An Acoustic Method

to Quantify Perceived Vocal Tremor

Phonology Project Technical Report #16

Marios Fourakis

July, 2010

Phonology Project, Waisman Center,
University of Wisconsin-Madison

Preparation of this report was supported by research grant DC00496 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health (Lawrence D. Shriberg, Principal Investigator)

Background

Phonology Project Technical Reports provide technical and substantive information on methods developed for a program of research in speech sound disorders of known and unknown origins. Primary goals of the Phonology Project are to identify etiologic origins, risk and protective factors, and diagnostic markers for eight subtypes of speech sound disorders of currently unknown origin (Shriberg, 2010).

The diagnostic instrument used in all Phonology Project studies is termed the Speech Disorders Classification System (SDCS: Shriberg et al., in press). The SDCS includes a typologic nosology for research and practice in speech sound disorders, and an etiologic nosology for the eight putative subtypes of speech sound disorders of currently unknown origin. Data reduction includes both perceptual and acoustic methods. Perceptual methods for narrow phonetic transcription of speech are based on extensions to the system described in *Clinical Phonetics* (Shriberg & Kent, 2003). Perceptual methods to code speakers' prosody and voice are based on extensions to the system described in The Prosody-Voice Screening Profiles (PVSP: Shriberg, Kwiatkowski, & Rasmussen, 1990) and Phonology Project Technical Report No. 1 (Shriberg, Kwiatkowski, Rasmussen, Lof, & Miller, 1992).

Recent methodological focus of the Phonology Project has been on identifying acoustic correlates for segmental domains (vowels, consonants), prosody domains (Phrasing, Rate, Stress) and voice domains (Loudness, Pitch, Laryngeal Quality, Resonance) heretofore transcribed and coded using only perceptual methods. The References section includes citations for these published and unpublished papers, including Technical Report No. 1 (Shriberg et al., 1992) which includes acoustic and psychometric information on the PVSP. Examples of research with the PVSP until 2001 are summarized in McSweeny and Shriberg (2001). More recent examples

include use of the PVSP in a study of motor speech disorders in galactosemia (Shriberg, Potter, & Strand, 2010), and in a study of the hypothesis of motor speech disorders in autism spectrum disorder (Shriberg, Paul, Black, & van Santen, 2010).

Statement of the Problem

The goal of the present technical report is to describe an acoustic procedure to validate and quantify perceived vocal tremor. The PVSP manual and accompanying audio samples provide exemplars for Prosody-Voice Code 26: Break/Shift/Tremulous. One use of this code is for utterances in which the PVSP coder perceives brief occurrences of modulation of the amplitude envelope, typically in a monophthongal vowel segment. This report includes rationale and procedures for an acoustics-based approach to confirm and quantify the frequency of perceived tremor in relatively short vowels produced by children.

Methodological Needs and Constraints

There are several reports in the literature on the acoustic analysis of tremor (e.g., Jiang, Lin, & Hanson, 2000; Schoentgen, 2002; Winholtz & Ramig, 1992; see Buder & Strand, 2003 for an extended review). As Buder and Strand describe, most approaches to quantifying vocal tremor apply some form of fast fourier transform (FFT) to the amplitude contour of a voiced signal. Buder and Strand proposed a multiple analysis technique that produces what they term a *modulogram*. The technique is applied to both the amplitude and F0 contours and produces information about the frequency of tremor (if present) in both the amplitude and F0 domains. Specifically, they divide the possible low frequency amplitude modulations into three categories: *wow* (0.2-2 Hz), *tremor* (2-10 Hz), and *flutter* (10-20 Hz). Buder and Strand use multiple FFT sizes adjusted for each frequency range and pad the FFT window with zeroes to reach the required spectral resolution.

Buder and Strand's useful method to quantify the three classes of F0 variability places a duration constraint on the eligibility of vowels with perceived vocal variability. To confirm and quantify wow (0.2-2Hz), vowels need to be a minimum of 500 ms in duration for the upper part of the range and a minimum of 5 seconds for the lower part of the range, assuming the amplitude contour is sampled at 1000 Hz. Similarly, for tremor, the minimum would be 100 ms for the upper limit and 500 ms for the lower limit. The quantification of flutter requires minimum durations of 50-100 ms. In the examples they use to show the effectiveness of the modulogram method, the voiced segments analyzed range from 8 seconds to as long as 40 seconds (cf. Figures 4, 5, and 6 in Buder & Strand). This duration is considerably longer than the 2-3 seconds of a prolonged vowel required by most other analyses for the acoustic analysis of tremor.

The vowel duration need poses a particular constraint for research with young children. When asked to produce sustained vowels, young children typically have difficulty understanding that they should try to keep the vowel as steady as possible (similar to the problem in asking young children to produce sequential and alternating motion rate tasks [i.e., diadochokinesis] keeping productions "as steady as possible"). In contrast, the continuous language or speech samples that are routinely obtained from speakers of all ages provide an excellent context for the potential occurrence of tremors. A constraint on use of such samples as the source for Buder and Strand's procedures, however, is that vowel segments from continuous speech typically are less than one second in duration. This constraint on confirming and quantifying perceived vocal tremor required development of a new method using vowel segments as short as approximately 250 ms.

Method

Participants

Two databases of digitally recorded audio samples from participants in recent studies completed in our laboratory were inspected to identify vowel tokens in conversational speech that were coded as having perceived vocal tremor. An experienced coder using Prosody-Voice Screening Profile (PVSP) procedures had assigned code PV26: Break/Shift/Tremulous to each candidate utterance. All candidate utterances were reviewed to confirm a perceived vocal tremor. One database included a group of thirteen 3-to 6-year-old participants with Speech Delay. The other included a group of seventeen 4-to 16-year-old participants associated with a metabolic disorder (galactosemia), some of whom had Speech Delay (GALT SD) and others who had Childhood Apraxia of Speech (GALT CAS).

Table 1 includes descriptive information on the five participants in the present study group who had at least one PV26 code in their conversational speech transcript or in a sentence repetition task termed Vowel Task 3. As indicated, the ages of participants, approximately 3 to 10 years, is representative of children assessed for speech sound disorders of known and unknown origin. The competence levels of participants (Percentage of Consonants Correct [PCC] scores) included one participant with Mild speech delay (85.8%), two participants with Mild-Moderate speech delay (84.0%, 81.9%), one participant with Moderate-Severe speech delay (62.3%), and one participant with Severe speech delay (46.5%).

Table 1. Demographic information for the five study participants.

Participant	Database	Gender	Age (Yrs:Mos)	Percentage of Consonants Correct-Revised (PCC)
1	Speech Delay	F	3:4	62.3
2	GALT CAS	M	5:2	46.5
3	GALT SD	M	6:10	85.8
4	GALT SD	M	8:5	81.9
5	GALT CAS	M	10:1	84.0

Procedures

Candidate utterances with PV26 codes were further examined to determine whether they included words with relatively long vowels preceded or followed by any consonant other than an approximant (/j̇/ or /w/), unless the combination of waveform and spectrogram can be used to segment the vowel reliably. This restriction was imposed to avoid the difficulty in segmenting vowels in the context of a preceding or following approximant. Words satisfying these criteria were extracted from the continuous speech file and saved as individual files for further processing. The free software Wavesurfer (available for download from www.kth.se) and the commercial sound editing software Adobe Audition 3.0 (Adobe Systems, Inc., 2007) were the two programs used to process the words as described below. Any software can, in fact, be used if it functions to complete two tasks: (a) compute a running amplitude contour (power plot) and save it as a text file, and (b) read in the text file as a time series and convert it to a "sound" file at the appropriate sampling rate.

The following five-step series describes the acoustic method developed to quantify perceived vocal tremor:

1. Each utterance on the PVSP log that has been assigned code PV26 is scanned for words containing vowels that show an envelope fluctuation. Each such word is then segmented from the utterance sound file and saved as a separate word sound file.
2. The beginning and end of the vowel are marked and its duration calculated. An RMS power plot of the word is created using the function available in Wavesurfer. The relevant parameters for this procedure are: no pre-emphasis, a 5 ms window and 1 ms step. The software asks for window size in terms of number of samples, so this must be calculated from the sampling rate of the original sound file. Thus, if the sampling rate was 44100 Hz, 5 ms would equal 220 points. The 1 ms step size will result in the power plot having a value for each ms of the waveform. The power plot is then saved as a text file.
3. The text file is loaded into the acoustic software's (e.g., Wavesurfer, Audition, Praat) waveform editor with the specification that it is a sound sampled at 1000 Hz. This effectively converts the power plot into a waveform—as if it were a sound.
4. At the beginning and end of each vowel, 1 second of silence (zeroes) is inserted before and after the section of the power plot waveform corresponding to the original section of vowel waveform.
5. A Hamming window is centered on the power plot section delimited by the zeroes on both sides and an FFT is performed. The resolution of the FFT is determined by the length of the window. The length of the window is defined by the number of samples. For this task, one sample corresponds to one millisecond, because the sampling rate was defined as 1000 Hz when the power plot was converted into an artificial waveform. Thus, for example, a 256 point FFT will produce an output roughly every 4 Hz; a 512

point FFT will produce an output roughly every 2 Hz, and; a 1024 point FFT will produce an output roughly every 1 Hz.

An example may help to consolidate these instructions. Consider a vowel that is 320 ms long. The corresponding section of the power plot when converted to a sound file will include 320 points. When a 512 point FFT is executed on this power plot waveform, the FFT algorithm will use the 320 available points plus it will extend to include 96 points of zero from each side (512-320=192, 192/2=96). This is the reason for padding the power plot section with zeroes. Thus, even though the vowel was only 320 ms long, the FFT will produce magnitude estimates every 2 Hz. The same rationale can be applied if the vowel is 600 ms long. The FFT window then can be 1024, producing a magnitude estimate every 1 Hz.

When this method is applied, the FFT produces a power spectrum delimited from 1 to 500 Hz. The resolution of the spectrum depends on the size of the FFT as described above. The spectrum shows a strong peak at low frequencies (less than 20 Hz) and another slightly weaker peak at a frequency corresponding to the fundamental frequency of the original vowel waveform.

Results

The method described in this report was applied to 13 sample words produced by the five participants described in Table 1. Table 2 is a summary of the findings. The entries under Spectral Peak are the frequencies of the lowest peak in the derived spectrum. These peaks averaged 8.3 Hz (SD: 3.1), placing the mean within the top end of the 2-10 Hz range that Buder and Strand designated as vocal tremor. Ten of the 13 (76.9%) spectral values in Table 2 were within this range, with the remaining three entries (“He” for Participant 2 [11Hz], “He’s” for Participant 3 [13Hz], and “Then” for Participant 5 [15Hz]) meeting Buder and Strand’s criterion for vocal flutter. The difference in the ages of the participants, 5:2, 6:10, and 10:1,

Table 2. Spectral peak findings for 13 words with perceived vocal tremor using the modified Buder and Strand (2003) procedure to quantify FO variability.

Participant	Word	Vowel Duration (ms)	FFT size (samples)	Spectral Peak (Hz)	Buder and Strand classification
1	Head	432	1024	6	Tremor
2	(S)he	765	512	11	Flutter
3	Dad	923	512	6	Tremor
3	He's	648	512	13	Flutter
3	Pool	858	512	6	Tremor
3	You	924	512	6	Tremor
4	Have	413	512	5	Tremor
4	He	428	1024	8	Tremor
5	Did	670	512	9	Tremor
5	Pool	723	1024	8	Tremor
5	Saw	515	1024	9	Tremor
5	Then	862	512	15	Flutter
5	We	737	1024	6	Tremor
			Mean SD	8.3 3.1	

respectively, does not support an age trend for vocal variability consistent with flutter compared to tremor.

At this point, it is necessary to collect more samples from different research collaborators so that the generality and reliability of the method can be assessed. There are steps in the method that require acoustic and procedural decisions, such as determining the onset and offset of the vowel segment, the peak determination in the final spectrum, and the choice of FFT size and the use/non-use of zero padding. A larger sample size, on the order of approximately 100 instances of perceived vocal tremor, would provide the needed reliability data and a useful database for additional study of potential linguistic and motor substrates of vocal tremor in pediatric speech sound disorders.

Summary

The method described in this report provides an acoustic means to confirm perceptual vocal tremor. The procedure should work optimally with relatively noise free recordings or, if there is noise, a constant noise which would add consistently to the RMS calculations. Often in recordings of children's speech there are transient noises superimposed on the speech signal coming from fidgeting, moving toys around, etc. The presence of these kinds of noise would render the recording non-usable. Another issue is the requirement of very long segments to determine amplitude modulations in the low range (0.2-2 Hz). A transcriber might hear this kind of modulation and mark it as tremor but there might be no vowels in the recorded speech sample long enough to use for determining the frequency of the modulation.

As noted above, additional studies of this acoustic procedure are needed to assess its validity and utility in studies with participants who meet criteria for a number of subtypes of speech sound disorders, including those in neurogenetic, neurological, complex

neurodevelopmental, and idiopathic contexts. The primary rationale for such studies is that a quantitative index of tremor or possibly flutter will provide a useful diagnostic marker of motor speech disorder.

Illustration of the Method and Sound Files used for this Report

The companion PowerPoint file for this report includes six slides that illustrate the procedures.

Slide 1 is the raw waveform of the vowel in the word “Did” produced by Participant 5. The red line traces the positive excursion of amplitude showing an amplitude modulation.

Slide 2 is the RMS power plot corresponding to the waveform shown in Slide 1. The power plot was created using an approximately 5 ms window moving in 1 ms steps.

Slide 3 is the result of a 512 point FFT using a Hamming window centered in the middle of the power plot “waveform.” The vertical cursor line is set at the highest amplitude low frequency peak, at roughly 9 Hz. The movement resolution of the mouse pointer on any given computer screen might result in a 1-2 Hz variability in taking this measure. The use of 512 points for the FFT resulted in a 2 Hz resolution in the spectrum. If there is no clear peak at this point, then zero padding of the power plot “waveform” and a 1024 point FFT can be used. This was necessary, in fact, for some of the utterances analyzed and reported in Table 2 and the accompanying discussion.

Slides 4-6 provide the audio files for the entries in Table 2. The underlined word in each utterance includes the vocal tremor.

References and Other Phonology Project Papers on Acoustic Methods

(see also http://www.waisman.wisc.edu/phonology/)

Buder, E.H. & Strand, E.A. (2003). Quantitative and graphic acoustic analysis of phonatory modulations: The modulogram. *Journal of Speech, Language, and Hearing Research, 46*, 475-490.

Jiang, J., Lin, E., & Hanson, D.H. (2000). Acoustic and airflow spectral analysis of vocal tremor. *Journal of Speech, Language, and Hearing Research, 43,* 191-204.

McSweeny, J. (1998). *Procedures to obtain extended conversational speech samples for prosody-voice analysis* (Tech. Rep. No. 7). Phonology Project, Waisman Center, University of Wisconsin-Madison.

McSweeny, J. L., & Shriberg, L. D. (2001). Clinical research with the prosody-voice screening profile. *Clinical Linguistics and Phonetics*, *15*, 508-528.

Schoentgen, J. (2002). Modulation frequency and modulation level owing to vocal microtremor. *Journal of Acoustical Society of America, 112*, 690-699.

Shriberg, L. (2010). *Childhood speech sound disorders: From post-behaviorism to the post-genomic era*. In R. Paul & P. Flipsen (Eds.), Speech sound disorders in children. (pp.1–34). San Diego, CA: Plural Publishing.

Shriberg, L. D., Ballard, K. J., Tomblin, J. B., Duffy, J. R., Odell, K. H., & Williams, C. A. (2006). Speech, prosody, and voice characteristics of a mother and daughter with a 7;13 translocation affecting *FOXP2*. *Journal of Speech, Language, and Hearing Research, 49*, 500-525.

Shriberg, L.D., Fourakis, M., Hall, S.D., Karlsson, H.B., Lohmeier, H.L., McSweeny, J.L., et al. (in press). Extensions to the Speech Disorders Classification System (SDCS). *Clinical Linguistics and Phonetics*.

Shriberg, L. D., Green, J. R., Campbell, T. F., McSweeny, J. L., & Scheer, A. (2003). A diagnostic marker for childhood apraxia of speech: The coefficient of variation ratio. *Clinical Linguistics and Phonetics*, *17*, 575-595.

Shriberg, L. D., & Kent, R. D. (2003). *Clinical phonetics* (3rd ed.). Boston: Allyn & Bacon.

Shriberg, L. D., Kwiatkowski, J., & Rasmussen, C. (1990). *The Prosody-Voice Screening Profile*. Tucson, AZ: Communication Skill Builders.

Shriberg, L. D., Kwiatkowski, J., Rasmussen, C., Lof, G. L., & Miller, J. F. (1992). *The Prosody-Voice Screening Profile (PVSP): Psychometric data and reference information for children* (Tech. Rep. No. 1). Phonology Project, Waisman Center, University of Wisconsin-Madison.

Shriberg, L. D., Paul, R., Black, L. M., & van Santen, J. P. (2010). *The hypothesis of apraxia of speech in children with autism spectrum disorder*. Manuscript submitted for publication.

Shriberg, L.D., Potter, N.L. & Strand, E.A. (2010). *Prevalence and phenotype of Childhood Apraxia of Speech in youth with galactosemia*. Manuscript submitted for publication.

Winholtz, W.S. & Ramig, L.O. (1992) Vocal tremor analysis with the vocal demodulator. *Journal of Speech, Language, and Hearing Research*, 35, 562-573.