Importance of the relative delay of glottal source harmonics

Ricardo Sousa[1], and Aníbal Ferreira[1] *

[1]*Universidade do Porto - Faculdade de Engenharia, Rua Dr. Roberto frias, s/n, 4200-465, Porto, Portugal*

Correspondence should be addressed to Aníbal Ferreira (a.j.ferreira@ieee.org)

ABSTRACT

In this paper we focus on the real-time frequency domain analysis of speech signals, and on the extraction of suitable and perceptually meaningful features that are related to the glottal source and that may pave the way for robust speaker identification and voice register classification. We take advantage of an analysis-synthesis framework derived from an audio coding algorithm in order to estimate and model the relative delays between the different harmonics reflecting the contribution of the glottal source and the group delay of the vocal tract filter. We show in this paper that this approach effectively captures the shape invariance of a periodic signal and may be suited to monitor and extract in real-time perceptually important features correlating well with specific voice registers or with a speaker unique sound signature. A first validation study is described that confirms the competitive performance of the proposed approach in the automatic classification of the breathy, normal and pressed voice phonation types.

1. INTRODUCTION

Speech is frequently modeled as a source-filter system where the source (in the case of voiced speech) consists of glottal pulses and the filter, normally an all-pole filter, represents the resonances of the vocal tract filter [1]. Both components are known to contribute to the sound signature of a speaker and to the voice register. Although it is usually acknowledged that the specific shape of the glottal pulses influences mainly the voice register [2], its importance concerning speaker identification remains largely unclear. The underlying reasons are strongly related to the practical difficulty in separating the source signal from the filter [1,2]. This paper represents an attempt to address this issue taking advantage of the phase information.

Speech elicits a rich and multidimensional analysis from the point of the human auditory system. If fact, it is a surprisingly simple task for a human with normal hearing, to infer from a short sentence as ‘Good Morning’, what was said, who said it and how it was pronounced (i.e., the phonation type). Each one of these three subtasks represents a tremendous challenge from the signal processing point of view [1]. In this paper we focus on the last two processing challenges: the real-time analysis of speech signals and the extraction of suitable and meaningful features paving the way for robust speaker identification and phonation type classification.

While speaker identification has been the object of intense research, also in the perspective of audio forensics, phonation type classification has received much less attention although is possesses a high interest in the context of automatic speech monitoring and segmentation. On the other hand, although real-time products exist already concerning speaker identification, their practical functionality and robustness to noise and other sources of interference (such as simultaneous speech sources or music background), are not commensurate with reliable operation or fast operation and, therefore, are not widely deployed when compared for example to biometric systems based on fingerprints. Interestingly, speaker identification algorithms frequently use the same tools and approaches as speech recognition (namely Mel-frequency Cepstral Coefficients, Gaussian Mixtures Models, Hidden Markov Models) [1] despite the fact that their purpose is quite opposite: if the former case the goal is to identify speech independently of the speaker, in the latter case the goal is to identify the speaker independently of

*This work was supported by the Portuguese Foundation for Science and Technology, an agency of the Portuguese Ministry for Science, Technology and Higher Education, under research project PTDC/SAU-BEB/104995/2008.

the speech. Most often, algorithms in these two application areas ignore the phase information from the speech signal which may be accounted for the huge difference between human performance and machine performance in sound analysis, interpretation and recognition. In this paper we explore specifically the phase information in order extract a new type of feature: the Normalized Relative Delay (NRD) of harmonic partials in voiced speech. As a first study meant to validate the pertinence of this approach, in this paper we use the NRD feature in an experiment of automatic identification of the phonation type (breathy, modal and pressed).

Three main related ideas and previous works motivate our approach.

- Contrarily to common a assumption, many illustrative examples show that phase plays a significant role on the perceptual signature of a periodic sound such as a sustained voiced vowel [3, page 128]. As an example that will be used in section 4.1, a sawtooth wave may be synthesized with the correct phase, with random phase or with the Schroeder phase rule minimizing the 'peak-factor' [3] of the periodic waveform. Despite the fact that the magnitude spectrum in all three cases is exactly the same, the shapes (or time envelopes) of the resulting periodic waveforms are different and, in fact, sound different to a human listener.

- Many experiments in time-scale modification of speech [4] or in pitch modification of speech [5] emphasize that in order to reach good quality for the modified speech and in order to preserve the acoustic signature of the speaker, the shape of the speech waveform around each pitch pulse onset must be invariant. This concept is known as 'shape-invariance' and is implemented by both time-domain algorithms (e.g., Pitch Synchronous Overlap-and-Add, [4]) and frequency domain algorithms [5, 6]. In the latter case, as pointed out by Laroche [5], a key aspect to insure shape invariance is to implement a correct phase synchronization between the harmonics. This is precisely the main purpose of this paper: a specific phase-related feature permitting analysis and synthesis of a desired harmonic phase synchronization, independently of the overall time delay of the waveform, and independently of its fundamental frequency (or pitch).

- The excitation of the vocal tract (i.e., the glottal pulse) influences very strongly the type of speech. In particular, it has been shown that the glottal pulse may be estimated from speech and suitable parameters may then be extracted from the estimated glottal waveform allowing for example to automatically identify the phonation type (breathy, normal, breathy) [7, 8]. Since the shapes of the glottal waveforms associated with these three different phonation types differ significantly, then the phase synchronization (or relative delay) among the harmonics must differ substantially in each case and this information must be embedded in the speech signal. Therefore, it should be possible to extract and use this information from the speech waveform. This is the challenge this paper tackles: to extract a meaningful phase-related feature from speech and to assess its ability and performance in identifying the phonation type.

This paper is structured as follows. In section 2 we highlight that the relative phase between the Fourier components of a non-periodic waveform are accessible when this waveform is made periodic by looking at the harmonics of a DFT analysis encompassing several periods of the waveform. In section 3 we describe the concept of NRD and its computation. In section 4 we describe the NRD algorithm and illustrate its operation with a few illustrative examples. In section 5 we describe the simulation tests implementing automatic classification of the phonation type and we characterize the results. Finally in section 6 we summarize the main results of the paper and we address future research and developments.

2. HARMONIC ANALYSIS AND SHAPE INVARIANCE

We assume in this section that $x[n]$ represents a non-periodic signal of length N and with smooth endings such as a single glottal pulse. Its DFT is obtained as

$$X[k] = \sum_{n=0}^{N-1} x[n] W_N^{kn} , \qquad (1)$$

where $W = e^{-j2\pi}$. Let us assume that $y[n]$ is obtained by repeating L times the period represented by $x[n]$. The DFT of the new signal has length LN and is given by

$$Y[k] = \sum_{n=0}^{LN-1} y[n] W_{LN}^{kn} . \qquad (2)$$

Since $y[n] = y[n+\ell N] = x[n]$ for $\ell = 1,2,\ldots,L-1$ and for $n = 0,1,\ldots,N-1$, (2) may be written as

$$Y[k] = \sum_{\ell=0}^{L-1} W_L^{k\ell} \sum_{n=0}^{N-1} x[n] W_{LN}^{kn} = L\delta[k-mL] \sum_{n=0}^{N-1} x[n] W_{LN}^{kn}\,, \tag{3}$$

which is equivalent to

$$Y[mL] = LX[m],\ m = 0,1,\ldots,N-1\,. \tag{4}$$

This result is well-known from multirate signal processing theory and reflects that oversampling in the frequency domain implies periodization in the time domain and *vice-versa*. In other words, in our context it also means that the details of a signal may be analyzed by picking the harmonics, in the frequency domain, of its periodic repetition in time. However, when dealing with real signals and practical algorithms such as the Fast Fourier Transform (FFT) and practical transform lengths (making that the length of $y[n]$ is frequently chosen as a power of two number), exact frequency sampling as denoted by (4) is never achieved in practice. Instead, leakage occurs due to the window used in the discrete Fourier analysis. Furthermore, the time window adds a phase contribution to $Y[k]$ which must be taken into consideration when extracting the phase of each harmonic component of the periodic signal. A detailed signal analysis must therefore be implemented as described in the next two sections.

3. NORMALIZED RELATIVE DELAY (NRD)

In this section, according to (5), we assume that a quasi-periodic signal $x[n]$ consists of L sinusoids that are approximately harmonic of a fundamental frequency ω_0.

$$\begin{aligned} x[n] &= A_0 \sin(n\omega_0 + \phi_0) + \sum_{\ell=1}^{L-1} A_\ell \sin(n\omega_\ell + \phi_\ell) \\ &= A_0 \sin\omega_0(n+n_0) + \sum_{\ell=1}^{L-1} A_\ell \sin\omega_\ell(n+n_\ell) \end{aligned} \tag{5}$$

In (5), A_ℓ, ϕ_ℓ and n_ℓ denote, respectively, the magnitude, phase and time delay of the ℓ^{th} sinusoid relative to a reference point in n. When $x[n]$ is transformed to the frequency domain using a complex uniform transform such as the DFT, the reference point depends on the time window that is used to multiply the data before DFT transformation. Its length N defines the time support of the time-frequency transformation and usually matches the length of the transform. Since most often the window is even symmetric, the natural reference point corresponds to the center of the window. In other words, it corresponds to the group delay of the filter whose impulse response is the time window. Therefore, if $X[k]$ represents the complex transform (of length N) of $x[n]$ after windowing, the phase of $X[k]$ denotes the time delay n_k relative to the center of the window. We may therefore omit in (5) the independent variable n so as to highlight the NRD concept and to emphasize that it may be estimated from the magnitude and phase information provided by $X[k]$. In this context and taking into consideration the discussion in section 2, if N is at least about three times the period of the fundamental frequency (i.e., if $N > 3\times P_0 = 3\times 2\pi/\omega_0$) then, the magnitude of $X[k]$ exhibits a discernible harmonic structure whose local maxima (or spectral peaks) denote the periods $P_\ell = 2\pi/\omega_\ell$ of the different sinusoids in (5). As will be described in section 4, these periods are estimated from the magnitude spectrum and the time delays are extracted from the phase spectrum. Both are instrumental to define the NRD:

$$\begin{aligned} x &= A_0 \sin\omega_0 n_0 + \sum_{\ell=1}^{L-1} A_\ell \sin\omega_\ell(n_\ell) \\ &= A_0 \sin 2\pi\frac{n_0}{P_0} + \sum_{\ell=1}^{L-1} A_\ell \sin 2\pi\frac{n_\ell}{P_\ell} \\ &= A_0 \sin 2\pi\frac{n_0}{P_0} + \sum_{\ell=1}^{L-1} A_\ell \sin 2\pi\frac{n_0+n_\ell-n_0}{P_\ell} \\ &= A_0 \sin 2\pi\frac{n_0}{P_0} + \sum_{\ell=1}^{L-1} A_\ell \sin 2\pi\left(\frac{n_0}{P_\ell}+\frac{n_\ell-n_0}{P_\ell}\right) \\ &= A_0 \sin 2\pi\frac{n_0}{P_0} + \sum_{\ell=1}^{L-1} A_\ell \sin 2\pi\left(\frac{n_0}{P_\ell}+\text{NRD}_\ell\right) \end{aligned} \tag{6}$$

In this equation NRD_ℓ represents the delay difference between the ℓ^{th} sinusoid and the fundamental frequency, relative to the period of the ℓ^{th} sinusoid. Since it is desired that the NRD is normalized, i.e., that $0.0 \leq \text{NRD}_\ell < 1.0$, the NRD is computed by following two steps:

1. $\text{NRD}_\ell = \frac{n_\ell - n_0}{P_\ell} + \lfloor\frac{n_0}{P_\ell}\rfloor$, where $\lfloor\cdot\rfloor$ denotes the largest integer, and
2. if $\text{NRD}_\ell < 0.0$, $\text{NRD}_\ell = \text{NRD}_\ell + 1$.

The meaning of the NRD concept may also be explained with the help of Fig. 1. In this figure, n_0 represents the

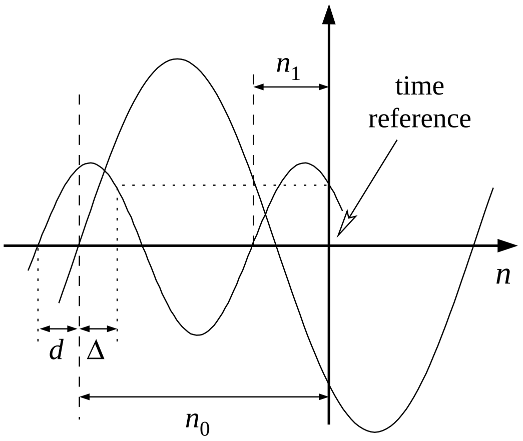

Fig. 1: Illustration of the relative delay d between a harmonic of a fundamental frequency sinusoid whose delay to a time reference is n_1, and the fundamental frequency sinusoid whose delay to a time reference is n_0.

delay of the fundamental frequency sinusoid relative to a time reference that in our context corresponds, as explained above, to the center of the time analysis window defining the time support of the time-frequency transformation. The delay of a harmonic sinusoid relative to the same time reference is represented by n_1. As illustrated in Fig. 1, n_0 is equal to Δ plus an integer number of periods of the harmonic sinusoid that fit within n_0. If the period of the harmonic sinusoid is represented by P_1, then $n_0 = \Delta + \lfloor n_0/P_1 \rfloor P_1$. The relative delay is therefore given by $d = n_1 - \Delta = n_1 - n_0 + \lfloor n_0/P_1 \rfloor P_1$. If $d < 0.0$, then $d = d + P_1$ so that d is always a positive number less than P_1. When d is divided by P_1 it becomes normalized between 0.0 and 1.0 and corresponds to the NRD. Due to its nature, the NRD has the same properties of phase and thus the notions of periodicity, wrapping and unwrapping, also apply.

Instead of describing a quasi-harmonic signal by using the triplet A_ℓ, ω_ℓ, and ϕ_ℓ, i.e., the magnitude, frequency and phase of all harmonic partials, the NRD concept allows to describe that signal by using the magnitude, the frequency and the delay of each partial of the harmonic structure relative to the fundamental frequency sinusoid. The advantage is that the only variable time information is the delay of the fundamental frequency sinusoid since the delays of all partials are relative to it. Thus, the NRD consists in an important signal feature characterizing the shape of the waveform of that signal, independently of its overall time shift, and independently of its fundamental frequency. This is extremely important not only for signal analysis and identification, but also for signal transformation since the NRD of one harmonic signal may be imprinted on any other harmonic signal of a different pitch frequency. Provided that the magnitude and frequency relations among all partials are preserved, shape invariance will also be preserved. Thus the NRD denotes the time waveform of a periodic signal and, therefore, implicitly denotes its time envelope. This suggests that the NRD has the potential to represent this perceptually meaningful characteristic of a complex tone, in a compact way.

The next section will describe the algorithm implementing the computation of the NRD concept and will describe its operation with a few illustrative examples.

4. NRD ALGORITHM

The NRD$_\ell$ coefficients are computed using the algorithm illustrated in Fig. 2. The signal $x[n]$ is first multiplied by the square root of a shifted Hanning window (which happens to correspond simply to the first half of a sine period and hence it is also known as sine window [9]) and is then transformed to the frequency domain by means of an Odd-frequency Discrete Fourier Transform (ODFT) [10]. These two steps are described in [11, 12] and are inherited from an audio coding algorithm [13]. The ODFT spectrum, $X_O[k]$, is then analyzed in detail so as to identify the most relevant harmonic structure of sinusoids [14] and in order to accurately estimate the frequency, magnitude and phase of each sinusoid relative to the center of the time analysis window. As illustrated in Fig. 2, only the frequency ω_ℓ and the phase ϕ_ℓ pertaining to each harmonic partial are used to compute the NRD coefficients. Accurate frequency estimation uses an algorithm presented in [12]. A similar algorithm adapted to

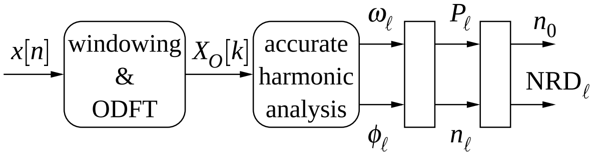

Fig. 2: Algorithm implementing the estimation of the NRD parameters and overall time shift (n_0) of the waveform.

the DFT has been recently presented in [15, 16]. Accurate phase estimation uses an algorithm similar to that presented in [17]. In fact, while in [17] the algorithm presumes a time reference corresponding to the beginning of the time window, a modification has been introduced such that the center of the time window is considered instead. It should be noted that both frequency estimation and phase estimation are performed in a non-iterative way using the magnitude and phase information available from a single audio frame. These constraints pave the way for the real-time operation of the algorithm which is a must considering our target application scenarios (including the biofeedback of the singing voice and real-time voice analysis for speaker identification purposes).

The period and delay of the ℓ^{th} sinusoid are given by (7) and (8), respectively.

$$P_\ell = \frac{2\pi}{\omega_\ell} \tag{7}$$

$$n_\ell = \frac{\phi_\ell}{\omega_\ell} \tag{8}$$

The NRD_ℓ coefficients are subsequently obtained as explained in section 3. It should be noted that shape invariance is represented by NRD_ℓ for $\ell = 1, 2, \ldots, L-1$ and is independent of the overall time shift (n_0) and pitch frequency of the waveform (ω_0). The next two sub-sections will illustrate a few examples of NRD analysis using synthetic and natural signals.

4.1. Examples with synthetic signals

In order to test the NRD algorithm, several synthetic signals for which the relative delays are known, have been used. One such signal is the sawtooth wave whose continuous time signal is easily reconstructed by using a desired number of terms of the Fourier series:

$$x(t) = -\sum_{\ell=1}^{L} \frac{2A}{\pi\ell} \sin \frac{2\pi}{T} \ell t\,. \tag{9}$$

In this equation A represents the amplitude, T represents the reciprocal of the fundamental frequency, and the sampling frequency has been set to 16 kHz. The number of sinusoids L has been set to 50. The resulting waveform, its magnitude spectrum, and the estimated NRD coefficients (only the first 30 are shown) are plotted in Fig. 3.

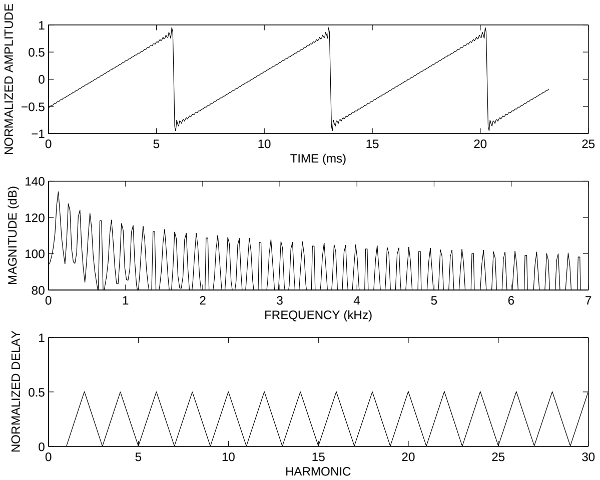

Fig. 3: Time representation of a sawtooth wave (upper figure) synthesized using 50 coefficients of the Fourier series, its magnitude spectrum (middle) and estimated NRD coefficients (lower figure).

The sawtooth waveform is interesting because the phases of all harmonic sinusoids are known. A simple analysis of 9 reveals that taking the null phase (or delay) of the fundamental frequency as a reference, then the phases of higher harmonics alternates between π, 0, π, …, which, in terms of Normalized Relative Delays, corresponds to $0.5, 0, 0.5, \ldots$. This is in fact the output of the algorithm as can be seen in Fig. 3 and, as expected, this output is stable and independent of the overall time shift of the waveform. Interestingly, if the minus sign is removed in (9) then all harmonics are in phase (relative to the null phase of the fundamental) and the output of the NRD algorithm is zero for all harmonics. Furthermore, if the input to the algorithm is according to (9) and if all NRDs are then forced to be zero, a reconstruction algorithm using the magnitude spectrum, n_0 (see Fig. 2) and the so modified NRDs, delivers an output waveform that is the symmetrical of that in Fig. 3, as expected. The possibilities opened by this analysis/synthesis approach using NRDs are promising and will be discussed in a future paper.

A perceptually interesting experiment has been performed as suggested by Schroeder [3, page 128], by altering the initial phases of all harmonics in (9) so as to minimize the 'peak-factor' [3] of the periodic waveform. Such as waveform and its corresponding analysis are displayed in Fig. 4. When comparing Fig. 4 and Fig. 3, it can be concluded that while the magnitude spectrum

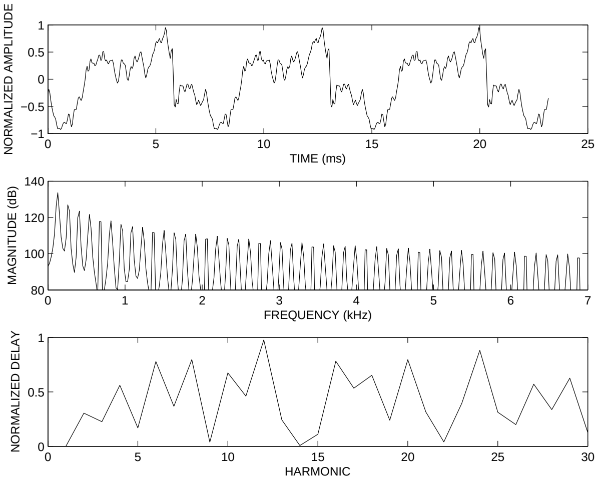

Fig. 4: Time representation of a sawtooth wave (top) synthesized using 50 coefficients of the Fourier series and whose initial phases have been modified according to a Schroeder rule [3] minimizing the 'peak-factor'. The figure in the middle represents the corresponding magnitude spectrum and the lower figure represents the estimated NRD coefficients.

is exactly the same in both cases, the NRD profiles are markedly different since the waveform shapes are quite different and indeed sound quite different. Yet, a signal analysis using Mel-Frequency Cepstral Coefficients (MFCC) for example (that only rely on magnitude spectrum) would led to the conclusion that both waveforms sound the same. Contrarily to a general assumption, this is a simple evidence that MFCCs must be completed with phase related information so as to reflect better human auditory perception.

An important synthetic signal has been tested that corresponds to a train of glottal pulses according to the Liljencrants-Fant model [18]. The Voicebox Matlab toolbox has been used to generate this signal. The corresponding analysis is depicted in Fig. 5. In this figure the first 30 partials of the magnitude spectrum are represented as well as the corresponding NRD coefficients (it should be noted that by definition the first NRD coefficient is always zero as it corresponds to the reference delay of the fundamental frequency). It is interesting to note that the NRD profile is regular and is wrapped. Unwrapping would lead to a straight line and this is perhaps a hint for the natural structure of the glottal excitation. This is an idea for further research on the possibility to reconstruct the glottal pulse using NRD information.

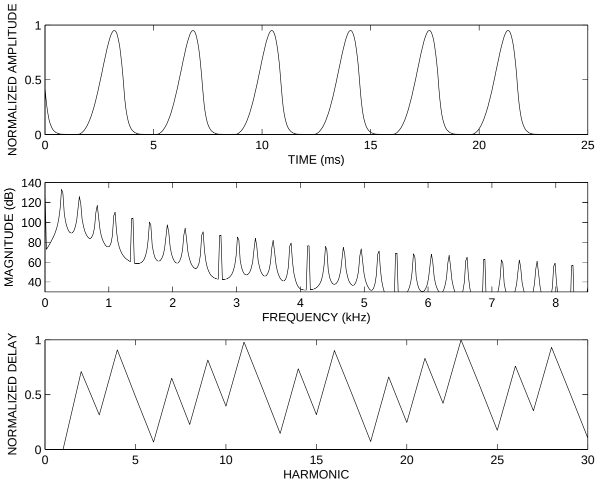

Fig. 5: Time representation of an ideal Liljencrants-Fant glottal pulse train (top), its corresponding magnitude spectrum (middle) and NRD representation (lower figure).

Interestingly, we did a reconstruction of the signal using a synthesis algorithm that is the reverse of that represented in Fig. 2 and by forcing a decay of the magnitude partials of 6db/oct and by zeroing all NRD coefficients. The result was a perfect sawtooth wave with the pitch of the glottal pulse train. This simple experiment is another evidence that in fact the NRD coefficients materialize the shape invariance concept. We also tried the converse: taking the sawtooth period signal, forcing a decay of the magnitude partials of 12 dB/oct and using the NRD 'signature' of Fig. 5. The resulting waveform was similar but not as smooth as the glottal pulse of Fig. 5. Further investigation revealed that the main problem is that the first few partials do not strictly follow a decay as high as 12 dB/oct as it is commonly admitted in the literature.

4.2. Examples with natural signals

In this section three electroglottographic (EGG) signals recorded from the same speaker are analyzed. The three signals correspond to three different phonation types of the same vowel. Both EGG and voice signals will be used in the automatic classification tests described in the next section.

Figures 6, 7 and 8 correspond to the phonation types breathy, normal and pressed, respectively. The EGG signals denote the area of contact of the vocal chords

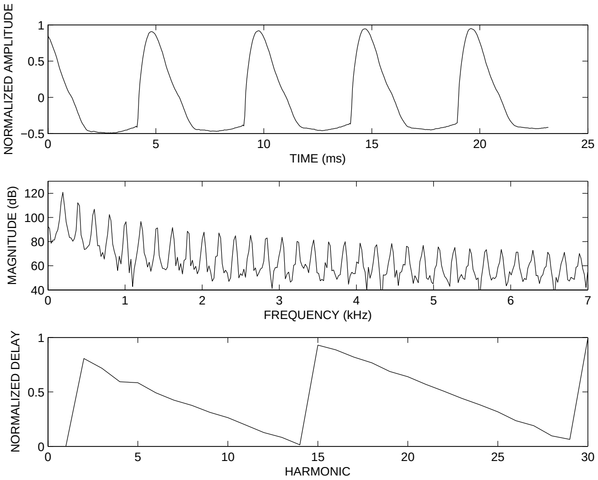

Fig. 6: Time representation of a natural EGG signal resulting from the breathy phonation of a sustained vowel (top), its corresponding magnitude spectrum (middle) and NRD representation (lower figure).

and thus do not represent directly the glottal airflow pulse shape. It can be seen that while the magnitude spectrum does not exhibit dramatic differences, the NRD data behaves smoothly in all three cases and the decay rate as a function of the harmonic number varies noticeably according to the phonation type. This suggests that the NRD may possess relevant discriminative information. This has been one of the motivations for the classification tests that are described in the next section.

5. SIMULATION TESTS AND RESULTS

In this section we describe the automatic classification tests that were conducted in order to assess the performance of the NRDs in the identification of the phonation type. A database of 39 audio records (kindly provided by Prof. Paavo Alku from the Aalto University School of Science and Technology) was used. This database includes the acoustic voice records as well as the EGG signals corresponding the phonation of vowel /a/ by 13 speakers, 7 males and 6 females, using three different phonation types: breathy, normal and pressed. Each record is about 200 ms long. The vowel /a/ was selected because it minimizes the interaction between the source (i.e., the glottal pulse) and the filter (i.e., the resonances of the vocal tract filter). In other words, the spectrum of an acoustic signal corresponding to vowel /a/ is more conveniently ‘illuminated’ as the harmonic partials till

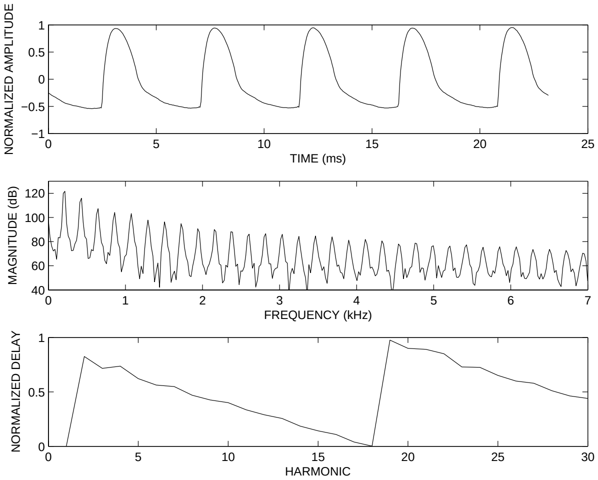

Fig. 7: Time representation of a natural EGG signal resulting from the normal phonation of a sustained vowel (top), its corresponding magnitude spectrum (middle) and NRD representation (lower figure).

about 1000 Hz are not as attenuated as they are for other vowels [19].

In order to benchmark our results we have considered an established method of glottal pulse estimation that is known as Iterative Adaptive Inverse Filtering (IAIF) [7]. Several perceptually relevant time parameters are then extracted from the estimated glottal pulses as well as their derivatives and are considered as input features for the classification of the phonation type. A similar research has been described in [8]. The Matlab software environment implementing the IAIF algorithm and computing the glottal pulse parameters is Aparat and is publicly available.

The glottal pulse parameters considered in our research are the open quotient 1 (OQ1), the open quotient 2 (OQ2), the quasi-opening quotient (QOQ) and the opening quotient ac (OQa) that characterize the glottal open phase; and the closing quotient (ClQ) that characterizes the glottal closing phase. For a detailed definition of these parameters please see [7, 20]. In addition, the amplitude quotient (AQ), the normalized amplitude quotient (NAQ), the speed quotient 1(SQ1) and the speed quotient 2 (SQ2) are computed from the glottal pulse and its derivative, and are also included as features. Local averages of these parameters are considered for groups of about five detected glottal pulses. This number has been found so that the total number of input vectors to

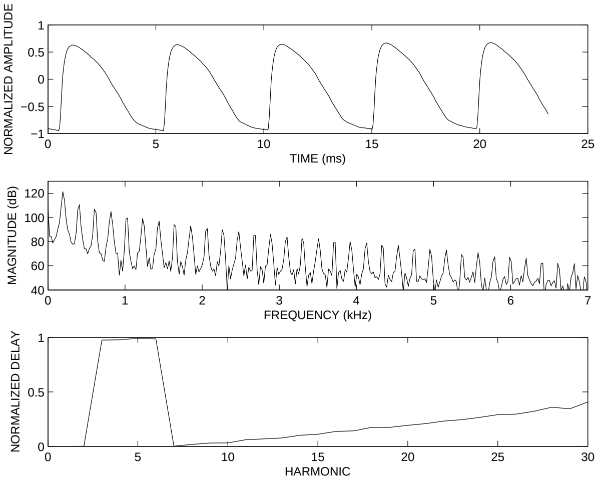

Fig. 8: Time representation of a natural EGG signal resulting from the pressed phonation of a sustained vowel (top), its corresponding magnitude spectrum (middle) and NRD representation (lower figure).

the classification environment (WEKA) is the same for the glottal pulse-based features and for the NRD-based features.

The NRD feature vector consists of the first 15 NRD coefficients estimated from the harmonic structure of the acoustic/EGG signal. As the sampling frequency of the signals is 22050 Hz and the NRD analysis involves frames with 1024 samples and 50% overlap, each 50ms-long frame for which a pitch is identified, generates an NRD feature vector that is used as input to the classification environment.

The selected pattern recognition and classification environment is WEKA and is publicly available [21]. In order to have fairly comparable conditions for glottal pulse-based classification and for NRD-based classification, WEKA has first been used to select the most efficient set of five features in both cases. These sets of features are identified in Table 1 and are presumed on all results presented in this section. Using the results of the feature selection stage, two scatterplots have been prepared to illustrate the discrimination ability provided by the two best features found for NRD and GLOT_F in the case of the acoustic signal. These scatterplots are represented in Fig. 9 for the NRD_1-NRD_2 feature pair, and in Fig. 10 for the NAQ-C1Q feature pair. These figures represent all the acoustic data. It can be seen in both cases that cluster sets can be found despite some dispersion degree.

Table 1: Selected NRD and glottal-based features (GLOT_F) for acoustic and EGG data.

Acoustic		EGG	
NRD	GLOT_F	NRD	GLOT_F
NRD_1	NAQ	NRD_1	NAQ
NRD_2	ClQ	NRD_2	AQ
NRD_6	AQ	NRD_3	QOQ
NRD_8	QOQ	NRD_4	OQa
NRD_{10}	OQ2	NRD_5	SQ1

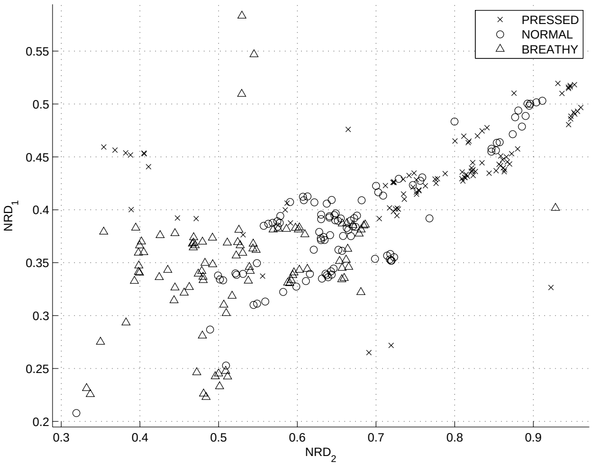

Fig. 9: Scatterplot for the first two NRD coefficients pertaining to all the breathy, normal and pressed voice data.

In fact, the data for the pressed and breathy phonation types are well separated with the normal data lying in-between which suggests that pressed and breathy data will be easier to classify than normal data. This is an indication that in both cases a discrimination ability exits that may lead to meaningful classification results.

In our classification tests we used the 10-fold cross-validation method and the Nearest-Neighbors classifier (NN-classifier) since it is simple and is known to provide good classification results [21]. The confusion matrix is represented in Table 2. This matrix reveals that the breathy and pressed phonation types are identified more successfully that the normal phonation type which confirms the preliminary conclusions extracted from Fig. 9 and from Fig. 10. On the other hand, the results suggest that the overall classification for the acoustic data is more successful than for the EGG data. In fact, the overall scores of correctly identified phonation types are 87.1%

Table 2: Confusion matrix. The columns indicate the number of instances pertaining to the phonation type indicated in the first column that have been identified as breathy (bre), normal (nor) or pressed (pre).

	Acoustic						EGG					
	NRD			GLOT_F			NRD			GLOT_F		
	bre	nor	pre	bre	nor	pre	bre	nor	pre	bre	nor	pre
bre	87	8	4	86	11	0	76	15	7	88	5	2
nor	15	80	5	10	87	3	14	70	16	9	78	8
pre	4	3	96	0	7	91	4	15	79	2	5	88

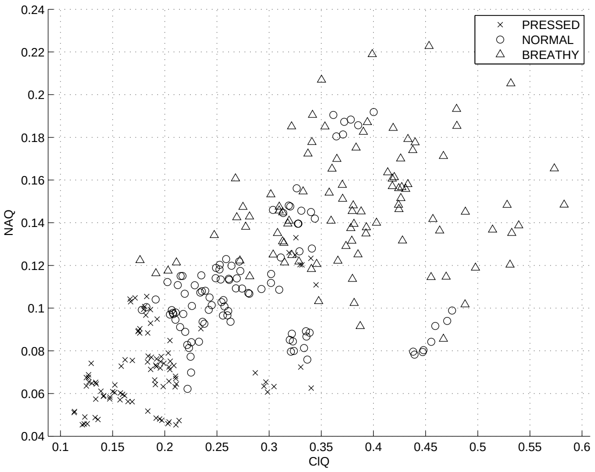

Fig. 10: Scatterplot for the C1Q and NAQ glottal pulse parameters as estimated by the IAIF algorithm and pertaining to all the breathy, normal and pressed voice data.

and 89.5% for acoustic data using NRD and GLOT_F features, respectively; and are 76.0% and 89.1% for EGG data using NRD and GLOT_F features, respectively.

More informative scores may be obtained by considering two intermediary performance measures, the True Positive Rate (TPR) defined as (10) and the Precision defined as (11), and a final performance measure, the F-Measure, defined as (12) [21].

$$\text{TPR} = \frac{\text{instances correctly classified as type X}}{\text{instances classified as type X}} \qquad (10)$$

$$\text{Precision} = \frac{\text{instances correctly classified as type X}}{\text{type X instances}} \qquad (11)$$

$$\text{F-Measure} = \frac{2 \times \text{TPR} \times \text{Precision}}{\text{TPR} + \text{Precision}} \qquad (12)$$

For a given phonation type, the TPR expresses the performance along a column of the confusion matrix, the Precision expresses the performance along a row of the confusion matrix, and the F-Measure combines both. Thus, the F-Measure is a more representative and informative performance measure.

Figure 11 represents the F-Measure scores for all the combinations of data (acoustic and EGG) and feature vector (NRD coefficients or glottal pulse-related parameters). Several conclusions may be extracted.

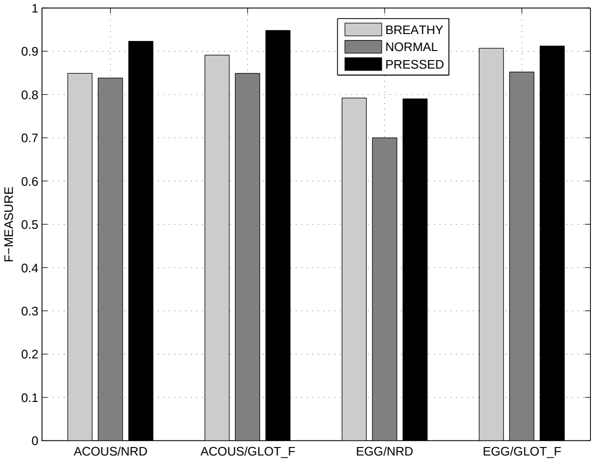

Fig. 11: F-Measure performance results for the automatic classification of the breathy, normal and pressed phonation types using as signal features the most efficient 5 NRD coefficients or the most efficient 5 glottal pulse-related parameters (GLOT_F) delivered by IAIF. The results were obtained with the Nearest-Neighbor classifier and the WEKA classification environment. Both acoustic (ACOUS) and EGG data are considered.

- The performance of the NRD features is remarkably close (and in same cases comparable) to the performance resulting from the glottal pulse features, which is surprising and very revealing. In fact, since the NRD features do not include magnitude information while the glottal pulse features implicitly consider both magnitude and phase information, reveals that the idea of using the relative delay between harmonics is perceptually pertinent and has the potential to lead to more complete models of auditory perception by incorporating phase information and by highlighting its psychophysical interpretation.

- The EGG data lead to better performance when the glottal pulse-based features are used than when the NRD features are used, which may indicate that for healthy voices the area of contact of the vocal folds faithfully reflects the shape the glottal pulse. Surprisingly, when the acoustic signal is used, the performance that is achieved using NRD features or glottal pulse features is quite comparable and is slightly higher than the performance obtained when EGG data is used, which may suggest that acoustic data conveys more reliable information regarding the glottal source waveform. A side conclusion is that since the NRD features are phase related only, they also have the potential to lead to better inverse filtering techniques, i.e., they have the potential to lead to more perfect glottal source estimation.
- It is true for all four combinations of data and test conditions, that pressed phonation can be identified reliably, followed by the breathy phonation, and only then by normal phonation. Thus, the pressed and breathy phonation types possess more extreme phase and magnitude features that facilitate their automatic identification.

6. CONCLUSION

In this paper we have proposed a new approach to the identification and modeling of the shape invariance of a periodic signal by capturing the relative delays between the harmonics. The concept has been described using theoretic considerations and using illustrative examples involving both synthetic and natural audio signals. A new signal feature based on normalized relative delays (NRDs) of the harmonic partials has been proposed, and first results have been presented characterizing its performance in a classification task involving three phonation types: breathy, normal and pressed. Theses results are assessed taking as a reference a well established method of inverse filtering and time-domain features obtained from the estimated glottal source signal. Results are quite encouraging and suggest that the proposed NRD concept and feature correlate well with perceptual information. Future developments will explore further the potential of NRDs in an extended analysis-synthesis context allowing to 1) reconstruct the glottal pulse from spectral information, 2) identify a richer variety of voice registers other than the phonation types addressed in this paper, 3) modify in real-time the sound signature of a speaker, and 4) improve the robustness and efficiency of existing speaker identification algorithms by combining spectral magnitude information and NRDs. Since NRDs depend very strongly on accurate instantaneous frequency and phase estimation algorithms, these will also be adapted for robust operation in real-time and under the influence of dynamic concurrent sounds.

Despite the fact that the NRD concept and associated features are presented here for the first time, and that first results are presented that demonstrate their discrimination potential, we firmly believe that their usefulness will be particularly relevant in forensic-related applications, namely in voice and speaker identification. Further research will be motivated by these objectives.

7. ACKNOWLEDGMENT

The authors would like to thank Prof. Paavo Alku from the Aalto University School of Science and Technology for providing the acoustic and EGG voice records that were used in the tests.

8. REFERENCES

[1] L. Rabiner and B-H Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., 1993.

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

[3] Manfred R. Schroeder, *Computer Speech - recognition, compression, synthesis*, Springer-Verlag, 1999.

[4] Thomas. F. Quatieri and Robert J. McAulay, “Shape invariant time-scale and pitch modification of speech,” *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, March 1992.

[5] Jean Laroche, "Frequency-domain techniques for high-quality voice modification," in *6th Int. Conf. on Digital Audio Effects (DAFx-03)*, 2003, pp. 1–5.

[6] Aníbal J. S. Ferreira, "A new frequency domain approach to time-scale expansion of audio signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 3577–3580.

[7] Matti Airas, "TKK aparat: Enviroment for voice inverse filtering and parameterization," *Logopedics Phoniatrics*, vol. 33, no. 1, pp. 49–64, 2008.

[8] Paavo Alku, "An automatic method to estimate the time-based parameters of the glottal pulseform," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1992, pp. II–29–II–32.

[9] Ted Painter and Andreas Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–513, April 2000.

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

[11] Aníbal J. S. Ferreira, "Combined spectral envelope normalization and subtraction of sinusoidal components in the ODFT and MDCT frequency domains," in *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 21-24 2001, pp. 51–54.

[12] Aníbal Ferreira and Deepen Sinha, "Accurate and robust frequency estimation in the ODFT domain," in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005, pp. 203–206.

[13] Aníbal Ferreira and Deepen Sinha, "Audio communication coder," *120th Convention of the Audio Engineering Society*, May 2006, Paper 6790.

[14] Aníbal J. S. Ferreira, "Perceptual coding of harmonic signals," *100th Convention of the Audio Engineering Society*, May 1996, Preprint n. 4177.

[15] Aníbal J. S. Ferreira and Ricardo Sousa, "DFT-based frequency estimation under harmonic interference," in *ISCCP 2010*, 2010.

[16] Ricardo Sousa and Aníbal J. S. Ferreira, "Non-iterative frequency estimation in the DFT magnitude domain," in *ISCCP 2010*, 2010.

[17] Aníbal J. S. Ferreira, "Accurate estimation in the ODFT domain of the frequency, phase and magnitude of stationary sinusoids," in *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 21-24 2001, pp. 47–50.

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

[19] Aníbal J. S. Ferreira, "Static features in real-time recognition of isolated vowels at high pitch," *Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 2389–2404, October 2007.

[20] Laura Lehto, Matti Airas, Eva Bjokner, Johan Sundberg, and Paavo Alku, "Comparison of two inverse filtering methods in parametrization of the glottal closing characteristics in different phonation types," *Journal of Voice*, vol. 21, no. 2, pp. 138–150, 2007.

[21] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques -2nd edition*, Morgan Kaufmann, 2005, `http://www.cs.waikato.ac.nz/∼ml/weka/`.