Acoustic measurement of overall voice quality: A meta-analysis[a)]

Youri Maryn[b)]
Department of Speech-Language Pathology and Audiology, Sint-Jan General Hospital, Ruddershove 10, 8000 Bruges, Belgium; Faculty of Health Care Vesalius, University College Ghent, Keramiekstraat 80, 9000 Ghent, Belgium; and Faculty of Social Health Sciences, University of Ghent, De Pintelaan 185, 9000 Ghent, Belgium

Nelson Roy
Department of Communication Sciences and Disorders, Division of Otolaryngology—Head and Neck Surgery, The University of Utah, 390 South 1530 East, Salt Lake City, Utah 84112-0252

Marc De Bodt
Department of Communication Disorders, University Hospital of Antwerp, Wilrijkstraat 10, 2650 Edegem, Belgium and Faculty of Social Health Sciences, University of Ghent, De Pintelaan 185, 9000 Ghent, Belgium

Paul Van Cauwenberge
Faculty of Medicine and Social Health Sciences, University of Ghent, De Pintelaan 185, 9000 Ghent, Belgium

Paul Corthals
Faculty of Health Care Vesalius, University College Ghent, Keramiekstraat 80, 9000 Ghent, Belgium and Faculty of Social Health Sciences, University of Ghent, De Pintelaan 185, 9000 Ghent, Belgium

(Received 27 July 2008; revised 10 August 2009; accepted 12 August 2009)

Over the past several decades, many acoustic markers have been proposed to be sensitive to and measure overall voice quality. This meta-analysis presents a retrospective appraisal of scientific reports, which evaluated the relation between perceived overall voice quality and several acoustic-phonetic correlates. Twenty-five studies met the inclusion criteria and were evaluated using meta-analytic techniques. Correlation coefficients between perceptual judgments and acoustic measures were computed. Where more than one correlation coefficient for a specific acoustic marker was available, a weighted average correlation coefficient was calculated. This was the case in 36 acoustic measures on sustained vowels and in 3 measures on continuous speech. Acoustic measures were ranked according to the strength of the correlation with perceptual voice quality ratings. Acoustic markers with more than one correlation value available in literature and yielding a homogeneous weighted r of 0.60 or above were considered to be superior. The meta-analysis identified four measures that met these criteria in sustained vowels and three measures in continuous speech. Although acoustic measures are routinely utilized in clinical voice examinations, the results of this meta-analysis suggest that caution is warranted regarding the concurrent validity and thus the clinical utility of many of these measures. © *2009 Acoustical Society of America.*
[DOI: 10.1121/1.3224706]

PACS number(s): 43.70.Jt, 43.72.Ar [AL] Pages: 2619–2634

I. INTRODUCTION

Evaluation of voice quality is considered an essential but controversial part of the assessment process in the field of voice pathology. In clinical as well as in research settings, two main approaches exist to describe the perceived severity of a voice disorder (Kreiman and Gerratt, 2000a). First, generic and/or global ratings such as "overall voice quality," also known as "G" (for "grade"), "severity of voice disorder," "severity of dysphonia," "overall abnormality," and "overall severity" have been used to capture a composite perceptual judgment of the degree of the perceived dysphonia. In contrast, other voice quality ratings pertain to single and very specific perceptual dimensions, the best known of which are roughness and breathiness. Recent evidence has suggested that perceptual rating of overall voice quality and other more specific perceptual dimensions is difficult, as such judgments depend on the listener's internal standard or scale for voice quality dimensions, on his/her sensitivity for this particular dimension, on fatigue, attention, exposure to various disordered voices, and training in perceptual evaluation of voice quality (Kreiman *et al.*, 1993; Eadie and Baylor, 2006). Furthermore, other aspects of voice quality judgments, such as type and range of the scale (Bele, 2005; Eadie and Doyle, 2002), or the type of sample to be evaluated, such

[a)]Portions of this work were presented in "Forty years of acoustic prediction of overall voice quality" at the 7th International Voice Symposium, Austrian Voice Institute, Salzburg, Austria, August 4–6, 2006.
[b)]Author to whom correspondence should be addressed. Electronic mail: youri.maryn@azbrugge.be

as sustained vowel versus continuous speech (Bele, 2005; Zraick *et al.*, 2005; Eadie and Baylor, 2006), can significantly affect the perceptual evaluation of voice quality.

In spite of these listener-related and other potential biases, many researchers have tried to correlate the outcome of acoustic-phonetic measures to vocal quality ratings and dysphonia severity. The replacement of analog recording systems with digital recording systems, the availability of automated analysis algorithms, and the non-invasiveness of acoustic measures, combined with the fact that acoustic parameters provide easy quantification of dysphonia improvement during the treatment process, have led to considerable interest in clinical voice quality measurement using acoustic analysis techniques.

In this regard, the correlation coefficient has emerged as the most frequently used index to determine the extent of the relationship or effect size between acoustic measures and listener judgments of dysphonia severity. The correlation coefficient as a measure of effect size measures the strength and direction of a linear relationship between two variables. In the voice quality literature, perceived overall voice quality is treated as the dependent variable with the objective acoustic measure treated as the independent variable. The degree of the linear relationship between dependent and independent variables (i.e., correlation) counts as an indication of validity, or the extent to which the score of a measurement (i.e., the acoustic parameter) can be regarded as a valid measure of the dependent variable (i.e., the perceptual rating). Consequently, the higher the absolute correlation coefficient, the more the acoustic measure is said to reflect the perception of overall voice quality, and vice versa. The correlation coefficient is thus an important and frequently used statistic in voice quality research, especially to validate acoustic measures.

Although the correlation coefficient is a frequent metric to assess the strength of the acoustic-perceptual relationship, at least 60 possible acoustic determinants of overall voice quality with varying predictive power have been identified in literature over the past 4 decades. Buder (2000) proposed a taxonomy of 15 signal processing-based categories to help manage the wide array of acoustic measures. The large numbers of studies reviewed by Buder (2000) clearly differ substantially in the number of participants and the magnitude of correlation with perceptual judgments of voice quality. Furthermore, the signal processing strategies vary from classic spectrography to sophisticated statistics on sound wave microstructure. Whereas some authors examined the predictive power of resonance-based aspects, the majority of investigators focused on glottal rather than on supraglottal phenomena, seeking correlates of overall voice quality in the distribution of fundamental frequency, in waveform perturbations, in various spectral parameters (including cepstral coefficients and noise content of the glottal sound source), in glottal air flow models obtained by inverse filtering, or in models based on non-linear dynamics theory.

Although an impressive body of research exists, which ostensibly assesses the utility of acoustic measurement to quantify voice quality and dysphonia severity, procedural differences in type and number of acoustic predictors, type of recorded material, analysis equipment, and measurement scales have made it almost impossible to qualitatively appraise the merits of these studies, and precisely define a subset of the most robust and sensitive acoustic measures. One approach to this seemingly intractable problem is to apply meta-analytic techniques. Meta-analysis refers to "the analysis of analyses," and is a statistical technique for amalgamating, summarizing, and reviewing previous quantitative research. Unlike traditional research methods, meta-analysis uses the summary statistics from individual studies as the data points for the purpose of integrating the findings. A key assumption of this analysis approach is that each study provides a different estimate of the underlying relationship within the population. By accumulating results across studies, one can gain a more accurate representation of the population relationship than is provided by the individual study estimators. In this way, meta-analyses permit confidence that the reported results are based on more than one study that found the same result (Frey *et al.*, 1991; Lipsey and Wilson, 2001).

Meta-analysis reviews findings in terms of effect sizes. Defining an effect size statistic that adequately represents the quantitative findings of an assortment of research reports in a standardized profile is essential to meta-analysis, as it permits meaningful numerical comparison and analysis (Lipsey and Wilson, 2001). The effect size provides information about the magnitude of the relationships observed across all studies and for subsets of studies. By treating individual correlation coefficients as indicators of effect size, meta-analysis can regroup study outcomes into homogeneous subsets and establish population effect sizes. The population effect size, i.e., the real relationship between an independent variable (a specific acoustic measure) and the dependent variable (a voice severity rating), is estimated by a "weighted" average of all correlations available for a particular acoustic marker. In addition to defining a weighted average of all effect sizes (i.e., correlation coefficients) in a meta-analysis, it is also important to know whether or not the various effect sizes all estimate the same population effect size. This is a question of homogeneity (or heterogeneity) of the effect size distribution, and a population effect size can only be interpreted reliably if the underlying data set is sufficiently homogeneous (Hunter *et al.*, 1982). When the variability of effect sizes around their weighted mean is no larger than the dispersion expected from sampling error alone, the effect size distribution is considered to be homogeneous. By comparison, in a heterogeneous distribution, individual effect sizes differ from the weighted mean by more than the sampling error (Lipsey and Wilson, 2001). Multiple correlation coefficients resulting in a homogeneous weighted mean correlation are considered to confirm each other, thereby increasing the generalizability of the findings.

Given the large body of research that relates acoustic measures to voice quality ratings, meta-analysis techniques can potentially reduce information overload, and distill this large literature into a manageable and/or tractable set of conclusions. Therefore, the aim of this meta-analysis is twofold: (1) to retrospectively appraise the acoustic-phonetic markers

for overall voice quality (e.g., dysphonia severity) and (2) to establish population relationship estimates for several acoustic measures.

II. METHOD

In most research on assessment of voice quality, measurements have been completed on sustained vowels as compared to continuous speech. This preference for sustained vowels over continuous speech in acoustic as well as perceptual measurements of voice quality has been motivated by several factors (Askenfelt and Hammarberg, 1986; Parsa and Jamieson, 2001), such as follows: (a) sustained vowels represent relatively time-invariant vocal phonation whereas continuous speech involves quick and continuous alterations of glottal and supraglottal mechanisms; (b) in contrast to continuous speech, sustained mid-vowel segments do not contain non-voiced phonemes, rapid voice onsets and offsets, or prosodic fundamental frequency and amplitude fluctuations; and (c) sustained vowels are not affected by speech rate, vocal pauses, phonetic context, and stress. However, sustained vowels may lack representation of daily speech and voice (Parsa and Jamieson, 2001; Eadie and Baylor, 2006), and continuous speech potentially contains perceptual cues, which are often considered to be decisive in vocal quality evaluations (Askenfelt and Hammarberg, 1986). Since both sample types offer valuable information in voice quality measurements, the present meta-analysis focused on studies of sustained vowels as well as connected speech.

A. Search strategy

Relevant scientific reports were identified by a systematic electronic search of the Medline database and the corpus of online publications by the American Speech-Language-Hearing Association. The combination of (a) keywords referring to composite perceptual voice evaluations and (b) keywords related to the concepts of measuring by means of acoustic markers was used as a guide. Using information derived from the titles and abstracts, an initial set of pertinent articles was generated. Subsequently, a manual search for references in relevant literature sources was launched using the same guide. This manual search started from the sources cited in the initial set of articles garnered from the electronic search and from periodicals, book chapters, and various bibliographies likely to contain relevant references and texts.

B. Inclusion and exclusion of literature sources

In order to be included, a study had to report sufficient mathematical detail on bivariate correlation coefficients establishing the relation between perceptual overall voice quality ratings of sustained vowels or continuous speech (the dependent variable) and one or more acoustic parameters derived from the same samples (the independent variables). Studies citing relevant correlation coefficients were included, whether or not significance levels were reported, and every study describing auditory-perceptual ratings of overall quality (e.g., dysphonia severity) was included, regardless of the type of rating scale used.

Investigations of acoustic correlates of specific perceptual dimensions such as breathiness and roughness were not included in the meta-analysis, as the present study concentrated on "composite" or "global" overall voice quality correlates. Furthermore, reports on non-acoustic or non-objective correlates, such as aerodynamic measures or electroglottographic parameters, were also excluded, as well as studies dealing with the relationship between the auditory-perceptual evaluation of overall voice quality and its visual-perceptual representation in narrowband spectrograms. Furthermore, since the present study aimed to focus on acoustic-auditory determinants of dysphonia severity, studies investigating the correlation between objective acoustic measures and visual inspection of spectrograms or other diagrams were excluded. Also, reports on parameters derived from synthesized vowel samples were not included in this study. Reports lacking sufficient quantitative and critical information, such as number of subjects or type of samples, were also excluded.

Methodological articles related exclusively to the use and development of perceptual rating scales or acoustic algorithms, which did not provide inferential statistics on the validity of the acoustic measure(s), were also excluded. Studies appraising the diagnostic value of acoustic parameters (i.e., the power of a diagnostic tool to discriminate between presence or absence of a voice disorder), expressed as sensitivity, specificity, positive predictive value, negative predictive, and/or area under the receiver operating characteristic (ROC) curve, or outcomes of studies based on comparative statistics between normal and pathologic voices, as expressed in chi-square tests, Mann-Whitney U tests, t tests, etc., were not included, because the present study concentrated on the correlation coefficient as population effect size.

In addition, reports on multivariate analyses were excluded, unless bivariate (zero-order) correlation coefficients were clearly identified (as in Wolfe and Martin, 1997; Wolfe *et al.*, 1997; Yu *et al.*, 2001; Eadie and Baylor, 2006; Ma and Yiu, 2006). The reason for excluding multiple regression studies is based on the assumption that some independent variables are dropped from the initial set of possible predictors as a result of co-linearity. A relevant independent variable, correlating well with the dependent variable, may be dropped when, in the presence of other independent variables, it does not substantially increase the amount of variance explained. This phenomenon makes it difficult to assess the separate contribution of each independent variable to the measurement of the dependent variable. Moreover, the algorithm of a multiple regression not only looks for a parsimonious equation, but also gives each remaining independent marker a coefficient that can only be interpreted in combination with the particular set of remaining independent markers in the rest of the equation. As a consequence, we had to exclude study results using multivariate statistics from the meta-analysis.

Finally, reliability of the auditory-perceptual ratings of voice quality, as an index on which an acoustic measure is validated, is also an important consideration (Kreiman and Gerratt, 2000a, 2000b; Kreiman *et al.*, 2007). Reliability of auditory-perceptual ratings is traditionally described in terms

of within and between listener reliability, consistency, agreement, or concordance. Such intra- and interrater reliability is considered an important prerequisite for validity. High reliability clearly and precisely defines the perceptual construct to be measured by an acoustic parameter. In contrast, listener unreliability increases "non-experimental" or "error" variance, thereby reducing the true variance in the perceptual construct that is to be accounted for by the acoustic measure. Thus, the increase in error variance due to listener unreliability should decrease the concurrent validity of the acoustic measure, as evaluated by a correlation coefficient. In single experiments, acceptable rater reliability is often considered an essential prerequisite before attempting to assess an acoustic measure's worth in estimating overall voice quality. However, across studies, many statistics have been used to measure rater reliability, for instance, Pearson's product-moment correlation coefficient, Cohen's kappa correlation coefficient, Cronbach's alpha correlation coefficient, and intraclass correlation coefficient, to mention only a few. Given the large number of studies reviewed in this meta-analysis, each using a variety of listeners (with differing levels of experience and training), and different scales with various interpretation guidelines so as to determine "adequate" reliability, we elected to treat listener reliability as a nuisance variable, and to not exclude any studies solely on the basis of their estimates of listener reliability. This decision is predicated on the assumption that listener unreliability essentially contributes to error variance, and necessarily attenuates any investigator's ability to identify significant correlations between listener ratings and specific acoustic measures. By treating listener reliability/unreliability as a nuisance variable, one that would necessarily vary between studies and differentially contribute to error variance, we assumed that across studies, the most compelling acoustic-perceptual relationships would eventually surface, having survived the potentially attenuating effects of listener unreliability.

Analogous to the listener reliability/unreliability, we also elected to treat between-study differences in data acquisition and processing methodology as a nuisance variable. Variety in room acoustics, microphone type and placement, software, analysis algorithms, etc., also creates error variance, and similarly decreases the variance in the perceptual construct that is to be explained by the acoustic measure. The large number of studies, each with its own acoustical configuration and hardware and software settings, clearly limits our ability to directly compare the outcomes of the studies. However, a guiding principle of meta-analysis is that the consistency of the significant results/conclusions across studies is paramount, and robust relationships should withstand such methodological "noise" (regardless of the source of the noise, e.g., listener unreliability, recording instrumentation and surroundings, computer software, etc). We therefore elected to consider methodological variations in recording conditions/settings, data acquisition and analysis algorithms, etc., as additional sources of error variance, and an inherent limitation of the meta-analysis. On the other hand, acoustic measures that yield consistent outcomes across a variety of study methods can be considered especially robust. As such, the inclusion of studies with varying methodology is considered advantageous in the present meta-analysis.

Originally, 85 reports were considered. Based on the aforementioned inclusion and exclusion criteria, however, many reports were excluded, producing a final corpus of 25 studies on which the meta-analysis was performed. Twenty-one studies involved measurements on sustained vowel samples (methodological aspects of these studies are summarized in Table I). Seven studies involved measurement on continuous speech samples (methodological aspects of these studies are similarly summarized in Table II). However, 3 studies contained information on both continuous speech and sustained vowels (Heman-Ackah *et al.*, 2002; Halberstam, 2004; Eadie and Baylor, 2006), thus leaving a total of 25 studies (i.e., $28-3=25$).

From these studies, a list of acoustic measures was generated. Subsequently, the measures were organized based on their description in the Method section of the original publication. The tabulation of Buder (2000) was chosen only as a loose framework to group the measures. Buder's (2000) tabulation was the first compilation of acoustic voice measures, as it presented a complete overview of acoustic measures in a comprehensive and consistently structured manner. It therefore served as a basis on which the measures of this meta-analysis were considered to be similar or different. In studies that analyzed sustained vowels, there were 69 acoustic markers identified, whereas in the connected speech studies, 26 acoustic measures were reported. Eighty-seven acoustic markers have been identified as measures of overall voice quality in the included studies. Table III lists these acoustic measures alphabetically and provides (for every measure) references expanding on the rationale and the digital signal processing underlying the measure.

C. Statistical analysis

Quantitative data from the selected scientific reports were analyzed using statistical software packages for personal computers, including Microsoft Office Excel 2003 and Meta-Analysis Programs version 5.3 (Ralf Schwarzer, Department of Psychology, Freie Universität Berlin, Germany). Meta-analyses on correlation coefficients according to the Schmidt-Hunter method (Hunter *et al.*, 1982) were performed on all acoustic voice quality correlates for which more than one effect size was available. This method is based on four statistics. The first statistic is the *number of effect sizes* (k) or the number of available bivariate correlation coefficients for a given acoustic measure. The second statistic is the total *number of subjects* (N). The third statistic is the *population effect size* or the weighted mean correlation coefficient ($\overline{r_w}$). Correlation coefficients based on studies with large sample sizes digress less from the population effect size and therefore more weight is assigned to large N effect sizes (Hunter *et al.*, 1982). If only one effect size is reported, a weighted effect size cannot be calculated. In this case, there is no meta-analysis and the discussion is based on the initial and solitary r-value. While there is no firm criterion or universal consensus for evaluating the magnitude of correlation coefficients (Frey *et al.*, 1991), we chose a corre-

TABLE I. Methodological features (number of subjects, type of voice recording, and organization and reliability of the perceptual ratings) of the 21 chronologically ordered studies included in this meta-analysis on sustained vowels.

Source	Subjects[a]			Voice sample[b]		Perceptual evaluation[b]				
	N	*P*	*T*	Vowel	Duration	No. of judges	Rating scale[c]	Perceptual construct	Intrarater reliability[d]	Interrater reliability[d]
Kojima *et al.* (1980)	28	30	58	/a/	NA	5	EAI (4)	Hoarseness	NA	NA
Yumoto *et al.* (1984)	0	87	87	/a/	3 s	8	EAI (4)	Hoarseness	NA	0.51–0.79 Sp
Hirano *et al.* (1986)	0	68	68	/e/	NA	NA	EAI (4)	G, grade	NA	NA
Prosek *et al.* (1987)	0	90	90	/a/	2 s	9	EAI (7)	Severity of voice disorder	0.90 Pe	0.82 Cr
	16	44	60			14		Hoarseness	NA	NA
Wolfe and Steinfatt (1987)	0	51	51	/a/ and /i/	1 s	8	EAI (7)	Severity of dysphonia Hoarseness	89% Ag	0.95 Cr
Feijoo and Hernández (1990)	64	57	121	/e/	NA	4	EAI (4)	G, grade	77.48% Ag	98.35% Ag
Kreiman *et al.* (1990)	0	18	18	/a/	1.67 s	10	EAI (7)	Overall abnormality	NA	NA
Wolfe *et al.* (1995)	20	60	80	/a/	1 s	22	EAI (7)	Overall severity	0.99 Cr	0.98 Cr
Dejonckere *et al.* (1996)	0	943	943	/a/	2 s	2	EAI (4)	G, grade	0.51 Co	0.87 Sp
Dejonckere and Wieneke (1996)	0	28	28	/a/	0.1 s	2	EAI (5)	Overall severity of hoarseness	NA	NA
De Bodt (1997)	98	634	732	/a/	3 s	1	EAI (4)	G, grade	NA	NA
Plant *et al.* (1997)	0	26	26	/i/	2 s	3	EAI (5)	Overall voice quality	0.86 NA	NA
Wolfe and Martin (1997)	0	51	51	/a/ and /i/	1 s	11	EAI (7)	Hoarseness	76% Ag	0.95 Cr
Wolfe *et al.* (1997)	0	51	51	/a/ and /i/	1 s	18	VAS	Hoarseness	0.80 Pe	0.94 Cr
Wolfe *et al.* (2000)	0	20	20	/a/	1 s	11	EAI (7)	Abnormality	0.81 Pe	0.98 Cr
Yu *et al.* (2001)	21	63	84	/a/	2 s	6	EAI (4)	G, grade	Cons	Cons
Heman-Ackah *et al.* (2002)	0	14	14	/a/	1 s	2	EAI (4)	G, grade	NA	0.83 Pe
Halberstam (2004)	0	60	60	/a/	1 s	2	EAI (7)	Hoarseness	0.89 NA	0.91 Cr
Eadie and Baylor (2006)	3	9	12	/a/	1 s	16	VAS	Overall severity	0.82–0.95 Pe	0.72–0.83 Pe
Gorham-Rowan and Laures-Gore (2006)	0	28^{ym}	28^{ym}	/a/	1 s	10	FMMEP	Hoarseness	−0.32 to 0.86 Pe	0.80 Cr
	0	28^{ew}	28^{ew}							
	0	28^{em}	28^{em}							
Yu *et al.* (2007)	38^{w}	270^{w}	308^{w}	/a/	2 s	4	VAS	G, grade	NA	NA
	20^{m}	121^{m}	141^{m}							

[a]*N*=number of normal subjects, *P*=number of pathological or dysphonic subjects, *T*=total number of subjects, ^{ym}=young men, ^{ew}=elderly women, ^{em}=elderly men, ^{m}=men, and ^{w}=women.

[b]NA=the information is not available in the original manuscript.

[c]EAI=equal-appearing interval scale with between parentheses the number of points on the scale, VAS=visual analog scale, and FMMEP=free modulus magnitude estimation paradigm.

[d]Sp=Spearman's rank-order correlation coefficient, Pe=Pearson's product-moment correlation coefficient, Co=Cohen's kappa correlation coefficient, Cr=Cronbach's alpha correlation coefficient, Ke=Kendall's coefficient of concordance, Ag=percentage of agreement/consistency between judgments, and Cons=consensus between listeners without quantitative measure of reliability.

lation coefficient (r or $\overline{r_w}$) of 0.60 as the cutoff to distinguish between strong and weak acoustic markers. Following the guidelines established by Franzblau (1958), this threshold intends to separate a "moderate" degree of correlation from a "marked" degree of correlation. It should be acknowledged, however, that other interpretations have been proposed, including Frey *et al.* (1991), for example, who recommended r of 0.70 to distinguish between moderate and marked correlations. We selected a less stringent correlation coefficient $r=0.60$ in light of our decision to treat listener unreliability and methodological/procedural differences as sources of error variance, which would potentially attenuate the strength of reported bivariate correlations across studies. The fourth statistic relates to the homogeneity or heterogeneity of the effect sizes. A population effect size can only be interpreted reliably if the underlying data set is sufficiently homogeneous (Hunter *et al.*, 1982). Here one can rely on several indicators: (1) the residual standard deviation, (2) the percentage of observed variance accounted for by the sampling error, and (3) the chi-square value. However, the preferred index for homogeneity is the population variance or its square root, called *residual standard deviation* (SD_{res}). This indicator, SD_{res}, is the variance left after the sampling error has been subtracted (Hunter *et al.*, 1982). Ideally, SD_{res} equals zero, meaning that all the observed variance is accounted for by sampling error and that the data set of correlations is completely homogeneous. If the analysis, however, failed to identify a source of systematic variation in the data, SD_{res} is indicative of heterogeneity. As a rule of thumb, a set of effect sizes can be considered homogeneous when SD_{res} is less than $\frac{1}{4}$ of $\overline{r_w}$ (Hunter *et al.*, 1982; Lipsey and Wilson, 2001).

III. RESULTS

A. Sustained vowels

Twenty-one studies meeting the selection criteria were identified; the majority originated from the *Journal of*

TABLE II. Methodological features (number of subjects, type of voice recording, and organization and reliability of the perceptual ratings) of the seven chronologically ordered studies included in this meta-analysis on continuous speech.

Source	Subjects[a]			Voice sample	Perceptual evaluation[b]				
	N	*P*	*T*		No. of judges	Rating scale[c]	Perceptual construct	Intrarater reliability[d]	Interrater reliability[d]
Askenfelt and Hammarberg (1986)	0	41	41	Voiced segments, 40 s of reading a story	6	EAI (6)	Overall voice quality	0.86–0.98 Pe	NA
Qi *et al.* (1999)	0	87	87	First and second sentences from rainbow passage	5	VAS	Overall voice quality	0.93–0.96 Pe	0.97 Cr
Heman-Ackah *et al.* (2002)	0	18	18	Second sentence from rainbow passage	2	EAI (4)	G, grade	NA	0.83 Pe
Halberstam (2004)	0	60	60	12 s from rainbow passage	2	EAI (7)	Hoarseness	0.93 NA	0.97 Cr
Eadie and Doyle (2005)	6	24	30	Second sentence from rainbow passage	12	DME	Overall severity	0.69 Pe	0.97 Cr
Eadie and Baylor (2006)	3	9	12	Second sentence from rainbow passage	16	VAS	Overall severity	0.80–0.97 Pe	0.84–0.91 Pe
Ma and Yiu (2006)	41	112	153	/ba ba d a bo /	4	EAI (11)	G, grade	≥0.90 Pe	0.86–0.91 Pe

[a]*N*=number of normal subjects, *P*=number of pathological or dysphonic subjects, and *T*=total number of subjects.
[b]NA=the information is not available in the original manuscript.
[c]EAI=equal-appearing interval scale with between parentheses the number of points on the scale, VAS=visual analog scale, and DME=direct magnitude estimation.
[d]Pe=Pearson's product-moment correlation coefficient and Cr=Cronbach's alpha correlation coefficient.

Speech (Language) and Hearing Disorders (6), *Journal of Voice* (3), and *Journal of Communication Disorders* (3). Other sources were *Acta Otorhinolaryngologica Belgica* (1), *Acta Otolaryngologica* (1), *Folia Phoniatrica et Logopaedica* (2), *Journal of Phonetics* (1), *Laryngoscope* (1), *ORL* (1), *Revue de Laryngologie-Otologie-Rhinologie* (1), and a chapter in volume VI of *Advances in Clinical Phonetics* (1). Relevant information concerning the methodology of the reports that were included in the meta-analysis can be found in Table I. All 21 studies reported on pathologic or dysphonic voices; however, only 8 studies also contained normal voices. The mean number of dysphonic voice samples was 115 (range 9–943). For the normal voices, the mean number was 34 (range 3–98). The total number of subjects was 116 on average and ranged from 12 to 943. In these studies, 146 distinct effect sizes (i.e., correlation coefficients) were reported, pertaining to 69 different acoustic measures as displayed in Table IV.

All acoustic parameters and data on sustained vowels were extracted from the central portion of the recordings. The length of the mid-vowel segment varied from 0.1 to 3 s with a mean duration of 1.5 s. 1 s was the modal duration, occurring in 50% of the studies (the duration was not specified in three studies). The vowels [a:], [i:], and [e:] were analyzed in 86%, 19%, and 10% of the studies, respectively. Substantial differences existed among the data acquisition systems that were used, which could potentially influence the outcome of acoustic measurements. For instance, recording equipment (e.g., type of microphone and microphone location relative to the sound source, and type of hardware), processing algorithms, measurement algorithms, and software settings such as sampling rate or method of fundamental period extraction varied among the studies and have been demonstrated to influence the outcome of acoustic measurements, particularly the outcomes of perturbation measures.

For the perceptual experiments, the number of judges ranged from 1 to 22, with a mean value of 8. The rating scale used was typically an equal-appearing interval scale, using 4, 5, or 7 points in 38%, 10%, and 33% of studies, respectively. In two studies (Wolfe *et al.*, 1997; Yu *et al.*, 2007), a visual analog scale was used. In Gorham-Rowan and Laures-Gore, 2006 a free modulus magnitude estimation paradigm was used. A variety of perceptual labels were used including hoarseness, G (from grade), severity of voice disorder, severity of dysphonia, overall abnormality, overall severity, overall severity of hoarseness, abnormality, and overall voice quality. A variety of estimates of inter- and intrajudge reliability estimates were used (see table entries). As mentioned previously, the variety and range of methods to determine reliability hamper comparisons between studies. In general, intrajudge reliability fluctuated from rather low, as in Dejonckere and Wieneke, 1996 and Gorham-Rowan and Laures-Gore, 2006, to very high, as in Wolfe *et al.*, 1995 and Prosek *et al.*, 1987. Similar variability was observed for interjudge reliability.

1. Meta-analysis on correlation coefficients

The results of the meta-analysis on sustained vowels are summarized in Table IV and Fig. 1. For 33 of the 69 acoustics measures (48%), there was only 1 correlation coefficient available and, consequently, no weighted mean correlation coefficient could be determined. For the remaining 36 acoustic determinants (52%), there was more than 1 correlation coefficient and the *k*-values ranged from 2 to 7. The most frequently investigated parameters were noise-to-harmonics ratio (NHR) from multi-dimensional voice program (MDVP) ($k=7$), and the vocal perturbation measures amplitude perturbation quotient, percent jitter, and percent shimmer ($k=6$). For these 36 markers, a $\overline{r_w}$ was calculated with Meta-Analysis Programs version 5.3. The organization of the meta-analysis on acoustic measures on sustained vowels is illustrated in Fig. 1.

TABLE III. The 87 acoustic measures included in this study, alphabetically ordered on the basis of their full name, with their respective sources/citations identified.

Acoustic measure	Sources included in study
Absolute jitter	Kreiman *et al.* (1990), De Bodt (1997), Wolfe *et al.* (1997), and Halberstam (2004)
Amplitude perturbation quotient	De Bodt (1997), Wolfe *et al.* (1997), Heman-Ackah *et al.* (2002), Halberstam (2004), and Gorham-Rowan and Laures-Gore (2006)
Amplitude perturbation quotient of residue signal	Prosek *et al.* (1987)
Area of voice range profile	Ma and Yiu (2006)
Breathiness index	Plant *et al.* (1997) and Wolfe *et al.* (2000)
Cepstral peak magnitude (also known as magnitude of first harmonic)	Dejonckere and Wieneke (1996)
Cepstral peak prominence	Wolfe and Martin (1997), Wolfe *et al.* (2000), Halberstam (2004), and Eadie and Baylor (2006)
Cepstrum of excitation signal	Feijoo and Hernández (1990)
Coefficient of excess	Prosek *et al.* (1987)
Coefficient of variation of fundamental frequency	Wolfe *et al.* (1997)
Coefficient of variation of jitter	Kreiman *et al.* (1990)
Coefficient of variation of period	Wolfe and Steinfatt (1987)
Coefficient of variation of shimmer	Kreiman *et al.* (1990)
Compression of relative frequency differences	Askenfelt and Hammarberg (1986)
Cycle-of-cycle variation of waveform	Feijoo and Hernández (1990)
Difference between frequencies of second and first formants	Kreiman *et al.* (1990)
Directional perturbation factor	Askenfelt and Hammarberg (1986)
Fluctuation in amplitude	Hirano *et al.* (1986)
Fluctuation in fundamental frequency	Hirano *et al.* (1986)
Frequency-domain harmonics-to-noise ratio	Eadie and Doyle (2005)
Frequency of first formant	Kreiman *et al.* (1990)
Frequency of second formant	Kreiman *et al.* (1990)
Frequency of third formant	Kreiman *et al.* (1990)
Fundamental frequency	Yu *et al.* (2001), Yu *et al.* (2007), and Ma and Yiu (2006)
Fundamental frequency range in voice range profile	Ma and Yiu (2006)
Harmonics-to-noise ratio from Kojima	Kojima *et al.* (1980)
Harmonics-to-noise ratio from Yumoto	Yumoto *et al.* (1984), Kreiman *et al.* (1990), and Wolfe *et al.* (1995)
Highest fundamental frequency in voice range profile	Ma and Yiu (2006)
Intensity range in voice range profile	Ma and Yiu (2006)
Jitter factor	Yu *et al.* (2001) and Yu *et al.* (2007)
Jitter from Yumoto	Yumoto *et al.* (1984)
Jitter ratio	Wolfe and Steinfatt (1987) and Dejonckere and Wieneke (1996)
Lowest fundamental frequency in voice range profile	Ma and Yiu (2006)
Lyapunov coefficient	Yu *et al.* (2001) and Yu *et al.* (2007)
Maximum intensity in voice range profile	Ma and Yiu (2006)
Mean harmonic emergence between 500 and 1500 Hz	Dejonckere and Wieneke (1996)
Minimum intensity in voice range profile	Ma and Yiu (2006)
Natural logarithm of standard deviation of period	Wolfe and Steinfatt (1987) and Kreiman *et al.* (1990)
Noise-to-harmonics ratio	Wolfe *et al.* (1997)
Noise-to-harmonics ratio from MDVP	Dejonckere *et al.* (1996), De Bodt (1997), Heman-Ackah *et al.* (2002), Halberstam (2004), Gorham-Rowan and Laures-Gore (2006), and Ma and Yiu (2006)
Normalized mean absolute period jitter	Feijoo and Hernández (1990)
Normalized mean absolute period shimmer	Feijoo and Hernández (1990)
Normalized noise energy	Feijoo and Hernández (1990)
No. of harmonics	Kreiman *et al.* (1990)
Partial period comparison	Kreiman *et al.* (1990)
Peakedness of relative frequency differences	Askenfelt and Hammarberg (1986)
Pearson r at autocorrelation peak	Wolfe *et al.* (2000)
Percent jitter	Kreiman *et al.* (1990), De Bodt (1997), Plant *et al.* (1997), Wolfe and Martin (1997), Wolfe *et al.* (1997), and Halberstam (2004)
Percent shimmer	Kreiman *et al.* (1990), Dejonckere *et al.* (1996), De Bodt (1997), Wolfe and Martin (1997), Wolfe *et al.* (1997), Halberstam (2004), and Ma and Yiu (2006)
Perturbation factor	Askenfelt and Hammarberg (1986)

TABLE III. (*Continued.*)

Acoustic measure	Sources included in study
Perturbation magnitude	Askenfelt and Hammarberg (1986)
Perturbation magnitude mean	Askenfelt and Hammarberg (1986)
Phonatory fundamental frequency range	De Bodt (1997), Yu *et al.* (2001), Halberstam (2004), and Yu *et al.* (2007)
Pitch amplitude	Prosek *et al.* (1987), Plant *et al.* (1997), and Eadie and Doyle (2005)
Pitch perturbation quotient	De Bodt (1997), Wolfe *et al.* (1997), and Halberstam (2004)
Pitch perturbation quotient of residue signal	Prosek *et al.* (1987)
Power spectrum ratio	Wolfe *et al.* (2000)
Ratio of amplitudes of first and second harmonics	Kreiman *et al.* (1990)
Ratio of frequencies of second and first formants	Kreiman *et al.* (1990)
Relative average perturbation	Wolfe *et al.* (1995), De Bodt (1997), Wolfe *et al.* (1997), Heman-Ackah *et al.* (2002), Halberstam (2004), and Ma and Yiu (2006)
Relative noise level	Hirano *et al.* (1986)
Residue signal power ratio	Plant *et al.* (1997)
Richness of high frequency harmonics	Hirano *et al.* (1986)
Shimmer in decibel	Wolfe *et al.* (1995), De Bodt (1997), Wolfe *et al.* (1997), and Halberstam (2004)
Signal-to-noise ratio	Yu *et al.* (2001) and Yu *et al.* (2007)
Signal-to-noise above 1000 Hz	Yu *et al.* (2001) and Yu *et al.* (2007)
Signal-to-noise ratio from Milenkovic	Wolfe and Martin (1997)
Signal-to-noise ratio from Qi	Qi *et al.* (1999) and Eadie and Doyle (2005)
Smoothed amplitude perturbation quotient	De Bodt (1997), Wolfe *et al.* (1997), and Halberstam (2004)
Smoothed cepstral peak prominence	Heman-Ackah *et al.* (2002), Halberstam (2004), and Eadie and Baylor (2006)
Smoothed pitch perturbation quotient	De Bodt (1997), Wolfe *et al.* (1997), Heman-Ackah *et al.* (2002), and Halberstam (2004)
Soft phonation index	De Bodt (1997)
Spectral distortion	Feijoo and Hernández (1990)
Spectral flatness of inverse filter	Prosek *et al.* (1987)
Spectral flatness of residue signal	Prosek *et al.* (1987) and Eadie and Doyle (2005)
Spectral noise level above and under 6000 Hz	Dejonckere and Wieneke (1996)
Spectral tilt	Eadie and Doyle (2005)
Spectral tilt of voiced segments	Eadie and Doyle (2005)
Standard deviation of cepstral peak prominence	Wolfe and Martin (1997)
Standard deviation of fundamental frequency	Wolfe *et al.* (1997)
Standard deviation of jitter	Kreiman *et al.* (1990) and Wolfe and Martin (1997)
Standard deviation of partial period comparison	Kreiman *et al.* (1990)
Standard deviation of period	Wolfe *et al.* (2000)
Standard deviation of relative frequency differences	Askenfelt and Hammarberg (1986)
Standard deviation of shimmer	Kreiman *et al.* (1990) and Wolfe and Martin (1997)
Standard deviation of signal-to-noise ratio from Milenkovic	Wolfe and Martin (1997)
Voice turbulence index	Halberstam (2004)

In the first subset there were 52 of the 69 acoustic measures on sustained vowels with a (weighted) correlation coefficient below 0.60. Weighted correlation coefficients ranged from 0.11 for coefficient of excess and voice turbulence index to 0.56 for amplitude perturbation quotient of residuals and harmonic-to-noise ratio from Yumoto. In this subset, there were 32 markers with a k-value of 2 or more. The SD_{res} statistics indicated heterogeneity for eight measures. For the remaining 24 acoustic correlates with $\overline{r_w} < 0.60$ and $k \geq 2$, SD_{res} statistics showed homogeneity. The second subset consisted of 17 acoustic measures with a (weighted) effect size equal to or above 0.60. In this subset of 17 measures, there were 4 markers with a k-value of 2 or more. Weighted correlation coefficients ranged from 0.62 for smoothed cepstral peak prominence to 0.75 for pitch amplitude. Statistical homogeneity testing (SD_{res}) indicated that these four $\overline{r_w}$-values were based on a set of homogeneous effect sizes, indicating that these effect sizes are consistently equal to or above 0.60 (smoothed cepstral peak prominence: $\overline{r_w}=0.62$, spectral flatness of residue signal: $\overline{r_w}=0.69$, Pearson r at autocorrelation peak: $\overline{r_w}=0.74$, and pitch amplitude: $\overline{r_w}=0.75$).

B. Continuous speech

Seven studies using continuous speech samples met the inclusion criteria of this meta-analysis. These studies were published in *Journal of Voice* (4), *Journal of Speech (Language) and Hearing Research* (1), *Journal of the Acoustical Society of America* (1), and *ORL* (1). As shown in Table V, there were 29 separate effect sizes pertaining to 26 distinct acoustic measures. Relevant information regarding the methodology of these seven reports is found in Table II. Whereas all seven studies used pathologic or dysphonic voice

TABLE IV. Summary of the meta-analytic findings for the individual acoustic measures of overall voice quality in sustained vowels. The acoustic measures are ordered according to their effect size (r or $\overline{r_w}$).

Acoustic measure	k [a]	r or $\overline{r_w}$ [b]	SD_{res} [c]
Fluctuation in fundamental frequency	1	0.00	/
Soft phonation index	1	0.01	/
Standard deviation of signal-to-noise ratio from Milenkovic	1	0.06	/
Frequency of second formant	1	0.07	/
Standard deviation of cepstral peak prominence	1	0.11	/
Voice turbulence index	2	0.11	He
Coefficient of excess	2	0.11	Ho
Frequency of third formant	1	0.14	/
Ratio of amplitudes of first and second harmonics	1	0.15	/
No. of harmonics	1	0.19	/
Fluctuation in amplitude	1	0.19	/
Richness of high frequency harmonics	1	0.19	/
Frequency of first formant	1	0.21	/
Standard deviation of percent jitter	2	0.22	Ho
Breathiness index	3	0.22	Ho
Spectral flatness of inverse filter	2	0.25	Ho
Fundamental frequency	3	0.28	Ho
Coefficient of variation of percent shimmer	1	0.28	/
Ratio of frequencies of second and first formants	1	0.32	/
Difference between frequencies of second and first formants	1	0.33	/
Coefficient of variation of amplitude	1	0.34	/
Standard deviation of period	2	0.37	Ho
Signal-to-noise ratio	3	0.38	He
Smoothed amplitude perturbation quotient	3	0.40	Ho
Residue signal power ratio	1	0.40	/
Relative noise level	1	0.40	/
Standard deviation of percent shimmer	2	0.41	Ho
Signal-to-noise ratio above 1000 Hz	3	0.42	He
Jitter factor	3	0.42	Ho
Power spectrum ratio	2	0.44	Ho
Amplitude perturbation quotient	6	0.45	Ho
Noise-to-harmonics ratio from MDVP	7	0.45	He
Shimmer in decibel	4	0.45	Ho
Absolute jitter	4	0.47	Ho
Standard deviation of fundamental frequency	2	0.47	Ho
Pitch perturbation quotient of residue signal	2	0.47	Ho
Coefficient of variation of percent jitter	1	0.48	/
Coefficient of variation of fundamental frequency	2	0.49	Ho
Percent jitter	6	0.49	Ho
Cepstral peak prominence	4	0.50	Ho
Natural logarithm of standard deviation of period	2	0.51	He
Percent shimmer	6	0.52	He
Spectral noise level above and under 6000 Hz	1	0.52	/
Relative average perturbation	5	0.52	Ho
Pitch perturbation quotient	3	0.52	Ho
Smoothed pitch perturbation quotient	4	0.53	Ho
Jitter ratio	2	0.53	Ho
Lyapunov coefficient	3	0.54	He
Phonatory fundamental frequency range	5	0.54	Ho
Harmonics-to-noise ratio from Yumoto	3	0.56	He
Amplitude perturbation quotient of residue signal	2	0.56	Ho
Mean harmonic emergence between 500 and 1500 Hz	1	0.58	/
Coefficient of variation of period	1	0.62	/
Smoothed cepstral peak prominence	3	0.63	Ho
Standard deviation of partial period comparison	1	0.67	/
Spectral flatness of residue signal	2	0.69	Ho
Partial period comparison	1	0.69	/
Jitter from Yumoto	1	0.71	/
Pearson r at autocorrelation peak	2	0.74	Ho
Normalized mean absolute period jitter	1	0.75	/
Pitch amplitude	3	0.75	Ho
Signal-to-noise ratio from Milenkovic	1	0.76	/
Cepstral peak magnitude	1	0.80	/
Cycle-to-cycle variation of waveform	1	0.83	/
Harmonics-to-noise ratio from Kojima	1	0.87	/
Normalized noise energy	1	0.88	/
Cepstrum of excitation signal	1	0.90	/
Spectral distortion	1	0.93	/
Normalized mean absolute period shimmer	1	0.93	/

[a] k=number of effect sizes available in the included literature.
[b] r=correlation coefficient (when $k=1$) and $\overline{r_w}$=mean weighted correlation coefficient (when $k>1$).
[c] SD_{res}=residual standard deviation, /=not applicable (when $k=1$), and Ho/He=homogeneous/heterogeneous r or $\overline{r_w}$ (when $k>1$).

samples, only three studies also investigated normal voices. The mean number of dysphonic voice samples was 50 (range 9–112). For the normal voices, the mean number was 7 (range 3–41). The mean number of subjects was 57 (range 12–153). All acoustic measures were extracted from recordings of continuous speech, most often from speakers reading from a text. With the exception of Askenfelt and Hammarberg (1986) and Ma and Yiu (2006), the so-called "rainbow passage" was read aloud and a portion (typically the second sentence) was extracted for further analysis.

As for auditory-perceptual evaluation, the mean number of judges employed across studies was 7 (range 2–16). In four studies the rating scale was an equal-appearing interval scale with 4, 6, 7, or 11 points. In two studies (Qi *et al.*, 1999; Eadie and Baylor, 2006), a visual analog scale was used. In another study (Eadie and Doyle, 2005), direct magnitude estimation was used. The following labels were used to designate the perceptual construct that was to be evaluated: hoarseness, G (for grade), overall severity, and overall voice quality. Estimates of reliability of listener judgments included two types of statistics: Pearson's product-moment correlation coefficient and Cronbach's coefficient alpha. To evaluate intrajudge reliability, Pearson's r-values were uniformly reported. Where a range of r-values was given (Askenfelt and Hammarberg, 1986; Qi *et al.*, 1999; Eadie and Baylor, 2006), the lowest r-value was chosen to calculate a weighted average of intrajudge correlation across reports (Table II). The intrajudge $\overline{r_w}$ was 0.81, which is indicative of homogeneous intrajudge reliability. It appears that listeners were generally consistent in their perceptual evaluations of continuous speech. Regarding interjudge reliability, only three studies provided a Pearson's r-value. Meta-analysis, again using the lowest r-value of the reported range, resulted in an interjudge $\overline{r_w}$ of 0.84, i.e., homogeneous interjudge reliability. This was corroborated by the three studies that used Cronbach's α, since they all mentioned an α-value of 0.97. Concerning the acoustic measures, Table V provides an overview of the determinants that were used to

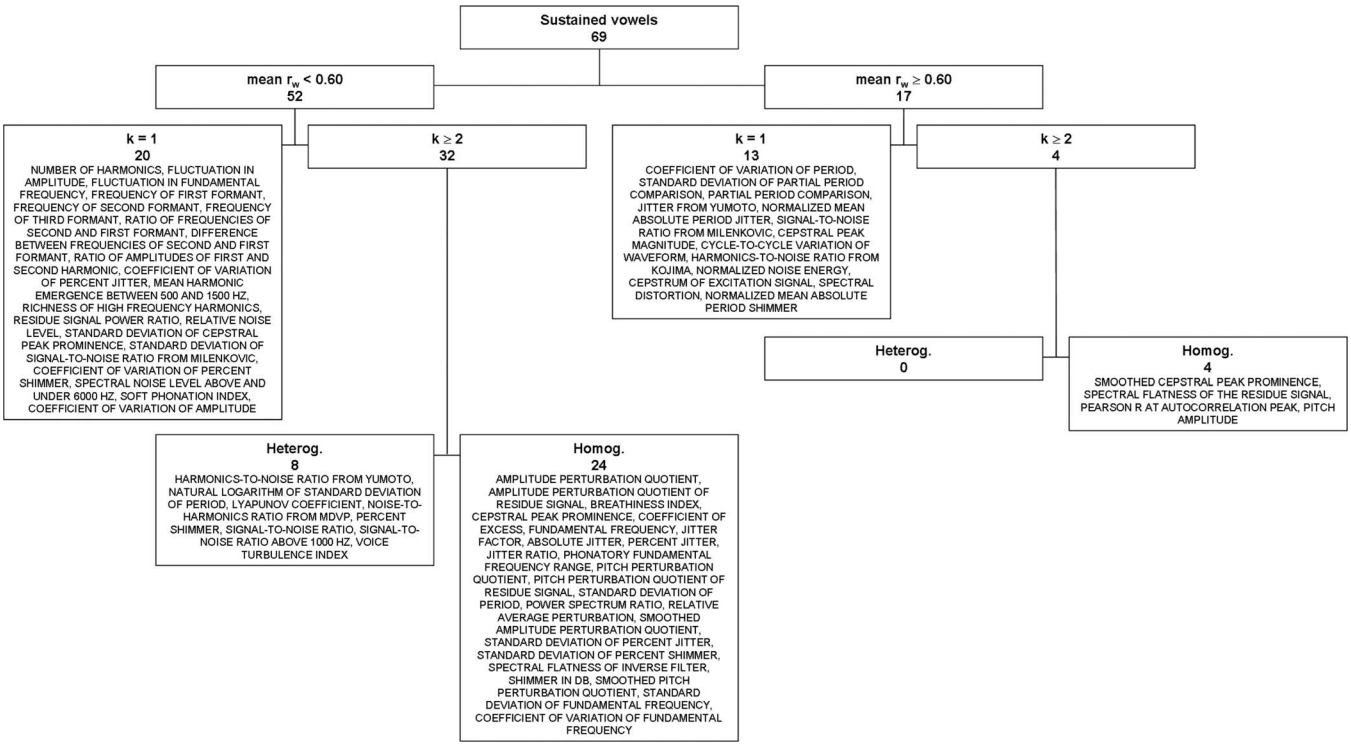

FIG. 1. Diagram illustrating the organization of the meta-analysis for acoustic measures on sustained vowels. The second line in every box contains the number of acoustic measures.

gauge overall voice quality. As was the case for sustained vowel studies, there were considerable differences between studies' recording equipment and settings.

1. Meta-analysis on correlation coefficients

The results of the meta-analysis on continuous speech data are summarized in Table V and Fig. 2. For 23 of the 26 (88%) acoustic measures cited, there was only 1 effect size available. For the remaining three acoustics determinants (cepstral peak prominence, smoothed cepstral peak prominence, and signal-to-noise ratio from Qi) there were two effect sizes ($k=2$). For these three markers, a $\overline{r_w}$ was calculated with Meta-Analysis Programs version 5.3. The organization of the meta-analysis on acoustic measures on continuous speech is illustrated in Fig. 2.

As in our meta-analysis on sustained vowel data, a correlation coefficient of 0.60 was chosen as the threshold to distinguish between marked and weak acoustic measures. In the first subset of 16 acoustic measures with a (weighted) effect size below 0.60, k was always equal to 1, and therefore no meta-analysis was performed. In the second subset consisting of ten acoustic measures with a (weighted) effect size equal to or above 0.60, there were three markers with $k=2$: signal-to-noise ratio from Qi, cepstral peak prominence, and smoothed cepstral peak prominence. Meta-analysis for these three measures yielded $\overline{r_w}$-values of 0.69, 0.88, and 0.88, respectively. Furthermore, SD_{res} indicated that these three $\overline{r_w}$-values were based on a set of homogeneous effect sizes.

IV. DISCUSSION

The present meta-analysis assessed the relationship between acoustic measures and perceptual judgments of overall voice quality. In Buder (2000) alone, more than 100 acoustic algorithms were cited and numerous microcomputer-based software systems offering various acoustic voice quality parameters have been developed. The fact that correlations between perception of overall voice quality and acoustic measures vary substantially (Kreiman and Gerratt, 2000a) raises questions regarding the validity and usefulness of these acoustic determinants. This meta-analysis represented an attempt to synthesize the corpus of algorithms and measures, and to establish a hierarchy of acoustic markers on a statistical basis. In total, 25 study reports were included. 21 studies reported on 150 correlation coefficients for 69 acoustic measures on sustained vowels. 7 studies identified 29 correlation coefficients for 26 acoustic measures on continuous speech.

In the context of the present meta-analysis, a homogeneous $\overline{r_w}$ exceeding 0.60 was judged to be a critical index. For instance, the amplitude perturbation quotient measure on sustained vowels was cited in five studies with 0.41, 0.54, 0.63, 0.50, 0.41, and 0.71 as coefficients of correlation. The single $r=0.71$ value, in particular, (Halberstam, 2004) seems to identify amplitude perturbation quotient as a valid acoustic marker for overall voice quality of sustained vowels. However, the r-values from the other studies are less persuasive and the meta-analysis resulted in a smaller homogeneous $\overline{r_w}$ of 0.45. In contrast to the amplitude perturbation quotient example wherein the meta-analysis resulted in a relatively weak $\overline{r_w}$ of 0.45, the meta-analysis outcome of

TABLE V. Summary of the meta-analytic findings for the individual acoustic measures of overall voice quality in continuous speech. The acoustic measures are ordered according to their effect size (r or $\overline{r_w}$).

Acoustic measure	k [a]	r or $\overline{r_w}$ [b]	SD_{res} [c]
Perturbation magnitude	1	0.01	/
Maximum intensity in voice range profile	1	0.02	/
Lowest fundamental frequency in voice range profile	1	0.09	/
Noise-to-harmonics ratio from MDVP	1	0.13	/
Fundamental frequency	1	0.18	/
Frequency-domain harmonics-to-noise ratio	1	0.26	/
Spectral flatness of residue signal	1	0.26	/
Spectral tilt of voiced segments	1	0.33	/
Highest fundamental frequency in voice range profile	1	0.34	/
Intensity range in voice range profile	1	0.35	/
Fundamental frequency range in voice range profile	1	0.37	/
Minimum intensity in voice range profile	1	0.38	/
Area of voice range profile	1	0.43	/
Spectral tilt	1	0.47	/
Pitch amplitude	1	0.58	/
Perturbation magnitude mean	1	0.59	/
Perturbation factor	1	0.62	/
Percent shimmer	1	0.62	/
Signal-to-noise ratio from Qi	2	0.69	He
Directional perturbation factor	1	0.71	/
Standard deviation of relative frequency differences	1	0.71	/
Compression of relative frequency differences	1	0.73	/
Peakedness of relative frequency differences	1	0.73	/
Relative average perturbation	1	0.75	/
Cepstral peak prominence	2	0.88	Ho
Smoothed cepstral peak prominence	2	0.88	Ho

[a] k=number of effect sizes available in the included literature.
[b] r=correlation coefficient (when $k=1$) and $\overline{r_w}$=mean weighted correlation coefficient (when $k>1$).
[c] SD_{res}=residual standard deviation, /=not applicable (when $k=1$), and Ho/He=homogeneous/heterogeneous r or $\overline{r_w}$ (when $k>1$).

studies related to smoothed cepstral peak prominence seems to suggest a much stronger association. For instance, although Halberstam's (2004) r-value of 0.55 for smoothed cepstral peak prominence does not provide strong support for smoothed cepstral peak prominence as a valid measure of overall voice quality; combining this result with the Heman-Ackah *et al.* (2002) and the Eadie and Baylor (2006) results of $r=0.80$ and $r=0.82$, respectively, the final $\overline{r_w}$ is 0.63, which supports smoothed cepstral peak prominence as a promising acoustic marker of overall voice quality. Based on the meta-analysis of sustained vowel studies, four measures satisfied the requirement of a homogeneous $\overline{r_w} \geq 0.60$: (1) Pearson r at autocorrelation peak, (2) pitch amplitude, (3) spectral flatness of residue signal, and (4) smoothed cepstral peak prominence. For continuous speech, three measures satisfied the criterion: (1) signal-to-noise ratio from Qi, (2) cepstral peak prominence, and (3) smoothed cepstral peak prominence. Consequently, these six measures are considered to be the most promising measures for the acoustic measurement of overall voice quality, as compared to the remaining 81 measures included in the original meta-analysis. The results of these six measures will be discussed in the next paragraphs.

The first of these six measures is Pearson r at autocorrelation peak. To obtain this measure, correlations are calculated between the voice signal and delayed versions of the same signal (i.e., autocorrelation) at time lags between the minimally and maximally expected fundamental periods. The Pearson moment-product correlation coefficient is computed at the highest peak of this autocorrelation function (i.e., the correlogram with "delay" or "time lag" on the abscissa and "correlation" on the ordinate). The rationale behind this measure is that more periodic voice signals display more prominent autocorrelation peaks, and vice versa. A perfectly periodic signal reveals a Pearson r at autocorrelation peak of 1.0, and the more the signal deviates from perfect periodicity, the more this correlation decreases (Hillenbrand and Houde, 1996). The correlation between overall voice quality and this measure of the autocorrelation function on the sound waveform has been investigated by Wolfe *et al.* (2000) in both male and female voices separately, yielding $k=2$ and $\overline{r_w}=0.74$. This result requires confirmation by other independent investigators to permit generalization. Although Hillenbrand and Houde (1996) indicated a correlation of 0.84 between breathiness ratings and Pearson r at autocorrelation peak for both sustained vowels and continuous speech, and

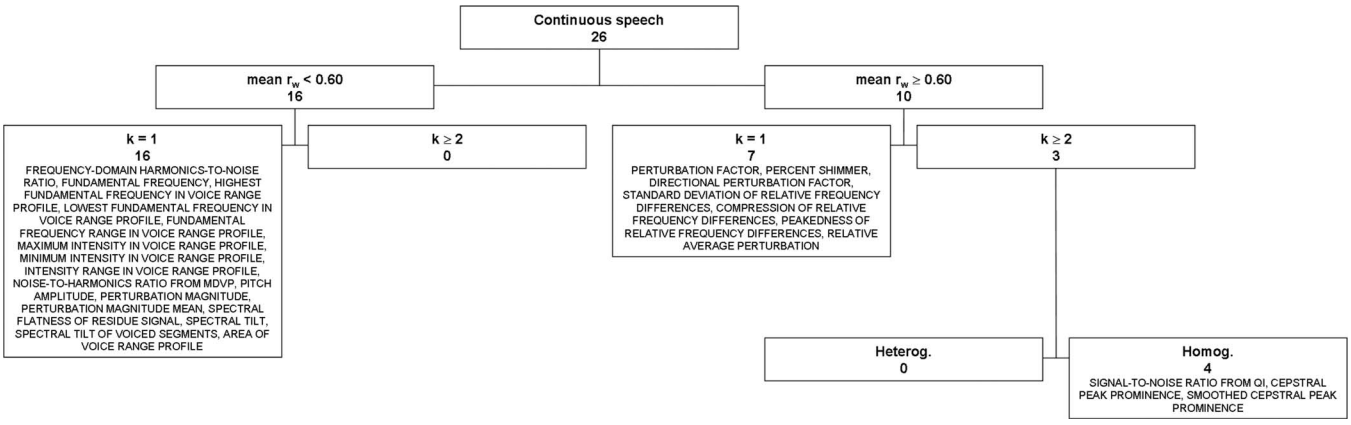

FIG. 2. Diagram illustrating the organization of the meta-analysis for acoustic measures on continuous speech. The second line in every box contains the number of acoustic measures.

concluded that Pearson r at autocorrelation peak is an accurate marker of breathiness, further corroboration of its concurrent validity is also needed.

The second measure is pitch amplitude. To acquire this measure, the radiated voice signal is first inverse filtered via a linear predictive coding algorithm. The result of this inverse filtering is a residue signal, i.e., a series of impulses theoretically showing the moment of vocal tract acoustic excitation provided by glottal closure (permitting investigation of the signal provided by the laryngeal source instead of the entire vocal tract). Second, the autocorrelation function of this residue signal is calculated. Pitch amplitude is the amplitude of the maximum correlation (i.e., traditionally corresponding with the pitch) in the correlogram and consequently is considered to be a measure of the strength of voice periodicity (Prosek *et al.*, 1987). Plant *et al.* (1997) investigated the correlation between this measure and overall voice quality, and Prosek *et al.* (1987) used pitch amplitude as a marker of both disorder severity and hoarseness in sustained vowels. These two independent studies resulted in $k=3$ and $\overline{r_w}=0.75$. Although Eadie and Doyle (2005) reported an r-value for pitch amplitude of only 0.58 when applied on continuous speech, further support for the value of measures based on inverse filtering is provided by Parsa and Jamieson (2001), who concluded that such measures are superior to perturbation measures for both continuous speech and sustained vowels. In part, Parsa and Jamieson (2001) arrived at their conclusion based on measures of diagnostic accuracy, which included the area under the ROC curve. In the case for pitch amplitude, the area under the ROC curve for sustained vowels was 0.977 (perfect diagnostic accuracy=1.00), and the rate of correct classification between normal and pathologic voices was 93.0%. For continuous speech there was an area under ROC curve of 0.953 and a correct classification rate of 88.9%. Parsa and Jamieson (2001) stated that pitch amplitude provided the best classification accuracy among all measures extracted from continuous speech samples.

Based on the results of the meta-analysis, the third acoustic measure, which produced respectable raw correlation results, was spectral flatness of residue signal. This measure also generates the residue signal as an output of the inverse filter. The spectrum is then derived from the residue signal, and finally the distribution of the frequencies in the spectrum is computed. The flatter the spectral distribution of the residue signal, the more the harmonics are considered to be masked by noise (Prosek *et al.*, 1987). Prosek *et al.* (1987) correlated spectral flatness of residue signal with both disorder severity and hoarseness separately ($k=2$) in sustained vowels. Our meta-analysis of these two correlation coefficients (derived from one study) yielded $\overline{r_w}=0.69$. Although there is no confirmation from other independently generated correlations, Parsa and Jamieson (2001), who used discriminant analyses, supported the utility of spectral flatness of residue signal as a valid discriminator between normal and pathological voices. For sustained vowels, spectral flatness of residue signal showed the largest area under ROC curve (0.996) and had the highest classification accuracy (96.5% correct). For continuous speech, the area under the ROC and the discrimination accuracy were 0.928% and 85.8%, respectively. Furthermore, Parsa and Jamieson (2001) concluded that more commonly used measures (as jitter, shimmer, and noise-to-harmonics ratio) did not perform as well as measures based on linear prediction modeling and inverse filtering.

The concurrent validity of the fourth measure, signal-to-noise ratio from Qi, was investigated in continuous speech only. This measure uses linear predictive coding and inverse filtering for the decomposition of speech samples into signal (i.e., waveform of the original signal) and noise (i.e., waveform of the signal with a typically random Gaussian distribution after removal of resonance-based and voice-based patterns). The ratio between the average root-mean-square amplitudes of the signal and the noise components can then be computed to quantify the acoustic properties of disordered voices (Qi *et al.*, 1999). Combining the independent results of Qi *et al.* (1999) and Eadie and Doyle (2005) leads to a homogeneous $\overline{r_w}=0.69$ ($k=2$), which is promising. This measure was also examined in the studies of Parsa and Jamieson (2001). Although not as robust as pitch amplitude and spectral flatness of residue signal, signal-to-noise ratio from Qi demonstrated acceptable diagnostic precision (distinguishing normophonic from dysphonic individuals) in both sustained vowels (area under ROC curve: 0.945; classification rate: 81.6%) and continuous speech (area under ROC curve: 0.903; classification rate: 79.6%). In summary, measures and algorithms based on inverse filtering and linear prediction modeling appear to be very promising and useful in clinical settings, where patients present with heterogeneous voice qualities and severities.

The meta-analysis outcome for the fifth and sixth measures, the cepstral markers of cepstral peak prominence and smoothed cepstral peak prominence, can be summarized as follows. To obtain these two measures, one constructs a cepstrum (i.e., a log power spectrum of a log power spectrum, resulting in a graph with "quefrency" on the abscissa and "cepstral magnitude" on the ordinate). The highest cepstral peak is identified between the minimally and maximally expected fundamental period, and a linear regression line is drawn, which relates quefrency to cepstral magnitude. The difference in amplitude between this cepstral peak and the corresponding value on the linear regression line exactly below the peak determines the cepstral peak prominence. Averaging (i.e., smoothing) of the cepstrum across time and across quefrency results in a smoothed cepstrum, and the difference between the highest peak and the corresponding value on the regression line in the smoothed cepstrum is called the smoothed cepstral peak prominence. The rationale behind these measures is that the more periodic a voice signal is, the more it displays a well-defined harmonic configuration in the spectrum, and, consequently, the more the cepstral peak will be prominent (Hillenbrand and Houde, 1996). For cepstral peak prominence on sustained vowels, meta-analysis of the Wolfe and Martin (1997), the Wolfe *et al.* (2000), and the Halberstam (2004) results yields a homogeneous $\overline{r_w}$ of 0.50 ($k=3$). On continuous speech, however, meta-analysis on the findings of Halberstam (2004) and Eadie and Baylor (2006) results in $\overline{r_w}=0.88$ ($k=2$). Heman-Ackah *et al.* (2002), Halberstam (2004), and Eadie and

Doyle (2005) investigated smoothed cepstral peak prominence applied on sustained vowels. Meta-analysis of these three independent studies results in $\overline{r_w}=0.63$ ($k=3$). Furthermore, Heman-Ackah *et al.* (2002) and Halberstam (2004) also provided correlation coefficients for continuous speech, which results in a homogeneous $\overline{r_w}=0.88$ ($k=2$) after meta-analysis. In summary, on the basis of this meta-analysis, the two cepstral measures, and smoothed cepstral peak prominence, in particular, can be viewed as potentially the most accurate acoustic algorithms or single correlates of overall voice quality. Additional evidence for the validity of cepstral measures can be found in Hillenbrand and Houde (1996), who found that cepstral peak prominence was among the most robust correlates of breathiness in sustained vowels as well as in continuous speech. Other studies confirming this conclusion were conducted by de Krom (1993), who stated that cepstrum-based harmonics-to-noise ratio was a strong marker of both roughness and breathiness in sustained vowels. Dejonckere and Wieneke (1996) found a correlation of 0.80 between overall severity of hoarseness and the amplitude of highest rahmonic (also known as cepstral peak magnitude). The magnitude of this correlation far exceeded the correlation of the other acoustic measures in their study (jitter ratio, relative noise level above 6 kHz, and mean harmonic emergence between 0.5 and 1.5 kHz). In a later study using factor analysis, Dejonckere (1998) reported that cepstral peak magnitude is negatively affected by irregularity in vocal fold vibration as well as by excessive glottal air leakage, bolstering the assertion that cepstral peak magnitude is sensitive to aspects that potentially contribute to overall dysphonia severity. Heman-Ackah *et al.* (2003) investigated the diagnostic validity of smoothed cepstral peak prominence on sustained vowels and continuous speech and of amplitude perturbation quotient, percent jitter, noise-to-harmonics ratio from MDVP, relative average perturbation, and smoothed pitch perturbation quotient on sustained vowels only. They concluded that the smoothed cepstral peak prominence measures are good correlates of dysphonia and that, on average, smoothed cepstral peak prominence on continuous speech performed better on measures of diagnostic precision such as sensitivity, specificity, positive predictive value, and negative predictive value, as compared to traditional time-based measures of perturbation. They concluded that smoothed cepstral peak prominence "are reliable measures that should become routine in objective voice analysis" (p. 332). Finally, Awan and Roy (2006) conducted a study in which they used a cepstral measure they called expected cepstral peak prominence in a multiple regression procedure. This measure actually is the ratio of the cepstral peak prominence to the expected amplitude of the cepstral peak based on linear regression. It is very similar to the cepstral peak prominence measure described by Hillenbrand and Houde (1996). Awan and Roy (2006) indicated that expected cepstral peak prominence "may be the most significant component" (p. 44) contributing to their four-factor model for measuring dysphonia severity. Collectively, measures derived from the cepstrum (such as cepstral peak prominence and smoothed cepstral peak prominence) can be used in sustained vowel as well as continuous speech samples because they do not rely on accurate fundamental period detection (Hillenbrand and Houde, 1996; Heman-Ackah *et al.*, 2003), and they can be easily implemented in clinical settings.

From the discussion above it is apparent that four of these six most promising acoustic measures share one interesting feature, i.e., they all measure what might be called "periodicity prominence." Cepstral peak prominence and smoothed cepstral peak prominence are two very similar measures initially introduced for pitch detection in speech signals via the cepstral method, and pitch amplitude and Pearson *r* at autocorrelation peak are two similar measures frequently applied for pitch determination via the autocorrelation function. Both the quefrency of the first rahmonic and the lag time of the highest autocorrelation peak within a specific analysis window correspond with the fundamental frequency; and, the height of these cepstral and autocorrelation peaks is related to the prominence of the fundamental frequency (i.e., periodicity) in the voice signal (Hillenbrand *et al.*, 1994). Based on the outcome of the present meta-analysis, it thus can be assumed that overall voice quality and dysphonia severity are mainly determined by periodicity dominance, and that factors attenuating periodicity of the voice signal also contribute to the perception of increased dysphonia. Furthermore, it is important to note that there are many other measures of periodicity prominence available to clinicians, such as the ubiquitous pitch and amplitude perturbation measures often generated automatically by most commercially available signal processing software. However, the important advantage of these four measures over the commonly used voice perturbation measures resides in their methods to assess periodicity, namely, they do not demand cycle boundary identification for fundamental frequency detection in the time-domain.

In addition to these six $k>1$ measures with $\overline{r_w}\geq 0.60$, there were many $k=1$ measures. However, because a high correlation in one study can be offset by a low correlation in another study (and vice versa), caution is warranted when interpreting the outcome of a solitary *r*, and this applies to all acoustic measures with $k=1$ (e.g., $r=0.93$ for normalized mean absolute period shimmer on sustained vowels or $r=0.26$ for frequency-domain harmonics-to-noise ratio on continuous speech). Without further confirmation, replication, or evidence rejecting the presented *r*-values, it is difficult to place these results in context, and impossible to draw firm conclusions regarding these $k=1$ measures at this point in time.

Interestingly, the measures that were investigated most often (with $k\geq 5$) were the perturbation measures (amplitude perturbation quotient, percent jitter, percent shimmer, and relative average perturbation), the noise measure noise-to-harmonics ratio from MDVP, and voice range profile measure phonatory fundamental frequency range (see Table V). Except for the phonatory fundamental frequency range, these time-domain perturbation measures also appeared to be most frequently used in voice clinics (De Bodt, 1997). Percent jitter is a measure of fundamental frequency or period perturbation. It measures the mean difference in fundamental frequency of adjacent periods relative to the mean fundamental frequency of all periods in the voice recording (Buder,

2000). Relative average perturbation is also a measure of fundamental frequency perturbation. This measure is similar to percent jitter, but uses a moving three-point smoothing and normalization to average the period before computing the mean deviation in period relative to the mean period of all periods (Buder, 2000). Percent shimmer, another measure similar to percent jitter, measures amplitude perturbations by computing the mean deviation in amplitude between adjacent cycles relative to the mean amplitude of all cycles (Buder, 2000). Amplitude perturbation quotient is another amplitude perturbation measure, but instead of working with adjacent cycles as percent shimmer, it first averages the amplitude of a moving number (i.e., an odd integer greater than 1) of successive cycles before calculating the mean deviation in amplitude between cycle groups relative to the mean amplitude of all cycles (Buder, 2000). These perturbation measures are traditionally linked to the measurement of irregular voice fold vibrations. The noise-to-harmonics ratio from MDVP is a spectral measure that computes the ratio of the between-harmonic spectral magnitudes in the range from 1500 to 4500 Hz to the harmonic spectral magnitudes in the range from 70 to 4500 Hz (Buder, 2000). This measure is classically associated with measurements of additive noise at the level of the glottis. The phonatory fundamental frequency range in the voice range profile is one of the measures of the dispersion of the fundamental frequency and consists of subtracting the lowest from the highest possible fundamental frequency (Buder, 2000). According to De Bodt (1997), who reviewed literature between 1991 and 1995, these are the most frequently mentioned measures in voice literature (except for F_0 and amplitude measures). Yet, on sustained vowels, these measures did not yield $\overline{r_w} \geq 0.60$. Regarding jitter, meta-analysis yielded homogeneous $\overline{r_w}$ of 0.47, 0.49, 0.52, and 0.52 for absolute jitter, percent jitter, relative average perturbation, and pitch perturbation quotient, respectively. Absolute jitter is the mean of the differences between the period and the fundamental frequency of adjacent cycles (Buder, 2000). Pitch perturbation quotient is the same as relative average perturbation, but with a smoothing factor of 5 cycles (Buder, 2000). Similarly, the meta-analysis for shimmer resulted in a homogeneous $\overline{r_w}$ of 0.45 for shimmer in decibel and amplitude perturbation quotient, and a heterogeneous $\overline{r_w}$ of 0.52 for percent shimmer. Regarding noise-to-harmonics ratio from MDVP, the measure most frequently encountered, a heterogeneous $\overline{r_w}$ of 0.45 was found. On continuous speech, there was a solitary correlation of 0.62 and 0.75 for percent shimmer and relative average perturbation, and 0.37 and 0.13 for phonatory fundamental frequency range and noise-to-harmonics ratio from MDVP, respectively. Measures related to the voice range profile yielded a $\overline{r_w}$ of maximally 0.43. In general, the results of this meta-analysis confirm the apparent inferiority of perturbation measures as compared to other measures that do not depend on accurate identification of cycle boundaries. This conclusion supports the findings of Parsa and Jamieson (2001), and is confirmed by Kreiman and Gerratt (2005), who concluded that "the associations between jitter, shimmer, and perceived voice quality are not sufficiently explanatory to justify continued reliance on jitter and shimmer as indices of voice quality" (p. 2209). As mentioned previously, F_0 and amplitude perturbation measures are especially susceptible for the influence of type of microphone and microphone location relative to the sound source, type of hardware, processing algorithms, measurement algorithms, and software settings such as sampling rate and fundamental period extraction. Furthermore, F_0 and amplitude perturbation measures are not sensitive to differences in glottal waveform shape and additive glottal noise, and appear only reliable in nearly periodic voice signals (Titze, 1995; Parsa and Jamieson, 2001).

This meta-analysis, combined with previous studies, seems to confirm that measures that do not rely on the extraction of the fundamental period in their calculation such as smoothed cepstral peak prominence, Pearson r at autocorrelation peak, and pitch amplitude produce stronger relationships with perceptual judgments of overall severity of dysphonia in sustained vowels as well as continuous speech, and deserve further attention in clinical circles (Hillenbrand and Houde, 1996; Parsa and Jamieson, 2001).

A. Caveats and limitations

There are several limitations regarding the present meta-analysis that not only restrict the generalizability of the findings, but also provide a direction for future research. It is important to acknowledge that current acoustic measures might not be sensitive measures of perceived voice quality because of limitations of their algorithms and the theoretical models on which they are based. First, this meta-analysis concentrated on the relationship between acoustic markers and overall voice quality. Additional meta-analytic research is needed to address the relationship between acoustic measures and specific vocal quality attributes, such as breathiness and roughness. Meta-analytic techniques may improve the resolution regarding which acoustic measures best track these specific voice qualities.

Second, the present meta-analysis was restricted to reports and findings based on correlation coefficients. Beyond the 69 acoustic measures on sustained vowels and the 26 measures on continuous speech, other acoustic measures have been discussed in literature. But because correlation coefficients were not available for these measures, the value of these markers in voice quality measurement and their relative validity in comparison to the aforementioned markers remains unclear. Future meta-analyses should potentially explore other effect size measures, aside from the correlation coefficient, to investigate the validity of these acoustic markers.

Third, overall voice quality can be investigated with measures other than acoustic measures. For example, certain aerodynamic measures could also be worth exploring within this context, and thus meta-analysis investigating the association between aerodynamic measures and perceptual voice quality measurement is recommended.

Fourth, the interpretation of the findings of the present meta-analysis is complicated by variability related to different data acquisition systems. While the influence of factors such as microphone type and placement, environmental noise, software, etc., on the outcomes of perturbation mea-

sures has already been investigated, the impact of these factors on other measures such as cepstral peak prominence and pitch amplitude remains unclear. Additional exploration of the impact of data acquisition systems and environments on the outcomes of these measures is warranted.

Fifth, the relationship between the auditory-perceptual rating and the acoustic measurement of overall voice quality relies greatly on the rationale and algorithm underlying the acoustic measure. However, as previously discussed, unreliability of listener ratings introduces perceptual noise and consequently tends to handicap the acoustic (or other) measurement of voice quality. While suggestions to improve rater reliability exist (e.g., Kreiman and Gerratt, 2000b; Eadie and Doyle, 2002; Bele, 2005; Eadie and Baylor, 2006; Yiu *et al.*, 2007; Kreiman *et al.*, 2007) few studies have estimated the true (absolute) impact of listener unreliability on the correlation between perception and acoustic measures. Furthermore, there is no universal standard distinguishing an acceptable from an unacceptable reliability estimate. Future research should address the criteria used to determine what precisely constitutes an acceptable level of listener reliability, and the impact of such criteria on the validation of acoustic voice quality measures.

Sixth, an important caveat is related to the low number of correlation coefficients (i.e., the k statistic) available for many of the acoustic measures in this meta-analysis. Because of $k=1$, no weighted average correlation coefficient could be calculated for 33 of the measures on sustained vowels (47.8%) and for 23 of the measures on continuous speech (88.5%). However, as long as the extant literature lacks corroborating evidence from multiple, independent correlational studies, no firm conclusions can be made regarding this very diverse and idiosyncratic set of $k=1$ acoustic measures. Although impressive correlations with dysphonia severity have been reported for some of these $k=1$ measures in sustained vowels (e.g., $r=0.88$ for normalized noise energy) as well as continuous speech (e.g., $r=0.75$ for relative average perturbation), equally poor correlations have been reported for others, such as $r=0.01$ for soft phonation index for sustained vowels, and $r=0.01$ for perturbation magnitude for connected speech. There were also several $k>1$ measures with restricted interpretative value, because although multiple r-values were reported, they were derived from the same study report. For instance, the two correlations on which the meta-analyses on Pearson r at autocorrelation peak (Wolfe *et al.*, 2000) and spectral flatness of residue signal (Prosek *et al.*, 1987) were based actually originated from the single studies of Wolfe *et al.* (2000) and Prosek *et al.* (1987), respectively. Furthermore, two of the three effect sizes in the meta-analysis on pitch amplitude also originated from the single study of Prosek *et al.* (1987). Because these are not replications *per se*, the generalizability of the meta-analytic evidence for these three measures is also limited. Therefore, like the $k=1$ scenario, further research/replication is also needed to corroborate the performance of these three acoustic measures.

V. CONCLUSIONS

The above-stated limitations notwithstanding, measures for which the meta-analysis resulted in a homogeneous $\overline{r_w}$ of at least 0.60, are Pearson r at autocorrelation peak, pitch amplitude, spectral flatness of residue signal, and smoothed cepstral peak prominence on sustained vowels; and signal-to-noise ratio from Qi, cepstral peak prominence, and smoothed cepstral peak prominence on continuous speech. However, only the smoothed cepstral peak prominence withstanded all criteria demanded by this meta-analytic approach: multiple r-values, derived from multiple study reports, leading to homogeneous $\overline{r_w} \geq 0.60$ in both sustained vowels and continuous speech. This cepstral metric thus can be regarded as the most promising and perhaps robust acoustic measure of dysphonia severity. Tables IV and V present a hierarchy of the numerous outcomes of acoustic markers measuring overall voice quality, but the reader is directed to the height of r or $\overline{r_w}$ as a quantity-based overview of the domain of acoustic voice quality measurement. Furthermore, the tables show the relative position of a given acoustic measure according to its concurrent validity as a measure of overall voice quality. In this regard, the present meta-analysis was able to effectively distil an extremely large number of potential acoustic measures to a subset of strong independent variables. This should be particularly informative for voice practitioners in clinical settings who are faced with software packages that automatically generate a daunting number of acoustic measures ostensibly aimed to quantify dysphonia severity and track voice change following intervention. The present meta-analysis confirmed that not all acoustic measures are created equal with respect to these clinical goals.

ACKNOWLEDGMENTS

The assistance by Jan Deman (Medical Library, Sint-Jan General Hospital, Bruges, Belgium) for library work and article retrieval is greatly appreciated. The authors also would like to credit the associate editor and the three anonymous reviewers for the numerous valuable comments on earlier versions of this manuscript.

Askenfelt, A. G., and Hammarberg, B. (**1986**). "Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures," J. Speech Hear. Res. **29**, 50–64.

Awan, S. N., and Roy, N. (**2006**). "Toward the development of an objective index of dysphonia severity: A four-factor acoustic model," Clin. Linguist. Phonetics **20**, 35–49.

Bele, I. V. (**2005**). "Reliability in perceptual analysis of voice quality," J. Voice **19**, 555–573.

Buder, E. H. (**2000**). "Acoustic analysis of voice quality: A tabulation of algorithms 1902–1990," in *Voice Quality Measurement*, edited by R. D. Kent and M. J. Ball (Singular, San Diego, CA), pp. 119–244.

De Bodt, M. (**1997**). "A framework of voice assessment: The relation between subjective and objective parameters in the judgment of normal and pathological voice," Ph.D. thesis, University of Antwerp, Antwerp, Belgium.

de Krom, G. (**1993**). "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," J. Speech Hear. Res. **36**, 254–266.

Dejonckere, P. H. (**1998**). "Cepstral voice analysis: Link with perception and stroboscopy," Rev. Laryngol. Otol. Rhinol. (Bord) **119**, 245–246.

Dejonckere, P. H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier-Buchman, L., and Millet, B. (**1996**). "Differentiated perceptual evaluation of pathological voice quality: Reliability and correlations with acoustic

measurements," Rev. Laryngol. Otol. Rhinol. (Bord) **117**, 219–224.
Dejonckere, P. H., and Wieneke, G. H. (**1996**). "Cepstra of normal and pathological voices: Correlation with acoustic, aerodynamic and perceptual data," in *Advances in Clinical Phonetics*, Studies in Speech Pathology and Clinical Linguistics Vol. **6**, edited by M. J. Ball and M. Duckworth (John Benjamins, Amsterdam), pp. 217–227.
Eadie, T. L., and Baylor, C. R. (**2006**). "The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice," J. Voice **20**, 527–544.
Eadie, T. L., and Doyle, P. C. (**2002**). "Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers," J. Acoust. Soc. Am. **112**, 3014–3021.
Eadie, T. L., and Doyle, P. C. (**2005**). "Classification of dysphonic voice: Acoustic and auditory-perceptual measures," J. Voice **19**, 1–14.
Feijoo, S., and Hernández, C. (**1990**). "Short-term stability measures for the evaluation of vocal quality," J. Speech Hear. Res. **33**, 324–334.
Franzblau, A. N. (**1958**). *A Primer of Statistics for Non-Statisticians* (Harcourt, Brace & Company, New York).
Frey, L. R., Botan, C. H., Friedman, P. G., and Kreps, G. L. (**1991**). *Investigating Communication: An Introduction to Research Methods* (Prentice-Hall, Englewood Cliffs, NJ).
Gorham-Rowan, M. M., and Laures-Gore, J. (**2006**). "Acoustic-perceptual correlates of voice quality in elderly men and women," J. Commun. Disord. **39**, 171–184.
Halberstam, B. (**2004**). "Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels," ORL **66**, 70–73.
Heman-Ackah, Y. D., Heuer, R. J., Michael, D. D., Ostrowski, R., Horman, M., Baroody, M. M., Hillenbrand, J., and Sataloff, R. T. (**2003**). "Cepstral peak prominence: A more reliable measure of dysphonia," Ann. Otol. Rhinol. Laryngol. **112**, 324–333.
Heman-Ackah, Y. D., Michael, D. D., and Goding, G. S. (**2002**). "The relationship between cepstral peak prominence and selected parameters of dysphonia," J. Voice **16**, 20–27.
Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. (**1994**). "Acoustic correlates of breathy vocal quality," J. Speech Hear. Res. **37**, 769–778.
Hillenbrand, J., and Houde, R. A. (**1996**). "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," J. Speech Hear. Res. **39**, 311–321.
Hirano, M., Hibi, S., Terasawa, R., and Masako, F. (**1986**). "Relationship between aerodynamic, vibratory, acoustic and psychoacoustic correlates in dysphonia," J. Phonetics **14**, 445–456.
Hunter, J. E., Schmidt, F. L., and Jackson, G. B. (**1982**). *Meta-Analysis, Cumulating Research Findings Across Studies* (Sage, Beverly Hills, CA).
Kojima, H., Gould, W. J., Lambiase, A., and Isshiki, N. (**1980**). "Computer analysis of hoarseness," Acta Oto-Laryngol. **89**, 547–554.
Kreiman, J., and Gerratt, B. (**2000a**). "Measuring vocal quality," in *Voice Quality Measurement*, edited by R. D. Kent and M. J. Ball (Singular, San Diego, CA), pp. 73–101.
Kreiman, J., and Gerratt, B. (**2005**). "Perception of aperiodicity in pathological voice," J. Acoust. Soc. Am. **117**, 2201–2211.
Kreiman, J., and Gerratt, B. R. (**2000b**). "Sources of listener disagreement in voice quality assessment," J. Acoust. Soc. Am. **108**, 1867–1876.
Kreiman, J., Gerratt, B. R., and Ito, M. (**2007**). "When and why listeners disagree in voice quality assessment tasks," J. Acoust. Soc. Am. **122**, 2354–2364.
Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., and Berke, G. S. (**1993**). "Perceptual evaluation of voice quality: Review, tutorial, and a framework for research," J. Speech Hear. Res. **36**, 21–40.
Kreiman, J., Gerratt, B. R., and Precoda, K. (**1990**). "Listener experience and perception of voice quality," J. Speech Hear. Res. **33**, 103–115.
Lipsey, M. W., and Wilson, D. B. (**2001**). *Practical Meta-Analysis* (Sage, Thousand Oaks, CA).
Ma, E., and Yiu, E. (**2006**). "Multiparametric evaluation of dysphonic severity," J. Voice **20**, 380–390.
Parsa, V., and Jamieson, D. G. (**2001**). "Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech," J. Speech Lang. Hear. Res. **44**, 327–339.
Plant, R. L., Hillel, A. D., and Waugh, P. F. (**1997**). "Analysis of voice changes after thyroplasty using linear predictive coding," Laryngoscope **107**, 703–709.
Prosek, R. A., Montgomery, A. A., Walden, E., and Hawkins, D. B. (**1987**). "An evaluation of residue features as correlates of voice disorders," J. Commun. Disord. **20**, 105–117.
Qi, Y., Hillman, R. E., and Milstein, C. (**1999**). "The estimation of signal-to-noise ratio in continuous speech for disordered voices," J. Acoust. Soc. Am. **105**, 2532–2535.
Titze, I. R. (**1995**). *Workshop on Acoustic Voice Analysis: Summary Statement* (National Center for Voice and Speech, Iowa City, IA).
Wolfe, V., Fitch, J., and Cornell, R. (**1995**). "Acoustic prediction of severity in commonly occurring voice problems," J. Speech Hear. Res. **38**, 273–279.
Wolfe, V., Fitch, J., and Martin, D. (**1997**). "Acoustic measures of dysphonic severity across and within voice types," Folia Phoniatr Logop **49**, 292–299.
Wolfe, V., and Martin, D. (**1997**). "Acoustic correlates of dysphonia: Type and severity," J. Commun. Disord. **30**, 403–416.
Wolfe, V. I., Martin, D. P., and Palmer, C. I. (**2000**). "Perception of dysphonic voice quality by naïve listeners," J. Speech Lang. Hear. Res. **43**, 697–705.
Wolfe, V. I., and Steinfatt, T. M. (**1987**). "Prediction of vocal severity within and across voice types," J. Speech Hear. Res. **30**, 230–240.
Yiu, E. M., Chan, K. M., and Mok, R. S. (**2007**). "Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation," Clin. Linguist. Phonetics **21**, 29–45.
Yu, P., Garrel, R., Nicollas, R., Ouaknine, M., and Giovanni, A. (**2007**). "Objective voice analysis in dysphonic patients: New data including nonlinear measurements," Folia Phoniatr Logop **59**, 20–30.
Yu, P., Ouaknine, M., Revis, J., and Giovanni, A. (**2001**). "Objective voice analysis for dysphonic patients: A multiparametric protocol including acoustic and aerodynamic measurements," J. Voice **15**, 529–542.
Yumoto, E., Sasaki, Y., and Okamura, H. (**1984**). "Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness," J. Speech Hear. Res. **27**, 2–6.
Zraick, R. I., Wendel, K., and Smith-Olinde, L. (**2005**). "The effect of speaking task on perceptual judgment of the severity of dysphonic voice," J. Voice **19**, 574–581.