Published in final edited form as:
*Biomed Signal Process Control*. 2015 March ; 17: 21–28. doi:10.1016/j.bspc.2014.11.007.

# Experimental Investigation on Minimum Frame Rate Requirements of High-Speed Videoendoscopy for Clinical Voice Assessment

**Dimitar D Deliyski**[1,2,3,4], **Maria EG Powell**[2,4], **Stephanie RC Zacharias**[2,4,5], **Terri Treman Gerlach**[6], and **Alessandro de Alarcon**[1,3]

[1] Division of Pediatric Otolaryngology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[2] Communication Sciences Research Center, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

[3] Department of Otolaryngology-Head and Neck Surgery, University of Cincinnati, Cincinnati, Ohio, USA

[4] Department of Communication Sciences and Disorders, University Cincinnati, Cincinnati, OH, USA

[5] Division of Speech-Language Pathology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

[6] Voice and Swallowing Center, Charlotte Eye Ear Nose and Throat Associates, Charlotte, NC, USA

## Abstract

This study investigated the impact of high-speed videoendoscopy (HSV) frame rates on the assessment of nine clinically-relevant vocal-fold vibratory features. Fourteen adult patients with voice disorder and 14 adult normal controls were recorded using monochromatic rigid HSV at a rate of 16000 frames per second (fps) and spatial resolution of 639×639 pixels. The 16000-fps data were downsampled to 16 other rate denominations. Using paired comparisons design, nine common clinical vibratory features were visually compared between the downsampled and the original images. Three raters reported the thresholds at which: (1) a detectable difference between the two videos was first noticed, and (2) differences between the two videos would result in a change of clinical rating. Results indicated that glottal edge, mucosal wave magnitude and extent, aperiodicity, contact and loss of contact of the vocal folds were the vibratory features most

Corresponding author: Dimitar D Deliyski, Ph.D. Communication Sciences Research Center Cincinnati Children's Hospital Medical Center 3333 Burnet Ave, MLC 15008, Cincinnati, OH 45229-3039 Tel: +1 513 803 5302 Fax: +1 513 803 1911 Dimitar.Deliyski@cchmc.org.

**Publisher's Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Portions of this study have been presented at the Annual Symposium: Care of the Professional Voice, Philadelphia, PA (2010-2012) and the International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, Firenze, Italy (2013).

sensitive to frame rate. Of these vibratory features, the glottal edge was selected for further analysis, due to its higher rating reliability, universal prevalence and consistent definition. Rates of 8000 fps were found to be free from visually-perceivable feature degradation, and for rates of 5333 fps, degradation was minimal. For rates of 4000 fps and higher, clinical assessments of glottal edge were not affected. Rates of 2000 fps changed the clinical ratings in over 16% of the samples, which could lead to inaccurate functional assessment.

**Keywords**

high-speed videoendoscopy; frame rate; clinical voice assessment; laryngeal imaging

## 1. Introduction

High-speed videoendoscopy (HSV) is emerging as a potentially valuable assessment tool for visualizing the vocal folds in motion. However, lack of practical guidelines for the use of HSV is one barrier to wide-spread implementation of HSV in clinical settings. Voice researchers have access to ultra-high-speed cameras that can capture vocal-fold vibrations at rates over 20000 frames per second (fps) [1]. However, there is a large disparity between high-speed technologies available in a research setting and commercial HSV systems available in the voice clinic, with frame rates usually ranging between 2000 and 5000 fps. Previous research suggests that an insufficient frame rate may bias the visual assessment of clinically relevant features [1-3]. Vibratory characteristics, such as mucosal wave and glottal edge, may become blurred or imperceptible at higher velocities due to temporal averaging inherent in HSV technology [1,3]. Additionally, cross-terms due to temporal aliasing may be substantial at lower frame rates [4,5].

Clinicians use visual perception when applying laryngeal imaging for functional voice assessment. Insufficient frame rate may alter a clinical feature the same way as a pathophysiological mechanism. This is due to several factors: 1) frame rates are limited by the camera sensitivity: sensitivity determines the minimum integration (exposure) time, thus, longer integration time leads to image blurring of the faster-moving components in the image [1]; 2) low frame rates may limit the visual information by offering an insufficient number of images per cycle; and 3) low frame rates also limit the speed of slow playback, causing it to fall below the threshold for the perception of apparent motion, which requires images to be presented at a rate of about 16 fps or higher to be perceived as continuous, flicker-free motion [6]. Therefore, if vocal-fold vibrations are captured with an insufficient frame rate, the accuracy and reliability of the visual evaluations of HSV may be jeopardized. Further, objective measures of clinically-relevant features are being developed and validated based on these clinically-established visual vibratory features. Objective measures would also be affected by the artifacts resulting from insufficient frame rates. However, before developing valid objective measures it is imperative to ensure that the visual measures used in clinical practice are not inaccurate and unreliable due to technical factors, such as insufficient frame rates. Therefore, it is highly important to first establish the technical accuracy and reliability of visual assessment of vocal-fold function.

The purpose of the current study is to investigate the threshold at which the HSV frame rate visually degrades clinically-relevant vocal-fold vibratory characteristics. The questions addressed in the study are: a) Which clinically-relevant vibratory features are most sensitive to the HSV frame rate?; b) Does higher fundamental frequency (Fo) require higher frame rates?; c) Which phonatory behaviors require higher frame rates?; d) Are the frame rate requirements different based on gender and pathology?; and e) What are the *recommended* and *minimum* frame-rate requirements for clinical voice assessment?

## 2. Method

### 2.1. Human Data

Fourteen adult vocally-normal speakers (7 male, 7 female) and 14 adults with voice disorders (7 male, 7 female) were recruited from the University of South Carolina (Columbia, SC) and the Charlotte Eye, Ear, Nose, and Throat Associates (Charlotte, NC). Data were collected using a monochrome HSV system at 16000 fps at the integration time of 61 μs. The HSV system was equipped with a monochrome high-speed camera Phantom v7.1 (Vision Research, Inc., Wayne, NJ), a 70-degree 10-mm rigid laryngoscope Model 49-4072 (JEDMED, St. Louis, MO), a 300-Watt xenon light source Model 7152A (KayPENTAX, Montvale, NJ), and a custom-developed 80-mm endoscopic lens adapter (Lighthouse Imaging Corporation, Portland, ME), Participants were recorded producing the vowel /i/ at six different phonatory behaviors (habitual pitch and loudness, high pitch, low pitch, breathy phonation, pressed phonation, and falsetto). A total of 168 recordings were collected. Nine recordings were excluded due to reduced visual quality; a total of 159 recordings were used. This study was approved by the Institutional Review Board of the University of South Carolina.

A substantial concern when designing an experimental study to establish the threshold of sensitivity to a single technical factor is the systematic influence and interaction of other technical factors. The dynamic range of the high-speed camera sensor, the sensor pixel sensitivity, the spatial pixel resolution and the color vs. monochromatic sensors are expected to interact with the visually-established thresholds of sensitivity to the integration time and frame rate. Moreover, the optical zooming, endoscope type, endoscopic angle, focal aperture and other factors can further influence the sensitivity to camera speed. It is not feasible to develop a research design to encompass all technical factors at the same time. Accordingly, the approach taken was to study one factor at a time by reducing within reason the influence of the other factors. Specifically, the spatial resolution of 639x639 pixels used in this study is considered high by HSV standards. Similarly, the use of monochromatic images provides much better structural precision than a color image with the same pixel resolution due to the detrimental effects of the Bayer filtering inherent to color sensors. The camera had true 12-bit dynamic range, i.e. 16 times higher than an 8-bit high-speed camera. The camera sensitivity of 4000 ISO was the highest on the market at the time of the study. Therefore, it is assumed that the results obtained would supersede any settings with lower precision in one or more of these technical factors, including those of color HSV systems.

### 2.2. Data Pre-Processing

Based on visual review of the 16000-fps recordings, a representative 62.5-ms (i.e. 1000-frame) token was manually selected from each recording and saved as an uncompressed video file with spatial resolution of 639x639 pixels. Further, each token was downsampled to form 17 frame-rate denominations ranging from 16000 to 200 fps (Table 1), resulting into a total of 2703 tokens for the study. Downsampling was performed by exactly emulating the capturing characteristics of high-speed cameras at lower frame rates with maximum integration time. This was achieved by summing adjacent frames to simulate continuous integration time, i.e. half the frame rate with twice the integration time, one third the frame rate with three times the integration time, and so forth (Table 1).

### 2.3. Vibratory Characteristics

Nine different vocal-fold vibratory characteristics were identified based on the current HSV clinical protocol, including: mucosal wave magnitude and extent, amplitude and phase asymmetry, aperiodicity, glottal edge, contact and loss of contact, and mucus bridges breaking (Table 2) [3].

### 2.4. Visual-Perceptual Assessment

Two speech-language pathologists (SLP) and one otolaryngologist (ENT), each with expertise in voice disorders and substantial experience with rating HSV images, were asked to compare two tokens of the same recording. Using custom-designed software with a specialized graphic user interface (Figures 1 and 2), raters were asked to compare the downsampled video on the right, to the reference video on the left, which remained at a constant 16000 fps. Raters were blinded to gender, status (normal vs. pathology), phonatory behavior, and frame rate of the downsampled video; however, they were aware that each subsequent token presented on the right was at a lower frame-rate than the previous token. The raters were instructed to mark the first token at which a detectable difference between the two videos was noticed, and subsequently mark the first token at which differences between the two videos would result in a change of clinical rating (or the token was too degraded to rate). Once a change in clinical rating was indicated, the next randomized series of video tokens began.

### 2.5. Experiments

The study was conducted in two stages. The purpose of the first stage was to identify which of the nine features was most sensitive to frame rate. The purpose of the second stage was to use the feature that was most sensitive to frame rate to determine the thresholds at which visual differences are first noticed and clinical ratings change.

**2.5.1. Stage 1: Investigate vibratory features' sensitivity to frame rate—**For each of the 9 vibratory features previously identified (Table 2), 2 male and 2 female subjects that represent that vibratory feature were selected for rating. That is, for some of the features the samples had to be preselected to ascertain the presence of the feature evaluated, because aperiodicity, mucosal wave, asymmetry or mucus bridges had to be present for these subjects in the both habitual-pitch and high-pitch conditions in order to be possible to rate the effect of the frame rate. From each of these 4 subjects’ 17 frame-rate denominations, the habitual-pitch and high-pitch phonatory behaviors were included, resulting in 136 tokens per vibratory feature. Three raters completed the thresholding sequence for each vibratory feature. This experiment was fully executed twice to assess the intra-rater reliability, i.e. the redundancy was 100%. The first trial sequences were presented at a playback rate of 60 fps, and the second was presented at 30 fps, as detailed in Table 1. An example of the software tool used in Stage 1 is shown in Figure 1. Video samples corresponding to the example in Figure 1 are provided in the electronic version of the article.

**2.5.2. Stage 2: Determine frame rate requirements—**Based on the results of the Stage 1 experiment, the most sensitive of the nine previously-defined features was determined. All 159 of the original HSV tokens were then used to rate the determined most sensitive feature for each of the 6 phonatory behaviors. The same thresholding sequence described in Table 1 was used for this stage. Ten percent randomized redundancy was built-in for assessing the intra-rater reliability. Thus, all 28 subjects’ 6 phonatory behaviors were represented in a sequence of 175 reference tokens each in 17 frame-rate denominations, resulting in 2975 tokens rated by each of the three raters. An example of the software tool used in Stage 2 is shown in Figure 2. As a concept, this software tool is very similar to the software used in Stage 1.

### 2.6. Analysis

In Stage 1, descriptive statistics were used to summarize the ratings. The maximum, mode and median fps-threshold values were analyzed among all 3 raters and were compared for each of the 9 vibratory features for each of the 3 rating trials. To gain complete understanding of the data, all fps-threshold distributions were inspected visually. The maximum fps-threshold values for each vibratory feature were reported. Direct agreement within one frame-rate denomination level was reported for the intra- and inter-rater reliability. The intra-rater reliability was measured at 100% redundancy between the two rating trials, i.e. each token was rated twice.

In Stage 2, Spearman's rho correlation was used to establish the relationship between Fo and the fps ratings of first noticed difference and clinical rating change. A 2×2×6 Analysis of Variance (ANOVA) was used to study the effects of gender, norm/pathology status and phonatory behavior on the fps-threshold ratings. Tukey post-hoc analysis was performed to study the relationships among the phonatory behaviors. The maximum values and the distributions of the rated fps thresholds were used to provide the recommended and minimum frame-rate requirements for clinical voice assessment. Direct agreement within one frame-rate denomination level was reported for the intra- and inter-rater reliability. The intra-rater reliability was measured at 10% randomized redundancy presented automatically by the software.

## 3. Results

### 3.1. Stage 1

Up to 7344 tokens were rated in Stage 1. Playback rate was determined to affect the judgments of the clinical feature aperiodicity. Therefore the playback rate for Stage 2 was defaulted to 60 fps to limit the number of tokens presented at playback rates below the threshold for the perception of apparent motion (see Table 1). No differences were detected at 8000 fps for any vibratory feature by any of the 3 raters. Noticeable differences were first noted at 5333 fps for 6 of the features (Table 3). From these 6 features, glottal edge was chosen as the feature to be assessed for Stage 2 for the following reasons: a) The definition of glottal edge is most consistent across clinicians; b) Intra-rater agreements for first noticeable difference were the highest for this feature, ranging from 88-100% among raters (92% combined); and c) This feature is always present and is visible in most tokens, allowing for use of the full dataset and without introducing a bias by prescreening for the presence of the feature.

### 3.2. Stage 2

Up to 8925 tokens were rated in Stage 2. The raw distribution of the visual rating thresholds as reported by the three raters per each frame-rate denomination is shown on Figure 3. The cumulative percentage of the visual rating difference/change per each frame-rate denomination is reported in Figure 4. No differences were detected at 8000 fps for any of the recordings by any of the raters. At 5333 fps differences were noted in 5% of the ratings, the large majority (70%) of which were ratings for tokens in falsetto register, pressed phonation and high pitch. Frame rates of 4000 fps and above did not report any clinical rating changes. A 0.2% change of clinical judgment was noted at 3200 fps (only for women), which increased to 2.5% at 2667 fps, to 6.1% at 2286 fps and to 16.2% at 2000 fps (Figure 4). At 95% confidence interval, the threshold for detecting first noticeable difference was 5333 fps and the threshold for clinical rating change was 2667 fps.

Intra-rater agreement was high, ranging from 94% to 100% for first noticeable difference, and moderate to high, ranging from 69% to 81% for change in clinical rating (Table 4). Inter-rater agreement for first noticeable difference was also high, ranging from 85% to 88%. For change in clinical rating, inter-rater agreement was higher between the two SLPs and lower between the SLP and the ENT professionals (Table 5).

A mild correlation was found between Fo and the frame rate at which image differences were first noted. The frame rate to Fo relationship revealed a moderate level of correlation for changes in clinical rating (Table 6).

The ANOVA was conducted across gender, norm/pathology status and phonatory behavior. For first noticeable difference, statistically significant main effects were reported for gender (F=11.8, p=0.001), norm/pathology status (F=11.1, p=0.001) and phonatory behavior (F=2.9, p=0.012), with a statistically significant interaction noted between gender and norm/pathology status (F=15.9, p<0.001). The Tukey post-hoc test for phonatory behavior revealed that only the category “breathy phonation” was statistically significantly different from “high pitch” and from “falsetto”.

Statistically significant main effects for change in clinical rating were noted across all independent variables gender (F=105.9, p<0.001), norm/pathology status (F=23.0, p<0.001) and phonatory behavior (F=23.9, p<0.001), with interaction noted between gender and norm/pathology status (F=10.4, p=0.001). The results from the Tukey post-hoc test for phonatory behavior are shown in Table 7.

## 4. Discussion

### 4.1. Stage 1

All nine vibratory features (Table 2) were assessed across two phonatory behaviors (habitual and high pitch), and no visual differences were noted at 8000 fps for any of these features. This finding suggests that 8000 fps were sufficient to accurately assess all clinically relevant vibratory features without concerns of image degradation. Glottal edge was found to be highly sensitive to frame rate, and was chosen for analysis in Stage 2 because, in addition to the frame rate high sensitivity, it is an easily-defined feature that is always present during the glottal cycle, and can be reliably rated. The vibratory features of contact and loss of contact demonstrated similar levels sensitivity (Table 3), because those are glottal-edge-based features, i.e. they are ratings of the contact and loss of contact of the glottal edge. Although less reliable among raters, within raters these two features were rated consistently with the glottal edge. Thus, we felt that choosing the glottal edge feature for further analysis represents well the features of contact and loss of contact, as well.

However, it should be noted that the two mucosal wave-related vibratory features, magnitude and extent, showed similar sensitivity to frame rate (Table 3) but were not selected for further analysis due to lower rater reliability and the lack of consistent presence. Although the results from Stage 1 showed very similar results among glottal edge and the two mucosal wave feature, there is a possibility that a Stage-2 analysis based on mucosal wave could deem slightly different results, especially for the clinical ratings change. However, the lower reliability in rating the mucosal wave represents a challenge and requires different research design of the Stage 2 experiment. We recognize this issue as one of the possible limitations of this study. A future follow-up study on mucosal wave could be an option to validate the generalization of our results.

According to Table 3, aperiodicity also showed high sensitivity to fame rates, however, our post-hoc investigation determined that this was mainly due to bias from videostroboscopy and sensitivity of the raters to the rate of playback, rather than actual percept of aperiodicity of the vocal folds. More specifically, when playback rates fell below 15 fps (see Table 1) some of the raters who have substantial clinical experience interpreting videostroboscopy reported first change noticed in aperiodicity. This is a bias of interpreting the image flicker due to low playback rates as “aperiodicity” because in videostroboscopy the instability of stroboscopic synchronization is often interpreted that way and can be perceived similarly, as a flicker [6]. The notion of flicker-based bias is supported by the fact that these raters reported first change noticed in aperiodicity for the 5333 fps denomination only in the second trial sequences, where the reference data were presented at a playback rate of 30 fps and the 5333 fps denomination was accordingly played back at 10 fps (Table 1). However, in the first trial, where references were presented at 60 fps, no difference was noticed at the

5333-fps denomination because the playback rate was 20 fps, and the sensitivity of those raters to aperiodicity dropped to 3200 fps, where the playback rates were 12 fps (Table 1). This is an interesting finding, demonstrating how the flicker of the video images, which appears once the presentation rates fall below the threshold of perceived continuous motion, can bias clinicians to perceive change in periodicity [6]. Clinicians experienced with videostroboscopy are accustomed to interpret a lack of continuous slow motion during vocal-fold vibration as aperiodicity of the vibration. Thus, when the percept of continuous motion in HSV was affected due to low playback rate, raters erroneously interpreted it as aperiodicity, as an artifact of their videostroboscopy training. Interestingly, raters who did not exhibit this bias with aperiodicity had significantly lower sensitivity to the frame rates of 2667 fps and below, which is consistent with the expectations that aperiodicity is the feature less sensitive to frame rate.

The features of amplitude asymmetry, phase asymmetry, and mucus bridges breaking were found to be less sensitive to the HSV frame rate.

### 4.2. Stage 2

As with Stage 1 findings, Figure 4 reports no visual differences noted at 8000 fps across gender, norm/pathology status and phonatory behavior, with Fo varying in a wide range (from 72 to 1000 Hz). Based on this finding, it is *recommended* that future clinical HSV systems allow for rates of 8000 fps. We feel confident generalizing the recommendation for the rate of 8000 fps across all 9 vibratory features studied, given that from up to 7344 ratings in Stage 1 and 8925 ratings in Stage 2, no single rating suggested that even a minor difference in the appearance of a feature was noticed at the 8000-fps denomination. To date, very few clinical studies have used frame rates of 8000 fps, or above. Thus, data on the effect of frame rate on clinical voice assessment in that rate range were not previously reported. Interestingly, the original videokymography system developed by Švec and Schutte used a scan rate of 7812.5 lines per second [7]. This is important, because a lot of knowledge about the vocal-fold vibratory characteristics has been gained using videokymography [8]. Our results suggest that these earlier findings should be free of artifacts due to insufficient scan rates.

The data in this study support the conclusion that currently existing systems utilizing frame rates above 5333 fps are essentially free of error caused by insufficient frame rates. At 5333 fps perceivable differences were noted in only 5% of the ratings (Figure 4), predominantly for tokens carrying higher pixel velocity, i.e. falsetto register, pressed phonation and high pitch. However, these were minor differences that had no effect on the clinical assessment ratings. For all practical purposes, in current systems utilizing frame rates such as 7200, 6000, or even 5500 fps, the degradation of the image due to temporal sampling can be considered negligible.

Our study also found that frame rates as low as 4000 fps did not affect the clinical assessment of glottal edge (Figure 4). Therefore, it is recommended that future clinical HSV systems allow for rates of 8000 fps with a *minimum* requirement of 4000 fps to assess the clinically-relevant glottal edge feature. As stated previously, other vibratory features also demonstrated similar sensitivity to frame rate in Stage 1 but were not analyzed further in

Stage 2. Based on the discussion of the results from Stage 1, it is expected that our results can be generalized to vibratory features that were determined to be less or similarly sensitive to frame rate. Specifically, the recommendations from Stage 2 satisfy the requirements for amplitude asymmetry, phase asymmetry, breaking of mucus bridges, aperiodicity, contact and loss of contact. Although, the results from Stage 1 suggested that mucosal wave magnitude and extent had similar sensitivity to frame rate, additional studies could be conducted to determine whether the minimum requirement of 4000 fps generalizes to those mucosal wave features. This caution is raised mainly because the lack of uniformity among clinicians about rating mucosal wave creates problems designing an easily-generalizable experiment, while at the same time some mucosal wave rating protocols could potentially deem greater sensitivity to fame rates that may shift the minimum requirement above 4000 fps.

**Effect of fundamental frequency—**As expected, a positive correlation was found between the Fo and the reported thresholds. For the change in clinical ratings the correlation was moderate, and for the first noticed difference, the correlation was mild. Strong correlations were not necessarily expected because the factor most sensitive to image degradation is expected to be the image pixel velocity. Fo is only one of the factors determining the image pixel velocity, but other factors such as vibratory amplitude, image magnification, and glottal configuration have essential role. Therefore, the correlations established by this study are partial correlations. Nevertheless, the moderate correlation of 0.64 (Table 6) suggests that the minimum required frame rates of 4000 fps should be especially enforced when the clinical protocols include higher Fo.

It is also important to mention that these results do not significantly alter our previously suggested rule of thumb that HSV frame rates need to warrant at least 16 frames per glottal cycle [1]. This general recommendation was established in the past based on: (1) the rationale that 16 snapshots during a single cycle are sufficient to represent the phases of the cycle (approximately 20 degrees per image sample); and (2) empirical evidence over several years of experience showing that for men (Fo≈125Hz) vibratory features appeared correct at 2000 fps, while for women (Fo≈250Hz) features appeared blurred and different at 2000 vs. 4000 fps. The results from this study do not reject this rule but refine it by demonstrating that the relationship between Fo and fps is not strictly linear and there are interactions with several other factors. The 16-frames-per-cycle rule still applies to warrant sufficient representation of the glottal cycle phases. However, this study also suggests that the minimum frame rates shall not fall below 4000 fps.

**Effect of gender pathology and phonatory behavior—**Despite several statistically significant ANOVA main effects and interactions, group-mean differences for first difference noticed are not substantial to affect the recommended rate of 8000 fps. The only substantial group-mean difference for change in clinical rating was gender. Interestingly, gender effects indicate minimum required frame rates can be reduced by up to 20% when assessing adult males (i.e. 3200 fps instead of 4000 fps). Although, it is not practical to design separate protocols for men and women, the ability to decrease frame rates without jeopardizing clinical rating can be used to improve image quality for males that present with overly dark

HSV images. This also means that pre-existing HSV data from male subjects at rates of 3200 fps and above should not pose reliability issues. As reported, norm/pathology status and phonatory behavior also demonstrated statistically significant differences. Although it was important to understand these effect, they did not demonstrate mean differences large enough to be taken into account in conclusion.

Intra- and inter-rater agreement was high due to the paired-comparisons design of the study. Findings from Table 5 suggest that ENT and SLP professionals have similar criteria when rating the first noticed difference, but there may be systematic discrepancies in clinical ratings, which can be explained by the differences in professional background and angle of clinical interest in functional voice assessment.

## 5. Conclusion

It is recommended that future clinical HSV systems allow for rates of 8000 fps with a minimum requirement of 4000 fps. For currently existing systems featuring frame rates above 5333 fps the image degradation is negligible and does not affect clinical assessment. HSV recordings at rates below 4000 fps for women, and 3200 fps for men should be interpreted with caution. Rates of 2000 fps changed the clinical ratings in over 16% of the samples, which could lead to inaccurate functional assessment. These recommendations and minimum requirements need to be especially taken into account for protocols including higher fundamental frequencies. Future studies could be conducted to determine whether the current findings generalize to other vocal-fold vibratory features and objective measures based on HSV.

The purpose of the current study was to investigate the thresholds at which the HSV frame rate visually degrades clinically-relevant vocal-fold vibratory characteristics, while minimizing the interactions with other technical factors. The derived recommendations were based on monochromatic HSV recordings. Future studies could address the effects of the Bayer filtering inherent to all color high-speed cameras [9], as well as the influence of the camera sensitivity, spatial pixel resolution, image dynamic range, optical zooming, endoscope type, endoscopic angle, and other factors that can further influence the requirements for camera speed.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

Funding for this study was provided by the National Institutes of Health — NIDCD: R01-DC007640 “Efficacy of Laryngeal High-Speed Videoendoscopy”. Special thanks to Habib Moukalled, Kimberly Hufnagel, Shannon Batson, and Heather Bonilha for contributions in early stages of this study.

## References

1. Deliyski, DD. Laryngeal high-speed videoendoscopy. In: Kendall, KA., Leonard, RJ., editors. Laryngeal Evaluation: Indirect Laryngoscopy to High-Speed Digital Imaging. Thieme Medical Publishers; New York: 2010. p. 243-270.

2. Shaw HS, Deliyski DD. Mucosal wave: A normophonic study across visualization techniques. Journal of Voice. 2008; 22(1):23–33. [PubMed: 17014988]

3. Deliyski DD, Petrushev PP, Bonilha HS, Gerlach TT, Martin-Harris B, Hillman RE. Clinical Implementation of Laryngeal High-Speed Videoendoscopy: Challenges and Evolution. Folia Phoniatrica et Logopaedica. 2008; 60:33–44. [PubMed: 18057909]

4. Ikuma T, Kunduk M, McWhorter A. Mitigation of temporal aliasing via harmonic modeling of laryngeal waveforms in high-speed videoendoscopy. Journal of the Acoustical Society of America. 2012; 132(3):1636–1645. [PubMed: 22978892]

5. Ikuma, T., McWhorter, A., Kunduk, M. Effects of frame rates and window size in objective analysis of high-speed videoendoscopy data using harmonic models. In: Deliyski, DD., editor. Proceedings of the 10th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research. AQL Press; Cincinnati, Ohio: 2013. p. 49-50.

6. Mehta DD, Deliyski DD, Hillman RE. Why laryngeal stroboscopy really works: Clarifying misconceptions surrounding Talbot's law and the persistence of vision. Journal of Speech, Language, and Hearing Research. 2010; 53(3):1263–1267.

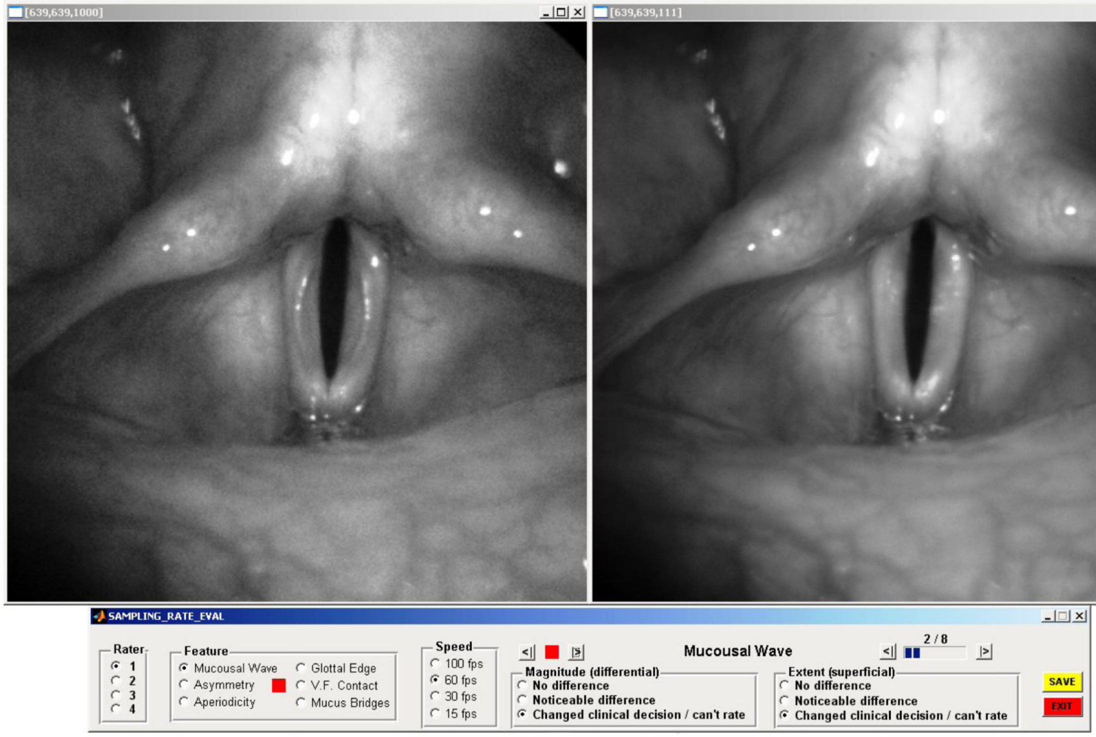
7. Švec JG, Schutte HK. Videokymography: High-speed line scanning of vocal fold vibration. Journal of Voice. 1996; 10(2):201–205. [PubMed: 8734395]

8. Švec JG, Šram F, Schutte HK. Videokymography in voice disorders: what to look for? Annals of Otology. Rhinology and Laryngology. 2007; 116(3):172–180.

9. Bayer, Bryce E. Color imaging array. US patent 3971065 issued 1976-07-20.

## HIGHLIGHTS

- This study relates to voice assessment using laryngeal imaging via high-speed videoendoscopy
- We investigated the impact of frame rates on the assessment of clinical vocal-fold vibratory features
- Results indicated that glottal edge, mucosal wave magnitude and extent, aperiodicity, contact and loss of contact of the vocal folds were the vibratory features most sensitive to frame rate
- Results suggest the recommended rates for laryngeal imaging should be 8000 frames per second
- Results suggest the minimum rates for laryngeal imaging should be 4000 frames per second



Rater: ● 1 ○ 2 ○ 3 ○ 4

Feature: ● Mucousal Wave ○ Asymmetry ○ Aperiodicity ○ Glottal Edge ○ V.F. Contact ○ Mucus Bridges

Speed: ○ 100 fps ● 60 fps ○ 30 fps ○ 15 fps

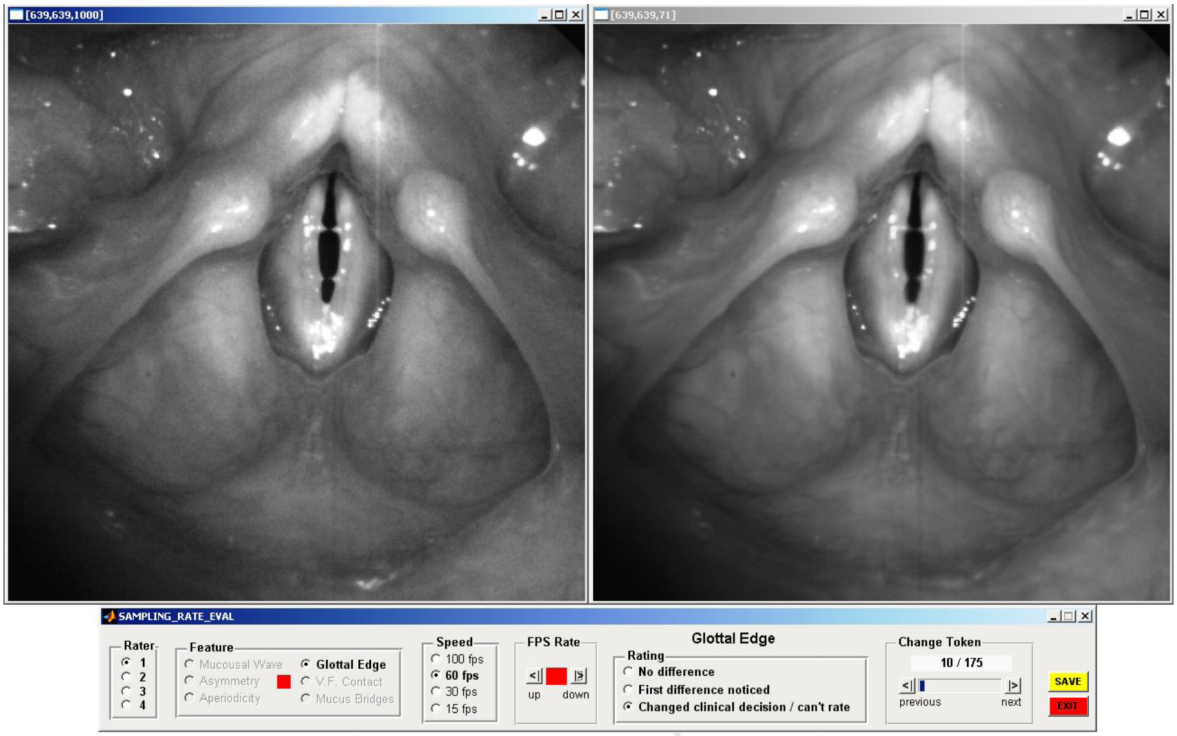
Mucousal Wave 2 / 8

Magnitude (differential): ○ No difference ○ Noticeable difference ● Changed clinical decision / can't rate

Extent (superficial): ○ No difference ○ Noticeable difference ● Changed clinical decision / can't rate

**Figure 1.**

Customized software graphic interface used in Stage 1 of the study, which allowed users to compare the reference video on the left (16000 fps) to the downsampled video on the right. Note the mucosal wave is visible on the left, but is not visible on the right. The electronic version of the article contains videos demonstrating the degradation of the mucosal wave feature shown in this figure, including a short video sequence of the 16000-fps reference and 3 corresponding downsampled denominations at: 8000 fps, 4000 fps and 2000 fps.



Rater: ● 1 ○ 2 ○ 3 ○ 4

Feature: ○ Mucousal Wave ○ Asymmetry ○ Aperiodicity 🟥 ● Glottal Edge ○ V.F. Contact ○ Mucus Bridges

Speed: ○ 100 fps ● 60 fps ○ 30 fps ○ 15 fps

FPS Rate: <| 🟥 |> up down

Glottal Edge

Rating: ○ No difference ○ First difference noticed ● Changed clinical decision / can't rate

Change Token: 10 / 175 <| |> previous next

**Figure 2.**

Customized software graphic interface used in Stage 2 of the study. Note the glottal edge in the downsampled (right) image appears blurred compared to the 16000-fps reference (left) image.

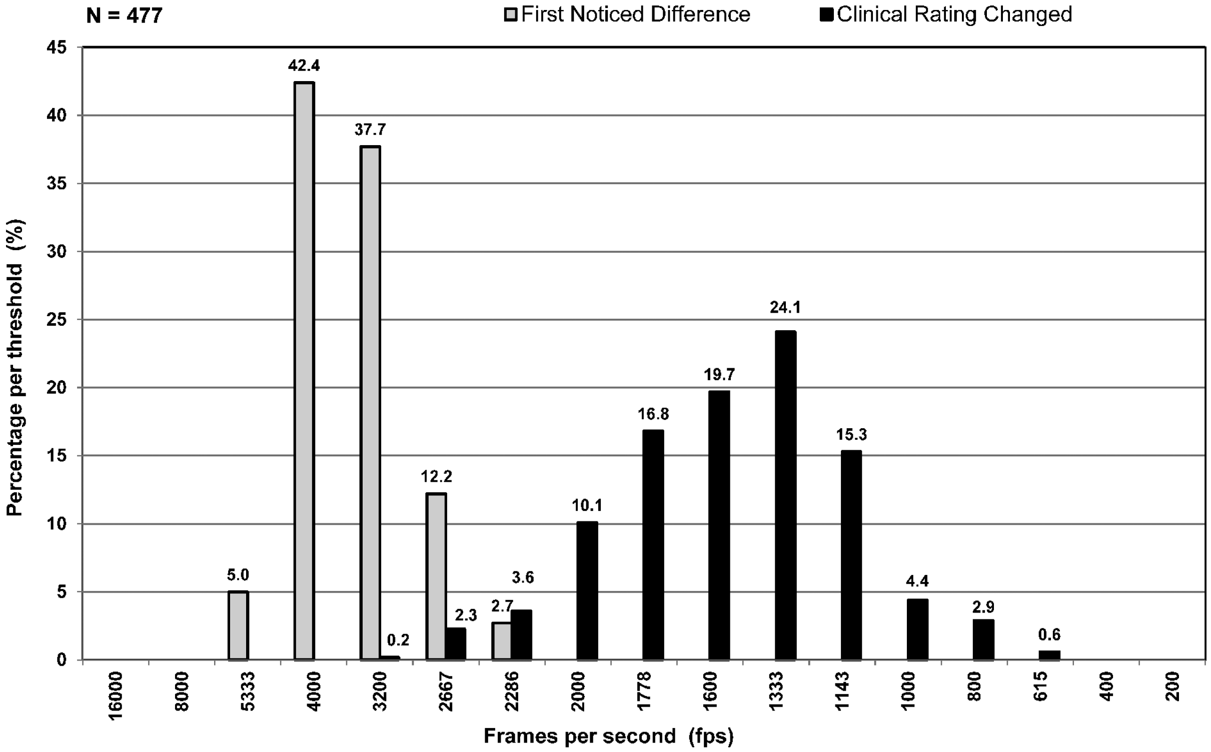


**Figure 3.**

Distribution of the visual rating thresholds for first noticed difference and change in clinical rating in the glottal edge vibratory feature. No difference was noted at 8000 fps by any of the raters for any of the 477 recordings rated. No clinical rating change was reported for frame rates of 4000 fps and above. Note that this figure reports only the raw percentage of the thresholds as established per each frame-rate denomination. For the cumulative effect of image degradation see Figure 4.

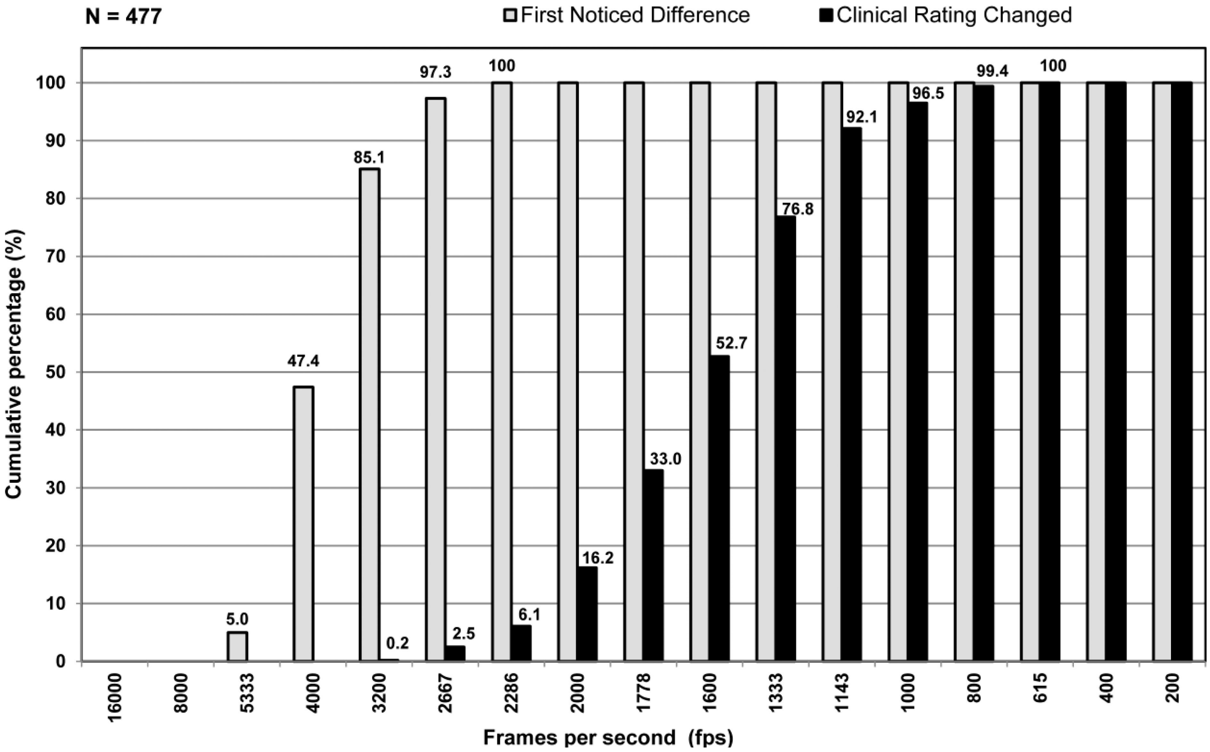


**Figure 4.**

Cumulative percentage of the visual rating differences/changes that have been reported for first noticed difference and clinical rating change in the glottal edge vibratory feature.

## Table 1

Downsampling scheme for 16000-fps tokens with the corresponding frame rates, resulting integration times and playback rates.

| Down-sampling Ratio | Frame Rate (fps) | Integration Time (Ms) | Number of Frames | Playback Rate Trial 1 (fps) | Playback Rates Trial 2 (fps) |
|---|---|---|---|---|---|
| 1 | 16000 | 61 | 1000 | 60 | 30 |
| 2 | 8000 | 123.5 | 500 | 30 | 15 |
| 3 | 5333 | 186 | 333 | 20 | 10[*] |
| 4 | 4000 | 248.5 | 250 | 15 | 7.5[*] |
| 5 | 3200 | 311 | 200 | 12[*] | 6[*] |
| 6 | 2667 | 373.5 | 167 | 10[*] | 5[*] |
| 7 | 2286 | 436 | 143 | 8.6[*] | 4.3[*] |
| 8 | 2000 | 498.5 | 125 | 7.5[*] | 3.8[*] |
| 9 | 1778 | 561 | 111 | 6.7[*] | 3.3[*] |
| 10 | 1600 | 623.5 | 100 | 6[*] | 3[*] |
| 12 | 1333 | 748.5 | 83 | 5[*] | 2.5[*] |
| 14 | 1143 | 873.5 | 71 | 4.3[*] | 2.1[*] |
| 16 | 1000 | 998.5 | 63 | 3.8[*] | 1.9[*] |
| 20 | 800 | 1248.5 | 50 | 3[*] | 1.5[*] |
| 26 | 615 | 1623.5 | 39 | 2.3[*] | 1.2[*] |
| 40 | 400 | 2498.5 | 25 | 1.5[*] | 0.8[*] |
| 80 | 200 | 4998.5 | 13 | 0.8[*] | 0.4[*] |

[*] Indicates playback rates that may fall below the threshold of perceived continuous motion, adding visual effects of flicker.

## Table 2

Definitions of the nine vibratory characteristics used to compare the downsampled tokens to the original 16000-fps recordings.

| Clinical Feature | Definition |
|---|---|
| Mucosal Wave Magnitude | Maximum difference between the lower and upper vocal fold margins during the closing phase (i.e. vertical phase difference or divergence angle) |
| Mucosal Wave Extent | Extent of lateral propagation of the mucosa during the closing phase |
| Amplitude Asymmetry | Amplitude difference between left-right vocal-fold vibration |
| Phase Asymmetry | Phase difference between left-right vocal-fold vibration |
| Aperiodicity | Variation of the period of the glottal vibratory cycle |
| Glottal Edge | Smoothness and shape of the vibrating vocal-fold edges |
| Contact | Realization of contact of the vocal folds during vibration |
| Loss of Contact | Loss of contact of the vocal folds during vibration |
| Mucus Bridges Breaking | Release of mucus strand bridging the glottis |

### Table 3

Stage 1 results of maximum sensitivity to first noticed difference and intra-rater agreement combined across the three raters for each vibratory feature. Glottal edge was among the 6 features most sensitive to frame rate, but also had the highest intra-rater reliability amid those features.

| Clinical Feature | First Difference Noticed (fps) | Intra-Rater Agreement (%) |
|---|---|---|
| Mucosal Wave Magnitude | 5333 | 88 |
| Mucosal Wave Extent | 5333 | 67 |
| Amplitude Asymmetry | 4000 | 71 |
| Phase Asymmetry | 4000 | 79 |
| Aperiodicity | 5333 | 50 |
| Glottal Edge | 5333 | 92 |
| Contact | 5333 | 75 |
| Loss of Contact | 5333 | 79 |
| Mucus Bridges Breaking | 4000 | 83 |

## Table 4

Stage 2 results for intra-rater agreement for the glottal edge vibratory feature (%)

| N=159 | First Difference Noticed | Clinical Rating Changed |
|---|---|---|
| Rater 1 (ENT) | 94% | 69% |
| Rater 2 (SLP) | 94% | 81% |
| Rater 3 (SLP) | 100% | 69% |

## Table 5

Stage 2 results for inter-rater agreement for the glottal edge vibratory feature (%)

| **N=159** | **First Difference Noticed** | | **Clinical Rating Changed** | |
|---|---|---|---|---|
| | **Rater 2 (SLP)** | **Rater 3 (SLP)** | **Rater 2 (SLP)** | **Rater 3 (SLP)** |
| Rater 1 (ENT) | 85% | 88% | 64% | 69% |
| Rater 2 (SLP) | | 87% | | 76% |

## Table 6

Spearman's rho correlations between frame rate and fundamental frequency for the glottal edge vibratory feature.

| N=159 | First Difference Noticed | Clinical Rating Changed |
|---|---|---|
| Spearman's Rho | 0.31 | 0.64 |
| p-value | 0.0000 | 0.0000 |

## Table 7

Statistically significant differences per phonatory behavior category determined using Tukey post-hoc test on the data for clinical ratings change for the glottal edge vibratory feature.

| Behavior Category | Statistically Different From |
|---|---|
| habitual pitch | high pitch, pressed phonation and falsetto |
| high pitch | habitual, low pitch, breathy phonation and falsetto |
| low pitch | high pitch, pressed phonation and falsetto |
| pressed phonation | habitual, low pitch and breathy phonation |
| falsetto | habitual, low, high pitch and breathy phonation |