CONTENTS

CHAPTER 4

Assessing Voice Characteristics of Depression among English- and Spanish-Speaking Populations

Gerardo M. González and Amy L. Ramos

Abstract

Here we examine the integration of computerized speech recognition and digital voice analyses (VIDAS) to assess depressed mood and symptoms in English- and Spanish-speaking populations. The findings show VIDAS consistency to administer reliable, valid, and culturally sensitive screening of depression in these populations. VIDAS has been implemented in high volume health care settings that serve diverse patient populations but lack bilingual personnel. VIDAS quickly and unobtrusively collects participant data, scores the data, and generates a report to inform health care staff of the participant's mood and symptoms. As a result, VIDAS assesses many individuals who are unlikely to initially seek out mental health services. However, further study needs to be accomplished in order to enhance and refine the VIDAS interview as a *viable* alternative method of assessment.

The relationship between the gender of the participant and choice of digitized voice showed a preference for a female

digitized voice. Several voice characteristics showed significant relationships to depression levels, such as vocal energy and variability; however, the findings have not been consistent across the various VIDAS studies. Shortcomings with the analysis of voice characteristics are discussed and a role of a baseline measurement is stressed as it may be difficult to discriminate between a person who is depressed and one who normally speaks with a monotonic voice, and because psychiatric comorbidity and medications also distort vocal markers for depression. Also gender, age, linguistic, and physical factors that interact with speech characteristics may require developing unique models of vocal emotional properties.

Assessing Voice Characteristics of Depression among English and Spanish Speakers

Depressive disorders afflict 6% to 7% of the general population in the United States (Smith & Weissmann, 1992). Major depressive disorder is the leading cause of disability in the United States and developing countries (World Health Organization, 2001). Many depressed individuals are treated at primary care medical settings, where up to 30% of the patients may be clinically depressed (Broadhead, Clapp-Channing, Finch, & Copeland, 1989). Primary health care settings, however, suffer from deficiencies in screening practices, high patient volume, and enormous time constraints that hinder the adequate assessment of depression. Pérez-Stable, Miranda, Muñoz, & Ying (1990) found that depression was accurately detected in only 36% of primary care medical patients.

Latinos constitute nearly 13% of the U.S. population, comprise the second largest ethnic group in the United States, and are the fastest growing ethnic group in the country (U.S. Census, 2000). Past research suggests that Latinos are at higher risk for depression than non-Latinos. For example, Kessler, McGonagle, Zhao, & Nelson (1994) found that Latinos reported an 8.1% prevalence rate for current affective disorders (7% is the norm). In fact, Mexican Americans reported a higher prevalence for affective disorders than their Mexican-born counterparts (Vega et al., 1998).

Latinos in the United States generally lack accessibility to culturally responsive and linguistically-compatible mental health services (González, 1997). An Epidemiological Catchment Area (ECA) study indicated that only 11% of Mexican Americans (vs. 22% of non-Hispanic Whites), who met the criteria for clinical depression, sought a mental health care provider for treatment (Hough et al., 1987; Shapiro et al., 1984). Latinos underutilize mental health services because of cultural, linguistic, financial, and service delivery barriers (Woodward, Dwinell, & Arons, 1992). Moreover, 40% of the U.S. Latino population primarily speaks Spanish or has limited English proficiency (U.S. Census, 2000).

The absolute number of Latino therapists in the United States (29 for every 100,000 Latinos compared to 173 clinicians per 100,000 non-Hispanic Whites) represents an insufficient number to feasibly meet the present mental health needs of U.S. Latino populations (Center for Mental Health Services, 2000). Clearly, more appropriately trained culturally sensitive bilingual mental health professionals are needed. Yet the growing disparity between the Latino population (estimated to increase over 50% in the next decade) and current pool of Latino clinical psychology doctoral students in the training pipeline (levels static since 1980) makes it unlikely that ample Spanish-speaking professionals will be available to provide necessary services (e.g., Bernal & Castro, 1994). Alternative strategies for delivering culturally responsive mental health assessment services for the detection of depression in Spanish-speaking communities are needed.

Computerized psychological assessment represents several major advantages in the structure, flexibility, and ease of test administration (Kobak, 1996). Structured computerized interviewing improves the quality, quantity, and integrity of clinical data by accurately transcribing, scoring, and storing patient responses, standardizing administration procedures, and minimizing errors attributable to human oversight (Erdman, Klein, & Greist, 1985). For example, a clinician may inadvertently omit up to 35% of clinically meaningful inquiries during an open-ended face-to-face (Climent, Plutchik, & Estrada, 1975). Many depressed patients report a preference for computer interactive interviews over face-to-face interviews, even when patients knew the clinician (e.g., Carr, Ghosh, & Ancill, 1983). One possible explanation for such a preference is that computerized interviewing may increase respondent self-disclosure because of discomfort with revealing sensitive issues (e.g., suicidal ideation) to a clinician (Levine, Ancill, & Roberts, 1989). Another appealing aspect of computerized assessment is that it produces a cost savings through the use of more efficient professional time to conduct assessment batteries and treatment (Butcher, 1987). Thus, computerized screening provides a cost-effective and efficient means for assessing depression.

Recent advances in computerized technology offer viable alternative screening methods for populations not reliably assessed with standard paper-and-pencil questionnaires (Starkweather & Muñoz, 1989). For example, illiterates or non-English speakers are less likely to utilize mental services because of written assessment or language barriers. For such populations, computerized technology has the potential to minimize the obstacles that contribute to the underidentification of depression. Among the technologies that have strong potential is computerized speech recognition. A computerized speech recognition application is capable of administering a discrete choice questionnaire by presenting an item (visually on a computer screen or aurally by a prerecorded prompt) and recognizing a spoken response. Based on the capabilities of speech recognition technology and the imminent need for alternative depression screening methods in English- and Spanish-speaking communities, González and colleagues developed bilingual computerized speech recognition applications for screening depression.

Research also indicates that voice analysis may improve the accuracy of detecting depression. Digital analysis of voice characteristics represents a powerful

methodology for the objective assessment of depression (Starkweather, 1992). Voice characteristics serve as useful clinical indices for depression symptoms because vocalizations (respiration, articulation, and tension or relaxation of larynx and oral muscles) are mediated by psychomotor disturbance stemming from neurophysiological and subcortical (mesolimbic) dysfunction (Flint, Black, Campbell-Taylor, Gailey, & Levington, 1993; Nilsonne, Sunberg, Ternstrom, & Askenfelt, 1988).

Research demonstrates that several quantitative voice characteristics are good predictors of depression, such as narrow variability in tone (monotone), low fundamental frequency (pitch), and low amplitude or loudness (Hargreaves & Starkweather, 1964; Vanger, Summerfield, Rosen, & Watson, 1992). Multilingual research has generated a model of depressed voice prosody (tempo and rhythm) represented by slower, flatter, and softer voice waves (Darby, Simmons, & Berger, 1984; Kuny & Stassen, 1993; Scherer & Zei, 1988). Cross-cultural studies also suggest that depressed individuals display distinctive speech patterns compared to nondepressed persons, including more pauses and fewer utterances (e.g., Friedman & Sanders, 1992; Stassen, Bomben, & Günther, 1991) and longer vocal response latency (vocal reaction time) to answer a presented item (e.g., Stout, 1981; Talavera, Sáiz-Ruiz, & García-Toro, 1994). Furthermore, changes in speech variables are better predictors of mood change for patients in treatment than psychiatrists' impressions (Siegman, 1987). Thus, voice analysis can help to discern between the acoustic characteristics of depressed and nondepressed persons.

Quantitative acoustic variables include speech rate (number of utterances per time frame), mean pitch (average fundamental frequency of utterances), pitch variability, changes in pitch, and vocal intensity (energy values of an utterance). For example, a sad mood displays identifiable vocal markers (e.g. slow, soft, monotonic speech) that are distinguishable from vocal effects in normal mood and other emotional states. Table 4–1 summarizes the general research findings on vocal characteristics for several emotional states (Murray & Arnott, 1993).

The two most common voice analyses of depression models are the structured speech and free-form speech approaches (Alpert, Pouget, & Silva, 2000). The structured speech approach requires the respondent to repeat a determined sound (please say "A") or to read text (please read the following paragraph). The recorded repetition or text is assessed for mood with short-time, long-time, and spectral analyses. The free-form speech approach involves the assessment of natural open-ended speech. The respondent is asked an open-ended question and the free-form response is recorded and analyzed. Also, it is common to obtain a pretest (baseline) of an individual's voice characteristics and a post-test (after treatment or intervention) to assess change in mood or emotion.

Speech Behavior

Gonzalez and colleagues initiated speech recognition research for investigating speech behavior to increase the detection of depression. Initially, the researchers explored speech behavior, such as vocal

Table 4–1. Summary of the research findings on vocal emotional effects relative to neutral speech

	Sadness	Fear	Disgust	Anger	Happiness
Speech rate	**Slightly slower**	Much faster	Very much slower	Slightly faster	Faster or slower
Pitch average	**Slightly lower**	Very much higher	Very much lower	Very much higher	Much higher
Pitch range	**Slightly narrower**	Much wider	Slightly wider	Much wider	Much wider
Pitch changes	**Downward inflections**	Normal	Wide, downward inflections	Abrupt, on stressed syllables	Smooth, upward inflections
Vocal intensity	**Lower**	Normal	Lower	Higher	Higher
Voice quality	**Resonant**	Irregular voicing	Grumbled, chest tone	Breathy, chest tone	Breathy, blaring
Articulation	**Slurring**	Precise	Normal	Tense	Normal

Note. From "Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion," by I. A. Murray and J. L. Arnott, 1993, *Journal of the Acoustical Society of America*, *93*, pp. 1097–1108.

response latency (VRL) and speech recognition accuracy (SRA), i.e., computer accuracy level for recognizing a participant's utterances. The researchers hypothesized that longer VRL and lower SRA would be related to depressed mood.

The speech recognition applications are based on the Center for Epidemiological Studies-Depression scale (CES-D). The CES-D is a 20-item self-report screening measure developed by the National Institute of Mental Health (NIMH) for assessing the frequency of depressive mood and symptoms during the past week (less than 1 day, 1–2 days, 3–4 days, 5–7 days). In the general population, a cut point score of 16 or greater suggests a high level of depressive symptoms (Radloff, 1977). The CES-D has strong psychometric sensitivity for identifying symptomatic individuals; well established normative, reliability, and validity data with English- and Spanish-speaking samples; and extensive testing with clinical and nonclinical populations (Mosciki, Locke, Rae, & Boyd, 1989; Myers & Weissman, 1980).

González, Costello, La Tourette, Joyce, & Valenzuela (1997) evaluated a bilingual speaker-dependent cellular telephone-assisted computerized speech recognition CES-D. In a single session counterbalanced design, 32 English (ES) and 23 Spanish speakers (SS) completed randomly ordered computer-telephone (CT) and face-to-face (FF) CES-D methods (0–7 days' response format), the Beck Depression Inventory (BDI) (Beck & Steer, 1993), and the Short Acculturation Scale (SAS). VRL and SRA were measured. The results suggested that the two CES-D methods displayed strong internal consistency

estimates (a > .85), good alternate forms reliabilities (> .85), and high correlations to the BDI (r > .80) for both language groups. The two groups rated both methods equally high, but the ES preferred the FF mode because it was more personable. Among SS, the correlation between depression and acculturation was not significant. For the CT method, depression scores directly correlated with VRL (.45) and inversely related to speech recognition accuracy (−.37) across both language groups. Thus, longer VRL and lower SRA (more recognition complications) served as general indices of depression.

González and colleagues conducted two studies with large Spanish- and English-speaking samples and collected retest data on participants in a second session. The purpose of the *two* one-year studies was to develop, test, and evaluate an English and Spanish continuous speaker-independent speech recognition CES-D application for screening depression symptoms by digital cellular telephone. A continuous speaker-independent system is designed to recognize natural continuous speech across multiple independent users. The system does not require template training; thus, interview time is significantly reduced. Also, the system presented the interview using a prerecorded digitized female or male voice selected by the participant. In previous prototypes, only a prerecorded digitized male voice was presented.

Study 1 assessed the psychometric congruence of two speech recognition CES-D methods (0–7 days' choices) for detecting depression levels in ES and SS. A 2 (language) × 2 (method) × 2 (session) repeated measures experimental design was employed. The CES-D was randomly administered to 82 ES and 85 SS in CT or FF form in two sessions (at least a 2-week interval). Additional measures included a structured demographic interview, the Bidimensional Acculturation Scale (BAS) (Marín & Gamba, 1996), and the BDI. VRL and SRA were also measured. The results suggested that both methods displayed strong psychometric properties. The means for the two methods were generally not significantly different for both ES and SS. The two methods demonstrated high inter-item consistencies (a range .83 to .94) and strong correlations to the BDI (range .68 to .88) for both languages. Test-retest reliabilities were very good (range .84 to .89); however, reliability of the ES CT method was moderate (.47). Although the two language groups rated both methods highly, both groups preferred the FF method. Analyses of the digitized interviewer gender showed that ES chose a female voice significantly more often in the first session while SS selected a female voice more frequently in the second session. FF VRL was positively correlated to depression scores for the ES sample in the first (.29) and second sessions (.46). SRA was negatively correlated with depression scores in the ES first session (−.28) and SS second session (−.45). In other words, depressed persons tended to experience more voice recognition complications during the computer interview requiring more repetitions of the items and more time to complete the interview (González et al., 2000).

Study 2 was a validation study of an English and Spanish telephone-assisted speaker-independent CES-D. The aim of Study 2 was to evaluate the sensitivity (detecting true depressives) and specificity (detecting true nondepressives) of the CES-D for assessing major depression in ES and SS. Presentation of the CES-D (0–7 days' choices) was refined based on

the findings of Study 1. The unique features of Study 2 included administering the Composite International Diagnostic Interview (CIDI) (Kessler, Nelson, McGonagle, & Liu, 1996) to identify depressed and nondepressed participants. The relationship of depression scores to the BDI-II, BAS, and VRL were also assessed. Study 2 utilized a 2 × 2 × 2 language × diagnosis × session repeated measures (test-retest) design. A total of 160 participants (80 ES and 80 SS) including diagnosis group (depressed and nondepressed) were interviewed.

Data analyses revealed that there were no significant language group differences for the means and variabilities of the CES-D across both sessions. The CES-D displayed strong internal consistency for both language groups in both sessions (a ranged from .88 to .94). Test-retest reliabilities were .85 and .64 for the SS and ES, respectively. There were strong convergent validity coefficients between the CES-D and the BDI-II in both sessions (.69 and .67 for SS and .64 and .87 for ES). CIDI analyses indicated that the CES-D displayed good sensitivity (.76) and specificity (.50) for the first session and similar sensitivity (.77) and specificity (58) in the second session. More than two thirds of all participants selected a female digitized voice for the first session. In the second session, 60% of the participants selected a female voice. Participants positively rated (1 = very uncomfortable to 6 = very comfortable) the CES-D on the first session (both group means over 4.3, no significant differences). VRL and CES-D total scores were positively related in the first session (r = .14) but not in the second session (González, 2000).

Digital voice analysis packages (e.g., Avaaz Interactive Voice Analysis System, IVANS) conduct complex short-time and long-time acoustic analyses for detecting emotion in voice characteristics. Short-time analysis examines a segment of a voice signal, such as a phoneme (basic sound). Long-time spectrum analysis assesses the entire voice signal. Spectral analyses of voice samples generate *spectrograms*, which are two- and three-dimensional visual representations plotted along various acoustic variables (time, frequency, and amplitude). Table 4–2 summarizes the definitions of common acoustic voice measures (Avaaz, 1998). Digital voice analysis was implemented in the next phase of research.

Voice-Interactive Depression Assessment System

Gonzalez and colleagues evaluated the Voice-Interactive Depression Assessment System (VIDAS) to detect depression symptoms (using the CES-D) among English and Spanish speakers. The researchers developed VIDAS using speaker-independent continuous speech recognition technology (Schalkwyk, Colton & Fanty, 1998). The researchers administered VIDAS to the participants using a Pentium laptop computer (Windows XP) with a microphone/speaker handset.

VIDAS presented a discrete choice questionnaire in English or Spanish by playing digitally recorded .wav audio files, recognizing a respondent's spoken answers, scoring the responses, and storing the data. Two bilingual male and female professionals fluent in both English and Spanish recorded the prompts, instructions, and items in a neutral tone to reduce potential biases from participant reactivity. VIDAS randomly ordered

Table 4–2. Definitions of acoustic voice measures

Measure	Definition
Long-Term Average Speech Spectrum (LTASS)	Summary of how energy in an utterance is distributed across frequency, on average, over the duration of the specified signal.
Spectral tilt	Rate at which the energy of the speech signal declines as frequency increases.
Flatness	Represents the flatness of the LTASS. For speech signals that have more noise content ("breathy" signals).
Centroid	A weighted measure that determines the effective fulcrum of the LTASS. For unvoiced sounds, the spectral centroid is usually around 2–3 kHz, while voiced sounds have a lower spectral centroid.
Skewness	Quantifies the spread of the LTASS. For a spectrum that has a Gaussian shape, the skewness is equal to zero. Positive skewness values indicate more energy in the high frequency region, while negative skewness values reflect low frequency spectra.
Kurtosis	Quantifies the shape of the spectrum. Lower kurtosis values indicate flat spectra, while higher values indicate spectra with varying peaks.
Speech Measures	Acoustic measures gathered from running speech.
Tilt	Similar to the spectral tilt parameter, except only the voiced segments of speech are included in the computation of the LTASS.
Harmonic-to-noise ratio	The effect of both pitch and amplitude perturbations. It also accounts for such conditions as the increased noise in the main formant frequency region, increased high frequency noise, and decreased higher harmonics.
Linear prediction signal-to-noise ratio (SNR)	Relies on linear prediction modeling of the input speech sample. The SNR measure is taken as the ratio of the input signal energy and the energy of the residual signal at the output of the linear prediction model. Normal talkers typically have high LP-SNR values, which reflect good linear prediction modeling performance.
Pitch amplitude	The amplitude of the second largest peak of the normalized autocorrelation function of the residual signal.
Spectral flatness ratio (SFR)	A measure of how successfully the LP technique was able to model the input signal. If the LP model is successful, the residual signal is made up of a series of impulses, one at each glottal excitation period.

Note. From *Interactive Voice Analysis System (IVANS) User's Guide*, by Avaaz Innovations Inc., 1998. Reprinted with permission of Avaaz Innovations Inc: London, Ontario, Canada.

the digitized male or female voice to which the participant vocally responded. VIDAS also recorded participant vocal data for subsequent voice analysis using IVANS. Table 4–3 summarizes the basic VIDAS interview sequence.

Table 4–3. Summary of VIDAS interview sequence

1. **Introduction**
   a. Interviewer instructs participant (English or Spanish) for completing VIDAS
   b. Interviewer asks participant to choose the gender of digitized interviewer voice (male or female)
   c. Interviewer initiates the VIDAS application
   d. Over the handset, VIDAS greets participant in primary language (English or Spanish) and presents brief instructions for completing a scale (randomized)
2. **Pretest Recording**
   VIDAS instructs participant to repeat a phrase, "This computer responds to my voice."
3. **CES-D Items**
   a. VIDAS presents brief instructions for completing the CES-D items orally
   b. VIDAS begins by presenting an item and waits for the participant's response
   c. Participant verbally responds to the item
   d. VIDAS registers and records the participant's recognized spoken response
   e. VIDAS continues to the next item until all the items are completed
   f. VIDAS proceeds to the conclusion
4. **Post-Test Recording**
   VIDAS instructs participant to repeat a phrase, "This computer responds to my voice."
5. **Conclusion**
   a. VIDAS thanks the participant, requests that the interviewer be advised, and terminates
   b. VIDAS scores and analyzes the responses
   c. VIDAS saves the results in a database
   d. VIDAS generates a brief interpretative report (summary of responses and interpretation)

Study 1 involved the development and *pilot* testing of bilingual telephone and microphone speech recognition VIDAS-1 prototypes. VIDAS-1 was a computer-telephone or computer-microphone (CM) form of the CES-D (0–7 days' choices). Fifty-eight English speakers and 60 Spanish speakers completed a randomly assigned CT or CM method. Other measures included demographics, BAS, the BDI-II, and CIDI.

The results suggested that the CT and CM methods did not significantly differ in total score means and variabilities. VIDAS-1 demonstrated good reliability ($a > .80$ for CT and CM in both language groups) and strong validity with the BDI-II (r range .69 to .73 for CT and CM in both languages). VIDAS-1 demonstrated good sensitivity (.83) and moderate specificity (.38) across language groups and methods. Although ES rated ($M = 4.5$) both VIDAS methods higher than SS ($M = 3.8$), there was no significant language and method interaction. ES and SS were significantly more likely (80%) to select a female digitized voice for the VIDAS interview.

A free-form approach for assessing participants' individual responses to the first two, middle two, and last two CES-D items was utilized. VRL was significantly longer

for depressed (M = 5.5 sec) than nondepressed participants (M = 3.3). VRL was also longer for the SS CT group M = 5.5 (SD = 6.3) than the ES CT group M = 3.1, (SD = 1.2) and the SS CM group M = 3.8 (SD = 3.1) than the ES CM group M = 2.1 (SD = .61), respectively. Correlations between CES-D total scores and VRL were examined by VIDAS method and language group. There were no significant correlations between CES-D total scores and VRL for either method or for Spanish speakers. However, a significant correlation was found between CES-D total score and VRL for English speakers, r (26) = .47, p < .05 (Ramos, G. M. González, P. González, Goldwaser, & Preble, 2002).

Study 2 compared VIDAS-1 and VIDAS-2. VIDAS-2 differed from VIDAS-1 in that new depression items (20) and response formats were designed for three subscales: subscale 1 (yes/no), subscale 2 (discrete choices, e.g., "All of the time"), and subscale 3 (open-ended response to questions, e.g., "How was your appetite?"). For the purpose of brevity, only VIDAS-2 subscale 2 data will be summarized. In total, 130 ES and 95 SS participants completed BAS, BDI-II, CIDI, VIDAS-1, and VIDAS-2.

VIDAS-2 demonstrated strong inter-item consistency (ES a .90 and SS a .80) and positive correlations to the BDI-II (ES .65 and SS .53). VIDAS-2 demonstrated strong sensitivity (.82) and moderate specificity (.39) across both language groups. In choosing the gender of digitized voice, 62% of ES and 83% of SS females and 84% of ES and 66% of SS males selected a female voice. Both language groups positively rated VIDAS-2 (scale 1 to 6), but ES had significantly higher levels of comfort (M = 4.4) than SS (M = 4.0). Using a free-form analysis of participant responses to selected subscale 2 items (first two, middle two, and last two), VIDAS-1 depression levels were significantly correlated to measures of voice intensity, such as spectral tilt (–.20) and speech tilt (–.34); thus, depressed individuals displayed less vocal energy. VIDAS-2 did not show significant relationships between voice properties and subscale 2 total scores (Ramos, Shriver, Reza, & González, 2003).

VIDAS-3 was a bilingual computerized speech recognition application for screening depression using two subscales based on CES-D and DSM-IV criteria. In this study, 128 English and 128 Spanish speakers completed a demographic interview, BAS, BDI-II, the CIDI-Short Form, and VIDAS-3. Recordings of participant repetitions of a phrase, "This computer responds to my voice," were obtained before (pretest) and after completion (post-test) of the CES-D.

The results suggested that VIDAS-3 subscales demonstrated high inter-item reliability (.81 to .92), strong criterion validity (.58 to .67), and adequate sensitivity (.64 to .87) and specificity (.44 to .71). Both language groups positively rated VIDAS-3. Male and female participants most often selected a digitized female voice to present VIDAS-3. Long-term average speech spectrum (LTASS) measures (kurtosis, flatness, skewness, and centroid) that assess the tone and pitch of an individual's vocal characteristics were used as the dependent variables in a multivariate analysis of variance (MANOVA).

The results revealed a significant main effect for depression in the participants' pretest recorded phrase across both language groups, [*Pillai's Trace* = .064, F (1, 236) = 6.81; p =. 016]. Separate follow-up ANOVAs for each dependent variable showed significant differences in cen-

troid, F (1, 241) = 9.55, p = .002, skewness, F (1, 241) = 5.11, p = .025, and kurtosis, F (1, 241) = 11.60, p = .001. Thus, depressed participants had less vocal energy than nondepressed participants. A MANOVA of LTASS measures revealed a significant main effect for depression in participants' post-test recording [*Pillai's Trace* = .076, F (6, 224) = 3.092; p = .006]. Separate ANOVAs for each dependent variable showed significant differences in flatness, F (1, 229) = 5.15, p = .024. Depressed participants' vocal responses were flatter than nondepressed participants. As with the pretest results, there were significant differences for centroid, F (1, 229) = 10.67, p = .001, skewness, F (1, 229) = 7.18, p = .008, and kurtosis, F (1, 229) = 12.74, p < .0001.

A MANOVA of speech measures [tilt, voiced tilt, harmonic-to-noise ratio, low pitch-to-signal-to-noise ratio (LP-SNR), pitch amplitude, and signal frequency ratio (SFR)] revealed a significant main effect for depression in participants' pretest recording [*Pillai's Trace* = .08, F (7, 235) = 2.91; p =. 006]. Separate ANOVAs for each dependent variable suggested that there were significant differences in harmonic-to-noise ratio, F (1, 241) = 5.48, p =. 02, and pitch amplitude, F (1, 241) = 7.40, p =. 007. Thus, nondepressed participants displayed less noise in their vocal sounds while depressed participants had more hoarse and breathy vocal responses.

Finally, to test for differences across time between depressed and nondepressed participants, dependent *t*-tests were conducted for LTASS and speech measures on variables that were found to show significant differences between depressed and nondepressed participants. There were more significant differences for nondepressed than depressed individuals across time. Specifically, for nondepressed individuals there was a significant difference across time in the skewness t (171) = −2.43, p = .016, harmonic-to-noise ratio t (171) = −4.26, p < .0001, and pitch amplitude t (171) = −4.97, p < .0001. However, for depressed individuals there was only a significant difference in pitch amplitude t (59) = −2.95, p = .005. In sum, nondepressed participants displayed greater vocal variability in their voice characteristics than depressed participants (González & Shriver, 2004).

Two studies evaluated VIDAS-4 for screening depression and anxiety symptoms in English and Spanish. Study 1 involved 48 ES and 45 SS. Study 2 involved 112 ES and 108 SS. Participants completed a demographic scale, BAS, BDI-II, BAI (Beck & Steer, 1993), CIDI-SF, and VIDAS-4 depression and anxiety subscales. The studies examined the psychometric properties, comfort ratings, and selection of digitized gender for VIDAS-4. Study 2 examined the sensitivity and specificity of VIDAS-4 to detect depression and anxiety levels among comorbid, depressed, anxious, and no-disorder groups. As with VIDAS-3, participant pre- and post-recordings were obtained. The studies found that VIDAS-4 subscales generally demonstrated adequate inter-item reliability (.80–.94), convergent validity (.62–.89), sensitivity (.84–.90), and specificity (.44–.69). Most participants regarded VIDAS-4 as comfortable. Three of four participants selected a female digitized voice. Comorbid participants reported the most severe levels of depression or anxiety.

Participants' pretest and post-test recorded phrases were analyzed using MANOVA to assess differences between the four diagnostic groups by language. Roy's Largest Root was the statistic used

instead of more traditional analyses such as Pillai's Trace because Roy's Largest Root is said to be the best statistic for dealing with differences among the groups when the difference is concentrated on the first discriminant function (which was the case, and the test for homogeneity of covariance matrices was positive). The results revealed a significant main effect for LTASS measures among the four diagnostic groups for English-speaking participants [Roy's Largest Root = .11, F (3, 100) = 2.66; p = .037] and Spanish-speaking participants [Roy's Largest Root = .2, F (3, 81) = 3.99; p = .005]. Follow-up analyses did not reveal any significant differences between the four diagnostic groups.

Speech measures assessing the amount of vocal variability and energy between the four diagnostic groups were used as dependent variables in a MANOVA. The results revealed a significant main effect among the four groups, regardless of language [Roy's Largest Root = .094, F (3,104) = 1.97; p = .009]. Follow-up analyses revealed no significant differences between the four diagnostic groups or for each language (Shriver, Ramos, & Gonzalez, 2003).

VIDAS-5 is a computerized speech recognition application for screening depression and anxiety symptoms in English and Spanish. Study 1 was a pilot study of 50 ES and 47 SS. Study 2 involved 108 ES and 109 SS in diverse settings. Participants completed a demographic scale, BAS, BDI-II, BAI, CIDI-SF, and VIDAS-5 (aural or visual methods). The audio portion of VIDAS was the same for the aural and visual methods. The difference between methods involved visual cues for the visual method, such as text and graphical messages to present the items and to reply to a recognized spoken answer. As in VIDAS-3 and -4, pre- and post-test participant recordings were obtained.

Studies 1 and 2 examined the psychometric properties and participant comfort ratings for VIDAS-5. Study 2 also examined psychometric sensitivity and specificity, and participant selection of digitized gender for VIDAS. The studies found that VIDAS-5 generally demonstrated a range of adequate inter-item reliability (.71–.91), convergent validity (.40–.86), sensitivity (.79–.1.0), and specificity (.39–.44). Discriminant validity results demonstrated high overlap between depression and anxiety scales (.31–.79). Several differences were observed in the psychometric properties of VIDAS subscales by language and method, such that the DAS and aural method displayed lower reliability and validity. Participants in both language groups favorably rated the two VIDAS methods but the visual method received higher positive reactions. Participant comfort ratings of digitized voice demonstrated an interaction such that the female *visual* voice and the male *aural* voice were rated more favorably.

In a preliminary analysis of voice characteristics among depressed and nondepressed individuals, correlations were conducted between BDI total scores, CES-D total scores, LTASS measures, and speech measures. Among English-speaking participants, four pretest LTASS measures were significantly correlated with the BDI, including harmonic-to-noise ratio (r = .314, p < .01), signal-to-noise ratio (r = –.267, p < .05), pitch amplitude (r = .262, p <.05), and signal frequency ratio (r = –.245, p < .05). None of the voice variables were significantly correlated with the BDI in Spanish or the CES-D in English or Spanish. Among Spanish-speaking participants, two post-test speech measures were significantly correlated with the BDI, such as tilt (r = -.344, p < .01) and voiced tilt (r = -.348, p < .01). There

were no significant correlations between any of the voice variables and the BDI in English or the CES-D in either language. Thus, depressed participants displayed less vocal intensity and variability (Gorzeman, Carter, & González, 2005).

Summary

The research presented here examined the integration of computerized speech recognition and digital voice analyses to assess depressed mood and symptoms in English and Spanish. The findings suggest that VIDAS is a feasible to administer, reliable, valid, positively acceptable, and culturally sensitive application for the screening of depression in English- and Spanish-speaking populations. The relationship between the gender of the participant and choice of digitized interviewer is complex, but most participants more often selected a female digitized voice. The preference for a female therapist has been documented in previous research (Kaplan, Becker, & Tenke, 1991). Most importantly, several voice characteristics demonstrate significant relationships to depression levels, such as vocal energy and variability; however, the findings have not been consistent across the various VIDAS studies.

There are several shortcomings with the analysis of voice characteristics. The literature reports complexities with differentiating between labile and transitional emotional states such as sadness, boredom, and indifference (Scherer, 1986). Moreover, without a baseline measurement, it may be difficult to discriminate between a person who is depressed and one who normally speaks with a monotonic voice. Psychiatric comorbidity also distorts vocal markers for depression. For instance, depressed persons may display mixed voice characteristics that represent both psychomotor retardation and agitation (Mandal, Srivastava, & Singh, 1990). In addition, psychotropic medications alter the vocal expression of depression symptoms, such as change in pitch and voice energy (Standke & Scherer, 1984). There are also gender, age, linguistic, and physical factors that interact with speech characteristics (Scherer, Banse, Wallbott, & Goldbeck, 1991). Differences between male and female voice ranges, age groups (children and geriatric populations), regional pronunciations, and speech impediments may require developing unique models of vocal emotional properties (Scherer, Ladd, & Silverman, 1984). Occasionally, the speech recognition system did not recognize vocal responses accurately. The speaker-independent speech recognition technology used in VIDAS is based on syntax and a phonetic structure. Such systems have limitations with the recognition of variations in vocal utterances. Differences between the respondent's pronunciation and the system's phonetic structure may significantly diminish recognition and affect the interaction between computer and user (Noyes & Frankish, 1994).

Obviously, the limitations of speech recognition and voice analyses need to be addressed. Overall, significant progress has been made toward developing a tool to increase the early and accurate detection of depression cases. VIDAS has been implemented in high volume health care settings that serve diverse patient populations, but lack bilingual personnel. VIDAS quickly and unobtrusively collects participant data, scores the data, and generates a report to inform health care staff of the participant's mood and symptoms.

As a result, VIDAS assesses many individuals who are unlikely to initially seek out mental health services. However, further study needs to be accomplished in order to enhance and refine the VIDAS interview as a *viable* alternative method of assessment.

By and large, past research has focused on standard voice variables, such as pitch, tempo, and speech rate. Acoustic variables that measure fine-grained variations in voice signals, such as shimmer (modulation in amplitude) and jitter (irregularity in vocal vibration), offer new insights into the relationship between mood and voice characteristics (Bachorowski & Owren, 1995). Advancements in experimental methodologies (structured, free-form, and pre-post designs) and digital voice analysis (short-time, long-time, and spectral analyses) can overcome the limitations in the evaluation of speech variables associated with the quality of voice sampling (Murray & Arnott, 1993). State-of-the-art voice analysis software packages that can detect subtle changes in voice properties will aid in evaluating vocal emotion. These new developments offer possibilities to develop a reliable and valid English and Spanish language voice analysis to accurately discern between depression and nondepression.

Acknowledgments. The primary author thanks Colby Carter, Gali Goldwaser, Patricia Gonzalez, Paul Hernandez, Jennifer Reza, Carlos Rodriguez, and Chris Shriver for their efforts in the data collection and data analyses. MBRS grant number MS4567 from the National Institute of General Medical Science (NIGMS) and the LRP program of the National Institutes of Health (NIH) supported the development of this manuscript.

References

Alpert, M., Pouget, E. R., & Silva, R. R. (2000). Reflections of depression in acoustic measures of the patient's speech. *Journal of Affective Disorders*, *66,* 59–69.

Avaaz Innovations Inc. (1998). *Interactive Voice Analysis System™ (IVANS) user's guide*. London: Avaaz Innovations.

Bachorowski, J. A., & Owren, M. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, *6*(4), 219–224.

Beck, A. T., & Steer, R. A. (1987). *Manual for the revised Beck Depression Inventory*. San Antonio, TX: Psychological Corporation.

Beck, A. T., & Steer, R. A. (1993). *Manual for the revised Beck Anxiety Inventory*. San Antonio, TX: Psychological Corporation.

Bernal, M. A., & Castro, F. G. (1994). Are clinical psychologists prepared for service and research with ethnic minorities? Report of a decade of progress. *American Psychologist*, *49*, 797–808.

Broadhead, W., Clapp-Channing, N., Finch, J., & Copeland, J. (1989). Effects of medical illness and somatic symptoms on treatment of depression in a family medicine residency practice. *General Hospital Psychiatry*, *11*(3), 194–200.

Butcher, J. (1987). Computerized clinical and personality assessment using the MMPI. In J. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 161–197). New York: Basic Books.

Carr, A., Ghosh, A., & Ancill, R. (1983). Can a computer take a psychiatric history? *Psychological Medicine*, *13*(1), 151–158.

Center for Mental Health Services. (2000). *Cultural competence standards in managed care mental health services: Four underserved/underrepresented racial/ethnic groups.* Retrieved July 26, 2001, from http://www.mentalhealth.org/publications/allpubs/SMA00–3457

Climent, C., Plutchik, R, & Estrada, H. (1975). A comparison of traditional and symptom-

CHAPTER 13

The Context of Voice and Emotion: A Voice-Over Artist's Perspective

Kathleen Antonia Tarr

Abstract

This chapter reflects the perspective of an actor on the creation and understanding of believable vocalized emotions.

Introduction

Radio ads and commentary. Audio books. Phone calls and voice mail. There are very few other venues that are home to voice and emotion without visual context. However, whether the audience sees the speaker or other visual cue, the body is key to emotional expression. As a voice-over artist, I cannot act the part vocally if physically I am disconnected from the emotion. If I am disconnected, the audience will be disconnected, too. If I am on stage, any personal disconnection from the emotion I intend to portray also disconnects the audience. The same is true of film, and the same is true whether one can see me or only hear my voice.

The Sense, the Context, and the Know-How

In addition to my emotional intention, the effort to produce emotions in voice requires a physical *context*. It is not enough to be a snapshot: lips upturned, a furrowed brow, a feeling of disgust. Context requires incorporation of the moments before and after, the bookends. Context in written form is how one knows the difference between *tear*—to rip—and *tear*—to cry. In visual form, an onlooker can look at a smile and try to decipher the supporting emotion of the smiler, but can fail to detect whether this person is truly happy—or perhaps instead complicit—if she doesn't know the situation that inspired the moment. Hence, one crucial key to understanding emotion in voice is understanding the context.

Take, for example, the quick image of me in a Sunsweet prunes (rather, "dried plums") commercial (still airing as of this printing: Editor). I look down at a dried plum held between my fingers and say, "*Wow*." That's it.

After this commercial aired during the 2006 *Golden Globe Awards*, I received several phone calls asking about the ad, always with the tag, "You sure were enjoying that prune!"

Was I? The context suggests so. Images of others bookended mine, all with more dialog, all truly enjoying Sunsweet dried plums. Although I am not shown biting into the prune, there seems to be something in my mouth, and with prune in hand, the "wow" is correctly interpreted as referring to that item. My nose isn't scrunched up, my eyes and brow are raised, and I am looking at the prune. Context. I could have meant, "*Wow*! This is the most distasteful thing I've eaten all day!" but in addition to the context of my own expression and others' in the commercial, audiences prejudge correctly that a company is not going to include the image of someone who hates its product in a promotional ad and thus correctly conclude that I like the taste. It is interesting that of those who watched the *Golden Globes*—wherein there was an overflow of enjoyment and "Wow!"—many made the observation, "You sure were enjoying that prune!" During previous airings, say, during *The Amazing Race*—a show with, yes, enjoyment, but also quite a bit of "ugh"—I did not receive characterizations of my dried plums performance that stressed the depth of my delight.

Contrast two other wows. A friend of mine who once at dinner accidentally put an entire chili pepper in her mouth, thinking it was a sweet green pepper, kept repeating "Wow" as she brushed her tongue with her napkin and followed up

with a gallon of water. My sister—whom I had forewarned about the foulness of fermenting yeast—put Vegemite on toast and took a bite. Her wows were interspersed with chugs of guava juice. In fact, both my friend and my sister were smiling during their ordeals. One snapshot of the action, and they could easily have taken my spot in the Sunsweet dried plums commercial and fooled an audience into thinking they were enjoying their moments. But they weren't. Quite the opposite. Why would the snapshot work?

Well, all three of us were sharing the feeling of "That's incredible." I was actually enjoying the flavor of my item, but my sense at that moment regarded the infusion of orange into the prune. Surprising! Delicious! My friend was also surprised but taken aback by "Pain! Help!" My sister predominated by "Yuck! My God, this is the most disgusting thing I have ever tasted!" And because they both have excellent senses of humor, smiles, raised eyes, and raised brows.

More of Context

Because I know my friend and sister, and because I was there for their entire taste disaster experiences, I knew that the smiles under the wows were not expressing happiness or enjoyment about the flavors consuming their consumption. Someone on the scene with less experience or detail about the players and the stories might not have figured it out as quickly. They might have wondered about why these two women were having so much dramatic fun. As they watched, they might have figured out that the moments actually involved quite a bit of displeasure, but only in the context of the players, two people who see comedy in almost everything.

It is easy to conclude, then, that understanding the emotion of a particular vocal moment is not only about external context but also about understanding one's own biases, expectations, and interpretive patterns. Biases can be benign or even insightful, as above. But here is also where interpretations of emotions and voice can take sickening turns.

Racism, sexism, homophobia, and other mental delusions all carry with them fallacies of interpretation of emotions and of the voice. The same phrase uttered by a white person may be interpreted by a racist with black bias as a casual statement in the former demographic scenario, but if uttered by an African American, as asserting superiority: "I disagree. I think all people are equal." Someone without racist delusions can hear that statement from someone of another race and not usually be threatened.

Someone with racist delusions most often cannot. It depends upon the context one projects. If it is simply an intellectual discussion, and *both parties have similar understanding*, disagreements are not typically threatening. If one party thinks that disagreement means the other doesn't know her "place" or thinks he is smarter—emotionally that the other feels confident when she should feel subservient or when he should feel deferential—then problems arise.

I used to play tackle football with men who were high school and college team standouts. Notoriously, a couple of them who had never played with or against me taunted me in the days before a game. "You're not even going to show up. Women can't play football. What are you? The cheerleader?" "I'll see you on

Sunday," was my common reply. "I'll see you on Sunday," infuriated many of them. They believed I was saying that they were weak, that they weren't really men, that *they* might as well play the cheerleader. What did I mean? That none of their posturing meant anything to me. Skill would be decided on the field, not by argument. Did I think I was the better player? Hell yes! But did I think they were weak, unmanly, cheerleader types? No. I wasn't thinking of them at all. I had no *feeling* for them at all. But for what they believed I thought and felt, I was threatened on several occasions with violence.

One's own bias and prejudgment are the great interpreters of voice and emotions. They are as, if not more, important than any other measurement. Compare the "problem" with people affected by physical disorders that impact vocal tone or physicality, including facial expression. The difficulties they face with connecting the emotion to the voice are in the *interpretation*, not in the manifestation. (More on this subject is provided in Chapter 8, 16, and 17 in Volume 2 of these series: Editor.)

The speaker has all of the emotions, *speaks* with all of the emotions supporting the speech. The failure in understanding the emotion is in the interpretation of the voice. Someone more familiar with the speaker will do a better job of understanding. It is like any language. Someone without fluency will not as easily get it.

The search for the perfect robotic display of vocal emotion is the search for the language of the masses. Attempts to create authentic computerized voices are not in fact efforts to make the speech more decipherable but instead more comfortable for the listener, somehow more sincere, warmer. If people took a moment, they would realize that the computerized voice pleasantly speaking to them has no feeling for them whatsoever, and in the end, it doesn't really matter whether it is monotonous or "kind." In fact, much of the fury over the delighted voices of multi-prompters seems to be the result of frustration over *not* being acknowledged by the voice on the other end of the phone, a voice that *sounds* warm, more human, less monotone, and computerized.

I'm on the phone having a fit because I'm on my 10th prompt transfer and "someone" is having a lovely day, telling me, "I'm sorry" yet again, and *with an audible smile*! I've been less put out by monotones prompting me to enter my account number over and over again. I can *hear* that they don't care. Hey, I admit it. I'm human, biased and deceived and led places by voices without real feeling like a lamb to the slaughter—which brings me back to my job.

Tricks

There are certain tricks I use to help audiences correctly interpret the emotions I intend to project. When I play restrained anger, I might smile with my mouth, but I am predatory with my eyes, scrunched, pinpointing attack.

When I play restrained derangement, I smile with my mouth, and my eyes remain neutral. The words then flow from these places. Anger, they do not leave the throat easily. Derangement, they flow smoothly. My face might transform a thousand compositions, but because the voice and eyes are specific, the emotion is fairly transparent. In fact, acting any part, voice only or not, I rely primarily

upon my throat and eyes. My face follows, but its expressions are only important in as much as they reflect the secondary emotion, i.e., the one my *character* is hoping to project in the scene. What the character is actually feeling, the emotion *I* primarily want understood, is not found there.

There is a scene in *Kill Bill, Vol. 2* in which Uma Thurman's character is going to be buried alive. The camera gives the audience a shot of her face as the gravedigger exclaims, "Look at those eyes. This bitch is furious!" She is still above ground during this scene, not yet confined to her casket six feet under. The shot is only of her eyes, partial forehead, bridge of nose. She doesn't look furious to me. She looks frightened. When her casket is being closed, the camera reveals a cut of the same sequence. The lighting is the same. The point of view the same. She looks frightened. It makes sense. The earlier usage is obviously the same take, nonsensical, above ground, only by virtue of the editor's effort to substitute for an otherwise unavailable shot. Had she spoken and sounded furious in the earlier shot, I might have then thought the character was deranged. When the eyes and voice don't match, insanity. Open the eyes wide in utter surprise and smile. No, I mean it. Go look at yourself in the mirror. Say "Hi there" or something. Next Bride of Chucky, huh?

But here again, it is my projection, my bias that does most of the emotional interpretation. A wide-eyed smile may just mean someone had really bad plastic surgery. I'm fairly astute, so usually I can detect bad surgery, and usually I simply know what someone is feeling, even if the emotion is only found in the voice and eyes. It's why people think I listen well. I hear more than what is being said. I can *reflect* more than what is said. But like I wrote, it is like any language, and I just happen to be very good at language.

Ability, Bias, and Observations

If one is unable to learn a new language, unable to precisely imitate tones and inflections and body movements, one undoubtedly has limitations in the ability to interpret emotions. If one has no baseline understanding of how an individual or group expresses itself, then the context is askew.

In language, I best learn new words by closely watching how sounds are used. *Watching* how they are used. My bias: someone walks into a room, if she speaks, she is probably going to say "Hello" or some similar greeting. Next, she will probably be asked how she is, and she will answer.

Observation: if she answers the equivalent to "I'm well," her voice will rise, she will smile. If she is not so well, she will support the words with emotional consistency, maybe heavy, slumping shoulders, down turned eyes, and a cringe for "awful." One can interpret more complicated responses accordingly. When I learn language, I come to understand not only the tone and thus vocabulary and grammar but also the presentation. More importantly, to *mimic* the presentation. The context. I very rarely do a word-for-word translation, even in my native American English, and likely few of us did when we were first learning to speak. It is the sound of the whole thought or phrase and the trumpeter's performance that teaches the voice associated with the feeling, and *that* becomes the music to play.

Job

What am I, a poet? Back to my job as a voice-over artist. I passed over it fairly quickly, but did you know that you can hear a smile? "Smile" is a frequent instruction in the recording studio, particularly for commercials, jingles, or spoken word. Is it actually the case that the smile itself changes the voice? Very little. It is the emotion behind smile that does most of the work. It is what causes you to smile that changes the tone of voice. I can smile and sound like I am about to enter a boxing ring. But if I think of something that makes me happy ("be happy" probably being the more precise direction than "smile"), I can then sing or talk and sell to an unsuspecting audience the notion that the product I'm singing or talking about is making me happy.

I must say that the best voice-over audition I ever had, one for which I did *not* book the job, was for one of George Lucas's animated projects. I auditioned for the role of an intelligent, older, woman whose temper was frightening and mystical. Let's suppose the dialog was:

> That's fine, but if you ever again persist in disregarding my direct order, I shall see to it that you fall into the deepest trenches of Hell and burn for an eternity without one hope of escape that is not suffocated by the stench of your damnation.

Wow. That's pretty wicked. I'm sure George Lucas wrote a kinder, gentler, script . . . but let's get on with it.

A good performance has contrasts, conflict, and intrigue. "That's fine." Spoken with a gentle, flowing vocal projection, intended to give the audience a (false) sense of safety, ease, and relaxation. Eyes soft. "But if you ever again persist" begins with a steadily widening eye squint and a steady increase in vocal volume and tempo—interlaced with normal spoken variants—that peaks at *stench* and rumbles to a stop at *damnation*.

For age effect, the baseline voice grumbled deeply in my throat. The result is a line that reads as a growl with small barks interlaced, building until one major threatening bark that then falls to a growl to conclude the thought. Maintaining some restraint in vocal projection, even at the peak, builds the threat by making the speaker seem on the verge of full out attack at any moment. This means for the audience that as scary as the character may sound, the situation can get scarier. Deepening the tone without altering the pitch builds the threat similarly. The audience begins relieved, then becomes tense, then ends assured that the immediate threat has passed but remains on edge that danger still lingers. Mission accomplished! . . . except for the booking, but honestly, just because a performer doesn't book a job, it doesn't mean the audition wasn't stupendous. Isn't that right, George?

Even without hearing my voice, you have an imagining about what I would sound like asking, "Isn't that right, George?," what I would look like while I speak. Perhaps you envision me laughing or smiling, at least in the eyes. You have given me a voice without hearing me or seeing me, and the emotion you attribute is the real measure. Perhaps even more interesting is the emotion with which you respond to the projection.

It's the feeling that counts.