Using Windowed Relative Deviation to Detect Possible Voice Pathology

J. Brandon Laflen, Ph.D., Cathy L. Lazarus, Ph.D., and Milan R. Amin, M.D.

Abstract— **A diagnostic method is presented that provides for analyzing pitch "jitter" in running speech. "Jitter" is typically measured with explicit voice tasks, namely sustained vowel phonation. However, some voice pathologies cannot be detected with sustained phonation. Further, it is not possible to ensure explicit voice productions from certain patients, including pediatric populations. In contrast, windowed relative deviation reports instantaneous pitch "jitter" as well as the overall "jitter" statistic commonly reported. Also, the width of the analysis window is related to the rate of pitch deviation, which provides a unique form of selectivity. Voice productions from a normal adult speaker and from an adult speaker with a known voice pathology were analyzed with this method. Voice productions from the normal speaker exhibited less than 1% pitch deviation during phonetic portions of the signal that were akin to sustained phonation. On the other hand, the speaker with a known pathology exhibited greater than 10% pitch deviation at quasi-periodic intervals within sustained phonation.**

I. INTRODUCTION

The evaluation of voice disorders is made difficult by a lack of objective voice measures with which to rate voice quality compared to normals. "Jitter" is one such objective measure that reports the average relative deviation of voice pitch. If the voice pitch at time t is $p(t)$ and the signal interval is $t \in [0, T]$, then "jitter" is

$$J = \frac{1}{\bar{p}\,T} \int_0^T |p(t) - \bar{p}|\, dt, \tag{1}$$

where $\bar{p} = 1/T \int_0^T p(t)dt$ is the mean pitch. Since "jitter" reports deviation from mean, it is effective regardless of the actual pitch frequency. Unfortunately, such measures require explicit voice production, such as sustained phonation for computing "jitter," which limits their clinical usefulness. In the clinic, it is not always possible to control against changing phonation in speech, often because the patient is not reliable, which is especially true in pediatric populations. Further, some effects can only be realized during transition, such as a vocal muscle spasm during a pitch change. It is important to control the recording interval to only capture the relevant voice behavior.

In contrast, perceptual ratings by trained speech-language pathologists can evaluate voice quality [1], [2]. Such ratings can serve as an indication of vocal tract or voice production pathologies. However, perceptual ratings do not provide an objective standard for normalizing data and may not offer an indication of the underlying pathology. Further, perceptual rating scales do not offer resolution for tracking incremental changes in pathological parameters, which can be useful for evaluating treatment efficacy.

This paper presents the windowed relative deviation spectrum as an alternative to the "jitter" constant. This spectral representation contains the "jitter constant" but also represents finer detail within smaller intervals of the pitch signal.

II. THEORY

Eq. 1 forumulates "jitter" over the entire interval of the pitch signal, $t \in [0, T]$. However, this statistic can be measured over any window interval of width w and center c, such that $0 \le c - w/2 \le t \le c + w/2 \le T$. The following define the windowed "jitter" over any such window:

$$\bar{p}_{c,w} = \frac{1}{w} \int_{c-w/2}^{c+w/2} p(t)dt \tag{2}$$

$$J_{c,w} = \frac{1}{\bar{p}_{c,w}\, w} \int_{c-w/2}^{c+w/2} |p(t) - \bar{p}_{c,w}|\, dt. \tag{3}$$

Thus, $J_{c,w}$ reports the relative deviation of the pitch signal within the window. As $w \to T$, this statistic approaches the coefficient of eq. 1; but for smaller w the statistic tends toward a more "instantaneous" measure of relative pitch deviation. For smoothly varying signals, the relative pitch deviation will approach zero as $w \to 0$, since $\lim_{w\to 0} \bar{p}_{c,w} = p(c)$ (see eq. 2).

Conceptually, windowed relative deviation (WRD) reports the extent that a signal varies within the interval from its mean and should therefore be sensitive to signal transitions. Two examples help illustrate this behavior. First, consider a signal changing proportionally in magnitude to its current magnitude over the window interval, such as $p(t) = A\exp(\alpha t)$. The WRD (eq. 3) is

$$J_{c,w} = \frac{2}{\alpha w}\left(\ln\left(\frac{2\sinh(\alpha w/2)}{\alpha w}\right) - 1\right) + \coth\left(\frac{\alpha w}{2}\right). \tag{4}$$

Notice that this expression does not depend upon A, is constant for all c, and increases with increasing α. Thus, the statistic is proportional to the extent of signal change but not to absolute magnitude. If $x = \alpha w/2$, then eq. 4 can be rewritten in terms of just the argument x: $J_{c,x2/\alpha} = (\ln(\sinh(x)/x) - 1)/x + \coth(x)$. Fig. 1 plots this function for this normalized parameter. Note that the deviation is approximately 1/2 when $w = 2/\alpha$ and asymptotically approaches 2 with increasing w. Windows with $w \ll 2/\alpha$ will report near-zero relative deviation, while windows with $w \gg 2/\alpha$ will

This work was not supported by any organization

JB Laflen is with the Faculty of Otolaryngology, Physiology and Neuroscience, and Mathematics, New York University, 550 First Ave., NBV 5 East 5, New York, NY 10016 j.brandon.laflen@ieee.org

CL Lazarus and MR Amin are with the Faculty of Otolaryngology and the NYU Voice Center, New York University, 550 First Ave., NBV 5 East 5, New York, NY 10016

1-4244-0033-3/06/$20.00 ©2006 IEEE.

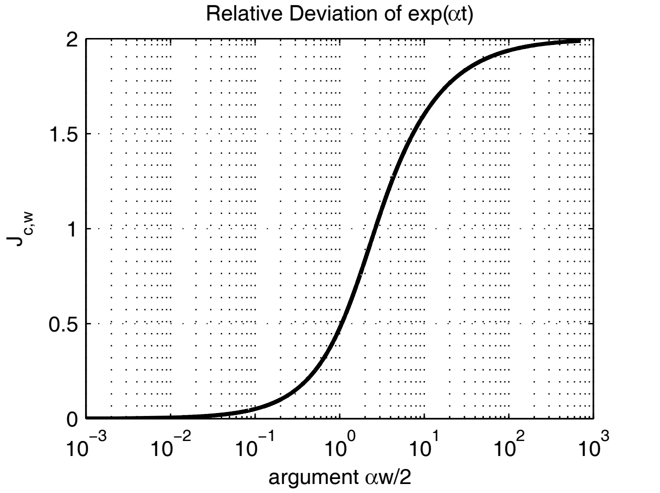

Fig. 1. The relative deviation of eq. 4 for normalized parameter $x = \alpha w/2$.

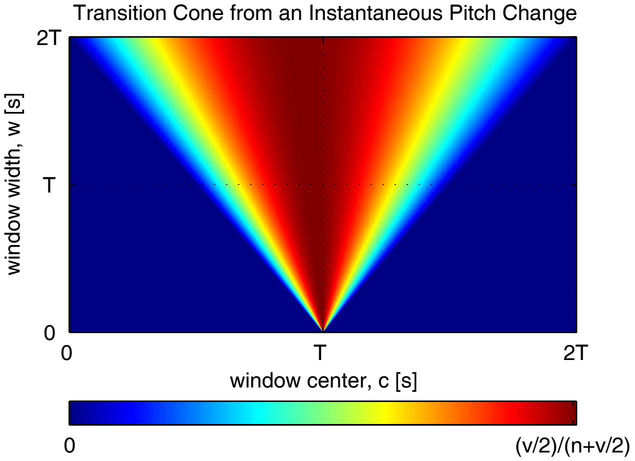

Fig. 2. The transition cone of an instantaneous pitch change (eq. 5).

approximately clip to a relative deviation of two. Essentially, w selects for α, the proportional rate of change. Second, consider the signal $p(t) = \eta + \nu \cdot 1_{\{t \geq \tau\}}$ (an instantaneous step from η to $\eta + \nu$ at time τ). If the window interval is $c - w/2 \leq \tau \leq c + w/2$, then

$$J_{c,w} = \frac{\left(1-2\left(\frac{c-\tau}{w}\right)\right)\left(1+2\left(\frac{c-\tau}{w}\right)\right)}{1+2\left(\frac{c-\tau}{w}\right)+\frac{2\eta}{\nu}}. \tag{5}$$

The WRD increases from zero at $c = \tau - w/2$, to a maximum at $c = \tau$, and then decreases back to zero at $c = \tau + w/2$ (it is zero $\forall |c - \tau| > w/2$). The maximum at $c = \tau$ is $(\nu/2)/(\eta + \nu/2)$, regardless of w; this is the ratio of the step-change to the average of the two signal values before and after the step. If the WRD is visualized as a two-dimensional spectrum with horizontal abscissa c and vertical abscissa w, the transition interval is always centered at $c = \tau$ while the width of the interval is w; this is the cone-shape illustrated in fig. 2. In general, the WRD depicts signal transitions as cones in the spectrum that are proportional to the extent of the signal transition.

III. METHOD

Voice productions for this project were recorded in a digital format. The implementation of eq. 3 for digital voice recordings is subdivided into the following components: pitch extraction, computing the spectrum, and displaying the spectrum in a diagnostically meaningful format.

A. Voice Recordings

Three specific voice tasks were performed for the construction of a normative database: /a/, /aba/, and "How are you?" Voice productions were recorded from subjects in quiet at the NYU Voice Center using the KayPENTAX Computerized Speech Lab 4500. The signals were digitized with a 50 kHz sampling rate using a high-quality A/D converter which included a front-end low-pass (anti-aliasing) filter with cutoff below 20 kHz (at 20 kHz, signal power was below -70 dB relative to signal peak).

B. Pitch Extraction

Pitch was estimated from the digitized voice recording as the fundamental frequency reported by windowed autocorrelation. Pitch was restricted to the range from 20 Hz to 1 kHz. Therefore, prior to pitch extraction, the voice recording was downsampled from the 50 kHz sampling rate to a more tractable 8 kHz sampling rate. Downsampling was performed with the Fast-Fourier Transform (FFT) by removing elements above 4 kHz and reconstructing a conjugate symmetric spectrum up to 8 kHz. The downsampled signal was acquired from this new spectrum through the inverse FFT. A 50 ms hamming window was applied to the downsampled signal and the FFT was taken on the resulting segment. The FFT was oversampled (zero padding) to ensure spectral sampling of not less than one sample per Hz. The power spectrum was computed as the magnitude-squared of the FFT signal – then the inverse FFT was taken to produce the autocorrelation signal. A cubic spline was fit to the autocorrelation signal within the interval $1/1000 \leq t \leq 1/20$ (limiting the pitch to the frequency range between 20 Hz and 1 kHz) and the time-location of the maximum peak of the spline was located. The fundamental frequency, and therefore the estimated pitch, was taken as the inverse of this time-location. This process was repeated over the entire digitized voice recording with a 95% window overlap. Panels A, B, and C of fig. 3 illustrate the intermediate results of this process.

C. Computing the Windowed Relative Deviation Spectrum

Since the pitch signal provided by the above method was discrete, a discrete analogue to eqs. 2 and 3 was used to compute the windowed relative deviation spectrum. Let $p(t) = \tau \sum_n p[n]\delta(t - n\tau)$ be the discrete pitch signal, where τ is the interval between samples. Further, let $W = w/\tau$ and $C = c/\tau$, with $\{C, W\} \in \mathbb{Z}^2$, be the discrete window width and center parameters, respectively. Then eqs. 2 and

3 respectively become

$$\bar{p}[C,W] = \sum_{n=\lfloor C-W/2+1\rfloor}^{\lceil C+W/2-1\rceil} \frac{p[n]}{2W} + \sum_{n=\lceil C-W/2\rceil}^{\lfloor C+W/2\rfloor} \frac{p[n]}{2W} \quad (6)$$

$$J[C,W] = \sum_{n=\lfloor C-W/2+1\rfloor}^{\lceil C+W/2-1\rceil} \frac{|p[n]-\bar{p}[C,W]|}{2W\bar{p}[C,W]} + \sum_{n=\lceil C-W/2\rceil}^{\lfloor C+W/2\rfloor} \frac{|p[n]-\bar{p}[C,W]|}{2W\bar{p}[C,W]} \quad (7)$$

(this form arises from edge effects when W is even, resulting in a half-integral over the impulse area at the edges).

The collection $\{J[C,W]\}$ was computed over the valid range of the input signal. If the pitch signal contained N samples ($1..N$), then legal values for W were $1..N$ and C was allowed integer values such that $1+\lfloor W/2\rfloor \leq C \leq N-\lfloor W/2\rfloor$. $J[C,W]$ was assigned a value of zero for illegal $\{C,W\}$. Panel D of fig. 3 illustrates the resulting spectrum.

D. Visualizing the Windowed Relative Deviation Spectrum in a Diagnostically Relevant Format

Although the collection $\{J[C,W]\}$ can be visualized as a two-dimensional spectrum with horizontal abscissa C and vertical abscissa W, it was transformed and presented in a more diagnostically relevant format. Generally, a voice pathology (such as vocal tremor) is diagnosed if a patient presents with overall "jitter" (see eq. 1) of 1% or greater (*i.e.*, $J \geq 0.01$). For more subtle pathology, "jitter" may be smaller than 1%. Thus, the windowed relative deviation (WRD) was more diagnostically meaningful when transformed with a logarithm to highlight subtle changes:

$$J_D[C,W] = \log_{10}(J[C,W] \times 100\%). \quad (8)$$

In this format, values in the range $[0..1]$ represent a 1% to 10% deviation, $[-1..0]$ a 0.1% to 1% deviation, etc., such that deviations on the order of 0.1% were represented with the same dynamic range as deviations on the order of 10%.

The resulting spectrum was topologically complex and required a visual enhancement to highlight diagnostically relevant features. Even normal subjects would exhibit pitch transition at some small level during sustained phonation (/a/), while the more complex voice productions (/aba/ and "How are you?") implicitly contained pitch transitions. Since a spectral cone occurred whenever there was a pitch transition, this lead to a topologically complex surface of distinct or overlapping transition cones of varying amplitude, proportional to the extent of each pitch transition. For diagnostic purposes, it was most useful to highlight certain features, namely 0.001%, 0.01%, 0.1%, 1%, and 10% deviation, so the topology was presented as a contour map with elevation lines at those points. Panel E of fig. 3 illustrates the resulting contour-log spectrum.

Finally, a subset of the valid WRD spectrum was presented. Since phonemes do not typically exceed 250 ms, larger window lengths (*i.e.*, $w = W\tau \geq 250$ ms) were not presented in the final spectrum. This allowed the spectrum to emphasize the pitch deviation within a phoneme, which was especially important for voice productions that contained multiple phonemes. Panel F of fig. 3 illustrates the final contour-log percent deviation spectrum.

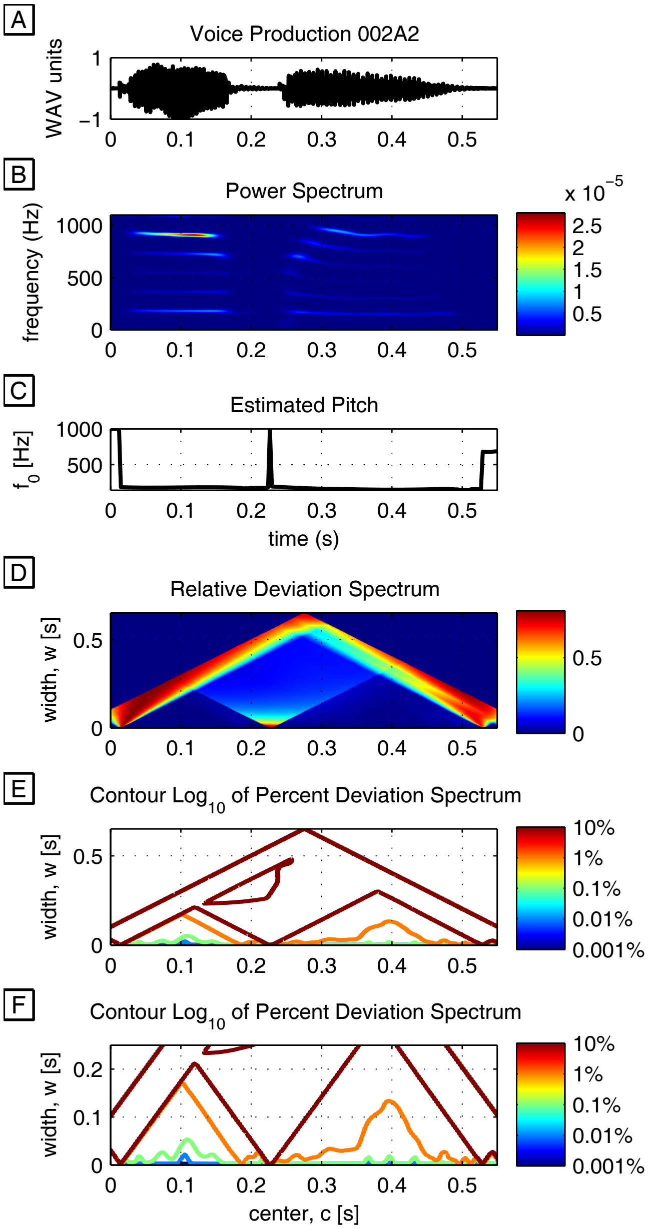

Fig. 3. The entire computational process of the windowed relative deviation (WRD) spectrum, displayed for a normal adult subject voicing /aba/. Note that the horizontal abscissa of the first three panels is time [s], while it is window center c [s] for the final three panels. A: original signal, /aba/. B: computed power spectrogram for a 50 ms moving hamming window. C: estimated pitch. D: WRD spectrum, $J[C,W]$, where $C = c/\tau$ and $W = w/\tau$ (see eq. 7). E: contour map of $\log_{10}(J[C,W] \times 100\%)$. F: contour map, with vertical axis scaled to only display window widths $0 \leq w \leq 250$ ms.

IV. RESULTS

This paper presents preliminary spectra from one normal adult speaker and one adult speaker with a known voice pathology. Currently, normative data is being collected and

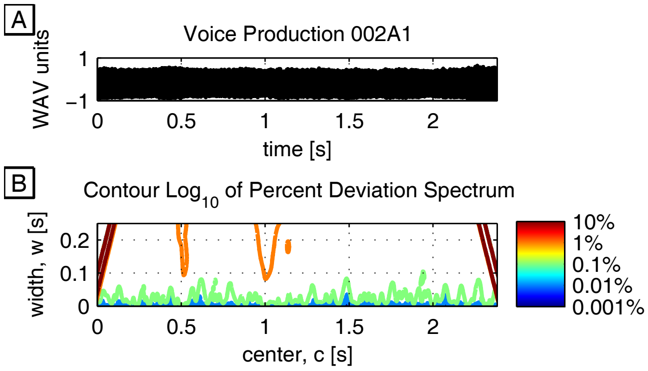

Fig. 4. Voice production of /a/ by the same subject as in fig. 3. Note that only the voice production (A) and the final spectrum (B) are displayed.

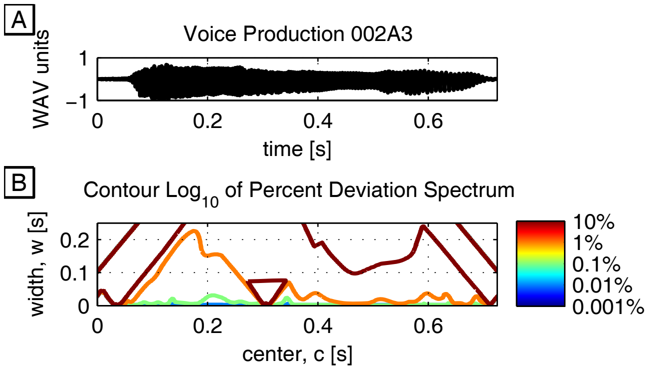

Fig. 5. Voice production of "How are you?" by the same subject as in figs. 3 and 4. Note that only the voice production (A) and the final spectrum (B) are displayed.

will include 20 normal adult speakers, 20 adult speakers with a known voice pathology, 20 normal infants, and 20 infants with a known voice pathology (in the case of infants, cry data is analyzed). Figs. 3, 4, and 5 respectively depict the contour log windowed percent deviation spectra from the normal speaker for /aba/, /a/, and "How are you?" Fig. 6 depicts the spectrum from the speaker with a known voice pathology for the sustained phonation /a/. In these spectra, orange lines indicate the locations where percent deviation crosses 1%, which is the common indicator of pathology for the "jitter" constant in sustained phonation (eq. 1). Red lines indicate $\geq 10\%$ deviation.

V. DISCUSSION

Although the deviation spectra from the normal speaker contain fluctuations, they are markedly different from the deviation spectrum of the speaker with a known voice pathology. Fig. 4 depicts the deviation spectrum of a sustained phonation by the normal speaker. Note that for window widths smaller than 100 ms, deviations are on the order of 0.1% (green lines) to 1% (orange lines). This is also generally true for /aba/ (fig. 3) during the vowels between 25 ms and 175 ms, and between 250 ms and 500 ms; and the beginning of "How are you?" (fig. 5) followed by a 10% pitch deviation cone at approximately 300 ms, which corresponds to the beginning of "you?" (After this cone the upwards pitch intonation of the question generates $\geq 1\%$ pitch deviation.) In contrast, the spectrum from the sustained phonation of the speaker with a known voice pathology contains deviations often in excess of 10% for widths below 100 ms. While the percent deviations $\geq 1\%$ indicate the speaker's known pathology, it is worth noting that the interval between these sharp transitions is often on the order of 100 ms and may be connected with the underlying physiological abnormality producing the voice pathology. Aside from these transitions, the 0.1% deviation structure (green lines) appears similar to the same topological feature in the normal speaker.

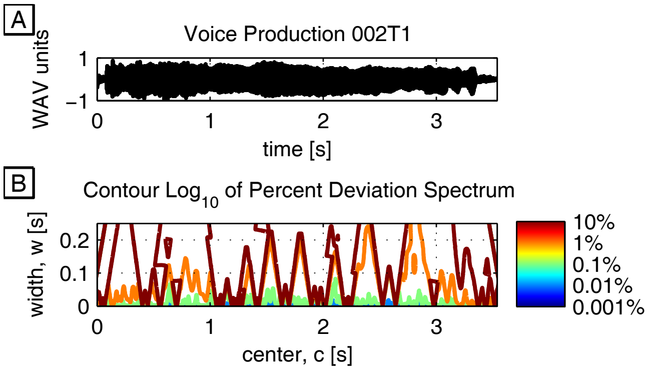

Fig. 6. Voice production of /a/ by an adult subject with a known voice pathology. Note that only the voice production (A) and the final spectrum (B) are displayed. Orange lines indicate the locations where percent deviation crosses 1%, the common indicator of pathology for the "jitter" constant in sustained phonations (eq. 1). Red lines indicate $\geq 10\%$ deviation.

VI. CONCLUSIONS AND FUTURE WORK

The presented spectra demonstrate the potential diagnostic value of the windowed relative deviation spectrum. Especially when visualized as a contour map of $\log_{10}$ percent deviation, the spectrum reveals the time locations and extent of pitch transitions, intonations, and more subtle structures that may be related to human voice control mechanisms. Spectral visualization of pitch deviation also provides a diagnostic tool for analyzing "jitter" in running speech, which potentially removes the constraint of sustained phonation while simultaneously allowing exploration of more subtle abnormalities in a more natural context.

Future work includes the analysis of a normative database to establish objective measures in running speech. In addition to possible effectiveness for adult patients, because this method does not require specific voice productions there is the potential to use this tool in infant populations as an early, non-invasive diagnostic method of certain voice pathologies.

REFERENCES

[1] E.M. Konst, T. Rietveld, H.F. Peters, and H. Weersink-Braks. "Use of a perceptual evaluation instrument to assess the effects of infant orthopedics on the speech of toddlers with cleft lip and palate." *Cleft Palate-Craniofacial Journal.* 40(6):597-605, 2003.

[2] A.L. Webb, P.N. Carding, I.J. Deary, K. MacKenzie, N. Steen, and J.A. Wilson. "The reliability of three perceptual evaluation scales for dysphonia." *European Archives of Oto-Rhino-Laryngology.* 261(8):429-34, 2004.