The Relationship Between Acoustic Signal Typing and Perceptual Evaluation of Tracheoesophageal Voice Quality for Sustained Vowels

Clapham, R.P.; van As-Brooks, C.J.; van Son, R.J.J.H.; Hilgers, F.J.M.; van den Brekel, M.W.M.

DOI
10.1016/j.jvoice.2014.10.002

Publication date
2015

Document Version
Final published version

Published in
Journal of Voice

The Relationship Between Acoustic Signal Typing and Perceptual Evaluation of Tracheoesophageal Voice Quality for Sustained Vowels

*,†Renee P. Clapham, †,‡Corina J. van As-Brooks, *,†Rob J. J. H. van Son, *,†Frans J. M. Hilgers, and *,†Michiel W. M. van den Brekel, *†Amsterdam, The Netherlands

Summary: Objectives. To investigate the relationship between acoustic signal typing and perceptual evaluation of sustained vowels produced by tracheoesophageal (TE) speakers and the use of signal typing in the clinical setting.
Methods. Two evaluators independently categorized 1.75-second segments of narrow-band spectrograms according to acoustic signal typing and independently evaluated the recording of the same segments on a visual analog scale according to overall perceptual acoustic voice quality. The relationship between acoustic signal typing and overall voice quality (as a continuous scale and as a four-point ordinal scale) was investigated and the proportion of inter-rater agreement as well as the reliability between the two measures is reported.
Results. The agreement between signal type (I–IV) and ordinal voice quality (four-point scale) was low but significant, and there was a significant linear relationship between the variables. Signal type correctly predicted less than half of the voice quality data. There was a significant main effect of signal type on continuous voice quality scores with significant differences in median quality scores between signal types I–IV, I–III, and I–II.
Conclusions. Signal typing can be used as an adjunct to perceptual and acoustic evaluation of the same stimuli for TE speech as part of a multidimensional evaluation protocol. Signal typing in its current form provides limited predictive information on voice quality, and there is significant overlap between signal types II and III and perceptual categories. Future work should consider whether the current four signal types could be refined.
Key Words: Automatic evaluation–Head and neck cancer–Perceptual evaluation–Acoustic signal typing–Tracheoesophageal speech–Laryngectomy.

INTRODUCTION

Functional voice assessment requires a multidimensional approach to evaluation, and data should allow a clinician to determine whether a voice is classified normal or pathologic, the severity and cause of pathology, and allow tracking changes in voice over time.[1] It is recommended that an evaluation protocol contain perceptual evaluation combined with acoustic, imaging, aerodynamic, and patient self-report measures.[1] A specialized protocol for voice assessment is required within the area of tracheoesophageal (TE) speech because the overall voice quality of substitute voicing should be compared with "near normal laryngeal voicing" rather than normal laryngeal voicing and performing acoustic evaluation can lead to unreliable and inaccurate measurements. This is because standard pitch-detection algorithms in general acoustic software fail when the speech signal has low or no fundamental frequency or high levels of noise.

Titze[2] introduced acoustic signal typing for laryngeal speakers as a decision making tool on whether the researcher/clinician could collect reliable acoustic data. Signal typing involves categorizing recorded speech samples based on visual characteristics observed on narrow-band spectrograms. Van As et al[3] adapted Titze's signal-typing technique for TE voice and identified four signal types based on the spectral characteristics of this speaker group. Although the use of signal typing is recommended as a decision making tool,[2,3] there is a relationship between signal type of sustained vowels and auditory-perceptual judgments of voice quality for running speech[3,4] and as such, signal typing has been proposed as an indicator of the overall perception of voice quality or of functional voice outcome.[3–5] The use of signal typing as part of a multidimensional evaluation of TE voice can be useful as it is estimated that 77% of TE speakers have a measurable fundamental frequency[3] and many acoustic measures will fail this population because of the lack of periodicity in the speech signal.

As noted by Van Gogh et al,[6] there is a subjective component when performing signal typing and reliability and agreement measures warrant reporting just as auditory-perceptual reliability, and agreement measures are generally reported. Many studies investigating signal type for TE speech, however, have used classifications from a single evaluator or do not include procedural information on who performed classifications and do not include reliability information.[3–7] The present study is unique in that we (a) consider the relationship of signal type and perceptual evaluation of the same stimuli and (b) use a scoring procedure that reflects the clinical setting. That is, rather than use mean scores of a large group of raters, we use consensus scores made by two speech pathologists.

This article explores the use of signal typing in its current form for TE voice and the relationship of signal type to

Accepted for publication October 2, 2014.
This research was supported in part by unrestricted research grants from Atos Medical (Horby, Sweden) and the Verwelius Foundation (Naarden, the Netherlands).
From the *Amsterdam Center for Language and Communication, University of Amsterdam (ACLC/UvA), Amsterdam, The Netherlands; †Department of Head and Neck Oncology and Surgery, The Netherlands Cancer Institute-Antoni van Leeuwenhoek, Amsterdam, The Netherlands; and the ‡Department of Marketing and Clinical Affairs, Atos Medical, Hörby, Sweden.
Address correspondence and reprint requests to Renee P. Clapham, Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 210, 1012 VT Amsterdam, The Netherlands. E-mail: r.p.clapham@uva.nl
Journal of Voice, Vol. 29, No. 4, pp. 517.e23-517.e29
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.10.002

perceptual scores of voice quality of the same stimuli. Our principal research line investigates the association between signal type and voice quality for the same stimuli and whether there is a predictive relationship between the two variables. Our secondary research line was to compare the inter-rater agreement and reliability of signal type evaluations with voice quality evaluations. The key variables are consensus acoustic signal type (ordinal data containing four categories) and consensus voice quality scores (continuous data 0–1000). We also use each rater's individual evaluations (ie, preconsensus evaluations) to report inter-rater agreement and reliability.

METHODS

Audio stimuli

Audio recordings were collected at the Netherlands Cancer Institute (Amsterdam, The Netherlands) as part of various research studies between 1996 and 2009. All speakers produced a sustained /a/ as part of the recording procedure. All speakers provided informed consent at the time of data collection and granted use of the recordings for research purposes. As the recording conditions, settings, and equipment varied across the past studies, for the present study, we digitalized analog recordings, and all recordings were converted to 44.1 kHz sampling rate with 16-bit signed integer PCM encoding. No compression had been used on the recordings. Where possible, we used original recordings, but in several cases, only 2-second segments of the vowels were available.

The collection contains recordings from 87 TE speakers. The majority of speakers were male (74 [85%]) and median age at time of laryngectomy was 57 years (range 38–85 years; age at time of laryngectomy was not recorded for one speaker). Age at the time of the recordings could be retraced for 37 of the speakers (43%; median age 66 years, range 46–81 years). As many speakers provided recordings for multiple studies, we selected the stimuli with the earliest recording date. For the recordings used in the present study, 83 speakers (95%) used a Provox1 or Provox2 prosthesis and the remaining 4 speakers (5%) used a Provox Vega prosthesis.

Acoustic signal typing

Procedure. The four signal types are type 1 (stable and harmonic), type II (stable and at least one harmonic), type III (unstable or partly harmonic), and type IV (barely harmonic). During the evaluation of 12 practice items, two speech pathologists (R.P.C. and C.J.V.A.-B.) discussed and adapted scale definitions. The signal typing criteria presented in a study by Van As et al[3] was adjusted to account for the minimum length of the presegmented stimuli and perceived ambiguity in the definition of "stable" (Table 1). For this present study, "stable" was defined as a continuous signal at the fundamental frequency harmonic. Note that the original signal typing criteria of 2 seconds was adjusted to 1.75 seconds as preedited 2-second recordings would have had missing margins in the spectrograms. Note also that the 2-second rule used in Van As was based on the minimum length of the stimuli.

Spectrograms were presented via a custom-made program termed the NKI TE-Voice Analysis tool (TEVA; English, German, and Dutch version available from www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/NKIcorpora/NKI_TEVA/), which runs as a Praat (download from www.praat.org) extension. The entire recording was visualized in a narrow-band spectrogram (window length 0.1 second; time step, 0.001 second; frequency step, 10 Hz; maximum frequency, 2 kHz), and raters were unable to play the sound file. Using the TEVA tool, each rater visually identified the most stable segment of the spectrogram and then classified this segment according to signal type. The raters were blind to speaker gender, speaker age, and prosthesis type. After individual classification, the raters came together and agreed upon the 1.75-second segment to be evaluated and the signal type of this segment.

Rater reliability and agreement. Table 2 lists the inter-rater agreement, and disagreement grouped according to consensus signal type. Raters agreed on signal type categorization in 50 cases (57%; permutation average 29% and standard deviation [SD] 4%) and were in close agreement for the remaining 31 cases (36%; permutation average 38% and SD 5%). The kappa for inter-rater agreement was statistically significant (weighted kappa: $k = 0.55$, $P < 0.001$, weights set at 0, 0.33, 0.66, 1.0). There was a statistically significant correlation between the two rater's evaluations (tau = 0.63, $P < 0.00$), and there was acceptable reliability between the raters (single-measure intraclass correlation coefficient (ICC) [consistency] using a two-way model: ICC = 0.73, 95% confidence interval, 0.62–0.82).

Auditory-perceptual evaluation

Procedure. Three months after performing signal typing evaluation, the same raters completed the auditory-perceptual

TABLE 1.
Criteria for Each of the Four Acoustic Signal Types

Acoustic Signal Type	Criteria
I. Stable and harmonic	• Stable signal for at least 1.75 s, and • Clear harmonics from 0 to 1000 Hz
II. Stable and at least one harmonic	• Stable signal for at least 1.75 s, and • At least one stable harmonic at the fundamental frequency for at least 1.75 s
III. Unstable or partly harmonic	• No stable signal for longer than 1.75 s, or • Harmonics in only part of the sample (for longer than 1 s)
IV. Barely harmonic	• No detectable harmonics or only short-term detectable harmonics for <1 s

TABLE 2.
Inter-Rater Agreement and Disagreement for Acoustic Signal Type (AST) Grouped According to Consensus Signal Type

AST	n	Exact Agreement (%)	Close Agreement (%)	Disagreement (%)
I	14	6 (43)	6 (43)	2 (14)
II	44	25 (57)	18 (41)	1 (2)
III	12	4 (33)	5 (42)	3 (25)
IV	17	15 (88)	2 (12)	0 (0)
Total	87	50 (57)	31 (36)	6 (7)

Notes: Agreement split into exact agreement (same category selected by raters), close agreement (categories differ by one type), and disagreement (categories differ by two types).

TABLE 3.
Inter-Rater Agreement and Disagreement for Voice Quality Scores Grouped According to Consensus Voice Quality Scores (Converted Into Ordinal Categories)

Voice Quality	n	Exact Agreement (%)	Close Agreement (%)	Disagreement (%)
Good	15	9 (60)	4 (27)	2 (13)
Fair	30	13 (43)	8 (27)	9 (30)
Moderate	23	7 (30)	7 (30)	9 (39)
Poor	19	7 (37)	6 (32)	6 (32)
Total	87	36 (41)	25 (29)	26 (30)

Notes: Agreement split into exact agreement (two scores ± 125), close agreement (two scores ± 250), and disagreement (two scores differ by >250).

evaluation task. The perceptual variables were based on scales used for the auditory-perceptual evaluation of running speech[3] and those developed for the INFVo (an abbreviation based on the scale's four parameters: impression, noise, fluency and voicing).[8] The raters discussed and adjusted scale definitions during the evaluation of 12 practice items. Although several additional parameters were included in the data collection, we restrict our analysis to the parameter "overall voice quality."

The raters were blind to all speaker information, including signal-type data. Stimuli were presented in a random order via an online self-paced experiment and raters listened to recordings via a headset. Stimuli were not represented. Raters recorded their evaluations on a computerized visual analog scale built within the TEVA tool. The response scale contained textual anchors at both extremes and did not display tick marks. Raters moved the cursor along the line to the desired location between the two anchors, and the cursor location was then saved as a value between 0 ("least similar to normal") and 1000 ("most similar to normal").

Scores that differed between the raters by more than 125 points were discussed and rescored in the consensus round. When scores were within the range of agreement, the mean score was considered the consensus score and these cases were not discussed. The value ±125 is derived from dividing the scale into four intervals, which corresponds with a four-point ordinal equal appearing interval scale. To aid scoring in the consensus round, major and minor tick marks were placed at every 10% and 5% scale distances, respectively. Numeric anchors were displayed at major tick marks.

Rater reliability and agreement. Table 3 lists the inter-rater agreement and disagreement grouped according to consensus voice quality scores (converted into four ordinal categories). The two rater's scores were in exact agreement (difference $\leq$125 points) in 36 cases (41%) and were in close agreement (difference $\leq$250 points) in the remaining 25 cases (29%). The strength of the correlation between the two raters' individual judgements was statistically significant (tau = 0.43, $P < 0.001$), and the reliability between the raters was acceptable (single-measure ICC [consistency] using a two-way model: ICC = 0.63, 95% confidence interval, 0.49–0.74).

Statistical analysis

All statistical analyses were completed with the statistics program R (available from www.r-project.org), and P value was set to $P < 0.05$ for testing main effects. A Bonferroni correction was applied for *post-hoc* comparisons. Although the evaluation task (ie, voice quality vs signal type) and stimuli (ie, auditory-perceptual vs visual) differed between the two measurements, where possible we used statistical tests for dependent samples as the stimuli were derived from the same recordings and the raters were the same for each task.

Relationship between the two variables. The chi-squared linear-by-linear test of association for ordinal data was used to test the association between consensus signal type categories and consensus voice quality categories. To do this, the visual analog scale was divided into four equal parts, and the consensus scores were coded into one of four ordinal categories: "good" (>750.75), "fair" (>500.5 and $\leq$750.75), "moderate" (>250.25 and $\leq$500.5), and "poor" ($\leq$250.25). To further understand the relationship between the two variables, a nonparametric analysis of variance (Kruskal-Wallis test) with Mann-Whitney test for post-hoc comparisons was used, and we evaluated whether voice quality was a significant predictor of signal type using linear regression.

Comparing proportions of agreement. To compare proportions of inter-rater agreement between the two measures, we used McNemar's nonparametric test for paired samples. That is, we completed two analyses: (1) signal type exact agreement with voice quality exact agreement and (2) signal type agreement (close + exact) with voice quality agreement (close + exact). We used a permutation method (data resampling without replacement, n = 100 000) to calculate the level of chance agreement within the data.

RESULTS

Relationship between signal type and voice quality

Ordinal scores of voice quality. Consensus voice quality scores were converted into a four-point ordinal scale by dividing the visual analog scale into four equal parts and labeled "good," "fair," "moderate," and "poor." The largest category

was for "fair" (30 cases), and the category with the least number of cases was for "good" (15 cases) (Tables 2 and 3). The relationship between consensus signal type and consensus ordinal voice quality scores is presented in Figure 1. Results of the kappa statistic indicate low, but statistically significant agreement between the two measures ($k = 0.22$, $P = 0.004$).

A test of the linear-by-linear association for ordinal variables indicates the association between the two variables was significant ($X^2(1, n = 87) = 29.71$, $P < 0.001$). If we consider signal type (I–IV) as a predictor of voice quality category (good to poor), signal type correctly predicts 38 cases (44%) when the perceptual scale is divided into four equal categories.

Continuous scores of voice quality. To further investigate the relationship between the consensus scores, we performed a nonparametric test of the effect of signal type (ordinal data) on the perceptual scores (continuous data). Kruskal-Wallis test shows that there is a significant main effect of signal type on perceptual scores of voice quality ($X^2 = 31.4$, $P < 0.05$). Mann-Whitney tests (P set to <0.0083 for multiple comparisons) indicate significant differences in median voice quality scores for signal type categories I–IV, I–III, and I–II.

If signal type is considered pseudo-continuous data, a linear regression analysis indicated that voice quality score significantly predicts acoustic signal types ($P < 0.001$) and explains a statistically significant proportion of the variation (multiple $R^2 = 0.37$, $F(1,85) = 49.8$, $P < 0.001$).

Comparing proportions of rater agreement

McNemar's test revealed no statistically significant difference in proportion of exact inter-rater agreement for perceptual voice quality scores (41%) and acoustic signal type (58%) (McNemar's $X^2(1, n = 87) = 3.84$, $P = 0.05002$). The difference in proportion of exact/close inter-rater agreement between voice quality measures (70%) and signal type (94%) was significant (McNemar's $X^2(1, n = 87) = 12.89$, $P < 0.001$).

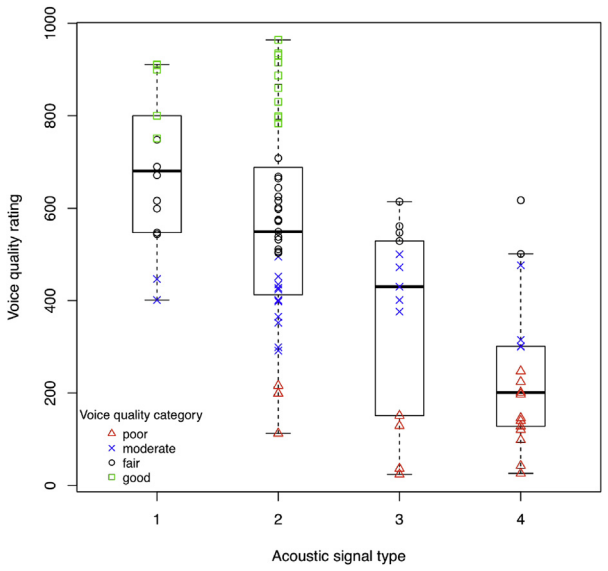

FIGURE 1. Voice quality scores by acoustic signal type. Voice quality data points overlay the boxplot and are coded according to boundaries for converting the continuous scores into categorical data ("good" n = 15 (17%): AST I n = 5, II n = 10; "fair" n = 30 (34%): AST I n = 7, II n = 17, III n = 4, IV n = 2; "moderate" n = 23 (28%): AST I n = 2, II n = 12, III n = 5, IV n = 4; "poor" n = 19 (22%): AST II n = 4, III n = 4, IV n = 11). AST, acoustic signal type.

DISCUSSION

Our primary research line was to investigate the association between consensus judgments of signal type and consensus judgments of voice quality for segments of sustained vowel /a/. In terms of data distribution, over half the stimuli (51%) was classified signal type II and the least frequent classification was for signal type III (14%). This distribution pattern is unlike that reported in Van As[3] (signal type IV was the most frequent and type I was the least frequent) and overlaps somewhat with the distribution pattern reported by D'Alatri et al[4] (signal type I was the most frequent and type III was the least frequently occurring category).

To allow comparison between signal type (ordinal data) and voice quality (continuous data), the visual analog scale was divided into four equal parts. More than 60% of stimuli fell in the central "fair" and "moderate" range with the most frequent category being "fair" (34%) and least frequent category being "good" (17%). Direct comparison of the category frequencies with those in the studies by Van As et al[3] and D'Alatri[4] is not possible as both these studies used a three-point ordinal rating scale. We elected to convert the visual analog scale into four parts as opposed to three in an attempt to maintain the sensitivity of the scale. Converting scores made on a continuous scale into an interval scale is a technique used in researcher (eg, studies by Eadie and Kapsner-Smith,[9] Kreiman and Gerratt,[10] Peterson et al,[11] and Wuyts et al[12]).

Figure 1 displays the relationship between signal type and ordinal voice quality scores. Stimuli with signal types III (unstable/some harmonics) and IV (no harmonics/mostly without harmonics) were never rated as having "good" voice quality. Likewise, stimuli with signal type I (stable with clear harmonics) were never rated as having "poor" voice quality. This pattern is similar to that reported by Van As[3] where the two extreme signal types never cooccurred with the opposite perceptual extreme when a three-point ordinal scale was used. In line with Van As' study comparing signal type of vowels and voice quality of running speech, we also found a statistically significant linear association between signal type and voice quality (ordinal scores) for the same stimuli. The strength of the agreement between the two variables, however, was low. It is clear from Figure 1 that signal type II co-occurs with a broad range of the quality scale (predominately "fair" to "poor"). Excluding stimuli with signal type II from the kappa analysis results in increased agreement (from $k = 0.22$ to 0.31).

Less than half of the ordinal voice quality scores can be correctly predicted by signal type. This highlights that our division of the perceptual scale into four equally spaced intervals may be too simplistic and an alternative division with unequal intervals may more accurately reflect severity (eg, studies by Lopes et al[13] and Yu et al[14]) and increase the strength of the

agreement between the two scales. However, we also completed exploratory analyses of signal type on continuous voice quality data and found a statistically significant main effect. The post-hoc analysis revealed that voice quality median scores differed for three of the signal type comparisons: only quality scores for signal type I could be differentiated from the other signal types.

As far as we are aware, only the study by D'Alatri et al[4] has found significant differences between adjoining signal types and that was only for types III and IV. We hypothesize that the broad definition of signal type II and the "and/or" criteria for signal type III results in high levels of variability in the data. This is in line with the result from our previous study in which signal types II and III were the most difficult to predict using acoustic measures.[15]

Our secondary research line was to compare the proportion of inter-rater agreement and the reliability of signal type evaluations with voice quality evaluations. Our primary argument for using the proportion of exact and proportion of close agreement as indices of agreement is that these measures can be applied to both continuous and ordinal data and allow us to directly compare proportions between the two scales. The drawback of this measurement method is that it does not take chance agreement between the two raters into consideration. We decided against converting individual voice quality scores into ordinal scores as this would not account for the situation where scores differ by a few points but fall either side of a cut-off point.

Before discussing the comparison results, we discuss the inter-rater agreement data for first signal type then voice quality. For signal type, the inter-rater disagreements were between signal types I–III (n = 4) and II–IV (n = 2). In no case was the disagreement larger than two categories (only possible for signal types I or IV). In no cases did the two raters disagree on signal type IV stimuli (Table 2). In terms of patterns of agreement, agreement was largest for signal types II and IV. This is most likely a reflection of the number of signal type II stimuli and that signal type IV is an easily identified category. However, because of the procedure used for data collection, we are unable to state whether the disagreement occurred because of differences in categorization (ie, identification of signal type) or because of differences in segment selection. In a future experiment, these two aspects might be separated by asking the raters to agree which 1.75 segment should be evaluated for signal type and only then do the individual ratings of signal type.

For voice quality, 61 (70%) of the rating pairs were in exact or close agreement. Scores in the center of the scale had higher counts of disagreement than scores at the extremes of the scale (Table 3). The strength of the association between the two raters' evaluations was statistically significant. Although the inter-rater agreement results indicated statistically significant levels of agreement and that the evaluations were made above chance level agreement, the results highlight that similar to perceptual evaluation, signal typing remains a subjective task and hence why consensus evaluations should be used in the clinical and research setting (for all subjective tasks) where possible.

Concerning differences in the proportion of inter-rater agreement between the two measures, although the proportions of agreement were higher for judgements of signal type than voice quality (exact agreement: 58% and 41% and close agreement: 94% and 70%, respectively), this difference was statistically significant for measures of close agreement. For measures of exact agreement, the results were just beyond the set level of statistical significance. That the proportions of agreement are larger for signal typing data is not an unexpected result; the signal typing task requires each rater to select one of four described categories (ie, 25% agreement due to chance), whereas in the voice quality task, the scale does not force the rater to select a category and textual anchors are only provided at the scale extremes. Although the proportion of close agreement on signal type was significantly higher than for voice quality, because of differences in the scales, the distances are not equal between the two variables and as such are difficult to compare directly. That is, close signal type agreement means that the scores differ by a maximum of ½ the "scale," whereas close voice quality agreement means that the scores differ by a maximum of ¼ of the scale.

Regarding the inter-rater reliability for the two measures, the reliability for both variables was significant but stronger correlations were found for signal type measures than voice quality measures. This is not surprising as the signal type variable requires the rater to make a forced choice from four options with each option having some criteria, whereas the voice quality scale is on a visual analog scale without textual anchors over the continuum of the scale. Compared with other studies of perceptual voice quality using ordinal scoring systems, the signal type results are similar to the average correlation value reported in a study by Shrivastav et al[16] for evaluations of breathiness on a five-point scale (average Tau, 0.64) and are lower than the coefficient reported in a study by Karnell et al[17] for the Grade scale on a four-point scale (Spearman = 0.85).

The ICC values for the two variables are stronger for signal type variable than the voice quality variable (0.73 and 0.63, respectively). Compared with other studies, the reliability results are lower than that reported in a study by Nemr et al[18] for a three-point scale to evaluate Grade from the GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) scale for healthy control and speakers with dysphonia (ICC = 0.88) and that reported in a study by Zraick et al[19] for the Grade part of the GRBAS (ICC = 0.66). Although agreement and reliability data are low, the procedure used to collect data (ie, consensus scores) is a technique that can be used in the clinical situation.

The results suggest that signal typing in its current form can be used as part of a multidimensional assessment of voice quality predominately as a way to categorize voice quality and serve as a decision making tool on acoustic analysis. We anticipate that future work will consider updating the signal type definitions by including signal subtypes for types II and III (eg, differentiation between types that contain continuous flat harmonics and types that contain continuous fluctuating harmonics). Part of this difficulty may be due to the variability in TE speech, that is, type III instability can be caused by

hypertonicity or hypotonicity, which both sound very different to a listener. However, in terms of signal typing serving as a basis for further acoustic analysis, type III would indicate that there is not stable fundamental frequency, and this should be considered when acoustic analyses are carried out.

This present study is part of our efforts to automate subjective evaluation of speech and voice quality so they can complement a clinician's evaluation. To this end, we have already begun work on automating signal type based on acoustic information[15]. We envisage that signal typing could be a useful component in the multidimensional evaluation of voice quality, and when paired with automatic acoustic data, predicted perceptual scores, and observed perceptual scores, a clinician can have a "voicegram" of the speaker that can be printed and kept in a patient's file for comparison with other patients and assessment of treatment results. We are currently developing a function within the TEVA application to produce a "voicegram" of a speaker, which contains several automated acoustic measures and can display the predicted acoustic signal type (Figure 2 for a concept voicegram).

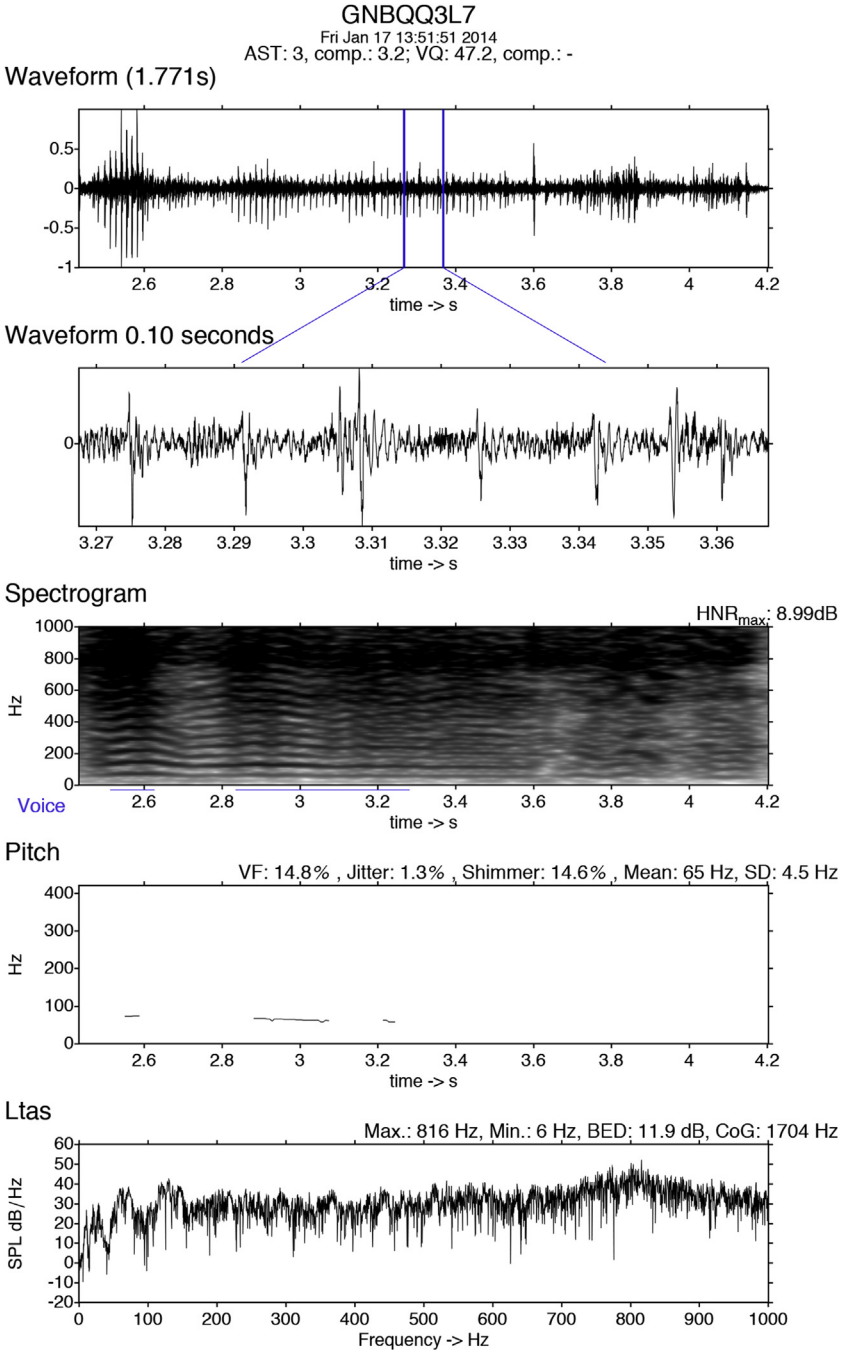

FIGURE 2. Concept version of the voicegram. The print displays (from top down) speaker code, date of print, observed signal type and voice quality score, computed signal type and voice quality score, waveform (box 1) and central 10 ms from waveform (box 2), spectrogram of predetermined segment used for signal typing and perceptual evaluation (box 3), pitch contour (box 4), and long-term average spectrum (Ltas) (box 5).

CONCLUSIONS

The results support the use of signal typing as part of a multidimensional evaluation of functional voice assessment. There is a statistically significant relationship between the two measures but signal typing in its current form provides limited predictive information on voice quality. The two extreme signal type categories are clear but there is a large overlap between signal types II and III and perceptual categories. However, signal typing can serve as a basis for determining further acoustic analysis (eg, type III would indicate that there is no stable fundamental frequency and fundamental frequency-based acoustic measures should be avoided. Our results have confirmed that while signal typing is a useful approach to evaluating voice quality, the definitions of the four existing signal types and inclusion of subtypes warrants further investigation.

Acknowledgments

The authors wish to thank all the researchers from the Netherlands Cancer Institute-Antoni van Leeuwenhoek (NKI-AVL) who collected the speech recordings.

REFERENCES

1. Dejonckere PH. Assessment of voice and respiratory function. In: Remacle M, Eckel HE, eds. *Surgery of Larynx and Trachea*. Berlin, Germany: Springer; 2010.
2. Titze I. Workshop on acoustic voice analysis: summary statement. *Natl Cent Voice Speech*. 1995;1–36.
3. Van As-Brooks CJ, Koopmans-van Beinum FJ, Pols LCW, Hilgers FJM. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *J Voice*. 2006;20:355–368.
4. D'Alatri L, Bussu F, Scarano E, Paludetti G, Raffaella Marchese M. Objective and subjective assessment of tracheoesophageal prosthesis voice outcomes. *J Voice*. 2012;26:607–613.
5. Sprecher A, Olszewski A, Jian JJ, Zhang Y. Updating signal typing in voices: addition of type 4 signals. *J Acoust Soc Am*. 2010;127:3710–3716.
6. Van Gogh CDL, Festen JM, Verdonck-de Leeuw IM, Parker AJ, Traissac L, Cheesman AD, Mahieu HF. Acoustical analysis of tracheoesophageal voice. *Speech Commun*. 2005;47:160–168.
7. Lawson G, Jamart J, Remacle M. Improving the functional outcome of Tucker's reconstructive laryngectomy. *Head Neck*. 2001;23:871–878.
8. Moerman M, Martens J-P, Crevier-Buchman L, et al. The INFVo perceptual rating scale for substitution voicing: development and reliability. *Eur Arch Otorhinolaryngol*. 2006;263:435–439.
9. Eadie TL, Kapsner-Smith M. The effect of listener experience and anchors on judgments of dysphonia. *J Speech Lang Hear Res*. 2011;54:430–447.
10. Kreiman J, Gerratt BR. Validity of rating scale measures of voice quality. *J Acoust Soc Am*. 1998;104:1598–1608.
11. Peterson EA, Roy N, Awan SN, Merril RM, Banks R, Tanner K. Toward validation of the cepstral spectral index of dysphonia (CSID) as an objective treatment outcomes measure. *J Voice*. 2013;27:401–410.
12. Wuyts FL, De Bodt MS, Van De Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice*. 1999;13:508–517.
13. Lopes LW, Barbosa Lima IL, Almeida LNA, Cavalcante DP, Figueirêdo de Almeida AA. Severity of voice disorders in children: correlations between perceptual and acoustic data. *J Voice*. 2012;26:819.e7–819.e12.
14. Yu P, Revis J, Wuyts FL, Zanaret M. Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatr Logop*. 2002;54:271–281.
15. Clapham R, Van As-Brooks C, Van den Brekel M, Hilgers F, Van Son R. Automatic tracheoesophageal voice typing using acoustic parameters. In Proceedings of the 14th Annual Conference of the International Speech Communication Association -Interspeech: 2013 Augustus 25-29; Lyon, France. ISCA; 2013: 2162–2166.
16. Shrivastav R, Sapienza C, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res*. 2005;48:323–335.
17. Karnell MP, Melton SD, Childes M, Todd JMC, Dailey SA, Hoffman HT. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *J Voice*. 2007;21:576–590.
18. Nemr K, Simões-Zenari M, Cordeiro GF, Tsuji D, Ogawa AI, Ubrig MT, Menezes MH. GRBAS and Cape-V scales: high reliability and consensus when applied at different times. *J Voice*. 2012;26:812.e17–812.e22.
19. Zraick RI, Kempster GB, Connor NP, Klaben BK, Bursac S, Thrush CR, Glaze LE. Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V). *Am J Speech Lang Pathol*. 2011;20:14–22.