Psychotherapy Research, 2013
Vol. 23, No. 4, 381–393, http://dx.doi.org/10.1080/10503307.2013.784421

Routledge
Taylor & Francis Group

Examining in-session expressions of emotions with speech/vocal acoustic measures: An introductory guide

DANIEL ROCHMAN[1] & OFER AMIR[2]

[1]*Department of Psychiatry, University of Alberta, Edmonton, Alberta, Canada* & [2]*Department of Communication Disorders, Sheba Medical Center, Israel*

(Received 27 July 2012; revised 13 February 2013; accepted 5 March 2013)

Abstract
Emotion-sensitive speech/vocal acoustic measures are now available to study emotions expressed during psychotherapy sessions. This paper highlights the feasibility of employing such acoustic parameters alongside well-known self-report and observer ratings of emotions. The acoustic method is presented together with findings of the acoustic profiles of discrete emotions. The primary goal, however, is to introduce researchers to this measurement procedure: an introductory guide provides examples of how to generate an acoustical analysis of emotionally loaded vocal expressions. The procedures have been adapted to the study of emotions typically aroused in psychotherapy such as anger and sadness.

Keywords: acoustic analysis; emotions; psychotherapy

Harnessing emotion-related information is a challenging undertaking. Not surprisingly, researchers have employed a number of measurement procedures to study emotional functioning. Fortunately, most of these procedures have been successfully adapted to psychotherapy research despite methodological challenges. Psychotherapy studies often require time-sensitive and non-intrusive measures. For example, clinicians interested in examining in-session emotional arousal prefer to assess emotions continuously, on a moment-by-moment basis, without interfering with the expression of emotions (i.e., thereby elevating ecological validity).

One critical issue, not sufficiently addressed in psychotherapy studies, is that clinically inclined researchers typically employ only one, or at the best two, emotion-sensitive measures in any given study. As explained below, such a one- or bi-dimensional approach can limit study conclusions. The decision *not* to employ additional measures is, however, understandable. Many psychotherapy researchers simply prefer the well-established, traditional measures commonly applied to the context of therapy (e.g., observer rating of emotions). Others may be concerned that new technologies, especially measures offering objective correlates of emotions, are too time-consuming and labor-intensive. These concerns are not unjustified.

Despite such considerations, this paper will advocate for the integration of vocal acoustical parameters into psychotherapy research. We then address the usefulness and reliability of the acoustic analytic approach, as evidenced by prior research, and offer examples of acoustical analyses. Finally, an introductory guide was assembled to offer step-by-step explanations leading to the extraction of various acoustic parameters. The acoustic measures presented here, and their associated calculation procedures, have been adapted specifically to the study of in-session emotional expressions.

A Multi-method Approach to Emotion Research

Why consider a new measurement procedure to study emotions aroused during psychotherapy? Emotional activation can be described on a number of dimensions. For example, once an emotion has been aroused, subjective feelings, physiological functioning and a variety of behavioral expressions come into play. Naturally, each one of these distinct response systems can potentially provide emotion-related information. Because such response systems can function more or less independently, experts have concluded that studying emotions should ideally "involve measurement across (...) multiple

Correspondence concerning this article should be addressed to Daniel Rochman, University of Alberta, Department of Psychiatry, 1E3.26 WMC, 8440-112 St, Edmonton, Alberta, T6G 2B7 Canada. Email: drochman@ualberta.ca

components simultaneously" because each individual measure reflects only one aspect of the targeted emotion (Larsen & Prizmic-Larsen, 2006, p. 338; also see Sloan & Kring, 2007, and Brandly & Lang, 2000, for a detailed discussion on this methodological issue).

Interestingly, the various emotional response systems (or dimensions; e.g., cognitive appraisal, facial expressions and physiological) seem to interact in complex ways producing, at best, only mild between-measures convergence (Larsen & Prizmic-Larsen, 2006). In fact, research findings speak to the *discordance* between measures captured simultaneously during the arousal of an emotion (Bradly & Lang, 2000; Nesse et al., 1985). In other words, each individual measurement procedure can provide unique, non-redundant information. Clearly, a multi-dimensional approach can potentially increase knowledge about emotions and emotional processing.

Despite their interest in emotion processing, clinicians have yet to heed the call for a multi-dimensional assessment of emotions proposed by basic emotion researchers. Indeed, in any given study, researchers have typically employed a one-dimensional approach, capturing, for example, either observer ratings or self-reports of emotional arousal (e.g., Elliott, 1986; Giese-Davis, Piemme, Dillon, & Twirbutt, 2005; Goldman, 1997; Goldman, Greenberg, & Pos, 2005; Greenberg, 1979; Kagan, 1980; Merten, 2005; Rasting & Beutel, 2005; Sexton, Hembre, & Kvarme, 1996).

This paper addresses the feasibility of and presents an introductory guide to the employment of emotion-sensitive speech and vocal acoustic parameters in the context of psychotherapy research. The usefulness of this method lies in its relative non-invasiveness, affording the option to collect measures concurrently with other emotion-sensitive measures. Recording voice only requires the speaker to be seated facing a microphone or wear a headset microphone. In recent psychotherapy studies, speech and voice-related measures have been found helpful in the identification of subtle or transitory yet clinically relevant changes in emotional processing (e.g., see G. M. Diamond, Rochman, & Amir, 2010, and Rochman, Diamond, & Amir, 2008, for studies examining anger and sadness).

Once digitally recorded, a voice signal can provide an array of speech/acoustic parameters (see the "Data Segmentation Procedures" section below). Each parameter quantifies a specific physical property of the sound wave or a temporal characteristic of speech production. For example, fundamental frequency (F0) is a parameter closely correlated with subjectively perceived pitch of the voice, and amplitude is a measure correlated with perceived loudness. Other measures quantify hardly audible changes in the quality of voice production. Such is the case of jitter or Pitch Perturbation Quotient (PPQ), which reflects F0 perturbation, and shimmer or Amplitude Perturbation Quotient (APQ), which reflects amplitude perturbation. Other commonly used speech measures are speaking rate or articulation rate. A more elaborate explanation of these measures will be provided in the "Commonly Employed Emotion-Sensitive Speech/Acoustic Measures" section. More technical or detailed information is presented in Appendix A. To facilitate the readability of the following sections, Table I provides a brief overview of some emotion-sensitive acoustic/speech parameters. The first measures presented in Table I are F0, mean F0 (mF0) and F0 range. Each one of these measures is fairly correlated with an audible property of the voice. F0 is closely correlated with perceived pitch and mF0 is associated with average

Table I. Acoustic and speech parameters: definition and perceptual correlates

Measures	Definition	Description/perceptual correlate
Fundamental frequency (F0)	Quantifies the rate at which vocal folds vibrate	Pitch
Mean fundamental frequency (mF0)	Average F0 of a segment of voice or speech	Average pitch
F0 range	Maximum F0 minus minimum F0 within a segment of voice or speech	Changes (dynamics) in levels of pitch
Amplitude range	Quantifies variability in intensity within a voice or speech signal	Loudness range
Jitter/PPQ* (pitch perturbation quotient)	Cycle-to-cycle variability of F0	Increased values might be perceived as hoarseness
Shimmer/APQ* (amplitude perturbation quotient)	Cycle-to-cycle variability of amplitude	Increased values might be perceived as hoarseness
Speaking rate	Spoken units (i.e., phonemes, syllables or words) per unit of time	Speed of speech
Articulation rate	Spoken units (i.e., phonemes, syllables or words) per unit of time, within fluent speech	Speed of speech

**Note*: PPQ and APQ calculations are similar to those of jitter and shimmer, respectively, but involve averaging successive vocal cycles or a "smoothing factor."

pitch production. F0 range would be perceived as pitch variability. MF0, F0-range and amplitude range (i.e., perceived as changes in loudness) can be grouped under the term *voice dynamic* measures. Jitter or PPQ and shimmer or APQ are often referred to as *voice quality-related* acoustic measures. These perturbation parameters are not directly correlated with perceptual properties of the voice unless they vary dramatically (Baken, 1987). Finally, the last two measures, speaking-rate and articulation-rate, can be referred to as *temporal characteristics* of speech.

Emotion-Specificity of Speech/Vocal Acoustic Parameters: Prior Findings

Is it justified to use speech/acoustic measures in emotion research? A considerable number of studies examining speech/acoustic features of emotional expressions have been conducted since the beginning of the 1990s. Over a hundred such studies have been reviewed by Juslin and Laukka (2003), who concluded that findings "strongly suggest (. . .) that there are emotion-specific patterns of acoustic cues that can be used to communicate discrete emotions" (p. 799). Their review also suggested the possibility that studies failing to find emotion-specific acoustic profiles might have involved an insufficient number of acoustic parameters. Juslin and Laukka (2003) also suggested that perception of speech rate, vocal loudness, pitch and voice quality (see below for a more comprehensive description of these vocal properties) are the vocal cues that are most likely to convey and differentiate among emotions (i.e., anger, sadness, happiness, fear, or tenderness). In addition to examining *vocal* expressions of emotions, the authors reviewed the acoustic features of *musical* expressions of emotions. They concluded that "music performance uses largely the same emotion-specific patterns of acoustic cues as does vocal expression" (p. 797). In other words, "musicians communicate emotions to listeners on the basis of the principles of vocal expression of emotion" (p. 799). For example, just as speakers transmit anger by increasing speech rate and loudness variability, music performers do so by increasing tempo and intensity variability levels. The fact that two distinct modalities of expression (i.e., speech/vocal and musical) converge on a similar set of cues to convey the same emotion lends validity to the assumption that discrete emotions can be identified on the basis of their acoustic profiles.

Perhaps the most convincing support for the existence of emotion-specific acoustic profiles is that listeners can correctly identify a person's affective state based on their voice alone with significantly higher-than-chance accuracy. Such accuracy has been estimated at a 60%, a figure five-fold higher than the expected 12% accuracy if listeners were to base their judgment on guessing or chance factors alone (Scherer, 1982). Since listeners are able to identify emotions based on those emotions' verbal expression, one should be able to find the acoustic cues that help determine listeners' judgments. Some verbally expressed emotions are more accurately recognized than others. Sadness and anger are the best-recognized emotions followed by fear (Johnstone & Scherer, 2000). Listeners' judgments of emotions, as would be expected, are indeed associated with acoustic features of the voices they have heard. Evidence indicates that a substantial proportion of variance in such judgments can be explained by a set of nine or 10 acoustic properties (see Banse & Scherer, 1996). Total multiple correlation coefficients obtained by regressing listeners' ratings of emotions on acoustical parameters are, for example, $R = .63$, $p = < .001$ for (hot) anger, $R = .38$, $p = < .001$ for anxiety, $R = .49$, $p = < .001$ for sadness, $R = .27$ $p = < .01$ for happiness and $R = .38$ $p = < .001$ for shame (Johnstone & Scherer, 2000).

Is it justified to study the acoustic correlates of emotions aroused during psychotherapy? Overwhelmingly, most research on emotion-sensitive speech/vocal acoustic parameters has not involved the expression of emotions elicited under therapy or therapy-like conditions. In other words, participants in such studies were not led to access, experience and express *genuine emotions associated with personally meaningful events in an unrestricted manner* (i.e., as they would in a therapy session). Indeed, in most studies, researchers requested lay or professional actors to *portray* specific emotional expressions (see Scherer, Johnstone, & Klasmeyer, 2003, for a review). By contrast, emotions aroused during psychotherapy are very personal in nature and are associated with the speaker's past or current events. Experiencing such emotions is likely to activate a number of psychological processes (e.g., attention, mental effort, attempts to moderate emotions) that are not activated, or at least not activated to the same degree, when people portray emotions. This is important because such psychological processes can affect physiological functioning and, thereby, the mechanics of voice production. Consequently, the study of emotions would do well to employ emotion-induction procedures that better correspond to the process of therapy. While examining emotions in a setting approximating therapy is challenging in terms of experimental control, there is a need for more ecologically valid research.

Rochman et al. (2008) were the first to adapt an acoustic method to the study of genuine emotional expressions with the explicit goal of enhancing generalizability to the context of psychotherapy. Specifically, Rochman et al. conducted two studies to examine speech and vocal acoustic profiles of emotions aroused, experienced and expressed in a manner similar to that of a therapy session. In the first of these studies, emotions were evoked during a mood-induction procedure that, although standardized, resembled as closely as possible the conditions of a therapy session. The second study further approximated the conditions of a therapy session by having individuals participate in an interview to talk about and experience anger and sadness (i.e., this procedure is considered an analogue therapy session). The results of both these studies converged to show that specific speech and acoustic parameters can be employed to identify and distinguish between (unresolved) anger and sadness. Anger evoked an increase relative to non-emotional (baseline) speech in dynamic and speech-temporal related measures (i.e., mF0, F0 range and articulation rate). On the other hand, sadness evoked an increase from baseline in a more subtle, voice quality-related measure (i.e., F0 perturbation, measured by PPQ [see Table I]).

In a more recent study, also involving participants with feelings of unresolved anger, G. M. Diamond et al. (2010) further examined speech and vocal acoustic properties in the context of therapy. This study examined the emotional impact of two distinct interventions delivered during the course of an analogue therapy session (i.e., guiding participants to express emotions as they would during a therapy session). The interventions were the relational reframe (G. S. Diamond, 2005; G. S. Diamond & Siqueland, 1998) and empty-chair dialogue (Perls, Hefferline, & Goodman, 1951), and the goal was to identify between-intervention differences in participants' emotions as indicated by the acoustic parameters of their voice. Results showed that both interventions evoked an increase in sadness-sensitive voice quality-related parameters (e.g., PPQ and APQ) relative to participants' baseline speech. However, other speech/acoustic parameters changed from their baseline values exclusively during the empty-chair dialogue, suggesting that this intervention evoked an arousal of fear or anxiety, which was evident in F0, F0 range and speaking rate (i.e., voice dynamic and temporal-speech measures). The vocal changes associated with sadness were subtle or hardly audible and, on the other hand, anxiety or fear resulted in perceptually noticeable changes in participants' pitch, variability of pitch and speaking rate.

The studies by Rochman et al. (2008) and G. M. Diamond et al. (2010) illustrated the utility and feasibility of employing speech and vocal acoustic measures to identify and describe in-therapy emotional processing. Table II provides examples of emotionally charged verbal expressions and their corresponding speech/acoustic profiles. These verbal expressions were extracted from the G. M. Diamond et al. (2010) database.

Regarding Table II excerpts, the reader is reminded that the relational reframe intervention involves participants *talking about* their vulnerability and that, by contrast, the empty chair involves participants *speaking to* their significant others (i.e., in the first person, in an imaginal confrontation procedure). Table II excerpts were translated from the original Hebrew. Values of the speech and acoustic parameters express change scores from baseline, non-emotional speech. Note that the relational reframes generated slightly decreased habitual level of pitch (reflected in mF0 values), slightly increased range of pitch variability (captured by the F0 range), mild-to-moderate increase in voice quality-related parameters (i.e., hardly audible changes captured by increases in PPQ and APQ), and a mild decrease in voice/speech fluency (i.e., reflecting increased frequency of speech

Table II. Changes from baseline values in acoustical and temporal-speech parameters during Relational Reframe and Empty-chair stages

Participant	Stage	Excerpt	mF0	F0 range	PPQ	APQ	Speech disfluency
A	RR	We simply could not stand each other, I never felt I had to say anything . . .. It wouldn't have changed anything.	−6.00	15.80	1.64	2.17	3.47
	ECH	You always think you are right, and I'm wrong . . . so you always tell me: "why are you not strong like your sister?"	19.36	69.40	0.26	0.95	−1.35
B	RR	I never before realized how angry I was at my dad . . . and now I miss the relationship we used to have.	−19.47	3.02	1.65	0.60	5.58
	ECH	During all those years your behavior was not OK, and despite everything, I still love you . . . and you should have put into this relationship as much effort as I did. Why are you not interested in your own daughter?	33.14	88.95	−0.29	−0.50	1.50

Note: RR = relational reframe. ECH = empty-chair

interruptions). This acoustic/speech profile has been associated with a state of sadness and vulnerability (i.e., G. M. Diamond et al., 2010; Rochman et al., 2008). By contrast, the empty-chair intervention generated mild-to-moderate increases in habitual pitch, moderate increases in the range of pitch variability, small voice quality changes and inconsistent and minor speech fluency changes. The highest increases in habitual pitch and pitch variability (i.e., measured by mF0 and F0 range respectively) were found during participant B's empty chair. This acoustic pattern is consistent with a state of anxiety, which appears to be associated with distress and a sense of abandonment and being neglected/rejected by the significant other. Participant B also expressed frustration and a longing for increased intimacy (e.g., "I still love you . . . and you should have put into this relationship as much effort as I did").

In sum, emerging research suggests that it is feasible and useful to integrate speech/vocal acoustic parameters into psychotherapy research of emotions. Although this measurement technique is relatively novel to psychotherapy research, it has proven instrumental in the identification of sadness, anger, and anxiety/fear; three emotions often expressed during therapy sessions, and of and shifts between them.

Obtaining Speech/Vocal Acoustic Measures: An Introductory Guide

The following is an introductory guide designed to familiarize psychotherapy researchers with the procedure of obtaining emotion-sensitive speech and acoustic parameters. By the end of the guide, the reader should have a basic understanding of speech and vocal acoustic analysis, including the basic knowledge necessary to obtain preliminary results reflecting vocal/acoustic changes associated with emotional arousal.

Microphones and Environmental Conditions

Ensuring appropriate voice recording conditions is the starting point of any study designed to examine speech/acoustic parameters. The full range of technicalities associated with microphones and the recording environment will not be given full consideration here. Interested readers should refer to Švec and Granqvist (2010). The microphone can be either fixed in front of the speaker or, alternatively, worn on a headset. A headset microphone is preferred if the speaker is likely to physically shift positions or rotate their face while talking (as sometimes occurs during emotional expression). The headset ensures a constant mouth-to-microphone relative position (i.e., distance and relative angle) during the recording, which is necessary to ensure optimal signal quality (i.e., minimal ambient noise). If a headset microphone is not viable, the alternative is to place a fixed microphone in front of the speaker (i.e., placed on a table or on a tripod). In either case, when the distance between the speaker's mouth and the microphone is too small, there is an increased chance of "sound clipping" occurring, especially if voice production becomes too loud. Sound clipping results in a loss of acoustic information due to truncation of the sound waveform during recording (see Appendix A for more information on clipping).

Ideally, the recording room should enable a minimal degree of background interfering noise. Because sound-treated rooms are generally unavailable to psychotherapy researchers, a quiet room should suffice. However, ambient sound should be reduced to a minimum by eliminating all possible sources of noise (e.g., air-conditioning, computer ventilation). As noted, an additional way to reduce interfering sound is to employ a headset microphone.

Recording

Currently, most research studies record speech signal digitally via a DAT (digital audio tape) recorder or direct recording to a computer. It is worth noting that the commonplace compressed signal formats (e.g., MP3) are generally not suitable for voice research because they eliminate parts of the original signal, resulting in loss of data. To ensure that the original digital format is preserved, researchers are advised to use a WAV file or equivalent format. It is not advisable to conduct acoustic analysis based on past recording technologies. Such recording methods (e.g., analog tape recorder) do not typically meet research standards of quality (e.g., Gilloire & Vetterli, 1992).

Data Segmentation Procedures

Once the recorded speech/voice signal has been obtained, the raw data need to be segmented prior to further analysis. The segmentation procedure is central to the definition of units of speech analysis. Such units are important because it is on their basis that computers perform the calculations to obtain acoustic and speech-temporal parameters. Well-defined segmentation criteria are a prerequisite to ensure reliability/validity of parameters. In addition, the segmentation procedure needs to be suited to the measures of interest and the purpose of the study. For example, if the target acoustic measure is F0 (correlated with subjectively perceived pitch), the unit of analyses can be either vowels produced in

isolation, spliced vowels extracted from continuous speech, or a larger unit such as an entire sentence or an utterance (i.e., a linguistically meaningful unit of approximately three to 10 words).

Selecting a segmentation procedure requires a level of technical knowledge beyond the scope of this paper. However, for a number of reasons, unrestrictedly expressed, emotionally charged speech (as typically occurs in psychotherapy) calls for "utterances" as the unit of analysis (e.g., G. M. Diamond et al., 2010). An utterance can be defined as a string of consecutive words that (a) communicate an idea, (b) are bounded by a simple intonation contour, (c) are grammatically acceptable, and (d) contain at least three consecutive words or five syllables (Hall, Amir, & Yairi, 1999). Utterances can be easily demarcated in fluent speech with minimal training.

Once a speech signal has been segmented, each individual segment is submitted to a computerized acoustic analysis. While a large number of acoustic measures have been suggested in the literature, only a limited number of parameters will be presented below, as a representative list of emotion-sensitive measures (e.g., Banse & Scherer, 1996; Juslin & Laukka, 2001, 2003; Laukka et al., 2008; Rochman et al., 2008). For further information including more technical details, interested readers can refer to Baken (1987) or Juslin and Scherer (2005).

Commonly Employed Emotion-Sensitive Speech/Acoustic Measures

F0 and amplitude. The F0 of the voice is a measure of the vibratory rate of the speaker's vocal folds during phonation (e.g., voice production). This parameter quantifies the number of vibratory cycles per second and is therefore measured in Hz. When a speaker's average F0 increases, listeners perceive increased pitch. Hence, pitch is referred to as the perceptual correlate of F0. To clarify, while F0 is a computer-calculated parameter that captures a physical property of a sound signal, the term "pitch" describes its subjectively perceived correlate. Women's vocal folds, for example, typically vibrate more quickly than those of men, thus generating higher F0 values. Therefore, women are perceived as having higher levels of pitch.

To obtain F0 values, the speech segment under consideration must contain sound produced by the vibratory activity of the vocal folds. Because all vowels involve such vibratory activity, F0 measures are considered most reliable in acoustic analysis of isolated vowels. During the production of an isolated vowel, the vocal mechanism functions in a relatively stable manner, thus offering the opportunity to capture reliable and valid F0 measures. For example, Figure 1 represents the vowel /a/ pronounced in isolation by a male speaker (recorded duration was

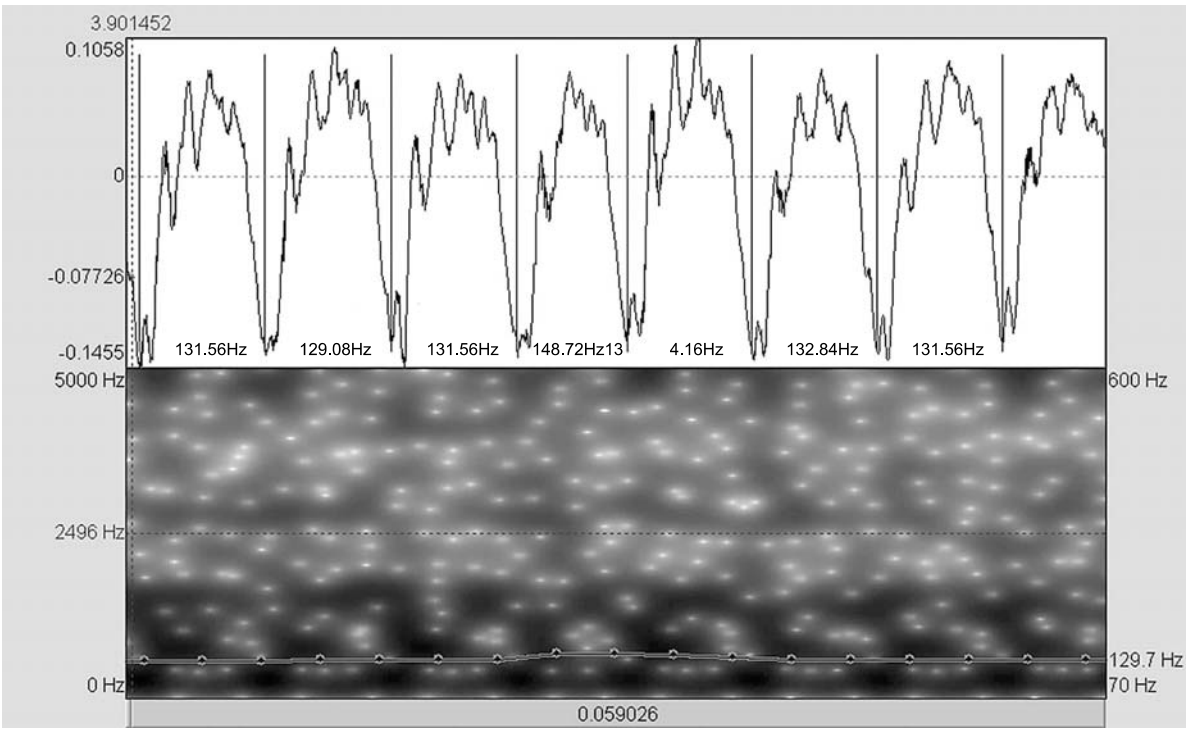

Figure 1. Sustained phonation of the /a/ vowel generated by a male speaker. Hertz (Hz, i.e., cycles per second) indicate F0 values for each vocal cycle.

59.026 ms; figure obtained using Praat software [Boersma & Weenink, 2010]). In this example, F0 values were calculated for each vibratory cycle and ranged from 129.08 to 148.72 Hz. Note that if the F0 values within this segment had been higher, the sound wave would appear "compressed," with a larger number of cycles being completed per unit of time. By contrast, a lower F0 would produce a "stretched" waveform, with each cycle spanning over a longer period of time. The lower section of the figure shows a dotted line representing a continuous measurement of F0 along the selected segment. In other words, this dotted line plots F0 values as a function of time across the segment.

As explained above, the faster the vocal folds vibrate, the higher the perceived level of pitch is. To obtain F0 values correlated with perceived pitch, a voice analyst would obtain mF0 values. This is done by averaging consecutive F0 values contained within the segment of voice/speech under consideration. In Figure 1, mF0 is 133.40 Hz, a value that is expected for men in general, and in Hebrew specifically (e.g., Baken, 1987; Most, Amir, & Tobin, 2000).

Amplitude is a physical measure of the intensity of the voice signal. It is calculated in decibels (dB) and correlates with the subjective perception of loudness (Fletcher & Munson, 1933). As air pressure provided by the lungs to the voice mechanism increases, so does the level of loudness and vocal effort. In emotion research, the measure of interest is amplitude range, which captures changes in voice signal intensity over the course of a segment of speech. Amplitude values are sensitive to a number of factors such as variability in mouth-to-microphone distance and angle, recording input levels and digitizing settings (Baken, 1987). Therefore, researchers should hold such variables constant throughout the recording session. For most purposes, the measure of interest is a *relative* change in amplitude values. A*bsolute* measures of amplitude require strict recording conditions and meticulous calibration procedures, the specifics of which will not be discussed. Appendix A illustrates how to calculate amplitude-range values.

Dynamic changes in F0 and amplitude have been linked to emotional arousal. Such dynamic changes are perceived as changes in a speaker's pitch and loudness, respectively. These changes are captured by F0 range and amplitude range.

Speaking rate and articulation rate. The calculation of speaking rate and of articulation rate involves counting the number of spoken units (i.e., phonemes, syllables or words) contained in a segment of speech, and then dividing that amount by its duration. The difference between these two measures is that while speaking rate is calculated on the basis of speech segments that *could* include pauses and other speech disruptions (Costello & Ingham, 1984; Howell, Au-yeung, & Pilgrim, 1999), articulation-rate calculations exclude such disfluencies (see Scherer, 1982, p. 151 for more detailed information). Of these two measures, speaking rate tends to better correlate with emotions. Indeed, it has been found that emotional arousal typically disrupts the stream of speech by introducing disfluencies, theoretically due to hesitation or other forms of cognitive effort (Bartolic, Basso, Schefft, Glauser, & Titanic-Schefft, 1999). Therefore, speaking rate is better suited for studies of emotions in psychotherapy (G. M. Diamond et al., 2010).

Frequency perturbation. Frequency-perturbation measures capture "micro" variations in F0, measured between successive voicing cycles (Baken, 1987). Typically, to obtain this measure a speaker is asked to produce a vowel as steadily as possible. The recorded signal is then inspected and F0 values are calculated for each cycle of a specified segment of the vowel (typically, the segment containing the steadiest expression of the vowel). F0 values will vary from cycle to cycle, and such variability, sometimes called *jitter*, is caused by various neurological, biomechanical, aerodynamic and acoustic sources (Titze, 2000). For any given sequence of voicing cycles, jitter reflects the deviation of F0 values relative to the average F0 across those cycles. In other words, *jitter* is the degree of F0 *un*steadiness within a voiced signal. The actual applicability of jitter to psychotherapy studies is, however, limited because asking a speaker to produce a steady phonation once an emotion has been aroused typically leads to a quick change in the speaker's emotional state (i.e., even before his/her voice can be recorded). By contrast, PPQ was found to be more suitable for psychotherapy because it allows for capturing F0 perturbation within samples of speech unrestrictedly produced (G. M. Diamond et al., 2010; Rochman et al., 2008). PPQ accounts more appropriately than *jitter* for prosody and other supra-segmental speech features that might artificially inflate F0 perturbation values. In other words, PPQ compensates for possible overestimation of F0 perturbation values. This compensation is accomplished by using a smoothing factor (i.e., typically of three or five factors, using a moving window technique; Baken, 1987), to render a more reliable measure.

Amplitude perturbation. Amplitude perturbation is defined as "micro" (i.e., cycle-to-cycle) variations in signal amplitude. Similar to frequency

perturbation, this measure was best applied for voice produced during a sustained vowel. In that case, the measure is called *shimmer*, and involves measuring each cycle's amplitude, identifying differences in amplitude between successive cycles, and then comparing these differences to the mean amplitude of the segment under consideration (Baken, 1987). Similarly to the procedure described for frequency perturbation, introducing a smoothing factor is the preferred calculation procedure when unrestrictedly generated (i.e., connected) speech is under consideration (Baken, 1987). APQ employs such a smoothing factor on eleven successive voice cycles.

Speech/Acoustic Profiles of Specific Emotions

Three discrete negative emotions are often aroused and expressed during psychotherapy sessions: anger, sadness and anxiety/fear (e.g., Greenberg & Pascual-Leone, 2006; Lieberman & Goldstein, 2006; Magnavita, 2006; Paivio & Shimp, 1998). Not surprisingly, these same emotions are associated with and perpetuate a number of psychological disorders (e.g., Apter et al., 1990; Beck, 1971; Jenkins & Oatley, 2000). Congruently, clinicians are often interested in emotion-related questions (e.g., Greenberg & Pascual-Leone, 2006; Sloan & Kring, 2007), the answers to which require emotion-sensitive measurement procedures. The following sections summarize speech/vocal acoustic properties of anger, sadness and anxiety/fear as per data most pertinent to psychotherapy (i.e., results most generalizable to unrestrictedly produced, connected speech of genuine emotional expressions).

Anger has been found to evoke increases relative to non-emotional speech in articulation rates (measured by words per minute; WPM), mF0, F0 range and amplitude range, with the first three of these parameters showing the most robust sensitivity to anger (Rochman et al., 2008). Prior (non-therapy) related literature has also associated anger with increased articulation rate, mF0 and/or voice dynamic parameters (Banse & Scherer, 1996; Scherer et al., 2003).

Sadness evokes increases relative to non-emotional speech in frequency perturbation measures such as PPQ (Rochman et al., 2008) and APQ, and decreases in speech-rate (G. M. Diamond et al., 2010). Prior literature corroborates these findings (Bartolic et al., 1999; Juslin & Laukka, 2001; Scherer, Banse, Wallbott, & Goldbeck, 1991). In addition, sadness has been associated with decreases in mF0, F0 range and amplitude range relative to non-emotional speech (Scherer et al., 2003; see Juslin & Laukka, 2003, for a review).

Finally, expressions of anxiety/fear, particularly those associated with feelings of vulnerability expressed in the context of therapy evoke increased mF0, F0 range, amplitude range, APQ and speech rate relative to non-emotional speech (G. M. Diamond et al., 2010). In prior research, anxiety/fear was also associated with increased mF0, but was linked with decreases rather than with increases of mF0 (an inconsistency perhaps attributable to differences in emotion-elicitation procedures across studies; Banse & Sherer, 1996; Juslin & Laukka, 2001). In a study involving genuine expression of anxiety/fear expressed by individuals suffering from social phobia, findings correlated increases in anxiety with increases in mean and maximum F0 and increases in the frequency of pauses in speech (Laukka et al., 2008).

Concluding Remarks

The goal of this paper was to introduce psychotherapy researchers to the basics of speech and voice analysis. The measures presented are speech/voice measures that reliably change from baseline levels (i.e., non-emotional levels) during the expression of specific emotions. As shown in prior psychotherapy studies, these speech and acoustic parameters can help understand the efficacy of emotion-focused interventions, thereby providing additional insight to the process of shifting between or regulating specific emotions (G. M. Diamond et al., 2010; Rochman et al., 2008).

On this same note, researchers are advised that speech and acoustic features of voice do not provide, in and of themselves, conclusive evidence that a certain emotion was aroused. As addressed in the introductory sections, emotions are, conceptually, multi-component by nature, with each individual measurement procedure potentially providing unique, non-redundant information. Accordingly, it is best to employ speech and voice parameters in conjunction with other indications of emotions, including, for example, observer rating of emotional expressions (to evaluate concurrent/divergent validity).

Self-report measures of emotional intensity can corroborate an individual's emotional state at the time of recording of speech. However, the past psychotherapy studies suggest that such measures do not correlate well with specific acoustic/speech measures (see Rochman et al., 2008, for more information). One reason for such poor correlation may be that self-report measures are necessarily obtained retrospectively, after the end of a session. It seems that when asked to rate the level of their own in-session emotional intensity, participants

provide a global score, representing an "averaged" emotional intensity. By contrast, acoustic measures capture changes associated with more transitory emotional states.

The calculation procedures necessary to obtain the various speech/acoustic measures were presented as introductory in this manuscript (see Appendix A for more information). The following example clarifies how to apply the speech/acoustic method to a hypothetical psychotherapy study: a researcher surmises that a specific intervention delivered during the course of a therapy session regularly evokes feelings of sadness, the acknowledgment and expression of which are known to lead to a positive, curative psychological process. Consequently, the researcher conducts a study to identify speech/vocal acoustic parameters of expressions that follow the sadness-inducing intervention. He or she would record the voice of a patient exposed to the intervention and a coder would then be instructed to identify the segment of speech of interest following establish criteria (ideally, the segment would last approximately a minute). The coder-identified sound file would be segmented into utterances and, for comparison, an additional baseline, non-emotional sound file would be obtained (possibly from the beginning of the interview, assuming that the patient was not expressing an emotion at that point in time). The baseline file would also be segmented into utterances. Thereafter, voice dynamic parameters such as mF0, F0 range and amplitude range would be calculated for each utterance. Speaking-rate measures would be calculated by counting syllables or words contained within each utterance and dividing that amount by the duration of the utterance being examined. Finally, perturbation measures (e.g., PPQ, APQ) would be extracted from within voiced sections of the utterance (i.e., sections for which F0 values were reliably obtained). Following averaging procedures, parameters would represent the speech/voice profiles corresponding to baseline and sadness, and discrepancies between these two conditions would be statistically evaluated for significance.

A general comment regarding the speech/acoustic method is that its application to the study of emotions expressed during therapy seems worthwhile. Although obtaining reliable speech/vocal acoustic measures requires considerable expertise, this manuscript addresses most basic difficulties a psychotherapy researcher may encounter. More specific technical information regarding parameter calculation procedures are included in Appendix A. At the same time, scholars should be aware that analyzing naturally occurring continuous (i.e., unrestricted) speech is, at this stage, cutting-edge research. Indeed, past research has relied almost exclusively on isolated vowels to identify emotion-related variability in speech/voice analytic measures. This implies that certain calculations of acoustic parameters need to be adapted to the typical variability of acoustic parameters in naturally occurring, continuous speech. This review has attempted to cover most basic knowledge required to approach a preliminary acoustical analysis. However, psychotherapy researchers are encouraged to consult or collaborate with speech scientists or speech/voice-pathology experts to corroborate their findings.

Acknowledgment

This manuscript was enhanced by the comments of Dr. Anthony S. Joyce and Dr. Lisa R. Rochman. The first author is grateful to Dianne and Irving Kipnes for supporting his post-doctoral fellowship at the Department of Psychiatry, University of Alberta.

References

Amir, O., Wolf, M., & Amir, N. (2009). A clinical comparison between two acoustic analysis softwares: MDVP and Praat. *Biomedical Signal Processing and Control, 4*, 202–205. doi: 10.1016/j.bspc.2008.11.002

Apter, A., van Praag, H.M., Plutchik, R., Sevy, S., Korn, M., & Brown, S. (1990). Interrelations among anxiety, aggression, impulsivity and mood: A serotonergically linked cluster? *Psychiatry Research, 32*, 191–199. doi:10.1016/0165-1781(90)90086-K

Baken, R.J. (1987). *Clinical measurement of speech and voice.* Boston, MA: College-Hill Press.

Banse, R., & Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*, 614–636. doi:10.1037/0022-3514.70.3.614

Bartolic, E. I., Basso, M. R., Schefft, B. K., Glauser, T., & Titanic-Schefft, M. (1999). Effects of experimentally-induced emotional states on frontal lobe cognitive task performance. *Neuropsychologia, 37*, 677–683. doi:10.1016/S0028-3932(98)00123-7

Beck, A.T. (1971). Cognition, affect, and psychopathology. *Archives of General Psychiatry, 24*, 495–500. doi:10.1001/archpsyc.1971.01750120011002

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer [Computer program]. Version 5.2.06. http://www.praat.org/.

Bradley, M.M., & Lang, P.J. (2000). Measuring emotion: Behavior, feeling and physiology. In R.D. Lane & L. Nadel (Eds.), *Cognitive neuroscience of emotion* (pp. 242–276). New York: Oxford University Press.

Costello, J.M., & Ingham, R.J. (1984). Assessment strategies for stuttering. In R. Curlee & W.H. Perkins (Eds.), *Nature and treatment of stuttering: New directions* (pp. 303–333). San Diego, CA: College-Hill Press.

Diamond, G.M., Rochman, D., & Amir, O. (2010). Arousing primary vulnerable emotions in the context of Unresolved Anger: "Speaking about" versus "Speaking to". *Journal of Counseling Psychology, 57*, 402–410. doi:10.1037/a0021115

Diamond, G.S. (2005). Attachment-based family therapy for depressed and anxious adolescents. In J.L. Lebow (Ed.),

Handbook of clinical family therapy (pp. 17–41). Hoboken, NJ: Wiley.

Diamond, G.S., & Siqueland, L. (1998). Emotions, attachment and the relational reframe: The first session. *Journal of Systemic Therapies*, *17*, 36–50.

Elliott, R. (1986). Interpersonal process recall (IPR) as a psychotherapy process research method. In L.S. Greenberg & W. Pinsof (Eds.), *The psychotherapeutic process* (pp. 503–527). New York: Guilford.

Fletcher, H., & Munson, W.A. (1933). Loudness, its definition, measurement, and calculation. *Journal of the Acoustical Society of America*, *5*, 82–108. doi:10.1121/1.1915637

Giese-Davis, J., Piemme, K.A., Dillon, C., & Twirbutt, S. (2005). Macrovariables in affective expression in women with breast cancer participating in support groups. In J.A. Harrigan, R. Rosenthal, & K.R. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 399–445). New York: Oxford University Press.

Gilloire, A., & Vetterli, M. (1992). Adaptive filtering in subbands with critical sampling: Analysis, experiment, and application to acoustic echo cancellation. *IEEE Transactions on Signal Processing*, *40*(8), 1862–1875. doi:10.1109/78.149989

Goldman, R.N. (1997). *Theme-related depth of experiencing and change in experiential psychotherapy with depressed clients*. Unpublished doctoral dissertation, York University, Toronto, Ontario, Canada.

Goldman, R.N., Greenberg, L.S., & Pos, A.E. (2005). Depth of emotional experience and outcome. *Psychotherapy Research*, *15*, 238–249. doi:10.1080/10503300512331385188

Greenberg, L.S. (1979). Resolving splits: Use of the two chair technique. *Psychotherapy: Theory, Research & Practice*, *16*, 316–324. doi:10.1002/jclp.20252

Greenberg, L.S., & Pascual-Leone, A. (2006). Emotion in psychotherapy: A practice-friendly research review. *Journal of Clinical Psychology*, *62*, 611–630. doi:10.1002/jclp.20252

Hall, K.D., Amir, O., & Yairi, E. (1999). A longitudinal investigation of speaking rate in preschool children who stutter. *Journal of Speech, Language, and Hearing Research*, *42*, 1367–1377.

Howell, P., Au-Yeung, J., & Pilgrim, L. (1999). Utterance rate and linguisticproperties as determinants of speech dysfluency in children who stutter. *Journal of the Acoustical Society of America*, *105*, 481–490. doi:10.1121/1.424585

Jenkins, J.M., & Oatley, K. (2000). Psychopathology and short-term emotion: The balance of affects. *Journal of Child Psychology and Psychiatry*, *41*, 463–472. doi:10.1111/1469-7610.00631

Johnstone, T., & Scherer, K.R. (2000). Vocal communication of emotion. In M. Lewis & J. Haviland (Eds.), *The handbook of emotion* (pp. 220–235). New York: Guilford.

Juslin, P.N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, *1*, 381–412. doi:10.1037/1528-3542.1.4.381

Juslin, P.N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*, 770–814. doi:10.1037/0033-2909.129.5.770

Juslin, P.N., & Scherer, K.R. (2005). Vocal expression of affect. In J.A. Harrigan, R. Rosenthal, & K.R. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 65–135). New York: Oxford University Press.

Kagan, N. (1980). Influencing human interaction-eighteen years with IPR. In A.K. Hess (Ed.), *Psychotherapy supervision: Theory, research and practice* (pp. 262–286). New York: Wiley.

Larsen, R.J., & Prizmic-Larsen, Z. (2006). Measuring emotions: Implications of a multimethod perspective. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 337–351). Washington DC: American Psychological Association.

Laukka, P., Linnman, C., Ahs, F., Pissiota, A., Frans, O., et al. (2008). In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech. *Journal of Nonverbal Behavior*, *32*, 195–214. doi:10.1007/s10919-008-0055-9

Lieberman, M.A., & Goldstein, B.A. (2006). Not all negative emotions are equal: The role of emotional expression in online support groups for women with breast cancer. *Psycho-Oncology*, *15*, 160–168. doi:10.1002/pon.932

Magnavita, J.J. (2006). The centrality of emotion in unifying and accelerating psychotherapy. *Journal of Clinical Psychology*, *62*, 585–596. doi:10.1002/jclp.20250

Merten, J. (2005). Facial microbehavior and the emotional quality of the therapeutic relationship. *Psychotherapy Research*, *15*, 325–333. doi:10.1080/10503300500091272

Most, T., Amir, O., & Tobin, Y. (2000). The Hebrew vowel system: Raw and normalized acoustic data. *Language and Speech*, *43*, 295–308. doi:10.1177/00238309000430030401

Nesse, R.N., Curtis, G.C., Thyer, B.A., McCann, D.S., Huber-Smith, M., & Knopf, R.F. (1985). Endocrine and cardiovascular responses during phobic anxiety. *Psychosomatic Medicine*, *47*, 320–332.

Paivio, S.C., & Shimp, L.N. (1998). Affective change processes in therapy for PTSD stemming from childhood abuse. *Journal of Psychotherapy Integration*, *8*, 211–209. doi:10.1023/A:1023265103791

Perls, F., Hefferline, R., & Goodman, P. (1951). *Gestalt therapy*. New York: Delta.

Rasting, M., & Beutel, M.E. (2005). Dyadic affective interactive patterns in the intake interview as a predictor of outcome. *Psychotherapy Research*, *15*, 188–198. doi:10.1080/10503300512331335039

Rochman, D., Diamond, G.M., & Amir, O. (2008). Unresolved anger and sadness: Identifying vocal acoustical correlates. *Journal of Counseling Psychology*, *55*, 505–517. doi:10.1037/a0013720

Scherer, K.R. (1982). Methods of research on vocal communication: Paradigms and parameters. In K.R. Scherer & P. Ekman (Eds.), *Handbook of methods in nonverbal behavior research* (pp. 136–198). Cambridge: Cambridge University Press.

Scherer, K.R., Banse, R., Wallbott, H.G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, *15*, 123–148. doi:10.1007/BF00995674

Scherer, K.R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R.J. Davidson, K.R. Scherer, & H.H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 433–456). Oxford: Oxford University Press.

Sexton, H.C., Hembre, K., & Kvarme, G. (1996). The interaction of the alliance and therapy microprocess: a sequential analysis. *Journal of Consulting and Clinical Psychology*, *64*, 471–480. doi:10.1037/0022-006X.64.3.471

Sloan, D., & Kring, A.M. (2007). Measuring changes in emotion during psychotherapy: conceptual and methodological issues. *Clinical Psychology: Science and Practice*, *14*, 307–322. doi:10.1111/j.1468-2850.2007.00092.x

Švec, J.G., & Granqvist, S. (2010). Guidelines for selecting microphones for human voice production research. *American Journal of Speech-Language Pathology*, *19*, 356–368. doi:10.1044/1058-0360(2010/09-0091)

Titze, I.R. (2000). *Principles of voice production* (second printing). Iowa City, IA: National Center for Voice and Speech.

Appendix A

Employing speech and acoustic analysis to examine emotional experiences is a novel technique in the context of psychotherapy. Psychotherapy researchers approaching this method should familiarize themselves with some technical details necessary to perform reliable calculations. As pointed out in the "Concluding Remarks" section, one recommendation is to recruit a speech analysis expert as a collaborator before embarking on a full-scale project. The following guide is a non-exhaustive summary of relevant concepts and measurement procedures. The goal is to help psychotherapy researchers gain proficiency in the jargon typically employed in vocal acoustic analysis.

Recording Speech

Software settings need to be properly adjusted to capture voice before a recording session begins. Softwares typically offer the option of sampling (i.e., recording) at a number of "rates." This rate is the frequency at which data are sampled per unit of time. For vocal acoustic research purposes, the sampling rate should be set at 44.1 or 48.0 kHz (i.e., this is particularly important for the calculation of voice quality-related perturbation measures). Lower sampling rates compromise the reliability of acoustic measures. Note that, however, on the other hand, increasing the sampling rate beyond 48.0 kHz does not further increase measurement accuracy and unnecessarily taxes computer memory capacity.

Clipping of the voice signal, a problem mentioned in this paper, is a condition in which the amplitude (i.e., perceptually, loudness) of the original voice signal exceeds the dynamic range of the recording equipment. Clipping is likely to occur during loud phonation or when the distance between the microphone and the speaker's mouth is too small. The consequence of clipping is that sections of the original signal are overlooked and not properly recorded. The original signal is distorted and its acoustic properties cannot be restored. It is important to identify clipping and adjust the recording settings in order to prevent this problem.

Basic Calculation Procedures

As explained in this paper, F0 serves as a basis of other measures of interest (i.e., PPQ, mF0 and F0 range). An important caveat applies to the interpretation of F0: computer-based algorithms can occasionally render inaccurate F0 values. A frequent and significant source of error is an "octave error," which results in an over- or underestimation of the true F0 value by a factor of two. This miscalculation occurs when the software fails to accurately detect the beginning and ending points of individual F0 cycles. To eliminate octave errors the analyst needs to visually monitor the contour of F0 values plotted against time and manually correct the analysis. Automatic corrections are possible to some extent, by setting lower and upper threshold values for F0 in the software.

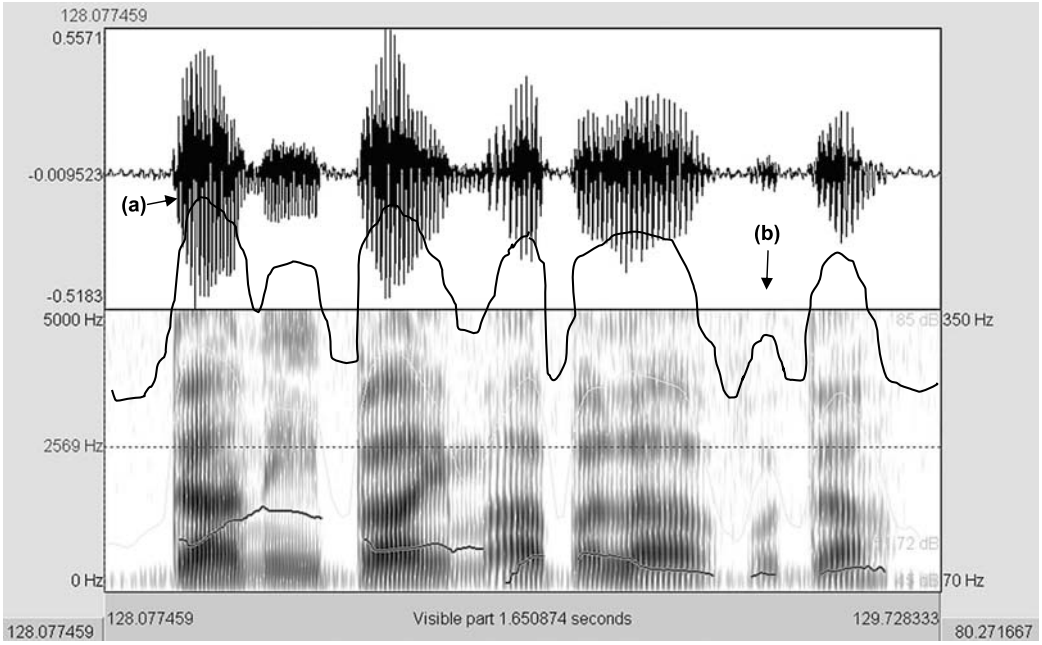

Figure 1A. Sustained phonation of the /a/ vowel, generated by a male speaker

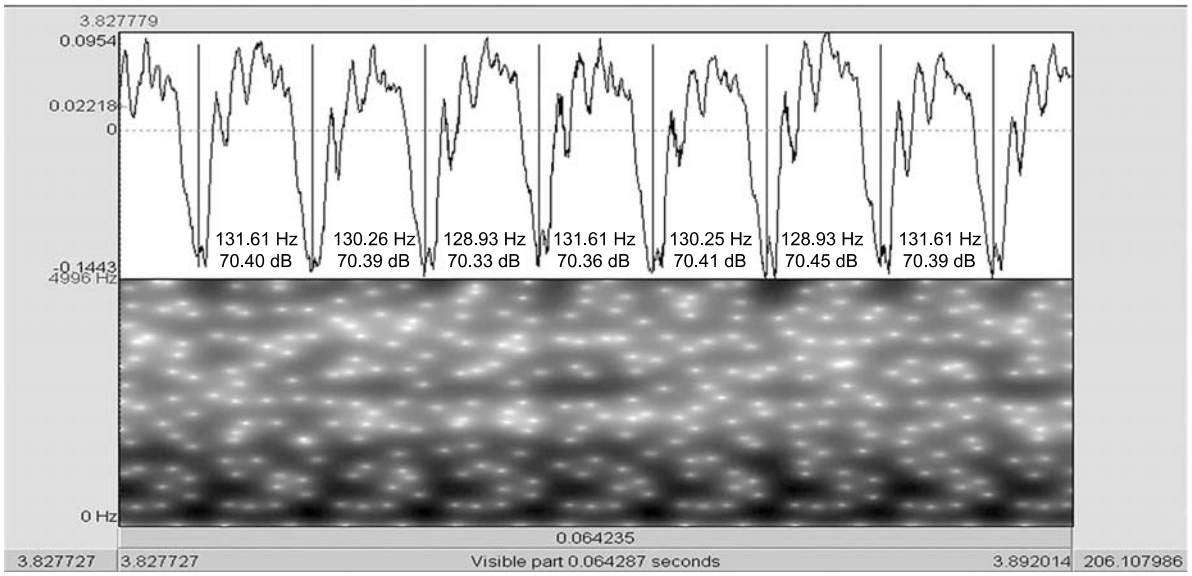

Figure 2A. A time-wave and a wide-band pectrographic display of a sustained production of the /a/ vowel produced by a male speaker. Fundamental frequency and amplitude values were indicated for each vocal cycle

This figure presents a waveform corresponding to a segment of speech. Note the thick line in the lower section of the figure, representing the Amplitude tracking display. In order to calculate amplitude-range, the minimum Amplitude value needs to be subtracted.

Acoustic analysis programs also display amplitude values. The measure of interest associated with amplitude is its dynamic variability or amplitude range (i.e., perceptually, changes in voice loudness). Figure 1A illustrates how to obtain amplitude range. The first step is to obtain the maximum amplitude value. To that end the cursor is placed on the highest point of the amplitude curve (point "a" in Figure 1A). The second step is to obtain the minimum amplitude value. This is done by placing the cursor on the lowest point (point "b"). In Figure 1A example, the maximum value is 80.01 dB, and the

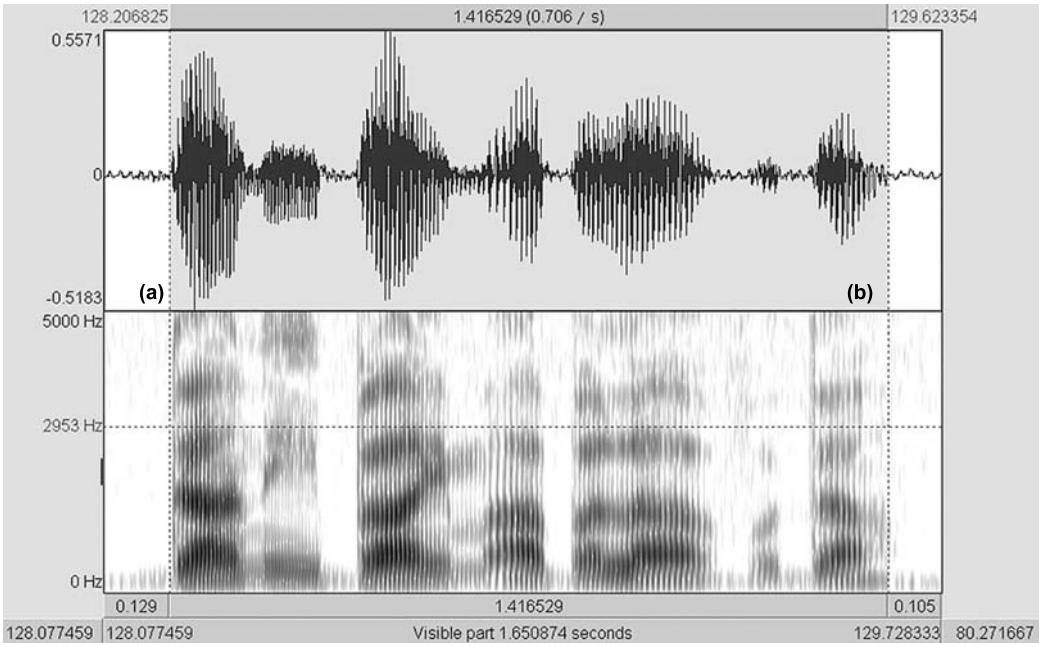

Figure 3A. A wide-band spectrographic and a time-wave display of a sentence

lowest is 62.73 dB, therefore amplitude range is 17.28 dB.

The F0-range calculation procedure is similar to that of amplitude range. To calculate F0 range, an acoustic analyst obtains the highest and lowest F0 values of a voice segment of interest, and then subtracts the lowest value from the highest.

Figure 2A was included to explain basic calculation procedures necessary to obtain frequency perturbation and amplitude perturbation values. Specifically, Figure 2A displays a 64 ms segment corresponding to the sustained production of the /a/ vowel produced by a male speaker. The vertical lines on the time-wave display mark the boundaries of each individual voicing cycle and the F0 and amplitude measures are presented at the bottom of the time-wave display for each individual voicing cycle in Hz and dB respectively. Notice that values present slight changes from cycle to cycle. Perturbation measures quantify this cycle-to-cycle variability around the average of the measure within the segment under consideration. For example, in Figure 2A, jitter was used as a measure of F0 perturbation and resulted in a value of 0.376%, which is considered within the expected normal range (e.g., Amir, Wolf, & Amir, 2009). Shimmer was employed as a measure of amplitude perturbation and rendered a value of 8.974%.

Figure 3A has been included to illustrate the measurement procedure for speaking rate, in this case expressed as number of syllables per second. The sentence under consideration in this example is / dani ba im aba laavoda/ (i.e., Hebrew for "Danny came to work with dad"). To calculate speaking rate the duration of the segment needs to be calculated. This is done by placing the cursor at the beginning and ending points, respectively (i.e., points "a" and "b") and subtracting beginning point in the timeline from its ending point. In this example, the duration of the sentences is 1.416529 seconds. Because the sentence consists of ten syllables, its speaking-rate value is 7.056 syllables/s (i.e., 10/1.416529).