# 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<br>• Clear harmonics from 0 to 1000 Hz |
| II. Stable and at least one harmonic | • Stable signal for at least 1.75 s, and<br>• 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<br>• 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 |