Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

PURPOSE AND APPLICATIONS

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) was developed as a tool for clinical auditory-perceptual assessment of voice. Its primary purpose is to describe the severity of auditory-perceptual attributes of a voice problem, in a way that can be communicated among clinicians. Its secondary purpose is to contribute to hypotheses regarding the anatomic and physiological bases of voice problems and to evaluate the need for additional testing.

CAPE-V is *not* intended for use as the only means of determining the nature of the voice disorder. It is not to be used to the exclusion of other tests of vocal function. Finally, it is not expected to demonstrate a 1:1 relation to results from other tests of vocal function.

ORIGIN

The CAPE-V was developed from a consensus meeting sponsored by the American Speech-Language-Hearing Association's (ASHA) Division 3: Voice and Voice Disorders, and the Department of Communication Science and Disorders, University of Pittsburgh, held in Pittsburgh on June 10-11, 2002. Attending this meeting were speech-language pathologists (SLPs) who specialize in voice disorders and invited experts in human perception (see appendix). The participants' charge was to develop standardized guidelines for auditory-perceptual evaluation of voice, based on theory and data in psychoacoustics, psychometric scaling, and voice perception. Clinical practicality and brevity of administration were also considered in developing these guidelines.

A working group was charged to formalize a consensus statement about minimal recommended standards for optimizing auditory-perceptual judgments in the clinical assessment of voice disorders by speech-language pathologists. The CAPE-V is the initial product. The hope is that wide-spread use of the current CAPE-V and its future development will encourage a more consistent approach and ultimately more research in the perceptual evaluation of voice disorders. The present document is the preliminary result of the consensus meeting. The ultimate goal is standardization of a reliable tool for clinical voice quality measurement.

DESIGN CONSIDERATIONS

The consensus was that the clinical evaluation of auditory-perceptual characteristics of voice should be derived from a tool with the following attributes: (a) perceptual dimensions should reflect a minimal set of clinically meaningful, perceptual voice parameters, identified by a group of expert clinicians; (b) procedures and results should be obtainable expediently; (c) procedures and results should be applicable to a broad range of vocal pathologies and clinical settings; (d)

ratings ultimately should be demonstrated to optimize reliability within and across clinicians , and (e) ultimately, exemplars should be available for training.

DESCRIPTION AND INSTRUCTIONS

General Description of the Tool: The CAPE-V indicates salient perceptual vocal attributes, identified by the core consensus group as commonly used and easily understood. The attributes are: (a) Overall Severity; (b) Roughness; (c) Breathiness; (d) Strain; (e) Pitch; and (f) Loudness. The CAPE-V displays each attribute accompanied by a 100 millimeter line forming a visual analog scale (VAS). The clinician indicates the degree of perceived deviance from normal for each parameter on this scale, using a tic mark. For each dimension, scalar extremes are unlabeled. Judgments may be assisted by referring to general regions indicated below each scale on the CAPE-V: “MI” refers to "mildly deviant," “MO” refers to “moderately deviant,” and “SE” refers to "severely deviant." A key issue is that the regions indicate *gradations* in severity, rather than discrete points. The clinician may place tick marks at any location along the line.” Ratings are based on the clinician’s direct observations of the patient’s performance during the evaluation, rather than patient report or other sources.

To the right of each scale are two letters, “C” and “I.” “C” represents "consistent" and “I” represents "intermittent" presence of a particular voice attribute. The rater circles the letter that best describes the consistency of the judged parameter. A judgment of “consistent” indicates that the attribute was continuously present throughout the tasks. A judgment of “intermittent” indicates that the attribute occurred inconsistently within *or* across tasks. For example, an individual may consistently exhibit a strained voice quality across all tasks, which include sustained vowels and speech. In this case, the rater would circle “C” to the right of the strain scale. In contrast, another individual might exhibit consistent strain during vowel production, but intermittent strain during one or more connected speech task. In this case, the rater would circle “I” to the right of the strain scale.

Definitions of Vocal Attributes:
OVERALL SEVERITY: Global, integrated impression of voice deviance.
Roughness: Perceived irregularity in the voicing source.
Breathiness: Audible air escape in the voice.
Strain: Perception of excessive vocal effort (hyperfunction).
Pitch: Perceptual correlateof fundamental frequency. This scale rates whether the individual's pitch deviates from normal for that person's gender, age, and referent culture. The direction of deviance (high or low) should be indicated in the blank provided above the scale.
Loudness: Perceptual correlateof sound intensity. This scale indicates whether the individual's loudness deviates from normal for that person's gender, age, and referent culture. The direction of deviance (soft or loud) should be indicated in the blank provided above the scale.

Blank scales and additional features: The six standard vocal attributes included on the CAPE-V are considered the minimal set of parameters for describing the auditory-perceptual characteristics of disordered voices. The form also includes two unlabeled scales. The clinician may use these to rate additional prominent attributes required to describe a given voice. The clinician may indicate the presence of other attributes or “positive signs” not noted elsewhere

under “Additional features.” If an individual is aphonic, this should be noted under "additional features" and no additional marks should be made on the scales.

Data collection: The individual should be seated comfortably in a quiet environment. The clinician should audio record the individual’s performance on three tasks: vowels, sentences, and conversational speech. Standard recording procedures should be used that incorporate a condensor microphone placed 45 degrees off from the front of the mouth and a 4 cm mike-to-mouth distance. Audio recordings are recommended to be made onto a computer with 16 bits of resolution and a signal sampling rate of no less than 20 KHertz (details included in Appendix; see http://www.ncvs.org/rescol/sumstat/sumstat.pdf).

Task 1: Sustained vowels: Two vowels were selected for this task. One is considered a lax vowel (/a/) and the other tense (/i/). In addition, the vowel, /i/, is the sustained vowel used during videostroboscopy. Thus, the use of this vowel during this task offers an auditory comparison to that produced during a stroboscopic exam.

The clinician should say to the individual, “The first task is to say the sound, /a/. Hold it as steady as you can, in your typical voice, until I ask you to stop.” (The clinician may provide a model of this task, if necessary) The individual performs this task three times for 3-5 sec each.“Next, say the sound, /i/. Hold it as steady as you can, in your typical voice, until I ask you to stop.” The individual performs this task three times for 3-5 sec each.

Task 2: Sentences: Six sentences were designed to elicit various laryngeal behaviors and clinical signs. The first sentence provides production of every vowel sound in the English language, the second sentence emphasizes easy onset with the /h/, the third sentence is all voiced, the fourth sentence elicits hard glottal attack, the fifth sentence incorporates nasal sounds, and the final sentence is weighted with voiceless plosive sounds.

The clinician should give the person being evaluated flash cards, which progressively show the target sentences (see below) one at a time. The clinician says, “Please read the following sentences one at a time, as if you were speaking to somebody in a real conversation.” (Individual performs task, producing one exemplar of each sentence.) If the individual has difficulty reading, the clinician may ask him or her to repeat sentences after verbal examples. This should be noted on the CAPE-V form. The sentences are: (a) The blue spot is on the key again; (b) How hard did he hit him? (c) We were away a year ago; (d) We eat eggs every Easter; (e) My mama makes lemon jam; (f) Peter will keep at the peak.

Task 3: Running speech: The clinician should elicit at least 20 seconds of natural conversational speech using standard interview questions such as, “Tell me about your voice problem." or "Tell me how your voice is functioning." ”

Data scoring: The clinician should have the individual perform all voice tasks—including vowel prolongation, sentenceproduction, andrunning speech, before completing the CAPE-V form. If performance is uniform across all tasks, the clinician should mark the ratings indicating overall performance for each scale. If the clinician notes a discrepancy in performance across tasks, he or she should rate performance on each task separately, *on a given line*. Only one CAPE-V form is used per individual being evaluated. In the case of discrepancies across tasks, tick marks should be labeled with the task number. Tick marks reflecting vowel prolongation should be

labeled #1 (see form). Tick marks reflecting running speech should be labeled #2. Tick marks reflecting story retelling should be labeled #3. In the rare event that the clinician perceives discrepancies within task type (for example, /a/ versus /i/), he or she may further label the ratings accordingly [for example, 1/a/ versus 1/i/ to reflect the different vowels, or 2(a)-(b)-(c)-(d)-(e)- or (f) for the different sentences]. Unlabeled tick marks indicate uniform performance. See examples below. [Note: Using labels to indicate discrepancies/variation across tasks in the severity of an attribute is different than indicating that an attribute is displayed intermittently (I). If an attribute is judged to have equal severity whenever it appears, but it is not present all the time, "I" should be circled to indicate that the attribute is intermittent and no additional labeling needs to be done.]

Scoring: After the clinician has completed all ratings, he or she should measure ratings from each scale. To do so, he or she should physically measure the distance in mm from the left end of the scale. The mm score should be written in the blank space to the far right of the scale, thereby relating the results in a proportion to the total 100 mm length of the line. The results can be reported in two possible ways. First, results can indicate distance in mm to describe the degree of deviancy, for example "73/100" on "strain." Second, results can be reported using descriptive labels that are typically employed clinically to indicate the general amount of deviancy, for example "moderate-to-severe" on "strain." We strongly suggest using both forms of reporting.

It is strongly recommended that for all rating sessions following the initial one, the clinician have a paper or electronic copy of the previous CAPE-V ratings available for comparison purposes. He or she should also rate subsequent examinations based on direct comparisons between earlier and current audio recordings. Such an approach should optimize the internal consistency/reliability of repeated sequential ratings within a patient, particularly for purposes of assessing treatment outcomes. Although difficult, clinicians are encouraged to make every effort to minimize bias in all ratings. We acknowledge that this solution is imperfect.

Other procedures: The clinician can indicate prominent observations about resonance phenomena under "Comments about resonance." Examples include, but are not limited to hyper- or hyponasality, and cul-de-sac resonance.

Cautions: Data available on the reliability of all rating scales for voice assessment indicate that both intra- and inter-judge agreement varies widely. Although we have attempted to limit sources of variability in the present tool, its reliability and validity have not yet been assessed. Future editions are projected to include referent voice recordings as "anchors" as well as training modules.

Examples: Refer to Example Form 1 for the following description. The patient displays the following ratings: Moderate to severe degree of overall dysphonia (78/100), moderate roughness (56/100), moderate to severe breathiness (74/100) and strain (62/100). Modal pitch (35/100) was judged to be mild to moderately low for the person's gender and age while the loudness (0/100) was judged to be normal. All voice attributes were judged as consistently present in this assessment. The patient in Example 1 also exhibits a positive sign for abnormal oral-pharyngeal resonance in the form of mild hypernasality.

Refer to Example Form 2 for the following description. The patient in Example 2 is status-post therapy, and displays the following ratings currently: Inconsistently mild degree of dysphonia (27/100), just noticeable roughness (3/100), mild to moderate inconsistent breathiness (38/100), inconsistently mild strain (9/100), consistently normal pitch (1/100), and inconsistently mild to moderate reduced loudness (29/100). He was further exhibits consistent mild to moderate asthenia (39/100). Most of the parameters are improved relative to earlier, pre-therapy ratings (Example Form 2).

Acoustic Recordings

Based on Titze (1994), http://www.ncvs.org/rescol/sumstat/sumstat.pdf

Many sections involve direct quotes

a. Use professional grade condenser microphone (omnidirectional or cardioid) with a minimum sensitivity of -60 dB (Titze & Winholtz, 1993)

b. For vowel utterances, the mouth-to-microphone distance should be held constant and less than 10 cm (preferably 3-4 cm) to avoid an artificial "wow" and to maintain a high signal-to-noise ratio; a miniature head-mounted microphone is recommended (Winhold & Titze, in press).

c. Close microphone distances require off-axis positioning (45 degrees to op degrees from the mouth axis) to reduce aerodynamic noise from the mouth in speech.

d. If samples are digitized into a computer, a 16-bit A/D converter or DAT recorder is recommended, but this must be accompanied by conditioning electronics (amplifiers, filters) that have signal-to-noise ratios in the 85-95 dB range (Doherty & Shipp, 1988).

e. If samples are digitized into a computer, sampling frequencies of 20-100kHz should be used, depending on the degree of interpolation between samples that the analysis software provides (Titze, Horii, & Scherer, 1987; Milenkovic, 1987; Deem et al., 1989).

f. If digitizing into a computer is used, manufacturers of workstations for acoustic voice analysis should be encouraged to provide DC coupling and low-frequency fidelity in acquisition hardware to accommodate physiologic signals (e.g., an electroglottograph, a flow mask) that augment the microphone signal. For all input signals, real-time feedback for clipping should be provided to avoid overloading the A/D converters. For DC coupling, there should be minimal drift and the drift should be reported and calibratable.

g. Line-level inputs (on the order of a few hundred millivolts) should be provided as a direct interface to the outputs of transducers, so that expensive high fidelity analog preamplifiers can be bypassed.

h. A digital audio tape (DAT) recorder should be used to store signals, unless A/D conversion is directly to the computer (Doherty & Shipp, 1988).

i. Recordings should be made in a sound-treated room (ambient noise < 50 dB); given that 120 Hz is very close to the average normal male speaking F_o, special care should be given to the removal of noise sources in the room that create 60 Hz hum and its associated harmonics. In general, one should specify the spectral weighting of the allowable noise in a sound-treated room. This is particularly important if inverse filtering from the microphone signal is attempted.