THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Automating quasi-stationary speech signal segmentation in sustained vowels

Citation for published version:
Tsanas, T, Triantafyllidis, AK & Arora, S 2021, 'Automating quasi-stationary speech signal segmentation in sustained vowels: application in the acoustic analysis of Parkinson's disease', Paper presented at 12th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications, Florence, Italy, 14/12/21 - 16/12/21 pp. 153-156.

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 06. Oct. 2022

AUTOMATING QUASI-STATIONARY SPEECH SIGNAL SEGMENTATION IN SUSTAINED VOWELS: APPLICATION IN THE ACOUSTIC ANALYSIS OF PARKINSON'S DISEASE

A. Tsanas[1], A. Triantafyllidis, S. Arora [3]

[1] Usher Institute, Medical School, University of Edinburgh, Edinburgh, UK
[2] Information Technologies Institute, Centre for Research and Technology Hellas
[3] Mathematical Institute, University of Oxford, Oxford, UK
(A. Tsanas): atsanas@ed.ac.uk; (A. Triantafyllidis): atriand@gmail.com; (S. Arora): arora@maths.ox.ac.uk

Abstract: **Acoustic analysis of sustained vowels is typically used to quantify perturbations in fundamental frequency (F0), amplitude, and deviations from periodicity, and associate these with clinical outcomes of interest. Computational and practical constraints suggest that 2-3 seconds are often sufficient to acoustically characterize a sustained vowel phonation. The question then is how to best determine a short quasi-stationary segment from a typical 20-30 seconds speech recording. We computed the F0 contour in 10 millisecond epochs using SWIPE, a state-of-the-art F0 estimation algorithm, which we had previously demonstrated is very competitive in F0 estimation for sustained /a/ vowels. Subsequently, we determined the two second signal segment that exhibits the smallest mean absolute successive F0 difference. We tested the segmentation algorithm on 100 randomly selected sustained vowel /a/ phonations from the Parkinson's Voice Initiative, where we had hand-labeled the quasi-stationary segments. We found the algorithm correctly identified the quasi-stationary segments in all cases, thus demonstrating it can be deployed at large scale studies automating further processing of sustained vowels. We also demonstrated that this pre-processing step can have a major influence in the acoustic characterization of the phonations.**

Keywords: **acoustic analysis, F0 estimation, speech signal segmentation, sustained vowels**

I. INTRODUCTION

The use of sustained vowels to assess voice disorders is well established in clinical practice [1]. Compared to conversational speech or reading out loud specific abstracts of phonetically rich text, sustained vowels have the advantage that they circumvent linguistic confounds and accent effects [1]. The acoustic analysis of sustained vowels towards the development of robust clinical decision support tools has received considerable research attention. Indicatively, we had previously used sustained vowel /a/ phonations to demonstrate: (i) almost 99% accurate differentiation of people diagnosed with Parkinson's Disease (PD) from Healthy Controls (HC) [2]; (ii) accurate replication of the most widely clinical tool assessing overall PD symptom severity reporting an error that is considerably lower than the inter-rater variability [3]–[6]; (iii) assessing PD voice rehabilitation [7]; and (iv) potential on early PD diagnosis/precursors [8], [9]. Researchers have also developed mechanistic models of speech articulation using sustained vowels, which may provide insights into the underlying vocal production mechanism and voice disorders in a physically interpretable way [10], [11].

In practice, the raw speech signal recordings typically include the prompt by the researcher/clinician, possibly some prior discussion, and one or more prolonged sustained vowel phonations by the study participant. Using the entire sustained vowel phonation (typically 20-30 seconds) is computationally demanding and may be prone to problems (e.g. participant coughing, running out of breath). Computational and practical constraints suggest that processing 2-3 seconds of the sustained vowel phonation are sufficient to acoustically characterize the sustained vowels [1], [12] and to develop mechanistic models [10].

The natural question then arises on how best to choose the short signal segment from the raw recording for further processing. Often, this is done manually by selecting the segment that 'looks best' (low amplitude and low frequency variation) or by selecting a pre-specified signal segment (e.g. the middle of the phonation because that would likely be a stable part of the phonation). For small datasets it may be possible to manually detect segments, however as we move on larger datasets, such as with the Parkinson's Voice Initiative (PVI) study with more than 18,000 sustained vowel phonations [13], [14], the need to develop an automated approach becomes obvious. Previous work in the context of speech signal analysis has focused on removing the non-sustained vowel segment of the recording (e.g. the prompts by investigators and silences). Surprisingly, to the best of our knowledge there is no published work on principled objective detection of short signal sustained vowel segments that

Please cite this paper as: A. Tsanas, A. Triantafyllidis, S. Arora: *Automating quasi-stationary speech signal segmentation in sustained vowels: application in the acoustic analysis of Parkinson's disease*, **12th International Workshop Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)**, pp. 153-156, Florence, Italy, 14-16 December 2021

would be best applicable towards further acoustic analysis. Moreover, this crucial pre-processing step is rarely reported in the research literature.

If we revisit the underlying principle of using sustained vowels, the aim is to elicit "stable" phonations and assess deviations from signal periodicity [1]. In practice, minor perturbations from maintaining constant amplitude and frequency are common even for people with no vocal pathologies, where larger fluctuations may be hinting towards a vocal pathology (which may be secondary e.g. to PD or other disorders) [1]. Hence, if we want to work on a short signal segment it would be reasonable to identify the most stable part of the phonation. Technically, that would be the most quasi-stationary segment, where stationarity suggests that the central order moments of the signal remain constant [15]. Relaxing the requirement of quantifying non-stationarity, we can instead aim to quantify changes in the fundamental frequency (F0), i.e. the F0 *contour*. The F0 is a key characteristic of speech and its computation is often a pre-requisite for many speech signal processing algorithms [1], [12].

The aim of this study is to develop an algorithmic approach towards automatically detecting the most quasi-stationary short signal segment from a speech recording that comprises a longer sustained vowel phonation which might also exhibit background noise (prompts, silence etc.). We demonstrate the effectiveness of the proposed approach towards the acoustic characterization of PD voices, although in principle the developed method is generalizable across applications focusing on sustained vowels.

II. METHODS

A. Data

We used data from the large PVI study [13], [14], which was set in seven major geographical locations. Participants were invited to call in a dedicated phone number and contribute two sustained vowel /a/ phonations along with basic demographic information (age, gender), and whether they had been clinically diagnosed with PD. The phonations were sampled at 8 kHz and stored on secure cloud servers. For the purposes of this study we have randomly selected phonations from 50 PD participants and from 50 control participants from the US cohort.

B. F0 estimation and signal segmentation

We had previously performed a thorough empirical comparison of multiple F0 estimation algorithms to establish the most accurate for the analysis of sustained vowels [16]. We had found that the Sawtooth Waveform Inspired Pitch Estimator (SWIPE) [17] was very competitive [16] and hence it was used in this study. We used 10 msec epochs to obtain the F0 contour in accordance to standard practice [1], [12], [16].

Following the computation of the F0 contour, we subsequently aimed to determine the short signal segment that exhibited the smallest mean absolute successive F0 differences (without loss of generality we searched for the best short segment of 2 seconds in duration). For convenience, we will simply use the term jitter later on to refer to the mean absolute successive F0 differences. We remark that alternative definitions of jitter variants (F0 perturbations) are possible [1], [4], [12]; here we wanted to explore the simplest approach.

C. Manual hand-labeling of quasi-stationary segments

We have manually hand-labeled the quasi-stationary segments of the 100 speech recordings by aural and visual inspection (e.g. that the quasi-stationary window appears between 4th to the 12th second). We assessed whether the 2-second segment determined by the proposed segmentation algorithm falls completely within the hand-labeled segments.

D. Acoustic analysis of speech segment

We used the Voice Analysis Toolbox which we had previously developed (open source MATLAB code, available at https://www.darth-group.com/software) for the analysis of sustained vowels [5], [12], [18]. We extracted 307 acoustic features which characterize the speech signal: broadly, these features quantify frequency changes (jitter variants), amplitude changes (shimmer variants), signal-to-noise ratio concepts, F0 variability using wavelets, and envelope modulation. For further information on the acoustic features, their algorithmic expression and their tentative interpretation please refer to the Voice Analysis Toolbox and the cited studies above. These features have been previously explored in detail in our PD work [4], [6], [9], [12].

We applied the algorithmic expressions for the computation of the acoustic features using two different segments for comparison: (i) the segment between 1-3 seconds, and (ii) the automatically determined 2-second segment with the algorithm in this study.

III. RESULTS

Fig. 1 presents an indicative sustained vowel recording and the F0 contour to visually illustrate the result of the segmentation algorithm. As a first step, we verified across all phonations used in the study that the automatically detected segment was indeed a short signal where the F0 variability appeared minimal and matched the hand-labeled quasi-stationary segments. Fig. 2 is the zoomed version of Fig. 1 focusing only on the selected signal segment. We can visually observe

from Fig. 1 that if we had pre-fixed a segment at the middle section of the phonation this would have included some large F0 fluctuations. This problem could have occurred at any point in the phonation, which cautions on the use of pre-fixed time segments for further acoustic analysis.

So far, we have demonstrated that the proposed segmentation algorithm correctly identified a short quasi-stationary segment within a speech recording. The next question is whether this makes any practical difference in the subsequent step with the acoustic characterization of the phonation. Table 1 provides summary statistics across some indicative acoustic features (selected to be representative of different acoustic feature families). We remark that some of the acoustic features exhibit considerable differences in the summary values, which indirectly suggests that this pre-processing segmentation step can have a major influence on the reported results.

IV. Discussion

We have developed a robust algorithmic approach towards detecting the quasi-stationary speech signal segment in sustained vowel /a/ phonations that exhibits the lowest F0 fluctuations. This was achieved by first estimating the F0 contour in 10 msec epochs (which is standard in F0 estimation), and subsequently determining the two consecutive seconds segment that exhibited the lowest jitter. We visually verified that in all cases the algorithm had correctly identified a short signal segment where F0 does not fluctuate considerably (see Fig. 2). Finally, we reported that the signal segment that is passed for further processing affects the computed acoustic features (see Table 1).

Although segmentation is a well-researched area in the signal processing and image processing research literature, we are not aware of any similar work that presents a principled approach towards determining a short speech segment within sustained vowels which would be a useful pre-processing step prior to further acoustic analysis. For example, Badawy et al. [19] attempted to correctly estimate the entire duration of the sustained vowel phonation, whereas here we aimed to determine the most quasi-stationary segment within a full recording. Other work has focused on removing silences in recordings [20] so that only the voiced segment could be presented to acoustic analysis algorithms. We remark that our algorithm can intrinsically automatically detect when the F0 fluctuations are above a maximum threshold of F0 fluctuations or unrealistic F0 ranges (e.g. silence recordings, background noise) and hence identify phonations of insufficient quality, prompting further investigation or rejecting those recordings from further processing.

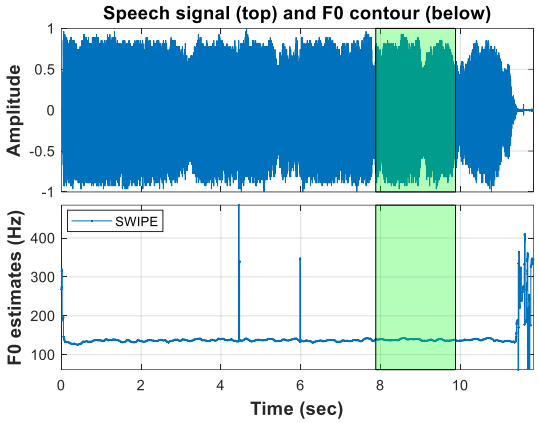

Fig. 1: *Indicative plot visually illustrating the selected signal segment (in transparent green) both in terms of the raw voice signal and the computed F0 contour*.

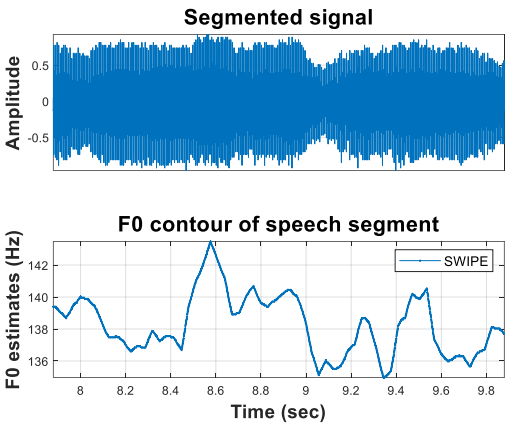

Fig. 2: *Focusing on the segmented signal and the F0 contour (zoomed in version from Fig. 1).*

Table 1: Summary statistics of indicative acoustic features for the phonations used in the study.

Indicative acoustic features	**Benchmark segment (1-3 sec)**	**Automatically determined segment**
Jitter	1.65±2.37	1.33±1.94
Shimmer	0.21±0.07	0.20±0.06
HNR	8.36±10.21	8.49±10.46
GNE	1.47±0.36	1.45±0.40
EMD-ER$_{NSR,TKEO}$	5.87±2.98	7.24±3.70
VFER$_{TKEO}$	0.72±0.59	2.51±1.74

The features are summarized in the form mean±standard deviation. HNR = Harmonics to Noise Ratio, GNE = Glottal to Noise Excitation, EMD-ER = Empirical Mode Decomposition Excitation Ratio, VFER = Vocal Fold Excitation Ratio. For the algorithmic definition of the features in the Table see [12].

This study focused exclusively on sustained vowels /a/ phonations. We remark that in principle these findings should generalize well in other settings with

sustained vowel phonations (e.g. the other two corner vowels /i/ and /u/), but that remains to be tested. So far, we are not aware of any work that has empirically extensively tested F0 estimation algorithms beyond /a/, and future work would likely also need to be done for other vowels or phonetically rich sounds used in clinical practice [1]. A seemingly very different speech signal analysis area to sustained vowels which is, perhaps surprisingly, intrinsically linked is processing of voice fillers. Voice fillers essentially exhibit similar properties to sustained vowels [21] even though they originate in conversational speech, which is a more generic setting where participants are not specifically instructed to produce a specific type of phonation. Previously, we had extracted the corresponding voice fillers for further acoustic analysis manually [21]; in principle, the presented algorithm herein should be generalizable.

We are currently working on extending our early work using the noisy speech data collected as part of the PVI project [13], [14]. This dataset presents considerable challenges because of its large size and the data have not been collected under carefully controlled acoustic conditions. This very challenging setting requires robust methodologies to extract clinically useful information, where automating segmentation and reducing the computations demands on acoustic characterization of phonations is crucial.

Collectively, these results provide a compelling argument that speech segmentation should be carefully considered and reported. This may also have important implications for real-time biomedical signal processing applications (e.g. processing on smartphones), where computational constraints need to be carefully considered. We envisage the proposed algorithm providing a convenient, robust approach to determining a short signal segment from a longer sustained vowel phonation towards standardizing acoustic analysis.

REFERENCES

[1] I. R. Titze, *Principles of voice production*. Iowa City: National Center for Voice and Speech, 2000.

[2] A. Tsanas *et al.*, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, 2012

[3] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, 2010

[4] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *J. R. Soc. Interface*, vol. 8, no. 59, pp. 842–855, 2011, doi: 10.1098/rsif.2010.0456.

[5] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity," in *International symposium on nonlinear theory and its applications (NOLTA)*, 2010, pp. 457–460.

[6] A. Tsanas, M. A. Little, and L. O. Ramig, "Remote assessment of Parkinson's disease symptom severity using the simulated cellular mobile telephone network," *IEEE Access*, vol. 9, pp. 11024–11036, 2021

[7] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 1, pp. 181–190, 2014

[8] S. Arora *et al.*, "Investigating voice as a biomarker for leucine-rich repeat kinase 2-associated Parkinson's disease," *J. Parkinsons. Dis.*, vol. 8, no. 4, pp. 503–510, 2018

[9] S. Arora, C. Lo, M. Hu, and A. Tsanas, "Smartphone speech testing for symptom assessment in rapid eye movement sleep behavior disorder and Parkinson's disease," *IEEE Access*, vol. 9, pp. 44813–44824, 2021

[10] P. Gómez-Vilda *et al.*, "Phonation biomechanics in quantifying parkinson's disease symptom severity," in *Recent Advances in Nonlinear Speech Processing*, vol. 48, 2016, pp. 93–102.

[11] A. Gómez *et al.*, "A neuromotor to acoustical jaw-tongue projection model with application in Parkinson's disease hypokinetic dysarthria," *Front. Hum. Neurosci.*, vol. 15, p. 622825, 2021

[12] A. Tsanas, "Accurate telemonitoring of Parkinson's disease using nonlinear speech signal processing and statistical machine learning," University of Oxford, 2012.

[13] S. Arora, L. Baghai-Ravary, and A. Tsanas, "Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice," *J. Acoust. Soc. Am.*, vol. 145, pp. 2871–2884, 2019.

[14] A. Tsanas and S. Arora, "Biomedical speech signal insights from a large scale cohort across seven countries: The Parkinson's voice initiative study," in *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2019, pp. 45–48.

[15] S. Theodoridis, K. Koutroumbas, *Pattern recognition*, Academic press, 4th ed., 2019

[16] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering.," *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2885–901, 2014

[17] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music.," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–52, Sep. 2008

[18] A. Tsanas, "Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms," in *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2013, pp. 37–40.

[19] R. Badawy *et al.*, "Automated quality control for sensor based symptom measurement performed outside the lab," *Sensors*, vol. 18, no. 4, pp. 1–22, 2018

[20] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB Approach*. Academic Press, 2014.

[21] E. San Segundo, A. Tsanas, and P. Gomez-Vilda, "Euclidean Distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics," *Forensic Sci. Int.*, vol. 270, pp. 25–38, 2017