Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech

van As, C.J.; van Beinum, F.J.; Pols, L.C.W.; Hilgers, F.J.M.

Published in:
Journal of Voice

DOI:
10.1016/j.jvoice.2005.04.008

Acoustic Signal Typing for Evaluation of Voice Quality in Tracheoesophageal Speech

*†Corina J. van As-Brooks, †Florien J. Koopmans-van Beinum, †Louis C.W. Pols, and *†‡Frans J.M. Hilgers

Amsterdam, The Netherlands

Summary: Because of the aperiodicity of many tracheoesophageal voices, acoustic analysis of the tracheoesophageal voice is less straightforward than that of the normal voice. This study presents the development and testing of an acoustic signal typing system based on visual inspection of a narrow-band spectrogram that can be used by researchers for classification of voice quality in tracheoesophageal speech. In addition to this classification system, a selection of acoustic measures [*median fundamental frequency, standard deviation of fundamental frequency, jitter, percentage of voiced (%Voiced), harmonics-to-noise ratio (HNR), glottal-to-noise excitation (GNE) ratio,* and *band energy difference (BED)*] was computed to provide more insight into the acoustic components of tracheoesophageal voice quality. For clinical relevance, relationships between the acoustic signal types and an overall judgment of the voice were investigated as well. Results showed that the four acoustic signal types form a good basis for performing more acoustic analyses and give a good impression of the overall quality of the voice.

Key Words: Acoustic analysis—Laryngectomy—Tracheoesophageal speech—Voice prosthesis.

INTRODUCTION

Voice quality is a perceptual phenomenon, and consequently, perceptual evaluations are considered the ''gold standard'' of voice quality evaluation. Disadvantages of perceptual evaluations are that listeners differ in their opinion about voice quality and that it is time consuming to acquire these judgments, because many raters are needed to obtain sufficient inter- and intrarater reliability.[1]

In clinical practice, perceptual evaluations play a prominent role in therapy evaluation purposes. Acoustic analyses are usually not routinely performed for clinical purposes. Acoustic measures do not show a one-to-one relationship with perceptual evaluation

Accepted for publication April 8, 2005.

Supported in part by a grant from the Amsterdam Center for Language and Communication, University of Amsterdam, the Netherlands; and from the Department of Speech and Language of the Radboud University Nijmegen, the Netherlands. The Maurits and Anna de Kock Foundation provided financial support for the equipment needed for the speech recordings, acoustic analyses, and the listening experiment.

From the *Department of Otolaryngology-Head & Neck Surgery, Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands, the †Institute of Phonetic Sciences, Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, The Netherlands, and ‡Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands.

Address correspondence and reprint requests to Frans J. M. Hilgers, Department of Otolaryngology-Head & Neck Surgery, The Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, Plesmanlaan 121, 1066CX Amsterdam, The Netherlands. E-mail: f.hilgers@nki.nl

Journal of Voice, Vol. 20, No. 3, pp. 355–368
0892-1997/$32.00

doi:10.1016/j.jvoice.2005.04.008

and therefore cannot serve as a substitute but merely as an adjunct to it. They can provide more insight in the acoustic characteristics of a voice. Several studies have described acoustic analyses of tracheoesophageal voice quality and have concluded that tracheoesophageal voices differ considerably from normal voices with respect to the acoustic measures.[2–5]

Acoustic measures computed digitally will always produce the same result for the same input, and when these measures are obtained correctly, they can form a valuable objective adjunct to perceptual evaluations in clinical practice. Unfortunately, acoustic analysis of the tracheoesophageal voice is less straightforward than that of the normal voice. It is because in many voices, aperiodicity is evidenced, and in some voices, the fundamental frequency is extremely low. Also, the acoustic measures must be perceptually relevant, which yet has to be shown.

In an earlier study of tracheoesophageal voice quality, moderate-to-strong correlations were found between the perceptual evaluations of a sustained /a/ and the acoustic measures, which are calculated on the same sustained /a/ with the *Multi Dimensional Voice Program* (*MDVP*) (Kay Elemetrics, Lincoln Park, NJ).[5] This study also showed that with *MDVP*, 30% of the voice samples could not be analyzed at all, or only very short parts were analyzable. Visual inspection of these voice samples showed that the patients had very low-pitched voices (and therefore fell outside the fixed pitch analysis range of *MDVP*) or had very aperiodic voices. Tracheoesophageal voices thus can be aperiodic to such an extent or can have such an extremely low pitch that the pitch detection algorithm fails, or even that there is no fundamental frequency present at all. It implies that acoustic measures based on pitch detection algorithms only provide reliable results for the tracheoesophageal voices with more regular periodicity. Narrow-band spectrograms to determine the overall acoustic character of the tracheoesophageal voice to be analyzed provide a good impression of the harmonic characteristics of the voice and consequently the ability to perform reliable periodicity-based acoustic measures.

Although for normal laryngeal voices, acoustic signal types (based on narrow-band spectrograms) are advised to be used by researchers as a visual information tool and as a decision-making tool for further acoustic analyses,[6,7] such a system has not been developed and presented for tracheoesophageal speech.

The aim of this study is to develop an acoustic signal typing system that is perceptually relevant, that can evaluate the entire range of tracheoesophageal voices, and that can serve as an underlying basis for further acoustic analyses. Acoustic measures are selected that enable calculations for the entire range of tracheoesophageal voice qualities, and subsequently, relationships between the acoustic signal types and the acoustic measures will be investigated to gain insight into the specific acoustic characteristics of the signal types. Furthermore, relationships between an overall perceptual judgment of voice quality and the acoustic signal types will be investigated. We will use these relationships to gain insight into the perceptual relevance of the acoustic signal types and to investigate whether they might form a valuable adjunct to perceptual evaluation in everyday clinical practice.

PATIENTS AND METHODS

Patients

Speech recordings were made of a total of 40 laryngectomized patients with tracheoesophageal speech by means of an indwelling voice prosthesis.[8] One speaker refused to produce a sustained /a/, which left 39 laryngectomized persons for the analyses. Twenty-nine of them were men, and 10 were women. Patient ages ranged from 47 to 82 years, with a mean of 66 years. Postoperative follow-up ranged from 1 to 18 years, with a mean of 6 years. More information about the speakers participating in this study is summarized in Table 1.

Speech material, recording, and processing

The speech material for the acoustic analyses consisted of three sustained vowels /a/ at a comfortable pitch and loudness level and a standard read-aloud text. The speech recordings were made in a quiet, sound-treated room. For the recordings, we used a DAT-recorder (Sony TCD-8; Sony Corporation, Tokyo, Japan), together with the hardware and software of the Computerized Speech Lab, Model 4300B (Kay Elemetrics, Lincoln Park, NJ). Via the external module of the Computerized Speech Lab, the speech data were digitally recorded on DAT

TABLE 1. *Patient Characteristics (n = 39)*

Parameter	Characteristic	Number
Sex	Male	29
	Female	10
Age (yrs)	Range	47–82
	Mean	67
	≤70 yrs	24
	>70 yrs	15
Postoperative (yrs)	Range	1–8
	Mean	6
	≤6 yrs	23
	>6 yrs	16
Extent of surgery	Standard	30
	Reconstruction	9
Radical neck dissection	No	20
	Uni/bilateral	19
Radiotherapy	Primary	17
	Postoperative	20
	None*	2
Myotomy	Yes	6
	No	23
Neurectomy	Yes	17
	No	12

*Subgroup too small and thus not included in statistical analyses.

tape. We used a headset microphone (AKG-c410, Kay Elemetrics) with the microphone located laterally at the corner of the mouth; the mouth-to-microphone distance was 2.5 cm. At the start of the speech recording, the recording level was adjusted for each speaker individually and then fixed to optimize the signal-to-noise ratio.

For analysis, the three sustained /a/'s were stored on a PC hard disk with a sampling frequency of 44100 Hz with a SoundBlaster card and the software program *Praat* (Version 3.8.68, website: http://www.praat.org, by P. Boersma and D. Weenink).[9] *Praat* is available at the above mentioned website.

The perceptual evaluations of the read-aloud text were carried out by four trained speech-language pathologists and are described in detail by van As et al.[1]

Acoustic signal typing

Similar to the acoustic signal typing systems of Yanagihara[6] and Titze[7] for normal voices, in this study, an acoustic signal typing system was developed for tracheoesophageal voices. This acoustic signal typing was chosen to achieve a visual impression of the acoustic content of the voice samples. An advantage of narrow-band spectrograms is that they can be obtained reliably for each speech sample and that they can indicate whether further acoustic analyses are appropriate. For instance, when the narrow-band spectrogram shows that the voice sample does not contain any harmonics, acoustic analysis of fundamental frequency should be omitted. Dividing those spectrograms into four signal types introduces some subjectivity, because researchers may differ in their opinion of which category to assign to the spectrogram. Thus, in this study, we used specific criteria to determine the signal type.

For signal typing, first, a narrow-band spectrogram (100-ms window) was made of each of the three sustained /a/'s of each speaker. Subsequently, the vowel with the best harmonic structure (most stable, least noise, most harmonics) was selected and a sample of the best 2 seconds of that vowel was stored for the signal typing. The 2-second time frame was chosen because a 2-second stable sample was available for all patients. Although the patients were asked to produce a stable sustained /a/ for 5 seconds, not all of them could do so.

Visual inspection of the narrow-band spectrograms of the sustained /a/'s of the tracheoesophageal speakers showed that one of the more obvious differences among the speakers lies in the ''harmonic strength'' of the signal. In some speakers, harmonics up to 1000 Hz are observed, whereas in other speakers, only one or two harmonics or even no harmonics at all are observed. Thus, the amount of spectral noise differs among patients. Also, the stability (both of pitch and signal amplitude) and continuation of the harmonic bands throughout the voiced signals are important characteristics that differed among patients from highly fluctuating to very stable, to absent or only partially present. Based on the visual spectrographic characteristics of the tracheoesophageal voice signals, four different types were defined. We used the following criteria for the typing of the tracheoesophageal voices:

Type I. Stable and harmonic (Figure 1)

—Stable signal for a full 2 seconds
—Clear harmonics up to at least 1000 Hz

Type II. Stable and at least one harmonic (Figure 2)

—Stable signal for a full 2 seconds
—At least one stable harmonic at the fundamental frequency for a full 2 seconds

Type III. Unstable or partly harmonic (Figure 3)

—Unstable signal with harmonics throughout full 2 seconds
—Absence of harmonics for less than 1 second

Type IV. Barely harmonic (Figure 4)

—Complete absence of harmonics
—Partial absence of harmonics for more than 1 second

The spectrogram with the clearest harmonic structure was selected out of the three available samples, and the speakers were divided into four subgroups based on the visual appearance of the acoustic signal in the narrow-band spectrogram. It was done as a consensus judgment of two authors.

Acoustic analysis

For acoustic analysis, seven voice-quality measures were chosen, all of which will be described in more detail. The measures chosen were meant to reflect the pitch (fundamental frequency) and quality (periodicity, harmonicity) of the voice. Five of these measures are regularly used by researchers for acoustic analysis of voice quality: fundamental frequency, standard deviation of fundamental frequency, jitter, HNR, and %Voiced speech. All of these measures are based on pitch extraction, and consequently, the first three can only be calculated for those voices with a clear fundamental frequency. The same yields true for other perturbation measures such as shimmer. In this study, only jitter was measured, because shimmer measures were at that time not available in the *Praat* software, and in an earlier study, the different perturbation measures were found to be highly correlated to one another.[5] Ideally, reliable acoustic measures should be available for the entire range of voice qualities. Although HNR and %Voiced are the results of pitch extraction as well, results can be obtained for all voice samples (for instance, a completely aperiodic sample is 0% voiced and has a low HNR), and they are, therefore, interesting measures. The other two measures (BED and GNE ratio) are frequency-independent measures. The BED has been described by Dejonckere and Lebacq[10] and has been used in laryngectomized patients by Debruyne et al.[2] The GNE ratio has been described by Michaelis et al.[11] The GNE ratio has not yet been described for tracheoesophageal voice quality. The advantage of these two measures is that they, unlike fundamental frequency-based measures, can be calculated for the entire range of tracheoesophageal voice qualities. They have, however, not yet been related to perceptual evaluations of tracheoesophageal voice, and thus their perceptual relevance needs to be studied.

For the acoustic analysis, 2 seconds of the most stable part of the selected vowel as observed in the narrow-band spectrogram, were analyzed with the software program *Praat*. The seven acoustic measures that were calculated are as follows:

Median fundamental frequency [F0-median (Hertz)]
The fundamental frequency as a function of time is measured with cross-correlation. We used default *Praat* settings, except for the pitch extraction range, which was set from 40 Hz to 250 Hz instead of 75 Hz to 600 Hz, and the voicing threshold, which was set to 0.40 instead of 0.45. In *Praat*, the length of the analysis window is based on the value of the lowest frequency of the pitch extraction range. To avoid pitch extraction errors, this frequency was increased for higher pitched voices (based on the harmonics in the narrow-band spectrogram). The slight decrease of the voicing threshold enables a ''voiced'' decision in a larger part of the voice sample; an even larger decrease of this voicing threshold seemed to introduce pitch extraction errors as confirmed by visual inspection. The median fundamental frequency is determined over all voiced segments of the 2-second interval.

Standard deviation of the fundamental frequency [F0-SD (Hertz)]
This measure is also derived from the calculations of fundamental frequency. It reflects the changes in fundamental frequency found in the 2-second voice sample of sustained /a/. It is determined over all voiced segments of the 2-second interval.

Jitter (%)
The percentage of jitter is calculated from the results of the pitch extraction. First, a so-called

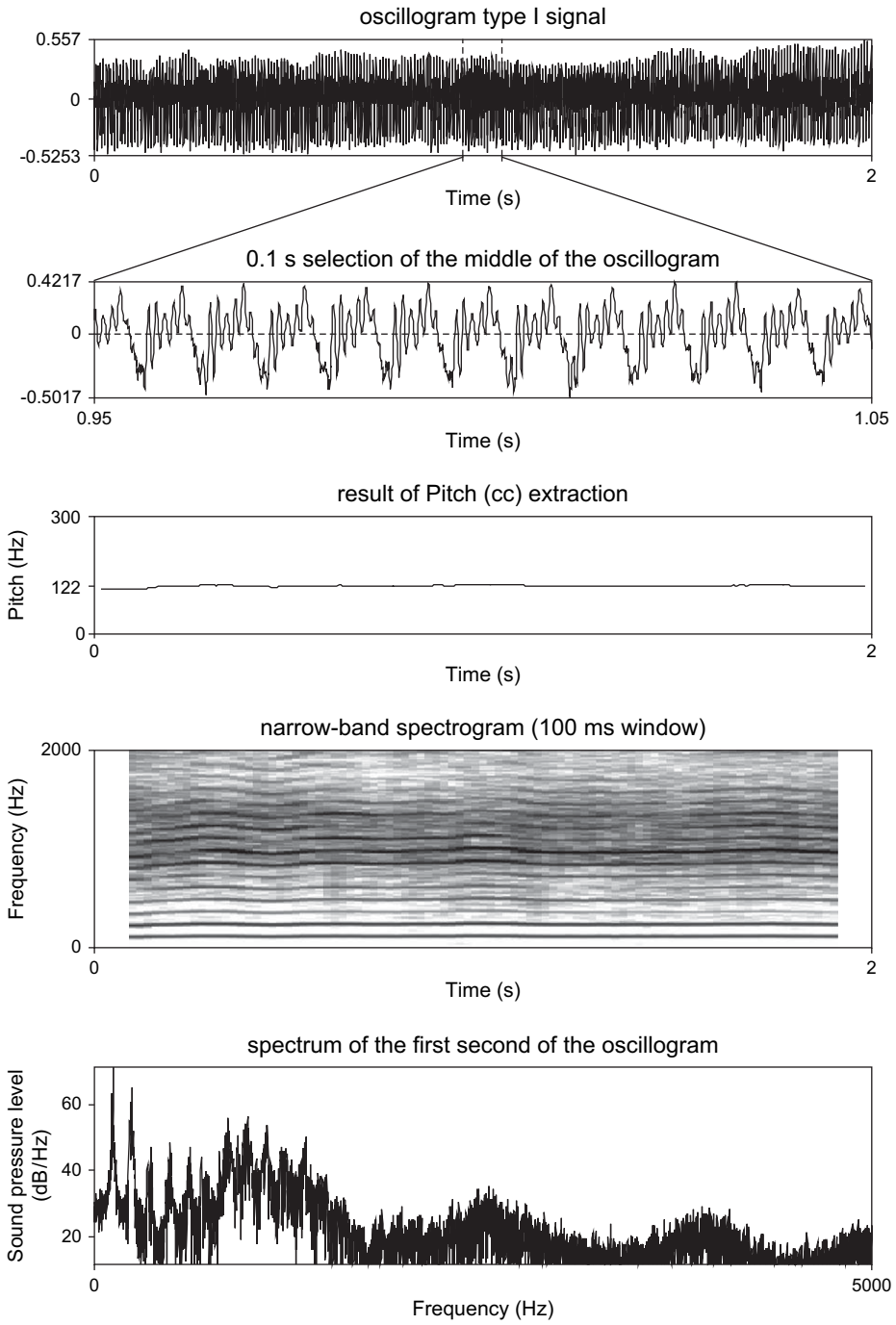

FIGURE 1. Example of a voice sample classified as type I. The oscillogram shows a stable signal, with stable loudness. The 100-ms selection of the oscillogram shows a clearly periodic pattern. The pitch contour shows a stable fundamental frequency (mean 122 Hz). In the narrow-band spectrogram (100-ms window), clear harmonics are observed up to 1500 Hz and for parts of the voice sample even up to 2000 Hz. The long-term average spectrum over 1 s also shows a clear harmonic structure in the lower frequencies and noise in the higher frequency region.

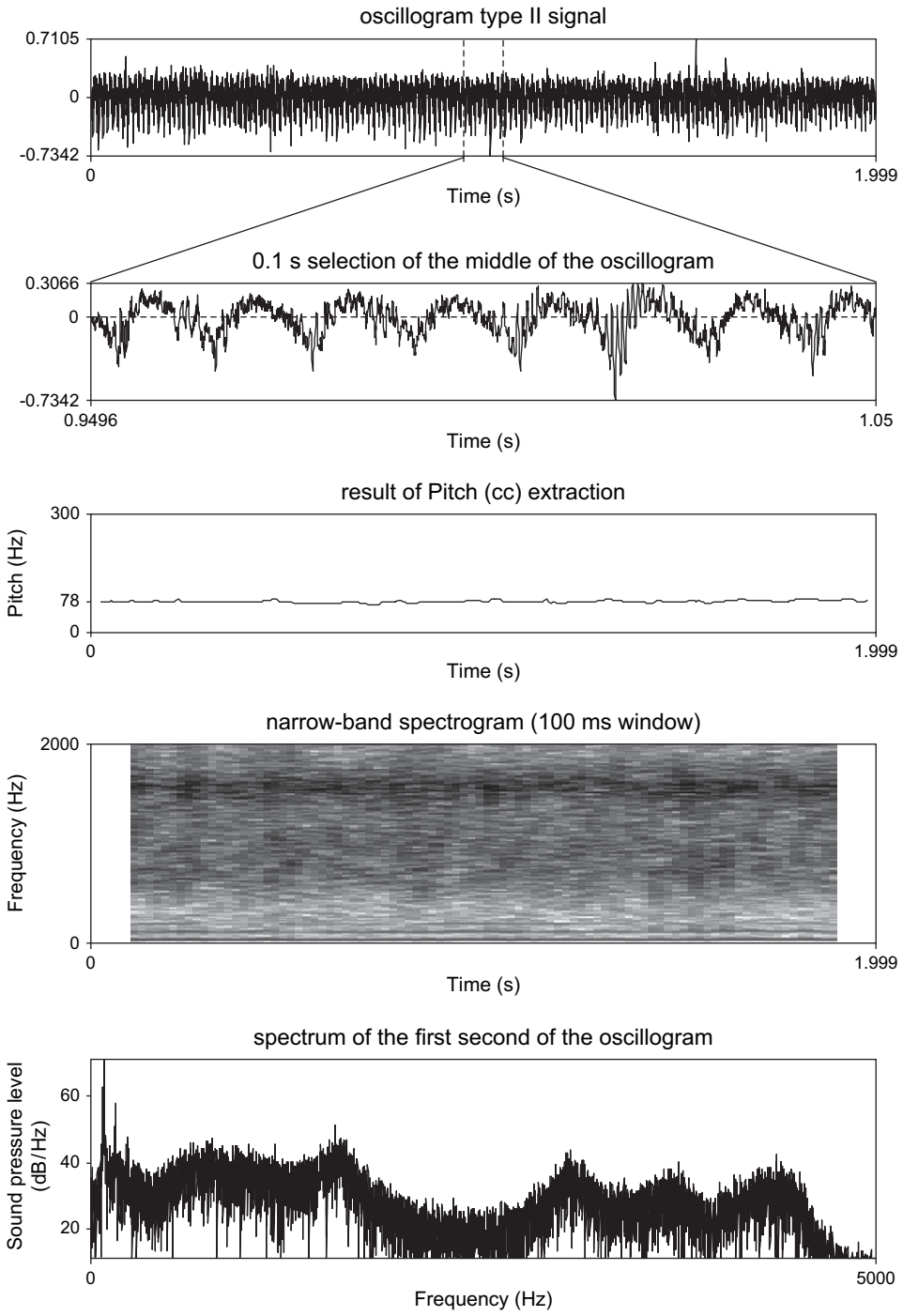

FIGURE 2. Example of a voice sample classified as type II. The oscillogram shows a stable signal, with stable loudness. The 100-ms selection of the oscillogram shows a periodic pattern, with noise. The pitch contour shows a stable fundamental frequency (mean 78 Hz). In the narrow-band spectrogram (100-ms window), the first harmonic is clearly visible and the second and third harmonic are visible in small parts of the spectrogram. In the long-term average spectrum over 1 s, also only three harmonics can be observed and the high-frequency noise is of a higher level than in the type I signal in Figure 1.

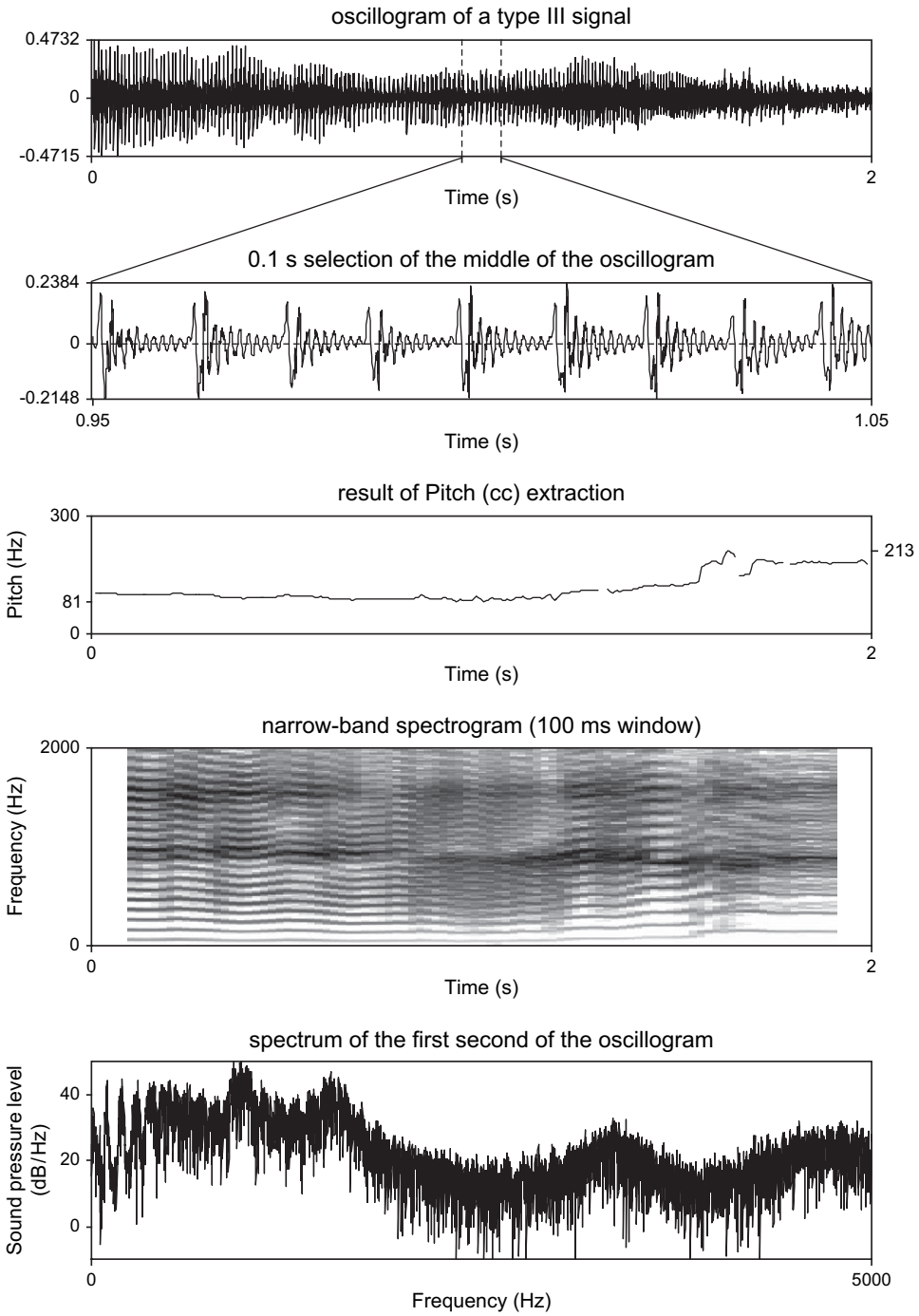

FIGURE 3. Example of a voice sample classified as type III. The oscillogram shows a signal of unstable loudness. The 100-ms selection of the oscillogram shows a clearly periodic structure. The pitch contour shows that the patient cannot produce a sustained /a/ at a stable pitch. The narrow-band spectrogram (100-ms window) shows that there are clear harmonics up to 2000 Hz, but that the voice signal is very unstable. In the long-term average spectrum over 1 s, only four harmonics are observed; because of the instability, the noise in the spectrum is high.

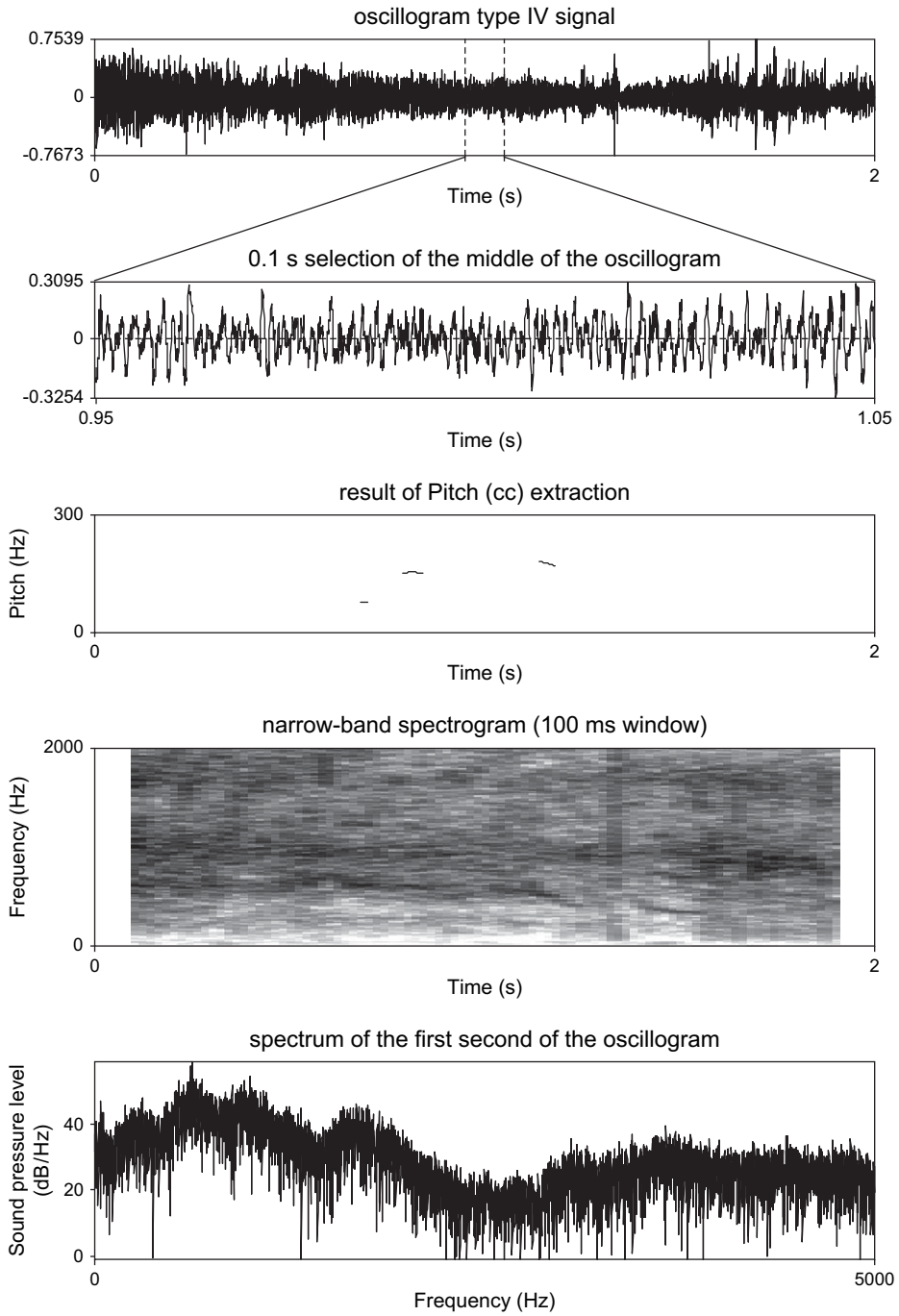

FIGURE 4. Example of a voice sample classified as type IV. The oscillogram shows a highly unstable signal. In the 100-ms selection of the oscillogram, no periodicity can be detected at all. It is reflected in the pitch contour, narrow-band spectrogram (100-ms window), and long-term average spectrum over 1 s in which no harmonicity is observed at all.

point process is created from the results of the pitch extraction. Then, the pulses extracted from the point process calculate the subsequent intervals. Jitter (%) is calculated with the following formula, in which T_i is the i-th interval and N is the number of intervals:

$$Jitter\ (\%) = \sum_{i=2}^{N-1} 2T_i - T_{i-1} - T_{i+1} \Big/ \sum_{i=2}^{N-1} T_i$$

The shortest possible interval to be considered was 0.1 ms. The longest possible interval considered was related to the lowest fundamental frequency found for that particular voice sample (for example, when the lowest F0 found was 50 Hz, this period was set to 20 ms). It is determined over all voiced segments of the 2-second interval.

%Voiced

This percentage is calculated on the relative number of unvoiced analysis windows found for the calculation of the fundamental frequency. In those voice samples for which no fundamental frequency was found at all, this percentage was thus zero.

HNR (decibels)

The HNR is calculated with cross-correlation. We used default *Praat* settings, except for minimum frequency for pitch extraction, which was set to 40 Hz, and for the silence threshold, which was set to 0. With the silence threshold set to 0, the HNR calculated is based on the entire voice sample and not only on the parts that can be considered voiced.

GNE ratio

This measure is calculated on a stable part of 0.25 s that was selected by hand from the 2-second voice sample. This measure indicates to what extent the voice excitation is caused by a pulse train or noise.[11] The minimum frequency was 500 Hz, the maximum frequency was 4500 Hz, the frequency band was 1000 Hz, and the frequency step was 80 Hz.

BED (decibels)

For calculation of this measure, the 0.25-s sample we used for calculation of the GNE ratio was sampled down to 10 kHz, after which a long-term average spectrum was made. In this spectrum, the difference in decibels between the mean spectral intensity in the band between 0 and 500 Hz and the mean spectral intensity in the band between 4000 and 5000 Hz is calculated. This measure can be considered an estimation of the relative amount of high-frequency noise in the spectrum of the voice.[2,10]

Perceptual evaluations

Four trained speech language pathologists (SLPs) judged voice quality in a read-aloud text and gave an overall judgment of voice quality as ''good,'' ''reasonable,'' or ''poor.'' This overall judgment was performed as part of a more detailed perceptual judgment of voice quality in tracheoesophageal speech, which is described by van As et al.[1] In summary, a ''good'' voice was defined as ''most similar to normal voice,'' a ''poor'' voice was defined as ''least similar to normal voice,'' and ''reasonable'' was defined for the group in between both extremes. There was a good correlation between the judgments of the four trained SLPs, and there was no voice sample judged as good by one rater and poor by any of the other three raters. A voice was considered ''good'' or ''poor'' when at least two raters gave that particular judgment. Consequently, the voices collecting three ''reasonable'' qualifications formed the ''reasonable'' group. There seemed to be 13 ''good,'' 14 ''reasonable,'' and 12 ''poor'' voices.

Statistical analysis

Means and standard deviations of the acoustic measures were calculated for the entire patient group. Before further statistical analyses were performed, the acoustic measures concerning fundamental frequency (*median fundamental frequency* and *standard deviation of fundamental frequency*) were logarithmically transformed for statistical reasons. The values in text and tables were transformed back to Hertz for clarity.

Analyses of variance (ANOVA) followed by *post hoc* Tukey tests with Bonferroni correction were performed to investigate the relationships between the acoustic signal types and the acoustic measures. When assumptions of normality could not be met (according to Q-Q plots), we used the nonparametric Kruskal–Wallis test followed by the nonparametric Mann–Whitney test, which was the case in *%Voiced* and *jitter.*

The relationship between the acoustic signal types and the overall perceptual judgment of tracheoesophageal voice quality by the trained raters (good-reasonable-poor) was investigated with a chi-squared test for linear-by-linear association.

Relationships between the eight clinical factors as specified in Table 1 (pharyngeal reconstruction, age, sex, postoperative follow-up, myotomy of the cricopharyngeal muscles, neurectomy of the pharyngeal plexus, primary or postoperative radiotherapy, and radical neck dissection) and the acoustic signal types were investigated with chi-squared tests.

Relationships between clinical factors and the acoustic measures obtained with *Praat* were investigated with *t* tests for two independent samples.

RESULTS

Acoustic signal typing

On the basis of the criteria proposed for the acoustic signal typing, four subgroups were formed. There seemed to be seven patients with a type I, 13 with a type II, 11 with a type III, and eight with a type IV signal.

Acoustic measures

Acoustical analysis based on pitch extraction on the complete signal or parts of it was possible in 30 of the 39 (77%) voice samples. Nine voice samples were considered almost completely unvoiced, and visual inspection of these nine voice samples indeed showed no clear periodicity. In Table 2, the results for the acoustic measures calculated with *Praat* are given.

In Table 2, it can be observed that for all acoustic measures, the range is wide and the standard deviation is high. The *median fundamental frequency* varies from 46 Hz to 229 Hz.

Acoustic signal typing versus acoustic measures

With regard to the acoustic measures, *%Voiced*, *GNE ratio*, *HNR*, and *BED* relationships could be calculated for all speakers. The acoustic measures *median fundamental frequency*, *standard deviation of fundamental frequency*, and *jitter* could be calculated for all type I signals, 12 of the 13 type II signals, 10 of the 11 type III signals, and only 1 of the 8 type IV signals. The *%Voiced* of this one type IV voice sample was 18%, which indicates that only a short part of this voice sample contained periodicity. The type IV signals were therefore left out of the analysis for these acoustic measures: The group size was too small and the fact that these measures generally cannot be calculated for this acoustic signal type confirms the suitability of signal types as a basis for further acoustic analyses.

TABLE 2. *Range Mean, and Standard Deviation of the Various Acoustic Measures*

Acoustic Parameter	N	Range	Mean	Standard Deviation
F0-median (Hz)	30	46 to 229	103	43
F0-SD (Hz)	30	0.09 to 35.8	6.31	7.17
Jitter (%)	30	0.57 to 27.48	6.78	6.29
% Voiced (%)	39	0 to 100	66	40
HNR (dB)	39	−2.0 to 15.3	4.30	4.45
GNE	39	0.57 to 0.98	0.78	0.11
BED (dB)	39	−37.3 to −2.5	−19.0	9.3

Notes: The entire patient group consists of 39 (29 men, 10 women) patients. *Median fundamental frequency* (F0-median), the *standard deviation of the fundamental frequency* (F0-SD), and *jitter* could be calculated for 30 patients (9 voice samples of patients were considered unvoiced). The remaining measures, *percentage of voiced* (% Voiced), *harmonics-to-noise ratio* (HNR), *glottal-to-noise excitation ratio* (GNE), and *band energy difference* (BED) could be calculated for all 39 patients.

The results of the analyses of variance are given in Table 3. From this table it becomes clear that the acoustic measures based on pitch extraction (*median fundamental frequency, standard deviation of fundamental frequency,* and *jitter)* differentiate type IV signals from the other signal types simply by the fact that these measures cannot be calculated for most of the type IV voice samples. The *standard deviation of the fundamental frequency* differentiates between type I and III, being lower in type I signals. *Jitter* differentiates between type I and type II signals, and between type I and type III signals, being lower in type I signals. None acoustic measure differentiates between the type II and III signals. The *%Voiced* differentiates between all signal types, except types II and III. The *HNR* separates the type I signals from the types II, III, and IV signals and the type III signals from the type IV signals. The *BED* differs between the type IV signals and the type I signals and between the type

TABLE 3. *Results of Analyses of Variance on the Subgroups of Acoustic Signal Typing With the Acoustic Measures*

Variable	*P* value	Acoustic signal typing			
		Type I	**Type II**	**Type III**	**Type IV**
F0-median (Hz)	0.894	101	96	99	Excluded from analysis; could be calculated for one voice sample only
F0-SD (Hz)	0.004	2.38	4.34	10.65	
			.003		
Jitter%*	0.057	2.9	7.7	8.4	
HNR (dB)	<0.001	10.14	3.85	4.25	–0.03
			.001		
			.003		.034
			.001		
%Voiced*	<0.001	100	80.4	74.5	2.3
			.001		
		.011			.001
			.001		.001
GNE	0.127	0.82	0.77	0.82	0.72
BED (dB)	0.002	–24.8	–22.9	–17.0	–10.0
				.004	
					.004

Notes: For the measures *percentage of voiced* (%Voiced), *harmonics-to-noise ratio* (HNR), *glottal-to-noise-excitation ratio* (GNE), and *band energy difference* (BED), the analyses are based on four subgroups of acoustic signal typing [I (n = 7), II (n = 13), III (n = 11), and IV (n = 8)]. For the measures *F0-median, F0-SD,* and *percentage of jitter*, the analyses are based on three subgroups of acoustic signal typing [I (n = 7), II (n = 12), and III (n = 10)]. For the acoustic measures with a significant *P* value after Bonferroni correction ($P < .007$), a *post hoc* Tukey test was performed, of which the *P* values are shown in the boxes attached to the arrows.
*Nonparametric (Kruskal–Wallis and Mann–Whitney) tests were used.

IV signals and the type II signals, but not between the type IV and the type III signals. Although the *BED* is lower in the type III signals, this difference is not significant. The *GNE ratio* does not differentiate between any of the four signal types.

Relationships between acoustic signal typing and overall perceptual judgment

In Table 4, the relationship between the acoustic signal typing into the four different types and the overall perceptual judgment of voice quality by the

TABLE 4. *Table of the Relationship Among the Four Acoustic Signal Types and the Perceptual Judgment of Overall Voice Quality for All 39 Speakers ($P < 0.001$)*

Acoustic Signal Typing	Perceptual Judgment of Overall Voice Quality			
	Good	Reasonable	Poor	Total
Type I	5	2	0	7
Type II	6	6	1	13
Type III	2	4	5	11
Type IV	0	2	6	8
Total	13	14	12	39

Note: Numbers represent number of patients.

trained expert raters as ''good,'' ''reasonable,'' or ''poor'' is shown. A chi-squared test for linear-by-linear association shows that there is a significant relationship ($P < 0.001$). Type IV signals are never perceived as ''good,'' whereas type I signals are never perceived as ''poor.'' Two patients that show type III signals are nevertheless perceived as ''good''; they received the qualification type III because their voice sample did not meet the criterion of visible harmonics for longer than 2 seconds. As can be seen, type II signals occur both in ''good'' and ''reasonable'' voices; apparently a voice sample with one stable harmonic and an otherwise noisy spectrum can still be perceived as a ''good'' tracheoesophageal voice.

Acoustic analyses related to patient characteristics

No significant relationships were found between the acoustic signal types and patient characteristics. However, with respect to the acoustic measures, patients with a standard total laryngectomy ($N = 30$) showed a higher *median fundamental frequency* ($P = 0.008$; 111 Hz versus 65 Hz), and a larger *BED* ($P = 0.022$; –20.78 dB versus –12.84 dB) than did the patients with more extensive resection and reconstruction.

The remaining patient characteristics were studied within the standard total laryngectomy group only and did not show any differences. It is especially noteworthy that no difference in fundamental frequency between male and female tracheoesophageal speakers was found. Moreover, one female tracheoesophageal speaker produced the lowest median fundamental frequency of 46 Hz, whereas one male speaker produced the highest one of 229 Hz.

DISCUSSION

The aim of this study was to develop an acoustic signal typing system that is perceptually relevant, that can evaluate the entire range of tracheoesophageal voices, and that can serve as an underlying basis for further acoustic analyses.

Acoustic signal typing was adopted from Titze[7] on the basis of narrow-band spectrograms and was adapted for acoustic signal typing of tracheoesophageal voice quality. The narrow-band signal typing was an important part of the study, the clear criteria allowed easy classification. The visual differences between the narrow-band spectrograms were obvious and provided direct insight into the acoustic characteristics of the entire range of tracheoesophageal voices.

For acoustic analysis, seven voice-quality measures were computed. Five of these measures are regularly used by researchers for acoustic analysis of voice quality: *fundamental frequency*, *standard deviation of fundamental frequency*, *jitter*, *HNR*, and *%Voiced*. All of these five measures are based on pitch extraction, and the first three can only be calculated for voices with sufficient periodicity. The same yields true for other perturbation measures such as shimmer. In this study, only jitter was measured, because shimmer measures are not available in the *Praat* software, and in an earlier study, the different perturbation measures were found to be highly correlated to one another.[5] Ideally, reliable acoustic measures should be available for the entire range of voice qualities. Although *HNR* and *%Voiced* are based on the results of pitch extraction as well, results can be obtained for all voice samples (for instance, a completely aperiodic voice sample is 0% voiced and has a low HNR), and they are, therefore, interesting measures. The other two measures (*BED*[10] and *GNE ratio*[2]) are fundamental-frequency-independent measures and thus can be calculated for the entire range of tracheoesophageal voice qualities.

With the software program *Praat*, fundamental frequency measures could be calculated reliably for 77% of the voice samples. Additionally, four acoustic measures could be calculated for the entire patient group (*HNR, %Voiced*, *GNE ratio,* and *BED*). It meant that even for the poor speakers, an objective measure of voice quality could be obtained. It should, however, be mentioned that the adjustments that can be made in *Praat* to optimize the pitch extraction results may also lead to inconsistencies if used by researchers incorrectly.

As in several other studies,[2–5] in this study, the acoustic measures showed a wide range (including standard deviations), pointing to considerable variability among the speakers. The mean fundamental frequency of 104 Hz found in this study is comparable with the mean fundamental frequency of normal male speakers, who have fundamental frequencies around 110 Hz.[5] Fundamental frequencies for tracheoesophageal speech found in other studies are 115 Hz,[5] 83 Hz,[3] 92 Hz,[12] between 50 Hz and 110 Hz,[2] and between 33 Hz and 121 Hz.[4] Although the mean fundamental frequency of normal male speech (110 Hz) and tracheoesophageal speech is comparable in male patients (109 Hz on average), for female patients the fundamental frequency (115 Hz on average) is obviously too low in comparison with that of normal female speech (220 Hz). We agree with Moon and Weinberg[4] that the high degree of intersubject variation in fundamental frequency is an important characteristic of tracheoesophageal speech. It is also certainly a result of the formation of the speaker group studied: It was meant to include the entire range of tracheoesophageal voice qualities, and even a few speakers after partial or full pharyngeal reconstruction were included. Apparently the interspeaker variability among tracheoesophageal speakers is larger than among ''normal'' speakers, because of a larger variability of the anatomy and morphology of the neoglottis compared with the vocal folds. The mean and standard deviation found for the *BED* are in concordance with the values found by Debruyne et al.[2]

Relationships between the acoustic signal types and the acoustic measures show that for the acoustic signal types, significant differences between each group could be found for the measures *standard deviation of fundamental frequency, HNR, %Voiced,* and *BED*. For only one out of the eight voice samples in the type IV group, fundamental frequency measures could be calculated, which confirms that indeed most of the type IV voice sample does not contain a periodic voice signal. In the type I group, the *%Voiced* was always 100%, which clearly separated this type from the other types. The *standard deviation of the fundamental frequency* is remarkably higher in the type III group, which corresponds with the fact that the type III group contains instable voice samples. The *percentage of jitter* is remarkably lower in the type I group, which indicates that the fundamental frequency is most stable in this group. The *HNR* and the *BED* show a clear relationship with the four acoustic signal types.

The relationships between the four acoustic signal types and the overall perceptual judgment of voice quality show that the voice quality of the acoustically better type I and type II signals is perceived better as the voice quality of the acoustically poorer signal types. A type I signal is never perceived as ''poor'' and a type IV signal never as ''good.'' Not only type I signals are perceived as ''good,'' but also 50% of the type II signals are perceived as ''good.'' Two type III signals were also perceived as ''good.'' Closer inspection of the narrow-band spectrograms of those two speakers showed that they almost met the criterion of a stable harmonic for the full 2 seconds. Apparently for the perception of voice quality in read-aloud text, this criterion is not that important and it could be discussed whether the 2-second criterion should have been shorter. This clear relationship between this acoustic signal typing and the perceptual impression of overall voice quality supports the usefulness of acoustic signal typing in clinical judgment of tracheoesophageal voice.

The finding that the voice quality of the patient group after standard total laryngectomy was better compared with the patient group that underwent a partial or full pharyngeal reconstruction is also reflected in the acoustic measures. The patient group after standard total laryngectomy had a higher *fundamental frequency* and a larger *BED* than the patient group after pharyngeal reconstruction. Regarding sex differences, the absence of a difference

between the fundamental frequency of male (109 Hz on average) and female (115 Hz on average) tracheoesophageal speech is confirmed. In normal voices, an evident difference between male and female voice would be recognized. Most studies of tracheoesophageal speech consider male patients only. The reason is mainly a practical one: More men have developed laryngeal cancer to date. Trudeau and Qi[13] studied 10 female esophageal speakers and concluded that characteristics of tracheoesophageal speech may be highly similar regardless of speaker gender. The psychosocial implications of such a low-pitched voice for a female speaker are self-evident.

It should be noted that the acoustic signal typing system presented in this study has been developed for voice quality of tracheoesophageal speech by means of a voice prosthesis. Obviously, this system could also evaluate traditional esophageal speech or compare esophageal with tracheoesophageal speech. However, one should keep in mind that especially the 2-second criterion for stable voice could be difficult for esophageal speakers to obtain, which would automatically place most of them in a lower category. In that case, the criteria could be reconsidered depending on the goal of the evaluation.

CONCLUSION

It can be concluded from this study that the four acoustic signal types give a good (visual) impression of the overall quality of tracheoesophageal voice as demonstrated by their significant relationship with the overall perceptual judgment of voice quality. Relationships between the acoustic signal types and selected acoustic measures have shown that the signal types are useful as an indication for the appropriateness of more detailed acoustic analyses. The results of this study imply that this acoustic signal typing system might be a valuable clinical tool for documentation, investigation, and follow-up of voice quality in tracheoesophageal speech.

Acknowledgments: We thank Guus Hart for his help with the statistics; Ank Oubrie for her help with processing of the speech recordings; Rianne Polak, Benita Scholtens, and Brigitte Boon-Kamma for their participation in the listening experiment; Paul Boersma for his help with implementing several acoustic measures in *Praat*; Paul Boersma and Ton Wempe for their help and guidance during the acoustic analyses; and all patients for their participation in this study.

REFERENCES

1. van As CJ, Koopmans-van Beinum FJ, Hilgers FJM, Pols LCW. Perceptual evaluation of tracheoesophageal speech by naive and experienced judges through the use of semantic differential scales. *J Speech Lang Hear Res*. 2003;46:947–959.
2. Debruyne F, Delaere P, Wouters J, Uwents P. Acoustic analysis of tracheo-oesophageal versus oesophageal speech. *J Laryngol Otol*. 1994;108:325–328.
3. Robbins J, Fisher HB, Blom ED, Singer MI. A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *J Speech Hear Disord*. 1984;49:202–210.
4. Moon JB, Weinberg B. Aerodynamic and myoelastic contributions to tracheoesophageal voice production. *J Speech Hear Res*. 1987;30:387–395.
5. van As CJ, Hilgers FJM, Verdonck-de Leeuw IM, Koopmans-van Beinum FJ. Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. *J Voice*. 1998;12:239–248.
6. Yanagihara N. Significance of harmonic changes and noise components in hoarseness. *J Speech Hear Res*. 1967;10:531–541.
7. Titze IR. *Workshop on Acoustic Voice Analysis. Summary Statement*. Iowa City, IA: National Center for Voice and Speech; 1994.
8. Hilgers FJM, Ackerstaff AH, Balm AJM, Tan IB, Aaronson NK, Persson J-O. Development and clinical evaluation of a second-generation voice prosthesis (Provox®2), designed for anterograde and retrograde insertion. *Acta Otolaryngol (Stockh)*. 1997;117:889–896.
9. Boersma P, Weenink D. *Praat, A System for Doing Phonetics by Computer*. Amsterdam, The Netherlands: Institute of Phonetic Sciences, University of Amsterdam, IFA Report 132; 1996.
10. Dejonckere Ph, Lebacq J. Harmonic emergence in formant zone of sustained <a> as a parameter for evaluating hoarseness. *Acta Otorhinolaryngol Belg*. 1987;41:988–996.
11. Michaelis D, Gramms T, Strube HW. Glottal-to-noise excitation ratio: a new measure for describing pathological voices. *Acta Acoust*. 1997;83:700–706.
12. Bertino G, Bellomo A, Miani C, Ferrero F, Staffieri A. Spectrographic differences between tracheal-esophageal and esophageal voice. *Folia Phoniatr Logop*. 1996;48:255–261.
13. Trudeau MD, Qi Y. Acoustic characteristics of female tracheoesophageal speech. *J Speech Hear Disord*. 1990;55:244–250.