Vocal Expression and Perception of Emotion

Jo-Anne Bachorowski[1]

Department of Psychology, Vanderbilt University, Nashville, Tennessee

Abstract

Speech is an acoustically rich signal that provides considerable personal information about talkers. The expression of emotions in speech sounds and corresponding abilities to perceive such emotions are both fundamental aspects of human communication. Findings from studies seeking to characterize the acoustic properties of emotional speech indicate that speech acoustics provide an external cue to the level of nonspecific arousal associated with emotional processes and, to a lesser extent, the relative pleasantness of experienced emotions. Outcomes from perceptual tests show that listeners are able to accurately judge emotions from speech at rates far greater than expected by chance. More detailed characterizations of these production and perception aspects of vocal communication will necessarily involve knowledge about differences among talkers, such as those components of speech that provide comparatively stable cues to individual talkers' identities.

Keywords

emotion; speech acoustics; vocal communication

The speech stream is a highly complex and variable signal that is most directly studied by analyzing its acoustic properties, or sound patterns. We know from everyday experience that talkers provide information about their emotional states through the acoustic properties of their speech. For instance, many of us have experienced talking in an unwittingly loud voice when feeling gleeful, speaking in an uncharacteristically high-pitched voice when greeting a sexually desirable person, or talking with marked vocal tremor while giving a public speech. In turn, listeners are seemingly adept at making accurate evaluations of emotional states—even in the absence of visual cues, as routinely occurs during telephone conversations. Production and perception phenomena are both facets of a broad research area concerned with understanding the ways in which speech acoustics provide personal information about talkers, such as gender and individual identity, independent of linguistic content. This article provides an overview of the links between speech acoustics and emotions (for more detailed reviews, see Pittam & Scherer, 1993, and Scherer, 1989). Some limitations of traditional approaches to this research area, and alternative ways of thinking about enduring problems, are also discussed.

SPEECH ACOUSTICS

The source-filter theory of speech production is helpful for understanding the ways in which speech acoustics might provide information about emotional state (see Kent, 1997, for a thorough introduction to speech acoustics). In this framework, speech sounds result from the combination of source energy, produced by vibration of the vocal folds (formerly referred to as the vocal cords), and the subsequent filtering of that energy by the vocal tract above the larynx.

Source-related acoustic cues refer to those aspects of speech sounds that are primarily associated with vocal-fold vibration. In emotions research, measures associated with F_0 (i.e., the fundamental frequency of speech, which corresponds to the rate of vocal-fold vibration and is perceived as vocal pitch) are the most commonly used. Other potentially important source measures include *jitter* and *shimmer,* which correspond to variability in the frequency and amplitude of vocal-fold vibration, respectively. Filter-related cues are examined less often by emotions researchers. However, these cues may be important for understanding emotional speech because facial expression (e.g., lip position) can influence filtering effects. Thus, a sentence spoken while smiling can sound different from the same sentence spoken while frowning. These kinds of acoustic differences are reflected in formants, which are vocal-tract resonances that correspond to the frequencies amplified through vocal-tract filtering. Another way of thinking about resonances is that they are the natural frequencies that are selectively reinforced because of the size and shape of the vocal tract (see Kent, 1997). Both source- and filter-related cues are sensitive to changes in vocal-production-related physiology, such as the fluctuations in respiration and muscle tension that can occur in conjunction with some emotions (Scherer, 1989).

VOCAL EMOTION FROM A PRODUCTION STANDPOINT

Most production-related investigations have been guided by the assumption that distinct patterns of acoustic cues will be found to be associated with discrete emotional states. Largely for practical reasons, these investigations have typically analyzed the emotional speech produced by small numbers of actors or naive subjects asked to portray various emotions while producing linguistically neutral utterances. For both theoretical and practical reasons, most analyses of emotional speech have focused on source-related acoustic cues. For these cues, a restricted yet fairly reliable pattern of findings has emerged. For example, Scherer, Banse, Wallbott, and Goldbeck (1991; also see Banse & Scherer, 1996; Leinonen, Hiltunen, Linnankoski, & Laakso, 1997) examined the acoustic features of neutral and emotional nonsense sentences spoken by four actors. In comparison with neutral speech, portrayals of fear, joy, and anger were each associated with a higher mean F_0, whereas portrayals of sadness were associated with a lower mean F_0. A corresponding pattern was observed for vocal intensity, or amplitude.

Across studies, portrayals of emotions associated with high levels of physiological arousal (e.g., anger, fear, anxiety, and joy) have been associated with increases in mean F_0, F_0 variability, and vocal intensity. Some acoustic differentiation among these emotions has been found by examining F_0 contours, or the pattern of F_0 changes over the course of an utterance. For example, F_0 has been noted to decrease over time during portrayals of anger, but to increase over time during portrayals of joy. In contrast, emotions associated with low levels of physiological arousal (e.g., sadness) are consistently associated with lower mean F_0, F_0 variability, and vocal intensity, as well as decreases in F_0 over time.

Rather than relying on acted portrayals, my colleague and I have studied the acoustic properties of natural speech produced by naive participants in the context of controlled emotion-induction procedures.[2] We have focused our acoustic analysis on very short vowel segments both because detailed measurement of source- and filter-related cues is possible with these sounds and because these speech samples are less likely than sentence-length utterances to be influenced by demand characteristics, such as culturally prescribed rules about how particular emotions ought to be conveyed in speech. In one such study (Bachorowski & Owren, 1995), positive and negative emotions were induced by giving participants "Good Job" and "Try Harder" feedback as they performed a difficult computerized spelling task. In reality, this feedback was not contingent on participants' performance. After each feedback presentation, subjects' speech was recorded as they announced the number (*n*) of the upcoming block of trials using the phrase "test *n* test." F_0, jitter, and shimmer were measured from 30 instances of the "eh" sound that occurred in the first utterance of "test" in each stock phrase. The results indicated that emotion expressed through the vocal channel depended not only on the valence (i.e., the relative pleasantness or unpleasantness) of the elicited emotion, but also on differences in the self-reported intensity with

which emotions were typically experienced (Bachorowski & Braaten, 1994).

We observed similar outcomes in an unpublished study that used a more standard emotion-induction paradigm in which naive participants described the thoughts and feelings evoked by emotion-eliciting slides. Notably, efforts to link both source- and filter-related acoustic cues with discrete emotions were largely unsuccessful. Instead, the overall pattern of results indicated that values of acoustic parameters were associated with nonspecific arousal and, to a lesser extent, emotional valence. Again, differences in emotional intensity mediated the relationships between acoustic measures and emotional states.

Although the expression and perception of emotion are salient aspects of human vocal communication, researchers have yet to fully characterize the ways in which speech acoustics provide cues to emotional states. The most parsimonious interpretation of production-related data is that speech acoustics provide an external cue to the level of nonspecific arousal associated with emotional processes. Less reliable differentiations are found when researchers look for associations between acoustic measures and either emotional valence or discrete emotion categories. Moreover, potentially important individual differences, including the identity of the talker and emotional intensity, are routinely found to mediate vocal expression of emotion. As Scherer (1986) has pointed out, there is an apparent contradiction between the difficulty in finding acoustic differentiation of emotional states and the comparative ease with which listeners are able to judge emotions from speech. Resolving this contradiction will likely involve an explicit understanding of the role that individual difference variables play in the production of emotional speech.

PERCEPTION OF VOCAL EMOTION

Tests of listeners' abilities to infer emotion from speech are critical for evaluating the perceptual importance of acoustic cues shown to be important from a production perspective, and help to inform research aimed at developing an acoustic typology of emotional speech. The standard perception paradigm is to have listeners choose which one of several emotion words best characterizes linguistically neutral utterances made by actors attempting to portray various emotions (e.g., Leinonen et al., 1997; Scherer, Banse, & Wallbott, 1998). Listeners are usually able to perceive the intended emotions at rates significantly better than those expected by chance. This general success in identifying emotions is typically interpreted to indicate that listeners associate particular patterns of acoustic cues with various discrete emotional states. Evidence for cross-cultural similarities in both perceptual accuracy and error patterns (Scherer et al., 1998) further suggests that the ability to infer emotion from speech is a fundamental component of human vocal communication.

In light of these findings, it is also important to note that error rates are also often quite high. A hint about the basis of detection failures comes from the fact that listeners are more accurate in inferring emotion from particular voices. Furthermore, for any given actor, listeners typically perceive some emotions more accurately than others. Although it is likely that some emotions may simply be more difficult to infer from voice than others, and that actors vary in the quality of their emotion portrayals, these effects also suggest that characteristic acoustic differences between voices play a role in perceptual evaluations of emotion from speech.

TOWARD A BROADER FRAMEWORK

A number of constraints have impeded the development of a detailed account of vocal-emotion-related phenomena. For instance, speech is complex, both in the number of potentially relevant acoustic cues related to emotional expression and in the multiplicity of other factors that influence the speech signal at any moment in time. More pragmatically, accurate and detailed acoustic analysis is time-consuming. From a methodological standpoint, the small number of participants typically studied and the reliance on acted portrayals have limited the generality of findings. Paradigms that involve collecting speech samples during the controlled induction of emotional states best balance the need for methodological rigor and real-world validity.

Although investigators have typically sought to identify invariant patterns of acoustic cues for various discrete emotional experiences, this strategy may be problematic for a number of reasons. For instance, this tactic generally fails to consider the talker–listener relationship and the "intended" impact of vocal signals on the listener's affective states. Some cues, especially those associated with the rate of vocal-fold vibration, are readily modifiable. They can be used, for example, to signal communicative intent or be recruited for the purposes of affective persuasion. Thus, treating these cues as honest readouts of emotional states ignores their other potential functions in emotion-related communication.

Incorporating a more talker-centered (i.e., idiographic) perspective may also help advance our understanding of emotional speech. Evaluations of emotional state are necessarily made against an acoustic backdrop of individually distinctive voice characteristics, and yet differences among talkers are usually treated as uninteresting variability in vocal-emotion research. However, everyday experience suggests that more accurate and detailed perceptual judgments of emotional state can be made for familiar than for unfamiliar talkers. For example, discriminations between related emotions, such as amusement and joy, are probably more accurate for speech samples from a close friend than those from a more casual acquaintance. Suggestive empirical support for the importance of talker characteristics comes from studies indicating that acoustic differences among talkers exert a powerful influence on cognitive operations such as linguistic processing and memory (e.g., Palmeri, Goldinger, & Pisoni, 1993). Thus, more detailed characterizations of the acoustic features of emotional speech might be found by examining fluctuations in acoustic cues against comparatively more stable but individually distinctive talker characteristics (see Bachorowski & Owren, 1998).

Research in vocal-emotion phenomena might also benefit from a reinterpretation of findings based on Russell and Feldman Barrett's (1998) distinction between affect and emotion. In their account, affect is always present and is best described by bipolar dimensions of arousal and valence. In contrast, prototypical emotion episodes happen more rarely and are associated with identifiable neurophysiological, behavioral, and cognitive processes. This distinction certainly sheds new light on vocal production and perception phenomena. In that most studies have arguably examined affect rather than emotion, it may have been unreasonable to expect that distinct acoustic patterns could be identified. Instead, there is remarkable consistency in support of the notion that the acoustic features of "emotional" speech are best described using dimensions of nonspecific arousal and affective valence, and that most vocal productions index affective rather than emotional experience.

The expression and perception of emotional states in speech acoustics are fundamental aspects of human communication. In fact, disturbances in either of these communication components can contribute to profound deficits in social relationships. By its very nature, research in vocal expression and perception of emotion is richly interdisciplinary—a circumstance that gives rise to both its inherent complexities and its considerable intellectual appeal. As a result of improved digital processing techniques as well as advances in the related disciplines of speech science, cognitive science, and acoustic primatology, findings obtained in the coming years should prove especially informative for our understanding of emotional expression through the vocal channel.

Recommended Reading

Bachorowski, J.-A., & Owren, M.J. (1995). (See References)
Kent, R.D. (1997). (See References)
Murray, I.R., & Arnott, J.L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America, 93,* 1097–1108.
Pittam, J., & Scherer, K.R. (1993). (See References)

Acknowledgments—Work on this article was completed while the author was generously hosted as a Visiting Scholar by the Department of Psychology at Cornell University. Funds in support of this work came from National Science Foundation (POWRE) and National Institute of Mental Health (B/START) awards, and from Vanderbilt University. Michael J. Owren provided valuable comments on an earlier version of this manuscript, and our collaborative work led to some of the ideas presented here.

Notes

1. Address correspondence to Jo-Anne Bachorowski, Department of Psychology, Wilson Hall, Vanderbilt University, Nashville, TN 37240; e-mail: j.a.bachorowski@vanderbilt.edu.

2. Preliminary results from work being conducted in other laboratories demonstrate that both standard emotion-induction paradigms and playful, gamelike paradigms are successful for eliciting speech samples that can be used to study vocal expression of emotion. Some investigators using these kinds of strategies include Arvid Kappas (arvid@psy.ulaval.ca), Gary Katz (gary.katz@csun.edu), and Tom Johnstone in Klaus Scherer's lab (johnstone@fapse.unige.ch).

References

Bachorowski, J.-A., & Braaten, E.B. (1994). Emotional intensity: Measurement and theoretical implications. *Personality and Individual Differences, 17,* 191–199.

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

Bachorowski, J.-A., & Owren, M.J. (1998). *Acoustic cues to gender and talker identity are present in a short vowel segment produced in running speech.* Manuscript submitted for publication.

Banse, R., & Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70,* 614–636.

Kent, R.D. (1997). *The speech sciences.* San Diego: Singular Publishing.

Leinonen, L., Hiltunen, T., Linnankoski, I., & Laakso, M.-L. (1997). Expression of emotional-motivational connotations with a one-word utterance. *Journal of the Acoustical Society of America, 102,* 1853–1863.

Palmeri, T.J., Goldinger, S.D., & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology, 19,* 309–328.

Pittam, J., & Scherer, K.R. (1993). Vocal expression and communication of emotion. In M. Lewis & J.M. Haviland (Eds.), *Handbook of emotions* (pp. 185–197). New York: Guilford Press.

Russell, J.A., & Feldman Barrett, L. (1998). *Affect and prototypical emotional episodes.* Manuscript submitted for publication.

Scherer, K.R. (1986). Vocal affect expression: A review and model for future research. *Psychological Bulletin, 99,* 143–165.

Scherer, K.R. (1989). Vocal measurement of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience: Vol. 4. The measurement of emotions* (pp. 233–259). New York: Academic Press.

Scherer, K.R., Banse, R., & Wallbott, H.G. (1998). *Emotion inferences from vocal expression correlate across languages and cultures.* Manuscript submitted for publication.

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