F0 Variability Measures Based on Glottal Closure Instants

Yu-Ren Chien, Michal Borsky, Jon Gudnason

Center for Analysis and Design of Intelligent Agents, Reykjavik University, Reykjavik, Iceland

yrchien@ntu.edu.tw, michalb@ru.is, jg@ru.is

Abstract

The periodicity of a voiced-sound signal can reflect physiological conditions such as identity, age, and voice disorder. One way to look into this periodicity is to measure the temporal variability of vocal fundamental frequency (F0). This paper proposes 2 measures of F0 variability based on glottal closure instant (GCI). GCI is essential to the detection of F0 when the signal waveform varies substantially between adjacent cycles, e.g., in breathy voice. Frequency-selective variability measurements are taken from an interpolated sequence of fundamental-period values based on GCIs, including certain spectral-shape parameters which constitute a multi-variate measure. The utility of these measures was demonstrated in two experiments designed for inter- and intra-speaker comparisons, respectively.

Index Terms: glottal closure instant, spline interpolation, F0 variability

1. Introduction

The periodicity of signals in one's voice has potential applications in the characterization of voice. For instance, it has been shown in the literature that signal periodicity measures provide effective features for the task of voice quality assessment [1, 2, 3, 4], which aims to estimate the severity of voice disorder from a recording of a person's voice. The variability of fundamental frequency (F0) between adjacent cycles of a quasi-periodic voice signal can provide some information about the aperiodicity of the signal because any thus measured variability apparently adds to the aperiodicity. In view of this, one would be interested in exploring an effective method for measuring F0 variability from a voice signal, which is the objective of this study.

Several F0 variability measures have been presented in the literature over the past decades. Lieberman [5] measured fluctuations in the F0 of normal speech, i.e., jitters, with the magnitude of duration difference between adjacent periods, with regard to the mean duration of periods. Hecker and Kreul [6] differentiated pathological and normal speakers with the maximum rate of F0 change, F0 perturbations, and F0 distribution. Pinto and Titze [7] presented a unifying measure which is based on some common elements of previous jitter measures. More recently, Schoentgen and Guchteneere [8] measured random, short-term perturbations in glottal cycle lengths by applying time-series whitening to the sequence of glottal cycle lengths. Vasilakis and Stylianou [9] estimated jitter as a spectral measurement of the relative movement between two coupled periodic phenomena that model jitter. A survey of acoustic perturbation measures was given by Baken [10].

Many of the aforementioned approaches rely on an assumed quasi-periodicity of analyzed signal to determine F0 values at various voiced time positions across the signal. However, at some low-periodicity voiced time positions, such as those exhibiting breathiness, the overall waveform periodicity could be so low that the estimation of F0 becomes highly unreliable. In this case, the detection of glottal closure instants (GCIs) provides opportunities for a more reliable approach to F0 estimation, which is the foundation of F0 variability measures proposed in this study. GCIs are time instants that each mark the completion of a glottal closure event occurring regularly across pitch cycles, once per cycle [11]. In GCI-based F0 estimation, one focuses on the repetitions of GCI sample points while ignoring cycle-to-cycle waveform variations at other sample points. This promises an improved robustness in F0 estimation and F0 variability measurement because GCIs can usually be detected reliably as long as the speech recording is not substantially corrupted by environmental noise.

The measures proposed in the current study quantify short-term F0 perturbations by using glottal-cycle durations derived from estimated GCIs. F0 variations are represented by a spline-interpolated signal constructed from a non-uniformly-spaced sequence of cycle durations, such that the resulting F0 variation signal has multiple uniformly spaced samples within each cycle. In addition to a variability value measured as a magnitude of energy in the F0 variation signal, several shape parameters are extracted from the F0 variation spectrum to give a multi-dimensional measure that quantifies not only the range of F0 variations, but also the spectral characteristics of F0 variations, such as the rate of F0 change.

2. F0 Variability Measures

Two measures are presented in this section: One represents the analyzed utterance as a sequence of audio frames and, for each time frame, quantifies the magnitude and spectral energy distribution of the observed instantaneous F0 variations. The other measure quantifies the total variability over the utterance without a time-frequency localization.

2.1. Frame-Based Spectral Measure

The measure considers a sequence of uniformly spaced analysis time positions distributed all over the analyzed utterance, and produces a measurement for each of these time positions. Here we use a hop size of 0.1 s for these time positions; i.e., any two adjacent analysis time positions are spaced 0.1 s apart. The typical hop size of 10 ms for speech processing is not used here because it would imply use of a short analysis time frame where possibly only one or two voiced-sound cycles can be observed. A hop size much larger than 0.1 s would be excessive as related to the time scale in which individual voiced sounds are usually sustained in continuous speech.

For each analysis time position, the measurement is based on calculating the spectrum of a fundamental-period signal derived from a short sequence of GCIs around the analysis time position. Here the number of GCIs analyzed is set to 18, which spans a time interval on the same order of magnitude as the hop size because the spanned interval ranges from 0.068 s to 0.17

Table 1: *Significance proportions of several F0 variability measures/parameters tested on speaker pairs. FSM = Frame-Based Spectral Measure; TVM = Total Variability Measure.*

FSM-Magnitude	FSM-Centroid	FSM-Spread	FSM-Skewness	FSM-Kurtosis	TVM	Jitter
0.71	0.55	0.52	0.51	0.40	0.75	0.77

s for a typical vocal F0 between 100 Hz and 250 Hz. Let the analysis time frame be denoted by $\{n_i\}_{i=1}^{N_c}$, where n_i denotes the ith GCI (measured in samples) in this frame, and $N_c = 18$. This sequence of GCIs gives, at the time position of each GCI, an estimate d_i of the instantaneous fundamental period:

$$d_i = n_{i+1} - n_i,\ i = 1, ..., N_c - 1. \tag{1}$$

Towards a representation of fundamental-period variations that is invariant to the average pitch level of the time frame, we calculate a normalized set of fundamental-period values by dividing each of the raw values $\{d_i\}_{i=1}^{N_c-1}$ by their median value. With these normalized values being regarded as samples taken (at the GCIs) from a continuous-time fundamental-period signal, we calculate a uniformly sampled discrete-time representation for this signal by interpolating (with a cubic spline) the 17 normalized values at 128 new time samples that are spaced uniformly across the 16 cycles inbetween. Thus, each cycle is represented by 8 new time samples on average, giving a cycle-synchronous representation of fundamental-period variations that is invariant to the average pitch. After removing the signal mean, the 128-point zero-mean fundamental-period signal is Hamming-windowed and discrete Fourier-transformed to calculate its squared magnitude spectrum, from which the spectral energy distribution of fundamental-period variations can be examined.

Five parameters of F0 variability are calculated from the fundamental-period (squared magnitude) spectrum for each analysis time position: magnitude, centroid, spread, skewness, and kurtosis. The magnitude of F0 variability is given by summing over all the non-DC frequency bins. The other parameters are based on a normalized version $\{D_k\}_{k=1}^{N_b}$ ($N_b = 64$) of the spectrum, which is calculated by dividing each frequency component by the magnitude of F0 variability. The normalized spectrum shares its form with probability mass functions, to which the definitions of mean (for the centroid), standard deviation (for the spread), skewness, and kurtosis can be applied to characterize the spectral shape [12, 13]. The spectral centroid μ is calculated as a mean:

$$\mu = \sum_{k=1}^{N_b} kD_k. \tag{2}$$

The spectral spread σ is calculated as a standard deviation:

$$\sigma = \sqrt{\sum_{k=1}^{N_b} k^2 D_k - \mu^2}. \tag{3}$$

The spectral skewness γ is given by

$$\gamma = \frac{\sum_{k=1}^{N_b} k^3 D_k - 3\mu\sigma^2 - \mu^3}{\sigma^3}. \tag{4}$$

The spectral kurtosis κ is calculated as

$$\kappa = \frac{\sum_{k=1}^{N_b} (k-\mu)^4 D_k}{\sigma^4}. \tag{5}$$

2.2. Total Variability Measure

In the total variability measure (TVM), an estimate of instantaneous fundamental period is again derived from any two adjacent GCIs as a difference between them, but only one scalar variability measurement is produced from a specific segment of voiced sound. Outlier period estimates are detected using the median absolute deviation method and replaced with local averages. A uniformly sampled instantaneous-period signal is constructed by interpolating the initial period estimates at sample time positions spaced with a rate of 2 kHz. The chosen rate of 2 kHz is intended to be much higher than the F0, allowing the interpolation to capture all important trends in the evolution of F0. Fig. 1 illustrates the process on a short segment from sustained phonation.

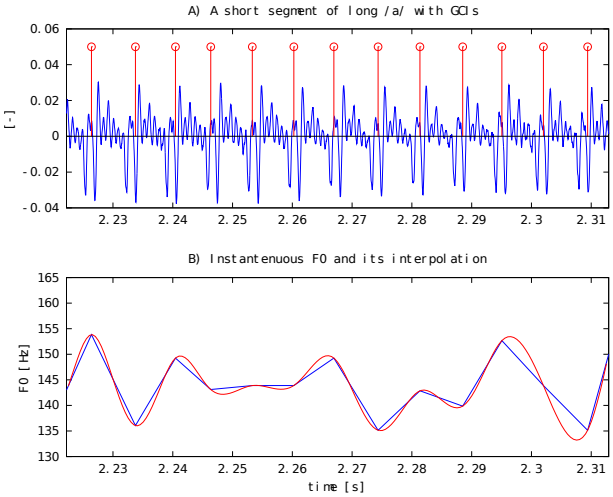

Figure 1: *An example of instantaneous F0 estimation for a short segment of long /a/.*

With the interpolated instantaneous-period signal being regarded as a superposition of a low-frequency trend and a high-frequency fluctuation, the measure decomposes the signal into the two components using low-pass and high-pass filters with a cut-off frequency of 6 Hz. The total variability is calculated as an energy ratio (in percent) of high frequency to low frequency over the analyzed segment of voiced sound. Since the average duration of vowel /a/ is 97.3 ms as measured from a set of read-speech utterances phonetically aligned with text, it is believed that the high-pass filter serves to remove the energy components associated with the mean F0, as well as an F0 trend that could be observed on an inter-phonemic time scale in continuous speech. The low-frequency component represents relatively controlled F0 variations, whereas the high-frequency component is intended to capture any involuntary F0 variability that can relate to a pathological condition in one's voice box.

3. Experimental Procedure

We demonstrate the utility of the proposed F0 variability measures by carrying out two experiments. One experiment compares F0 variability measurements between speakers. The other experiment compares sustained-vowel measurements against read-speech measurements. GCIs were extracted with the YAGA algorithm [11] from utterances of sustained /a/ and read speech, which were used in these experiments. Vowel segments in the reading utterances were isolated for F0 variability measurement by means of forced alignment of each reading utterance with its corresponding Icelandic text. To this end, an automatic speech recognition (ASR) system trained for Icelandic was used [14, 15].

3.1. Inter-Speaker Comparison

To see the potential of the proposed F0 variability measures for distinguishing different voice qualities occurring between individual speakers, we collected 884 pairs of read-speech utterances, each pair produced by two speakers reading the same Icelandic text paragraph. Different utterance pairs were produced by different pairs of speakers, with the read text shared among all utterance pairs. The pairs of speakers are collectively composed of 1,028 speakers.

Since a common text underlies all the reading utterances in this experiment, each utterance can be divided into a fixed number of segments that correspond to the same phoneme sequence derived from the text, which is achieved by forced alignment of the text with each utterance. With this common segmentation structure, we analyze each utterance by calculating an F0 variability measurement for each of the 267 vowel segments in the utterance, and compare the measurements between any two speakers by applying a paired significance test to all the measurements paired across the two utterances (267 pairs) to give a p-value. The test used here is the paired, two-sided Wilcoxon signed rank test, which checks if the median difference between the first sample data and the second sample data is nonzero with statistical significance. To measure the performance of each F0 variability measure in distinguishing the voice quality of one speaker from that of the other in each speaker pair, we calculate a *significance proportion* over all the 884 speaker pairs, which is defined as the proportion of speaker pairs for which the resulting p-values are less than 0.05.

In this experiment, we compare results among the frame-based spectral measure, the total variability measure, and Praat's "ppq5" jitter [16]. For the frame-based spectral measure, each vowel-segment measurement is calculated by taking the median value of frame-wise measurements over all the analysis time positions falling within the vowel segment. For the other two measures, each vowel-segment measurement results from applying the measure directly to the signal segment (and the GCI sub-sequence for the total variability measure). In case that a vowel segment is too short for Praat to produce a jitter value, the analyzed signal segment is successively dilated in length by a factor of 2 until the dilated segment is sufficiently long.

3.2. Comparison Between Sustained Vowels and Continuous Speech

The second experiment focused on comparing TVM measurements from utterances of sustained /a/ and from occurrences of vowel /a/ in reading utterances. The average durations of the two types of utterances across the database were 4 and 64 seconds, respectively. For each reading utterance, the experiment analyzed only the vowel segments which were classified by the ASR as phonetically identical to the vowel from sustained phonation. This approach allowed us to account for alternative pronunciations of the same grapheme. A median measurement was calculated for each reading utterance from all its available segments of vowel /a/. The comparison was carried out for 1,443 speakers. We compared the two types of measurements by calculating their respective modes over all the speakers.

4. Results

Significance proportions from the inter-speaker comparisons are listed in Table 1. Since the jitter, the total variability measure, and the magnitude parameter of the frame-based spectral measure all concern the total magnitude of F0 variability, it is of particular interest to compare results among these 3 parameters. All the significance proportions for the 3 parameters are between 0.7 and 0.8. Whereas the difference between jitter and total variability is 0.02, the proportion for the frame-based magnitude is lower than for the total variability measure by a gap of 0.04. The relatively low significance proportion for the frame-based measure suggests that the 18-GCI frame size may have been so large for some vowel segments in some utterances, that an analysis time frame can sometimes include a substantial amount of non-voiced signal in addition to an intended vowel segment. This in turn suggests possible performance optimization for this measure in future investigation through a reduction of frame size. Regarding the 4 spectral-shape parameters, it is encouraging that they are each capable of distinguishing at least 40% of the speaker pairs. This suggests that not only the magnitude of F0 variability, but also the spectral characteristics of F0 variations, are useful in distinguishing personal voice qualities. When the 5 frame-based parameters are used altogether as a multi-dimensional measure, they will promise a better characterization of F0 variability than any one of them, in that they are orthogonal to each other in the sense that they describe the area and 4 shape dimensions of the spectral curve respectively. This is in stark contrast to the traditional scalar representation of F0 variability achieved by jitter measurement.

In terms of the centroid parameter, an example analysis is presented here for the speaker pair with the lowest p-value among all the 884 pairs, which is $p = 0.0$. For a majority of vowel segments, Speaker 1 of this pair exhibited a higher median centroid value than Speaker 2. To illustrate this comparison, a vowel segment is selected from the 267 vowel segments, such that its difference in centroid value is the 134th highest, which is $4.7 - 3.7 = 1.0$. For this segment, Speaker 1's median centroid value 4.7 is calculated from only one time frame, while Speaker 2's value 3.7 is calculated from two time frames (with centroid values 3.6 and 3.8). The lengths of segment are 0.12 s and 0.18 s respectively for Speakers 1 and 2. Note that the unwanted effect of analysis frames including some signal outside the vowel segment could be alleviated by the use of Hamming window, which focuses the analysis on the central part of each time frame to a certain degree. Normalized fundamental-period signals for the time frames with centroid values 4.7 (from Speaker 1) and 3.6 (the first frame from Speaker 2) are shown in Fig. 2, where Speaker 1 exhibits a faster change in fundamental period resulting in the higher value of spectral centroid.

Histograms of TVM measurements from sustained-vowel and reading utterances are plotted in Fig. 3. Both the histograms show a distribution that is approximately log-normal. Even so,

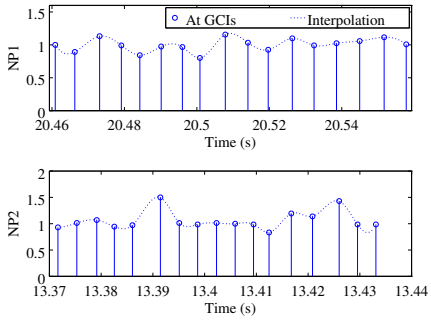

Figure 2: *Normalized fundamental-period signals plotted for the lowest-p-value speaker pair in terms of the centroid parameter in the frame-based spectral measure. NP1/2 = Normalized Fundamental-Period Signal From Speaker 1/2.*

if we examine how many speakers had a sustained-/a/ or reading measurement that is lower than 0.5, it can be seen that more speakers had such a low TVM measurement from sustained /a/ than from reading. The modes of the two types of measurements were 0.45 for sustained /a/ and 0.65 for reading. In other words, sustained /a/ exhibited a lower F0 variability than the /a/ segments in read speech. Moreover, a positive mean difference of 0.17 was obtained by subtracting each read-speech measurement from its sustained-phonation counterpart, which could have resulted from a small number of speakers having an exceptionally high TVM measurement from sustained vowel. This dominance by outliers was resolved when we normalized each individual difference by the reading measurement, which gave a negative mean of -38%.

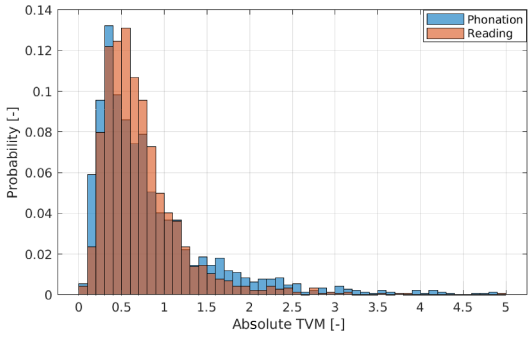

Figure 3: *Histograms of total variability measure (TVM) for sustained phonation and read speech.*

5. Conclusions

Two measures of F0 variability have been presented in this paper, which estimate the evolution of F0 from detected GCIs. Results of inter-speaker comparisons showed that the proposed measures, as well as an existing F0 variability measure, were able to distinguish over 70% of the tested speaker pairs. Results of intra-speaker comparisons showed that sustained-vowel utterances tend to have a lower total variability than read speech. As a direction for future study, use of the proposed spectral-shape parameters in conjunction with the magnitude of variability can be investigated in applications such as voice quality assessment and speaker recognition.

6. References

[1] R. A. Prosek, A. A. Montgomery, B. E. Walden, and D. B. Hawkins, "An evaluation of residue features as correlates of voice disorders," *Journal of Communication Disorders*, vol. 20, no. 2, pp. 105–117, 1987.

[2] V. I. Wolfe, D. P. Martin, and C. I. Palmer, "Perception of dysphonic voice quality by naive listeners," *Journal of Speech, Language, and Hearing Research*, vol. 43, pp. 697–705, 2000.

[3] Y. Maryn, P. Corthals, P. V. Cauwenberge, N. Roy, and M. D. Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels," *Journal of Voice*, vol. 24, no. 5, pp. 540–555, 2010.

[4] S. N. Awan, N. Roy, M. E. Jetté, G. S. Meltzner, and R. E. Hillman, "Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V," *Clinical Linguistics & Phonetics*, vol. 24, no. 9, pp. 742–758, Sep. 2010.

[5] P. Lieberman, "Perturbations in vocal pitch," *The Journal of the Acoustical Society of America*, vol. 33, no. 5, pp. 597–603, 1961.

[6] M. H. L. Hecker and E. J. Kreul, "Descriptions of the speech of patients with cancer of the vocal folds. part I: Measures of fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 49, no. 4B, pp. 1275–1282, 1971.

[7] N. B. Pinto and I. R. Titze, "Unification of perturbation measures in speech signals," *The Journal of the Acoustical Society of America*, vol. 87, no. 3, pp. 1278–1289, 1990.

[8] J. Schoentgen and R. de Guchteneere, "Time series analysis of jitter," *Journal of Phonetics*, vol. 23, no. 1, pp. 189–201, 1995.

[9] M. Vasilakis and Y. Stylianou, "Spectral jitter modeling and estimation," *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 183–193, 2009.

[10] R. J. Baken, *Clinical measurement of speech and voice*. Boston: College-Hill Press, 1987.

[11] M. R. P. Thomas, Jon Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[12] E. H. Vanmarcke, "Properties of spectral moments with applications to random vibration," *ASCE Journal of the Engineering Mechanics Division*, vol. 98, no. 2, pp. 425–446, 1972.

[13] R. F. Dwyer, "Detection of non-Gaussian signals by frequency domain kurtosis estimation,," in *Proc. ICASSP*, 1983.

[14] Anna B. Nikulásdóttir, Inga R. Helgadóttir, Matthías Pétursson, and Jón Guðnason, "Open ASR for Icelandic: Resources and a baseline system," in *Proc. LREC 2018*, 2018.

[15] Inga Run Helgadottir, R. Kjaran, Anna Bjork Nikulasdottir, and Jon Gudnason, "Building an ASR corpus using Althingi's parliamentary speeches," in *Proc. Interspeech 2017*, 2017, pp. 2163–2167. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-903

[16] P. Boersma, "Should jitter be measured by peak picking or by waveform matching?" *Folia Phoniatrica et Logopaedica*, vol. 61, pp. 305–308, 2009.