, K.K.: Evaluation of the modified group delay feature for isolated word recognition. Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005. pp. 715–718. IEEE (2005).
10. Pobloth, H., Kleijn, W.B.: On phase perception in speech. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings. pp. 11–14 (1999).
11. Kim, D.: On the perceptually irrelevant phase information in sinusoidal representation of speech. IEEE Transactions on Speech and Audio Processing. 9, 900–905 (2001).
12. Banno, H., Lu, J., Nakamura, S., Shikano, K., Kawahara, H.: Efficient representation of short-time phase based on group delay. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98. pp. 861–864. IEEE (1998).
13. Alsteris, L.D., Paliwal, K.K.: Short-time phase spectrum in speech processing: A review and some experimental results. Digital Signal Processing. 17, 578–616 (2007).
14. Murthy, H.A., Gadde, V.R.R.: The modified group delay function and its application to phoneme recognition. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings.(ICASSP'03). pp. 2–5 (2003).
15. Padmanabhan, R., Parthasarathi, S., Murthy, H.A.: Robustness of phase based features for speaker recognition. Proc. Interspeech 2009. pp. 3–6 (2009).
16. Kua, J.M.K., Epps, J., Ambikairajah, E., Choi, E.: LS regularization of group delay features for speaker recognition. Proc. Interspeech 2009. pp. 2887–2890 (2009).
17. Hegde, R.M., Murthy, H.A., Gadde, V.R.R.: Application of the modified group delay function to speaker identification and discrimination. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. p. I–517–20. IEEE (2004).
18. Murty, K., Yegnanarayana, B.: Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Processing Letters. 13, 52–55 (2005).
19. Zheng, N., Lee, T., Ching, P.C.: Integration of Complementary Acoustic Features for Speaker Recognition. IEEE Signal Processing Letters. 14, 181–184 (2007).
20. Wang, L., Minami, K., Yamamoto, K., Nakagawa, S.: Speaker identification by combining MFCC and phase information in noisy environments. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. pp. 4502–4505. IEEE (2010).
21. Masuko, T., Hitotsumatsu, T., Tokuda, K., Kobayashi, T.: On the security of HMM-based speaker verification systems against imposture using synthetic speech. Proceedings of the European Conference on Speech Communication and Technology. pp. 1223–1226. Citeseer (1999).
22. De Leon, P.L., Apsingekar, V.R., Pucher, M., Yamagishi, J.: Revisiting the security of speaker verification systems against imposture using synthetic speech. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1798–1801. IEEE (2010).
23. Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sanchez, J., Sainz, I.: Evaluation of pitch detection algorithms under real conditions. IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. pp. 1057–1060 (2007).

24. ITU-R: ITU-R BS 1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, (1994).
25. Saratxaga, I., Hernáez, I., Erro, D., Navas, E., Sanchez, J.: Simple representation of signal phase for harmonic speech models. Electronics Letters. 45, 381–383 (2009).
26. De Leon, P.L., Pucher, M., Yamagishi, J., Hernaez, I., Saratxaga, I.: Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. IEEE Transactions on Audio, Speech, and Language Processing. 20, 2280–2290 (2012).
27. Saratxaga, I., Erro, D., Hernáez, I., Sainz, I., Navas, E.: Use of Harmonic Phase Information for Polarity Detection in Speech Signals. Proc. Interspeech 2009. 1075–1078 (2009).
28. Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., Erro, D.: Using Harmonic Phase Information to Improve ASR Rate. Proc. Interspeech 2010. pp. 1185–1188 (2010).
29. Hernáez, I., Saratxaga, I., Sanchez, J., Navas, E., Luengo, I.: Use of The Harmonic Phase in Speaker Recognition. Proc. Interspeech 2011. pp. 2757–2760 (2011).
30. De Leon, P.L., Hernáez, I., Saratxaga, I., Pucher, M., Yamagishi, J.: Detection of synthetic speech for the problem of imposture. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 4844–4847 (2011).
31. Saratxaga, I., Hernaez, I., Pucher,, M., Navas, E., Sainz, I.: Perceptual Importance of the Phase Related Information in Speech. Proc. Interspeech 2012. ISCA, Portland, OR (2012).
32. Saratxaga, I., Luengo, I., Navas, E., Hernáez, I., Sánchez, J., Sainz, I.: Detección de pitch en condiciones adversas. Proc. V Jornadas en Tecnología del Habla. pp. 13–18 (2006).