



Pitch-based speech perturbation measures using a novel GCI detection algorithm: Application to pathological voice classification

Khalid Daoudi, Ashwini Jaya Kumar

INRIA Bordeaux-Sud Ouest, GEOSTAT team

33405 Talence. France.

<http://geostat.bordeaux.inria.fr>

khalid.daoudi@inria.fr, ashwini.jaya-kumar@inria.fr

Abstract

Classical pitch-based perturbation measures, such as Jitter and Shimmer, are generally based on detection algorithms of pitch marks which assume the existence of a periodic pitch pattern and/or rely on the linear source-filter speech model. While these assumptions can hold for normal speech, they are generally not valid for pathological speech. The latter can indeed present strong aperiodicity, nonlinearity and turbulence/noise. Recently, we introduced on a novel nonlinear algorithm for Glottal Closure Instants (GCI) detection which has the strong advantage of not making such assumptions. In this paper, we use this new algorithm to compute standard pitch-based perturbation measures and compare its performances to the widely used tool PRAAT. We address the task of classification between normal and pathological speech, and carry out the experiments using the popular MEEI database. The results show that our algorithm leads to significantly higher classification accuracy than PRAAT. Moreover, some important statistical features become significantly discriminative, while they are meaningless when using PRAAT (in the sense that they have almost no discrimination power).

Index Terms: Perturbation measures, Pitch marks, Jitter, Shimmer, Pathological speech classification.

1. Introduction

Most of the classical approaches in speech processing are based on linear techniques that may not adequately capture the complex dynamics of speech. Indeed, in normal speech, some of the well-known examples of non-linear phenomena include: the existence of turbulent sound source in production process of unvoiced sounds, the existence of a time spread and turbulent component for the excitation source of plosives (which is idealized as an impulse in the linear framework) and the evidences regarding characterization of voiced sounds by highly complex air flows like jets and vortices. In pathological speech, linear methods are definitely not enough to characterize the strong aperiodicity, the turbulence and breathy noise that can be present in such signals.

In particular, dysphonia analysis is generally performed using pitch-based perturbation measures such as Jitter, Shimmer and their variants. Most of the algorithms to compute these measure use linear methods at the frame-basis level and implicitly assume some kind of periodicity constraints. As a canonical example, the widely used tool PRAAT [1] uses a short-term autocorrelation method [2] for the detection of pitch marks. Such implicit periodicity assumptions can not hold for pathological speech, particularly speech with severe dysphonia.

In [3], we recently developed a nonlinear algorithm for Glottal Closure Instants (GCI) detection. This algorithm has the strong advantages of not making any assumption about the signal (such as periodicity of the source), it does not operate a frame-basis and is computationally efficient. The algorithm has been compared to state-of-the-art methods and showed similar performances for clean (normal) speech and significantly better ones on noisy (normal) speech. Precise detection of GCI has many applications in speech technology: closed phase Linear Prediction (LP) analysis [4, 5, 6, 7], pitch synchronous speech processing for converting the pitch and duration of speech [8], prosody modification [9], synthesis [10, 11], dereverberation [12], casual-anticausal deconvolution [13, 14] and glottal flow estimation [15]. In this paper, we use our GCI detection algorithm in the framework of pathological voice classification. Indeed, motivated by our good results in noisy (normal) speech, we could fairly expect that our algorithm may also have an interesting behavior in the case of dysphonic speech. We thus proceed to compute classical pitch-based perturbation measures using the GCI locations detected by our algorithm as pitch marks. We then analyze the results we obtain and compare them with one obtained using PRAAT on the task of classification between normal and pathological speech. We carry out the experiments using the popular MEEI-KayPENTAX Voice Disorders database (KPdb) [16]. The results clearly show that going beyond linear method can indeed lead to much better insight and performances in this task.

The paper is organized as follows. In the next section we give a brief description of our GCI detection algorithm. Then, in section 3, we list the perturbation measures we consider in this paper. The core of the paper is in section 4 where we present the experimental results and their analysis. We then provide a conclusion in section 5.

2. GCI detection

In this section, we briefly recall the description of our GCI detection methodology and then, in order to make the paper self-contained, provide the algorithm. The details can be found in [3].

Our GCI detection algorithm is based on a novel framework called the Microcanonical Multiscale Formalism (MMF) [17]. MMF allows the study of local geometrico-statistical properties of complex signals from a multiscale perspective. It is based on precise computation of local parameters called the Singularity Exponents (SE) at every point in signal domain. When correctly defined and estimated, these exponents alone can provide valuable information about local dynamics of complex signals and

has recently proven to be promising in many signal processing applications ranging from signal compression to inference and prediction in a quite diverse set of scientific disciplines such as satellite imaging [18, 19, 20, 21], adaptive optics [22, 23], computer graphics [24] and natural image processing [25, 26]. In the field of speech processing, besides GCI detection [3], we have also successfully used MMF in phonetic segmentation [27, 28, 29]. An important subset of SEs is called the Most Singular Manifold (MSM), which is defined as the set of points having the smallest SE values. Indeed for a given point, the smaller the value of SE is, the higher unpredictability is at this point [17]. It has been established that the critical transitions of the system occur at these points, and this fact has been used in many signal processing applications [20, 30]. MSM constitutes the core of our algorithm for GCI detection which consists in the following steps. Let $s[n]$ be a discrete time speech signal with a sampling frequency f_s . The multi-scale integral of the following scale r_i dependent functional is first defined as:

$$\Gamma_{r_i}(s[n]) = |2s[n] - s[n - r_i] - s[n + r_i]| \quad (1)$$

Then the singularity exponent $h[n]$ at the discrete time instance n is computed as:

$$h[n] = \sum_{i=1}^I h_i[n] \quad (2)$$

where

$$h_i[n] = \frac{\log(\Gamma_i(s[n]))}{\log(r_i)} \quad (3)$$

I is the number of scales used for estimation and $r_i = i/f_s$.

We then define the regularity-drop functional $\mathcal{D}_L[n]$ that measures the change in local average of SEs before and after any time instant n , on two adjacent windows of length T_L :

$$\mathcal{D}_L[n] = \sum_{k=n-T_L}^{n-1} h[k] - \sum_{k=n}^{n+T_L} h[k] \quad (4)$$

The final algorithm¹ is given below:

Algorithm 1 : GCI detection

- 1: Calculate $h[n]$ and $\mathcal{D}_L[n]$.
 - 2: In $\mathcal{D}_L[n]$, for any positive-going zero-cross time instant n_{pos} , find the next negative-going zero-cross n_{neg} .
 - 3: $n_{peak} \leftarrow \underset{n}{\operatorname{argmax}} \mathcal{D}_L[n]$, $n \in [n_{pos}, n_{neg}]$.
 - 4: MSM formation: take n_1, n_2, n_3 having the lowest values of $h[n]$ in $n \in \{n_{pos}, n_{neg}\}$.
 - 5: $n_{msm} \leftarrow \underset{n_i}{\operatorname{argmin}} |n_i - n_{peak}|$
 - 6: $n_{gci} \leftarrow [(n_{peak} + n_{msm})/2]$
-

This algorithm has two free parameters: I and T_L which can be tuned on development data.

3. Perturbation measures

The features we consider in this paper are all derived from pitch marks, that we denote $To^{(i)}$, which is the duration of the i -th (estimated) pitch period. We use the default parameters for PRAAT. That is, the Period Floor and Period Ceiling parameters are set at 50Hz and 10000Hz respectively. For our feature computation algorithm, we do not use any threshold. The features we compute are Jitter, Shimmer and their classical variants.

¹A Matlab implementation of this algorithm is made publicly available in <http://geostat.bordeaux.inria.fr/>

1. Jitter is computed by:

$$jitter = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |To^{(i)} - To^{(i+1)}|}{To} \quad (5)$$

where N is the number of extracted pitch marks and $To = \frac{1}{N} \sum_{i=1}^N To^{(i)}$.

RAP, PPQ and sPPQ are computed similarly to Jitter but with 3, 5 and 55 pitch cycles respectively (see PRAAT website www.praat.org).

2. The $To^{(i)}$ standard deviation σ is computed by:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N-1} (To - To^{(i)})^2} \quad (6)$$

where:

3. The $To^{(i)}$ variance vTo is computed by:

$$vTo = \sigma \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{1}{N} \sum_{j=1}^N To^{(j)} - To^{(i)} \right)^2}}{To} \quad (7)$$

4. The Skewness is computed by:

$$skewness = \frac{\frac{1}{N} \sum_{i=1}^N (To^{(i)} - To)^3}{\left(\frac{1}{N} \sum_{i=1}^N (To^{(i)} - To)^2 \right)^{3/2}} \quad (8)$$

5. Shimmer is computed by:

$$shimmer = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |A^{(i)} - A^{(i+1)}|}{\frac{1}{N} \sum_{i=1}^N A^{(i)}} \quad (9)$$

where: $A^{(i)}$, $i = 1, 2, \dots, N$ are the extracted peak-to-peak amplitude and N is number of extracted pitch marks.

APQ and sAPQ are computed similarly to Shimmer but with 5 and 55 peak-to-peak cycles.

4. Experimental results

4.1. The MEEI-KayPENTAX Voice Disorders database (KPdb)

The MEEI-KayPENTAX Voice Disorders database [16] was released in 1994 and has been developed by the MEEI Voice and Speech lab and the Kay Elemetrics (now KayPENTAX) Corp. The recordings consist in sustained phonation of the vowel /ah/ (53 normal and 657 pathological) and utterance of the first sentence of the rainbow passage (53 normal and 662 pathological). All normal vowels and 77 pathological vowels are sampled at 50 kHz, while the remaining 580 pathological vowels are sampled at 25 kHz. 36 of the normal rainbow sentences are sampled at 25 kHz and 17 at 10 kHz. 648 of the pathological sentences are sampled at 25 kHz, 13 at 10 kHz and one at 50 kHz. More details about KPdb can be found in [16]. In the last years, KPdb has been the most widely used dataset for research in pathological voice classification. In this paper, we use the full dataset of sustained vowels in the experiments.

4.2. Results analysis

We use a Random Forest classifier with leave-one-out cross validation method. The performance of the classifier are evaluated in terms of percentage of true positives (TP), i.e. when a disordered subject is correctly assigned to the disordered class and true negatives (TN), i.e. when a normal subject is correctly assigned to the normal class. The overall performance is the percentage of correctly classified subjects in both classes. Classification scores are presented in Table 1. The first remark is that GCI significantly outperforms PRAAT for Jitter and its variants (RAP, PPQ and sPPQ) in both TP and TN (and hence overall scores). For Shimmer, PRAAT outperforms GCI in TN but the GCI's overall score is higher. For Shimmer variants (APQ and sAPQ), GCI significantly outperforms PRAAT. This suggests that GCI is definitely a better choice than PRAAT for the most widely used pitch perturbation features.

More importantly, the second and third order statistics show a very interesting behavior. It is worth noting that these statistics are generally not used as perturbation measures. The explanation become obvious when looking at PRAAT's scores with these features ($vT0$, σ and $skewness$). Indeed, they have very low scores on TN (besides for σ which has an acceptable one) which implies that they have no discriminative power. On the other hand, when computed by GCI, these features become discriminative and even significantly outperform Jitter and its variants sometimes. In particular, σ yields a (relatively) very high score and the best one. This shows that GCI captures the expected deviation from the mean of pitch periods in pathological voices. Moreover $skewness$, which is a measure of asymmetry in the distribution of $To^{(i)}$, yields a good score which is significantly higher than Jitter and its variants (in TN). This shows that GCI reveals an asymmetry in the distribution of pitch periods around the mean in pathological voices. To the best of our knowledge, this property has not been observed before in pathological voice analysis. This also suggests that other statistics may be worth analyzing or other new features can be conceived in order to reveal other interesting properties or to achieve higher discriminative power.

Another observation is that PRAAT is unable to process some speech signals sometimes, because of the implicit constraints it imposes in the computation of perturbation measures. On the other hand, thanks to the fact that no constraints or assumptions are imposed by our method, GCI can process any signal. Table 2 presents the number of files missed by PRAAT for each feature. The first 2 columns in "processed files" present the number of files which have been used to compute the scores: the files unprocessed by PRAAT was not considered in score calculation for a fair comparison.

Finally, we emphasize these interesting results were obtained by a "brut" version of our GCI detection algorithm. Indeed, pathological speech was not considered in the conception of the latter. Therefore, we can fairly expect that if we take this option in consideration, we could come up with a more accurate pathological GCI detection algorithm, and hence a better estimation of perturbation measures. Also, a finer analysis would allow us to choose appropriate thresholds to remove outliers (as in PRAAT) and improve classification accuracy.

5. Conclusion

Using our recently developed GCI detection algorithm, we evaluated some standard pitch-based perturbation measures on the task of classification between normal and pathological speech.

Table 1: Classification scores on KPdb

	TP [%]		TN [%]		Overall [%]	
	GCI	PRAAT	GCI	PRAAT	GCI	PRAAT
jitter	89.94	87	47.17	45.23	86.7	83.83
RAP	90.87	88.85	52.83	50.94	87.98	85.98
PPQ	89.77	88.37	56.6	50.94	87.25	85.53
sPPQ	87.18	84.03	49.06	41.5	84.1	80.58
shimmer	92.57	89.16	52.83	58.49	89.56	86.84
APQ	90.03	90.8	60.38	50.94	87.77	87.77
sAPQ	89.77	88.11	54.72	47.12	86.95	84.83
$vT0$	87.13	84.96	54.72	22.64	84.67	80.23
σ	94.89	93.03	73.58	58.49	93.27	90.41
skewness	89.77	83.41	60.38	9.43	87.53	77.79

Table 2: Number of files processed/missed by GCI and PRAAT

	processed files		missed files			
	total files	total files	GCI	PRAAT	GCI	PRAAT
	normal	pathol	normal	normal	pathol	pathol
jitter	53	646	0	0	0	11
RAP	53	646	0	0	0	11
PPQ	53	645	0	0	0	12
sPPQ	53	601	0	0	0	56
shimmer	53	646	0	0	0	11
APQ	53	642	0	0	0	15
sAPQ	53	606	0	0	0	51
$vT0$	53	645	0	0	0	12
σ	53	646	0	0	0	11
skewness	53	645	0	0	0	12

The results first showed that Jitter/Shimmer and their variants have significantly better classification performances with our GCI than PRAAT. More importantly, the results showed that second and third order statistics (of pitch periods) are highly discriminative and informative when using GCI, while they have almost no discriminative power and convey no information when using PRAAT (at least on this database and this task). This suggests in particular that other new features can be conceived in order to reveal other interesting properties and/or to achieve higher discriminative power. This is the objective of our ongoing work, as well as the improvement of our GCI detection algorithm to make it more adapted and robust to pathological speech.

6. References

- [1] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," *Computer program*, 2013.
- [2] —, "Praat: a system for doing phonetics by computer," *Report of the Institute of Phonetic Sciences of the University of Amsterdam*, 1996.
- [3] V. Khanagha, K. Daoudi, and H. Yahia, "Detection of glottal closure instants based on the microcanonical multiscale formalism," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1941–1950, 2014.
- [4] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27 (4), pp. 309–319, 1979.
- [5] E. N. Pinson, "Pitch synchronous time domain estimation of formant frequencies and bandwidths," *Journal of the Acoustical Society of America*, vol. 35 (8), pp. 1264–1273, 1963.
- [6] K. Steiglitz and B. Dickinson, "The use of time-domain selection for improved linear prediction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25 (1), pp. 34–39, 1977.

- [7] P. A. Naylor, "Estimation of glottal closure instants in voiced speech using the dyspa algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15 (1), pp. 34–43, 2007.
- [8] T. Ewender and B. Pfister, "Accurate pitch marking for prosodic modification of speech segments," in *Proceedings of INTERSPEECH*, 2010.
- [9] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453 – 467, 1990.
- [10] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Interspeech conference*, 2010.
- [11] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9 (1), pp. 21–29, 2001.
- [12] N. Gaubitch and P. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *15th International IEEE Conference on Digital Signal Processing*, 2007.
- [13] B. Bozkurt and T. Dutoit, "Mixed-phase speech modeling and formant estimation, using differential phase spectrums," in *ISCA Voice Quality: Functions, Analysis and Synthesis*, 2003.
- [14] T. Drugman, "Advances in glottal analysis and its applications," Ph.D. dissertation, University of Mons, 2011.
- [15] D. Wong, J. Markel, and A. J. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 4, pp. 350–355, aug 1979.
- [16] *Voice disorders database, (Version 1.03 cd-rom)*. Lincoln Park, NJ: Kay Elemetrics Corp., 1994.
- [17] A. Turiel, H. Yahia, and C. Pérez-Vicente., "Microcanonical multifractal formalism: a geometrical approach to multifractal systems. part 1: singularity analysis," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, p. 015501, 2008.
- [18] J. Grazzini, A. Turiel, H. Yahia, and I. Herlin, "Edge-preserving smoothing of high-resolution images with a partial multifractal reconstruction scheme," in *International Society for Photogrammetry and Remote Sensing (ISPRS)*, 2004.
- [19] J. Grazzini, A. Turiel, and H. Yahia, "Multifractal Formalism for Remote Sensing: A Contribution to the Description and the Understanding of Meteorological Phenomena in Satellite Images," in *Complexus Mundi. Emergent Patterns in Nature*, M. M. Novak, Ed. World Scientific Publishing Co. Pte. Ltd., 2006, pp. 247–256.
- [20] H. Yahia, J. Sudre, C. Pottier, and V. Garçon, "Motion analysis in oceanographic satellite images using multiscale methods and the energy cascade," *Journal of Pattern Recognition*, 2010, to appear. doi:10.1016/j.patcog.2010.04.011.
- [21] H. Yahia, J. Sudre, V. Garçon, and C. Pottier, "High-resolution ocean dynamics from microcanonical formulations in non linear complex signal analysis," in *AGU FALL MEETING*. San Francisco, États-Unis: American Geophysical Union, Dec. 2011.
- [22] S. K. Maji, O. Pont, H. Yahia, and J. Sudre, "Inferring information across scales in acquired complex signals," in *European Conference on Complex Systems (ECCS)*, 2012.
- [23] S. K. Maji, H. M. Yahia, O. Pont, J. Sudre, T. Fusco, and V. Michau, "Towards multiscale reconstruction of perturbed phase from hartmann-shack acquisitions," in *AHS*, 2012, pp. 77–84.
- [24] H. Badri, "Computer graphics effects from the framework of reconstructible systems," Master's thesis, Rabat faculty of science-INRIA Bordeaux Sud-Ouest, 2012.
- [25] A. Turiel and A. del Pozo, "Reconstructing images from their most singular fractal manifold," *IEEE Transactions on Image Processing*, vol. 11, pp. 345–350, 2002.
- [26] A. Turiel and N. Parga, "The multi-fractal structure of contrast changes in natural images: from sharp edges to textures," *Neural Computation*, vol. 12, pp. 763–793, 2000.
- [27] V. Khanagha, K. Daoudi, O. Pont, and H. Yahia, "A novel text-independent phonetic segmentation algorithm based on the microcanonical multiscale formalism," *INTERSPEECH*, 2010.
- [28] —, "Improving text-independent phonetic segmentation based on the microcanonical multiscale formalism," *ICASSP*, 2011.
- [29] —, "Phonetic segmentation of speech signal using local singularity analysis," *Digital Signal Processing journal*, vol. 35, pp. 86–94, 2014.
- [30] O. Pont, A. Turiel, and C. J. Pérez-Vicente, "Description, modeling and forecasting of data with optimal wavelets," *Journal of Economic Interaction and Coordination*, vol. 4, no. 1, June 2009.