Simple representation of signal phase for harmonic speech models

I. Saratxaga, I. Hernáez, D. Erro, E. Navas and J. Sánchez

A novel representation of the phase information in harmonic speech models is proposed. A transformation from instantaneous phases to initial phase shift differences with respect to the fundamental frequency provides a clear insight into the structure of the phase information and largely simplifies the manipulation of this information.

Introduction: Harmonic analysis considers the speech signal as a sum of a periodic component and a non-periodic one, both components are modelled separately [1]. While several representations are used for the non-periodic part, the periodic part is always modelled by means of a sum of sinusoids at the harmonic frequencies of the fundamental frequency or pitch, weighted by certain amplitudes and shifted by certain phases.

In these models, harmonic amplitudes have a clear and distinct meaning for they reflect the spectral energy and, thus, capture basic perceptual information of the signal. However, phase information has been traditionally considered of little importance from a perceptual point of view. It is also intrinsically difficult to be modelled and represented. This is why this information has often been disregarded by some applications (e.g. speech recognition), or some speech synthesis methods (e.g. MBROLA). Nevertheless, there are a number of studies which demonstrate the importance of the phases to maintain the perceived naturalness and quality of the voice [2].

In this Letter we present a new method for transforming the instantaneous phases that are usually obtained by the harmonic analysis methods into the relative phase shifts of the harmonics against a common reference. This transformation removes the linear phase contribution (owing to the frequency of every harmonic) and allows phase structure to arise, facilitating its modelling and further manipulation.

From instantaneous phase to relative phase shift: Harmonic analysis models a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency:

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

where N is the number of bands, A_k are the amplitudes, $\phi_k(t)$ is the instantaneous phase, f_0 the pitch or fundamental frequency and θ_k is the initial phase shift of the kth sinusoid. Usually the parameters of the model are obtained by minimising the mean squared error of the estimated signal. The methods to solve this problem give the whole instantaneous phase of every sinusoid, $\phi_k(t)$, instead of the initial phase shift θ_k. This instantaneous phase change depends on the analysis instant as well as on the frequency of the harmonic, owing to the linear phase term $2\pi k f_0$. However, the initial phase shift (θ_k) is constant while the waveform shape is stable under the assumption of local stationarity, regardless of the time instant chosen for the analysis. θ_k can be regarded as the combined phase response of the glottal and vocal tracts.

It is worth noting that, if the analysis instant is at the same point of every period, the instantaneous phase values of every harmonic will be equally affected by the linear phase term. This is one of the reasons why pitch synchronous analysis is interesting. However, finding these special points is not trivial and different techniques have been applied to determine them: centre of gravity [3], pitch onset times [4], or detection of glottal closure instants.

Our approach tries to tackle the problem from another angle. Phase data reflect two features of the signal waveform shape and time synchronicity. The waveform shape depends only on the differences between the initial phase shifts of the components, which we will call relative phase shifts (RPSs). These RPSs are 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.

The second feature of the signal contained in the phase information, time synchronicity, requires knowing the instantaneous phase, because it anchors the signal to the time reference. We only need to keep the instantaneous phase of the reference component, the f_0, because the instantaneous frequency of the rest of the components can be obtained from the RPSs.

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

$$x_1(t) = \cos(2\pi f_1 t + \theta_1) \qquad x_k(t) = \cos(2\pi f_k t + \theta_k) \tag{2}$$

where $x_1(t)$ is the reference sinusoid with frequency f_1 and $x_k(t)$ another sinusoid with frequency $f_k > f_1$. θ_k is the initial phase shift and t stands for time. For the sake of simplicity we 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:

$$\varphi_1(t_a) = 2\pi f_1 t_a \qquad \varphi_k(t_a) = 2\pi f_k t_a + \theta_k \tag{3}$$

From these two expressions we can get the RPS:

$$\theta_k = \varphi_k(t_a) - \frac{f_k}{f_1}\varphi_1(t_a) \tag{4}$$

In the case of harmonic analysis, f_1 will be the fundamental frequency (f_0), the frequencies of the two sinusoids will be harmonically related, so $f_k = kf_1$, and thus:

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

Finally, the phase difference is wrapped to values in the $[-\pi, \pi]$ interval. The obtained values of θ_k allow the reconstruction of the shape of the signal, starting at any point of its period. The instantaneous phase of the reference signal $\phi_1(t_a)$ is also kept in order to enable synchronous reconstruction of the original signal.

Signal reconstruction: With the RPS representation, the reconstruction of the harmonic part of the signal is simple and flexible. Its expression for the ith frame is:

$$\hat{h}^{(i)}(t) = \sum_{k=1}^{N} A_k^{(i)}(t) \cos\left(\varphi_k^{(i)}(t)\right) \tag{6}$$

where $A_k^{(i)}(t)$ are the time evolving coefficients (linearly interpolated from frame i to frame $i+1$) and $\phi_k^{(i)}(t)$ stands for the instantaneous phase.

Owing to the harmonic relationship between the components the reference sinusoid can be generated first and its instantaneous phase can be used as a reference for the rest of the harmonic sinusoids. This assures phase continuity and the desired interharmonic phase difference. The control of this reference, the fundamental frequency, is the keystone of the process, as the rest of the instantaneous phases of all the harmonics rely on it. The instantaneous phase for this fundamental frequency can be written as

$$\varphi_1^{(i)}(t) = 2\pi \int_0^t f_1^{(i)}(\tau) d\tau + \varphi_1^{(i)}(0) + t\Delta\theta_1 \tag{7}$$

where $\phi_1^{(i)}(0)$ is the initial phase of the signal at the beginning of the frame, $f_1^{(i)}(t)$ is the time dependent fundamental frequency and $\Delta\theta_1$ is a correction term which allows synchronisation with the instantaneous phases of the original signal. It is calculated as:

$$\Delta\theta_1 = \frac{1}{T}\left[\varphi_1^{(i+1)} - \varphi_1^{(i)} - 2\pi \int_0^T f_1^{(i)}(\tau) d\tau\right] \tag{8}$$

where $\phi_1^{(i)}$ and $\phi_1^{(i+1)}$ are the instantaneous phases of the fundamental frequency of the ith and $(i+1)$th frames, and T is the length of one frame.

This $\Delta\theta_1$ term is useful in order to get a fully synchronised signal for copy synthesis. It can be avoided with the only effect being that the reconstructed waveform will be locally out of phase with the original one, but the overall waveform shape will still be kept.

For the rest of the harmonics getting the instantaneous phase at instant t is straightforward:

$$\varphi_k^{(i)}(t) = \varphi_1^{(i)}(t)k + t\Delta\theta_k \qquad 1 < k \leq N \tag{9}$$

where $\Delta\theta_k$ is the RPS correction term. As it evolves smoothly in time, it is simply calculated by linear interpolation between the RPSs at the beginning and end of the frame ($\theta_k^{(i)}$, $\theta_k^{(i+1)}$):

$$\Delta\theta_k = \frac{1}{T}\left(\theta_k^{(i+1)} - \theta_k^{(i)}\right) \qquad 1 < k \leq N \tag{10}$$

Discussion: The RPS transformation has notable advantages compared with the instantaneous phases. Its major feature is that it reveals a meaningful and structured pattern in the phase information of the voiced segments, which remains hidden when using instantaneous phases. Both magnitudes can be represented by means of a 'phasegram' which, as its counterpart the spectrogram, shows the evolution with time of the phase information for each frequency. The phasegrams for a voiced speech segment of two sustained vowels |ea| are shown in Fig. 1.

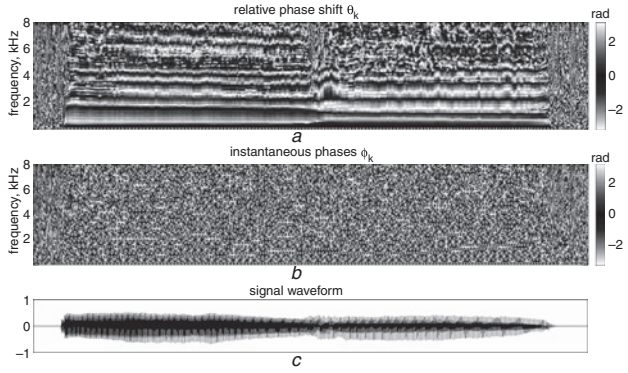

Fig. 1 *Phasegrams of voiced speech segment /ea/ sampled at 16 kHz*
a Relative phase shift (θ_k)
b Instantaneous phases (ϕ_k)
c Signal waveform

As can be seen in Fig. 1, RPSs are stable along the time axis and are distinct for both sounds. Preliminary studies on vowels show evidence that some features of the phase difference patterns remain unchanged from speaker to speaker. This information could be used in many areas such as ASR, speech segmentation and voiced-unvoiced detection. Furthermore, RPSs evolve smoothly across the frequency axis. Fig. 2 shows the unwrapped phase differences for several adjacent frames pertaining to the |e| sound. This smooth evolution both in time and in frequency allows the use of linear interpolation of the initial RPS from frame to frame to reconstruct the signal. Also, linear interpolation of phases can be done in the frequency axis. This avoids the use of complex envelope estimation methods when performing pitch modification on the original signal.

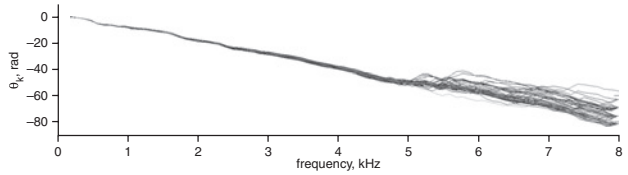

Fig. 2 *Unwrapped phase shift of consecutive frames of /e/ sound*

The use of RPSs also makes it trivial to perform the phase adjustments for shape invariant time and pitch scale modifications. When using instantaneous phases it is necessary to calculate the new resulting phases when changing either the length or the pitch of the original signal [4, 5]. Using the RPS, the fundamental frequency or the length of the frame can be changed without worrying about phases, because the waveform will stretch or shrink unchanged for the required length as long as the RPS is kept constant. In fact, the transformation separates the phase difference information that determines the waveform shape from the instantaneous phase of the reference harmonic which anchors the waveform to a definite time instant, allowing independent control of the waveform shape and its time-evolution.

Conclusions: The proposed RPS transformation is simple and flexible. It can be applied independently of the phase determination technique used (error minimisation, spectral phases), effectively allows pitch asynchronous analysis and its information can be used for several synthesis methods. It implies no loss of information with respect to the usual representation of instantaneous phases and clarifies phase structure, making phase manipulation and modelling easier.

Acknowledgments: This work was partially supported by the Avivavoz project, MEC (TEC2006-13694-C03-02) and the ANHITZ program of the Basque Government (IE06-185).

12 December 2008
doi: 10.1049/el.2009.3328

I. Saratxaga, I. Hernáez, D. Erro, E. Navas and J. Sánchez (*Aholab, Signal Processing Laboratory, Department of Electronics and Telecommunications, High Technical School of Engineering, University of the Basque Country, Urkijo zum., Bilbao 48498, Spain*)

E-mail: ibon.saratxaga@ehu.es

References

1 Laroche, J., Stylianou, Y., and Moulines, E.: 'HNM: a simple, efficient harmonic + noise model for speech'. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, USA, 1993, pp. 169–172
2 Paliwal, K.K., and Alsteris, L.: 'Usefulness of Phase Spectrum in Human Speech Perception'. Proc. Eurospeech, Geneva, Switzerland, 2003, pp. 2117–2120
3 Stylianou, Y.: 'Removing linear phase mismatches in concatenative speech synthesis', *IEEE Trans. Speech Audio Process.*, 2001, **9**, (3), pp. 232–239
4 Quatieri, T.F., and Mcaulay, R.J.: 'Shape invariant time-scale and pitch modification of speech', *IEEE Trans. Signal Process.*, 1992, **40**, (7), pp. 497–510
5 Erro, D., Moreno, A., and Bonafonte, A.: 'Flexible harmonic/stochastic speech synthesis'. 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, 2007, pp. 194–199