# Critical Review: How Accurate are Voice Accumulators for Measuring Vocal Behaviour?

Lauren Greenwood
M.Cl.Sc. (SLP) Candidate
University of Western Ontario: School of Communication Sciences and Disorders

This critical review examines the accuracy of voice accumulators for measuring vocal behaviour in laboratory and field settings. Study designs include single group, between group (nonrandomized), and single subject. Overall, research supports the accuracy of voice accumulators for measuring aspects of vocal behaviour in different settings, however, more research is needed to determine the capabilities of these devices for measuring the vocal behaviour of individuals with deviant voice qualities.

## *Introduction*

An important clinical issue in the study of voice disorders is determining how individuals use their voices outside of the clinical setting (e.g., Ohlsson, Brink, & Löfqvist, 1989). It has been difficult to draw the connection between vocal behaviour and voice disorders in an ecologically valid way due to a limited number of resources available to measure vocal behaviour in natural contexts. The *voice accumulator* has been developed for this purpose (e.g., Szabo, Hammarberg, Håkansson, & Södersten, 2001).

Vocal parameters such as *phonation time (PT)*, *fundamental frequency* (*$f_o$*), and speech *sound pressure level (SPL)* are important clinical measures of vocal behaviour. PT is the length of time that the vocal folds vibrate, *$f_o$* is the lowest frequency produced by the voice, and SPL is the pressure of the voice, often interpreted as speech loudness level (Granqvist, 2003). Voice accumulators are designed to collect data based on these parameters (Cheyne, Hanson, Genereux, Stevens, & Hillman, 2003). These devices may be more useful than traditional recording devices (e.g., digital audiotape recorders) because they do not capture actual speech, therefore, they protect the privacy of the user (Ryu, Komiyama, Kannae, & Watanabe, 1983).

Given the potential clinical applications of voice accumulators, it is important to determine whether they are accurate in measuring vocal behaviour both within and outside of clinical settings. If they accurately capture vocal parameters of interest, they may be used to help clinicians to acquire information about their clients' vocal behaviours in their everyday settings, and address these behaviours within the clinical context.

## *Objectives*

The primary objective of this paper was to provide a critical evaluation of the existing literature regarding the accuracy of voice accumulators for measuring vocal behaviour in various contexts. The secondary objective was to offer recommendations for the clinical use of these devices and ideas for future research.

## *Methods*

### Search Strategy

Articles related to the topic of interest were found using the following computerized databases: PsycINFO, PubMed, Medline, and CINAHL. Keywords used for the database searches were as follows:
[(voice) or (speech) and (accumulator)]
[(voice accumulator) and (vocal accumulator) and (speech accumulator)].

### Selection Criteria

Studies selected for inclusion in this critical review were required to investigate the accuracy of voice accumulators for measuring vocal behaviour in any setting and with any participant group.

### Data Collection

The literature search yielded six articles congruent with the selection criteria. Three of these articles described two studies each, yielding a total of nine studies. Five studies employed a single group design, three studies employed a between group (nonrandomized) design, and the last study employed a single subject design.

## *Results*

### *Testing the Device in Field Settings*

Ryu et al. (1983) conducted a between groups study (nonrandomized; level two evidence) to determine the capabilities of the voice accumulator for measuring the speaking time of 11 subjects in different occupations. Within one day, the voice accumulator collected a large range of data representing the total speaking time of each subject (i.e., 33 minutes to 182 minutes). These data matched the authors' expectations, and therefore, they inferred that the device was useful and accurate for measuring vocal behaviour over the course of one day.

There are some methodological weaknesses of this study. The authors did not discuss the sampling procedure used or the demographic properties of the sample, therefore, it is difficult to generalize the study results. Furthermore, there were no statistical analyses done; the results were presented at a descriptive level. There were also inconsistencies in the data presentation methods; subjects were grouped together in some analyses and described individually in others. Given these concerns, the applicability of this study to clinical practice is questionable. Results are deemed equivocal and should be interpreted with considerable caution.

Beukers, Bierens, Kingma, and Marres (1995) conducted a between subjects study (nonrandomized; level two evidence) to determine the accuracy of a voice accumulator for measuring PT and SPL of 72 professionals (representing 12 different professions) in their work environments over 12 hours. The authors stated that wearing the device for longer periods of time resulted in more valid data. The investigators used a sound level meter to collect data about vocal intensity, and the data had good agreement with those from the voice accumulator. The authors compared the PT and SPL of different professionals, and some of the results agreed with their predictions, (i.e., teachers spoke for a longer period of time and at a higher intensity than all other professionals groups) and some did not (i.e., speech therapists, receptionists, and telephone operators spoke at lower intensities than expected). Selected subjects were instructed to record the types of activities in which they participated and their perceived vocal loudness at specified times. Comparisons were made between these subjective ratings and the data from the voice accumulator. Some discrepancy existed among these results. Authors reported that there was good agreement between these subjective ratings and the data from the voice accumulator. They also reported that vocal intensity, as measured by the voice accumulator, was lower than the subjects' self-ratings. Overall, the authors concluded that the device was accurate in collecting vocal data.

A calibration measure was done with a sound level meter in order to confirm proper placement of the device during recordings. This showed that the device was prepared the same way for all subjects. There are some weaknesses with respect to study results. First, the authors did not give an explanation for why the device was more valid in measuring longer recordings, nor did they describe what was considered to be a longer recording. The discrepancy in results, as described above, is another issue. Given these concerns, the results are deemed suggestive and should be interpreted with caution when applying them to clinical practice.

***Testing the Device in Laboratory and Field Settings***

Ohlsson et al. (1989) conducted two studies to investigate the accuracy of a voice accumulator for measuring $f_o$ and PT. The first study was a single group (level three evidence) *validation* study to test the device in a laboratory setting, and the second study was a between groups (nonrandomized; level two evidence) *application* study to test the device in natural contexts.

In the first study, authors assessed the ability of the device to measure PT by having 3 female subjects read a standardized passage several times over the course of one day while vocal information was collected by the voice accumulator and a tape recorder. In order to determine the relationship between the measurements from the two devices, a pearson product-moment correlation ($r$=0.35) and regression analysis (slope of 0.51; intercept of 10.92) were conducted. The authors did not report a $p$-value, therefore, it is unclear whether the relationship was significant. Overall, the proportion of time phonating (i.e., PT) was found to be 30% lower as measured by the voice accumulator than the tape recorder. The authors suggested that this measurement error was due to lower-amplitude segments being missed by the device. They explained that these omissions should only pose a problem when short segments of speech are being captured and a small number of values are used to calculate the average. In the next part of this study, authors assessed the ability of the voice accumulator to measure $f_o$. Four male and 4 female subjects read a standardized passage and an electroglottography device collected information about $f_o$ simultaneously with the accumulator. A pearson product-moment correlation ($r$=1) and a regression analysis (slope of 1; intercept of 2.86) were used to determine the relationship between the recordings. Results indicated that the two methods were highly correlated. Overall, the authors concluded that the voice accumulator was suitable for measuring $f_o$ and PT.

Although these results suggest agreement between the voice accumulator and comparison measures, there are some methodological concerns. The use of a different sample and a different comparison device for assessing PT and $f_o$ makes it difficult to determine whether the voice accumulator was more accurate in capturing $f_o$ than PT, as the correlations would suggest. Another concern is that the authors did not discuss the validity of the comparison measurements; therefore, it may be difficult to meaningfully interpret the correlations.

In the second study by Ohlsson et al. (1989), authors measured the vocal behaviour of 10 female speech-language pathologists (SLPs) and 10 female nurses over the course of two work days. Due to work-related voice use and vocal training in the former group, the authors hypothesized that the SLPs would have a lower $f_o$ and a longer PT than the nurses. Average group data of $f_o$ and PT as well as the subjects' predictions of their PT were compared and analyzed using the Wilcox rank sum test ($p < 0.5$). Results indicated that the SLPs had a significantly lower average $f_o$ than the nurses. The SLPs had higher values of PT than the nurses, but this difference was not significant. Both $f_o$ and PT varied according to work activity, as expected. Based on these results, the authors concluded that the device was useful for measuring vocal behaviour in field settings.

The authors used appropriate, nonparametric statistical analyses and reported a significance level, which allowed for meaningful interpretation of the results. One weakness of this study is that the theory on which the authors based their hypotheses was not well described. Furthermore, the researchers did not control for the length of time that the device was used which may have contributed to variations in vocal behaviour.

Although many of the results from the studies by Ohlsson et al. (1989) indicated that the device accurately measured vocal parameters in certain circumstances (e.g., longer-term recordings), there were some unknown variables (e.g., sampling procedure, validity of comparison tool, effects of length of use) that should be considered when interpreting the results. The results of these studies are deemed suggestive. It is recommended that they be interpreted with some caution when making decisions about the use of this device both within and outside of the clinical setting.

Szabo et al. (2001) evaluated the accuracy of a voice accumulator for measuring $f_o$ and PT in two studies. The first study was a single group laboratory study and the second study was a single group field study (level three evidence). The device used was a revised model of an earlier device described by Ohlsson et al. (1989).

Four subjects participated in the first study; 2 of the 4 subjects were employed in voice professions and the other 2 subjects had no history of vocal training. One microphone was connected to the voice accumulator and a second was connected to a computer program, which was considered a valid comparison measure. Differences between the measures from the accumulator and the computer program were expressed as a percentage. For $f_o$, results showed high agreement between the voice accumulator and the computer program for 3 of 4 subjects. There was more variation with respect to PT. For every second of speech, high correlations were found between measures from the voice accumulator and computer program. Overall, results indicated that the voice accumulator was accurate in measuring $f_o$ and PT for long-term recordings, in comparison to the computer program.

There are both strengths and weaknesses of this study. First, the authors discussed the sample characteristics in detail, which increases the ability to generalize study results. The authors also ensured that a standardized microphone placement procedure was used with all subjects. The validity of the comparison measure was also clearly explained, which made the comparison analyses meaningful. Additionally, the authors used appropriate correlations to analyze test results. Although the pearson correlation values were high, the authors did not report an alpha level, which may have been useful in confirming their significance. Furthermore, authors used percent difference values to determine the difference between the voice accumulator and the computer program. Their analysis may have been strengthened if they used a statistical procedure and reported results with a level of significance.

In their second study, Szabo et al. (2001) tested the voice accumulator with 2 female SLPs and 2 male engineers at their work. Voice accumulator data revealed that mean $f_o$ and PT values were comparable to other field studies (e.g., Ohlsson et al., 1989).

The authors clearly presented their results, and provided recommendations for future uses of the accumulator. It may have been beneficial to include a comparison measure in this study in order to determine the relative accuracy of the device in a field setting.

Overall, the results from both the laboratory and field studies are deemed compelling and should be used to support the use of this voice accumulator in clinical situations.

Szabo, Hammarberg, Granqvist, and Södersten (2003) conducted two studies to determine the accuracy of a voice accumulator for measuring $f_o$ and PT in comparison to a digital audiotape (DAT) recording. The first study was a single case laboratory study (level four evidence) and the second study was a single group field study (level three evidence).

One female SLP participated in the first study. She read a text passage four times with a normal voice, breathy voice, strained voice, and creaky voice. She also sustained the vowel /a/ with an increasing frequency and intensity. Voice accumulator values were reported descriptively as a percentage relative to DAT values.

There was high agreement between the voice accumulator and the DAT with respect to $f_o$ and PT for most of the recordings. The exceptions were the creaky voice recording, and the highest frequency (i.e., 440 Hz) and softest intensity of /a/. The authors concluded that based on these results, the voice accumulator was accurate for measuring the $f_o$ and PT of most voice qualities, as compared to the DAT.

This article has many strengths. First, the inclusion of a variety of recording samples (i.e., different voice qualities, frequencies, and intensities) allowed authors to analyze the range of capabilities of the accumulator. Furthermore, the DAT is a reliable comparison measure, allowing for meaningful interpretation of the calculated difference scores.

Despite these strengths, this study also presented with limitations. It is difficult to generalize results from single subject studies to the broader population. Furthermore, despite the fact that the SLP had extensive voice training, her production of various voice qualities may be different from individuals who speak with breathy, strained, or creaky voices. Therefore, it is difficult to conclude with certainty that the device can accurately measure the vocal behaviour of individuals with deviant voice qualities. Additionally, the authors did not provide any explanations for why the device did not register high frequencies and low intensities. This may pose a problem for collecting voice information from certain individuals (i.e., those with hypophonia). As with Szabo et al. (2001), authors reported the relationship between the devices as percent differences; their results may have been strengthened if they used a statistical test and reported a significance value.

In their second study, Szabo et al. (2003) tested the accuracy of the voice accumulator with 3 female pre-school teachers at work. The accumulator and the DAT were used to record subjects over the course of one work day. An additional program was used to eliminate background noise from voice recordings. Spearman non-parametric correlations were used to determine the relationship between the two recording methods. For both $f_o$ and PT, correlations between the two methods were high for two of three subjects ($r_s \geq 0.80$). The difference ranges between the methods were less for $f_o$ than PT, meaning that $f_o$ may have been captured more accurately. Authors felt that the data were different for the third subject because she had subcutaneous tissue on her neck which affected microphone placement.

This study has both strengths and weaknesses. First, unlike other studies included in this review, Szabo et al. (2003) used a program to eliminate background noise from their data and this is important because background noise is a factor in most field settings. The authors used a reliable comparison device to determine the relative accuracy of the voice accumulator. They also used appropriate statistical procedures to analyze their data. Even though the spearman correlations were high, the authors did not report a significance level which may have strengthened their results. Another potential limitation is that the sample was very homogenous which may limit generalizability.

Given the strengths and weaknesses of both studies by Szabo et al. (2003), the results are deemed suggestive and should be regarded with some caution when considering the use of this device in clinical practice.

Cheyne et al. (2003) developed a voice accumulator and conducted a single group study (level three evidence) to determine the accuracy of the device for measuring $f_o$, PT, and SPL. The device used an accelerometer placed on the neck that tracked vocal fold vibrations. The authors reported some advantages of using an accelerometer over a microphone (e.g., less obtrusive and more immune to background noise).

Stationary recordings, short-term ambulatory recordings, and long-term ambulatory recordings were used to test the device. A total of 87 subjects participated in the stationary recordings; 67 of these subjects had voice disorders that were associated with dysphonia and the remainder of the subjects had normal voice qualities. Simultaneous recordings were made with an accelerometer and a microphone attached to the DAT. The subjects were asked to produce several different speech samples, including productions with varying pitch and loudness ranges. Eight subjects with normal voices participated in the short-term ambulatory recordings. The accelerometer signals were analyzed from the DAT while subjects engaged in their usual work activities. Four subjects with normal voices participated in the long-term ambulatory recordings while wearing the voice accumulator at work.

The accelerometer was able to analyze all acceleration data, regardless of severity of dysphonia. It was also capable of detecting soft and loud speech signals (i.e., 20 dB to 150 dB). The authors used a linear least-squares fit to determine the relationship between the SPL captured by the microphone and the accelerometer signal ($r$=0.80, $p < 0.001$). These results indicated that there was good agreement between the microphone and accelerometer recordings. Specifically, as SPL increased, the acceleration signal of vocal fold vibrations also increased. The long-term recordings made with the accumulator lasted between 7.1 and 11.2 hours, which was slightly less than the goal of 12 hours.

The authors used an appropriate statistical test to analyze the relationship between the microphone and accelerometer. They also explained that regardless of the correlation between the methods, the accelerometer may be more accurate in capturing important underlying vocal fold information (e.g., vocal fold collision factors), than the microphone method. This may have implications for understanding individuals' use of the vocal mechanism. The authors used a different sample for each recording; therefore, it is unclear whether the device was equally accurate in all three recording scenarios. Given the results and these discussion points, this study is deemed compelling. Results should be highly regarded when making decisions about using the accelerometer and voice accumulator method within and outside of clinical settings.

### *Discussion*

The recent literature provides evidence for the accuracy of voice accumulators for measuring vocal behaviour both within and outside of clinical settings. Studies have evolved since the earliest forms of the voice accumulator (Ryu et al., 1983), and demonstrate the increasing strengths and applications of these devices. It is important to recognize that most of the authors who conducted the studies in this review helped to create the devices that they tested. Because of this, a level of bias may be present and should be considered when interpreting the test results.

Researchers have identified the use of voice accumulators for measuring $f_o$, PT, and SPL. In the laboratory studies, results showed that data from the voice accumulator correlated with data from other related measurement devices (e.g., electroglottography, DAT, computer programs; Ohlsson et al., 1989; Szabo et al., 2001; 2003). Some common themes that arose were that proper placement of the microphone was important for accurate measurements (Cheyne et al., 2003), earlier versions of the voice accumulator underestimated PT (Ohlsson et al., 1989), and that the device was typically more accurate in measuring vocal behaviour in longer-term recordings (Beukers et al., 1995). The device had difficulty capturing some forms of deviant voice quality, as well as high frequencies and low intensities (Szabo et al., 2003). These deviant vocal properties may be a result of differences in the way an individual's vocal folds vibrate. A potential solution to address this issue is to use an input method other than a microphone (e.g., accelerometer), which may be more accurate in capturing information about vocal fold activity (Cheyne et al., 2003).

In field settings, the $f_o$ and PT outputs of the voice accumulator were typically congruent with other field studies (e.g., Szabo et al., 2001). Furthermore, vocal differences found between individuals in different professional groups typically matched authors' expectations (e.g., Ryu et al., 1983).

It is important to note that the voice accumulator is not the only device that can be used to capture voice data in different settings. Airo et al. (2000) devised a method to collect voice data in a laboratory setting by using two microphones and subtracting background noise from a speech signal in order to determine speech SPL and PT. The device was accurate in collecting this information in a moderate amount of background noise. Other researchers have also devised methods to capture vocal information in ecologically valid ways (e.g., Granqvist, 2003). Despite having relevancy to the topic area, these studies were not included in this review because they used a different recording method. They should, however, be considered when making decisions about the use of measurement devices in clinical contexts.

### *Conclusion*

At present, the literature reveals that voice accumulators are accurate for measuring vocal behaviour in comparison to other measurement devices. These devices may be more accurate for capturing vocal information in longer-term recordings and with individuals whose voice qualities are within the normal frequency and intensity range.

### *Recommendations*

Based on the results of this review, it is recommended that further research be conducted to clarify the capabilities and accuracy of voice accumulators. Future studies should consider:

a) Employing a larger and more variable sample (e.g., subjects with deviant voice qualities), such that results can be generalized and applied to clinical practice;
b) Conducting a laboratory and field study with the same group of subjects and the same comparison device. This will allow authors to determine whether the voice accumulator is equally accurate in capturing voice information in both settings;
c) Recording speech samples of varying lengths to determine the optimal length of time for recording, such that measurement error does not affect the data;
d) Expanding the outer limits of the voice accumulator for measuring fundamental frequency and intensity;
e) Using the voice accumulator in a clinical context to collect pre-treatment and post-treatment data. This will allow researchers to determine the accuracy of the device for capturing voice related changes due to treatment effects.

### *Clinical Implications*

The evidence available on voice accumulators provides merit for their use in clinical practice related to voice and voice disorders. They may be useful for measuring the vocal behaviour of individuals with mild deviations in fundamental frequency and intensity. At present, the voice accumulator may not be useful for measuring the voices of individuals with certain degrees of hypophonia or severely deviant voice qualities.

### ***References***

Airo, E., Olkinuora, P., Sala, E. (2000). A method to measure speaking time and speech sound pressure level. *Folia Phoniatrica et Logopaedica, 52,* 275-288.

Buekers, R., Bierens, E., Kingma, H., & Marres, E. H. (1995). Vocal load as measured by the voice accumulator. *Folia Phoniatrica et Logopaedica, 47,* 252-261.

Cheyne, H. A., Hanson, H. M., Genereux, R. P., Stevens, K. N., & Hillman, R. E. (2003). Development and testing of a portable vocal accumulator. *Journal of Speech, Language, and Hearing Research, 46,* 1457-1467.

Granqvist, S. (2003). The self-to-other ratio applied as a phonation detector for voice accumulation. *Logopedics, Phoniatrics, Vocology, 28,* 71-80.

Ohlsson, A-C., Brink, O., & Löfqvist, A. (1989). A voice accumulator – validation and application. *Journal of Speech and Hearing Research, 32,* 451-457.

Pegoraro Krook, M. I. (1988). Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis. *Folia Phoniatrica et Logopaedica, 40,* 82-90.

Ryu, S., Komiyama, S., Kannae, S., & Watanabe, H. (1983). A newly devised speech accumulator. *Journal of Oto-Rhino-Laryngology, Head and Neck Surgery, 45,* 108-114.

Szabo, A., Hammarberg, B., Granqvist, S., & Södersten, M. (2003). Methods to study pre-school teachers' voice at work: Simultaneous recordings with a voice accumulator and a DAT recorder. *Logopedics, Phoniatrics, Vocology, 28,* 29-39.

Szabo, A., Hammarberg, B., Håkansson, A., & Södersten, M. (2001). A voice accumulator device: Evaluation based on studio and field recordings. *Logopedics, Phoniatrics, Vocology, 26,* 102-117.