A HOLISTIC GLOTTAL PHASE-RELATED FEATURE

Aníbal J. Ferreira *

Department of Electrical and Computers Engineering
Faculty of Engineering - University of Porto
Porto, Portugal
ajf@fe.up.pt

José M. Tribolet

Department of Computer Science and Engineering
Instituto Superior Técnico - University of Lisbon
Lisbon, Portugal
jose.tribolet@inesc.pt

ABSTRACT

This paper addresses a phase-related feature that is time-shift invariant, and that expresses the relative phases of all harmonics with respect to that of the fundamental frequency. We identify the feature as Normalized Relative Delay (NRD) and we show that it is particularly useful to describe the holistic phase properties of voiced sounds produced by a human speaker, notably vowel sounds. We illustrate the NRD feature with real data that is obtained from five sustained vowels uttered by 20 female speakers and 17 male speakers. It is shown that not only NRD coefficients carry idiosyncratic information, but also their estimation is quite stable and robust for all harmonics encompassing, for most vowels, at least the first four formant frequencies. The average NRD model that is estimated using data pertaining to all speakers in our database is compared to that of the idealized Liljencrants-Fant (L-F) and Rosenberg glottal models. We also present results on the phase effects of linear-phase FIR and IIR vocal tract filter models when a plausible source excitation is used that corresponds to the derivative of the L-F glottal flow model. These results suggest that the shape of NRD feature vectors is mainly determined by the glottal pulse and only marginally affected by either the group delay of the vocal tract filter model, or by the acoustic coupling between glottis and vocal tract structures.

1. INTRODUCTION

DFT-based phase processing of speech and musical sounds has been addressed since the birth of signal processing, early in the 60s of the 20th century. As a strong motivation, the theory of Fourier analysis of continuous and discrete signals was already well established, in particular concerning periodic signals, whose spectrum consists of a harmonic structure of sinusoidal components. However, owing to i) the discrete nature of the DFT and its underlying circular properties, ii) the specificity of popular and practical optimization metrics which emphasize quadratic measures, and iii) to a belief that to a considerable extent the 'human ear is insensitive to phase', phase processing in DFT analysis has not received as much attention as magnitude-based processing. A clear evidence of this reality is given by the simple fact that most front-ends for speech recognition and even speaker identification rely on the extraction of acoustic features that are based on spectral magnitude information only. Another reason explaining this reality involves the meaning of phase, especially the meaning in a psychoacoustic sense. Here again, the psychoacoustic meaning that is associated with the spectral magnitude is quite obvious and appealing: for example, it helps to explain pitch (i.e. the fundamental frequency), timbre, dark sounds (low-pass signals) and bright sounds (high-pass signals). On the contrary, phase was never associated with such a clear psychoacoustic interpretation and, in a large number of signal processing applications, such as spectral subtraction in noise reduction, phase is either ignored or simply discarded.

In this paper, we provide an illustrated motivation to the importance of phase as a relevant holistic feature for locally periodic signals, and we focus on its importance to characterize the periodic component of the glottal excitation. Although an in-depth treatment will be addressed in a forthcoming paper, here we use both synthetic and natural voice signals, notably vowel sounds, in order to illustrate holistic phase patterns that reflect idiosyncratic traits due mainly to the periodic glottal source, to illustrate the human diversity in vocal fold operation, and to evaluate how close popular models of the glottal pulse are to practical results.

In this section, we will briefly mention how phase has been looked at and acted upon notably in such areas as speech coding [1, 2] and time-scale modification of speech [3, 4].

Work in speech coding, during the 60e and 70s of the 20th century, especially in the area of frequency-domain coding of speech, has regarded phase as a frequency-domain parameter that could be quantized and coded or replaced by a synthetic phase, on a DFT coefficient basis. With the help of real transforms, such as the Discrete Cosine Transform (DCT), phase was even avoided -at least explicitly- and the focus was rather concentrated on adaptive quantization schemes defining how coarsely or finely the DCT coefficients should be quantized such as to minimize an objective distortion, or such as to minimize the perceptual impact of the quantization and coding noise. Later on, in the 70s, 80s, and 90s, these same principles were applied to wideband speech and high-quality audio coding. In this context, explicit phase-based processing was also avoided by using the Modified DCT [5].

An important class of speech algorithms dealing directly with the DFT phase representation involve time-scale and pitch modification of voiced regions in speech [6, 7]. Although first methods were oriented to phase processing on a DFT coefficient by coefficient basis, the associated subjective quality was considered poor as it was characterized by signal smearing, reverberation and 'phasiness' [8]. Techniques addressing this problem implemented phase modification while preserving certain phase relationships among neighboring DFT channels (or bins) in the region of a local maximum in the magnitude spectrum, a technique known as 'phase locking' [8, 9]. Another category of phase modification involved the harmonics of a periodic waveform. The goal was to preserve the local shape of the waveform even when its duration is artificially modified while preserving the fundamental frequency, or when its fundamental frequency is modified while pre-

* This work was financed by FEDER - Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020-Operacional Programme for Competitiveness and Internationalization (POCI), and by Portuguese funds through FCT-Fundação para a Ciência e a Tecnologia in the framework of the project POCI-01-0145-FEDER-029308.

serving the duration. To a significant extent, shape-invariance was implemented in order to avoid the typical poor subjective quality of vocoders and other frequency-domain methods that focused on magnitude modification in the Fourier domain. Phase processing tried as much as possible to preserve the local phase relationships among harmonic frequencies, especially near pitch pulse onset times, because these instants were believed to represent the time 'at which sine waves add coherently' [7, 497], i.e. when they are presumed to be in phase. To our knowledge, this assumption was never really demonstrated and in fact chances are that at pitch pulse onset times the different harmonic frequencies are combined with the same phase relationship, but not necessarily in phase. Furthermore, these methods also depended on robust phase-unwrapping algorithms [10], not only to estimate pitch, but also to create extended phase models allowing to modify the time and frequency scales of a periodic waveform.

With exception of a few works including Di Federico [11] and Saratxaga [12] that we will address in the next section, those phase locking rules, as well as the shape-invariant harmonic phase modification criteria, were not framed as an interpretable holistic phase-related feature, or model, that is amenable to statistical analysis, modification and re-synthesis.

The same remark can also be made regarding the use of phase-related information in speaker recognition. Attempts have been made to include phase directly extracted from a DFT analysis of the speech signal [13], or by first processing it such as to compute a Group Delay Function (GDF) [14]. However, even in this case, phase has been looked at as an additional signal feature conveying information that complements that already provided by classic Mel-Frequency Cepstral Coefficients (MFCCs) [15], and that authors believed to be linked to the glottal source excitation. Yet, a psychophysical meaning was not attached to those phase-related features. In addition, it is quite intriguing that phonetic-oriented segmentation is typically not used to govern phase estimation in this context, which would be particularly meaningful in voiced regions of the speech.

In this paper, we briefly describe and illustrate, with the help of practical examples, a holistic phase-related feature, or model, that is linked to the harmonic phases of a periodic waveform, and that is (time) shift-invariant and independent on the pitch.

The reminder of this paper is organized as follows. In Sec. 2 we explain the nature of NRD and we illustrate it with a simple practical example. In Sec. 3 we illustrate NRD estimation with real vowel sounds. In Sec. 4 we use synthetic and natural signals to characterize the influence of the vocal tract filter on the phase characteristics of the glottal excitation. Section 5 discusses NRD models that may be used to describe the periodic part of the glottal excitation of humans. Finally, Sec. 6 summarizes the main results of this paper and discusses future work.

2. A SHIFT-INVARIANT PHASE-RELATED FEATURE

The holistic phase feature we focus on in this paper emerges directly from the Fourier analysis of the harmonics of a periodic wave. A meaningful way to introduce it is by means of a simple practical example [16]. We use the well known sawtooth waveform which is synthesized using the Fourier series comprising L terms:

$$x(t) = \sum_{\ell=1}^{L} \frac{A}{\pi\ell} \sin \frac{2\pi}{T}\ell t \, , \tag{1}$$

where A represents amplitude and T represents the reciprocal of the pitch. Although the NRD coefficients can be found directly form any periodic wave, for illustration purposes we use the derivative of the sawtooth waveform which can be easily obtained as

$$\frac{d}{dt}x(t) = \sum_{\ell=1}^{L} \frac{2A}{T} \cos \frac{2\pi}{T}\ell t = \sum_{\ell=1}^{L} \frac{2A}{T} \sin\left(\frac{2\pi}{T}\ell t + \frac{\pi}{2}\right) \, . \tag{2}$$

This form is very convenient because it highlights the phase at the *sinusoidal onset* of each harmonic. Let us now split this result in a part consisting of the fundamental frequency, and another part grouping all harmonics:

$$\begin{aligned} \frac{d}{dt}x(t) &= \frac{2A}{T} \sin\left(\frac{2\pi}{T}t + \phi_0\right) + \sum_{\ell=2}^{L} \frac{2A}{T} \sin\left(\frac{2\pi}{T}\ell t + \phi_\ell\right) \\ &= \frac{2A}{T} \sin \frac{2\pi}{T}(t + t_0) + \sum_{\ell=2}^{L} \frac{2A}{T} \sin \frac{2\pi}{T}\ell\,(t + \, t_\ell) \, , \end{aligned}$$

where $t_0 = T\phi_0/(2\pi)$ and $t_\ell = T\phi_\ell/(2\pi\ell)$ represent the absolute time-shifts of the different terms of the Fourier series. If we concentrate on the second part of this development, we may conveniently introduce a relative time-shift:

$$\begin{aligned} & \sum_{\ell=2}^{L} \frac{2A}{T} \sin \frac{2\pi}{T}\ell\,(t + t_0 + (t_\ell - t_0)) \\ = & \sum_{\ell=2}^{L} \frac{2A}{T} \sin\left(\frac{2\pi}{T}\ell(t + t_0) + 2\pi\frac{(t_\ell - t_0)}{T/\ell}\right) \\ = & \sum_{\ell=2}^{L} \frac{2A}{T} \sin\left(\frac{2\pi}{T}\ell(t + t_0) + 2\pi\frac{(\phi_\ell - \ell\phi_0)}{2\pi}\right) \\ = & \sum_{\ell=2}^{L} \frac{2A}{T} \sin\left(\frac{2\pi}{T}\ell(t + t_0) + 2\pi\mathrm{NRD}_\ell\right) \, . \end{aligned} \tag{3}$$

In Eq. (3), NRD_ℓ denotes Normalized Relative Delay (NRD) and expresses a relative delay between harmonic ℓ and the fundamental, which is further normalized by the period of the harmonic [17]. Although the acronym is reminiscent of the way each NRD coefficient is computed in practice, NRDs reflect simply a normalized value in the range $[0.0,\ 1.0[$ which depends on a difference involving the phase of the harmonic and the phase of the fundamental. Thus, the number of NRD coefficients equals the number of harmonics. Other important properties of the NRD coefficients are as follows:

- as a relative phase-related feature, the NRD of the fundamental is zero by definition,
- because NRDs express phase differences, the concepts of phase wrapping and phase unwrapping also apply, in this paper unwrapped NRDs are used since this facilities modeling and understanding,
- NRDs are intrinsically time-shift invariant, and are also independent on the fundamental frequency.

Hence, NRDs express phase relationships that, in addition to the magnitude of the harmonics, explain the shape of a specific periodic waveform, and thus completely define its shape invariance.

The NRD concept has been independently introduced in [17], and has found practical application in singing voice analysis [18], glottal source modelling [19], speaker identification [16], parametric audio coding [20] and dyspohonic voice reconstruction [21]. It was recently brought to our attention that a similar concept (Relative Phase Delay) had been presented in 1998 by Di Federico [11]. Other smooth phase descriptors for harmonic signals that are similar to NRD were also proposed by Stylianou in 1996 (phase envelope [22, page 44]) and Saratxaga in 2009 (Relative Phase Shift -RPS [12]). Our NRD estimation is closer to the method proposed by Di Federico [11] (that estimates $(t_\ell - t_0)/(T/\ell)$) than that proposed by Saratxaga [12] (that estimates $\phi_\ell - \ell\phi_0$).

To complete the illustration using our example, we use the phase values in Eq. (2) to obtain $\text{NRD}_\ell = \frac{\pi/2 - \ell\pi/2}{2\pi} = \frac{1-\ell}{4}$, $\ell = 2, \ldots, L$. We have synthesized Eq. (2) using $L = 20$ harmonics and 22050 Hz sampling frequency (FS). We obtained the NRD numerical results using the algorithm described in [17] and they are represented in Fig. 1. This algorithm uses phase unwrapping and it can be seen that results are as expected. In particular, for $\ell = 20$, the NRD becomes -4.75. This figure also represents the experimental results regarding the waveform described by Eq. (1), in which case all NRDs are clearly zero. We conclude the illustration of the NRD concept using another synthetic signal alternative. Taking the negative of Eq. (2), we obtain

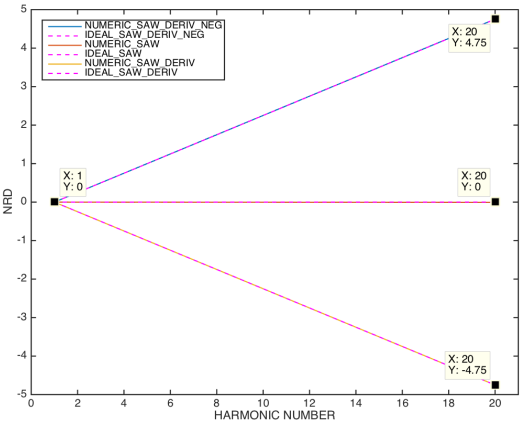

Figure 1: *Unwrapped NRD estimation results for the sawtooth wave, its derivative and its negative derivative. Ideal (analytical) and experimental results are overlapped.*

$$-\frac{d}{dt}x(t) = -\sum_{\ell=1}^{L} \frac{2A}{T} \cos\frac{2\pi}{T}\ell = \sum_{\ell=1}^{L} \frac{2A}{T} \sin\left(\frac{2\pi}{T}\ell t - \frac{\pi}{2}\right), \tag{4}$$

which highlights that the phases at the sinusoidal onset of all harmonics, are all equal to $-\pi/2$. It follows that $\text{NRD}_\ell = \frac{-\pi/2 + \ell\pi/2}{2\pi}$, or $\text{NRD}_\ell = \frac{\ell - 1}{4}$, $\ell = 2, \ldots, L$. In particular, for $\ell = 20$, the NRD becomes 4.75. This result is also illustrated in Fig. 1. The experimental results are also shown and it can be seen that the agreement is clear.

Although the NRD concept is a simple one to grasp, the actual computation, or estimation, is less trivial. The major difficulty is that the phases at sinusoidal onsets are not readily available from the DFT or similar transform. What is available is phase information that is referred to a time instant (or sample) corresponding to the delay of the DFT filter bank, and which also depends on the influence of the time analysis window prior to DFT transformation. Thus, this influence must first be compensated for, then phase information (ϕ_ℓ) is converted into time delays (n_ℓ) which are made relative to the time delay of the fundamental (n_0), and further wrapped using the period of each harmonic (P_ℓ). Finally, a normalization by each harmonic period is applied [17]. Fig. 2 illustrates the NRD estimation algorithm. We use the Odd-DFT [23] instead of the plain DFT due to a number of interesting properties which facilitate accurate estimation of the frequencies, phases and magnitudes of the sinusoidal components that exist in a signal. Thus, accurate frequency and phase estimation of each individual sinusoidal component [24] is very important to the reliability, accuracy and robustness of the NRD estimation algorithm.

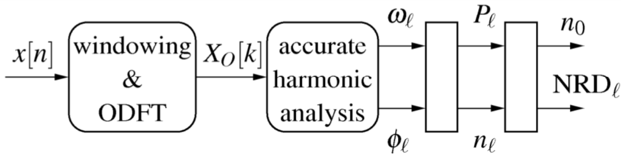

Figure 2: *NRD estimation algorithm [17].*

3. A HOLISTIC PHASE DESCRIBING VOICED SOUNDS

In this section, we present first results for a holistic phase-related feature that consists of unwrapped NRD coefficients. These coefficients are obtained from the accurate frequency analysis, as described in Sec. 2, of the spectrum of voiced vowel signals. The signals correspond to sustained vowel utterances produced by 37 subjects of which 20 are female, and 17 are male. The recordings that are included in the data base were obtained for forensic purposes, focusing on speaker identification, and are described in [25]. Figure 3 represents the magnitude spectrum of an /a/ vowel segment uttered by a female speaker (upper panel), and an overlay of all NRD vectors that are estimated in a sustained vowel region (lower panel) lasting about 1 second. The harmonic structure is signaled in the magnitude spectrum by means of vertical triangles. The dashed line in this figure represents the LPC model (order 22) of the spectral envelope defined by the peaks of all harmonics.

The overlay of NRD vectors suggest a few interesting conclusions. First, a region of consistent and stable NRD coefficients is apparent that involves the first 20 harmonics. These harmonics happen to be the strongest before the spectral valley located at around 4500 Hz. When harmonics have a very small magnitude or are close to the noise floor, then accurate frequency, phase and magnitude estimation is adversely affected in a significant way. Higher order harmonics are also more prone to estimation inaccuracies because their period is quite short, in the order of 3 speech samples or less. Since the period of each harmonic is individually estimated, accounting for some degree of inharmonicity, then shorter periods are more likely to be affected by noise or interferences and, thus, the phase estimation also becomes more unreliable. The impact in terms of unwrapped NRD estimation is a spreading of the NRD values as illustrated in Fig. 3 which may generate visually appealing patterns. However, this spreading is not problematic mainly for two reasons. First, the most important voice formant frequencies are typically accommodated by the NRD region that is stable. Secondly, and this is especially impor-

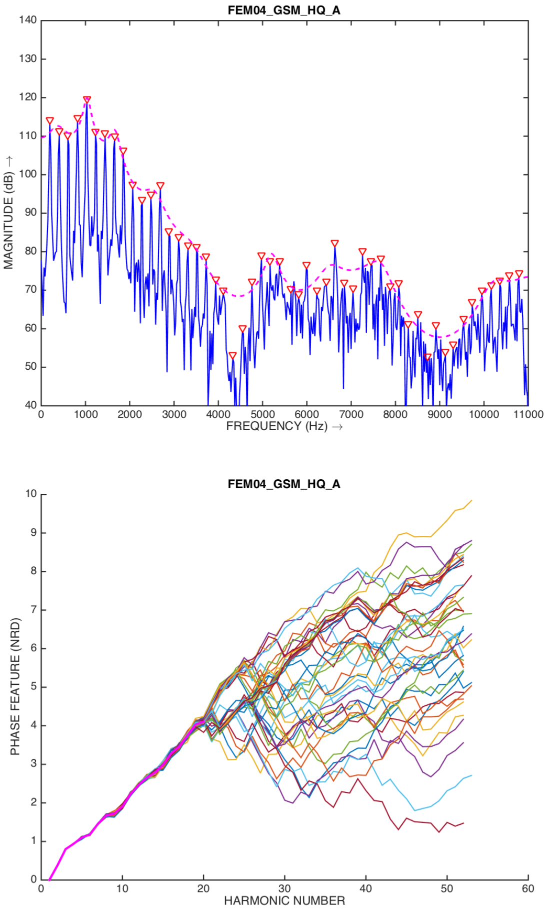

Figure 3: *Magnitude spectrum of a voiced /a/ vowel segment uttered by a female speaker (upper figure). The vertical triangles signal the harmonic structure. The lower figure represents all (unwrapped) NRD vectors found in a sustained /a/ vowel region and that includes the represented magnitude spectrum. The thick magenta line represents the average NRD vector up to harmonic 19.*

tant for synthesis purposes -which is not discussed in this paper-, the NRDs in the 'wild' region, i.e. the region where an exuberant NRD spreading can be observed, can be replaced by the new NRDs that are extrapolated from the stable NRD region.

Figure 4 represents a magnitude spectrum and a peculiar overlay of NRD vectors pertaining to a /u/ vowel uttered by a female. Since this is a back vowel whose two relevant formants have a very low frequency, then the NRD vector is stable only for the first few harmonics, five in this case. Although for other speakers, the stable NRD region may be wider even for this difficult vowel, that has no real relevance as just the first few harmonics define the vowel, both linguistically and in terms of quality.

Figure 5 illustrates the magnitude spectrum and a overlay of NRD vectors pertaining to a /o/ vowel uttered by a male. Since the pitch is about one octave lower than in the case of a female voice, the harmonic density is higher and NRD vectors may have as many as 100 coefficients within the Nyquist range. It can be confirmed in Fig. 5 that the first 4 formant frequencies are represented by the first 42 harmonics, which corresponds to the stable NRD region.

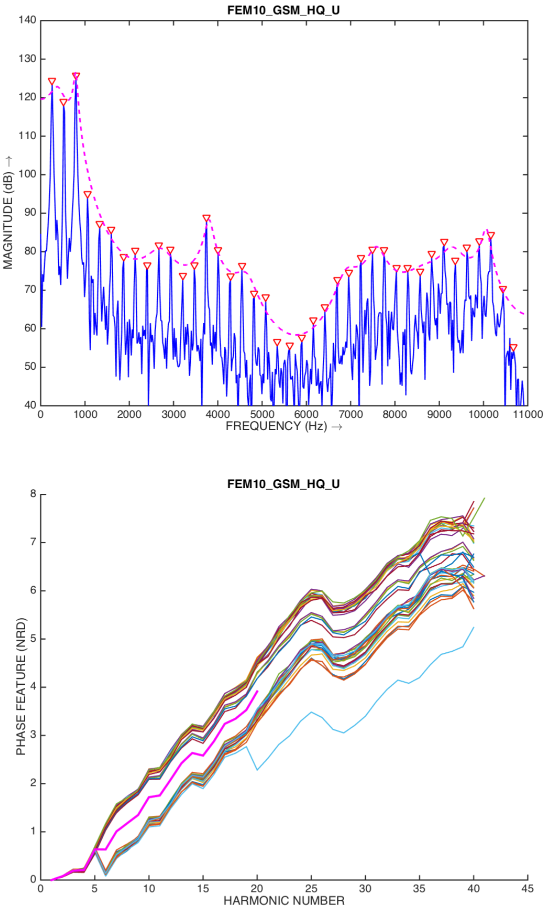

Figure 4: *Magnitude spectrum of a voiced /u/ vowel segment uttered by a female speaker (top). The vertical triangles signal the harmonic structure. The lower figure represents an overlay of all (unwrapped) NRD vectors found in the /u/ vowel region. The thick magenta line represents the average NRD vector up to partial 19.*

The above results suggest that, in most cases, it is safe to assume that the first 19 coefficients represent stable NRD vectors. Figure 6 illustrates the average NRD vectors for sustained vowel regions pertaining to five different vowels uttered by a male speaker. Results are presented for two repetitions of the same vowel exercise. It can be seen that the profile of the different average NRD vectors are in good agreement, which suggests that there is a trend that is common even for different vowels uttered by the same speaker. Rather than the vocal tract filter, which varies from vowel to vowel realization, what is really common in these situation is the glottal excitation which is mainly characterized by a periodic part due to the vibration of the vocal folds. Thus, the NRDs appear to be mainly determined by the shape of the glottal pulse. It should be noted however that for some speakers, the NRD vectors estimated from /i/ or /u/ vowel regions may deviate from the NRD trend defined by the remaining vowels. As explained above, this may be due to the fact that certain harmonics are very weak, such as in the case of the /i/ vowel which has the largest F1-F2 formant separation, or in the case of the /u/ vowel whose harmonics decay quite strongly just after the F1 and F2 formants.

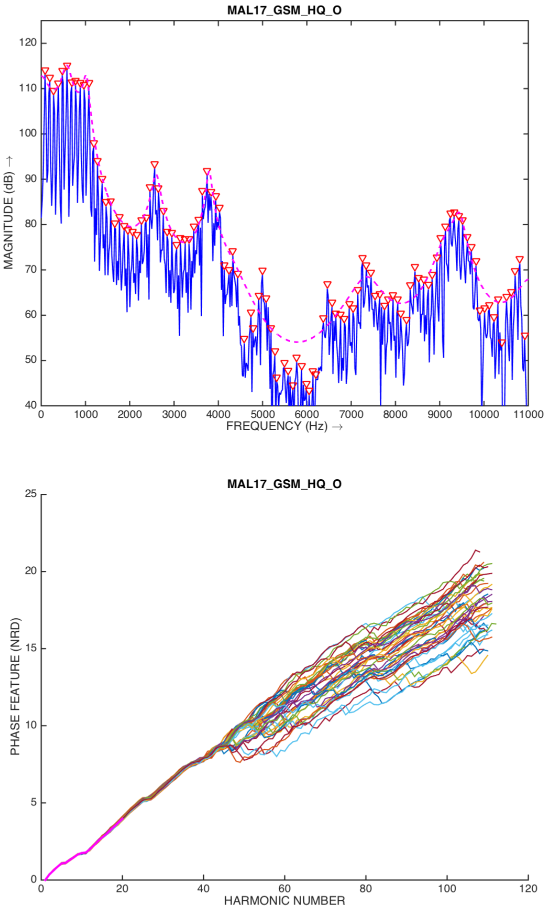

Figure 5: *Magnitude spectrum of a voiced /o/ vowel segment uttered by a male speaker (top). The vertical triangles signal the harmonic structure. The lower figure represents an overlay of all (unwrapped) NRD vectors found in the /o/ vowel region. The thick magenta line represents the average NRD vector up to partial 19.*

4. VOCAL TRACT FILTER PHASE EFFECTS USING SYNTHETIC AND NATURAL VOICED SOUNDS

According to the ideal source-filter model of voice production [26, 27], the signal generated at the glottis is the source signal and includes a stochastic and a periodic part. The supralaryngeal structures, including the oral and nasal cavities, shape the source signal in time and frequency such as to convey a desired linguistic message. This time and frequency shaping, which is mainly influenced to the vocal tract resonant frequencies -also commonly referred to as formants-, is modeled as a filter which may be considered as stationary for sustained sounds, or locally stationary in running speech considering the average syllabic duration, in the order of 10 to 20 ms. Most frequently, the filter is modeled as an all-pole filter; in our experiments, as indicated in Sec. 3, we use a 22nd-order LPC model. The filter may also include the radiation effect due mainly to the lips and nostrils. The radiation effect is usually modeled as a time differentiation operation that converts the air flow into sound pressure.

A very interesting issue that to our knowledge has never been clarified in the literature, deals with the phase contribution due to the source excitation, and that due to the filter. The combined effects are known to be additive in terms of phase or, equivalently, in terms of group delay. However, the clarification of how much the phase contribution -or group delay- due to the filter modifies the phase of the source signal is an open issue.

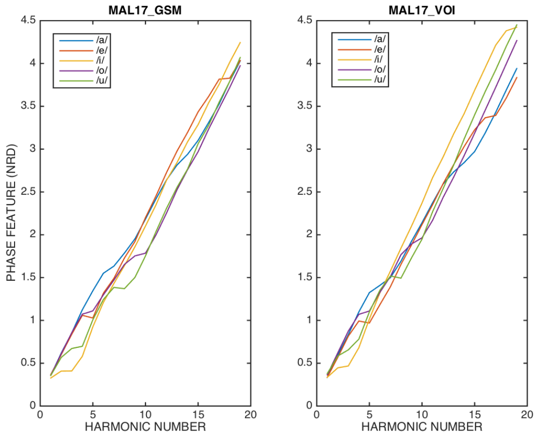

Figure 6: *Average NRD vectors (19-dimensional) obtained from sustained vowels uttered by a male speaker. Results are presented for 5 vowels produced during two different conversations.*

Using the NRD concept and using the results that were illustrated in the previous section, we may shed some light on the issue. In that regard, we will assume as a plausible periodic glottal source excitation, the derivative of the Liljencrants-Fant model (L-F) of glottal flow [28]. A 210 Hz fundamental frequency glottal excitation using the L-F model has been conveniently generated using the freely available Voicebox Matlab toolbox (FS=22050 Hz).

Figure 7 illustrates a few periods of the L-F glottal flow derivative (upper panel), the corresponding magnitude spectrum with all harmonics signaled by means of vertical triangles (middle panel), and the unwrapped NRD coefficients up to harmonic 50. This fig-

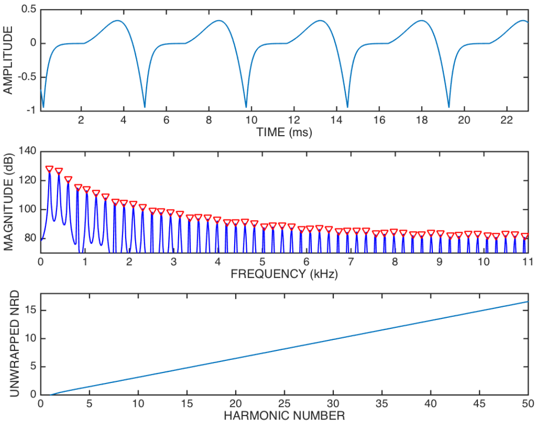

Figure 7: *Analysis of the derivative of the L-F glottal flow model. The top panel represents the time waveform and its magnitude spectrum is represented in the middle panel. The harmonics are signaled by red triangles and the unwrapped NRD coefficients pertaining to the first 50 harmonics are represented in the lower panel.*

ure suggest that the NRD feature vector may be faithfully approximated by means of a simple first order model that is given by

$$\text{NRD}_\ell = -0.207431 + 0.335465\ell, \; \ell = 2, \ldots, L. \qquad (5)$$

As indicated previously, by definition $\text{NRD}_\ell = 0, \; \ell = 1$.

Concerning the filter model, we took advantage of all the LPC models (order 22) that were obtained for all vowels from all speakers. Figures 3, 4 and 5 represent examples of the magnitude frequency responses of the IIR filters corresponding to those models. We took the average of all models separately for vowels /a/, /e/ and /i/. We considered female models only as the formant frequencies characterizing a given vowel, are typically higher in female voices than in male voices (due to anatomical differences between male and female speakers). Then, using the average power spectral density (PSD) of those models, we designed a linear-phase FIR filter (500 taps) and an IIR filter (order 22) having a magnitude frequency response approximating that PSD. The FIR filter has been obtained using a single band Parks-MccClellan optimal equiripple design. The IIR has been obtained using the Levinson-Durbin recursion and after the autocorrelation coefficients are obtained from the PSD using the Wiener-Khintchine theorem. Figure 8 represents the PSD of the average /e/ vowel, as well as the magnitude frequency responses of the FIR and IIR filters. It can be seen that both filters approximate well the PSD. An obvious (and intented) difference lies however in the phase response of both filters. In fact, the linear-phase FIR has a constant group delay response (249.5 samples) while the IIR exhibits a non linear group-delay response that is represented in Fig. 9. Assuming the L-F model as a plausible excitation to the filter, we want to assess how much the NRD coefficients at the output of the filter are affected by the group delay of the filter, according to the two alternatives: linear-phase FIR and IIR filter modeling. In other words, how much are the phase properties of the source excitation affected by the phase properties of the filter ?

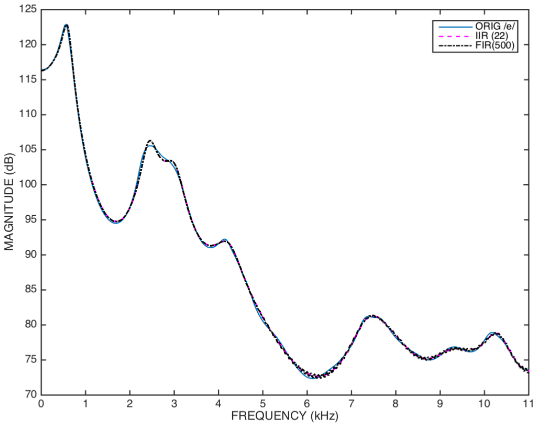

Figure 8: *Average model of the PSD of vowel /e/ uttered by female speakers, and magnitude frequency responses of a linear-phase FIR filter (order 500), and an IIR filter (order 22) approximating that PSD.*

To answer this question we filtered the source excitation illustrated in Fig. 7 using the two alternative filters and then, in each case, we extracted the NRD feature vector of the output signals.

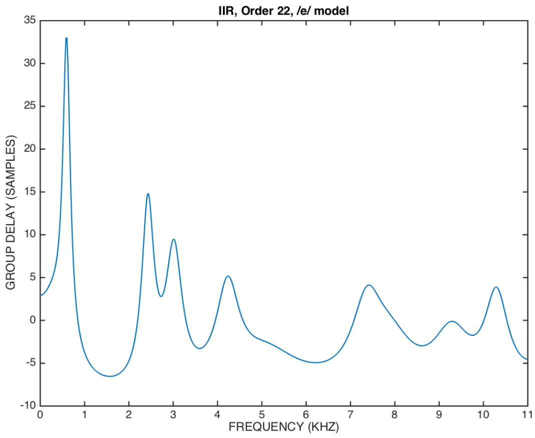

Figure 9: *Group delay of the 22nd-order IIR filter approximating the average PSD of the /e/ vowel uttered by female speakers.*

As indicated above, we repeated the experiment for vowels /a/, /e/ and /i/. In rigour, prior to this operation, we should have compensated the spectral magnitude of the excitation by its spectral tilt such that the signal at the output of the filter exhibits a PSD which corresponds to that of the original vowel PSD. Ignoring this step has however no consequences regarding phase, is just produces an output PSD which has a stronger spectral tilt than the original.

Figure 10 illustrates the NRD feature vector at the output of both filters and taking as a reference the original NRD feature of the excitation. It can be seen that, as expected, in the case of the linear-phase FIR filter, because the group delay is constant, then no modification takes place. However, in the case of the IIR filter, then visible modifications take place, although these do not represent a dramatic modification of the trend defined by the source excitation, exception for vowel /i/. In this case, a plausible explanation is that the group delay of the corresponding filter is such that it modifies significantly the NRD trend of the source excitation. Further research is required to clarify this. Considering however that this vowel represents an exception, it is interesting to compare these results that presume a synthetic excitation signal, and the results displayed in Fig. 6 that were obtained for real natural voices. In both situations, results suggest the vocal/nasal tract filter modifies the phase properties of the glottal excitation although not too strongly as the overall trend in the NRD feature vector of the source excitation is essentially preserved. We believe this is an innovative result that emerges from experimental data with NRDs. It can also be argued that the deviations to the excitation NRD feature vector, after the filter, may be due to the acoustic coupling between the glottis and the vocal/nasal tracts for different configurations of the latter and which modify slightly the shape of the glottal pulse. Clarifying this hypothesis would however imply complex and invasive experiments capturing the signal near the vocal folds.

5. A MODEL OF THE HUMAN GLOTTAL PHASE

In this section, we discuss NRD models that may be used to describe the holistic phase structure of the periodic part of the human glottal excitation.

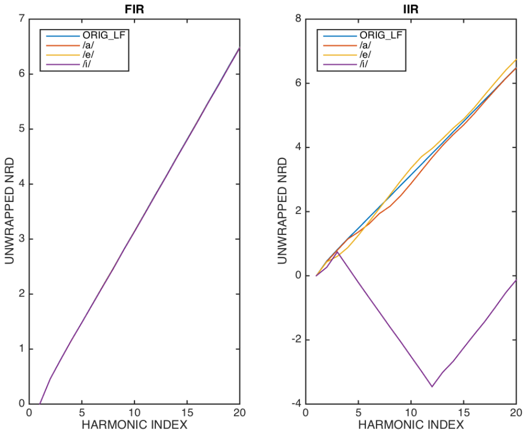

Figure 10: *Illustration of NRD modification of the source excitation due to the phase properties of the filter modeling the PSD of three vowels: /a/, /e/ and /i/. When the filter is a linear-phase FIR filter, no modification exists. When the filter is a 22nd-order IIR filter, its group delay modifies slightly the original NRD feature vector. A strong deviation is observed in the case of vowel /i/.*

Figure 11 represents an overlay of all the average NRD feature vectors that were obtained from the 5 vowel realizations by each speaker. As our data base includes 37 speakers and each speaker has produced two independent realizations for each vowel, Fig. 11 represents 74 true human average NRD data. This figure also represents the NRD feature vectors that have been obtained experimentally from synthetic signals consisting of the derivative of the ideal L-F glottal flow model, and the derivative of the Rosenberg glottal flow model. Both models were generated using the Voicebox toolbox. The L-F NRD model is well approximated by Eq. (5) and has already been illustrated in Figs. 7 and 10. Figure 11 also represents the average NRD model of all human vowel realizations, its first order best approximation is given by

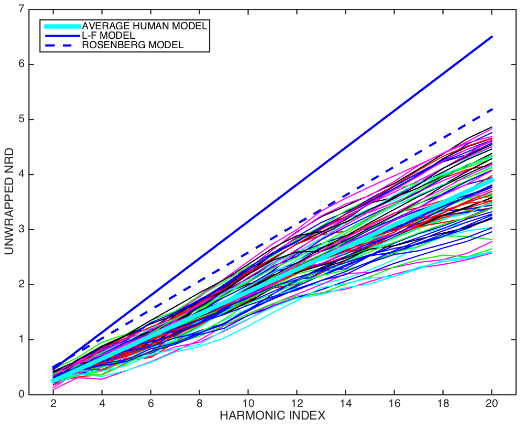

Figure 11: *Overlay of all the average NRD feature vectors for the 5 vowels uttered by each one of the 37 speakers in our data base. The experimental NRD vector of the derivative of both ideal Rosenberg and L-F glottal flow models are also represented.*

$$\mathrm{NRD}_\ell = -0.1522222 + 0.2025505\ell,\ \ell = 2, \ldots, L\,. \quad (6)$$

For the sake of completeness, the first order best approximation to the Rosenberg NRD model is given by

$$\mathrm{NRD}_\ell = -0.014001 + 0.259785\ell,\ \ell = 2, \ldots, L\,. \quad (7)$$

It can be seen that the L-F NRD model deviates more from the experimental average human NRD model than the Rosenberg model.

To conclude this section, we present a verifiable example of the capability of NRDs in representing the holistic phase properties of any periodic wave. We prepared two `.mat` Matlab files, one of them (`LFmag.mat`) contains the first 20 harmonic magnitudes of the derivative of the L-F glottal flow model, and another one (`LFNRD.mat`) contains the first 20 NRD values pertaining to the corresponding harmonics, including the fundamental. These experimental-based magnitude and NRD values are used to synthesize the derivative of the L-F glottal flow model using

$$\mathrm{dgf}(t) = \sum_{\ell=1}^{L} \mathrm{LFmag}_\ell \cdot \sin\left(\frac{2\pi}{T}\ell t + 2\pi \cdot \mathrm{LFNRD}_\ell\right)\,. \quad (8)$$

In this synthesis we use a fundamental frequency of 210 Hz and FS=22050 Hz. The resulting signal is represented in Fig. 12. We may then replace the accurate NRD coefficients LFNRD_ℓ by the approximate first-order model given by Eq. (5). The result of this approximation is also represented in Fig. 12. It can be concluded that the resulting wave is a faithful approximation to the original.

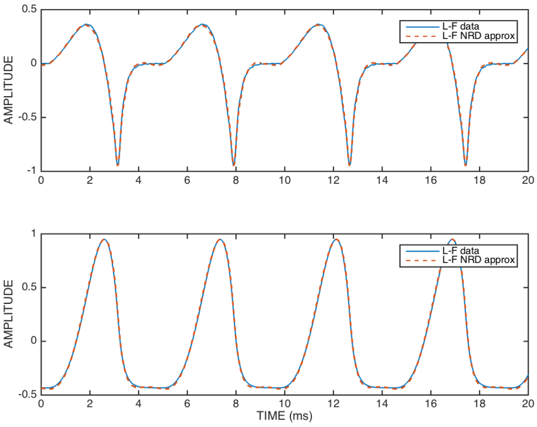

Figure 12: *L-F idealized glottal flow wave and its derivative using experimental data concerning the first 20 harmonic magnitudes and NRD coefficients. Versions of these signals are also shown that use a first-order NRD approximation. The Matlab code allowing to generate this figure is available.*

On the other hand, it is known from basic Fourier theory that if $X(j\Omega)$ is the Fourier transform of $x(t)$, then the Fourier transform of the integration of $x(t)$ is given by $X(j\Omega)/(j\Omega)$. This means that the magnitude of the Fourier transform is divided by the frequency, and $\pi/2$ is subtracted to the phase. Thus, the glottal flow model, by integrating Eq. (8) and except for a scaling factor, is simply given by

$$\mathrm{gf}(t) = \sum_{\ell=1}^{L} \frac{\mathrm{LFmag}_\ell}{\ell} \cdot \sin\left(\frac{2\pi}{T}\ell t + 2\pi \cdot \mathrm{LFNRD}_\ell - \frac{\pi}{2}\right) \ . \quad (9)$$

This result, as well as its version when LFNRD_ℓ is approximated by its first-order model are also represented in Fig. 12. In order to facilitate the reproducibility of these results, the Matlab code generating Fig. 12 is available [1].

We have shown that we know how the holistic phase of the periodic part of the human glottal excitation looks like, future research will leverage on this result to more accurately estimate the spectral magnitude of the human glottal excitation.

6. CONCLUSION

We described in this paper how the NRD phase-related feature and that is extracted from the harmonics of a periodic waveform, effectively acts as an important holistic glottal feature that carries idiosyncratic information. NRD coefficients were shown to be moderately affected by the group delay of the vocal/nasal tract filters, or by the acoustic coupling between glottis and supra-laryngeal structures. We also identified several relevant first-order NRD approximation models, one of which represents the average NRD feature of the glottal excitation of a human speaker. Future work will include further research on phase effects of the vocal tract filter, the modeling of the glottal excitation spectral magnitude, and the application of the NRD features in such areas as speaker identification, high-quality parametric speech coding and dysphonic voice reconstruction.

7. REFERENCES

[1] J. M. Tribolet and R. E. Crochiere, “Frequency domain coding of speech,” *IEEE Trans. on Acoustics, Speech and Sig. Proc.*, vol. 27, no. 5, pp. 512–530, Oct. 1979.

[2] Andreas S. Spanias, “Speech coding: A tutorial review,” *Proc. of the IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.

[3] Eric Moulines and Jean Laroche, “Non-parametric techniques for pitch-scale and time-scale modification of speech,” *Speech Com.*, vol. 16, pp. 175–205, 1995.

[4] Salim Roucos and Alexander M. Wilgus, “High quality time-scale modification of speech,” in *IEEE ICASSP*, 1985, pp. 13.6.1–13.6.4.

[5] T. Painter and A. Spanias, “Perceptual coding of digital audio,” *Proc. of the IEEE*, vol. 88, no. 4, pp. 451–513, 2000.

[6] M. R. Portnoff, “Time-scale modification of speech based on short time Fourier analysis,” *IEEE Trans. on Ac., Speech and Sig. Proc.*, vol. 29, no. 3, pp. 374–390, June 1981.

[7] T. F. Quatieri and R. J. McAulay, “Shape invariant time-scale and pitch modification of speech,” *IEEE Trans. on Signal Proc.*, vol. 40, no. 3, pp. 497–510, March 1992.

[8] Jean Laroche and Mark Dolson, “Improved phase vocoder time-scale modification of audio,” *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 3, pp. 323–331, May 1999.

[9] Aníbal J. S. Ferreira, “An Odd-DFT based approach to time-scale expansion of audio signals,” *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 4, pp. 441–453, July 1999.

[10] José M. Tribolet, “A new phase unwrapping algorithm,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 2, pp. 170–177, April 1977.

[11] Riccardo Di Federico, “Waveform preserving time stretching and pitch shifting for sinusoidal models of sound,” in *COST-G6 Digital Audio Effects Workshop*, 1998, pp. 44–48.

[12] I. Saratxaga, I Hernaez, D. Erro, E. Navas, and J. Sanchez, “Simple representation of signal phase for harmonic speech models,” *Electronic Letters*, vol. 45, no. 381, 2009.

[13] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, “Speaker identification by combining MFCC and phase information in noisy environments,” in *IEEE ICASSP*, 2010.

[14] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, “Using group delay functions from all-pole models for speaker recognition,” in *Proc. of Interspeech*, 2013.

[15] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. on Ac., Speech and Sig. Proc.*, vol. 28, no. 4, pp. 357–366, August 1980.

[16] Aníbal Ferreira, “On the possibility of speaker discrimination using a glottal pulse phase-related feature,” in *IEEE ISSPIT*, December 2014, Noida, India.

[17] Ricardo Sousa and Aníbal Ferreira, “Importance of the relative delay of glottal source harmonics,” in *39th AES Int. Conf. on Audio Forensics*, 2010, pp. 59–69.

[18] Ricardo Sousa and Aníbal Ferreira, “Singing voice analysis using relative harmonic delays,” in *Interspeech*, 2011.

[19] Sandra Dias and Aníbal Ferreira, “Glottal pulse estimation - a frequency domain approach,” in *Speech Proc. Conf.*, July 2014, Tel-Aviv, Israel.

[20] Aníbal Ferreira and Deepen Sinha, “Advances to a frequency-domain parametric coder of wideband speech,” *140th Convention of the AES*, May 2016, Paper 9509.

[21] Aníbal Ferreira, “Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information,” in *ISIVC*, 2016, pp. 159–166, Tunis, Tunisia.

[22] I. Stylianou, *Harmonic + noise models for speech, combined with statistical methods, for speech and speaker modification*, Ph.D. thesis, École Nat. Sup. Télécom., France, 1996.

[23] Maurice Bellanger, *Digital Processing of Signals*, John Willey & Sons, 1989.

[24] Aníbal Ferreira and Deepen Sinha, “Accurate and robust frequency estimation in the ODFT domain,” in *IEEE WASPAA*, Oct. 2005, pp. 203–206.

[25] Aníbal Ferreira and Vânia Fernandes, “Consistency of the F0, Jitter, Shimmer and HNR voice parameters in GSM and VOIP communication,” in *DSP 2017*, 2017.

[26] G. Fant, *Acoustic Theory of Speech Production*, The Hague, 1970.

[27] Gunnar Fant, “Glottal flow: models and interaction,” *Journal of Phonetics*, vol. 14, no. 3/4, pp. 393–399, 1986.

[28] G. Fant, J. Liljencrants, and Q. Lin, “A four-parameter model of glottal flow,” *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[1] `http://www.fe.up.pt/~ajf/DAFx18_AJF_JMT.zip`