Singing Voice Analysis Using Relative Harmonic Delays

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

[1] Department of Electrical and Computer Engineering, University of Porto-FEUP, Portugal
rsousa@fe.up.pt, ajf@fe.up.pt

Abstract

In this paper we introduce new phase-related features denoting the delay between the harmonics and the fundamental frequency of a periodic signal, notably of voiced singing. These features are identified as Normalized Relative Delay (NRD) and denote the phase contribution to the shape invariance of a periodic signal. Thus, NRDs are amenable to a physical and psychophysical interpretation and are structurally independent of the overall time shift of the signal, an important property that is shared with the magnitude spectrum in the case of a locally stationary signal. We describe the NRD and report on preliminary studies testing the discrimination capability of NRDs applied to singing signals.

Index Terms: Singing voice, phase and delay features.

1. Introduction

Time and frequency-domain techniques are commonly used to analyze the singing voice which exhibits specific acoustic characteristics and thus requires specific signal processing techniques [1]. For example, inverse filtering techniques are used to extract the glottal excitation of the singing voice [2], and real or complex filter banks are used to analyze the frequency content of singing voice, typically in the form of a magnitude spectrum or a spectrogram [3]. Signal features based on the magnitude spectrum are commonly used in speech processing. On the other hand, the phase spectrum is rarely used because it is difficult to interpret and manipulate. The phase spectrum is frequently ignored despite evidence showing that signals with the same magnitude spectrum but with different phase spectrum, in fact, sound differently to a human listener [3].

In this paper we focus on the extraction of suitable features based on the phase spectrum for classification or identification purposes. A preliminary study was conducted in order to validate the pertinence of our approach through experiments involving automatic classification of sung vowels and involving singer identification. Three main ideas and previous results support our approach. First, the phase information has a significant importance in the perceptual identification of a periodic sound such as a sustained voiced sound [3]. On the other hand, experiments in speech modification reveal that speech quality depends on the correct phase synchronization between the harmonics [4]. In this context, phase-related features that are independent of time delay and of the fundamental frequency, are desired. Finally, we hypothesize that phase-related features may show a useful discrimination capability in automatic vowel classification and singer identification tasks. The structure of this paper is as follows. In section 2 we present the NRD concept and the associated estimation algorithm. In section 3 we describe the classification tests that have been performed and the results that have been obtained. Finally, in section 4 we address the main results of the paper as well as future research.

2. Normalized Relative Delay

In this section, the NRD concept and its estimation algorithm are described. First, we assume that a quasi-periodic signal consists of M sinusoids harmonically related according to a fundamental frequency ω_0:

$$\begin{aligned} s[n] &= A_0 \sin(n\omega_0 + \phi_0) + \sum_{i=1}^{M-1} A_i \sin(n\omega_i + \phi_i) \\ &= A_0 \sin\omega_0(n+n_0) + \sum_{i=1}^{M-1} A_i \sin\omega_i(n+n_i)\,. \end{aligned} \tag{1}$$

The A_i, ω_i, ϕ_i and n_i parameters denote, respectively, the magnitude, the frequency, the phase and the time delay of the i^{th} harmonic. If $s[n]$ is multiplied by a time window $h[n]$ before time-frequency transformation, by means of a DFT for example, one has access to the spectral coefficients $X[k] = DFT\{x[n]\}$ where $x[n] = s[n].h[n]$. Provided that there is sufficient spectral resolution, it can be shown [5] that the phases of the harmonic peaks in $X[k]$, are related to the arguments of the sine functions in (1) and depend on the time index n in the sense that it corresponds to the group delay of the time window $h[n]$. Since this window is chosen to be even-symmetric, the group delay is constant and is independent of the frequency. In fact, the group delay corresponds to the center of the window, i.e. if N is the length of the time window (and also the length of the DFT), the group delay is simply given by (N-1)/2 samples. Hence, when analyzing the harmonic peaks in $X[k]$, the time index n takes the value of a constant time reference corresponding to the group delay and affecting equally all frequencies. In order to focus our analysis on the delays of the harmonics relative to the same time reference ([6] provides a more detailed explanation), we may ignore the influence of n in (1). For simplicity, making this time reference equal to zero, (1) becomes

$$\begin{aligned} s &= A_0\, sin(n_0\omega_0) + \sum_{i=1}^{M-1} A_i\, sin(n_i\omega_i) \\ &= A_0\, sin(2\pi\frac{n_0}{P_0}) + \sum_{i=1}^{M-1} A_i\, sin(2\pi\frac{n_i}{P_i}) \\ &= A_0\, sin(2\pi\frac{n_0}{P_0}) + \sum_{i=1}^{M-1} A_i\, sin[2\pi(\frac{n_0}{P_i} + \frac{n_i - n_0}{P_i})] \\ &= A_0\, sin(2\pi\frac{n_0}{P_0}) + \sum_{i=1}^{M-1} A_i\, sin[2\pi(\frac{n_0}{P_i} + NRD_i)]. \end{aligned} \tag{3}$$

The NRD_i parameter represents the relative delay difference between the i^{th} harmonic and the fundamental frequency sinusoid, and is normalized by the i^{th} sinusoid period P_i. Thus, $\mathrm{NRD_i}$ are simply defined as

$$NRD_i = \frac{n_i - n_0}{P_i}. \tag{4}$$

The NRDs coefficients inherit the underlying phase properties, hence notions such as periodicity and wrapping also apply. The NRD_i characterize the signal waveform, independently of the overall time shift and of the fundamental frequency. This property is useful for signal analysis, identification and transformation. The *NRD* algorithm is illustrated in Figure 1.

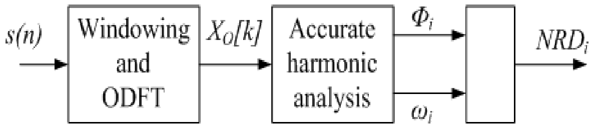

Figure 1: *Computation of the NRD_i coefficients.*

The speech signal $s[n]$ is first multiplied by a sine window (square root of a shifted Hanning window) [5] and then is transformed to the frequency domain using an Odd-frequency Discrete Fourier Transform (ODFT) [7]. The next step consists of a harmonic analysis that accurately identifies the harmonics in the harmonic structure by estimating the amplitude, frequency and phase using a peak picking algorithm and subsequent interpolation [5]. Only the frequency ω_i and phase φ_i are used to compute the NRD coefficients. The period and delay of the i^{th} sinusoid are given respectively by:

$$P_i = \frac{2\pi}{\omega_i}, \tag{5}$$

$$n_i = \frac{\phi_i}{\omega_i}. \tag{6}$$

Finally, the NRD_i are obtained as explained earlier in this section. It should be noted that both frequency and phase are estimated non-iteratively using the $X_O[k]$ magnitude and phase information. These characteristics are important when real-time constraints are mandatory (e.g., biofeedback in singing).

3. Classifications tests

Automatic classification tests were performed in order to assess the discrimination capability of the NRD_i using singing. A database of 40 singing voice recordings was used which includes short 200 ms segments of sung vowels [a], [e], [i], [o] and [u]. These segments were extracted from *arpeggio* vocalizations performed by 8 singers (identified as S1, S2, S3, S4, S5, S6, S7 and S8). The sampling frequency is 22050 Hz.

For comparison purposes, our tests used two well-established set of features with very high discrimination capability. One set of features consists of MFCC coefficients that are widely used in speech recognition. We have used an MFCC analysis software that is publicly available [8]. The second set of features consists of time parameters that are extracted from the estimated glottal pulses and their derivatives. In general, these parameters characterize the glottal flow phases as well as their dynamics where most of perceptual related phenomena occur. The glottal pulse estimation is performed by the Iterative Adaptive Inverse Filtering algorithm that is implemented in a Matlab software environment [9] that is also publicly available. The time parameters consist of the open quotient (OQ), the open quotient 2 (OQ2), the quasi-open quotient (QOQ), open quotient a (OQa), closed quotient (ClQ), speed quotient 1 (SQ1), speed quotient 1 (SQ2), amplitude quotient (AQ) and the normalized amplitude quotient (NAQ) [9]. These features are identified as GLOT_F. The NRD and MFCC features are extracted from the singing signal using 1024-sample frames with 50% overlap. In the case of the GLOT_F features, a local average for each parameter was considered using about five glottal pulses.

In order to characterize the discrimination capability of those features, we used the WEKA [10] [11] pattern recognition and classification environment. This task consisted of three main steps. First, the MFCC, GLOT_F and NRD were extracted from each frame and assumed as an instance for classification training and testing. A combination of NRD and MFCC was also tested. Then, a feature selection algorithm (included in the WEKA environment) was used in order to select the most significant features. The last step corresponds to the classification and relative performance assessment. The Nearest Neighbor classifier and the 10-fold cross-validation method were chosen among the available methods since it has shown better results in preliminary tests. The F-measure was chosen to characterize the identification performance. This measure is computed using (9) and involves the Precision and True Positive Rate (TPR) parameters that are computed using (7) and (8), respectively. The Precision is the proportion of correctly identified signals among all instance classified as a class, and the TPR reflects how much of the class was correctly captured.

$$Precision=\frac{instances\_correctly\_classified\_as\_type\_X}{instances\_classified\_as\_type\_X} \tag{7}$$

$$TPR=\frac{instances\_correctly\_classified\_as\_type\_X}{number\_of\_type\_X\_instances} \tag{8}$$

$$F=\frac{2\times TPR\times Precision}{TPR+Precision} \tag{9}$$

Therefore, F-measure is an overall and more informative measure of performance than just Precision or TPR.

4. Results

This section presents the results of our tests regarding the discrimination capability for vowel and singer using different feature sets. Table 1 exhibits the five most efficient NRDs, glottal (GLOT_F) and MFCC features that were selected for vowel classification.

NRD	GLOT_F	MFCC
NRD1	OQ2	MFCC1
NRD2	ClQ	MFCC2
NRD3	OQa	MFCC3
NRD4	QOQ	MFCC4
NRD5	SQ1	MFCC5

Table 1: *Selected NRD, glottal and MFCC features.*

While the low-order MFCC coefficients convey coarse spectral shape information, the NRDs coefficients convey phase information pertaining to the first few (and typically stronger) harmonics and thus convey time shape information.

Tables 2 to 4 present the confusion matrices of the vowel classification tests for the NRD, GLOT_F and MFCC features. In general, the confusion matrices show high scores of correctly classified vowels.

Vowel	NRD Classified as: (%)				
	[a]	[e]	[i]	[o]	[u]
[a]	**87,5**	4,7	4,7	3,1	0
[e]	3,1	**79,7**	7,8	4,7	4,7
[i]	8,1	3,2	**83,9**	3,2	1,6
[o]	1,6	0	3,2	**90,5**	4,8
[u]	3,1	0	4,7	1,6	**89,1**

Table 2: *Confusion matrix of vowel classification using NRD features.*

The results show that NRDs have a surprisingly competitive discrimination capability in comparison to the other sets of features.

Vowel	GLOT_F Classified as: (%)				
	[a]	[e]	[i]	[o]	[u]
[a]	**96,7**	0	3,3	0	0
[e]	1,7	**88,2**	8,5	1,7	0
[i]	1,7	6,8	**81,4**	8,5	1,7
[o]	8,5	6,8	3,4	**74,6**	6,8
[u]	0	1,7	3,5	8,6	**86,2**

Table 3: *Confusion matrix of vowel classification using GLOT_F features.*

Vowel	MFCC Classified as: (%)				
	[a]	[e]	[i]	[o]	[u]
[a]	**87,7**	0	0	8,8	3,5
[e]	0	**94,9**	3	0	0
[i]	0	5,1	**94,9**	0	0
[o]	8,6	3,5	0	**82,8**	5,2
[u]	0	0	0	5,2	**94,8**

Table 4: *Confusion matrix of vowel classification using MFCC coefficients.*

In particular, the F-measure resulting from the NRD feature set is more uniform than in the case of the GLOT_F or the MFCC feature sets. The fact that GLOT_F features are able to discriminate among vowels is a surprising result since GLOT_F should reflect only source-related information, not filter-related information. A plausible explanation is that the inverse filtering process is not effective removing filter information. The overall performance rate of correctly classified instances is 86,3%, 85,4 and 91,1% for the NRD, GLOT_F and MFCC features sets, respectively. Figure 2 represents the F-measure scores of the vowel classification tests.

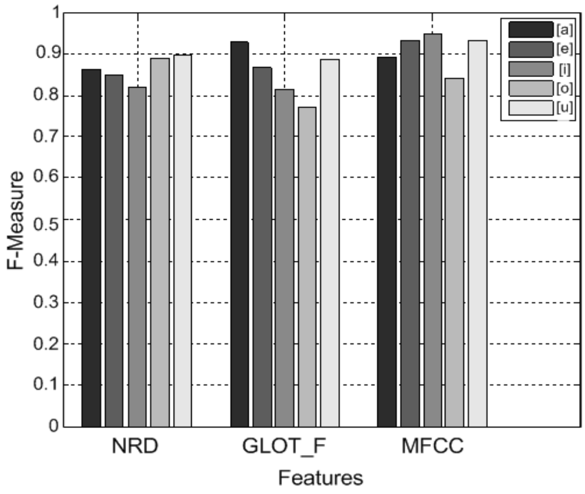

Figure 2: *F-Measure parameter for each vowel and for each feature set.*

Table 5 shows the five most significant features in the singer identification test. Table 5 shows essentially the same behavior obtained in the tests with vowels, i.e., the spectral information in the low frequency region appears to present more discrimination capability.

NRD	GLOT_F	MFCC
NRD1	OQ1	MFCC2
NRD2	OQ2	MFCC3
NRD3	AQ	MFCC4
NRD4	OQa	MFCC6
NRD13	QOQ	MFCC7

Table 5: *Selected NRD, MFCC and glottal features.*

The confusion matrices regarding the singer identification task are not shown due to its considerable size.

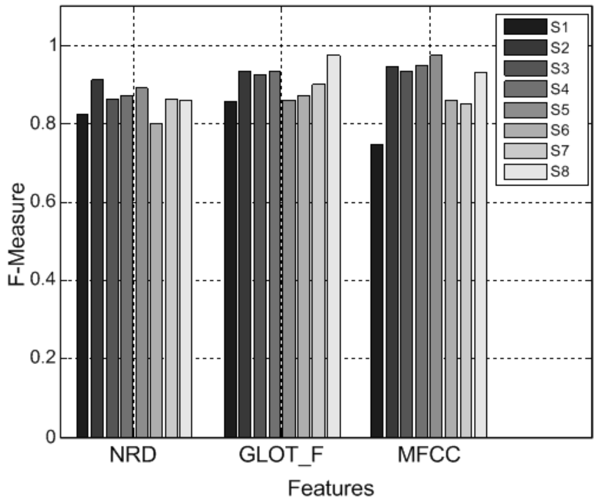

Figure 3: *F-Measure parameter for each speaker and for each feature set.*

Figure 3 shows that NRD also presents competitive results for singer identification. In this case, GLOT_F and MFCC have a small advantage over the NRD coefficients. On the other hand, the NRDs present more uniform F-measure scores considering all subjects. The overall performance rate for the NRD, GLOT_F and MFCC feature set is 86,1%, 90,9% and 90,3%, respectively. The high score obtained by the GLOT_F feature set just confirms that source information is strongly related to speaker individuality.

The combination of NRD and MFCC features presents an overall performance rate of 94,6% for vowel classification, and 97,9% for subject identification. In terms of coefficients for vowel classification, NRD1 to NRD4 and MFCC1, MFCC3, MFCC4 and MFCC6 to MFCC8 were the most discriminative coefficients. NRD2, NRD4, NRD11 and MFCC1 to MFCC4 are the most significant in the subject identification task.

These results are very stimulating and overall, they confirm that the inclusion of phase information improves recognition. It is also possible to conclude that the NRD have a strong impact in subject identification when they are combined with MFCCs.

5. Conclusion

A new set of phase-related features (NRDs) was proposed to be used in the analysis of the singing voice. The phase information may be used for classification and preliminary tests show it can be quite competitive in vowel and singer classification, when compared to time features such as glottal features or magnitude spectrum features such as MFCCs. In future work, the robustness of NRDs features will be evaluated with running singing and speech in recognition tasks and real-time scenarios. The combination with others set of features like the MFCC may also increase significantly the efficiency of the classification.

6. Acknowledgements

This work has been developed in the context of a doctoral program supported by the Portuguese Foundation for Science and Technology under grant SFRH/BD/24811/2005 and project PTDC/SAU-BEB/104995/2008.

7. References

[1] Sundberg J., “Perception of Singing”, Quarterly Progress and Status Report, 20(1):1-48, 1979.

[2] Lindestad, P-Å., Södersten, M., Merker, B. and S. Granqvist, “Voice Source Characteristics in Mongolian Throat Singing Studied with High-Speed Imaging Technique, Acoustic Spectra, and Inverse Filtering”, Journal of Voice, 15(1):78-85, 2001.

[3] Schoeder, M. R., Computer Speech – recognition, compression, synthesis, Springer-Verlag, 1999.

[4] Quatieri, T. F. and McAuley, R. J., “Shape invariant time-scale and pitch modification of pitch”, IEEE Transactions on Signal Processing, 40(3):497-510, 1992.

[5] Ferreira, A., “Accurate estimation in the ODFT domain of the frequency, phase and magnitude of stationary sinusoids,” in 2001 IEEE Workshop on Applications of the Signal Processing to Audio and Acoustics, USA, pp. 47-50, October 2001.

[6] Sousa, R. and Ferreira, A., “Importance of the relative delay of Glottal Source Harmonics”, Proceedings of the AES 39th International conference, Denmark, pp. 59-69, June 2010.

[7] Bellanger, M., Digital Processing of Signals, John Willey &Sons, 1989.

[8] MFCC software. Available at: http://www.mathworks.de/matlabcentral/fileexchange/23119, 2010.

[9] Airas, M., “TKK Aparat: Environment for voice inverse filtering and parametrization“, Logopedics Phoniatrics, 33(1): 49-64, 2008.

[10] Witten, I.H., and Frank, E., Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[11] WEKA. Available at: http://www.cs.waikato.ac.nz, 2010.