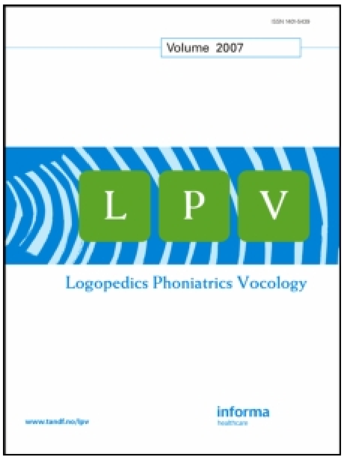


## Logopedics Phoniatrics Vocology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713713058

# Not just sound: Supplementing the voice range profile with the singer's own perceptions of vocal challenges

Anick Lamarche [a]; Sten Ternström [a]; Stellan Hertegård [b]
[a] Kungliga Tekniska Högskolan, School of Computer Science and Communication, Department of Speech, Music and Hearing, Stockholm, Sweden [b] Karolinska Institute, Department of Logopedics and Phoniatrics, Ear Nose Throat Cochlea Clinic, Stockholm, Sweden

First Published on: 13 August 2008

*Logopedics Phoniatrics Vocology*
2008, 1–8, iFirst article



**ORIGINAL ARTICLE**

# Not just sound: Supplementing the voice range profile with the singer's own perceptions of vocal challenges

ANICK LAMARCHE[1], STEN TERNSTRÖM[1] & STELLAN HERTEGÅRD[2]

[1]*Kungliga Tekniska Högskolan, School of Computer Science and Communication, Department of Speech, Music and Hearing, Stockholm, Sweden,* [2]*Karolinska Institute, Department of Logopedics and Phoniatrics, Ear Nose Throat Cochlea Clinic, Stockholm, Sweden*



**Abstract**
A commercial phonetograph was complemented with a response button, such that presses resulted in marked regions in the voice range profile (VRP). This study reports the VRP data of 16 healthy female professionally trained singers (7 mezzo-sopranos and 9 sopranos). Subjects pressed the button to indicate sensations of vocal instability or reduced control during phonation. Each press thereby marked potential areas of difficulty. A method is presented to quantify the consistency of button use for repeated tasks. The pattern of button presses was significantly consistent within subjects. As expected, the singers pressed at the extremes of VRP contours as well as at register transitions. These results and the potential of the method for the assessment of vocal problems of singers are discussed.

**Key words:** *Evaluation tool, self-perception, singing voice, voice assessment, Voice Handicap Index, phonetogram, voice range profile*

## Introduction

Computerized phonetograms or voice range profiles (VRP) are now easily accessible and are often part of standard clinical equipment. Current VRP systems often augment the range data with additional metrics that describe voice quality (e.g. crest factor, jitter, and shimmer) (1), and they provide these data not only on the bounding contours but also over the interior of the VRP. The phonetograph is an appealing tool for voice assessment (2), as it provides a summarizing voice image in which the specific interactions are depicted between an observed entity on the one hand (usually, the ability to phonate), and level and frequency on the other. In clinical settings, the VRP is commonly used as an objective acoustic measure, in combination with other subjective measures. The perceptual aspect of the clinical evaluation of voice is two-sided: as perceived by the patient and as perceived by the therapist.

Firstly, the modern and widely accepted health definition proposed by the World Health Organization (WHO) states that:

> Health is a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity (3).

This shift in health definition has strongly impacted on the health care system. Clinical approaches increasingly value and measure patient self-perception and experience as an integral part of the overall evaluation process. In voice clinic environments, instruments for self-reporting such as the Voice Handicap Index (VHI) introduced in 1997, are commonly included in protocols (4). Much attention is directed to the patient's own vocal experience of the reported problem. Some researchers have attempted to adapt the VHI approach to the specific reality/concerns of the professional singer (5,6). It is recognized that there is a need to assess patient self-perceptions and also to adapt this assessment in response to the particular needs of certain groups of patients. Furthermore, a survey of the literature reveals increasing interest in the relationship of subjective assessment to other

Correspondence: Anick Lamarche (MMus). Fax: +46-8-790-7854. E-mail: anick@nada.kth.se

*(Accepted 23 May 2008)*

ISSN 1401-5439 print/ISSN 1651-2022 online 
DOI: 10.1080/14015430802239759

measures (acoustic or perceptual) common in clinical practice (7).

Secondly, external perception of voice quality is crucial to voice evaluation. Instruments such as the Grade, Roughness, Breathiness, Asthenia, Strain (GRBAS) (8) and Consensus Auditory Perceptual Evaluation–Voice (CAPE–V) (9) protocols serve to objectify and standardize the clinician's perception in evaluation (10–12). Despite some unresolved issues in this area (13,14), training perception provides satisfactory and fairly robust results (15). However, clinicians mostly train their perceptual judgment with the motivation to recognize phonatory failures in spoken voice. The voice problems of professional singers are often very specific and not always detected by mainstream voice assessment protocols, which typically are designed for speech. To singers, voice effort and pitch are well known dimensions and their voice problems typically occur at certain combinations of intensity and pitch. Therefore, the level-versus-pitch map of the VRP should be well suited for isolating problematic phonation. However, vocal problems might be so subtle that, even if they are perceived by the singer, they do not show up in the acoustic features displayed in the VRP, or indeed in any acoustic dimension.

By tapping into the singer's own perceptions, it might be possible to augment the VRP with non-acoustic but singer-relevant information. One way of doing this could be to allow the singer to signal some aspect of his/her production in a way that is automatically registered in context by the phonetograph. Such perceptual data could help bridge the gaps that exist between the singer's experience of voice production and what can be perceived or measured by the clinician.

Here, the use of a simple push-button for combining subjective immediate self-perceptual information and objective vocal measurements was investigated. Singers were asked to press the button whenever they felt that they did not have adequate control of their voice and/or when they felt discomfort. Each press of the button was registered and displayed as a black mark at the corresponding point in the VRP. Expectations for this group of singers were that button presses would generally be located at VRP extremes and would be mostly incidental or transitory. This paper describes some of the issues encountered with this approach, and how the reliability and the validity of such marker data might be assessed.

## Methods

### *Signal acquisition*

A computerized phonetograph, *Phog* (Version 2.00.10, Hitech Development AB, Sweden) was used in combination with a digital signal processing (DSP) sound card (BlueWaves LSI-PC/C32 board). The phonetograph was modified by author ST and Svante Granqvist to record also presses of an external hand-held button.

Each down-press of the button generated a 73 ms pulse, regardless of how long the button was held down or of how hard it was pressed. The duration of the button press was discarded, as subjects in development trials would sometimes hold the button pressed for a second or two—for example, over the ends of tones with large drops in sound pressure level (SPL)—resulting in irrelevant smears in the data and ambiguities in the subsequent interpretation. As illustrated in Figure 1, the binary pulses were recorded in a vacant channel, in parallel with the phonetograph's fundamental frequency ($F_0$), SPL, and voice quality parameters. Only button presses that were made during phonation were mapped into the VRP display, since their position would otherwise be undefined.

All recordings were conducted in a sound-treated and isolated but not anechoic room (volume 45 m$^3$, ceiling height 3 m, reverberation time, $T_{30} = 0.1$ s, reverberation radius $> 1.2$ m across the spectrum, and 0.5 m deep absorbents). Singers were asked to perform in a singing stance at 30 cm mouth-to-microphone distance. For a few subjects, this distance was increased to 1 m (see Discussion) when the singer's SPL would exceed the 120 dB limit of the phonetograph; for these, the calibration of the sound file was consequently altered by a factor of 10.5 dB. A condenser microphone (Brüel & Kjaer, model 4003, Denmark) was used with a preamplifier (Brüel & Kjaer, model 2812) and a line amplifier (Nyvalla-DSP Audio Interface Box). Singers used a single earphone piece (Bassonic-Champion 4939, USA) to hear prompting tones during tasking.

The *Phog* system's criterion for detecting voicing is not a level threshold but a phonation period-time stability threshold. The running standard deviation in period-time over seven consecutive cycles is computed, and if the standard deviation is small enough, voicing is detected. This threshold was set to 0.2% or 75 cents standard deviation, given that even with a large vibrato, $F_0$ does not change by as much as 75 cents in seven glottal cycles. For this reason, tones with vibrato were reliably tracked. The

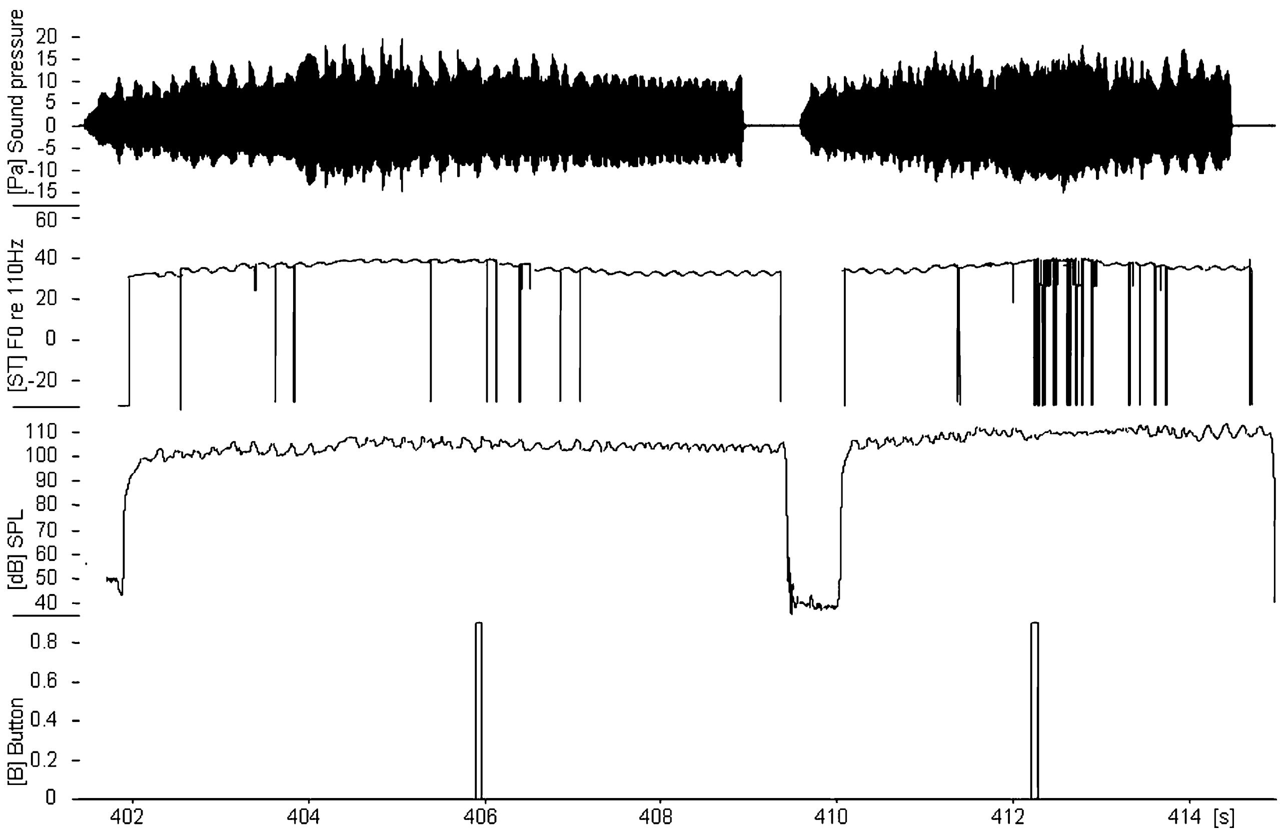


Figure 1. Signal file including, from the top, channels with audio, $F_0$, sound pressure level, and the button state.



phonetograph's resolution (cell size) in $F_0$ is one semitone, and in SPL it is one decibel. There was also a threshold for accumulated time: a cell was included in the VRP if it had been visited for a total of at least 25 milliseconds. This choice of time threshold meant that a single excursion of a vibrato cycle would be sufficiently long to be included for display and analysis.

### *Procedures and subjects*

The subjects could communicate with the investigator by intercom, and visual contact through a window was possible. The subjects were not able to see the phonetograph display. This prevented them from being distracted by visual concurrent feedback as they performed the singing tasks and thereby enhanced an introvert locus of attention as they used the button. Singers were asked to perform the tasks on the phoneme /a:/. The three tasks of this experiment were as follows.

Task 1: A performance voice range profile was recorded. For this kind of VRP, subjects were asked to use a performance voice with their habitual vibrato at all times and to phonate as they deemed musically acceptable. The task was designed to resemble a typical vocalise, with a minor or major triad carrier. In a first step, subjects were asked to perform a *messa di voce* (a gradual rise and fall of musical dynamic on one stable frequency) on a comfortable tone in order to exercise their full performance mode dynamic range. Following this exercise, subjects sang the ascending and descending triad carrier in *pianissimo* as well as in *mezzo forte* and *fortissimo* (soft, medium, and loud) musical dynamics. Singers could break as they pleased and were given freedom in structuring their performance (phrasing, breathing, and pace). In order to test the consistency of behaviour, singers replicated Task 1 later in the procedure.

Task 2: A performance VRP was recorded for a discrete pitch exercise. A prompting pitch was played to the singer in an earphone. The singer was then asked to sing this tone in *mezzo forte*, *pianissimo* and *fortissimo* (medium, soft, and loud dynamics). The prompted intervals followed recommendations by Schutte and Seidner (16). The frequencies equivalent to the musical notes C-E-G-A in several octaves were tested across the singer's range. Again, performance voice only was required.

Task 3: Singers performed their best audition aria with lyrics. Concerning the use of the button, the singers were given the following instructions:

As you sing, press the button at any time you feel vocal instability or discomfort. Aim at communicating your sensations during your performance.

A total of 23 classical female singers were recorded; 5 subjects took part in an initial pilot phase of the experiment. All subjects had 4 years or more of professional training. Subjects 8 and 21 were excluded from analysis on the basis of vocal problems reported in a questionnaire. The analysis, excluding the pilot data, was thus conducted on the remaining 16 recordings of which there were 7 mezzo-sopranos and 9 sopranos. Finally, subject 7 needed to be excluded from the replicated task analysis due to technical difficulties. All 16 female participants were involved also in another gender-specific project, so for convenience this study reports results obtained with female singers only.



### *Button data validity*

In this study, subjects were asked to perform two motor tasks simultaneously, one of them corresponding to a perceptual judgment of internal performance experience. It may be expected that such a combination of tasks and the request for explicit self-awareness might lead to various kinds of errors in the position of the button marks in the VRP.

### *Sources of errors in the button timing*

*Reaction time.* Simply said, reaction time is the time taken between a stimulus and a movement. This time could also include a choice before the execution of the movement. Reaction time depends on nerve connections and signal pathways. Reaction times to sounds are similar to reaction times found for touch stimuli and are in the order of 140–160 ms (17,18). Despite differences in reactions to different stimuli, the time for motor preparation and response is constant for all types of reaction time tasks. Indeed, reaction time is linked to processing ('the space bar task') (19). Due to reaction delays, there is a possibility that when the subject presses the button, she is already near the end of the tone (SPL is descending rapidly) or she has already moved on to the next tone ($F_0$ may have changed), in which cases the mark will be erroneously placed.

While this study is not concerned with reaction time, its consequences need to be taken into account. The tasks were therefore executed at a pace that would allow the singer ample time to react. By asking the singer to phonate a minimum of 2–3 seconds per token, an average reaction time of 150 ms was accommodated into the task. A short training session prior to recording was included to decrease the singer's reaction time. This training session included either 'Ridente la calma' by Mozart or 'Somewhere over the rainbow' by Arlen. Initial phrases were sung in widely different keys in order to provoke some feeling of discomfort that would make the subject press the button.

*Vibrato.* The frequency and amplitude modulations incurred by the vibrato will introduce some uncertainty into the precise location of the button marks. Sources report different frequency extent values when it comes to the assessment of vibrato. On average a typical frequency swing can vary from 71 cents up to 128 cents (20,21). The level swing induced by vibrato was very variable, both from tone to tone and from singer to singer; according to this study's data observations, it could easily range from 0 to 5 dB. This seems to agree with previous reports on amplitude variation and vibrato (21). However, it is not practical to try to adapt the VRP to each singer's vibrato. The instant of pressing the button is probably quite unrelated to the vibrato cycle, so the vibrato will add a small random component to the SPL and $F_0$ co-ordinates of the button markers. This uncertainty needs to be borne in mind when examining the button marks in the VRP. In principle, the vibrato may be filtered out by technical means, but this would noticeably increase the response time of the device.

### *Post-task validation by the singers*

As part of a post-recording questionnaire, singers rated how well the button markings on their VRPs reflected their performance experience and typical areas of vocal challenges. It seemed important to cross-check validity by giving the subjects themselves the chance to evaluate the instrument. Our aim was to find out if the display made sense to them with the experience still vivid in their minds. This type of post-task questionnaire included a definition of the VRP and was answered in writing. A visual presentation of the subjects' VRP for the repeated task was available to help the evaluation. When asked 'Do the button presses relate well to your own singing experience today?', visual analogue scale (a 10-cm line) ratings were very high, with 94% of the singers rating 7.5 and above.

These results suggest that, on the whole, the singers found the button marks to be consistent with their recollections of their performance. Similar results were obtained, with 91% answering in the affirmative to 'Are the highlighted portions of your VRP typical areas of vocal difficulty or/and limits?' These positive results suggest that singers viewed the button as a possible way to communicate their perceptions, and provide some support for the validity of the button data. The use of the button was examined for each individual and was not generalized to the group.

*Assessment of reliability*

Previous studies report long-term and short-term subject variability in recording VRPs (22). Some degree of short-term variability in the overall VRP recordings as well as in the use of the button was therefore expected. If the new button data are to be useful, they need to be both valid and reproducible. When a subject repeats a task we expect the two sets of responses to be fairly similar. Since vocal difficulties can be transitory, especially in healthy singers, and because the singer's attention can wander, responses would never be identical. Visual inspection revealed similarities in most cases, but gave no quantification. However, if the button reports are reliable, repeating a task should give greater similarity to the first response than to the outcome of a random process with the same number of button presses. A procedure based on this requirement was devised to quantify the similarity of button presses in two VRPs, henceforth called VRP-A and VRP-B. It assessed only the similarity of the button data, not the similarity of VRPs A and B as a whole. For each recording, the *Phog* system saves both a file containing the VRP data matrix and a multitrack signal file containing the audio with a host of extracted parameters. For the custom analysis needed here, Matlab scripts were created to read the signal file and reconstruct the VRP matrix including the button data. Figure 2 depicts such a reconstruction. For reasons explained in Appendix A, a region around each button mark was constructed, as shown in Figure 3. The percentage of overlap of button regions in the VRPs A and B was then computed by simply counting coinciding and non-coinciding cells. This percentage of overlap was then used as a similarity score. For further detail concerning the four-step analysis elaborated for this study, see Appendix A.

## Results

The similarity scores and *p*-values for 15 button-VRP pairs are shown in Table I. For 13 of the 15 subjects, the real button overlap percentages obtained were significantly higher than the ones obtained for the mean of 20 iterations of a randomized distribution of presses. The average similarity score was 19.3%.

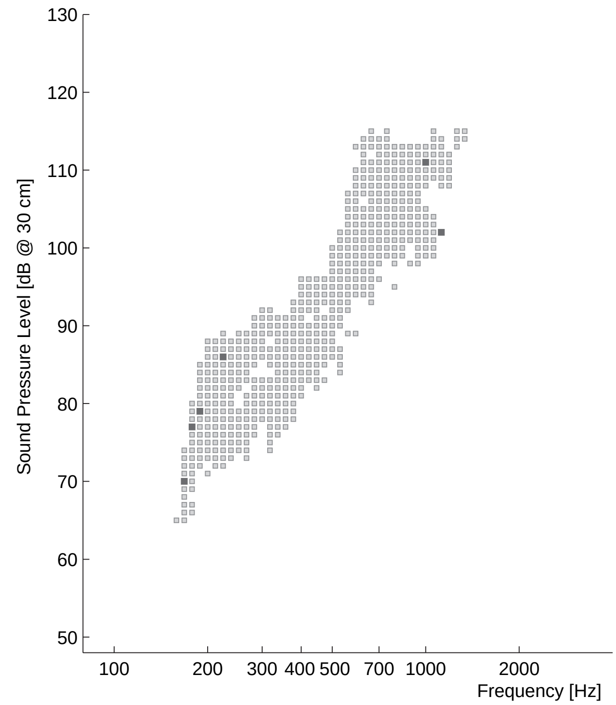


Figure 2. A voice range profile (VRP) which includes the button presses acquired during performance (darker cells).

Results for the similarity scores between Task 1 and Task 2 are given in Table II. When button presses were compared across different tasks, the

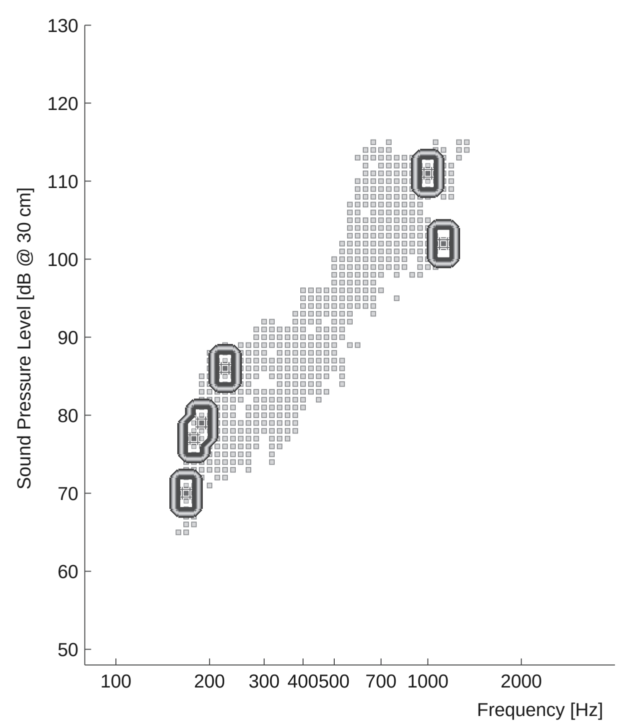


Figure 3. A Matlab reconstruction of a voice range profile (VRP) displaying button presses and button regions. As illustrated, healthy singers tend to press mostly at the boundaries of their voice: the extremes of the VRP.

Table I. Similarity of trials A and B for a repeated task (Task 1). Column 2 gives the similarity score: the percentage of overlap of button regions for trial A and B. Column 3 gives the mean overlap of 20 randomly redistributed A with B trials (an estimate of a Poisson distribution λ). Column 4 gives the probability, assuming the Poisson distribution, of the observed button overlap being an outcome of a random process. Bolded *p*-values are significant, indicating that the subject was replicating presses at higher than chance level.

| Subject | Original overlap% | Random mean overlap% | *p*-value ($p<0.05$) |
|---|---|---|---|
| 6 | 28 | 14.3 | **<0.001** |
| 9 | 9 | 1.6 | **<0.001** |
| 10 | 10 | 6.8 | **0.010** |
| 11 | 30 | 14.1 | **<0.001** |
| 12 | 20 | 8.7 | **<0.001** |
| 13 | 9 | 4.8 | **0.024** |
| 14 | 15 | 5.1 | **<0.001** |
| 15 | 8 | 12.1 | 0.848 |
| 16 | 19 | 12.5 | **0.030** |
| 17 | 8 | 1.5 | **<0.001** |
| 18 | 20 | 7.3 | **<0.001** |
| 19 | 17 | 2.0 | **<0.001** |
| 20 | 35 | 28.9 | 0.110 |
| 22 | 8 | 3.6 | **0.012** |
| 23 | 51 | 8.0 | **<0.001** |

average similarity score dropped to 11.6%, and only 8 of 16 subjects were consistent with themselves across tasks. Real button overlap percentages were nearly always higher than the ones obtained for the mean of 20 iterations of a randomized distribution of presses, but not all of these differences were statistically significant according to our criterion of non-randomness. An exception can be observed in Table II in the case of subjects 9, 14, and 17 where the similarity percentage reported was 0 and where the mean of a random distribution of presses held a higher percentage. In these cases, very few presses of the button were registered for Task 2, hence limiting the chances of an actual overlap with Task 1. When the presses from Task 1 were distributed randomly 20 times, possibilities for overlap became higher and, in turn, the *p*-values reported were amongst the highest in this analysis.

Table II. Similarity of different tasks (Task 1 and Task 2). Column 2 gives the similarity score: the percentage of overlap of button regions for Task 1 and Task 2. Column 3 gives the mean overlap of 20 randomly redistributed Task 1 with Task 2 (an estimate of a Poisson distribution λ). Column 4 gives the probability, assuming the Poisson distribution, of the observed button overlap being an outcome of a random process. Bolded *p*-values are significant.

| Subject | Original overlap% | Random mean overlap% | *p*-value ($p<0.05$) |
|---|---|---|---|
| 6 | 26 | 16.2 | **0.008** |
| 7 | 15 | 11.0 | 0.092 |
| 9 | 0 | 0.7 | 0.489 |
| 10 | 23 | 9.4 | **<0.001** |
| 11 | 23 | 17.6 | 0.084 |
| 12 | 4 | 1.0 | **0.004** |
| 13 | 16 | 7.9 | **0.003** |
| 14 | 0 | 2.5 | 0.918 |
| 15 | 4 | 4.5 | 0.459 |
| 16 | 9 | 6.4 | 0.117 |
| 17 | 0 | 3.3 | 0.962 |
| 18 | 10 | 4.9 | **0.012** |
| 19 | 12 | 4.1 | **<0.001** |
| 20 | 23 | 13.2 | **0.005** |
| 22 | 7 | 4.4 | 0.076 |
| 23 | 13 | 1.9 | **<0.001** |

## Discussion

With a microphone distance of 1 m, room acoustic effects can be of concern even in a heavily treated room. Because such a room (with isolated walls and ceiling) has a single dominating floor reflection, a random variation in SPL of about $\pm 0.8$ dB at 30 cm and $\pm 2$ dB at 100 cm can be predicted. This was closely confirmed by comparing the SPL-time contours in the *Phog* signal files, for real sessions reproduced at 30 and 100 cm using a high-quality loud-speaker in place of the singer. However, this random source of variation, although undesirable, can only make the similarity scores lower and not higher. Hence the similarity tests reported here are slightly more conservative than what would have been the case with a true anechoic room and a truly fixed microphone distance.

The validity of the button as a new device was investigated by testing the consistency of button presses for a singer. Overall, the similarity scores confirmed the subjective visual impression that the button information was not random, but was repeatable and therefore can be assumed to reflect actual difficulties experienced by the singers. Results for this group of subjects are encouraging as they attest to the applicability of the button-mediated responses as a new metric. For the replicated task, the singers demonstrated a significantly consistent behaviour in the use of the button. Expectations formulated at the onset of the experiment were met: button presses were in general located at VRP extremes, and, according to the singers' informal reports, presses were mostly related to momentary vocal difficulties (unprepared onset, phlegm on the folds, vocal limits, etc.). Because the subjects in this group were vocally healthy, yet asked to communicate problems during performance, button presses might be expected to be infrequent and/or range-specific. For a group of injured singers, one might expect the similarity of replicated tasks to be higher and more problem-specific.

When assessed across tasks, the similarity scores were lower but continued to support consistent behaviour for half of the subject group. There were no expectations that the button presses would be particularly reproducible across tasks, but the above-random effect for a part of the group is nonetheless worth mentioning. The reduction in similarity scores is not surprising since the performance VRP's overall shape for each task would be different. It follows that areas highlighted by button presses for one task might not even be registered in a different task. Consequently, if this method were used, say, for assessing pre-and post-therapy, the task would have to be the same, pre and post.

All in all, this new method could have clinical potential to document the performance experiences of singers. An advantage of this method is that the perceptual judgment is instantaneous and most likely intrinsically related to the experience of the moment (23). If the aim is to evaluate the singing voice and understand its failures in relation to stage performance and injury, then it is of interest to identify the phonatory conditions (pitch and effort level) that invoke a problem. For the singer who has trained his/her kinaesthetic sensitivity and is a vocal athlete, the instants of vocal problems caused by injury are very specific. The button, in this case could allow the depiction of those problematic areas and possibly assist effectively the evaluation and/or even the rehabilitation process.

It may be noted that, in soprano singing, fundamental frequency tuning to the first formant ($F_0$–$F_1$) practically eliminates sound pressure level (SPL) variability across vowels (24). Nonetheless, the phoneme /a:/ (and its variations) was used for all tasks with the exception of the aria excerpt. This decision was taken for the sake of comparability with other studies in the literature and to follow recommendations by Gramming and Sundberg (25).

*Conclusion and future work*

A VRP was augmented with a button to tap into the singer's perceptions as he/she performs. In using this button, singers met initial expectations that healthy singers would use the button to communicate transitory difficulties and press mostly at the extremes of their performance range. An attempt to quantify the reproducibility demonstrated that singers reproduced their use of the button at least more closely than would a random process. It seems reasonable to conclude that singers, to some degree, can communicate their perceptions with a button as they perform.

Our intention is to test the use of the button-augmented VRP with injured singing voices, as an integral part of the voice evaluation process. It remains to be seen if the use of the button in this case becomes more consistent and problem-specific. It appears that the combination of subjective self-perception and the objective VRP has the potential to offer a new layer of understanding of singing voice, in the research laboratory, the clinic, and the singing studio.

***Declaration of interest:*** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

## References

1. Pabon JPH. Objective acoustic voice-quality parameters in the computer phonetogram. J Voice. 1991;5:203–16.
2. Holmberg EB, Ihre E, Södersten M. Phonetograms as a tool in the voice clinic: Changes across voice therapy for patients with vocal fatigue. Logoped Phoniatr Vocol. 2007;32:113–27.
3. World Health Organization. International Classification of Impairments, Disabilities and Handicaps. Geneva: World Health Organization; 1980.
4. Jacobson BH, Johnson A, Grywalski C, Silbergleit A, Jacobson G, Benninger MS, et al. The Voice Handicap Index (VHI): development and validation. Am J Speech Lang Pathol. 1997;6:66–70.
5. Morsomme D, Simon CH, Jamart J, Remacle M, Verduyckt I. A proposal to adapt the voice handicap index to the singing voice]. Rev Laryngol Otol Rhinol (Bord). 2005;126:305–13.
6. Rosen CA, Murry T. Voice handicap index in singers. J Voice. 2000;14:370–7.
7. Lehto L, Laaksonen L, Vilkman E, Alku P. Occupational voice complaints and objective acoustic measurements—do they correlate? Logoped Phoniatr Vocol. 2006;31:147–52.
8. Hammarberg B, Gauffin J. Perceptual and Acoustic Characteristics of Quality Differences in Pathological Voices as Related to Physiological Aspects. In: Fujimura O, Hirano M, editors. Vocal Fold Physiology, Voice Quality Control. San Diego: Singular Publishing Group; 1995. p. 283–303.
9. American Speech-Language-Hearing Association. Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V): ASHA Special Interest Division 3, Voice and Voice Disorders. 2002–2006. Available from: http://www.asha.org/NR/rdonlyres/C6E5F616-972F-445A-AA40-7936BB49FCE3/ 0/CAPEVprocedures.pdf (accessed May 2008).
10. Oates JM, Bain B, Davis P, Chapman J, Kenny D. Development of an auditory-perceptual rating instrument for the operatic singing voice. J Voice. 2006;20:71–81.
11. Kenny DT, Mitchell HF. Acoustic and perceptual appraisal of vocal gestures in the female classical voice. J Voice. 2006;20: 55–70.
12. Gelfer MP. Perceptual attributes of voice: Development and use of rating scales. J Voice. 1988;2:320–36.
13. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial and a framework for future research. J Speech Hear Res. 1996;36: 21–40.
14. Shrivastav R, Sapienza CM, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. J Speech Lang Hear Res. 2005;48:23–35.
15. Hammarberg B, Fritzell B, Gauffin J, Sundberg J, Wedin L. Perceptual and acoustic correlates of abnormal voice qualities. Acta Otolaryngol. 1980;90:441–51.

16. Schutte HK, Seidner W. Recommendation by the Union of European Phoniatricians (UEP): standardizing voice area measurements/phonetography. Folia Phoniatr. 1994;35: 286–8.
17. Galton F. On instruments for (1) testing perception of differences of tint and for (2) determining reaction time. J Royal Anthr Inst. 1899;19:27–9.
18. Brebner JT, Welford AT. Introduction: an historical background sketch. In: Welford AT, editor. Reaction Times. New York: Academic Press; 1980. p. 1–23.
19. Miller JO, Low K. Motor processes in simple, go/no-go, and choice reaction time tasks: a psychophysiological analysis. J Exp Psychol Hum Percept Perform. 2001;27:266.
20. Prame E. Vibrato extent and intonation in professional Western lyric singing. J Acoust Soc Am. 1997;102:616–21.
21. Nawka T, Seidner W, Cebulla M. Experiments on intentionally influencing the singer's vibrato. In: Friberg A, Iwarsson J, Jansson E, Sundberg J, editors. SMAC93. Stockholm: Stockholm Music Acoustics Conference, Royal Swedish Academy of Music; 1993. p. 221–7.
22. Gramming P, Sundberg J, Akerlund L. Variability of phonetograms. Folia Phoniatr (Basel). 1991;43:79–92.
23. Darrow AA. Examining the validity of self-report: middle-level singers' ability to predict and assess their sight-singing skills. International Journal of Music Education. 2006;24:21–9.
24. Lamarche A, Ternström S. An Exploration of Skin Acceleration Level as a Measure of Phonatory Function in Singing. J Voice. 2008;22:10–22.
25. Gramming P, Sundberg J. Spectrum factors relevant to phonetogram measurement. J Acoust Soc Am. 1988;83: 2352–60.



## Appendix A: *Details on the statistical approach*

Four main steps define the analysis.

1) *Button region definition.* To account for proximity without actual overlap, and also for vibrato-induced variations as described in the article (see Methods), a surrounding region was 'bled' around the co-ordinates of every button press. Based on previous reports and on observations of vibrato behaviour of the recorded singers, the region was chosen to range $\pm 2$ dB in height and $\pm 1$ semitone in width. As seen in Figure 3, each cell marked by button presses becomes the centre of a larger overlap rectangle of $5 \times 3$ cells. This region is somewhat analogous to a proximity weighting function as used in image correlation calculations.

2) *Overlap calculation.* Task 1 was replicated in broken practice style which meant that for each subject, there were two performance voice range profiles (VRPs) of Task 1 with button information available for comparison. The button regions in the VRP-A were overlapped with the button regions in VRP-B, and the percentage of total button region overlap was calculated.

3) *Randomization of the original button region obtained in A.* A high percentage of overlap is not in itself a good measure of similarity, since the degree of random overlap will be higher if the button marks are dense rather than sparse. Rather, we needed to know if the observed overlap in each case was higher than would be expected by chance. Therefore in a third step, the button regions in VRP-A were uniformly repositioned within the total VRP-A area at random, and the overlap with B was recalculated. This randomization was iterated 20 times for each pair of VRPs A and B. The average overlap and the standard deviation for the 20 iterations per subject were calculated in a final step.

4) *Cumulative distribution function.* For small to moderate amounts of overlap, the distribution of overlap outcomes of repeated random trials can be modelled by a Poisson distribution. The Poisson distribution is a discrete probability distribution returning only values greater than or equal to zero. In this context, the cell overlap is discrete: it happens in an integer number of cells. The parameter lambda ($\lambda$) is equivalent to the mean of the Poisson distribution. For each A-B pair of button maps, the average percentage overlap of the 20 random iterations was used as an estimate of the $\lambda$ parameter of the Poisson distribution for that pair. Using the cumulative Poisson distribution function, the probability was calculated that the random overlap percentage for each subject was less than or equal to the real button overlap percentage. Significance was defined as alpha $= 0.05$ for $p$, where $p$ is the probability that the observed percentage of button overlap could have been the outcome of the simulated random process.

Similarity scores were computed also for *different* tasks, comparing button presses from Task 1 to those from Task 2.