23

VOCAL EXPRESSION OF EMOTION

Klaus R. Scherer, Tom Johnstone,
and Gundrun Klasmeyer

This chapter reviews theoretical models and empirical evidence on the effects of emotion on vocalization, in particular on human speech. Of all expressive manifestations of emotional arousal, vocalization represents the most phylogenetically continuous modality. The neural control structures, the voice production mechanisms, and the characteristics of vocal emotion signals are comparable across many species of mammals, including humans (Hauser, 1996; Marler & Tenaza, 1977; Morton, 1977; Scherer, 1985; see chapter 24, this volume). In consequence, the study of emotional vocalization constitutes a prime tool to investigate the relationships between the physiological production of affect vocalization and its use as a signal in social interaction and communication. The basis for the following discussion is a modified version of the Brunswikian lens model (Figure 23.1; see also Kappas, Hess, & Scherer, 1991; Scherer, 1978, 1982). This model allows one to clearly distinguish between the *expression* (or encoding) of emotion on the sender side, the *transmission* of the sound, and the *impression* (or decoding) on the receiver side, resulting in emotion inference or attribution. Consequently, the model encourages research on the complete process of emotion communication by (1) determining which distal characteristics of the sound wave emerging from the mouth of an affectively aroused speaker (i.e., the acoustic parameters) are produced by the underlying emotion, (2) determining how these cues are transmitted from sender to receiver through the vocal-acoustic-auditory channel, and (3) determining how these proximal cues (the representation of the distal voice characteristics in the sensorium and the central nervous system) are used by the receiver to infer the emotion of the sender. As will be shown, most research in this area has focused either on encoding or decoding. Our aim in organizing the chapter by a model that charts the entire process is to encourage future research to attend to all of these aspects and their interrelationships.

Encoding: Expression of Emotion in Voice and Speech

The first step described by the Brunswikian lens model is the encoding of emotion in the acoustic properties of the speech signal, through a selective modification of speech production. It has been suggested that the vocal characteristics of emotional utterances are determined by both push and pull effects (Scherer, 1985, 1994). *Push effects* refer to the direct effects of the physiological changes characterizing many emotional responses on the voice and speech production system (see Scherer, 1989). *Pull effects* reflect the fact that vocalization is, as is other expressive behavior, often closely monitored and regulated (or sometimes even expressly produced) for strategic reasons. When pull effects operate, voice production targets are determined, at least in part, by normatively proscribed or conventionalized acoustical signal patterns (see also Caffi & Janney, 1994). In this chapter we will mostly focus on

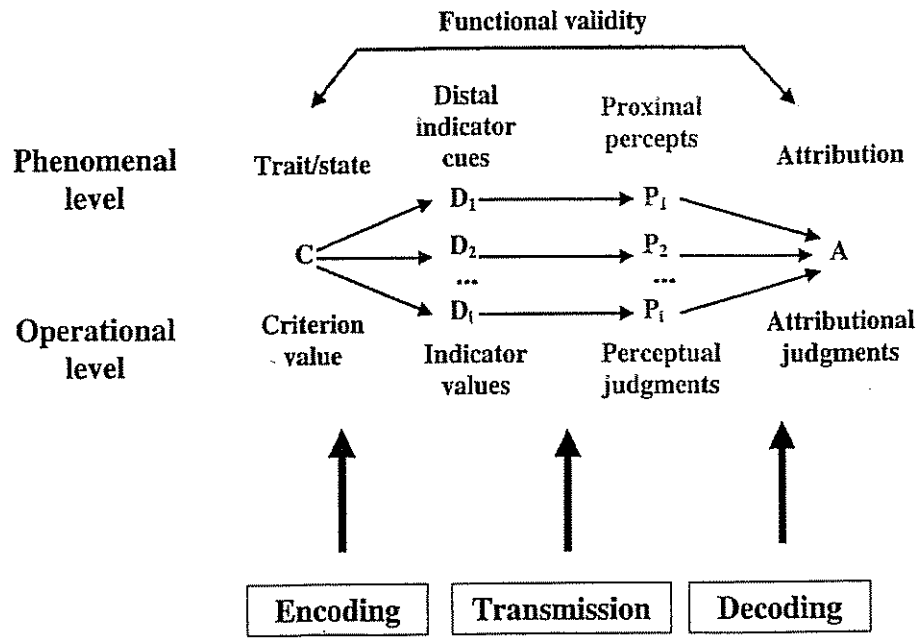

Figure 23.1 Modified form of the Brunswikian lens model applied to the process of the vocal communication of emotion. Adapted from Scherer (1982).

push effects, although pull effects will be mentioned in places.

Theoretical Predictions of Competing Emotion Models

Much of the work in this area has been atheoretical, searching to empirically determine which changes in voice and speech production, and which corresponding patterns of acoustic parameters, will be produced by inductions of stress or specific emotions in the speaker. As one might expect, this approach entails the problem of making sense of a multitude of different, often non-replicated results. In particular, it has been difficult to determine which type of emotional state has been induced in a study and to decide whether it can be reasonably compared with similarly labeled states in other studies. It can be shown (see Banse & Scherer, 1996; Scherer, 1986) that the term *anger* may cover anything from a mild irritation produced by a manipulation of experimenter rudeness to violent rage, as portrayed by an actor. In consequence, recent work has attempted to ground research on theoretical models of vocalization changes in emotion and to examine concrete predictions. In what follows we review the respective predictions of three major theoretical traditions: arousal theories, discrete emotion theories, and appraisal theories (see Scherer, 2000a, and the introduction to this volume).

Arousal Theories

Proponents of this approach have generally neglected vocal behavior. For example, arousal models of emotion (e.g., Duffy, 1962; Thayer, 1989) have focused on predicting physiological response in the autonomic nervous system (ANS), not being much concerned with vocal or facial expression. These models suggest that emotions are mostly differentiated along a continuum from low to high sympathetic activation (often linked to an active-passive dimension in terms of the corresponding feeling states). Even though these theorists have not produced explicit predictions of vocal expression patterns, it is relatively straightforward to extrapolate the vocal characteristics to be expected for certain emotions on the basis of the physiological predictions for sympathetic arousal (e.g., deeper and faster respiration, increased cardiovascular activity, increased muscle tension). The corresponding acoustical effects are increases in central tendency and range of fundamental frequency (F0) and intensity, as well as an increase in harmonic energy and in the rate of articulation. In consequence, if arousal models of emotion explain emotion differentiation with a sufficient degree of accuracy, one would expect vocal parameters that are directly tied to sympathetic arousal to provide an exhaustive set of predictors for the vocal patterning of different affect states. It has indeed long been presumed that vocal expression mostly, if not exclusively, indicates the arousal component of emotion, being unable to convey differences in valence or qualitative differences between emotion categories (see Banse & Scherer, 1996; Scherer, 1979, 1986, for a more detailed discussion).

One-dimensional arousal theories are less popular today, even though in many studies on vocal expression of emotion, and especially of stress, they seem to still implicitly determine the interpretation of the data by researchers.

Most modern dimensional researchers combine the arousal or activity dimension with the valence or pleasantness dimension (Lang, Greenwald, Bradley, & Hamm, 1994; Russell, 1980). As was the case of earlier arousal theorists, these modern dimensionalists have unfortunately not made any direct predictions about vocal expression (but see Bachorowski, 1999; Bachorowski & Owren, 1995, for a recent dimensional approach on vocal expression).

Discrete Emotion Theories

After the waning of the popularity of arousal models in the 1960s, discrete emotion models, as suggested by Tomkins (1962, 1984) and popularized by Ekman (1972, 1992) and Izard (1971, 1977), dominated the field of emotion (as well as the textbooks). Following Darwin's (1872/1998) pioneering effort to define a limited number of basic emotions, Tomkins suggested that specific response patterns for these fundamental, discrete emotions were produced by innate neuromotor programs. While Tomkins, just as Darwin, insisted on the importance of specific vocal expressions for basic emotions, he did not develop specific hypotheses, contrary to the facial expression modality for which he, and in particular his followers, Ekman, Friesen, and Izard developed very elaborate predictions of facial patterning (as based on the presumed action of neuromotor programs; Ekman & Friesen, 1978; Izard, 1971). In consequence, it is difficult to extrapolate the predictions that discrete-emotion theorists might make for the vocal expression of these basic emotions. However, given the underlying philosophy of this approach, it seems reasonable to assume that discrete-emotion theorists would postulate clearly distinct and highly emotion-specific patterns of vocal parameter configurations, completely predictable on the basis of the type of emotion a particular token is expected to exemplify.

Appraisal Theories

The most recent type of emotion models is represented by appraisal theories that assume that emotion differentiation is determined by the outcomes of event evaluation processes as based on a set of appraisal criteria (see chapter 29, this volume; Scherer, 1999a; Scherer, Schorr, & Johnstone, 2001). Many of these theorists assume that the efferent response patterns (including physiological changes and facial/vocal expression) are produced by specific appraisal results and serve as adaptive responses to the need for information processing and action tendencies (Scherer, 1984, 1992; Smith & Ellsworth, 1985). For example, Smith (1989) showed a correlation between the appraisal of goal obstructiveness and innervation of the corrugator muscle. Thus, while subscribing, like discrete-emotion theorists, to the assumption that emotion-specific patterning constitutes an adaptive response, appraisal theorists do not assume unitary mechanisms, such as neuromotor programs, for a small number of "basic" emotions. Rather, they propose that the specificity in the patterning can be best predicted as the cumulative result of the adaptive changes produced by a specific appraisal profile (see Scherer, 1986, 1992, 2000b, 2001; Smith & Scott, 1997, for detailed reviews of this hypothetical patterning mechanism).

On the basis of these assumptions, Scherer (1986) has produced an extensive set of predictions concerning the physiological changes and the ensuing consequences for the voice production mechanism that can be expected for *appraisal results on specific dimensions* (based on functional considerations). The complexity of these predictions does not allow us to present them in detail here. Briefly put, the justification for the predictions rests in using the assumed functional consequence of a particular appraisal result to predict the pattern of peripheral physiological arousal. Then, the effects of the respective physiological pattern on the voice production process are estimated and the acoustic concomitants are inferred. For example, appraisal of goal obstructiveness requires strong action (e.g., fighting), which should lead to high sympathetic arousal with the consequent changes for respiration and muscle tension, and thereby changes in phonation (higher F0, different glottal pulse shape producing energy changes in the spectrum). Similarly, it is predicted that an appraisal of high coping potential (e.g., power to deal with an obstacle) will lead to orofacial changes evolutionarily linked to biting behavior. The configuration of the vocal tract produced by this setting will privilege certain filter characteristics of the vocal tract (see Ladefoged, 1975; Scherer, 1982) and will thus affect energy distribution in the spectrum. Based on a theoretical analysis of emotion-specific appraisal profiles, Scherer (1986) has proposed a number of detailed hypotheses concerning the vocal parameter configurations to be expected for a number of *specific modal emotions*. The interested reader is invited to read further details of the underlying model and the predictions it generates in Scherer (1986) or Johnstone, van Reekum, and Scherer (2001).

In what follows, we review the empirical evidence available to date with respect to the predictions emanating from these three theoretical traditions. Specifically, three questions are addressed: (1) Is the notion of sympathetic arousal sufficient to describe vocal emotion expression differentiation (as would be held by arousal models)? (2) Is there evidence for highly prototypical vocal parameter configurations that might be produced by invariable neuromotor programs (as held by discrete emotion theorists)? (3) What is the evidence for the more molecular predictions from appraisal theory, either for the effects of results on specific appraisal dimensions or for the combined predictions for modal emotions? Sympathetic arousal theories imply that emotion-specific variance in the acoustic properties of speech are explainable in terms of a single

arousal factor, and that emotions that are qualitatively different but similar with respect to sympathetic arousal (e.g., rage and elation) are vocally indistinguishable. Both discrete-emotion theories and appraisal theories posit the existence of more specific vocal expressions. Proponents of discrete-emotion theories would predict that specific vocal patterns would be found in well-defined clusters, each corresponding to a different discrete emotion. In contrast, appraisal theories suggest the existence of a highly differentiated set of vocal profiles, since the change in outcome of any single appraisal check is expected to result in changes to vocal production. In consequence, even for emotional states that are labeled with the same word, one may find somewhat varying vocal patterns, depending, for example, on the degree of goal obstruction or whether a reaction is considered to be urgent.

Evidence from Empirical Research

The empirical studies conducted over the last six decades can be classified into three major categories, based on the criterion of the type of speech material used: natural vocal expression, induced emotional expression, and simulated emotional expression (see also Campbell, 2000).

Studies using *natural vocal expression* have made use of recorded material that was taped during naturally occurring emotional states of various sorts, such as dangerous flight situations for pilots, journalists reporting emotion-eliciting events, affectively loaded therapy sessions, and so on (see Table 7 in Scherer, 1986, for details on earlier studies; see also Bachorowski, 1999; Frick, 1985; Murray & Arnott, 1993). In recent years, affectively toned speech samples recorded off the air (e.g., from TV game or reality shows) have been used (Douglas-Cowie, Cowie, & Schröder, 2000; Roach, Stibbard, Osborne, Arnfield, & Setter, 1998). While the use of naturally occurring voice changes in emotionally charged situations has the obvious advantage of high ecological validity, there are serious drawbacks. In many cases, only brief voice samples for a very small number of speakers, often suffering from bad recording quality, are available. Furthermore, it is not always obvious which emotion the speakers have really experienced in the situation. While researchers using this type of material often infer emotional quality on the basis of the type of situation, appraisal theorists point out that the same event may provoke very different emotions in different individuals and insist on the need to study the subjective appraisal of each individual, even in apparently similar predicaments (see Roseman & Smith, 2001; chapter 29, this volume). In addition, it is possible that there are strong effects of self-presentation or display rules, especially for material recorded from TV shows, that may render the ecological validity of such material suspect.

In a number of experimental studies, vocal expression during *induced emotions* has been studied (again, more detailed information can be found in Table 7 in Scherer, 1986; see also Bachorowski, 1999; Frick, 1985; Murray & Arnott, 1993; for empirical research using these types of vocal emotional expression, see Bachorowski & Owren, 1995; Johnstone, 2001; Kappas, 1997; Karlsson et al., 2000; Scherer, Johnstone, Klasmeyer, & Bänziger, 2000; Tolkmitt & Scherer, 1986). The induction procedures have varied from study to study, including stress induction (e.g., difficult tasks to be completed under time pressure), presentation of emotion-inducing films or slides, imagery methods, or computer games. While the experimentally controlled induction procedures allow the production of comparable voice samples for all participants, these procedures often produce only relatively weak affect. Furthermore, in spite of using the same procedure for all participants, it is not certain, as mentioned just above, that the same affective states are produced in all individuals, given individual differences in event appraisal (see chapter 29, this volume).

By far the largest number of studies in this area have made use of *simulated vocal expressions* as produced by professional or lay actors (again see Table 7 in Scherer, 1986, for details on earlier studies; see also Bachorowski, 1999; Frick, 1985; Murray & Arnott, 1993; for empirical research using these types of vocal emotional expression see Banse & Scherer, 1996; Juslin & Laukka, 2001; Klasmeyer, 1999; Paeschke, Kienast, & Sendlmeier, 1999; Tischer, 1993; Wallbott & Scherer, 1986). While studies using simulated vocal portrayal of emotions can be expected to yield much more intense, prototypical expressions than either induced states or natural emotions that can be publicly observed (Scherer, 1986, p. 159), there is a danger that actors overemphasize relatively obvious cues and miss more subtle ones that might appear in natural expression of emotion (Scherer, 1986, p. 144). Clearly, such simulated speech samples reflect not only the push effects of emotion on the voice but also sociocultural norms that provide the actor with a target vocal pattern that pulls the voice in a particular direction. However, there is reason to believe (see Johnstone & Scherer, 2000; Scherer, 1986) that such vocal expressions still strongly resemble purely push-related vocal expressions of emotion: If the two were to diverge too much, the acted version would lose its credibility (as may sometimes be the case with bad actors). Yet it cannot be excluded that actors use conventionalized stereotypes of vocal expression (possibly in part dictated by specific acting schools) that allow them to differentially communicate the intended emotions in a way that corresponds only partly to natural expression.

Classification and Measurement of Vocal Parameters

In this section, we discuss the nature of the vocal parameters that have served as indicators of emotional change

and that will be used to summarize the findings. As in many other areas of psychological functioning, it is difficult to observe the mechanisms that underlie the effect of emotional arousal on voice and speech. The mental processes involved in speech production, including the planning of motor commands, can hardly be accessed at all. Using appropriate instruments, one can measure movements of the articulators (e.g., lips, tongue) or the behavior of the voice source (e.g., the vibrations of the vocal folds). But apart from the fact that specialized instruments such as flow masks, laryngographs, palatographs, or articulographs are rarely available to researchers interested in the vocal expression of emotions, there is another problem: If we subject a speaker to this type of intrusive measurement in order to study articulation or voice source dynamics, we can hardly expect him or her to produce natural emotional speech. This is why analysis of acoustic parameters of the speech signal resulting from the production process, which can be simply recorded with a microphone, has been the method of choice for researchers in this area, using analog measurement devices in the earlier periods and digital speech analysis procedures in recent years (see Ladefoged, 1975, for a general introduction to these measures; more technical treatment of these issues is provided in O'Shaughnessy, 2000; Rosen & Howell, 1991).

Historically, acoustic signal dimensions such as duration, amplitude (intensity), fundamental frequency of vocal fold vibration (F0), and energy distribution in the frequency spectrum were the first parameters to be objectively measured. They are relatively easy to define and can be measured fairly automatically. In contrast, parameters requiring more or less extensive phonetic interpretation often lack consensual definition and cannot easily be measured automatically. It so happens that the rather straightforward, automatically analyzable parameters are mostly indicative of arousal, making it necessary to employ more complex, derived parameters if one wants to go beyond the arousal dimension. As shown later, a large number of different acoustic parameters can be currently obtained by appropriate digital analysis of recorded speech. Table 23.1 summarizes the acoustic-phonetic and psychoacoustic parameters described throughout the chapter.

In order to interpret the acoustic parameters in a theoretical context, we need to define the acoustic parameters in terms of the physiology of speech production and the hypothesized links with physiological response patterns found for different emotions (see Scherer, 1986). While there is no one-to-one relationship, links with the underlying speech production mechanisms can be shown for many acoustic-phonetic parameters. Traditionally, the speech production process is classified into respiration, phonation, and articulation. In terms of acoustics, the effects of articulator positions and movements in the vocal tract can be regarded as a filter with time-varying filter characteristics applied to the (also time-varying) source signal produced by respiration and phonation (Fant, 1960; see also Ladefoged, 1975; O'Shaughnessy, 2000). Therefore, it would be useful to obtain separate indicators for source and filter characteristics. Unfortunately, it is not a trivial task to distinguish clearly between the influence of articulation and the characteristics of the source signal in acoustic analyses of recorded speech signals. A further analytical subdivision between the acoustic effects of respiration and phonation as determinants of the source signal is even more difficult (see Sundberg, 1994, for details). This is why acoustic models often regard respiration and phonation as just one component of the speech production process.

Voice source parameters can be classified into *general voice quality* variables (such as type of phonation; Laver, 1980), which more or less reflect a generalized "tension state" of specific muscles in the larynx, and *prosodic* parameters that reflect voluntary, dynamic changes in the tension of specific larynx muscles during speech production (giving rise to intonation that has both syntactic and pragmatic functions). Scherer (1986) hypothesized that the general tension state of larynx muscles is affected directly by the outcomes of event evaluation and thus by the resulting emotional response characteristics. These tension states are assumed to appear independently of the subject's decision whether to speak or not. If speech is produced, dynamic changes in specific larynx muscles are superimposed on the general tension states.

Articulation parameters can be classified into those that are mandatory for the production of intelligible speech (e.g., those that determine formant positions) and those that might reflect extralinguistic factors such as facial expressions (e.g., Massaro, 2000; Ohala, 1980; de Gelder & Vroomen, 2000). Such detailed interpretations require a great deal of phonetic knowledge or make assumptions about the speech production process. A complicating factor is the nature of individual speaker characteristics, such as those related to form, size, and mass of vocal organs and their state of health, as well as to personal habits concerning their use (which Laver, 1991, distinguishes as "vocal equipment" and "vocal settings").

A further factor that renders the investigation of acoustic parameters so complicated is the fact that speech signals are transient signals: All parameters change constantly over time. In earlier studies on vocal expression, time-varying parameter values within the utterance were often aggregated into means and variability coefficients, but there is increasing evidence that these statistical descriptions of time series are not sufficient to understand the process of emotional expression in speech. More or less static parameters describing general voice quality, for example, can be measured more reliably in individual phonemes such as the open-vowel / a /, than in long-term average measurements because for these segments the vocal tract is open and articulatory effects have less effect

Table 23.1. Overview of Major Acoustic-Phonetic and Psychoacoustic Parameters

1. Acoustic-phonetic parameters	
Speech rate	Number of speech segments per time unit
F0	Fundamental frequency (vibration rate of vocal folds)
F0 perturbation	Slight variations of duration of glottal cycle
F0 contour	Fundamental frequency values plotted over time (intonation or "speech melody")
Intensity	Average squared amplitude within a predefined time segment ("energy")
Spectral energy distribution	Relative amount of energy within predefined frequency bands
Spectral slope	Linear regression of energy distribution in the frequency band above 1 kHz
Laryngalization	Sudden change of oscillation mode of vocal folds (usually to double glottal cycle duration)
Tremor	Regular modulation of glottal cycle duration
Jitter	Regular or irregular variation of glottal cycle duration
Shimmer	Regular or irregular variation of amplitude maxima in subsequent glottal cycles
HNR	("Harmonic-to-noise-ratio") Ratio between harmonic and aperiodic signal energy
GNE	"Glottal-noise-excitation," a measure of hoarseness
Inverse filtered glottal pulse	Transglottal airflow estimated by inverse filtering techniques
Formant	(Time-varying) resonance of vocal tract (significant energy concentration in the spectrum)
F1	First formant—important for vowel identification
F2	Second formant—important for vowel identification
Formant bandwidth	Width of the spectral band containing significant formant energy (− 3 dB threshold)
Formant precision	Degree to which formant frequencies attain values prescribed by phonological system of a language
2. Psychoacoustic parameters	
Perceived loudness	Calculated from weighted energy distributions in specific frequency bands
Perceived pitch	Calculated from the F0 contour taking into consideration the linear and differential glissando-threshold
Perceived rhythm	Rhythmic events calculated from the perceived loudness contour

Note: In many cases, these parameters are aggregated over segments of speech by using measures of central tendency and distribution statistics such as mode, mean, median, variance, standard deviation, and upper or lower 5(1)% of the distribution.

on the voice source signal. To extract dynamic features from time contours of acoustic parameters, such as F0 contours or formants, for example, it is helpful to modify measured contours into simplified, but functionally equivalent forms. Unfortunately, even among linguists and phoneticians there is little agreement on how to stylize F0 contours and how to separate meaningful from irrelevant information (see Mertens, Beaugendre, & d'Alessandro, 1997).

Summary of the Empirical Results

Table 23.2 provides an overview of the major effects of emotion on vocal expression that have been empirically identified. Column 1 lists the acoustic parameters generally used to measure the effects of emotion, annotated with comments. Columns 2 to 7 summarize the research results of the effects of arousal, happiness, anger, sadness, fear, and boredom on the respective acoustic parameters. The entries in the table refer to differences on the respective parameter compared to "normal" speech. In order to maximize the number of studies upon which to base the inference, we list only the emotions that have been most frequently studied in this field. Other emotions such as pride, jealousy, love, and humor (laughter) have been studied considerably less frequently. In consequence, the evidence so far is extremely sparse and in urgent need of replication. Given the number of studies, the manifold differences in emotions studied and acoustic parameters measured, as well as inconsistencies in the data, it is impossible to document in detail how the entries in Table 23.2 have been excerpted from the literature. The authors

Table 23.2. Synthetic Review of the Empirical Findings Concerning the Effect of Emotion on Vocal Parameters

Acoustic Parameters	Arousal/Stress	Happiness/ Elation	Anger/Rage	Sadness	Fear/Panic	Boredom
Speech Rate and Fluency						
Number of syllables per second	>	>=	<>	<	>	<
Syllable duration	<	<=	<>	>	<	>
Duration of accented vowels	>=	>=	>	>=	<	>=
Number and duration of pauses	<	<	<	>	<>	>
Relative duration of voiced segments			>		<>	
Relative duration of unvoiced segments			<		<>	
Voice Source—F0 and Prosody						
F0 mean[3]	>	>	>	<	>	<=
F0: 5th percentile[3]	>	>	=	<=	>	<=
F0 deviation[3]	>	>	>	<	>	<
F0 range[3]	>	>	>	<	<>	<=
Frequency of accented syllables	>	>=	>	<		
Gradient of F0 rising and falling[3,6]	>	>	>	<	<>	<=
F0 final fall: range and gradient[3,4,7]	>	>	>	<	<>	<=
Voice Source—Vocal Effort and Type of Phonation						
Intensity (dB) mean[5]	>	>=	>	<=		<=
Intensity (dB) deviation[5]	>	>	>	<		<
Gradient of intensity rising and falling[2]	>	>=	>	<		<=
Relative spectral energy in higher bands[1]	>	>	>	<	<>	<=
Spectral slope[1]	<	<	<	>	<>	>
Laryngealization		=	=	>	>	=
Jitter[3]		>=	>=		>	=
Shimmer[3]		>=	>=		>	=
Harmonics/Noise Ratio[1,3]		>	>	<	<	<=
Articulation—Speed and Precision						
Formants—precision of location	?	=	>	<	<=	<=
Formant bandwidth	<		<	>		>=

Notes:
1. depends on phoneme combinations, articulation precision or tension of the vocal tract
2. depends on prosodic features like accent realization, rhythm, etc.
3. depends on speaker-specific factors like age, gender, health, etc.
4. depends on sentence mode
5. depends on microphone distance and amplification
6. for accented segments
7. for final portion of sentences

In specific phonemes, < "smaller," "lower," "slower," "less," "flatter," or "narrower"; = equal to "neutral"; > "bigger," "higher," "faster," "more," "steeper," or "broader"; <= smaller or equal, >=bigger or equal; <> both smaller and bigger have been reported

have used their best judgment in synthesizing the emerging pattern of data from the studies cited earlier. It should be noted that for some of the more rarely measured acoustic parameters the table entry is based on only very few studies, and in some cases a single one. In consequence, Table 23.2 should be viewed as a set of empirical expectations rather than an authoritative summary of established results.

It is impossible to summarize or synthesize the mass of findings presented in Table 23.2. However, we will try to draw some preliminary conclusions with respect to how well the evidence available so far supports the three theoretical models outlined at the beginning. In order to do so, we will review the three questions posed earlier.

1. Are sympathetic effects sufficient to describe the vocal emotion expression differentiation that has been empirically observed? Based on the results reviewed, there can be no doubt that vocal parameters are very powerful indicators of physiological arousal (as stated in prior reviews by Frick, 1985; Murray & Arnott, 1993; Pittam &

Scherer, 1993; Scherer, 1979, 1986). High-stress or high mental workload conditions have generally been found to lead to raised values of F0, greater intensity, and faster speech rate than low-stress situations. Similarly, compared with neutral speech, vocal expressions of high-arousal emotions such as anger, fear, and elation have been measured as having high mean F0, high F0 variability, high intensity, and increased speaking rate. Conversely, sad and bored vocal expressions have been found to have low F0, low F0 variability, low intensity, and decreased speaking rate. On the basis of such combined results, it is evident that where there has been considerable consistency in the findings, it has usually been related to arousal, regardless of the specific quality of the emotion under investigation.

However, while there has never been any doubt as to the important role of arousal, there is debate as to whether sympathetic arousal is sufficient to account for the vocal differentiation of emotion, as unidimensional arousal theories of emotion would hold. Scherer and his collaborators (Banse & Scherer, 1996; Johnstone, Van Reeckum, & Scherer, 2001; Scherer, 1986) argue that while only very few studies of induced or real emotional speech identified acoustic patterns that could unambiguously differentiate the major non-arousal dimensions of emotion such as valence and control, there are good reasons for favoring a more differentiated model of emotional expression. Although comparisons between fairly extreme emotions, such as elation, rage, or fear on the one hand and sadness or boredom on the other, reflect mainly the effects of arousal, less extreme forms such as happiness, irritation, anxiety, and disappointment do not show such a consistent relationship with arousal. For example, anxious speech is often low in intensity and irritated speech has been found to be low in F0. The fact that these emotions can still be accurately perceived by listeners (as will be discussed in the section on decoding studies) implies that some aspects of the acoustic signal other than those related to arousal serve as a marker for the emotion. Since most studies have limited themselves to measuring those acoustic parameters that have a clear physiological connection with arousal, such as F0 and intensity, it is not surprising that nonarousal markers of emotion in speech have not been consistently identified so far.

Table 23.2 also provides evidence that certain emotions are accompanied by specific types of phonation (Klasmeyer & Sendlmeier, 1997), which cannot be explained by simple arousal models. For example—voicing irregularities that appear when the speaker is almost weeping or retching, or voice breaks, found for panic fear—are never found in angry utterances even when produced with extreme arousal. From these observations one may conclude that some parameters, such as voicing irregularities, can be best explained by emotion-specific innervation of the muscles in the larynx.

In addition to voice source parameters, the effect of emotion on articulation patterns (apart from elaboration or reduction phenomena) is difficult to explain with simple arousal models. These effects are likely to be emotion-specific rather than arousal-related. This is clearly seen in the effects of facial expression on acoustic parameters. For example, both the specific actions of the *m. zygomaticus* in smiling and the *m. orbicularis oris* in anger expression affect the energy distribution in the spectrum. While the respective contribution of phonetic-linguistic and emotional factors to the joint determination of vowel formant frequencies and bandwidth is difficult to assess given major differences in articulation patterns between individuals and dialectal forms, this group of parameters seems highly promising for future research.

Perhaps the most compelling evidence so far for the existence of acoustic profiles that distinguish not only between high and low arousal but also between different emotions with comparable levels of arousal comes from a recent study by Banse and Scherer (1996), in which 14 emotions were expressed (in two nonsense sentences) by 12 professional theater actors using a scenario-based Stanislaski technique. Acoustic analyses revealed quite specific acoustic profiles for a number of emotions. Emotions such as rage, panic, and elation, all high in arousal, had distinct acoustic profiles (as summarized by Johnstone & Scherer, 2000, table 3). When the acoustic parameters were used in discriminant analysis to classify each spoken token, the classification accuracies were well above chance levels for all emotions and corresponded closely to the classification accuracies of human judges as well as showing similar patterns of confusions.

2. Is there evidence for highly prototypical vocal parameter configurations that might be produced by invariable neuromotor programs as postulated by discrete emotion theories? While the evidence is certainly consistent with this position, the strength and prototypicality of the patterning that would be required to confirm the existence of neuromotor programs has not been found so far. On the contrary, as already shown, there are important differences in vocal patterning between members of the same emotion family. For example, as shown by Banse and Scherer (1996), "hot anger" is characterized by a strong increase in F0 and F0 variability, fast speech rate, and a strong increase in high frequency energy, but this is not true for "cold anger" (which seems to be encoded through subtler cues, possibly including intonation). This suggests that there are many factors contributing to the production of the vocal expression of a particular emotional state. Further research will need to identify these factors and explain their interactions. Unfortunately, discrete-emotion theorists have not, so far, specified concrete hypotheses for vocal patterning that could serve as guidance for a comparative test of different theoretical predictions.

3. What is the evidence for the more molecular predic-

tions from appraisal theory, either for the effects of results on specific appraisal dimensions, or for the combined predictions for modal emotions? Appraisal theory predictions for vocal patterning are relatively recent (Scherer, 1986) and consequently, there is little experimental evidence to date. However, the study by Banse and Scherer (1996) aimed at a systematic test of the respective predictions made for modal emotions. Johnstone, van Reekum, and Scherer (2001) have summarized the results in a table reproduced here (see Table 23.3). While many of the predictions were supported by the data, there were also a number of marked discrepancies. Similar results have been found in a more recent test of Scherer's predictions by Juslin and Laukka (2001). As pointed out by Johnstone, van Reekum, and Scherer, it is difficult to identify the source of such discrepancies because the predictions reflect the composite of a number of individual appraisal predictions. Although it is possible to speculate about the effects of individual appraisals by looking at differences between emotion pairs on a single appraisal dimension, it is obvious that such speculation needs to be backed up by direct empirical testing.

The predictions tested in the study by Banse and Scherer (1996) concerned the combined predictions, based on complete appraisal profiles for modal emotions (Scherer, 1994). What is needed to confirm the general approach is experimental confirmation for specific effects of individual appraisal predictions. This has been attempted in more recent research, using computer games and tasks to induce emotional vocal responses (Johnstone, 1996, 2001; Kappas, 1997). The computer games and tasks were manipulated so as to provoke specific appraisal outcomes that were theorized by Scherer (1986) to lead to specific vocal changes. For example, one such game included characters that either hindered or helped the player through situations that were accompanied by pleasant or unpleasant sounds. Such a game was designed to manipulate the player's appraisals of goal conduciveness and intrinsic pleasantness, respectively. During the gameplay, players were requested to say standard phrases, which were recorded for later acoustical analysis. The results showed that the acoustic patterns of induced vocal changes could not be explained by a single arousal dimension, since the pattern of changes across acoustic parameters was different for different manipulations. For example, Johnstone (2001) found that varying the intrinsic pleasantness of an event caused changes to spectral energy distribution, but not to overall energy, F0 level, or fluency. In contrast, changes to the conduciveness of an event produced changes to the latter set of variables, but not to spectral energy distribution. Although a single-dimension arousal model could be modified to fit such data (for example, by positing different non-monotonic relationships between arousal and individual acoustical parameters), a more parsimonious explanation is that emotional changes to the voice reflect two or more dimensions, presumably reflecting two or more underlying mechanisms.

Summary of the Evidence on the Vocal Patterning of Emotion

This review of the pertinent research results does confirm the important role of arousal but strongly argues against the claim that a unidimensional arousal theory can account for the data in a satisfactory manner. While the data are consistent with discrete-emotion theories, there is little evidence so far that favors the more constraining predictions concerning prototypical patterns for a limited number of "basic" emotions over the more molecular predictions of appraisal theory. Similarly, the partial support for the appraisal-theory generated predictions in the work by Banse and Scherer (1996), Johnstone (1996, 2001), and Juslin and Laukka (2001) is also consistent with the assumption made by discrete-emotion theories that there are qualitative differences between emotions. In addition, they do not contradict dimensional theories, since psychophysiological arousal and subjective valence of emotional experience are components of both discrete and appraisal models. Although it is difficult to differentially test the theoretical views, it would seem desirable to abandon the atheoretical stance that has characterized this field and to plan and conduct further studies in this area with more

Table 23.3. Predicted and Measured Standardized Vocal Parameters for 12 Emotions as Reported by Banse and Scherer (1996)

	Contempt	Boredom	Happiness	Anxiety	Shame	Sadness	Disgust	Cold Anger	Despair	Hot Anger	Panic	Elation
F0	**0−**	−	−	**+−**	**+−**	00	**+0**	00	++	**0+**	++	++
Energy	**+−**	**0−**	−	?0	**?−**	−	**+−**	++	++	++	++	++
LF energy	**−0**	0+	++	**−+**	−0	0+	−0	−	−	−	−0	00
Duration	?0	?+	+−	?−	?0	++	?0	?0	**−0**	−	−	−

Note: LF energy = Low frequency energy. In each cell, the first symbol represents the predicted value, the second symbol represents the measured value. −, 0, + : low, medium, high values, respectively. Symbols in bold indicate a significant difference between the predicted and measured values. ? indicates that no prediction was made. From Johnstone, van Reekum, & Scherer (2001), copyright © Oxford University Press, Inc. Used by permission of Oxford University Press, Inc.

explicit attempts at theory-driven investigation. This would not only allow a critical examination of conflicting predictions made by competing theories but also an accumulation of empirical findings to a systematic framework. Recent efforts in this direction have been made for facial expression of emotion (e.g., Wehrle, Kaiser, Schmidt, & Scherer, 2000), with mixed success. As mentioned at the start of this section, there are, in principle, some opposing claims made by discrete-emotion theories and appraisal theories that could be empirically tested. Given a large database of emotional speech recordings, it might be possible to use cluster analysis and factor analytic techniques to determine whether the acoustic characteristics of emotional speech are characterized by a small number of concentrated vocal prototypes, or by a more diffuse arrangement explainable by a number of dimensions corresponding to different appraisals. In addition, it might be that more neuroscientifically oriented research, as will be discussed later in the chapter, ultimately will indicate which of the theories presented earlier is better able to predict and explain the vocal patterning of emotional expression.

Transmission of Voice Sounds and Perception by the Listener

One of the important features of the Brunswikian lens model is the consideration given to the transmission of the distal signals from the sender to the receiver/listener who perceives these signals as subjective, proximal cues. The Brunswikian lens model is useful to model these processes on a theoretical level. It can also be used for quantitative modeling (e.g., using path analysis or structural models; e.g., Scherer, 1978) for subsets of variables that are linked by linear functions. In modeling the transmission part of the model, two major aspects need to be taken into account: (1) the transmission of sound through space and (2) the transform functions in perception, as determined by the nature of human hearing mechanisms. These effects, which are highly nonlinear and thus require more complex statistical approaches, will be briefly described.

1. The transmission of sound through space (and consequently the nature of the cues that reach the listener's ear) is influenced by many environmental factors, including the distance between sender and receiver and the presence of other sounds and background noise. A consideration of such factors is important in understanding the physiological mechanisms that underlie both the expression and perception of emotional speech. If, for example, speech has to carry over a long distance to reach a listener, the mode of production will be different than that used for close-range communication. In particular, the speaker needs to produce more intense speech, requiring more vocal effort. However, the result of greater vocal effort is not simply an increase in intensity; a large number of acoustic characteristics related to voice production at the larynx will also be affected. In addition, articulation has to be rather precise in order to avoid misunderstanding, particularly since other nonverbal cues that normally support the verbal message, such as lip movements, gestures, eye contact, or facial expression, may not be be visible, or only marginally so. Thus, the need to communicate over a long distance will impose constraints on the use of certain vocal features for emotional expression. The absence of a number of nonvocal cues that accompany the affective vocal signal when the listener is close to the speaker but cannot be used at a distance will affect the characteristics of emotional vocal expression. For example, the effects of facial expression and gestures, especially head movements or body movements toward or away from the listener, on the intensity and spectral distribution of the acoustic signal (see Laver, 1991) may become increasingly smaller with increasing distance.

The importance of factors affecting the transmission of emotional vocal expressions between sender and receiver is twofold. First, transmission constrains the way the voice can be used for emotional signaling, possibly influencing the vocal expression of some emotions more than others. Second, if vocal communication evolved largely as part of an emotional signaling system, it is likely that the mechanisms that underlie vocal production and perception evolved in such a way as to exploit such constraints. A greater understanding of the transmission stage of the Brunswikian lens model is thus likely to lead to insights about the design and functioning of vocal production and perception systems.

2. The proximal cues available to the listener are determined by the transform functions for auditory stimuli that are built into the human hearing system. Psychoacoustic research has demonstrated that the perceptual representation of sounds does not correspond in a one-to-one fashion to the sound's objectively measured acoustic properties. For example, the perceived loudness of voiced speech signals correlates more strongly with the amplitude of a few harmonics or even a single harmonic than with its overall intensity (Gramming & Sundberg, 1988; Titze, 1992). Furthermore, listeners apparently use their knowledge about the specific spectral distribution of harmonics and noise in loud and soft voices to decide the vocal effort with which a voice was produced. This judgment is still reasonably correct when both loud and soft voices are presented at the same perceived loudness level, with the soft voice having more than six times higher overall intensity than the loud voice (Klasmeyer, 1999, p. 112).

Psychoacoustic effects also play a role in the perception of F0 contours. F0 movements have to cross the glissando threshold before they can be perceived as melodic movements (see d'Allessandro & Mertens, 1995). The glissando threshold for a uniform F0 change with constant slope is

frequency dependent (about 0.16 times the square of F0 in synthetic stimuli and even higher in fluent speech), but changes in the slope (differential glissando threshold) are perceived more easily. If F0 does not change monotonically but oscillates regularly or irregularly around a fixed or slowly varying value, these changes are perceived as voice quality effects (tremor, jitter) rather than melodic movements.

Furthermore, F0 can influence the perceived duration of a spoken utterance. Higher F0 or an F0 rise at the end of an utterance is perceived as a faster speaking rate (Kohler, 1995). Spectral changes in fluent speech, which are caused either by fast articulatory movements or rapid F0 changes or both, result in a decrease of perceived phoneme duration, speech rate, and utterance duration (Klasmeyer, 1999, p. 49). There is only a weak correlation between physical signal duration and perceived duration of fluent speech segments. This also applies to perceived rhythm. Rhythm perception is based on the subsequence of accented syllables. But while time intervals between accented syllables in fluent speech might vary a great deal, the rhythm is generally perceived as being much more regular (Allen, 1975; Fraisse, 1963; Jakobson, Fant, & Halle, 1951). Speech rate and rhythm are used in synchronizing speaker and listener (Byers, 1976; Condon, 1986). This explains why the perceptual "equalization" of durations serves an important function in communication situations and why listeners are not very good at judging objective durations in fluent speech in a very precise manner.

All of these psychoacoustic properties affect which acoustic characteristics of emotional speech are likely to be the most perceptually salient. For example, if just a few harmonics of voiced speech determine its perceived loudness, these harmonics are likely to be particularly salient for the perception of emotions such as rage and elation. Similarly, the influence of F0 on perceived duration and speech rate implies that F0 will be an important perceptual cue for expressions of emotions such as anxiety and fear, which are perceived as involving more rapid speech.

The role of voice sound transmission and the transform functions specific to the human hearing system have been rarely studied in this area of research, mostly because researchers have focused exclusively on either encoding or decoding studies, rarely taking the complete vocal communication process into account. This is regrettable for theoretical reasons but also because we may have missed parameters that can be useful to differentiate the vocal expression of different emotions. For example, if one assumes that vocal emotion expressions need to be very distinctive to listeners because of their communicative function, it may pay to transform the objective signal characteristics into their perceptual counterparts using known psychoacoustic functions (e.g., perceived loudness, perceived pitch, perceived rhythm, or Bark bands in spectral analysis; see Zwicker, 1982).

Decoding: Listener Inference of Emotion from Voice and Speech

The large majority of studies in the area of the vocal communication of emotion have focused on the decoding or inference part of the Brunswikian lens model. Generally, the portrayal paradigm has been used, that is, asking actors to act out or simulate a number of different emotions, generally with standard content utterances. Groups of lay judges are then requested to infer the emotions portrayed in a series of vocal stimuli, mostly on rating sheets with preestablished lists of emotion labels. The data analysis consists of the computation of the percentage of stimuli per emotion that were correctly recognized, which is then compared with the percentage of responses that would be expected to be correct on the basis of chance guessing. In some studies, confusion matrices are reported showing the patterns of errors.

Empirical Evidence

The question of whether judges can recognize emotion solely on the basis of vocal cues has interested psychologists and psychiatrists from the beginning of the 20th century and was especially popular between 1950 and 1980 (see Scherer, 1979). As is generally the case in this research, a number of different emotions were vocally portrayed by professional or lay actors (using nonsense syllables or standard speech samples) and judges were asked to identify the emotion expressed. A review of approximately 30 of these early studies in which normal voice portrayals were used (excluding studies with pathological voice samples and filtered speech) yielded an average accuracy percentage of about 60%, or about five times higher than what would be expected by chance (Scherer, 1989). Since many of these early studies often used very short voice samples and included rarely studied emotions such as pride or jealousy (as compared to "basic" emotions such as anger, joy, sadness, or fear), this level of accuracy is quite remarkable. More recent studies reported similar levels of average recognition accuracy across different emotions. Van Bezooijen (1984) reports a mean accuracy of 65% for voice samples of disgust, surprise, shame, interest, joy, fear, sadness, and anger. Scherer, Banse, Wallbott, and Goldbeck (1991), who studied fear, joy, sadness, anger, and disgust as portrayed by professional radio actors, reported a mean accuracy of 56% across a variety of different types of listener and age groups.

Many of these studies can be criticized for using only a relatively small number of emotions, with often only one positive exemplar, thus constituting *discrimination* studies (deciding between alternatives) rather than *recognition* studies (recognizing a particular category in its own right). Therefore, it has been suggested to correct the accuracy

Table 23.4. Accuracy (%) of Facial and Vocal Emotion Recognition in Studies in Western and Non-Western Countries

	Neutral	Anger	Fear	Joy	Sadness	Disgust	Surprise	Mean
Facial/Western (20)		78	77	95	79	80	88	78
Vocal/Recent Western (11)	74	77	61	57	71	31		62
Facial/Non-Western (11)		59	62	88	74	67	77	65
Vocal/Non-Western (1)	70	64	38	28	58			52

Note: Empty cells indicate that the respective emotions have not been studied in these regions. Numbers in parentheses in column 1 indicate the number of countries studied.

Source: From Scherer (1999b).

coefficients for guessing by taking the number of given alternatives and the response distribution in the margins into account (Wagner, 1993). However, it is difficult to determine the appropriate procedure for such a correction, each method having a different type of drawback (Banse & Scherer, 1996). Alternatively, one can compute different comparison levels for chance guessing of different emotions. For facial expression, Ekman (1994) has suggested using different levels of chance accuracy for positive and negative emotions, given their differential frequency in the stimulus material. While this may be useful for facial expression where the *m. zygomaticus* action is often a giveaway for positive emotions, it may be less adequate for vocal expression studies since exuberant joy or elation shares many vocal characteristics with anger (see Banse & Scherer, 1996; Scherer et al., 1991).

In addition, none of these studies took into account that there are often different instantiations or variants of specific emotions, such as hot and cold anger, which can be seen as different members of the same family of emotions (see also Ekman, 1992). These variants may have rather different vocal characteristics, a fact which could explain some of the lack of replication observed in the literature (see Scherer, 1986). To deal with this problem, Banse and Scherer (1996) used a large stimulus set, consisting of 14 emotions (in several cases, two members of the same emotion family were used) portrayed by 12 professional theater actors. The average accuracy level was found to be 48% (as compared to 7% expected by chance if all 14 categories were weighted equally). If, in an attempt to compare the data to earlier studies, one computes the agreement *between families* only for those emotions where there were two variants (yielding 10 categories) the accuracy percentage increases to 55% (as compared to 10% expected by chance). In reviewing the evidence from the studies to date, one can conclude that the recognition of emotion from standardized voice samples, using actor portrayals, attains between 55% and 65% accuracy, about five to six times higher than what would be expected by chance.

This figure is about 15% lower than what has been found for the recognition of facially expressed emotions (see Ekman, 1994). The difference is mostly due to facial expressions of joy and disgust being recognized with close to 100% accuracy, based on facial actions that are highly specific for these emotions (smiling for joy and nose wrinkling for disgust; see Ekman & Friesen, 1978). Table 23.4 shows a comparison of accuracy figures for vocal and facial portrayal recognition obtained in a set of studies having examined a comparable range of emotions in both modalities (the data in this table are based on reviews by Ekman, 1994, for facial expression, and Scherer, Banse, & Wallbott, 2001, for vocal expression). The table shows that the figure for average percent accuracy hides some rather large differences between emotions. Sadness and anger are generally best recognized vocally, followed by fear. Joy has rather mixed accuracy percentages in different studies, possibly due to differences with respect to quiet happiness versus elated joy being portrayed by the actors. Vocal disgust portrayals are recognized very badly, often barely above chance level. Johnstone and Scherer (2000) discuss a number of potential explanations of these differences, based on the evolutionary pressure toward accurate vocal communication for different emotions. For example, there is clear adaptive advantage in being able to warn (in fear) or threaten (in anger) others in an unambiguous fashion over large distances, something for which vocal expression is ideally suited. In contrast, warning others of rotten food (as in disgust) may be most functional to conspecifics eating at the same place as the signaler, in which case facial expression may be more appropriate.

These differences in the vocal communication functions of different emotions underline the need to analyze the recognizability of these emotions separately. In addition, confusion matrices should be reported regularly, as errors are not randomly distributed and as the patterns of misidentification provide important information on the judgment process. Banse and Scherer (1996), analyzing the confusion patterns in their recognition data in great detail, showed that while disgust portrayals are generally confused with almost all other negative emotions, there are more specific patterns for the other emotions. As one might expect, there are many errors between members of the same emotion family (for example, hot anger is confused consistently only with cold anger and contempt). Other frequent errors occur between emotions with similar valence (for example, interest is confused more often with pride and happiness than with the other 11 emotions

taken together). Johnstone and Scherer (2000) argue that rather than considering confusions as errors, they should be interpreted as indicators of the similarity or proximity between emotion categories, taking into account quality, intensity, and valence. Emotions that are similar on one or more of these dimensions are obviously easier to confuse. An additional complication is due to individual differences in the use of emotion labels.

The analysis of confusion patterns is particularly important in cross-cultural studies of vocal emotion recognition. While such data are still rare in this area (especially compared with the abundance of cross-cultural studies on facial expression), there is some evidence (see Frick, 1985; van Bezooijen, Otto, & Heenan, 1983) that vocal expressions by members of one culture, for at least some emotions, are universally recognized by members of other cultures. Recently, Scherer, Banse, and Wallbott (2001) reported the findings of a series of cross-cultural studies including eight European and one Asian country. Vocal emotion portrayals (joy, anger, fear, sadness, neutral) by professional German radio actors were used. Local collaborators recruited college students to listen to these samples in small groups (using identical equipment in all countries) and to judge the intended emotions in a standardized rating procedure. The data show an overall accuracy percentage of 66% across all emotions and countries, with a high of 74% in Germany and a low of 52% in Indonesia. Joy was inferred with much lower accuracy (42%) than the other emotions (around 70%). Generally, accuracy decreased with increasing language dissimilarity from German in spite of the use of language-free speech samples (which suggests that portrayals may be characterized by culture and language-specific paralinguistic patterns that influence the decoding process). However, one of the striking results was that the patterns of confusion were similar across all countries, including Indonesia. This was interpreted as evidence for the existence of universal inference rules from vocal characteristics to specific emotions across cultures.

Identifying the Acoustic Cues Used in Emotion Inference from Voice

In order to understand the nature of such universal inference rules, it is important to identify the acoustic cues perceived and used by listeners in the process of attributing emotion to a speaker. While the issue of the accuracy of recognition has dominated this field so far, a more comprehensive process model (like the Brunswikian lens model described earlier) requires detailed study of the vocal cues that are attended to in voice perception and how they are interpreted with respect to the underlying emotional state. We review studies in which attempts were made to systematically isolate or measure specific acoustic cues in order to determine their role in the inference process used by judges trying to identify the emotion expressed in a series of voice samples. The purpose of this type of research is to identify the vocal characteristics that judges use to identify the expressed emotion.

Johnstone and Scherer (2000) provide a detailed overview of the different research strategies that have been used to determine the role of various acoustic cues in the judgment process. One of the most frequently used techniques is the isolation of specific acoustic cues, such as pitch or rhythm, via filtering, masking, or speech synthesis or resynthesis. Scherer, Feldstein, Bond, and Rosenthal (1985) and Friend and Farrar (1994) have compared different masking techniques (filtering, randomized splicing, playing backwards, pitch inversion, and tone-silence coding). Each of these techniques removes and/or preserves different combinations of acoustic characteristics of a vocal expression. Speech intelligibility is removed by all of these procedures, allowing the use of natural speech from "real-life" rather than artificially posed emotions. For example, Scherer, Ladd, and Silverman (1984) used a corpus of affectively laden utterances produced by civil servants interacting with citizens to determine which acoustic cues are used by listeners to infer speaker emotion and attitude. One of the interesting findings was that politeness, as a speaker attitude, was still recognizable in the most severely masked speech samples.

Researchers in this area have been very interested in methods that allow a rigorously experimental approach in which acoustic variables are systematically varied to determine their effects, in isolation and in interaction, on listener judgment of emotional and attitudinal speaker state. Acoustic *synthesis* methods have been used early on to study the impressions produced by different acoustic variables (see Lieberman & Michaels, 1962; Scherer & Oshinsky, 1977), allowing researchers to determine the effects of parameters such as amplitude variation, pitch level, contour and variation, tempo, envelope, harmonic richness, tonality, and rhythm on emotion attributions. The development of easily available speech resynthesis techniques in the 1970s made it possible to take neutral natural voices and systematically change different cues via digital manipulation of the sound waves. In a number of studies by Scherer and his collaborators, F0 level, contour variability and range, intensity, duration, and accent structure of real utterances have been systematically manipulated in order to observe the effect of listener attributions of emotion and speaker attitude (Ladd, Silverman, Tolkmitt, Bergmann, & Scherer, 1985; Tolkmitt, Bergmann, Goldbeck, & Scherer, 1988). With the availability and quality of synthesis and resynthesis methods having greatly increased in recent years, it is not surprising that a fairly large number of such studies have been reported. Recent progress in speech technology has multiplied the tools available for a systematic, experimental study of vocal cue use in emotion attribution (see Abadjieva, Murray & Ar-

nott, 1995; Burkhardt & Sendlmeier, 2000; Cahn, 1990; Carlson, Granström, & Nord, 1992; Granström, 1992; Heuft, Portele, & Rauth, 1996; Mozziconacci, 1995; Murray & Arnott; 1995; see also the special issue of the journal *Speech Communication*, 2002).

Apart from masking or resynthesis methods, it is possible to use *acoustic analyses* and/or voice experts to measure the acoustic and/or phonatory-articulatory characteristics of vocal emotion portrayals (acted or natural), and to correlate these with the listeners' judgments of underlying emotion or attitude of the speaker. Several studies of this type have yielded information concerning the vocal characteristics that determine judges' inference (van Bezooijen, 1984; Scherer et al., 1991; Wallbott & Scherer, 1986). The extensive data in the study conducted by Banse and Scherer (1996) statistically regressed the judges' emotion inferences on the various acoustic variables that had been obtained by digital speech analysis in order to determine which acoustic characteristics best predict the judges' attributions.

The highly significant results showed that a sizable proportion of the variance was explained by a set of about 9 to 10 variables, including mean F0, standard deviation of F0, mean energy, duration of voiced periods, proportion of energy up to 1000 Hz, and energy drop-off in the spectrum. For hot anger 36% of the variance could be explained by this set of variables; for the majority of the remaining emotion categories, 10 to 25% of the variance could be accounted for. To infer hot anger, judges used the cues of high F0 and strong F0 variation as well as a strong difference between energy in the lower and higher frequency bands. Examples for other cue-inference associations are low intensity = sadness; F0 increase, low F0 variability, and low intensity = anxiety; low F0, low intensity, and slow speech rate = boredom. Banse and Scherer (1996) also compared the performance of human judges to statistical classification algorithms such as discriminant analysis and jackknifing in order to determine whether the acoustic measures correlated with listener judgments do indeed allow the discrimination of the emotions used. They found that the discrimination rates were similar for human judges and statistical algorithms. Unexpectedly, they also found a high degree of correspondence between the confusion matrices, suggesting that the human inference rules mirror the underlying patterns of association between acoustic parameters and emotion class (as portrayed by actors).

While the analysis of the correspondence between human judgment and statistical classification algorithms is interesting, it does not replace an analysis based on the Brunswikian lens model, using path analysis or structural modeling. This kind of analysis would allow us to determine which distal vocal cues characterize specific emotions, how these cues are transmitted and proximally represented in listeners, and which patterns of inferences are based on those cues. Unfortunately, no comprehensive Brunswikian analysis of the vocal emotion communication process has yet been published.

Neglected Issues

While the empirical research on the recognition of emotion from vocal expression has yielded interesting results, some aspects of the process have been neglected. One intriguing issue concerns the dimensions of affect that listeners tend to infer from the voice under natural conditions. In most studies, listener-judges are asked to identify discrete emotions, generally by checking one of several categories provided by the researcher, thus being forced to infer a category (see the controversy between Ekman, 1994, and Russell, 1994). In normal social interaction, this might not be the case. Listeners might tend to use the vocal characteristics to infer arousal of the speaker, an action tendency of the speaker (e.g., impending aggression), or the speaker's appraisal of an event (e.g., perceived urgency of action).

The issue of what is being expressed by emotional expression has been hotly debated. Thus, Fridlund (1994) has proposed that facial expressions do not express emotions but must be considered as social signals. However, researchers in the area of emotional expression, starting with Darwin, have always recognized that one of the essential functions of emotional expression is the social signaling of the expressor's reaction and action tendency (see Scherer, 1984, 1985). As suggested by Bühler (1934; see Scherer, 1988), all signs, including emotional expressions, have representation (meaning), symptom (expression), and appeal (social signaling) functions at the same time. What is more interesting than a sterile debate about what is being expressed is the issue of what is being inferred by listeners in particular contexts. Future research in this area may find this a rewarding question to ask.

Selected Topics of Major Research Interest

We next deal with a number of special issues that show that the study of vocal emotion expression is even more complex than what has been intimated previously. These complexities are mostly due to the fact that we generally express emotion when we produce *speech*. Linguistic and expressive cues are intimately intertwined. This is particularly clear for prosody, which plays a major role in both language and expression. Another interesting issue is provided by tone languages that use F0, normally a mainstay for expression, for semantic purposes.

The Role of Prosody

In the section on the encoding of emotion in the voice, it was noted that few vocal characteristics had been identi-

fied that successfully differentiated emotions of similar arousal. Prosody or intonation, which is often considered a prime carrier of such affective information, has been almost completely neglected by researchers in this area (with a few notable exceptions, e.g., Fonagy, 1981; Fonagy & Madgics, 1963, Ladd et al., 1985; Paeschke, Kienast, & Sendlmeier, 1999; Scherer et al., 1984). One of the reasons for this neglect is the lack of a consensual definition of "intonational form" (see, for example, Beckman, 1995; Ladd, 1995; Möbius, 1995), specifying how intonation should be empirically measured.

Scherer et al. (1984) suggested two general principles underlying the coding of emotional information in speech, covariation and configuration. The *covariation principle* assumes a continuous but not necessarily linear relationship between some aspect of the emotional response and a particular acoustic variable. Thus, if F0 is directly related to physiological arousal, F0 will be higher in rage as compared to mild irritation. Such continuous relationships can be assessed by standard statistical covariance measures.

In contrast, almost all linguistic descriptions assume that intonation involves a number of categorical distinctions, analogous to contrasts between segmental phonemes or between grammatical categories. In consequence, the *configuration principle* implies that the specific affective meaning conveyed by an utterance is actively inferred by the listener based on the total configuration of the linguistic choices in the context, based on phonological categories such as "falling intonation contour." In consequence, statistical assessment of the affective coding based on the configuration principle requires combinatorial analysis of category variables rather than scalar covariance.

Voice quality is generally coded according to the covariation principle and phonemic structure according to the configuration principle. The central acoustic variable that underlies intonation—F0—may be coded according to both types of principles, depending on specific features of its dynamic change over time. For example, *final pitch movements* are coded by the configuration principle: final-rise versus final-fall patterns of F0 in themselves do not carry emotional meanings, but they are linked to sentence modes such as question versus non-question. However, it can be shown that context, such as type of sentence, affects interpretation. While a falling intonation contour is judged as neutral in a WH-question, it is judged as aggressive or challenging in a yes/no question (Scherer et al., 1984). In contrast, F0 range shows a covariance relationship with attitudinal and affective information (Ladd et al., 1985). Further evidence for covariance coding of F0 has been found in other studies as well. For example, newborn infants (who do not know about phonological categories) are able to decode simple emotional meanings from intonation patterns (Papousek, 1994).

Ladd et al. (1985) suggested that *overall F0 range and voice quality* might reflect arousal, while differences of *prosodic contour type* signal differences of more cognitively based speaker attitudes. Alternatively, it could be hypothesized that continuous variables are linked to push effects (externalization of internal states), while configurations of category variables are more likely to be linked to pull effects (specific normative models for affect signals or display). More empirical data are required to test these assumptions experimentally.

In the last two decades a few studies of emotion effects on prosody were conducted using either analysis of emotional speech material or digital resynthesis to test the perceived effect of specific modifications. For example, Ladd et al. (1985) in a perception study systematically manipulated F0 range and contour to investigate the effects on emotion inference. To prepare the stimuli, a linguistic theory of intonation was first used to identify relative peaks on accented syllables (Garding & Bruce, 1981; Ladd, 1983; Liberman & Pierrehumbert, 1984). Then a logarithmic model was used to manipulate F0 range and F0 contour (uptrend vs. downtrend) by shifting relative F0 peaks in the accented syllables while preserving the declination line. These resynthesized stimuli were presented to listeners who were asked to judge the emotional message conveyed in the different representations of the sentence. The results showed strong main effects for these vocal cue manipulations: Wide F0 range and uptrend F0 contours were seen as signals of arousal, annoyance, and involvement. Narrow F0 range was seen as a sign of sadness or of absence of specific speaker attitudes. Uptrend contours also signaled greater emphasis, stronger contradiction, and less cooperativeness. Of all variables studied, F0 range had the most powerful effect on judgments. Relatively few effects due to interactions between the manipulated variables were found. This implies that the synthesized variables independently influenced judges' ratings. Only very minor effects for speaker and utterance content were found, indicating that the results are likely to generalize over different speakers and utterances. In subsequent perception studies of a similar nature (Goldbeck, Tolkmitt, & Scherer, 1988; Ladd, Scherer, & Silverman, 1986) modifications of mean F0, F0 perturbation, average intensity, "uptrend" versus "downtrend" contours, accent height in sentence-final (second) accent position, relative height of subsequent local maxima in the F0 contour, and durations of accented syllable were manipulated. The results showed that emotion inferences are systematically related to changes in these parameters.

Analysis studies, assessing prosodic features in natural or simulated speech material (Klasmeyer, 1999; Paeschke et al., 1999), have mainly focused on (1) frequency (expressed in percentage) of unstressed, weakly stressed, moderately stressed, and heavily stressed segments (as based on auditory impression (of lay or expert judges); (2)

F0 stylizations (composition of straight, rising, and falling lines), particularly the frequency (%) and the gradients of rise/straight/fall segments; (3) F0 contours (on the basis of a complete phrase expressed in stylized syllables steps, identifying the location of F0 maxima and main stresses); (4) F0 contours on syllable basis comparing relative locations of local F0 maxima; and (5) microprosody (e.g., intonation changes in small segments).

Evidence from Prosody for the Predictions of the Three Emotion Theories

All emotional information with respect to prosody should be coded exclusively in average F0 and the variability of contours if arousal accounted for all emotion-caused vocal differentiation. However, Table 23.2 provides some evidence that specific prosodic patterns can be used to code discrete emotional meanings or appraisal results of specific situations. For example, in monosyllabic utterances, F0 maxima seem to appear earlier in the case of happiness and later in the case of anger expressions (Klasmeyer, 1999). Thus, the limited evidence to date (as well as persuasive case studies, e.g., Fonagy, 1983) point to the existence of prosodic patterns that can differentiate emotional states in a qualitative fashion. As mentioned above, it would seem that discrete-emotion theorists should postulate a relatively small number of standard patterns (although there my be blends) for basic emotions. So far, there is limited evidence for such a restricted number, even within any one language. Given the interaction between emotion-generated prosodic changes and language- or dialect-specfic prosodic features, it seems unlikely that emotion-specific prosodic patterns for the standard basic emotions will be found. Appraisal theories do not introduce such constrained predictions. However, few concrete predictions for prosodic effects of individual appraisal checks have so far been ventured, let alone tested. More than in most other areas described in this chapter, further research on emotion effects on prosody is urgently needed.

Choice of Intonational Models

This future work will have to develop prosody or intonation *models* that reduce the more or less continuous variety of possible intonation realizations to a limited number of more or less well-defined categories and link these to emotional variations. Unfortunately, a large number of rather different intonation models are proposed currently, and there is no established procedure for comparative evaluation, providing clear criteria for preferring one model over another.

Linguists classify intonation models into two major categories: (1) *hierarchically organized models* that interpret F0 contours as a complex pattern resulting from the superposition of several components. For example, F0 curves can be considered as the superposition of relatively fast pitch movements on a slowly falling F0 line, the declination line or "baseline." Models of this type are called two-component models of intonation. One component represents the global aspect of the pitch curve, the declination line and relative pitch level; the other component represents pitch-movement variations and is related to pitch range; (2) *accent models* that interpret F0 as a sequence of phonologically distinctive tones or categorically different pitch accents that are locally determined and do not interact. This type of intonation model considers F0 targets rather than pitch movements. This is a matter of lively debate, too technical to be reproduced. As always, each model has both advantages and disadvantages for the purpose at hand.

The Study of Tone Languages

The issues concerning prosody that have been discussed are intimately tied with the role of F0 in intonation as part of Western languages. This makes it difficult to examine the effects of psychobiological push and sociolinguistic pull factors on this fundamental aspect of emotional vocalization. For this reason, it is interesting to study non-Western languages in which F0 has yet another function, that of semantic differentiation. Tone languages, which are common in parts of East Asia and Africa, are those that use pitch variation to convey lexical information, as opposed to most Indo-European languages that do not. For example, in Thai, there are five contrastive tones used to distinguish among different syllables otherwise consisting of the same sequence of consonants and vowels. Because pitch is the principal carrier of prosody, including affective prosody, a potential conflict between prosodic and tonal-lexical information exists in tone languages. One might expect that the use of pitch to convey affective information is restricted in tone languages since pitch is already being used for lexical purposes. Giving credence to this hypothesis, some research has shown less use of short-term changes in fundamental frequency to express emotion in tone languages (Ross, Edmondson, & Seibert, 1986). In their research, Ross et al. performed acoustic analyses of a standard sentence spoken with happy, sad, angry, surprised, and neutral affective prosody by five speakers each of Mandarin, Taiwanese, Thai, and American English. They discovered that for all of the tone languages, the use of F0 variation and F0 slope to express affect was reduced compared to American English. In the three tone languages, the shape of the intonation contours of affective utterances differed less from neutral utterances than was the case for American English. It seems that the use of a particular acoustic feature (in this case, F0) in spoken language can constrain its use for the communication of emotion. It remains an open question whether this constraint reduces the ability to prosodically express

affect in tone languages. One possibility is that parameters other than F0, such as intensity or speech rate, are used more in tone languages than in non-tone languages to compensate for reduced availability of F0. Ross et al., however, found no systematic differences between American English and the three tone languages for other measured parameters, such as intensity and rate of speech.

One might also argue that the lexical use of tones might render the language more susceptible to interference from expressed affect. Indeed, if one accepts that the effects of emotion on vocal production are due at least in part to involuntary physiological changes, for example, to the respiratory and laryngeal musculature, one might expect the expression of emotion in tone languages to interfere with lexical coding. Such interference would be manifest in increased speech comprehension errors by listeners to affective speech in tone languages. To our knowledge, no research has directly addressed this question.

Neuropsychological Approaches to the Study of Vocal Communication

Neuropsychology, in particular the study of specific communicative deficits resulting from localized damage to the brain, provides evidence directly pertinent to questions of how the brain processes affective information in voice and speech. We will briefly point to the comparative perspective in neuropsychological assessment of vocalization since there is much to suggest that the characteristics of affective speech in humans reflect a phylogenetically continuous evolution of the affect system, including vocal affect (Hauser, 1996; Robinson, 1972; Scherer, 1985; see chapter 24, this volume). The assumption that the human emotion system is the result of continuous modification of and addition to a more primitive, phylogenetically older emotion system forms the basis of many modern emotion theories (LeDoux, 1996; Panksepp, 1998). It follows from such an assumption that similarities will exist between human affective signaling and signaling observed in other animals, particularly those phylogenetically close to humans. Furthermore, because other animals lack spoken language and, to a certain extent, the explicit cognitive abilities of humans (which are presumably responsible for many of the pull effects on affective speech), the study of these animals should lead to insights into the way in which emotion directly affects vocal production.

Jürgens (1979, 1988, 1994) has suggested that the functional evolution of vocalization is reflected in three hierarchically organized neural systems, the second of which—consisting of the mid-brain periaqueductal gray, parts of the limbic system including the hypothalamus, midline thalamus, amygdala and septum, and the anterior cingulate cortex—is also central to the generation and regulation of emotion. This system is thought to be responsible for the initiation of vocalization and selection from a number of different vocal patterns, presumably as part of an emotional signaling system making up part of a more general emotion response system. Consistent with the ideas of Jürgens, the "polyvagal theory of emotion" proposed by Porges (1997) posits a special role of the ventral vagal complex, which includes both somatic and visceral efferent neural fibers of the cranial nerves, in both the rapid regulation of metabolism to meet environmental challenges and the production of emotional expressions, including vocalizations. Organized neural structures have also been identified in the midbrain periaqueductal gray (PAG) that are involved in differentiated emotional responses and are also fundamental to the control of respiratory and laryngeal functioning during vocal production (Bandler & Shipley, 1994; Davis, Zhang, Winkworth & Bandler, 1996). In work with cats, stimulation of such PAG nuclei leads to specific types of hissing, mewing, or howling vocal patterns, depending on exactly which parts of the PAG are stimulated. When these results are integrated, a picture develops of an emotional expression system that makes up part of a more general emotional response system, in which specific brain stem nuclei, activated via pathways from the limbic system, coordinate the activation of groups of laryngeal and respiratory muscles leading to specific vocal patterns. Although such a signaling system seems to support the theory of emotion-specific, motor-expressive programs as posited by Ekman (1992) and Izard (1971), it remains to be seen if such neural mechanisms are organized around specific emotions or around appraisals, as suggested by Scherer (1986) and other appraisal theorists (see Smith & Scott, 1997). The extent to which such mechanisms are the exclusive result of automatic processing or might (particularly in humans) also be engaged in a controlled manner remains unclear. Furthermore, whether the neural mechanisms underlying animal vocalizations, which resemble most closely human affect bursts (see Scherer, 1994), are also responsible for the prosodic modification of speech, remains unanswered.

In the case of the human brain, it has long been generally accepted that processing of segmental, linguistic and grammatical speech information is concentrated in specialized centers in the left hemisphere. The most well-researched left-hemisphere speech centers are the left inferior frontal area (Broca's area) and the left temporo-parieto-occipital area (Wernicke's area), lesions on which lead to fluency and syntactic, and comprehension and lexical deficits, respectively. Due largely to a relative dearth of research, no such consensus exists for the localization (if it indeed exists) of nonlinguistic processing of affective information in speech. Hughlings-Jackson (1915) observed that patients with severe linguistic deficits due to brain damage still had the ability to communicate emotions through voice, and he suggested that such functions might be subserved by the right hemisphere. It was not until the 1970s that neurological evidence of a right-hemisphere

specialization for affective speech comprehension was forthcoming (Heilman, Scholes, & Watson, 1975; Tucker, Watson, & Heilman, 1977). Since then, a number of studies comparing listeners with unilateral right-hemisphere brain damage (RHD) to listeners with left-hemisphere damage (LHD) have reported a deficit in the perception of affective prosody in RHD listeners (e.g., Bowers, Coslett, Bauer, Speedie, & Heilman, 1987; Heilman, Bowers, Speedie, & Coslett, 1984, Peper & Irle, 1997; Ross, 1981).

Not all studies support such a right-hemisphere lateralization for decoding emotional prosody, however, with several reporting similar performance by patients with right-hemisphere damage to those with left-hemisphere damage (e.g., Cancelliere & Kertesz, 1990; Tompkins & Flowers, 1985). Based on a detailed analysis of how the prosodic acoustic parameters were used by RHD and LHD patients in classifying (and misclassifying) emotional speech stimuli, Van Lancker and Sidtis (1992) suggested that the appearance of a right-hemisphere specialization for affective prosody might be the result of a more general hemispheric specialization for processing certain types of acoustic information. According to this view, disruption in RHD patients of processing of pitch level and variability, which are known to be major carriers of emotion information in speech (Banse & Scherer, 1996; Scherer, 1986), leads to deficient decoding of affective prosody. This explanation fits well with studies that have shown a right-hemisphere specialization for the processing of analog information—in particular, continuously changing acoustic parameters such as fundamental frequency and spectral energy distribution—in contrast to left-hemisphere specialization for the processing of categorical and temporal information (Fitch, Miller, & Tallal, 1997; Robin, Tranel, & Damasio, 1990). A corollary of this hypothesis is that types of information other than emotion that are conveyed in an analog manner, such as speaker identity, should also be decoded predominantly in the right hemisphere. Van Lancker, Kreiman, and Cummings (1989) found such a lateralization in the recognition of familiar voices but not in the discrimination between unfamiliar voices. The explanation given for these and similar results with other types of stimuli was that a right-hemisphere specialization exists for the processing of personally relevant and familiar information (Van Lancker, 1991). This position contradicts the theory that hemispheric lateralization reflects prosodic structure rather than affective content (see also Baum & Pell, 1999).

The lack of consistent neuropsychological findings regarding the localization of the perception of affective and nonaffective prosody is probably due in part to the variability in lesion size and location in brain-damaged patients. Lesions not only commonly span multiple cortical regions but also include subcortical areas, making it difficult to determine precisely which areas are implicated in prosodic processing. Lesion studies can also be criticized on the grounds that they do not directly address questions concerning the mental processes of healthy, non-brain-damaged individuals. Functional brain imaging techniques present an alternative method, allowing greater spatial and temporal isolation of functionally implicated brain regions, as well as an examination of prosodic processing in non-brain-damaged listeners.

To date, only a small number of prosodic imaging studies have been reported. In a recent positron emission tomography (PET) study, George et al. (1996) reported greater right prefrontal activity during processing of the emotional prosody than during processing of the emotional propositional content of spoken sentences. Pihan, Altenmüller, and Ackermann (1997) reported a right-hemisphere lateralization in DC components of the scalp electroencephalography (EEG) signal for the perception of both temporal (accented syllable duration) and frequency (F0 range) mediated emotional prosody. Imaizumi, Mori, Kiritani, Hosoi, and Tonoike (1998), using magnetoencephalography (MEG), found evidence supporting the existence of prosody-specific right-hemisphere processing, as well as the involvement of certain left hemisphere centers in both linguistic and prosodic processing. In a separate study, Imaizumi et al. (1997) found that regional cerebral blood flow, as measured by PET, indicated that a number of areas of both cerebral hemispheres were differentially involved in the identification of speaker versus the identification of emotion in spoken words. This latter study hints that some of the contradictory results obtained in lesion studies of the localization of affective and nonaffective prosodic perception might be resolved with techniques that have more precise spatial resolution, such as PET and functional magnetic resonance imaging (fMRI). Of these two techniques, fMRI has the advantage of allowing the functional image to be accurately overlaid upon a concurrently acquired structural image of the brain. FMRI is also noninvasive, in that no radioactive tracers are injected into the subject's body.

In summary, there is little consensus as to the extent and functional significance of the localization of decoding of affective and nonaffective prosody in speech. Certain researchers claim a special link between decoding of affect and hemispheric specialization (e.g., Ross, Homan, & Buck, 1994), while others propose that perception of prosody is a more multifaceted process that depends both on the structure of the parts of the speech signal being processed, as well as the information contained therein (Van Lancker, 1991).

Applied Research: Vocal Indicators of Affective Disturbance and Therapy Outcomes

The analysis of vocal emotion expression is used increasingly in applied settings, including health psychology,

consumer psychology, speech technology, media psychology, and many other areas in which speech plays a major role in daily life. Since it would be impossible to cover the entire gamut of applications in the context of this chapter, we are focusing one specific area: clinical psychology and psychiatry.

The study of the nonlinguistic aspects of speech, such as the vocal expression and perception of emotion, is directly pertinent to a number of applied issues in clinical domains. Clinical psychologists and psychiatrists have long used the tone of voice of a patient as a valuable indicator of the patient's mental state, but have had to rely on subjective and intuitive assessments rather than objective measures of the patient's vocal characteristics. Attempts were made to use acoustic measures of voice and speech measurement as early as the beginning of the 20th century in order to evaluate the diagnostic value of vocalization for the study of depression (Isserlin, 1925; Scripture, 1921; Zwirner, 1930). Following these pioneers, there have been many efforts to investigate this issue empirically (see reviews in Darby, 1981; Maser, 1987).

One of the clinical syndromes that has been most frequently studied with respect to vocal characteristics is depression, often using speech rate and F0 as indices. Normally, rate of speech tends to slow down in sad or depressed states, and this is indeed what is found in virtually all of the studies on emotion encoding (portrayals by actors or laymen; see Pittam & Scherer, 1993, for a review). Furthermore, in studies in which depressed states are induced experimentally, rate or tempo of speech goes down (e.g., Natale, 1977). Empirically, reduced rate or tempo of spontaneous speech in depression is a frequently reported finding and seems to be a stable phenomenon (see review by Siegman, 1987). On the whole, there is rather good evidence that patients in acute and severe depressive states are likely to speak more slowly and with longer pauses. Therefore, an increase in speech tempo and shortening of pauses may well indicate therapeutic success or remission from a depressive state. What is less clear is the mechanism that underlies this phenomenon.

With respect to F0, most studies have reported a rather low mean F0 for depressives in relation to normals, or decreased F0 in an acute state of depression, although there are reports of an increase in F0 with the severity of depression (see review in Scherer, 1987). The opposite pattern was found in studies on vocal changes following therapy. Most studies have found a decrease in F0 after therapy or during positive mood states (only one study suggested an increase in F0 and no significant effects were found in two other studies (see review in Scherer, 1987). The apparent discrepancy in the results may be explained at least in part by the fact that in most studies no clear distinction between manic and depressed states seems to have been made.

In a large-scale longitudinal study of depressive disorders, Ellgring and Scherer (1996) showed, as predicted, that an increase in speech rate and a decrease in pause duration are powerful indicators of mood improvement in the course of therapy (remission from depressive state). For F0, there were interesting sex differences. In female but not in male patients, a decrease in minimum fundamental frequency of the voice predicted mood improvement. The authors suggest that these differences may be due to differences in the emotions underlying depression. There are some indications that these could consist of suppressed anger in men and resignation in women. These and other data show the great promise of using vocal analysis as unobtrusive markers of affective change in patients suffering from emotional disorders.

Conclusions

In closing, we want to reiterate the need for a more comprehensive model as a grounding for research on the process of the vocal communication of emotion, such as the Brunswikian lens model according to which this chapter was structured. As mentioned repeatedly, most studies have focused on only one aspect of the total process, generally encoding or decoding.

As has become abundantly clear in reviewing the evidence to date, such a one-sided approach implies the danger of neglecting the important interactions between expression and impression. Future theorizing and research may profit from modeling the vocal communication process in its complete form. This may help to better understand the recursive relationships between push and pull effects in encoding and decoding. For animal communication, Leyhausen (1967) has provided a powerful demonstration of how the reception requirements of the receiver (pull effects) can shape the expression of the sender (push effects). Similarly, the fact that objective signal characteristics are transformed by the hearing mechanism of the receiver requires one to study the complete process of encoding of speaker states in distal form, transmission of the signal, and perception and interpretation by the receiver.

Many of the inconsistencies in the findings in the literature to date might be interpretable upon having access to the missing pieces of the puzzle. In addition, studying vocal parameters that are based on perceptual cues rather than objective acoustic-phonetic signals, as has been the case in research in the past, may also help to better understand the process of emotion inference from the voice, including individual, group, and cultural differences.

Apart from studying the vocal communication process as a whole, it may also be time to drop the assumption of separate linguistic and nonlinguistic channels, together with the hermetic separation of the respective research traditions. As we have shown, there is much evidence that

a large part of emotion signaling in voice and speech is dually coded, in both linguistic and nonlinguistic features. Thus, a rapprochement between researchers interested in expression and those interested in language (see chapter 27, this volume) is highly desirable, as is a more intensive interaction between researchers studying vocal and facial expression, two research areas that have had little contact so far, even though they have a common origin in the underlying emotion and are often interpreted as a Gestalt by the perceiver. The evolution of an affective science may help to create a context for theory and research that will encourage such a rapprochement.

REFERENCES

Abadjieva, E., Murray, I. R., & Arnott, J. L. (1995). Applying analysis of human emotional speech to enhance synthetic speech. *Proceedings Eurospeech, 95*, 909–912.

Allen, G. D. (1975). Speech rhythm: Its relation to performance universals & articulatory timing. *Journal of Phonetics, 3*, 75–86.

Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8*(2), 53–57.

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*(4), 219–224.

Bandler, R., & Shipley, M. T. (1994). Columnar organization in the mid-brain periaqueductal gray: Modules for emotional expression? *Trends in Neuroscience, 17*, 379–389.

Banse, R., & Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636.

Baum, S. R., & Pell, M. D. (1999). The neural bases of prosody: Insights from lesion studies and neuroimaging. *Aphasiology, 13*(8), 581–608.

Beckman, M. E. (1995). Local shapes and global trends. *Proceedings XIIIth ICPhS, 95*, Stockholm, Sweden, Vol. 2, pp. 100–108.

Bowers, D., Coslett, H. B., Bauer, R. M., Speedie, L. J., & Heilman, K. M. (1987). Comprehension of emotional prosody following unilateral hemispheric lesions: Processing defect versus distraction defect. *Neuropsychologia, 25*, 317–328.

Burkhardt, F., & Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland, September 5–7, 2000. Raleigh, NC: International Society for Computers and Their Applications.

Bühler, K. (1934). *Sprachtheorie* [Theory of speech]. Jena: Fischer (new ed. 1984).

Byers, P. (1976). Biological rhythms as information channels in interpersonal communication behavior. In P. P. G. Bateson & P. H. Klopfer (Eds.), *Perspectives in ethology*, Vol. 2 (pp. 135–164). New York: Plenum.

Caffi, C., & Janney, R. W. (1994). Toward a pragmatics of emotive communication. *Journal of Pragmatics, 22*, 325–373.

Cahn, J. (1990). The generation of affect in synthesised speech. *Journal of the American Voice I/O Society, 8*, 1–19.

Campbell, N. (2000). Databases of emotional speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland, September 5–7, 2000, 114–121. Raleigh, NC: International Society for Computers and Their Applications.

Carlson, R., Granström, B., & Nord, L. (1992). Experiments with emotive speech—acted utterances and synthesized replicas. *Proceedings ICSLP, 92*(1), 671–674.

Cancelliere, A., & Kertesz, A. (1990). Lesion localization in acquired deficits of emotional expression and comprehension. *Brain and Cognition, 13*, 133–147.

Condon, W. S. (1986). Communication: Rhythm and structure. In J. Evans & M. Clynes (Eds.), *Rhythm in psychological, linguistic, and musical processes* (pp. 55–77). Springfield, IL: Charles C. Thomas.

d'Alessandro, C., & Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language, 9*, 257–288.

Darby, J. K. (Ed.) (1981). *Speech evaluation in psychiatry*. New York: Grune & Stratton.

Darwin, C. (1872/1978). *The expression of emotions in man and animals* (3rd ed.). P. Ekman (Ed.). London: HarperCollins.

Davis, P. J., Zhang, S. P., Winkworth, A., & Bandler, R. (1996). Neural control of vocalization: Respiratory and emotional influences. *Journal of Voice, 10*, 23–38.

de Gelder, B., & Vroomen, J. (2000). Bimodal emotion perception: Integration across separate modalities, cross-modal perceptual grouping or perception of multimodal events? *Cognition and Emotion, 14*, 321–324.

Douglas-Cowie, E., Cowie, R., & Schröder, M. (2000). A new emotion database: Considerations, sources and scope. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland, September 5–7, 2000. Raleigh, NC: International Society for Computers and Their Applications.

Duffy, E. (1962). *Activation and behavior*. New York: Wiley.

Ekman, P. (1972). Universals and cultural differences in facial expression of emotion. In J. R. Cole (Ed.), *Nebraska Symposium on Motivation* (pp. 207–283). Lincoln: University of Nebraska Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3/4), 169–200.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin, 115*, 268–287.

Ekman, P., & Friesen, W. V. (1978). *The Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Ellgring, H., & Scherer, K. R. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior, 20*, 83–110.

Fant, G. M. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fitch, R. H., Miller, S., & Tallal, P. (1997). Neurobiology of speech perception. *Annual Review of Neuroscience, 20*, 331–353.

Fonagy, I. (1981). Emotions, voice and music. In J. Sundberg (Ed.). *Research aspects on singing* (pp. 51–79). Stockholm: Royal Swedish Academy of Music; Paris: Payot.

Fonagy, I. (1983). *La vive voix [The speaking voice]*. Paris: Payot.
Fonagy, I., & Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift für Phonetik, 16*, 293–326.
Fraisse, P. (1963). *The psychology of time*, New York: Harper & Row.
Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin, 97*, 412–429.
Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. New York: Academic Press.
Friend, M., & Farrar, M. J. (1994). A comparison of content-masking procedures for obtaining judgments of discrete affective states. *Journal of the Acoustical Society of America, 96*(3), 1283–1290.
Garding, E., & Bruce, G. (1981). A presentation of the Lund model for Swedish intonation. In T. Fretheim (Ed.), *Nordic prosody II* (pp. 33–39). Trondheim: TAPIR.
George, M. S., Parekh, P. I., Rosinsky, N., Ketter, T. A., Kimbrell, T. A., Heilman, K. M., Herscovitch, P., & Post, R. M. (1996). Understanding emotional prosody activates right hemisphere regions. *Archives of Neurology, 53*, 665–70.
Goldbeck, T., Tolkmitt, F., & Scherer, K. R. (1988). Experimental studies on vocal communication. In K. R. Scherer (Ed.), *Facets of emotion* (pp. 119–138). Hillsdale, NJ: Erlbaum.
Gramming, P., & Sundberg, J. (1988). Spectrum factors relevant to phonetogram measurement, *Journal of the Acoustical Society of America, 83*, 2352–2360.
Granström, B. (1992). The use of speech synthesis in exploring different speaking styles. *Speech Communication, 11*, 347–355.
Hauser, M. D. (1996). *The evolution of communication*. Cambridge, MA: MIT Press.
Heilman, K. M., Bowers, D., Speedie, L., & Coslett, H. B. (1984). Comprehension of affective and nonaffective prosody. *Neurology, 34*, 917–921.
Heilman, K. M., Scholes, R., & Watson, R. T. (1975). Auditory affective agnosia: Disturbed comprehension of affective speech. *Journal of Neurology, Neurosurgery and Psychiatry, 38*, 69–72.
Heuft, B., Portele, T., & Rauth, M. (1996). Emotions in time domain synthesis. *Proceedings ICSLP 96*(3), 1974–1977.
Hughlings-Jackson, J. (1915). On affections of speech from diseases of the brain. *Brain, 38*, 106–174.
Imaizumi, S., Mori, K., Kiritani, S., Hosoi, H., & Tonoike, M. (1998). Task-dependent laterality for cue decoding during spoken language processing. *NeuroReport, 9*, 899–903.
Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., Itoh, K., Kato, T., Nakamura, A., Hatano, K., Kojima, S., & Nakamura, K. (1997). Vocal identification of speaker and emotion activates different brain regions. *NeuroReport, 8*, 2809–2812.
Isserlin, M. (1925). Psychologisch-phonetische Untersuchungen. II. Mitteilung. (Psychological-phonetic studies. 2nd communication). *Zeitschrift für die Gesamte Neurologie und Psychiatrie, 94*, 437–448.
Izard, C. E. (1971). *The face of emotion*. New York: Appleton-Century-Crofts.
Izard, C. E. (1977). *Human emotions*. New York: Plenum.
Jakobson, R., Fant, C., & Halle, M. (1951). *Preliminaries to speech analysis*. Cambridge, MA: MIT Press.
Johnstone, T. (1996). Emotional speech elicited using computer games. *Proceedings ICSLP 96*(3), 1985–1988.
Johnstone, T. (2001). *The communication of affect through modulation of non-verbal vocal parameters*. Ph.D. Thesis. University of Western Australia.
Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. In M. Lewis & J. Haviland (Eds.). *Handbook of emotion* (2nd ed.) (pp. 220–235) New York: Guilford Press.
Johnstone, T., van Reekum, C. M., & Scherer, K. R. (2001). Vocal correlates of appraisal processes. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.). *Appraisal processes in emotion: Theory, methods, research* (pp. 271–284). New York: Oxford University Press.
Jürgens, U. (1979). Vocalization as an emotional indicator. A neuroethological study in the squirrel monkey. *Behaviour*, 69, 88–117.
Jürgens, U. (1988). Central control of monkey calls. In D. Todt, P. Goedeking, & D. Symmes (Eds.). *Primate vocal communication* (pp. 162–170). Berlin: Springer.
Jürgens, U. (1994). The role of the periaqueductal grey in vocal behaviour. *Behavioural Brain Research, 62*, 107–117.
Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1*(4), 381–412.
Kappas, A. (1997). His master's voice: Acoustic analysis of spontaneous vocalizations in an ongoing active coping task. *Psychophysiology, 34*, S5. (abstract)
Kappas, A., Hess, U., & Scherer, K. R. (1991). Voice and emotion. In B. Rimé & R. S. Feldman (Eds.), *Fundamentals of nonverbal behaviour* (pp. 200–238) New York: Cambridge University Press.
Karlsson, I., Banziger, T., Dankovicova, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., & Scherer, K. (2000). Speaker verification with elicited speaking styles in the VeriVox project. *Speech Communication, 31*(2–3), 121–129.
Klasmeyer, G. (1999). Akustische Korrelate des stimmlich emotionalen Ausdrucks in der Lautsprache [Acoustical correlates of emotional expression in voice.]. In H.-W. Wodarz, G. Heike, P. Janota, & M. Mangold (Eds.), *Forum Phoneticum, 67* (pp. 1–238). Frankfurt am Main: Hector.
Klasmeyer, G., & Sendlmeier, W. F. (1997). The classification of different phonation types in emotional and neutral speech. *Forensic Linguistics, 4*, 104–124.
Klasmeyer, G., & Sendlmeier, W. F. (1999). Voice and emotional states. In R. Kent & M. Ball (Eds.), *Voice quality measurement* (pp. 339–359). San Diego, CA: Singular Publishing Group.
Kohler, K. J. (1995). *Einführung in die Phonetik des Deutschen* [Introduction to the phonetics of German] (2nd ed.). Berlin: Erich Schmidt.
Ladefoged P. (1975). *A course in phonetics*. New York: Harcourt Brace Jovanovich.
Ladd, D. R. (1983). Phonological features of intonational peaks, *Language, 59*, 721–759.
Ladd, D. R. (1995). Linear and overlay descriptions: An autosegmental-metrical middle-way. *Proceedings XIIIth ICPhS 95*, Stockholm, Sweden, 2, 116–123.
Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *Journal of the Acoustical Society of America, 78*, 435–444.
Ladd, D. R., Scherer, K. R., & Silverman, K. E. A. (1986).

An integrated approach to studying intonation and attitude. In C. Johns-Lewis (Ed.), *Intonation in discourse* (pp. 125–138). London: Croom Helm.

Laver, J. (1980). *The phonetic description of voice quality.* Cambridge: Cambridge University Press.

Laver, J. (1991). *The gift of speech.* Edinburgh: Edinburgh University Press.

LeDoux, J. (1996). *The emotional brain: The mysterious underpinnings of emotional life.* New York: Simon & Schuster.

Leyhausen, P. (1967). Biologie von Ausdruck und Eindruck (Teil 1). [The biology of expression and impression. Part 1]. *Psychologische Forschung, 31,* 113–176.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oehrle (Eds.). *Language sound structures* (pp. 157–233). Cambridge, MA: MIT Press.

Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *Journal of the Acoustical Society of America, 34,* 922–927.

Marler, P., & Tenaza, R. (1977). Signaling behavior of apes with special reference to vocalization. In T. A. Sebeok (Ed.), *How animals communicate* (pp. 965–1033). Bloomington: Indiana University Press.

Maser, J. D. (Ed.). (1987). *Depression and expressive behavior.* Hillsdale, NJ: Erlbaum.

Massaro, D. W. (2000). Multimodal emotional perception: Analogous to speech processes. In *Proceedings of the ISCA Workshop on Speech and Emotion,* Newcastle, Northern Ireland, September 5–7, 2000, 114–121.

Mertens, P., Beaugendre, F., & d'Alessandro, C. (1997). Comparing approaches to pitch contour stylization for speech synthesis. In J. P. H. van Santen, R. W. Sproat, J. P. Olive, & J. Hirschberg (Eds), *Progress in speech synthesis* (pp. 347–363). New York: Springer.

Möbius, B. (1995). Components of a quantitative model of German intonation, *Proceedings XIIIth ICPhS 95,* Stockholm, Sweden, 2, 108–117.

Morton, E. S. (1977). On the occurrence and significance of motivational-structural rules in some bird and mammal sounds. *American Naturalist, 111,* 855–869.

Mozziconacci, S. J. L. (1995). Pitch variations and emotions in speech. *Proceedings XIIIth ICPhS 95,* Stockholm, Sweden, 1, 178–181.

Murray, I. R., & Arnott, J. L. (1993). Toward a 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.

Murray, I. R., & Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication, 16,* 369–390.

Natale, M. (1977). Effects of induced elation-depression on speech in the initial interview. *Journal of Consulting and Clinical Psychology, 45,* 45–52.

Ohala, J. J. (1980). The acoustic origin of the smile. *Journal of the Acoustic Society of America, 68,* 33 (Abstract).

O'Shaughnessy, D. (2000). *Speech communication: Human and machine.* New York: IEEE Press.

Paeschke, A., Kienast, M., & Sendlmeier, W. F. (1999). F0-Contours in emotional speech, *Proceedings ICPhS 99,* San Francisco, Vol. 2, 929–933.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions.* New York: Oxford University Press.

Papousek, M. (1994). *Vom Schrei zum ersten Wort: Anfänge der Sprachentwicklung in der vorsprachlichen Kommunikation* [From the infant cry to the first word: The bases of speech development in preverbal communication.] Bern: Hans Huber.

Peper, M., & Irle, E. (1997). Categorical and dimensional decoding of emotional intonations in patients with focal brain lesions. *Brain and Language, 57,* 233–264.

Pihan, H., Altenmüller, E., & Ackermann, H. (1997). The cortical processing of perceived emotion: A DC-potential study on affective speech prosody. *NeuroReport, 8,* 623–627.

Pittam, J., & Scherer, K. R. (1993). Vocal expression and communication of emotion. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 185–198). New York: Guilford Press.

Porges, S. W. (1997). Emotion: An evolutionary by-product of the neural regulation of the autonomic nervous system. *Annals of the New York Academy of Sciences, 807,* 62–77.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S., & Setter, J. (1998). Transcription of prosodic and paralinguistic features of emotional speech. *Journal of the International Phonetic Association, 28,* 83–94.

Robin, D. A., Tranel, D., & Damasio, H. (1990). Auditory perception of temporal and spectral events in patients with focal left and right cerebral lesions. *Brain and Language, 39,* 539–555.

Robinson, B. W. (1972). Anatomical and physiological contrasts between human and other primate vocalizations. In S. L. Washburn & P. Dolhinow (Eds.), *Perspectives on human evolution* (pp. 438–443). New York: Holt, Rinehart & Winston.

Roseman, I., & Smith, C. (2001). Appraisal theory: Overview, assumptions, varieties, controversies. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.). *Appraisal processes in emotion: Theory, methods, research* (pp. 3–19). New York: Oxford University Press.

Rosen, S., & Howell, P. (1991). *Signals and systems for speech and hearing.* New York: Harcourt Brace Jovanovich.

Ross, E. D. (1981). The aprosodias: Functional-anatomical organization of the affective components of language in the right hemisphere. *Archives of Neurology, 38,* 561–569.

Ross, E. D., Edmondson, J. A., & Seibert, G. B. (1986). The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice. *Journal of Phonetics, 14,* 283–302.

Ross, E. D., Homan, R. W., & Buck, R. (1994). Differential hemispheric lateralization of primary and social emotions: Implications for developing a comprehensive neurology for emotions, repression, and the subconscious. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 7,* 1–19.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39,* 1161–1178.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin, 115,* 102–141.

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology, 8,* 467–487.

Scherer, K. R. (1979). Non-linguistic indicators of emotion

and psychopathology. In C. E. Izard (Ed.), *Emotions in personality and psychopathology* (pp. 495–529). New York: Plenum.

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). New York: Cambridge University Press.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 293–318). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1985). Vocal affect signaling: A comparative approach. In J. Rosenblatt, C. Beer, M. Busnel, & P. J. B. Slater (Eds.), *Advances in the study of behavior* (pp. 189–244). New York: Academic Press.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143–165.

Scherer, K. R. (1987). Vocal assessment of affective disorders. In J. D. Maser (Ed.), *Depression and expressive behavior* (pp. 57–82). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1988). On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology, 7*, 79–100.

Scherer, K. R. (1989). Vocal correlates of emotion. In H. Wagner & A. Manstead (Eds.), *Handbook of psychophysiology: Emotion and social behavior* (pp. 165–197). London: Wiley.

Scherer, K. R. (1992). What does facial expression express? In K. Strongman (Ed.), *International review of studies on emotion* (Vol. 2, pp. 139–165). Chichester, England: Wiley.

Scherer, K. R. (1994). Affect bursts. In S. H. M. van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 161–196). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1999a). Appraisal theories. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion* (pp. 637–663). Chichester, England: John Wiley.

Scherer, K. R. (1999b). Universality of emotional expression. In D. Levinson, J. Ponzetti, & P. Jorgenson (Eds.), *Encyclopedia of human emotions* (pp. 669–674). New York: Macmillan.

Scherer, K. R. (2000a). Psychological models of emotion. In J. Borod (Ed.), *The neuropsychology of emotion* (pp. 137–162). New York: Oxford University Press.

Scherer, K. R. (2000b). Emotions as episodes of subsystem synchronization driven by nonlinear appraisal processes. In M. Lewis & I. Granic (Eds.), *Emotion, development, and self-organization* (pp. 70–99). New York/Cambridge: Cambridge University Press.

Scherer, K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.). *Appraisal processes in emotion: Theory, methods, research* (pp. 92–120). New York: Oxford University Press.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76–92.

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

Scherer, K. R., Feldstein, S., Bond, R. N., & Rosenthal, R. (1985). Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research, 14*, 409–425.

Scherer, K. R., Johnstone, T., Klasmeyer, G., & Bänziger, T. (2000). Can automatic speaker verification be improved by training the algorithms of emotional speech? *Proceedings ICSLP2000*, Beijing, China.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America, 76*, 1346–1356.

Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion, 1*, 331–346.

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. New York: Oxford University Press.

Scripture, E. W. (1921). A study of emotions by speech transcription. *Vox, 31*, 179–183.

Siegman, A. W. (1987). The pacing of speech in depression. In J. D. Maser (Ed.), *Depression and expressive behavior* (pp. 83–102). Hillsdale, NJ: Erlbaum.

Smith, C. A. (1989). Dimensions of appraisal and physiological response in emotion. *Journal of Personality and Social Psychology, 56*, 339–353.

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology, 48*, 813–838.

Smith, C. A., & Scott, H. S. (1997). A componential approach to the meaning of facial expressions. In J. A. Russell & J. M. Fernández-Dols (Eds.), *The psychology of facial expression* (pp. 229–254). Cambridge: Cambridge University Press.

Sundberg, J. (1987). *The science of the singing voice*. DeKalb, IL: Northern Illinois University Press.

Sundberg, J. (1994). *Vocal fold vibration patterns and phonatory modes*. STL-QPRS 2–3, KTH Stockholm, Sweden, 69–80.

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. New York: Oxford University Press.

Tischer, B. (1993). *Die vokale Kommunikation von Gefühlen*. [The vocal communication of emotions]. Weinheim: Psychologie-Verlags-Union.

Titze, I. (1992). Acoustic interpretation of the voice range profile (Phonetogram). *Speech Hearing Research, 35*, 21–34.

Tolkmitt, F., Bergmann, G., Goldbeck, T., & Scherer, K. R. (1988). Experimental studies on vocal communication. In K. R. Scherer (Ed.), *Facets of emotion: Recent research* (pp. 119–138). Hillsdale, NJ: Erlbaum.

Tolkmitt, F. J., & Scherer, K. R. (1986). Effects of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 302–313.

Tompkins, C. A., & Flowers, C. R. (1985). Perception of emotional intonation by brain damaged adults: The influence of task processing levels. *Journal of Speech and Hearing Research, 28*, 527–583.

Tomkins, S. S. (1962). *Affect, imagery, consciousness. Vol. 1. The positive affects*. New York: Springer.

Tomkins, S. S. (1984). Affect theory. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 163–196). Hillsdale, NJ: Erlbaum.

Tucker, D. M., Watson, R. T., & Heilman, K. M. (1977). Discrimination and evocation of affectively intoned speech in patients with right parietal disease. *Neurology, 27*, 947–950.

van Bezooijen, R. (1984). *The characteristics and recognizability of vocal expression of emotions.* Dordrecht, The Netherlands: Foris.

van Bezooijen, R., Otto, S., & Heenan, T. A. (1983). Recognition of vocal expressions of emotions: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology, 14*, 387–406.

Van Lancker, D. (1991). Personal relevance and the human right hemisphere. *Brain and Cognition, 17*, 64–92.

Van Lancker, D., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology, 11*, 665–674.

Van Lancker, D., & Sidtis, J. J. (1992). The identification of affective-prosodic stimuli by left- and right-hemisphere-damaged subjects: All errors are not created equal. *Journal of Speech and Hearing Research, 35*, 963–970.

Wagner, H. L. (1993). On measuring performance in category judgment studies on nonverbal behavior. *Journal of Nonverbal Behavior, 17*(1), 3–28.

Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology, 51*, 690–699.

Wehrle, T., Kaiser, S., Schmidt, S., & Scherer, K. R. (2000). Studying dynamic models of facial expression of emotion using synthetic animated faces. *Journal of Personality and Social Psychology, 78*(1), 105–119.

Zwicker, E. (1982). *Psychoacoustics.* New York: Springer.

Zwirner, E. (1930). Beitrag zur Sprache des Depressiven. [Contribution on the speech of depressives.] *Phonometrie III. Spezielle Anwendungen I.*, 171–187. Basel: Karger.