

# Using Harmonic Phase Information to Improve ASR Rate

Ibon Saratxaga, Inma Hernandez, Igor Odriozola, Eva Navas, Iker Luengo, Daniel Erro

Aholab Signal Processing Laboratory, University of the Basque Country, Bilbao

{ibon, inma, igor, eva, iker, derro}@aholab.ehu.es

## Abstract

Spectral phase information is usually discarded in automatic speech recognition (ASR). The Relative Phase Shift (RPS), a novel representation of the phase information of the speech, has features which seem to be appropriate to improve the ASR recognition rate. In this paper we describe the RPS representation, discuss different ways to parameterize this information in a suitable way for the HMM modelling, and present the results of the evaluation experiments. WER improvements ranging from 12 to 22% open promising perspectives for the use of this information jointly with the classical MFCC parameterization.

**Index Terms:** ASR, phase spectrum, harmonic analysis

## 1. Introduction

Automatic Speech Recognition (ASR) employs short time spectral information to model appropriate speech units. The spectrum of a signal is complex and has magnitude and phase components, but usually only magnitude information is used by ASR systems. Magnitude information is directly related with spectral power density, and therefore with the formant structure and intelligibility, and it can be parameterized in a relatively easy way. By the contrary, contribution of the short-time phase information to the speech intelligibility is controversial (see [1] for an extensive review) and phase information is intrinsically tricky and difficult to model mainly due to the wrapping problem.

Several attempts to include phase information in ASR have been reported. Murthy and Hegde [2][3] proposed a transformation on the group delay function (the negative frequency derivative of the instantaneous phase of the signal) and used this modified group delay function (MODGDF) jointly with the MFCC coefficients in several ASR tasks. This MODGDF representation has also been extensively studied by Alsteris and Paliwal [1], who tested it with clean and noisy signals. The results showed that MODGDF in isolation did not provide improvement over the MFCCs, and when used jointly with the MFCCs there were some slight improvements but they were not consistent across the different experiments.

More group delay based representations were proposed by Bozkurt and Couvert [4] who obtained slightly better results in ASR tasks using joint MFCC and phase information than the MFCC only baseline. Group delay information has also been used in several related applications like: speaker recognition [5], speaker identification [6][7], gender identification [8] and others.

In the context of our work in harmonic analysis of the speech we have proposed a new representation of the harmonic phase information [9], called Relative Phase Shift (RPS). This representation of the difference in the phase shifts of the pitch harmonic components shows clear patterns across time and frequency which lead naturally to check its applications in the ASR field.

This paper describes the RPS representation and its application to a simple speech recognition task. The paper is structured as follows: In the next section the Relative Phase Shift representation is developed. In section 3 several features of the RPSs are studied in order to obtain a parameter set appropriate for HMM modelling. Section 4 describes the experiments and results carried out. The final conclusions and future works are summarized in the last section.

## 2. Relative Phase Shift Representation

The Relative Phase Shift is a representation for the harmonic phase information and was described in [9]. Harmonic analysis models each frame of a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency.

$$h(t) = \sum_{k=1}^N A_k \cos(\varphi_k(t)) \quad \varphi_k(t) = 2\pi k f_0 t + \theta_k \quad (1)$$

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  $\theta_k$  is the initial phase shift of the  $k$ -th sinusoid.

Usually the term ‘‘phase’’ is applied to the whole instantaneous phase of every sinusoid,  $\varphi_k(t)$ , instead of the initial phase shift  $\theta_k$ . This instantaneous phase changes depending on the analysis instant as well as on the frequency of the harmonic, due to the linear phase term  $2\pi k f_0 t$ . On the contrary, the initial phase shift ( $\theta_k$ ) is constant while the waveform shape is stable under the assumption of local stationarity, regardless of the time instant chosen for the analysis.

The initial phase shift determines the waveform shape of the signal. For a given set of harmonic sinusoids the resulting waveform shape depends only on the differences between the initial phase shifts ( $\theta_k$ ) of the components, which we call Relative Phase Shift (RPS). These RPSs are also constant as long as the initial phase shifts are so. Thus, they can be calculated at any analysis point wherever local stationarity conditions can be assumed, avoiding the necessity of determining any special point for the analysis. Being relative, the RPSs are computed using a common reference. The fundamental frequency,  $F_0$ , being the basic harmonic component, constitutes the natural one.

We have developed an expression to obtain the relative differences of the initial phase shifts from the measured instantaneous phases. Let us consider two sinusoids:

$$x_1(t) = \cos(2\pi f_1 t + \theta_1) \quad x_k(t) = \cos(2\pi f_k t + \theta_k) \quad (2)$$

where  $x_1(t)$  will be the reference sinusoid with frequency  $f_1$  and  $x_k(t)$  another sinusoid with frequency  $f_k > f_1$ .  $\theta_k$  is the initial phase shift and  $t$  stands for time. For the sake of simplicity we will consider  $\theta_1=0$ , which implies setting the time origin at the point where  $x_1(t)$  has instantaneous phase 0. For any arbitrary analysis point ( $t_a$ ) the instantaneous phases are:

$$\varphi_1(t_a) = 2\pi f_1 t_a \quad \varphi_k(t_a) = 2\pi f_k t_a + \theta_k \quad (3)$$

In the case of harmonic analysis,  $f_1$  will be the fundamental frequency ( $f_0$ ) and the frequencies of the two sinusoids will be harmonically related, so  $f_k = kf_1$ . Applying this condition, we get the RPS:

$$\theta_k = \varphi_k(t_a) - k\varphi_1(t_a) \quad (4)$$

Finally the RPS is wrapped to values in the  $[-\pi, \pi]$  interval.

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. Figure 1 shows a phasegram of the voiced speech segment of five sustained vowels [aeiou], where the stable pattern of every vowel can be clearly distinguished. These patterns led us to think that RPS information could be useful to improve ASR recognition rate.

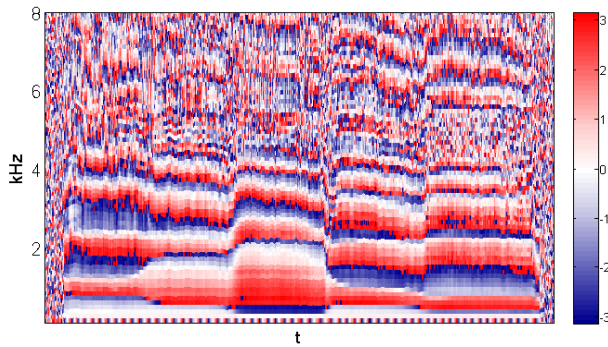


Figure 1: RPS phasegram of a voiced speech signal /aeiou/.

### 3. Parameterization of the RPS data

Despite the defined and clear looking of the RPS patterns, they cannot be directly employed as parameters in a Hidden Markov Model, HMM, based ASR system. There are several points that need to be addressed, and they are discussed next in this section.

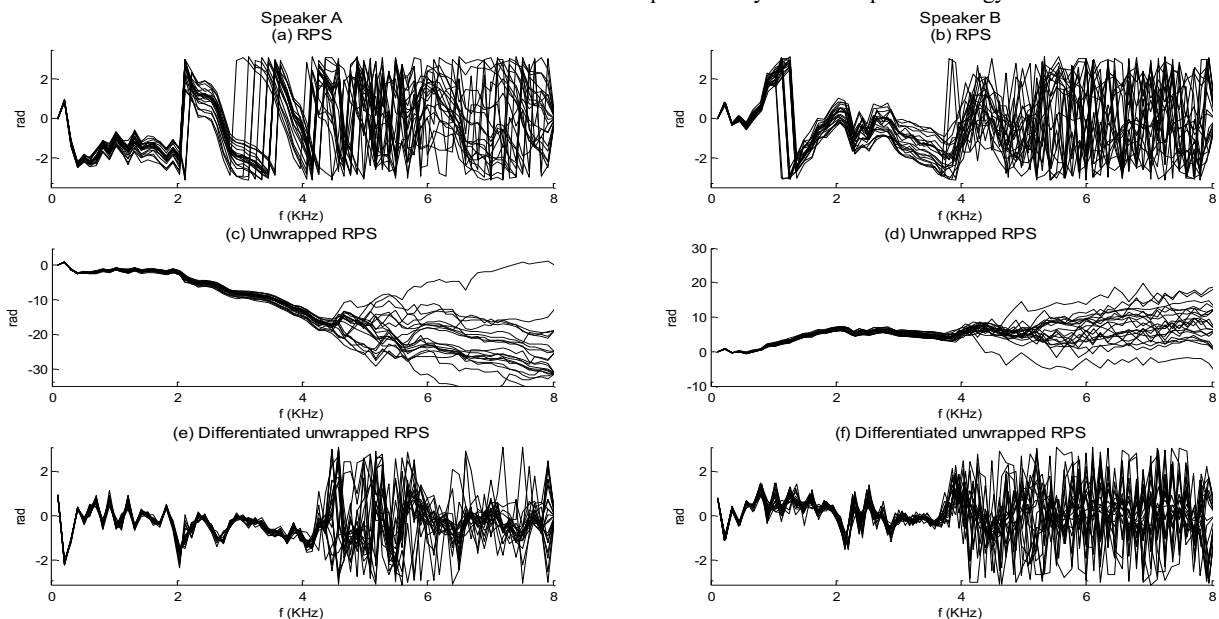


Figure 2: RPS information for two sustained |i| speech segments of 200ms (20 frames) by two male speakers: (a-b) RPS, (c-d) unwrapped RPS, (e-f) differentiation of the unwrapped RPS.

### 3.1. Unvoiced speech frames

There is no meaningful RPS for unvoiced speech. In the unvoiced segments of the speech there is no valid reference ( $F_0$ ) to base the RPS calculations. Furthermore, in unvoiced segments the excitation signal is supposed to be a random signal, whose phases will be random, and this will mask the phase structure of the vocal tract.

This lack of meaningful information in the unvoiced segments is not only a problem of the RPS transformation, but also an inherent problem of the random phase signals which hinder the effect of the phase response of the filter.

The workaround has been to define an arbitrary reference  $F_0$  (100Hz) for the unvoiced segments and to obtain the RPSs from the instantaneous phases at the multiples of this  $F_0$ . This is a convenient solution because it allows homogeneous treatment of the voiced and unvoiced segments, using both of them in the ASR system, but it does not mean that the phase of the unvoiced segments has any meaning. In fact it could be assumed that the unvoiced parts of the speech are excluded from the scope of the technique described in the paper.

### 3.2. Parameter number and dimensionality reduction

The number of RPS values varies from frame to frame as it is dependent on the number of pitch harmonic components that fit in the analyzed spectral bandwidth, which, obviously, varies with the pitch value for the frame. For usual pitch values, the number of values is too high, forcing high dimensionality models if used directly.

Two approaches have been tested to cope with this situation. First we tried resampling of the unwrapped RPS ‘‘envelopes’’ (Fig. 2c-d) at fixed frequencies, in order to get not only a fixed number of parameters, but also at fixed and homogenous frequencies so that they could be correctly modelled. This resampling was done by means of a linear interpolation of the linear frequency envelope at the desired frequencies. The other approach was to use a normalized Mel filter bank to filter the RPS data. This operation produces a fixed number of outputs (corresponding to the number of filters) and introduces perceptual information that has been proved very useful in spectral energy based ASR.

### 3.3. RPS envelope features

In order to parameterize the RPS data by frame it is necessary to analyze the usual shape of the RPS function across the frequencies in different segments of different speakers.

First the problem of wrapping arises: the RPS values are wrapped values given in the range  $[-\pi, \pi]$ . While this wrapping does not greatly affect the temporal smoothness of the RPS values, it produces discontinuities in the frequency axis representation of a RPS frame, as it can be observed in Fig. 2.a and 2.b. Unwrapping is thus necessary to get a smooth function that can be parameterized by a reduced number of values. But unwrapping is an ambiguous operation that can produce very different results for similar data. An example is shown in Fig 2.c and 2.d where the unwrapped RPSs for two different speakers are significantly different (note the vertical scale change in both graphics).

The unwrapped RPS values usually evolve smoothly in frequency with larger negative phase jumps nearby the poles of the vocal tract, i.e. the formants. This behaviour suggests the use of differentiated unwrapped RPSs as source information for the parameterization as it shows these phase jumps in a more prominent way. This is shown in Figures 2.e and 2.f. where the apparent divergences of the unwrapped RPS of both speakers disappear and common features can be appreciated. This is clearer in Figure 3, which displays the mean of the overlapped frames of differentiated unwrapped RPS, where the major negative peaks at the frequencies of formants can be clearly seen.

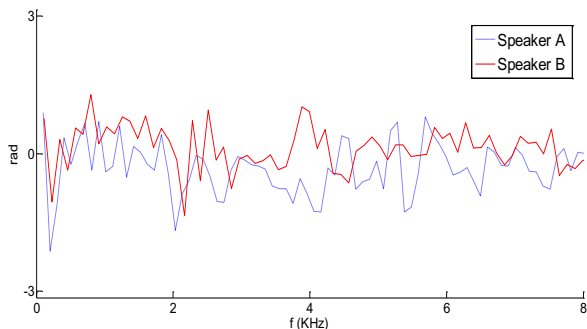


Figure 3: Mean of the differentiated unwrapped RPS for Speakers A and B.

The Discrete Cosine Transform has been frequently employed in the literature to parameterize both magnitude and phase data to be used in ASR models. This transformation successfully reduces the number of parameters needed to model the speech. We also use the DCT even though the spiky shape of the differentiated unwrapped RPS worsens the DCT modelling capability.

### 3.4. Mean normalization

The analysis of RPS information of different speakers shows that the overall unwrapped envelope (Fig. 2.c-d) has a slope whose tilt seems to change from speaker to speaker, or even from utterance to utterance. In any case, in the differentiated RPSs this slope appears as an offset (see Fig. 3) which can be removed by means of a mean normalization.

### 3.5. Signal polarity

It is necessary to notice that the RPS information is sensitive to the signal polarity (i.e. the signal can be up-down inverted

or not), so uniform polarity has to be assured in the training and testing signals. A polarity detection algorithm [10] which uses the RPS information to calculate the polarity of a signal has been used to correct the polarity of every signal before the extraction of the phase information.

## 4. Performance Evaluation

To check the effect of including RPS information in an ASR system, we have devised a comparative experiment consisting on an isolated word recognition task of a limited domain dictionary. Recognition results with different parameterization techniques for the RPSs are compared against the standard MFCC based baseline results to check out the relative improvement. As the recognition task only implies isolated words, the need of a grammar model is avoided.

The experiments have been carried out using the Cambridge Hidden Markov model (HMM) Toolkit HTK [11]. We have also used HTK to calculate the MFCC features. The RPS derived parameters have then been merged with these MFCCs to obtain the full parameter files used for training and testing.

### 4.1. Training and evaluation databases

We have used a Basque language speech database to test the RPS influence on ASR. It is a multi-speaker database (230 speakers, 127 female and 103 male) recorded in home and office environments. The signals are recorded at 16 kHz and 16 bits.

This database has been divided into two parts for training and testing. The training section includes 43,953 utterances. The testing section consists of 11,859 utterances with a dictionary of 435 words.

### 4.2. Experiments

Different sets of parameters have been tried to check the performance of every arrangement. Through all the experiments the training process starts from the audio files and the phonetic transcription. The training process creates context dependent triphone models with 3 emitting left-right no skips topology and one Gaussian by state. We have not used more complex models because the recognition rate is high enough with this setting and, in any case, the aim is not to obtain the best possible absolute rate but to compare the rates with and without the phase information, thus seeking the highest relative improvement.

The baseline recognition rate has been obtained using 13 MFCC coefficients (including MFCC0), calculated by the HTK toolkit. We have merged these MFCC parameters with different parameters derived from the RPS. The first half of Table 1 shows the results of these experiments:

- In MFCC+RPS we added to the 13 MFCC, the 20 first values obtained from DCT transformation of the frequency interpolated and unwrapped RPS envelope.
- In MFCC+  $\Delta$ RPS the parameter set was composed of the 13 MFCC coefficients and the 20 first values from the DCT transformation of the mean normalized difference in the frequency axis of the frequency interpolated and unwrapped RPS envelope.
- In MFCC +  $\Delta$ RPS<sub>mel</sub> the parameter set was composed of the 13 MFCC coefficients and the 20 first values from the DCT transformation of the Mel filtered (32 filters) mean normalized difference in the frequency axis of the original unwrapped RPS envelope.

The baseline system (shaded) has quite a high recognition score (86.12% or 13.88% WER), thus the WER reduction has to be judged by its relative weight compared to the error rate.

	Recognition Rate	WER	WER reduction
MFCC	86.12	13.88	0.00
MFCC + RPS	87.48	12.52	9.80
MFCC + $\Delta$ RPS	86.6	13.4	3.46
MFCC + $\Delta$ RPS mel	89.27	10.73	<b>22.69</b>
MFCC+ $\Delta$ + $\Delta^2$	96.1	3.9	0.00
MFCC+ $\Delta$ + $\Delta^2$ +RPS	96.26	3.74	4.10
MFCC+ $\Delta$ + $\Delta^2$ + $\Delta$ RPS	96.19	3.81	2.31
MFCC+ $\Delta$ + $\Delta^2$ + $\Delta$ RPS mel	96.58	3.42	<b>12.31</b>

Table 1. Results of the experiments (%)

The results show that the phase information can improve the recognition rate of the system. The improvement rate suggests that the additional information added by the phase is partially already present in the MFCCs. This is coherent with the fact that the major phase changes are located at the formant positions, as it happens with the energy part of the spectrum.

Regarding the different parameterization strategies the best results are produced by the Mel-filtered, mean normalized, frequency-differenced original RPS envelope, were a 3,15 point improvement in the WER (a reduction of the 22%) is obtained (this result proved to be statistically significant, with  $p < 0.05$ ). This is also consistent with the ASR results with MFCC, where the perceptual filtering has proven its usefulness in these tasks. This result also confirms the initial hypothesis of the convenience of doing mean normalization in the differentiated RPS to eliminate some of the speaker and channel dependent information.

The crucial role of the DCT transform to get a compact and suitable parameter set for the HMM modelling has to be remarked, even if the data does not look appropriate for the DCT. Some other experiments using none or other transformations have shown a poorer performance.

Additional tests have been executed to check the influence of the RPS derived information in a more realistic setup, where not only static MFCCs were employed, but also dynamic ones: deltas and accelerations of the MFCCs. In this case the baseline system (second half of Table 1, shaded MFCC+ $\Delta$ + $\Delta^2$ ) has a very high recognition rate, 96.1%. The improvements obtained including the RPS information are smaller than in the previous set of experiments. This suggests that the dynamic information of the MFCCs overlaps with some of the information previously provided by the RPSs.

The results of these additional tests are consistent with those of the previous experiments, with the different parameterizations performing comparatively in the same way. The best results are again those corresponding to the Mel filtered difference in the frequency axis of the original unwrapped RPS envelope, with a 12% reduction in the WER (also significant with  $p < 0.05$ ).

Finally, we have also tested the performance of the RPS parameters by themselves in the ASR task. We have just added the first of the MFCC coefficients, the MFCC0, to the  $\Delta$ RPSmel parameter set, because phase information is not sensitive at all to the energy of the signal. This energy information is absolutely necessary at the beginning of the training process to correctly segment the signals, distinguishing between silence and speech parts. The result has been a recognition rate of 67.1 %, which can be

considered a good result taking into account that the RPSs carry no information for the unvoiced sounds.

## 5. Conclusions

The main goal of this work has been the study the possible use of the RPS information in an ASR task, and evaluate its potential to improve the recognition results over the magnitude based standard methods. To do so, a suitable parameterization has been developed and different recognition experiments have been done. The results show significant improvements in the WER (reductions from 12.31 to 22.69%). Although no direct comparison with other published results is possible, the RPS seems to perform better than other proposed phase representations (improvements from 1.47% to 6.32% in comparable experiments by Alsteris and Paliwal [1] with MODGDF or 10.53% in Bozkurt with CGDZP [4]).

Having found an appropriate parameterization for the RPS data, future research will try to extensively evaluate its usefulness, with different databases, acoustic conditions and languages and with more complex ASR tasks.

## 6. Acknowledgements

The work presented in this paper has been partially funded by the Spanish Government under grant TEC2009-14094-C04-02 (BUCEADOR project) and by the Basque Government under grant IE09-262 (BERBATEK project).

## 7. References

- [1] Alsteris, L. and Paliwal, K., "Short-time phase spectrum in speech processing: A review and some experimental results", *Digital Signal Processing*, 17:578-616, 2007.
- [2] Murthy, H. and Gadde, V., "The modified group delay function and its application to phoneme recognition", 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2-5, 2003.
- [3] Hegde, R., Murthy H. and Ramana Rao, G., "Speech processing using joint features derived from the modified group delay function", 2005 IEEE International Conference on Acoustics, Speech and Signal Processing, 541-544, 2005.
- [4] Bozkurt, B. and Couvreur, L., "On the use of phase information for speech recognition," *Proc. EUSIPCO*, 2-5, 2005.
- [5] Thiruvanan, T., Ambikairajah, E. and Epps, J., "Group delay features for speaker recognition", 6th International Conference on Information, Communications & Signal Processing, 2: 1-5, 2007.
- [6] Hegde, R., Murthy, H. and Rao, G., "Application of the modified group delay function to speaker identification and discrimination," 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 517-520, 2004.
- [7] Wang, L., Minami, K., Yamamoto, K. and Nakagawa, S., "Speaker identification by combining MFCC and phase information in noisy environments," 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, 4502-4505, 2010.
- [8] Lee, K., Kang, S., Song, J. and Chang, J., "Group delay function for improved gender identification," *Proc. Interspeech* 2008, 1513-1516, 2008.
- [9] Saratxaga, I., Hernandez, I., Erro, D., Navas, E. and Sanchez, J., "Simple representation of signal phase for harmonic speech models", *Electronics Letters*, 45: 381-383, 2009.
- [10] Saratxaga, I., Erro, D., Hernandez, I., Sainz, I. and Navas, E., "Use of harmonic phase information for polarity detection in speech signals," *Proc. Interspeech* 2009, 1075-1078, 2009.
- [11] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V. and Woodland, P. "The HTK book", Cambridge University Engineering Department, 2009.