UNIVERSITÉ DE GENÈVE

Archive ouverte UNIGE
https://archive-ouverte.unige.ch

Article scientifique | Article | 2003

Published version | Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

Vocal communication of emotion: A review of research paradigms

Scherer, Klaus R.

How to cite

SCHERER, Klaus R. Vocal communication of emotion: A review of research paradigms. In: Speech Communication, 2003, vol. 40, n° 1-2, p. 227–256. doi: 10.1016/S0167-6393(02)00084-5

This publication URL: https://archive-ouverte.unige.ch//unige:102083
Publication DOI: 10.1016/S0167-6393(02)00084-5

© This document is protected by copyright. Please refer to copyright holder(s) for terms of use.

Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Speech Communication 40 (2003) 227–256

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Vocal communication of emotion: A review of research paradigms

Klaus R. Scherer *

Department of Psychology, University of Geneva, 40. Boulevard du Pont d'Arve, CH-1205 Geneva, Switzerland

Abstract

The current state of research on emotion effects on voice and speech is reviewed and issues for future research efforts are discussed. In particular, it is suggested to use the Brunswikian lens model as a base for research on the vocal communication of emotion. This approach allows one to model the complete process, including both encoding (expression), transmission, and decoding (impression) of vocal emotion communication. Special emphasis is placed on the conceptualization and operationalization of the major elements of the model (i.e., the speaker's emotional state, the listener's attribution, and the mediating acoustic cues). In addition, the advantages and disadvantages of research paradigms for the induction or observation of emotional expression in voice and speech and the experimental manipulation of vocal cues are discussed, using pertinent examples drawn from past and present research.
© 2002 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Der Aufsatz gibt einen umfassenden Überblick über den Forschungsstand zum Thema der Beeinflussung von Stimme und Sprechweise durch Emotionen des Sprechers. Allgemein wird vorgeschlagen, die Forschung zur vokalen Kommunikation der Emotionen am Brunswik'schen Linsenmodell zu orientieren. Dieser Ansatz erlaubt den gesamten Kommunikationsprozess zu modellieren, von der Enkodierung (Ausdruck), über die Transmission (Übertragung), bis zur Dekodierung (Eindruck). Besondere Aufmerksamkeit gilt den Problemen der Konzeptualisierung und Operationalisierung der zentralen Elemente des Modells (z.B., dem Emotionszustand des Sprechers, den Inferenzprozessen des Hörers, und den zugrundeliegenden vokalen Hinweisreizen). Anhand ausgewählter Beispiele empirischer Untersuchungen werden die Vor- und Nachteile verschiedener Forschungsparadigmen zur Induktion und Beobachtung des emotionalen Stimmausdrucks sowie zur experimentellen Manipulation vokaler Hinweisreize diskutiert.
© 2002 Elsevier Science B.V. All rights reserved.

Résumé

L'état actuel de la recherche sur l'effet des émotions d'un locuteur sur la voix et la parole est décrit et des approches prometteuses pour le futur identifiées. En particulier, le modèle de perception de Brunswik (dit "de la lentille" est proposé) comme paradigme pour la recherche sur la communication vocale des émotions. Ce modèle permet la modélisation du processus complet, de l'encodage (expression) par la transmission au décodage (impression). La conceptualisation et l'opérationalization des éléments centraux du modèle (l'état émotionnel du locuteur, l'inférence de cet état par l'auditeur, et les indices auditifs) sont discuté en détail. De plus, en analysant des exemples de la recherche dans le

* Tel.: +41-22-705-9211/9215; fax: +41-22-705-9219.
E-mail address: klaus.scherer@pse.unige.ch (K.R. Scherer).

0167-6393/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(02)00084-5

domaine, les avantages et désavantages de différentes méthodes pour l'induction et l'observation de l'expression émotionnelle dans la voix et la parole et pour la manipulation expérimentale de différents indices vocaux sont évoqués.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Vocal communication; Expression of emotion; Speaker moods and attitudes; Speech technology; Theories of emotion; Evaluation of emotion effects on voice and speech; Acoustic markers of emotion; Emotion induction; Emotion simulation; Stress effects on voice; Perception/decoding

1. Introduction: Modeling the vocal communication of emotion

The importance of emotional expression in speech communication and its powerful impact on the listener has been recognized throughout history. Systematic treatises of the topic, together with concrete suggestions for the strategic use of emotionally expressive speech, can be found in early Greek and Roman manuals on rhetoric (e.g., by Aristotle, Cicero, Quintilian), informing all later treatments of rhetoric in Western philosophy (Kennedy, 1972). Renewed interest in the expression of emotion in face and voice was sparked in the 19th century by the emergence of modern evolutionary biology, due to the contributions by Spencer, Bell, and particularly Darwin (1872, 1998). The empirical investigation of the effect of emotion on the voice started at the beginning of the 20th century, with psychiatrists trying to diagnose emotional disturbances through the newly developed methods of electroacoustic analysis (e.g., Isserlin, 1925; Scripture, 1921; Skinner, 1935).

The invention and rapid dissemination of the telephone and the radio also led to increasing scientific concern with the communication of speaker attributes and states via vocal cues in speech (Allport and Cantril, 1934; Herzog, 1933; Pear, 1931). However, systematic research programs started in the 1960s when psychiatrists renewed their interest in diagnosing affective states via vocal expression (Alpert et al., 1963; Moses, 1954; Ostwald, 1964; Hargreaves et al., 1965; Starkweather, 1956), non-verbal communication researchers explored the capacity of different bodily channels to carry signals of emotion (Feldman and Rimé, 1991; Harper et al., 1978; Knapp, 1972; Scherer, 1982b), emotion psychologists charted the expression of emotion in different modalities (Tomkins, 1962; Ekman, 1972, 1992; Izard, 1971, 1977), linguists and particularly phoneticians discovered the importance of pragmatic information in speech (Mahl and Schulze, 1964; Trager, 1958; Pittenger et al., 1960; Caffi and Janney, 1994), and engineers and phoneticians specializing in acoustic signal processing started to make use of ever more sophisticated technology to study the effects of emotion on the voice (Lieberman and Michaels, 1962; Williams and Stevens, 1969, 1972). In recent years, speech scientists and engineers, who had tended to disregard pragmatic and paralinguistic aspects of speech in their effort to develop models of speech communication for speech technology applications, have started to devote more attention to speaker attitudes and emotions—often in the interest to increase the acceptability of speech technology for human users. The conference which has motivated the current special issue of this journal (ISCA Workshop on Voice and Emotion, Newcastle, Northern Ireland, 2000) and a number of recent publications (Amir and Ron, 1998; Bachorowski, 1999; Bachorowski and Owren, 1995; Banse and Scherer, 1996; Cowie and Douglas-Cowie, 1996; Erickson et al., 1998; Iida et al., 1998; Kienast et al., 1999; Klasmeyer, 1999; Morris et al., 1999; Murray and Arnott, 1993; Mozziconacci, 1998; Pereira and Watson, 1998; Picard, 1997; Rank and Pirker, 1998; Sobin and Alpert, 1999) testifies to the lively research activity that has been sprung up in this domain. This paper attempts to review some of the central issues in empirical research on the vocal communication of emotion and to chart some of the promising approaches for interdisciplinary research in this area.

I have repeatedly suggested (Scherer, 1978, 1982a) to base theory and research in this area on a modified version of Brunswik's functional lens

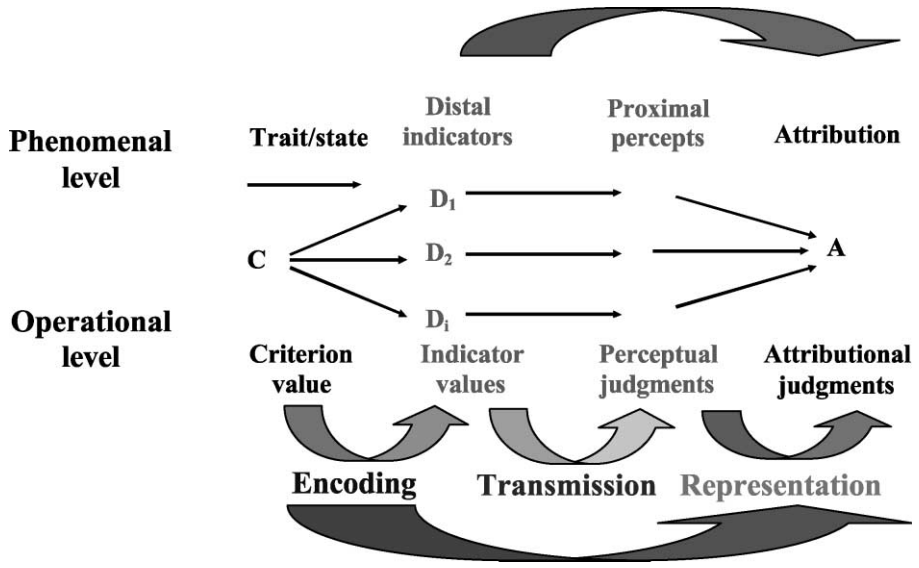


Fig. 1. A Brunswikian lens model of the vocal communication of emotion.

model of perception (Brunswik, 1956; Gifford, 1994; Hammond and Stewart, 2001). Since the detailed argument can be found elsewhere (see Kappas et al., 1991; Scherer et al., in press), I will only briefly outline the model (shown in the upper part of Fig. 1, which represents the conceptual level). The process begins with the encoding, or expression, of emotional speaker states by certain voice and speech characteristics amenable to objective measurement in the signal. Concretely, the assumption is that the emotional arousal of the speaker is accompanied by physiological changes that will affect respiration, phonation, and articulation in such a way as to produce emotion-specific patterns of acoustic parameters (see (Scherer, 1986) for a detailed description). Using Brunswik's terminology, one can call the degree to which such characteristics actually correlate with the underlying speaker state *ecological validity*. As these acoustic changes can serve as cues to speaker affect for an observer, they are called *distal cues* (distal in the sense of remote or distant from the observer). They are transmitted, as part of the speech signal, to the ears of the listener and perceived via the auditory perceptual system. In the model, these perceived cues are called *proximal cues* (proximal in the sense of close to the observer).

There is some uncertainty among Brunswikians exactly how to define and operationalize proximal cues in different perceptual domains (see (Hammond and Stewart, 2001) for the wide variety of uses and definitions of the model). While Brunswik apparently saw the proximal stimulus as close but still outside of the organism (Hammond and Stewart, 2001), the classic example given for the distal–proximal relationship in visual perception—the juxtaposition between object size (distal) and retinal size (proximal)—suggests that the proximal cue, the pattern of light on the retina, is already inside the organism. Similarly, in auditory perception, the fundamental frequency of a speech wave constitutes the distal characteristic that gives rise to the pattern of vibration along the basilar membrane, and, in turn, the pattern of excitation along the inner hair cells, the consequent excitation of the auditory neurons, and, finally, its representation in the auditory cortex. Either phase in this input, transduction, and coding process could be considered a proximal representation of the distal stimulus. I believe that it makes sense to extend this term to the neural representation of the stimulus information as coded by the respective neural structures. The reason is twofold: (1) It is difficult to measure the raw input (e.g., vibration

of the basilar membrane) and thus one could not systematically study this aspect of the model in relation to others; (2) The immediate input into the inference process, which Brunswik called *cue utilization*, is arguably the neural representation in the respective sensory cortex. Thus, the proximal cue for fundamental frequency would be perceived pitch. While fraught with many problems, we do have an access to measuring at least the conscious part of this representation via self-report (see Bänziger and Scherer, 2001).

One of the most important advantages of the model is to highlight the fact that objectively measured distal characteristics are not necessarily equivalent to the proximal cues they produce in the observer. While the proximal cues are based on (or mimick) distal characteristics, the latter may be modified or distorted by (1) the transmission channel (e.g., distance, noise) and (2) the structural characteristics of the perceptual organ and the transduction and coding process (e.g., selective enhancement of certain frequency bands). These issues are discussed in somewhat greater detail below.

The decoding process consists of the inference of speaker attitudes and emotions based on internalized representations of emotional speech modifications, the proximal cues. The fit of the model can be ascertained by operationalizing and measuring each of its elements (see operational level in Fig. 1), based on the definition of a finite number of cues. If the attribution obtained through listener judgments corresponds (with better than chance recognition accuracy) to the criterion for speaker state (e.g., intensity of a certain emotion), the model describes a functionally valid communication process. However, the model is also extremely useful in cases in which attributions and criteria do not match since it permits determination of the missing or faulty link of the chain. Thus, it is possible that the respective emotional state does not produce reliable externalizations in the form of specific distal cues in the voice. Alternatively, valid distal cues might be degraded or modified during transmission and perception in such a fashion that they no longer carry the essential information when they are proximally represented in the listener. Finally, it is possible that the proximal cues reliably map the valid distal cues but that the inference mechanism, i.e., the cognitive representation of the underlying relationships, is flawed in the respective listener (e.g., due to lack of sufficient exposure or inaccurate stereotypes). To my knowledge, there is no other paradigm that allows to examine the process of vocal communication in as comprehensive and systematic fashion. This is why I keep arguing for the utility of basing research in this area explicitly on a Brunswikian lens model.

Few empirical studies have sought to model a specific communication process by using the complete lens model, mostly due to considerations involving the investment of the time and money required (but see Gifford, 1994; Juslin, 2000). In an early study on the vocal communication of speaker personality, trying to determine which personality traits are reliably indexed by vocal cues and correctly inferred by listeners, I obtained natural speech samples in simulated jury discussions with adults (German and American men) for whom detailed personality assessments (self- and peer-ratings) had been obtained (see Scherer, 1978, 1982a). Voice quality was measured via expert (phoneticians) ratings (distal cues) and personality inferences (attributions) by lay listeners' ratings (based on listening to content-masked speech samples). In addition, a different group of listeners was asked to rate the voice quality with the help of a rating scale with natural language labels for vocal characteristics (proximal cues). Path analyses were performed to test the complete lens model. This technique consists of a systematic series of regression analyses to test the causal assumptions in a model containing mediating variables (see Bryman and Cramer, 1990, pp. 246–251). The results for one of the personality traits studied are illustrated in Fig. 2. The double arrows correspond to the theoretically specified *causal paths*. Simple and dashed arrows correspond to non-predicted direct and indirect effects that explain additional variance. The graph shows why *extroversion* was correctly recognized from the voice: (1) extroversion is indexed by objectively defined vocal cues (ecologically valid in the Brunswikian sense), (2) these cues are not too drastically modified in the transmission process, and (3) the listeners' inference structure in decoding mirrors

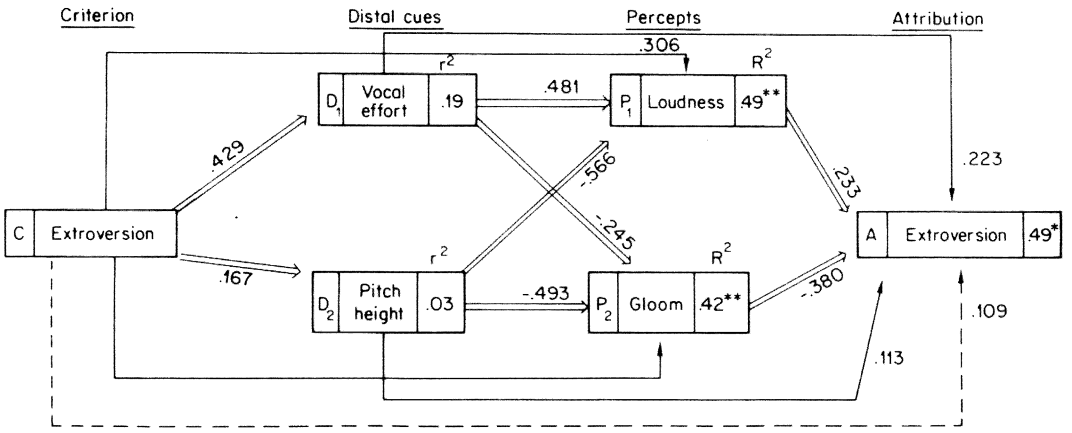


Fig. 2. Two-dimensional path analysis model for the inference of extroversion from the voice (reproduced from Scherer, 1978). The dashed line represents the direct path, the double line the postulated indirect paths, and the single lines the indirect paths compatible with the model. Coefficients shown are standardized coefficients except for r_{CD1} and r_{CD2}, which are Pearson rs (the direction shown is theoretically postulated). R^2s based on all predictors from which paths lead to the variable. $^{*}p < 0.05$; $^{**}p < 0.01$.

the encoding structure. These conditions were not met in the case of *emotional stability*. While there were strong inference structures, shared by most listeners (attributing emotional stability to speakers with resonant, warm, low-pitched voices), these do not correspond to an equivalent encoding structure (i.e., there was no relationship between voice frequency and habitual emotional stability in the sample of speakers studied; see (Scherer, 1978) for the data and further discussion).

Clearly, a similar approach could be used with emotions rather than personality traits as speaker characteristics. Unfortunately, so far no complete lens model has been tested in this domain. Yet this type of model is useful as a heuristic device to design experimental work in this area even if only parts of the model are investigated. The curved arrows in Fig. 1 indicate some of the major issues that can be identified with the help of the model. These issues, and the evidence available to date, are presented below.

2. A review of the literature

2.1. Encoding studies

The basis of any functionally valid communication of emotion via vocal expression is that different types of emotion are actually characterized by unique patterns or configurations of acoustic cues. In the context of a Brunswikian lens model that means that the identifiable emotional states of the sender are in fact externalised by a specific set of distal cues. Without such distinguishable acoustic patterns for different emotions, the nature of the underlying speaker state could not be communicated reliably. Not surprisingly, then, there have been a relatively large number of empirical encoding studies conducted over the last six decades, attempting to determine whether elicitation of emotional speaker states will produce corresponding acoustic changes. These studies can be classified into three major categories: natural vocal expression, induced emotional expression, and simulated emotional expression.

2.1.1. Natural vocal expression

Work in this area has made use of material that was recorded during naturally occurring emotional states of various sorts, such as dangerous flight situations for pilots, journalists reporting emotion-eliciting events, affectively loaded therapy sessions, or talk and game shows on TV (Johannes et al., 2000; Cowie and Douglas-Cowie, 1996; Duncan et al., 1983; Eldred and Price, 1958; Hargreaves et al., 1965; Frolov et al., 1999; Huttar, 1968; Kuroda et al., 1979; Niwa, 1971; Roessler and

Lester, 1976, 1979; Simonov and Frolov, 1973; Sulc, 1977; Utsuki and Okamura, 1976; Williams and Stevens, 1969, 1972; Zuberbier, 1957; Zwirner, 1930). The use of naturally occurring voice changes in emotionally charged situations seems the ideal research paradigm since it has very high ecological validity. However, there are some serious methodological problems. Voice samples obtained in natural situations, often only for a single or a very small number of speakers, are generally very brief, not infrequently suffering from bad recording quality. In addition, there are problems in determining the preise nature of the underlying emotion and the effect of regulation (see below).

2.1.2. Induced emotions

Another way to study the vocal effects of emotion is to experimentally induce specific emotional states in groups of speakers and to record speech samples. A direct way of inducing affective arousal and studying the effects of the voice is the use of psychoactive drugs. Thus, Helfrich et al. (1984) studied the effects of antidepressive drugs on several vocal parameters (compared to placebo) over a period of several hours. Most induction studies have used indirect paradigms that include stress induction via difficult tasks to be completed under time pressure, the presentation of emotion-inducing films or slides, or imagery methods (Alpert et al., 1963; Bachorowski and Owren, 1995; Bonner, 1943; Havrdova and Moravek, 1979; Hicks, 1979; Karlsson et al., 1998; Markel et al., 1973; Plaikner, 1970; Roessler and Lester, 1979; Scherer, 1977, 1979; Scherer et al., 1985; Skinner, 1935; Tolkmitt and Scherer, 1986). While this approach, generally favoured by experimental psychologists because of the degree of control it affords, does result in comparable voice samples for all participants, there are a number of serious drawbacks. Most importantly, these procedures often produce only relatively weak affect. Furthermore, in spite of using the same procedure for all participants, one cannot necessarily assume that similar emotional states are produced in all individuals, precisely because of the individual differences in event appraisal mentioned above. Recently, Scherer and his collaborators, in the context of a large scale study on emotion effects in automatic speaker verification, have attempted to remedy some of these shortcomings by developing a computerized induction battery for a variety of different states (Scherer et al., 1998).

2.1.3. Simulated (portrayed) vocal expressions

This has been the preferred way of obtaining emotional voice samples in this field. Professional or lay actors are asked to produce vocal expressions of emotion (often using standard verbal content) as based on emotion labels and/or typical scenarios (Banse and Scherer, 1996; Bortz, 1966; Coleman and Williams, 1979; Costanzo et al., 1969; Cosmides, 1983; Davitz, 1964; Fairbanks and Pronovost, 1939; Fairbanks and Hoaglin, 1941; Fonagy, 1978, 1983; Fonagy and Magdics, 1963; Green and Cliff, 1975; Höffe, 1960; Kaiser, 1962; Kienast et al., 1999; Klasmeyer and Sendlmeier, 1997; Klasmeyer and Meier, 1999; Klasmeyer and Sendlmeier, 1999; Klasmeyer, 1999; Kotlyar and Morozov, 1976; Levin and Lord, 1975; Paeschke et al., 1999; Scherer et al., 1972a,b, 1973; Sedlacek and Sychra, 1963; Sobin and Alpert, 1999; Tischer, 1993; van Bezooijen, 1984; Wallbott and Scherer, 1986; Whiteside, 1999; Williams and Stevens, 1972). There can be little doubt that simulated vocal portrayal of emotions yields much more intense, prototypical expressions than are found for induced states or even natural emotions (especially when they are likely to be highly controlled; see (Scherer, 1986, p. 159)). However, it cannot be excluded that actors overemphasize relatively obvious cues and miss more subtle ones that might appear in natural expression of emotion (Scherer, 1986, p. 144). It has often been argued that emotion portrayals reflect sociocultural norms or expectations more than the psychophysiological effects on the voice as they occur under natural conditions. However, it can be argued that all publicly observable expressions are to some extent "portrayals" (given the social constraints on expression and unconscious tendencies toward self-presentation; see (Banse and Scherer, 1996)). Furthermore, since vocal portrayals are reliably recognized by listener-judges (see below) it can be assumed that they reflect at least in part "normal" expression patterns (if the two were to diverge too much, the acted version

would lose its credibility). However, there can be little doubt that actors' portrayals are influenced by conventionalised stereotypes of vocal expression.

2.1.4. Advantages and disadvantages of methods

As shown above, all of the methods that have been used to obtain vocal emotion expression samples have both advantages and disadvantages. In the long run, it is probably the best strategy to look for convergences between all three approaches in the results. Johnstone and Scherer (2000, pp. 226–227) have reviewed the converging evidence with respect to the acoustic patterns that characterize the vocal expression of major modal emotions. Table 1 summarizes their conclusions in synthetic form.

Much of the consistency in the findings is linked to differential levels of arousal or activation for the target emotions. Indeed, in the past, it has often been assumed that contrary to the face, capable of communicating qualitative differences between emotions, the voice could only mark physiological arousal (see (Scherer, 1979, 1986) for a detailed discussion). That this conclusion is erroneous is shown by the fact that judges are almost as accurate in inferring different emotions from vocal as from facial expression (see below). There are two main reasons why it has been difficult to demonstrate qualitative differentiation of emotions in acoustic patterns apart from arousal: (1) only a limited number of acoustic cues has been studied, and (2) arousal differences within emotion families have been neglected.

As to (1), most studies in the field have limited the scope of acoustic measurement to F0, energy, and speech rate. Only very few studies have looked at the frequency distribution in the spectrum or formant parameters. As argued by Scherer (1986), it is possible that F0, energy, and rate may be most indicative of arousal whereas qualitative, valence differences may have a stronger impact on source and articulation characteristics.

As to (2), most investigators have tended to view a few basic emotions as rather uniform and homogeneous, as suggested by the writings of discrete emotion theory (see Scherer, 2000a). More recently, the concept of emotion families has been proposed (e.g., Ekman, 1992), acknowledging the fact that there are many different kinds of anger, of joy, or of fear. For example, the important vocal differences between the expression as hot anger (explosive rage) and cold, subdued or controlled, anger may explain some of the lack of replication for results concerning the acoustic patterns of anger expression in the literature. The interpretation of such discrepancies is rendered particularly difficult by the fact that researchers generally do not specify which kind of anger has been produced or portrayed (see Scherer, 1986). Thus, different instantiations or variants of specific emotions, even though members of the same family, may vary appreciably with respect to their acoustic expression patterns.

Banse and Scherer (1996) have conducted an encoding study, using actor portrayal, which has attempted to remedy both of these problems. They used empirically derived scenarios for 14 emotions, ten of which consisted of pairs of two members of a same emotion family for five modal emotions, varying in arousal. In addition to the classical set of acoustic parameters, they measured a number

Table 1
Synthetic compilation of the review of empirical data on acoustic patterning of basic emotions (based on Johnstone and Scherer, 2000)

	Stress	Anger/rage	Fear/panic	Sadness	Joy/elation	Boredom
Intensity	↗	↗	↗	↘	↗	
F0 floor/mean	↗	↗	↗	↘	↗	
F0 variability		↗		↘	↗	↘
F0 range		↗	↗(↘)	↘	↗	↘
Sentence contours		↘		↘		
High frequency energy		↗	↗	↘	(↗)	
Speech and articulation rate		↗	↗	↘	(↗)	↘

of indicators of energy distribution in the spectrum. Banse and Scherer interpret their data as clearly demonstrating that these states are acoustically differentiated with respect to quality (e.g., valence) as well as arousal. They underline the need for future work that addresses the issue of subtle differences between the members of the same emotion family and that uses an extensive set of different acoustic parameters, pertinent to all aspects of the voice and speech production process (respiration, phonation, and articulation).

One of the major shortcomings of the encoding studies conducted to date has been the lack of theoretical grounding, most studies being motivated exclusively by the empirical detection of acoustic differences between emotions. As in most other areas of scientific inquiry, an atheoretical approach has serious shortcomings, particularly in not being able to account for lack of replication and in not allowing the identification of the mechanism underlying the effects found. It is difficult to blame past researchers for the lack of theory, since emotion psychology and physiology, the two areas most directly concerned, have not been very helpful in providing tools for theorizing. For example, discrete emotion theory (Ekman, 1972, 1992; Izard, 1971, 1977; see Scherer, 2000a), which has been the most influential source of emotion conceptualization in this research domain, has specified the existence of emotion-specific neural motor patterns as underlying the expression. However, apart from this general statement, no detailed descriptions of the underlying mechanisms or concrete hypotheses have been forthcoming (except to specify patterns of facial expression—mostly based on observation; see (Wehrle et al., 2000)).

It is only with the advent of componential models, linked to the development of appraisal theories, that there have been attempts at a detailed hypothetical description for the pattern of motor expressions to be expected for specific emotions (see Scherer, 1984, 1992a; Roseman and Smith, 2001). For the vocal domain, I have proposed a comprehensive theoretical grounding of emotion effects on vocal expression, as based on my component process model of emotion, yielding a large number of concrete predictions as to the changes in major acoustic parameters to be expected for 14 modal emotions (Scherer, 1986). The results from studies that have provided the first experimental tests of these predictions are described below.

In concluding this section, it is appropriate to point out the dearth of comparative approaches that examine the relative convergence between empirical evidence on the acoustic characteristics of animal, particularly primate, vocalizations (see (Fichtel et al., 2001) for an impressive recent example) and human vocalizations. Such comparisons would be all the more appropriate since there is much evidence on the phylogenetic continuity of affect vocalization, at least for mammals (see Hauser, 2000; Scherer, 1985), and since we can assume animal vocalizations to be direct expressions of the underlying affect, largely free of control or self-presentation constraints.

2.2. Decoding studies

Work in this area examines to what extent listeners are able to infer actual or portrayed speaker emotions from content-free speech samples. Research in this tradition has a long history and has produced a large body of empirical results. In most of these studies actors have been asked to portray a number of different emotions by producing speech utterances with standardized or nonsense content. Groups of listeners are given the task to recognize the portrayed emotions. They are generally required to indicate the perceived emotion on rating sheets with standard lists of emotion labels, allowing to compute the percentage of stimuli per emotion that were correctly recognized. A review of approximately 30 studies conducted up to the early 1980s yielded an average accuracy percentage of about 60%, which is about five times higher than what would be expected if listeners had guessed in a random fashion (Scherer, 1989). Later studies reported similar levels of average recognition accuracy across different emotions, e.g., 65% in a study by van Bezooijen (1984) and 56% in a study by Scherer et al. (1991).

One of the major drawbacks of all of these studies is that they tend to study *discrimination* (deciding between alternatives) rather than *recog-*

nition (recognizing a particular category in its own right). In consequence, it is essential to correct the accuracy coefficients for guessing by taking the number of given alternatives and the empirical response distribution in the margins into account (Wagner, 1993). Most of the studies reviewed above have corrected the raw accuracy percentages to account for the effects of guessing. Unfortunately, there is no obvious way to compute a single correction factor (see Banse and Scherer, 1996). In the area of facial expression research, Ekman (1994) has suggested to use different comparison levels of chance accuracy for positive and negative emotions, given that there are usually fewer positive than negative emotion exemplars in the facial stimulus sets used. While this is also true for most vocal stimulus sets, there may be a less clear-cut positive–negative distinction for acoustic parameters as compared to facial muscle actions (see Banse and Scherer, 1996; Scherer et al., 1991).

As in the case of facial expression, vocal expressions of emotion by members of one culture are, on the whole, recognized with better than chance accuracy by members of other cultures (see Frick, 1985; van Bezooijen et al., 1983; Scherer, 1999b). Recently, Scherer et al. (2001a), reported data from a large scale cross-cultural study (using German actor portrayals and listener-judges from 9 countries in Europe, Asia, and the US) showing an overall accuracy percentage of 66% across all emotions and countries. These findings were interpreted as evidence for the existence of universal inference rules from vocal characteristics to specific emotions across cultures particularly since the patterns of confusion were very similar across all countries, including Indonesia. Yet, the data also suggest the existence of culture- and/or language-specific paralinguistic patterns in vocal emotion expression, since, in spite of using language-free speech samples, the recognition accuracy decreased with increasing dissimilarity of the language spoken by the listeners from German (the native language of the actors).

One of the issues that remain to be explored in greater detail is the differential ease with which different emotions can be decoded from the voice. In this context, it is interesting to explore whether the frequencies of confusion are comparable for vocal and facial emotion expressions. Table 1 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, adapted from (Scherer, 1999b), are based on reviews by Ekman (1994), for facial expression, and Scherer et al. (2001a), for vocal expression). As this compilation shows there are some major differences between emotions with respect to their recognition accuracy scores. In addition, there are major differences between facial and vocal expression. While joy is almost perfectly recognized in the face, listeners seem to have trouble identifying this emotion unequivocally in the voice. In the voice, sadness and anger are generally best recognized, followed by fear. Disgust portrayals in the voice are recognized barely above chance level. Johnstone and Scherer (2000) have attempted to explain the origin of these differences between facial and vocal expression on the basis of evolutionary pressure towards accurate vocal communication for different emotions. Preventing others from eating rotten food by emitting disgust signals may be most useful to conspecifics eating at the same place as the signaler. In this context, facial expressions of disgust are most adequate. In addition, they are linked to the underlying functional action (or action tendency) of regurgitating and blocking unpleasant odors. According to Darwin, such adaptive action tendencies (''serviceable habits'') are at the basis of most expressions (see Scherer, 1985, 1992a). In contrast, in encountering predators, there is clear adaptive advantage in being able to warn friends (in fear) or threaten foes (in anger) over large distances, something for which vocal expression is ideally suited. The differences in the recognizability of the vocal expression of different emotions suggest that it is profitable to separately analyze the recognizability of these emotions.

As is evident in Table 2, the recognition accuracy for vocal expressions is generally somewhat lower than that of facial expression. On the whole, 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

Table 2
Accuracy (in %) of facial and vocal emotion recognition in studies in Western and Non-Western countries (reproduced from Scherer, 2001)

	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 following the slash in column 1 indicate the number of countries studied.

would be expected by chance. In comparison, facial expression studies generally report an average accuracy percentage around 75%. There are many factors that are potentially responsible for this difference, including, in particular, the inherent dynamic nature of vocal stimuli which is less likely to produce stable patterns, contrary to facial expression where basic muscle configurations seem to identify the major emotions (Ekman, 1972, 1994), reinforced by the frequent use of static stimulus material (photos). Another important factor could be that members of a similar *emotion family* (see above) are more distinct from each other in vocal as compared to facial expression. For example, the important vocal differences between the expression of exuberant joy and quiet happiness may explain some of the lack of replication for results concerning the recognition of joy that are observed in the literature (since researchers generally do not specify which kind of joy has been used; see (Scherer, 1986)). In the above mentioned study by Banse and Scherer (1996) two members of the same emotion family for five modal emotions (i.e., 10 out of a total of 14 target emotions) with variable levels of activation or arousal were used. Listeners were asked to identify the expressed emotion for a total of 224 selected tokens, yielding an average accuracy level of 48% (as compared to 7% expected by chance if all 14 categories were weighted equally). If the agreement *between families* is computed for those emotions where there were two variants (yielding 10 categories), the accuracy percentage increases to 55% (as compared to 10% expected by chance). This indicates that, as one might expect, most of the confusions for these emotion categories occurred within families.

So far, the discussion has focussed on accuracy, overall or emotion-specific. However, the confusions listeners make are arguably even more interesting as errors are not randomly distributed and as the patterns of misidentification provide important information on the judgment process. Thus, confusion matrices should be reported as a matter of fact, which is generally not the case for older studies and, unfortunately, also for many more recent studies. The research by Banse and Scherer (1996), reported above, can serve as an example for the utility of analyzing the confusion patterns in recognition data in great detail. These authors found that each emotion has a specific confusion pattern (except for disgust portrayals which were generally confused with almost all other negative emotions). As mentioned above, 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 eleven emotions taken together). Johnstone and Scherer (2000) have suggested that confusion patterns can help to identify the similarity or proximity between emotion categories taking into account quality, intensity, and valence. It is likely that emotions similar on one or more of these dimensions are easier to confuse.

In addition, analysis of similarities or differences in the confusion patterns between different populations of listeners can be profitably employed to infer the nature of the underlying inference processes. For example, Banse and Scherer (1996) found that not only the average accuracy per emotion but also the patterns of confusion were

highly similar between human judge groups and two statistical categorization algorithms (discriminant analysis and bootstrapping). They interpreted this as evidence that the judges are likely to have used inference mechanisms that are comparable to optimized statistical classification procedures. In a different vein, as reported above, Scherer et al. (2001a,b) used the fact that confusion patterns were similar across different cultures as an indication for the existence of universal inference rules. This would not have been possible on the basis of comparable accuracy coefficients since judges might have arrived at these converging judgments in different ways. Clearly, then, it should be imperative for further work in this area to routinely report confusion matrices.

In closing this section I will address an issue that is often misunderstood—the role of stereotypes. It is often claimed that decoding studies using actor portrayals are fundamentally flawed since actors supposedly base their portrayals on cultural stereotypes of what certain emotions should sound like. These stereotypes are expected to match those that listeners use in decoding the expression intention. Since the term "stereotype" has a very negative connotation, implying that the respective beliefs or inference rules are false, it is commonly believed that stereotypes are "bad". This is not necessarily true, though, and many social psychologists have pointed out that stereotypes often have a kernel of truth. This is one of the reasons why they are shared by many individuals in the population. Stereotypes are often generalized inference rules, based on past experience or cultural learning, that allow the cognitive economy of making inferences on the basis of very little information (Scherer, 1992b). Thus, unless the inference system is completely unrelated to reality, i.e., if there is no kernel of truth, neither the socially shared nature nor the tendency of overgeneralization is problematical. On the contrary, it is a natural part of the human information processing system. It is intriguing to speculate on the origin and development of such stereotypes. Are they based on self- or other-observation, or both? Do they represent the average of many different emotional expression incidents with many different intensities or are they representative of extreme exemplars (in terms of intensity and prototypicality)? What is the role of media descriptions? One of the potential approaches to such issues consists in the study of how the ability to decode affect from the voice develops in the child (see Morton and Trehub, 2001). Obviously, the answers to these issues have important implications on research in this area.

It is a different issue, of course, whether actor portrayals that use such stereotypical inference structures provide useful information for decoding studies. Clearly, if the portrayals are exclusively based on and mimick socially shared inference rules to a very large extent, then the utility of the finding that listeners can recognize these portrayals is limited to increasing our understanding of the nature of these inference rules (and, in the case of cross-cultural comparison, the degree of their universality). However, it is by no means obvious that all actor portrayals are based on shared inference rules. If actors use auto-induction methods, e.g., Stanislavski or method acting procedures, as they are often directed to do by researchers in this area, or if they base their portrayals on auto-observation of their own emotional expressions, the distal cues they produce may be almost entirely independent of shared inference rules used by observers. In this case they can be treated as a valid criterion for speaker state in the sense of the Brunswikian lens model. Since it is difficult to determine exactly how actors have produced their portrayals, especially since they may well be unaware of the mechanisms themselves, one needs to combine natural observation, induction, and portrayal methods to determine the overlap of the expressive patterns found (see below).

2.3. Inference studies

While the fundamental issue studied by decoding studies is the ability of listeners to recognize the emotional state of the speaker, inference studies are more concerned with the nature of the underlying voice–emotion inference mechanism, i.e., which distal cues produce certain types of emotion inferences in listeners. As one might expect, the procedure of choice in such studies is to systematically manipulate or vary the acoustic cues in

sample utterances in such a way as to allow determination of the relative effect of particular parameters on emotion judgment. Three major paradigms have been used in this area: (1) cue measurement and regression, (2) cue masking, and (3) cue manipulation via synthesis.

(1) *Cue measurement and statistical association.* The simplest way to explore the utilization of distal cues in the inference process is to measure the acoustic characteristics of vocal emotion portrayals (acted or natural), and then to correlate these with the listeners' judgments of the underlying emotion or attitude of the speaker. Such studies generate information on which vocal characteristics are likely to have determined the judges' inference (Scherer et al., 1972a,b; van Bezooijen, 1984). Banse and Scherer (1996), in the study reported above, used multiple regression analysis to regress the acoustic parameters measured on the listeners' judgments. The data showed that a large proportion of the variance in judgments of emotion was explained by a set of about 9–10 acoustic measures, including mean F0, standard deviation of F0, mean energy, duration of voiced periods, proportion of energy up to 1000 Hz, and spectral drop-off. Table 3 (modified from Banse and Scherer, 1996, Table 8, pp. 632) shows the detailed results. The direction (i.e., positive or negative) and strength of beta-coefficients indicates the information value and the relative importance for each parameter. The multiple correlation coefficient R in the last row indicates the percentage of the variance in the listener judgments accounted for by the variables entered into the regression model.

While such results are interesting and useful, this method is limited by the nature of the voice samples that are used. Obviously, if certain parameters do not vary sufficiently, e.g., because actors did not produce a sufficiently extreme rendering of a certain emotion, the statistical analysis is unlikely to yield much of interest. In addition, the results are largely dependent on the precision of parameter extraction and transformation.

(2) *Cue masking.* In this approach verbal/vocal cues are masked, distorted, or removed from vocal emotion expressions (from real life or produced by actors) to explore the ensuing effects on emotion

Table 3
Multiple correlation coefficients and beta-weights resulting from regressing judges' emotion ratings on residual acoustical parameters

	HAng	CAng	Pan	Anx	Des	Sad	Ela	Hap	Int	Bor	Sha	Pri	Dis	Con
MF0	0.183*	−0.14	0.331***	0.287**	0.39***	0.114	0.178*	−0.162	−0.016	−0.409***	0.153	−0.357***	−0.177	−0.358***
SdF0	0.156**	0.044	−0.12	−0.359***	−0.18**	0.009	0.073	0.116	0.029	−0.038	0.029	0.137	0.054	0.142*
MElog	0.085	0.193	0.223	−0.421**	−0.220	−0.538***	0.279	−0.130	−0.075	0.408**	−0.405**	0.15	0.110	0.101
DurVo	−0.093	−0.012	−0.019	−0.129	0.229**	0.115	−0.061	−0.143	−0.184	0.328***	−0.079	−0.04	−0.007	0.007
HammI	0.194**	0.0	−0.09	0.057	0.167*	0.118	−0.103	−0.045	−0.014	−0.135	0.1	−0.16	−0.099	−0.053
PE1000	−0.083	−0.091	0.125	0.018	0.150	0.313*	0.180*	0.066	0.061	0.008	−0.025	−0.064	−0.294*	−0.149
DO1000	−0.007	0.030	0.188	−0.046	−0.136	0.05	−0.01	−0.084	−0.158	0.228	0.091	−0.024	−0.041	−0.133
R	0.63***	0.16	0.39***	0.38***	0.44***	0.49***	0.39***	0.27*	0.18	0.40***	0.38***	0.28*	0.17	0.36***

MF0: mean fundamental frequency, SdF0: standard deviation, MElog: mean energy, DurVo: duration of voiced periods, HammI: Hammarberg index, PE1000: proportion of voiced energy up to 1000 Hz, DO1000: slope of voiced spectral energy above 1000 Hz; HAng: hot anger, CAng: cold anger, Pan: panic fear, Anx: anxiety, Des: despair, Sad: sadness, Ela: elation, Hap: happiness, Int: interest, Bor: boredom, Sha: shame, Pri: pride, Dis: disgust, Con: contempt.

* $p \leqslant 0.05$.
** $p \leqslant 0.01$.
*** $p \leqslant 0.001$.

inferences on the part of judges listening to the modified material. This technique has been used early on, particularly in the form of low-pass filtering, by many pioneers in this research domain (Starkweather, 1956; Alpert et al., 1963). More recent studies by Scherer et al. (1984, 1985), and Friend and Farrar (1994) have applied different masking techniques (e.g., filtering, randomized splicing, playing backwards, pitch inversion, and tone-silence coding), each of which removes different types of acoustic cues from the speech signal or distorts certain cues while leaving others unchanged. For example, while low-pass filtering removes intelligibility and changes the timbre of the voice, it does not affect the F0 contour. In contrast, randomized splicing of the speech signal (see Scherer et al., 1985) removes all contour and sequence information from speech but retains the complete spectral information (i.e., the timbre of the voice).

The systematic combination of these techniques allows one to isolate the acoustic cues that carry particular types of emotional information. Intelligibility of verbal content is removed by all of these procedures, permitting the use of natural, "real life" speech material without regard for the cue value of the spoken message. For example, in the study by Scherer et al. (1984), highly natural speech samples from interactions between civil servants and citizens were used. Using a variety of masking procedures allowed determining which type of cues carried what type of affective information (with respect to both emotions and speaker attitudes). Interestingly, judges were still able to detect politeness when the speech samples were masked by all methods jointly. The data further showed that voice quality parameters and F0 level covaried with the strength of the judged emotion, conveying emotion independently of the verbal context. In contrast, different intonation contour types interacted with sentence grammar in a configurational manner, and conveyed primarily attitudinal information.

(3) *Cue manipulation via synthesis.* The development of synthesis and copy synthesis methods in the domain of speech technology has provided researchers in the area of vocal expression of emotion with a remarkably effective tool, allowing to systematically manipulate vocal parameters to determine the relative effect of these changes on listener judgment. In a very early study, Lieberman and Michaels (1962) produced variations of F0 and envelope contour and measured the effects on emotion inference. Scherer and Oshinsky (1977) used a MOOG synthesizer and systematically varied, in a complete factorial design, amplitude variation, pitch level, contour and variation, tempo, envelope, harmonic richness, tonality, and rhythm in sentence-like sound sequences and musical melodies. Their data showed the relative strength of the effects of individual parameters, as well as of particular parameter configurations, on emotion attributions by listeners. More recent work (Abadjieva et al., 1995; Breitenstein et al., 2001; Burkhardt and Sendlmeier, 2000; Cahn, 1990; Carlson, 1992; Carlson et al., 1992; Heuft et al., 1996; Murray and Arnott, 1993; Murray et al., 1996) shows great promise as also demonstrated in other papers in this issue. Not surprisingly, much of this work is performed with the aim of adding emotional expressivity to text to speech systems.

Copy synthesis (or resynthesis) techniques allow one 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, this technique has been used to systematically manipulate F0 level, contour variability and range, intensity, duration, and accent structure of real utterances (Ladd et al., 1985; Tolkmitt et al., 1988). Listener judgments of apparent speaker attitude and emotional state for each stimulus showed strong direct effects for all of the variables manipulated, with relatively few effects due to interactions between the manipulated variables. Of all variables studied, F0 range effected the judgments the most, with narrow F0 range signaling sadness and wide F0 range being judged as expressing high arousal, negative emotions such as annoyance or anger. High speech intensity was also interpreted as a sign of negative emotions or aggression. Fast speech led to inferences of joy, with slow speech judged as a mark of sadness. Only very minor effects for speaker and utterance content were found, indicating that the results are likely to generalize over different speakers and utterances.

It is to be expected that the constant improvement of speech synthesis will provide the research community with ever-more refined and natural sounding algorithms that should allow a large number of systematic variations of acoustic parameters pertinent to emotional communication. One of the developments that are most eagerly awaited by researchers in this domain is the possibility to use different glottal source characteristics for the manipulation of voice timbre in formant synthesis. There have been some recent reports on work in that direction (Burkhardt and Sendlmeier, 2000; Carlson, 1992; Lavner et al., 2000).

2.4. Transmission studies

The Brunswikian lens model acknowledges the need to separately model the transmission of the distal signals from the sender to the receiver/listener where they are represented on a subjective, proximal level. Most studies in the area of the vocal communication of emotion have considered this voyage of the sound waves from the mouth of the speaker to the ear or brain of the listener as a 'quantité négligeable' and have given little or no consideration to it. However, a Brunswikian approach, trying to systematically model all aspects of the communication process forces the researcher to separately model this element since it may be responsible for errors in inference. This would be the case, for example, if the transmission process systematically changed the nature of the distal cues. In this case, the proximal cues could not be representative of the characteristics of the distal cues at the source. A well-known example is the filtering of the voice over standard telephone lines. Given the limited frequency range of the transmission, much of the higher frequency band of the voice signal is eliminated. The effect on the listener, i.e., the systematic biasing of the proximal cues, is to hear the pitch of the voice as lower as it is—with the well known ensuing effect of overestimating the age of the speaker at the other end (often making it impossible to keep mother and daughter apart). Scherer et al. (in press) discuss two of the contributing factors in detail: (1) the transmission of sound through space and electronic channels, and (2) the transform functions in perception, as determined by the nature of human hearing mechanisms. Below, the issues will be briefly summarized.

(1) The transmission of vocal sound through physical space is affected by many environmental and atmospheric factors including the distance between sender and receiver (weakening the signal), the presence of other sounds such as background noise (disturbing the signal), or the presence of natural barriers such as walls (filtering the signal). All of these factors will lessen the likelihood that the proximal cues can be a faithful representation of the distal cues with respect to their acoustic characteristics. Importantly, the knowledge about such constraints on the part of the speaker can lead to a modification of the distal cues by way of an effort to modify voice production to offset such transmission effects. For example, a speaker may shout to produce more intense speech, requiring more vocal effort, and greater articulatory precision, if communicating over large distances. Greater vocal effort in turn, will affect a large number of other acoustic characteristics related to voice production at the larynx. Furthermore, the distance between speaker and listener may have an effect on posture, facial expression and gestures, all of which are likely to have effects on the intensity and spectral distribution of the acoustic signal (see Laver, 1991).

Just as distance, physical environment, and atmospheric conditions in the case of immediate voice transmission, the nature of the medium may have potentially strong effects on proximal cues in the case of mediated transmission by intercom, telephone, Internet, or other technical means. The band restrictions of the line, as well as the quality of the encoding and decoding components of the system can systematically affect many aspects of the distal voice cues by disturbing, masking, or filtering and thus render the proximal cues less representative of the distal cues. While much of this work has been done in the field of engineering, only little has been directly applied to modeling the process of the vocal communication of emotion (but see Baber and Noyes, 1996). It is to be hoped that future research efforts in the area will

pay greater attention to this important aspect of the speech chain.

(2) Another factor that is well known but generally neglected in research on vocal communication is the transformation of the distal signal through the transfer characteristics of the human hearing system. 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 and Sundberg, 1988; Titze, 1992). In addition, listeners seem to use an internal model of the specific spectral distribution of harmonics and noise in loud and soft voices to infer the vocal effort with which a voice was produced. Klasmeyer (1999, p. 112) has shown that this judgment is still reasonably correct when both loud and soft voices are presented at the same perceived loudness level (the soft voice having more than six times higher overall intensity than the loud voice). Similar effects can be shown for the perception of F0 contours and the perceived duration of a spoken utterance (see (Scherer et al., in press) for further detail).

Given space limitations I could only very briefly touch upon the issue of voice perception which is centrally implicated in the distal–proximal relationship. Recent research paints an increasingly more complex picture of this process which is characterized by a multitude of feedforward and feedback processes between modalities (e.g., visual–vocal), levels of processing (e.g., peripheral–central), and input versus stored schemata. There is, for example, the well-documented phenomenon of the effects of facial expression perception on voice perception (e.g., Borod et al., 2000; de Gelder, 2000; de Gelder and Vroomen, 2000; Massaro, 2000). Things are further complicated by the fact that voice generally carries linguistic information with all that this entails with respect to the influence of phonetic-linguistic categories on perception. A particularly intriguing phenomenon is the perception and interpretation of prosodic features which is strongly affected by centrally stored templates. There is currently an upsurge in neuropsychological research activity in this area (e.g., Pell, 1998; Shipley-Brown et al., 1988; Steinhauer et al., 1999; see also Scherer et al., in press).

The fact that 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, is not only regrettable for theoretical reasons (preventing one from modeling the communication process in its entirety) but also because one may have missed parameters that are central in the differentiation of the vocal expression of different emotions. Thus, rather than basing all analyses on objective measures of emotional speech, it might be fruitful to employ, in addition, variables that reflect the transformations that distal cues undergo in the representation as proximal cues after passage through the hearing mechanism (e.g., perceived loudness, perceived pitch, perceived rhythm, or Bark bands in spectral analysis; see (Zwicker, 1982)).

2.5. Representation studies

The result of sound transmission and reception by the listener's hearing system is the storage, in short term memory, of the proximal cues or percepts of the speaker's vocal production. Modeling of the voice communication process with the Brunswikian lens model requires direct measurement of these proximal cues, i.e., the subjective impressions of the vocal qualities. Unfortunately, this is one of the most difficult aspects of the research guided by this model. So far, the only possibility to empirically measure these proximal cues is to rely on listeners' verbal report of their impressions of the speaker's voice and speech. As is always the case with verbal report, the description is constrained by the semantic categories for which there are words in a specific language. In addition, since one wants to measure the impressions of naïve, untrained listeners, these words must be currently known and used, in order to ensure that all listeners studied understand the same concept and the underlying acoustic pattern by a particular word. In most languages, including English, there are only relatively few words describing vocal qualities and only a subset of those is frequently used in normal speech or in literature.

Only very few efforts have been made to develop instruments that can be used to obtain proximal voice percepts. One of the pioneers in

this area has been John Laver (1980, 1991) who suggested a vocal profile technique to describe voice quality on the basis of phonatory and articulatory settings and to obtain reliable judgments on these settings. In some ways, of course, Laver attempts to get at the distal cues via the use of perceived voice quality rather than obtaining an operationalized measure of the subjective impression (in other words, he wants his judges to agree whereas a psychologist interested in proximal cues would be happy to see differences in perceived voice quality). Scherer (1978, 1982a,b) designed a rating sheet to obtain voice and speech ratings reflecting such subjective impressions from naïve listeners. This effort was later continued by an interdisciplinary group of researchers (Sangsue et al., 1997). One of the main problems is to obtain sufficient reliability (as measured by interrater agreement) for many of the categories that are not very frequently employed, such as nasal, sharp, or rough. In our laboratory, Bänziger is currently involved in another attempt to develop a standard instrument for proximal voice cue measurement in the context of Brunswikian modeling (see Bänziger and Scherer, 2001).

Representation studies, of which there are currently extremely few (see (Scherer, 1974) for an example in the area of personality), study the inference structures, i.e., the mental algorithms that are used by listeners, to infer emotion on the basis of the proximal cues. The inference algorithms must be based on some kind of representation, based on cultural lore and/or individual experience, as to which vocal impressions are likely to be indicative of particular types of emotion.

3. Issues for future research

3.1. How should emotion be defined?

One of the major problems in this area is the lack of a consensual definition of emotion and of qualitatively different types of emotions. The problem already starts with the distinction of emotions from other types of affective states of the speaker, such as moods, interpersonal stances, attitudes or even affective personality traits. Scherer (2000a) has suggested using a design feature approach to distinguish these different types of affective states. Table 4 reproduces this proposal. It is suggested to distinguish the different states by the relative intensity and duration, by the degree of synchronization of organismic subsystems, by the extent to which the reaction is focused on an eliciting event, by the degree to which appraisal has played a role in the elicitation, the rapidity of state changes, and the extent to which the state affects open behavior. As the table shows, the different types of affective states do vary quite substantially on the basis of such a design feature grid and it should be possible to distinguish them relatively clearly, if such an effort were consistently made.

It should be noted that emotions constitute quite a special group among the different types of affective states. They are more intense, but of shorter duration. In particular, they are characterized by a high degree of synchronization between the different subsystems. Furthermore, they are likely to be highly focused on eliciting events and produced through cognitive appraisal. In terms of their effect, they strongly impact behavior and are likely to change rapidly. One of the issues that is very important in the context of studying the role of voice and speech as affected by different types of affective states is the question as to how often particular types of affective states are likely to occur and in which contexts. This is particularly important for applied research that is embedded in speech technology contexts. Obviously, it may be less useful to study the effects of very rare affective events on voice and speech if one wants to develop a speech technology instrument that can be currently used in everyday life.

Another issue that must not be neglected is the existence of powerful regulation mechanisms as determined by display rules and feeling rules (Gross and Levenson, 1997; see Scherer, 1994a, 2000b for a detailed discussion). Such regulation efforts are particularly salient in the case of very intense emotions, such as hot anger, despair, strong fear, etc. It is unfortunate, that it is exactly those kinds of emotions, likely to be highly regulated and masked under normal circumstances in

Table 4
Design feature delimitation of different affective states

Type of affective state: brief definition (*examples*)	Intensity	Duration	Synchronization	Event focus	Appraisal elicitation	Rapidity of change	Behavioral impact
Emotion: relatively brief episode of synchronized response of all or most organismic subsystems in response to the evaluation of an external or internal event as being of major significance (*angry*, *sad*, *joyful*, *fearful*, *ashamed*, *proud*, *elated*, *desperate*)	+ +–+ ++	+	+ + +	+ + +	+ + +	+ + +	+ + +
Mood: diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (*cheerful*, *gloomy*, *irritable*, *listless*, *depressed*, *buoyant*)	+–+ +	++	+	+	+	++	+
Interpersonal stances: affective stance taken toward another person in a specific interaction, colouring the interpersonal exchange in that situation (*distant*, *cold*, *warm*, *supportive*, *contemptuous*)	+–+ +	+–+ +	+	++	+	+ + +	++
Attitudes: relatively enduring, affectively coloured beliefs, preferences, and predispositions towards objects or persons (*liking*, *loving*, *hating*, *valueing*, *desiring*)	0–+ +	+ +–+ ++	0	0	+	0–+	+
Personality traits: emotionally laden, stable personality dispositions and behavior tendencies, typical for a person (*nervous*, *anxious*, *reckless*, *morose*, *hostile*, *envious*, *jealous*)	0–+	+ + +	0	0	0	0	+

0: low, +: medium, ++: high, + + +: very high, –: indicates a range.

everyday life, that are consistently studied by research in this area. While such research is of great interest in trying to understand the psychobiological basis of emotional expression in the voice, it is not necessarily the best model to look at the effect of affective states on voice and speech in everyday life with the potential relevance to speech technology development. Rather, one would think that the kinds of affective modifications of voice and speech that are likely to be much more pervasive in everyday speaking, such as the effects of stress (see Baber and Noyes, 1996; Tolkmitt and Scherer, 1986), deception (see Anolli and Ciceri, 1997; Ekman et al., 1991), relatively mild mood changes, attitudes or interpersonal stances or styles vis-à-vis a particular person (see Ambady et al., 1996; Granström, 1992; Ofuka et al., 2000; Scherer et al., 1984; Siegman, 1987a,b; Tusing and Dillard, 2000), may be of much greater import to that kind of applied question.

3.2. Which theoretical model of emotion should be adopted?

This issue, which has received little attention so far, has strong implications for the types of emotions to be studied, given that different psychological theories of emotion have proposed different ways of cutting up the emotion domain into different qualitative states or dimensions. The issue of the most appropriate psychological model of emotion is currently intensively debated in psychology. Scherer (2000a) has provided an outline of the different types of approaches and the relative advantages and disadvantages. The choice of model also determines the nature of the theoretical predictions for vocal patterning. While past research in this area has been essentially atheoretical, it would be a sign of scientific maturity if future studies were based on an explicit set of hypotheses, allowing more cumulative research efforts.

The two theoretical traditions that have most strongly determined past research in this area are discrete and dimensional emotion theories. Discrete theories stem from Darwin and are defended by Tomkins (1962), Ekman (1992) and Izard (1971). Theorists in this tradition propose the existence of a small number, between 9 and 14, of basic or fundamental emotions that are characterized by very specific response patterns in physiology as well as in facial and vocal expression. Most studies in the vocal effects of emotion area have followed this model, choosing to examine the effects of happiness, sadness, fear, anger and surprise. More recently, the term ''the big six'' has gained some currency, implying the existence of a fundamental set of six basic emotions (although there does not seem to be agreement on which six these should be). Unfortunately, to date there has been no attempt by discrete emotion theorists to develop concrete predictions for vocal patterning.

Another model that has guided quite a bit of research in this area is the dimensional approach to emotion, which goes back to Wundt (1874/1905). In this tradition different emotional states are mapped in a two or three-dimensional space. The two major dimensions consist of the valence dimension (pleasant–unpleasant, agreeable–disagreeable) and an activity dimension (active–passive) (see Scherer, 2000a). If a third dimension is used, it often represents either power or control. Researchers using this model in the investigation of emotion effects on voice often restrict their approach to studying positive versus negative and active versus passive differences in the emotional states, both with respect to production and measurement of listener inferences. Recently, Bachorowski and her collaborators (Bachorowski, 1999; Bachorowski and Owren, 1995) have called attention to this approach. However, a more elaborate theoretical underpinning of this position and the development of concrete predictions remains to be developed.

There is some likelihood that a new set of psychological models, *componential models of emotion*, which are often based on appraisal theory (see Scherer et al. 2001b), is gaining more influence in this area. The advantage of such componential models is that the attention is not restricted to subjective feeling state, as is the case with dimensional theories, nor to a limited number of supposedly basic emotions, as is the case with discrete emotion theory. Componential models are much more open with respect to the kind of emotional states that are likely to occur, given that they do not expect emotions to consist of a limited number of relatively fixed neuromotor patterns, as in discrete emotion theory. And rather than limiting the description to two basic dimensions, these models emphasize the variability of different emotional states, as produced by different types of appraisal patterns. Furthermore, they permit to model the distinctions between members of the same emotion family, an issue that, as described above, plays a very important role in this research area.

The most important advantage of componential models, as shown, for the example, by Scherer's component process model described above, is that these approaches provide a solid basis for theoretical elaborations of the mechanisms that are supposed to underlie the emotion–voice relationship and permit to generate very concrete hypotheses that can be tested empirically. Table 5 provides an example for the hypotheses on the effect of appraisal results on the vocal mechanisms and the ensuing changes in acoustic parameters. These predictions can be examined directly by systematically manipulating the appraisal dimensions in laboratory studies (see below). Furthermore, this set of predictions also generates predictions for complete patterns of modal emotions (see (Scherer, 1994b) for the difference to basic emotions) since there are detailed hypotheses as to which profiles of appraisal outcomes on the different dimensions produce certain modal emotions. For example, a modal emotion that is part of the fear-family is predicted to be produced by the appraisal of an event or situation as obstructive to one's central needs and goals, requiring urgent action, being difficult to control through human agency, and lack of sufficient power or coping potential to deal with the situation. The major difference to anger-producing appraisal is that the latter entails a much higher evaluation of controllability and available coping potential.

Based on these kinds of hypotheses, shared among appraisal theorists, the predictions in Table 5

Table 5
Component patterning theory predictions for voice changes (adapted from Scherer, 1986)

Novelty check	
Novel	Not novel
Interruption of phonation, sudden inhalation, ingressive (fricative) sound with glottal stop (noise-like spectrum)	No change
Intrinsic pleasantness check	
Pleasant	Unpleasant
Faucal and pharyngeal expansion, relaxation of tract walls, vocal tract shortened by mouth corners retracted upward (increase in low frequency energy, F1 falling, slightly broader F1 bandwidth, velopharyngeal nasality, resonances raised) "*Wide voice*"	Faucal and pharyngeal constriction, tensing of tract walls, vocal tract shortened by mouth corners retracted downward (more high frequency energy, F1 rising, F2 and F3 falling, narrow F1 bandwidth, laryngopharyngeal nasality, resonances raised) "*Narrow voice*"
Goal/need significance check	
Relevant and consistent	Relevant and discrepant
Overall relaxation of vocal apparatus, increased salivation (F0 at lower end of range, low-to-moderate amplitude, balanced resonance with slight decrease in high-frequency energy) "*Relaxed voice*"	Overall tensing of vocal apparatus, decreased salivation (F0 and amplitude increase, jitter and shimmer, increase in high frequency energy, narrow F1 bandwidth, pronounced formant frequency differences) "*Tense voice*"
If event conducive to goal: *relaxed voice + wide voice*	If event conducive to goal: *tense voice + wide voice*
If event obstructive to goal: *relaxed voice + narrow voice*	If event obstructive to goal: *tense voice + narrow voice*
Coping potential check	
Control	No control
(As for relevant and discrepant) "*Tense voice*"	Hypotonus of vocal apparatus (low F0 and restricted F0 range, low amplitude, weak pulses, very low high-frequency energy, spectral noise, format frequencies tending toward neutral setting, broad F1 bandwidth) "*Lax voice*"
High power	Low power
Deep, forceful respiration, chest register phonation (low F0, high amplitude, strong energy in entire frequency range) "*Full voice*"	Head register phonation (raised F0, widely spaced harmonics with relatively low energy) "*Thin voice*"
Norm/self-compatibility check	
Standards surpassed	Standards violated
Wide voice + full voice (+ relaxed voice if expected or + tense voice if unexpected)	Narrow voice + thin voice (+ lax voice if no control or + tense voice if control)

can be synthesized to predicted patterns of acoustic changes that can be expected to characteristically occur during specific modal emotions. These predicted patterns are shown in Table 6. This table also provides information on which predictions have been supported (predicted change occurring in the same direction) and which have been directly contradicted (actual change going into the opposite direction from prediction) in the study by Banse and Scherer (1996). Another test of these predictions has recently been performed by Juslin and Laukka (2001). These researchers again found

Table 6
Predictions for emotion effects on selected acoustic parameters (based on Table 4 and appraisal profiles; adapted from Scherer, 1986)

	ENJ/ HAP	ELA/ JOY	DISP/ DISG	CON/ SCO	SAD/ DEJ	GRI/ DES	ANX/ WOR	FEAR/ TER	IRR/ COA	RAG/ HOA	BOR/ IND	SHA/ GUI
F0												
Perturbation	<=	>			>	>		>		>		
Mean	< ✓	> ✓	>	<>	<> ✓	> ✓	>?	>> ✓	<> ✓	<>	< ✓	>?
Range	<=	>			<	>		>>	<	>>		
Variability	<	>			<	>?		>>?	<	>> ✓		
Contour	<	>			<	>	>	>>	<	=		>
Shift regularity	=	<						<		<	>	
Formants												
F1 Mean	<	<	>	>	>	>	>	>	>	>	>	>
F2 Mean			<	<	<	<	<	<	<	<	<	<
F1 Bandwidth	>	<>	<<	<	<>	<<	<	<<	<<	<<	<	<
Formant precision		>	>	>	<	>	>	>	>	>		>
Intensity												
Mean	< ✓	> ✓	>?	>>?	<< ✓	> ✓		> ✓	> ✓	>> ✓	<>	
Range	<=	>			<			>	>	>		
Variability	<	>			<			>		>		
Spectral parameters												
Frequency range	>	>	>	>>	>	>>		>>	>	>	>	
High-frequency energy	<	<> ✓	>	>	<>	>> ✓	>?	>>	>>	>> ✓	<>	>
Spectral noise					>							
Duration												
Speech rate	<?	> ✓			< ✓	>		>> ✓		> ✓		
Transition time	>	<			>	<		<		<		

Note: ANX/WOR: anxiety/worry; BOR/IND: boredom/indifference; CON/SCO: contempt/scorn; DISP/DISG: displeasure/disgust; ELA/JOY: elation/joy; ENJ/HAP: enjoyment/happiness; FEAR/TER: fear/terror; GRI/DES: grief/desperation; IRR/COA: irritation/cold anger; RAGE/HOA: rage/hot anger; SAD/DEJ: sadness/dejection; SHA/GUI: shame/guilt; F0: fundamental frequency; F1: first formant; F2: second formant; >: increase; <: decrease. Double symbols indicate increased predicted strength of the change. Two symbols pointing in opposite directions refer to cases in which antecedent voice types exert opposing influence. (✓) prediction supported, (?) prediction contradicted by results in (Banse and Scherer, 1996).

support for many predictions as well as cases of contradiction indicating the need to revise the theoretical predictions.

3.3. How should emotional speech samples be obtained?

Part of this question is obviously linked to the psychological model of emotion that is being adopted and the considerations that follow there from. Another issue concerns the aim of the respective study, for example the relative link to applied questions for which a high degree of ecological validity is needed. However, the most important issues concern the nature of the emotion induction for the context in which the voice samples are obtained. As mentioned above, most research has used the portrayal of vocal emotions by lay or professional actors. This approach has been much criticized because of the potential artifacts that might occur with respect to actors using stereotypical expression patterns. However, as discussed above, the issue is much more complex. It can be argued, that actors using auto-induction can produce valid exemplars of vocal expression. With respect to regulation attempts such as display rules, it has been argued that most normal speech can be expected to be more or less "acted" in a way that is quite similar to professional acting (see above).

On the whole, it would seem that many of the issues having to do with portrayal have been insufficiently explored. For example, there is good evidence that the use of laymen or even lay actors is potentially flawed since these speakers do not master the ability to voluntarily encode emotions in such a way that they can be reliably recognized. It is generally required that portrayed emotions be well recognized by listeners, representing at least what can be considered as a shared communication pattern for emotional information in a culture. Using professional actors mastering Stanislavski techniques or method acting, as well as providing appropriate instructions for the portrayals, including realistic scenarios, can alleviate many of the difficulties that may be associated with acting. Yet, obviously, one has to carefully investigate to what extent such acted material corresponds to naturally occurring emotional speech. Unfortunately, so far there has been no study in which a systematic attempt has been made to compare portrayed and naturally occurring vocal emotions. This would seem to be one of the highest priorities in the field. A first attempt is currently made by Scherer and his collaborators in the context of the ASV study mentioned above: In addition to the experimental induction of different states through a computerized induction battery, speakers are asked to portray a number of comparable states and of discrete emotions, allowing to compare natural and portrayed affect variations in the voice within and across speakers (Scherer et al., 2000).

With the increasing availability of emotional material in the media, especially television, there has been a tendency to rely on emotional speech to be recorded off the air. However, this procedure has also many obvious drawbacks. As to the presumed spontaneity of the expression, one cannot, unfortunately, exclude that there have been prior arrangements between the organizers of game or reality shows and the participants. One general question that can never be answered satisfactorily is to what extent the fact of being on the air will make speakers act almost like an actor would in a particular situation. This danger is particularly pronounced in cases in which normal individuals are on public display, as in television shows. Not only is their attention focussed on how their expressions will be evaluated by the viewers, encouraging control of expression (display rules; Ekman, 1972), in addition there are strong "feeling rules" in the sense of what they think is required of them as an appropriate emotional response (Hochschild, 1983). Furthermore, it is not always obvious which emotion the speakers have really experienced in the situation; since in many cases speakers cannot be interrogated on what they really felt in the situation (especially in cases in which material is taped off the radio or television). Generally, researchers infer emotional quality on the basis of the type of situation, assuming that everybody would react in the same fashion to a comparable event. However, as appraisal theorists of emotion have convincingly shown (see overviews in (Scherer, 1999a; Scherer et al., 2001b)) it is

the subjective appraisal of a situation, which may be highly different between individuals, that determines the emotional reaction. Finally, in many cases, it is impossible to obtain a more systematic sampling of vocal production for different types of emotions for the same speaker and a comparable set for other speakers. This, however, is absolutely essential, given the great importance of individual differences in the vocal encoding of emotion (see Banse and Scherer, 1996). In consequence, the hope of speech researchers to use emotional expression displays recorded from television game shows or other material of this sort as a corpus of "natural expressions" may well reside on a somewhat hollow foundation.

One of the methods to obtain emotional voice samples that has been rather neglected in this research domain is the experimental induction of emotion. There are several possibilities for changing the state of speakers through experimental manipulations that have been shown to yield rather consistent results. One of the most straightforward possibilities is to increase the cognitive load of the speaker while producing speech, generally through some kind of rather effortful cognitive tasks. This procedure is often used to induce cognitive stress and has been shown to reliably produce speech changes (Bachorowski and Owren, 1995; Scherer et al., 1998; Karlsson et al., 2000; Tolkmitt and Scherer, 1986). Another possibility is to use the presentation of agreeable and disagreeable media material, such slides, videotapes, etc. that can also be shown to produce relatively consistent emotional effects in many subjects (Westermann et al., 1996). This procedure has also been shown reliably to produce vocal changes in a number of studies. It is often the method of choice in studies of facial expressions and one could assume, in consequence that it could provide a useful method for vocal researchers well.

Psychologists have developed quite a number of induction techniques of this type. One of the more frequently used techniques is the so-called *Velten* induction technique (Westermann et al., 1996). In this method speakers read highly emotionally charged sentences over and over again which seems to produce rather consistent affective changes. Another possibility that is used rather frequently in the domain of physiological psychology is to produce some rather realistic emotional inductions through the behavior of experimenters in the laboratory. For example, Stemmler and his collaborators (Stemmler et al., 2001) have angered subjects through rather arrogant and offensive behavior of the experimenter. The size of the experimental physiological reactions showed the degree to which that manipulation resulted in rather intense emotions. Thomas Scherer (2000) has analyzed vocal production of the subjects under this condition, demonstrating rather sizable effects for several vocal parameters. Finally, a number of researchers have experimented with computer games as a means of emotion induction in highly motivated players (see Johnstone et al., 2001; Kappas, 1997). For example, in order to examine some of the detailed predictions in Scherer's (1986) component process model, Johnstone (2001) manipulated different aspects of an arcade game (e.g., systematically varying a player's power by varying shooting capacity) and observed the direct effect of emotion antecedent appraisal results on vocal parameters. A number of significant main effects, and particularly interaction effects for different appraisal dimensions, illustrate the high promise of this approach.

Obviously, the experimental manipulation of emotional state is the preferred method for obtaining controlled voice and speech variations over many different speakers allowing one to clearly examine the relative importance of different emotions as well as individual differences between speakers. Unless this possibility of investigating within-speaker differences for different emotions and between-speaker differences for set of emotions is provided, our knowledge of the mechanisms that allow emotions to mediate voice production and to affect acoustic parameters will remain very limited indeed.

Another issue that is of great importance in this context is the validation of the effect of a portrayal or manipulation. In many recent studies, it seems to be taken for granted that if one requires the speaker, either a lay or professional actor, to produce a certain portrayal the effect is likely to be satisfactory. As can be shown in a large number of studies, one has to use rather sophisticated selection

methods to select those portrayals that do actually communicate the desired emotional state. In consequence, in this context it is not advisable to do acoustic parameter extraction for vocal portrayals for which it has not been clearly demonstrated that judges can reliably judge the emotion to be portrayed, using appropriate statistical techniques (see Banse and Scherer, 1996). Similarly, in studies that record natural emotional speech from the media, more efforts need to be made to insure that the presumed emotional arousal is actually present. As mentioned above, it has generally been taken for granted that a certain situation will provoke certain types of emotions. Appraisal theorists have shown (Scherer, 1999a; Roseman and Smith, 2001) this is not necessarily the case. Rather, there is a large likelihood that different individuals will react very differently to a similar situation. To accept a researcher's impression of what emotion underlies a particular expression as a validation is clearly methodologically unsound. At the very least, one would require a consensual judgment of several individuals to agree on what the likely meaning of the expression is. In most experimental studies using induction methods, the validation is automatically produced through the use of manipulation checks, probably the best method to insure that a particular emotional state has been at the basis of a particular type of voice production.

3.4. How to measure emotional inferences?

Most studies in this area have used the rather convenient procedure of presenting listeners with lists of emotion that correspond to the types of stimuli that are used in the study, requiring them to judge which emotion best describes a particular stimulus (categorical choice). In some studies, judges are given the option to indicate on the rating sheet the relative intensity of several emotions (emotion blends). In the psychology of emotion there has recently been quite a debate about the adequacy of different types of methods to obtain judgments on expressive material. In the domain of the facial expression of emotion, Russell (1994) has strongly attacked the use of fixed lists of alternatives, demonstrating potential artifacts with this method. Critics have pointed out that some of the demonstrations by Russell are clearly biased (Ekman, 1994). The method of providing fixed alternatives may not yield results that are too different from asking subjects to use free report (Ekman, personal communication). However, the debate has shown that one cannot take a particular method for granted and that it is necessary to examine potential problems with different methods (see Frank and Stennett, 2001).

In addition, there are serious problems of statistical analysis, especially in the case where judges have to provide intensity ratings for several emotions. In many cases, there are large numbers of zero responses, which changes the nature of the statistical analyses to be used. Furthermore, there is no established method for analyzing *mixed or blended emotions*, both with respect to the production and to the inference from expressive material. All these issues pose serious problems for research in this area of which so far remain largely unacknowledged (see (Scherer, 1999a) for a review of these issues).

3.5. Which parameters should be measured?

As has been shown above, part of the problem in demonstrating the existence of qualitative differences between the vocal expression for different types of discrete emotions and the hypothesis that the voice might only reflect arousal, has been due to a limitation with respect to the number of acoustic parameters extracted. Recent research using a larger number of parameters (Banse and Scherer, 1996; Klasmeyer, 2000) has shown that spectral parameters play a major role in differentiating qualitative differences between emotions. In consequence, it is absolutely essential that a large variety of parameters be measured in future research. One of the more important aims would seem to be to achieve some kind of standardization with respect to the types of parameters extracted. In order to insure accumulation of findings and a high degree of comparability of results, it is necessary to measure similar sets of parameters (including similar transformations and aggregations) in studies conducted by different investigators in multiple laboratories.

3.6. New questions and research paradigms

In recent years there have been enormous technological advances in speech technology, physiological recording, and brain imagery. The quality of the methods and their widespread availability will hopefully encourage future research attempting to come to grips with the underlying mechanisms in voice production and voice perception. As mentioned above, there is currently a strong increase in work on the neurophysiological grounding of the perception of vocal and prosodic features, especially as linked to emotional expression. Unfortunately, there has been less neuroscience research on voice production in states of emotional arousal. There is little question that the study of the brain mechanisms underlying the production and perception of emotional speech will be one of the major growth areas in this field. The same is true for speech technology. While the engineers working in this area showed little interest in emotional speech until fairly recently, this situation has drastically changed (as demonstrated by the ISCA conference in Newcastle upon which this special issue is based; see editors' introduction). Unfortunately, there still is a tendency on their part to neglect the underlying mechanisms in the interest of fast algorithmic solutions.

In spite of the methodological advances and the much increased research activity in this area, one major shortcoming—the lack of interaction between researchers from various disciplines, working on different angles and levels of the problem—remains unresolved. If one were able to overcome this isolation, allowing neuroscientists, for example, to use high-quality synthesis of emotional utterances, systematically varying theoretically important features, one could rapidly attain a much more sophisticated type of research design and extremely interesting data. However, beneficial interdisciplinary effects are not limited to technologically advanced methodology. Neuroscientists and speech scientists could benefit from advances in the psychology of emotion, allowing them to ask more subtle questions than that of different acoustic patterning of a handful of so-called basic emotions.

4. Conclusion

In the past, studies on the facial expression of emotion have far outnumbered work on vocal expression. There are signs that this may change. Given the strong research activity in the field, there is hope that there will soon be a critical mass of researchers that can form an active and vibrant research community. The work produced by such a community will have an extraordinary impact since it is, to a large extent, interdisciplinary work, given the origin of the researchers coming from acoustics, engineering, linguistics, phonetics, and psychology, and contributing concepts, theories, and methods from their respective disciplines. Given the ever-increasing sophistication of voice technology, both with respect to analysis and synthesis, in the service of this research, we are likely to see some dramatic improvement in the scope and quality of research on vocal correlates of emotion and affective attitudes in voice and speech. However, in our enthusiasm to use the extraordinary power of voice technology, we need to be careful not to compromise the adequacy of the overall research design and the sophistication of emotion definition and measurement for the elegance of vocal measurement and synthesis. To this end, and to arrive at a truly cumulative process of building up research evidence, we need a thorough grounding of present and future research in past work and its achievements. In line with the position taken throughout this paper, I strongly encourage the continuous refinement of theoretical models based on established literature in all of the disciplines pertinent to the issue of emotion effects on the voice. It can serve as a strong antidote to one-shot, quickly designed studies that are more likely than not to repeat the errors made in the past.

References

Abadjieva, E., Murray, I.R., Arnott, J.L., 1995. Applying analysis of human emotional speech to enhance synthetic speech. In: Proc. Eurospeech 95, pp. 909–912.

Allport, G.W., Cantril, H., 1934. Judging personality from voice. J. Soc. Psychol. 5, 37–55.

Alpert, M., Kurtzberg, R.L., Friedhoff, A.J., 1963. Transient voice changes associated with emotional stimuli. Arch. Gen. Psychiatr. 8, 362–365.

Ambady, N., Koo, J., Lee, F., Rosenthal, R., 1996. More than words: Linguistic and nonlinguistic politeness in two cultures. J. Pers. Soc. Psychol. 70 (5), 996–1011.

Amir, N., Ron, S., 1998. Towards an automatic classification of emotions in speech. In: Proc. ICSLP 98, Vol. 3, pp. 555–558.

Anolli, L., Ciceri, R., 1997. The voice of deception: Vocal strategies of naive and able liars. J. Nonverb. Behav. 21 (4), 259–284.

Baber, C., Noyes, J., 1996. Automatic speech recognition in adverse environments. Hum. Fact. 38 (1), 142–155.

Bachorowski, J.A., 1999. Vocal expression and perception of emotion. Current Direct. Psychol. Sci. 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. Psychol. Sci. 6 (4), 219–224.

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70 (3), 614–636.

Bänziger, T., Scherer, K.R., 2001. Relations entre caractéristiques vocales perçues et émotions attribuées. In: Proc. Journées Prosodie, Grenoble, 10–11 October 2001.

Bonner, M.R., 1943. Changes in the speech pattern under emotional tension. Amer. J. Psychol. 56, 262–273.

Borod, J.C., Pick, L.H., Hall, S., Sliwinski, M., Madigan, N., Obler, L.K., Welkowitz, J., Canino, E., Erhan, H.M., Goral, M., Morrison, C., Tabert, M., 2000. Relationships among facial, prosodic, and lexical channels of emotional perceptual processing. Cogn. Emot. 14 (2), 193–211.

Bortz, J., 1966. Physikalisch-akustische Korrelate der vokalen Kommunikation. Arbeiten aus dem psychologischen Institut der Universität Hamburg, 9 [Reports of the Psychology Department, University of Hamburg, Vol. 9].

Breitenstein, C., Van Lancker, D., Daum, I., 2001. The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. Cogn. Emot. 15 (1), 57–79.

Brunswik, E., 1956. Perception and the Representative Design of Psychological Experiments. University of California Press, Berkeley.

Bryman, A., Cramer, D., 1990. Quantitative Data Analysis for Social Scientists. Taylor and Francis, Hove, Great Britain.

Burkhardt, F., Sendlmeier, W.F., 2000. Verification of acoustical correlates of emotional speech using formant-synthesis. In: Proc. ISCA Workshop on Speech and Emotion, 5–7 September 2000, Newcastle, Northern Ireland, pp. 151–156.

Caffi, C., Janney, R.W., 1994. Toward a pragmatics of emotive communication. J. Pragmat. 22, 325–373.

Cahn, J., 1990. The generation of affect in synthesised speech. J. Amer. Voice I/O Soc. 8, 1–19.

Carlson, R., 1992. Synthesis: Modeling variability and constraints. Speech Communication 11 (2–3), 159–166.

Carlson, R., Granström, B., Nord, L., 1992. Experiments with emotive speech—acted utterances and synthesized replicas. In: Proc. ICSLP 92, Vol. 1, pp. 671–674.

Coleman, R.F., Williams, R., 1979. Identification of emotional states using perceptual and acoustic analyses. In: Lawrence, V., Weinberg, B. (Eds.), Transcript of the 8th Symposium: Care of the Professional Voice, Part I. The Voice Foundation, New York.

Cosmides, L., 1983. Invariances in the acoustic expression of emotion during speech. J. Exp. Psychol.: Hum. Percept. Perform. 9, 864–881.

Costanzo, F.S., Markel, N.N., Costanzo, P.R., 1969. Voice quality profile and perceived emotion. J. Couns. Psychol. 16, 267–270.

Cowie, R., Douglas-Cowie, E., 1996. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: Proc. ICSLP'96, Philadelphia, pp. 1989–1992.

Darwin, C., 1872. The Expression of Emotions in Man and Animals. John Murray, London (third ed., P. Ekman (Ed.), 1998, Harper Collins, London).

Davitz, J.R., 1964. The Communication of Emotional Meaning. McGraw-Hill, New York.

de Gelder, B., 2000. Recognizing emotions by ear and by eye. In: Lane, R.D., Nadel, L. (Eds.), Cognitive Neuroscience of Emotion. Oxford University Press, New York, pp. 84–105.

de Gelder, B., Vroomen, J., 2000. Bimodal emotion perception: Integration across separate modalities, cross-modal perceptual grouping or perception of multimodal events? Cogn. Emot. 14, 321–324.

Duncan, G., Laver, J., Jack, M.A., 1983. A psycho-acoustic interpretation of variations in divers' voice fundamental frequency in a pressured helium–oxygen environment. Work in Progress, 16, 9–16. Department of Linguistics, University of Edinburgh, UK.

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

Ekman, P., 1992. An argument for basic emotions. Cogn. Emot. 6 (3/4), 169–200.

Ekman, P., 1994. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. Psychol. Bull. 115, 268–287.

Ekman, P., O'Sullivan, M., Friesen, W.V., Scherer, K.R., 1991. Face, voice, and body in detecting deception. J. Nonverb. Behav. 15, 125–135.

Eldred, S.H., Price, D.B., 1958. A linguistic evaluation of feeling states in psychotherapy. Psychiatry 21, 115–121.

Erickson, D., Fujimura, O., Pardo, B., 1998. Articulatory correlates of prosodic control: Emphasis and emotion. Lang. Speech 41, 399–417.

Fairbanks, G., Hoaglin, L.W., 1941. An experimental study of the durational characteristics of the voice during the expression of emotion. Speech Monogr. 8, 85–90.

Fairbanks, G., Pronovost, W., 1939. An experimental study of the pitch characteristics of the voice during the expression of emotion. Speech Monogr. 6, 87–104.

Feldman, R.S., Rimé, B. (Eds.), 1991. Fundamentals of Nonverbal Behavior. Cambridge University Press, New York.

Fichtel, C., Hammerschmidt, K., Jürgens, U., 2001. On the vocal expression of emotion. A multi-parametric analysis of different states of aversion in the squirrel monkey. Behaviour 138 (1), 97–116.

Fonagy, I., 1978. A new method of investigating the perception of prosodic features. Lang. Speech 21, 34–49.

Fonagy, I., 1983. La vive voix. Payot, Paris.

Fonagy, I., Magdics, K., 1963. Emotional patterns in intonation and music. Z. Phonetik 16, 293–326.

Frank, M.G., Stennett, J., 2001. The forced-choice paradigm and the perception of facial expressions of emotion. J. Pers. Soc. Psychol. 80 (1), 75–85.

Frick, R.W., 1985. Communicating emotion: The role of prosodic features. Psychol. Bull. 97, 412–429.

Friend, M., Farrar, M.J., 1994. A comparison of content-masking procedures for obtaining judgments of discrete affective states. J. Acoust. Soc. Amer. 96 (3), 1283–1290.

Frolov, M.V., Milovanova, G.B., Lazarev, N.V., Mekhedova, A.Y., 1999. Speech as an indicator of the mental status of operators and depressed patients. Hum. Physiol. 25 (1), 42–47.

Gifford, R., 1994. A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. J. Pers. Soc. Psychol. 66, 398–412.

Gramming, P., Sundberg, J., 1988. Spectrum factors relevant to phonetogram measurement. J. Acoust. Soc. Amer. 83, 2352–2360.

Granström, B., 1992. The use of speech synthesis in exploring different speaking styles. Speech Communication 11, 347–355.

Green, R.S., Cliff, N., 1975. Multidimensional comparison of structures of vocally and facially expressed emotion. Percept. Psychophys. 17, 429–438.

Gross, J.J., Levenson, R.W., 1997. Hiding feelings: The acute effects of inhibiting negative and positive emotion. J. Abnorm. Psychol. 106 (1), 95–103.

Hammond, K.R., Stewart, T.R. (Eds.), 2001. The Essential Brunswik: Beginnings, Explications, Applications. Oxford University Press, New York.

Hargreaves, W.A., Starkweather, J.A., Blacker, K.H., 1965. Voice quality in depression. J. Abnorm. Psychol. 70, 218–220.

Harper, R.G., Wiens, A.N., Matarazzo, J.D., 1978. Nonverbal Communication: The State of the Art. Wiley, New York.

Hauser, M.D., 2000. The sound and the fury: Primate vocalizations as reflections of emotion and thought. In: Wallin, N.L., Merker, B. (Eds.), The Origins of Music. The MIT Press, Cambridge, MA, USA, pp. 77–102.

Havrdova, Z., Moravek, M., 1979. Changes of the voice expression during suggestively influenced states of experiencing. Activitas Nervosa Superior 21, 33–35.

Helfrich, H., Standke, R., Scherer, K.R., 1984. Vocal indicators of psychoactive drug effects. Speech Communication 3, 245–252.

Herzog, A., 1933. Stimme und Persönlichkeit [Voice and personality]. Z. Psychol. 130, 300–379.

Heuft, B., Portele, T., Rauth, M., 1996. Emotions in time domain synthesis. In: Proc. ICSLP 96, Vol. 3, Philadelphia, USA, pp. 1974–1977.

Hicks, J.W., 1979. An acoustical/temporal analysis of emotional stress in speech. Diss. Abstr. Intern. 41 (4-A).

Hochschild, A.R., 1983. The Managed Heart: The Commercialization of Human Feeling. University of California Press, Berkeley.

Höffe, W.L., 1960. Über Beziehungen von Sprachmelodie und Lautstärke [On the relationships between speech melody and intensity]. Phonetica 5, 129–159.

Huttar, G.L., 1968. Relations between prosodic variables and emotions in normal American English utterances. J. Speech Hear. Res. 11, 481–487.

Iida, A., Campbell, N., Iga, S., Higuchi, F., Yasumura, M., 1998. Acoustic nature and perceptual testing of corpora of emotional speech. In: Proc. ICSLP 98, Sydney, Vol. 4, pp. 1559–1562.

Isserlin, M., 1925. Psychologisch-phonetische Untersuchungen. II. Mitteilung [Psychological-phonetic studies. Second communication]. Z. Gesamte Neurol. Psychiatr. 94, 437–448.

Izard, C.E., 1971. The Face of Emotion. Appleton-Century-Crofts, New York.

Izard, C.E., 1977. Human Emotions. Plenum Press, New York.

Johannes, B., Petrovitsh Salnitski, V., Gunga, H.C., Kirsch, K., 2000. Voice stress monitoring in space – possibilities and limits. Aviat. Space Environ. Md. 71 (9, Section 2, Suppl.), A58–A65.

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: Lewis, M., Haviland, J. (Eds.), Handbook of emotion, second ed.. Guilford, New York, pp. 220–235.

Johnstone, T., van Reekum, C.M., Scherer, K.R., 2001. Vocal correlates of appraisal processes. In: Scherer, K.R., Schorr, A., Johnstone, T. (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, New York and Oxford, pp. 271–284.

Juslin, P.N., 2000. Cue utilization in communication of emotion in music performance: Relating performance to perception. J. Exp. Psychol.: Hum. Percep. Perform. 26, 1797–1813.

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.

Kaiser, L., 1962. Communication of affects by single vowels. Synthese 14, 300–319.

Kappas, A., 1997. His master's voice: Acoustic analysis of spontaneous vocalizations in an ongoing active coping task.

In: Thirty-Seventh Annual Meeting of the Society for Psychophysiological Research, Cape Cod.

Kappas, A., Hess, U., Scherer, K.R., 1991. Voice and emotion. In: Rimé, B., Feldman, R.S. (Eds.), Fundamentals of Nonverbal Behavior. Cambridge University Press, Cambridge and New York, pp. 200–238.

Karlsson, I., Bänziger, T., Dankovicova, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K.R., 1998. Speaker verification with elicited speaking styles in the VeriVox project. In: Proc. RLA2C, Avignon, pp. 207–210.

Karlsson, I., Banziger, T., Dankovicova, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K.R., 2000. Speaker verification with elicited speaking styles in the VeriVox project. Speech Communication 31 (2–3), 121–129.

Kennedy, G., 1972. The Art of Rhetoric in the Roman World. 300 BC–AD 300. Princeton University Press, Princeton, NJ.

Kienast, M., Paeschke, A., Sendlmeier, W.F., 1999. Articulatory reduction in emotional speech. In: Proc. Eurospeech 99, Budapest, Vol. 1, pp. 117–120.

Klasmeyer, G., 1999. Akustische Korrelate des stimmlich emotionalen Ausdrucks in der Lautsprache. In: Wodarz, H.-W., Heike, G., Janota, P., Mangold, M. (Eds.), Forum Phoneticum, Vol. 67. Hector, Frankfurt am Main.

Klasmeyer, G., 2000. An automatic description tool for time-contours and long-term average voice features in large emotional speech databases. In: Proc. ISCA Workshop Speech and Emotion, Newcastle, Northern Ireland, pp. 66–71.

Klasmeyer, G., Meier, T., 1999. Rhythm in Emotional Speech, DAGA-Tag der ASA-Tagung 99 in Berlin, CD-Rom zur ASA-EAA-DEGA Gemeinschaftstagung Berlin 99.

Klasmeyer, G., Sendlmeier, W.F., 1997. The classification of different phonation types in emotional and neutral speech. Forensic Linguist. 4, 104–124.

Klasmeyer, G., Sendlmeier, W.F., 1999. Voice and emotional states. In: Kent, R., Ball, M. (Eds.), Voice Quality Measurement. Singular Publishing Group, San Diego, CA, pp. 339–359.

Knapp, M.L., 1972. Nonverbal Communication in Human Interaction. Holt Rinehart and Winston, New York.

Kotlyar, G.M., Morozov, V.P., 1976. Acoustical correlates of the emotional content of vocalized speech. J. Acoust. Acad. Sci. USSR 22, 208–211.

Kuroda, I., Fujiwara, O., Okamura, N., Utusuki, N., 1979. Method for determinating pilot stress through analysis of voice communication. Aviat. Space Environ. Md. 47, 528–533.

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. J. Acoust. Soc. Amer. 78, 435–444.

Laver, J., 1980. The Phonetic Description of Voice Quality. Cambridge University Press, Cambridge.

Laver, J., 1991. The Gift of Speech. Edinburgh University Press, Edinburgh.

Lavner, Y., Gath, I., Rosenhouse, J., 2000. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. Speech Communication 30 (1), 9–26.

Levin, H., Lord, W., 1975. Speech pitch frequency as an emotional state indicator. IEEE Trans. Systems, Man Cybernet. 5, 259–273.

Lieberman, P., Michaels, S.B., 1962. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. J. Acoust. Soc. Amer. 34, 922–927.

Mahl, G.F., Schulze, G., 1964. Psychological research in the extralinguistic area. In: Sebeok, T.A., Hayes, A.S., Bateson, M.C. (Eds.), Approaches to Semiotics. Mouton, The Hague, pp. 51–124.

Markel, N.N., Bein, M.F., Phillis, J.A., 1973. The relation between words and tone of voice. Lang. Speech 16, 15–21.

Massaro, D.W., 2000. Multimodal emotion perception: Analogous to speech processes. In: Proc. ISCA Workshop on Speech and Emotion, 5–7 September 2000, Newcastle, Northern Ireland, pp. 114–121.

Morris, J.S., Scott, S.K., Dolan, R.J., 1999. Saying it with feeling: Neural responses to emotional vocalizations. Neuropsychologia 37 (10), 1155–1163.

Moses, P., 1954. The Voice of Neurosis. Grune and Stratton, New York.

Mozziconacci, S.J.L., 1998. Speech variability and emotion: Production and perception. Ph.D. Thesis, Technical University Eindhoven, Eindhoven.

Morton, J.B., Trehub, S.E., 2001. Children's understanding of emotion in speech. Child Develop. 72 (3), 834–843.

Murray, I.R., Arnott, J.L., 1993. Toward a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J. Acoust. Soc. Amer. 93, 1097–1108.

Murray, I.R., Arnott, J.L., Rohwer, E.A., 1996. Emotional stress in synthetic speech: Progress and future directions. Speech Communication 20 (1–2), 85–91.

Niwa, S., 1971. Changes of voice characteristics in urgent situations. Reports of the Aeromedical Laboratory, Japan Air Self-Defense Force, 11, pp. 246–251.

Ofuka, E., McKeown, J.D., Waterman, M.G., Roach, P.J., 2000. Prosodic cues for rated politeness in Japanese speech. Speech Communication 32 (3), 199–217.

Ostwald, P.F., 1964. Acoustic manifestations of emotional disturbances. In: Disorders of Communication, Vol. 42. Research Publications, pp. 450–465.

Paeschke, A., Kienast, M., Sendlmeier, W.F., 1999. F0-contours in emotional speech. In: Proc. ICPhS 99, San Francisco, Vol. 2, pp. 929–933.

Pear, T.H., 1931. Voice and Personality. Chapman and Hall, London.

Pell, M.D., 1998. Recognition of prosody following unilateral brain lesion: Influence of functional and structural attributes of prosodic contours. Neuropsychologia 36 (8), 701–715.

Pereira, C., Watson, C., 1998. Some acoustic characteristics of emotion. In: Proc. ICSLP 98, Sydney, Vol. 3, pp. 927–934.
Picard, R.W., 1997. Affective Computing. MIT Press, Cambridge, MA.
Pittenger, R.E., Hockett, C.F., Danehy, J.J., 1960. The First Five Minutes: A Sample of Microscopic Interview Analysis. Martineau, Ithaca, NY.
Plaikner, D., 1970. Die Veränderung der menschlichen Stimme unter dem Einfluss psychischer Belastung. [Voice changes produced by psychic load]. Ph.D. Thesis, University of Innsbruck, Austria.
Rank, E., Pirker, H., 1998. Generating emotional speech with a concatenative synthesizer. In: Proc. ICSLP 98, Sydney, Vol. 3, pp. 671–674.
Roessler, R., Lester, J.W., 1976. Voice predicts affect during psychotherapy. J. Nerv. Mental Disease, 163, 166–176.
Roessler, R., Lester, J.W., 1979. Vocal pattern in anxiety. In: Fann, W.E., Pokorny, A.D., Koracau, I., Williams, R.L. (Eds.), Phenomenology and Treatment of Anxiety. Spectrum, New York.
Roseman, I., Smith, C., 2001. Appraisal theory: Overview, assumptions, varieties, controversies. In: Scherer, K.R., Schorr, A., Johnstone, T. (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, New York and Oxford, pp. 3–19.
Russell, J.A., 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychol. Bull. 115, 102–141.
Sangsue, J., Siegwart, H., Grosjean, M., Cosnier, J., Cornut, J., Scherer, K.R., 1997. Developpement d'un questionnaire d'évaluation subjective de la qualité de la voix et de la parole, QEV. Geneva Stud. Emot. Comm. 11 (1), Available from: <http://www.unige.ch/fapse/emotion/genstudies/genstudies.html>.
Scherer, K.R., 1974. Voice quality analysis of American and German speakers. J. Psycholing. Res. 3, 281–290.
Scherer, K.R., 1978. Personality inference from voice quality: The loud voice of extroversion. Europ. J. Soc. Psychol. 8, 467–487.
Scherer, K.R., 1979. Non-linguistic indicators of emotion and psychopathology. In: Izard, C.E. (Ed.), Emotions in Personality and Psychopathology. Plenum Press, New York, pp. 495–529.
Scherer, K.R., 1982a. Methods of research on vocal communication: Paradigms and parameters. In: Scherer, K.R., Ekman, P. (Eds.), Handbook of Methods in Nonverbal Behavior Research. Cambridge University Press, Cambridge, UK, pp. 136–198.
Scherer, K.R. (Ed.), 1982b. Vokale Kommunikation: Nonverbale Aspekte des Sprachverhaltens. [Vocal Communication: Nonverbal Aspects of Speech]. Beltz, Weinheim.
Scherer, K.R., 1984. On the nature and function of emotion: A component process approach. In: Scherer, K.R., Ekman, P. (Eds.), Approaches to Emotion. Erlbaum, Hillsdale, NJ, pp. 293–318.
Scherer, K.R., 1985. Vocal affect signalling: A comparative approach. In: Rosenblatt, J., Beer, C., Busnel, M., Slater, P.J.B. (Eds.), Advances in the Study of Behavior. Academic Press, New York, pp. 189–244.
Scherer, K.R., 1986. Vocal affect expression: A review and a model for future research. Psychol. Bull. 99 (2), 143–165.
Scherer, K.R., 1989. Vocal correlates of emotion. In: Wagner, H., Manstead, A. (Eds.), Handbook of Psychophysiology: Emotion and Social Behavior. Wiley, London, pp. 165–197.
Scherer, K.R., 1992a. What does facial expression express? In: Strongman, K. (Ed.), International Review of Studies on Emotion, Vol. 2. Wiley, Chichester, pp. 139–165.
Scherer, K.R., 1992b. On social representations of emotional experience: Stereotypes, prototypes, or archetypes. In: von Cranach, M., Doise, W., Mugny, G. (Eds.), Social Representations and the Social Bases of Knowledge. Huber, Bern, pp. 30–36.
Scherer, K.R., 1994a. Affect bursts. In: van Goozen, S.H.M., van de Poll, N.E., Sergeant, J.A. (Eds.), Emotions: Essays on Emotion Theory. Erlbaum, Hillsdale, NJ, pp. 161–196.
Scherer, K.R., 1994b. Toward a concept of "modal emotions". In: Ekman, P., Davidson, R.J. (Eds.), The Nature of Emotion: Fundamental Questions. Oxford University Press, New York/Oxford, pp. 25–31.
Scherer, K.R., 1999a. Appraisal theories. In: Dalgleish, T., Power, M. (Eds.), Handbook of Cognition and Emotion. Wiley, Chichester, pp. 637–663.
Scherer, K.R., 1999b. Universality of emotional expression. In: Levinson, D., Ponzetti, J., Jorgenson, P. (Eds.), Encyclopedia of Human Emotions. Macmillan, New York, pp. 669–674.
Scherer, K.R., 2000a. Psychological models of emotion. In: Borod, J. (Ed.), The Neuropsychology of Emotion. Oxford University Press, Oxford/New York, pp. 137–162.
Scherer, K.R., 2000b. Emotional expression: A royal road for the study of behavior control. In: Perrig, W., Grob, A. (Eds.), Control of Human Behavior, Mental Processes, and Consciousness. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 227–244.
Scherer, K.R., Oshinsky, J.S., 1977. Cue utilization in emotion attribution from auditory stimuli. Motiv. Emot. 1, 331–346.
Scherer, K.R., Koivumaki, J., Rosenthal, R., 1972a. Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech. J. Psycholinguist. Res. 1, 269–285.
Scherer, K.R., Rosenthal, R., Koivumaki, J., 1972b. Mediating interpersonal expectancies via vocal cues: Different speech intensity as a means of social influence. Europ. J. Soc. Psychol. 2, 163–176.
Scherer, K.R., London, H., Wolf, J., 1973. The voice of confidence: Paralinguistic cues and audience evaluation. J. Res. Person. 7, 31–44.
Scherer, K.R., Ladd, D.R., Silverman, K.E.A., 1984. Vocal cues to speaker affect: Testing two models. J. Acoust. Soc. Amer. 76, 1346–1356.
Scherer, K.R., Feldstein, S., Bond, R.N., Rosenthal, R., 1985. Vocal cues to deception: A comparative channel approach. J. Psycholinguist. Res. 14, 409–425.

Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck, T., 1991. Vocal cues in emotion encoding and decoding. Motiv. Emot. 15, 123–148.

Scherer, K.R., Johnstone, T., Sangsue, J., 1998. L'état émotionnel du locuteur: facteur négligé mais non négligeable pour la technologie de la parole. Actes des XXIIèmes Journées d'Etudes sur la Parole, Martigny.

Scherer, K.R., Johnstone, T., Klasmeyer, G., Bänziger, T., 2000. Can automatic speaker verification be improved by training the algorithms on emotional speech? In: Proc. ICSLP2000, Beijing, China.

Scherer, K.R., Banse, R., Wallbott, H.G., 2001a. Emotion inferences from vocal expression correlate across languages and cultures. J. Cross-Cult. Psychol. 32 (1), 76–92.

Scherer, K.R., Schorr, A., Johnstone, T. (Eds.), 2001b. Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press, New York and Oxford.

Scherer, K.R., Johnstone, T., Klasmeyer, G., in press. Vocal expression of emotion. In: Davidson, R.J., Goldsmith, H., Scherer, K.R. (Eds.), Handbook of the affective sciences, Oxford University Press, New York and Oxford.

Scherer, T., 2000. Stimme, Emotion und Psyche: Untersuchungen zur emotionalen Qualität der Stimme [Voice, emotion, and psyche: Studies on the emotional quality of voice]. Ph.D. Thesis, University of Marburg, Germany.

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

Sedlacek, K., Sychra, A., 1963. Die Melodie als Faktor des emotionellen Ausdrucks [Speech melody as a factor of emotional expression]. Folia Phonatrica 15, 89–98.

Shipley-Brown, F., Dingwall, W., Berlin, C., Yeni-Komshian, G., Gordon-Salant, S., 1988. Hemispheric processing of affective and linguistic intonation contours in normal subjects. Brain Lang. 33, 16–26.

Siegman, A.W., 1987a. The telltale voice: Nonverbal messages of verbal communication. In: Siegman, A.W., Feldstein, S. (Eds.), Nonverbal Behavior and Communication, second ed.. Erlbaum, Hillsdale, NJ, pp. 351–433.

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

Simonov, P.V., Frolov, M.V., 1973. Utilization of human voice for estimation of man's emotional stress and state of attention. Aerospace Md. 44, 256–258.

Skinner, E.R., 1935. A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness. Speech Monogr. 2, 81–137.

Sobin, C., Alpert, M., 1999. Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. J. Psycholinguist. Res. 28 (4), 347–365.

Starkweather, J.A., 1956. Content-free speech as a source of information about the speaker. J. Pers. Soc. Psychol. 35, 345–350.

Steinhauer, K., Alter, K., Friederici, A., 1999. Brain potentials indicate immediate use of prosodic cues in natural speech processing. Nature Neurosci. 2, 191–196.

Stemmler, G., Heldmann, M., Pauls, C.A., Scherer, T., 2001. Constraints for emotion specificity in fear and anger: The context counts. Psychophysiology 38 (2), 275–291.

Sulc, J., 1977. To the problem of emotional changes in the human voice. Activitas Nervosa Superior 19, 215–216.

Tischer, B., 1993. Die vokale Kommunikation von Gefühlen [The Vocal Communication of Emotions]. Psychologie-Verlags-Union, Weinheim.

Titze, I., 1992. Acoustic interpretation of the voice range profile (Phonetogram). J. Speech Hear. Res. 35, 21–34.

Tolkmitt, F.J., Scherer, K.R., 1986. Effects of experimentally induced stress on vocal parameters. J. Exp. Psychol.: Hum. Percept. Perform. 12, 302–313.

Tolkmitt, F., Bergmann, G., Goldbeck, Th., Scherer, K.R., 1988. Experimental studies on vocal communication. In: Scherer, K.R. (Ed.), Facets of Emotion: Recent Research. Erlbaum, Hillsdale, NJ, pp. 119–138.

Tomkins, S.S., 1962. Affect, Imagery, Consciousness. The Positive Affects, Vol. 1. Springer, New York.

Trager, G.L., 1958. Paralanguage: A first approximation. Stud. Linguist. 13, 1–12.

Tusing, K.J., Dillard, J.P., 2000. The sounds of dominance: Vocal precursors of perceived dominance during interpersonal influence. Hum. Comm. Res. 26 (1), 148–171.

Utsuki, N., Okamura, N., 1976. Relationship between emotional state and fundamental frequency of speech. Reports of the Aeromedical Laboratory, Japan Air Self-Defense Force, 16, 179–188.

van Bezooijen, R., 1984. The Characteristics and Recognizability of Vocal Expression of Emotions. Foris, Dordrecht, The Netherlands.

van Bezooijen, R., Otto, S., Heenan, T.A., 1983. Recognition of vocal expressions of emotions: A three-nation study to identify universal characteristics. J. Cross-Cult. Psychol. 14, 387–406.

Wagner, H.L., 1993. On measuring performance in category judgment studies on nonverbal behavior. J. Nonverb. Behav. 17 (1), 3–28.

Wallbott, H.G., Scherer, K.R., 1986. Cues and channels in emotion recognition. J. Pers. Soc. Psychol. 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. J. Pers. Soc. Psychol. 78 (1), 105–119.

Westermann, R., Spies, K., Stahl, G., Hesse, F.W., 1996. Relative effectiveness and validity of mood induction procedures: A meta-analysis. Europ. J. Soc. Psychol. 26 (4), 557–580.

Whiteside, S.P., 1999. Acoustic characteristics of vocal emotions simulated by actors. Perc. Mot. Skills 89 (3, Part 2), 1195–1208.

Williams, C.E., Stevens, K.N., 1969. On determinating the emotional state of pilots during flight: An exploratory study. Aerospace Md. 40, 1369–1372.

Williams, C.E., Stevens, K.N., 1972. Emotions and speech: Some acoustical correlates. J. Acoust. Soc. Amer. 52, 1238–1250.

Wundt, W., 1874/1905. Grundzüge der physiologischen Psychologie, [Fundamentals of physiological psychology, orig. pub. 1874] fifth ed. Engelmann, Leipzig.

Zuberbier, E., 1957. Zur Schreib- und Sprechmotorik der Depressiven. [On the motor aspect of speaking and writing in depressives]. Z. Psychother. Medizin. Psychol. 7, 239–249.

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

Zwicker, E., 1982. Psychoacoustics. Springer, New York.