*Journal of Phonetics* (2000) **28,** 335–342
doi.10.006/jpho.2000.0120
Available online at http://www.idealibrary.com on IDEAL®

# Theoretical and methodological development in the study of pathological voice quality

**Bruce R. Gerratt and Jody Kreiman**

*Bureau of Glottal Affairs, Division of Head/Neck Surgery,*
*UCLA School of Medicine, U.S.A.*

*Accepted 3rd August 2000*

This special issue focuses on the development of theory in the study of communication disorders and its contribution to a broad theoretical framework for describing speech and language processes. In the case of the study of pathological voice quality, little attention has been paid to theoretical development or to issues of measurement validity. Instead, researchers have generally appropriated scales and categories traditionally used in the study of normal vocal quality. In this paper, we describe some theoretical and methodological problems that we think plague the study of pathological vocal quality as a result of the application of methodology first developed to characterize normal behavior and then applied to clinical phenomena. We argue that application of methods from the study of normal phenomena has actually hindered rather than advanced this clinical research area, and describe an alternative approach that addresses issues of the validity and utility of acoustic and perceptual measurements. Because pathological vocal qualities encompass the full range of possible vocal sounds, more systematic, theoretically motivated studies of pathological voice quality can serve as a fertile source of inspiration (and data) for studies of normal processes in voice production and perception. © 2000 Academic Press

The formal study of voice quality has a long intellectual history, and dates at least from Roman times (see, e.g., Gray, 1943; Laver, 1981, for reviews). However, the modern study of voice has become fragmented, with different aspects or functions of voice becoming the province of different academic disciplines. For example, law enforcement and security officials have an interest in determining the accuracy of speaker identifications (e.g., Hammersley & Read, 1996). Musicians examine vocal register and artistic expression in singing voices (Sundberg, 1987; Scherer, 1995). Linguists are interested in how changes in voice quality can signal changes in meaning (Ladefoged, Maddieson & Jackson, 1988). Psychologists are concerned with the perception of emotion and other personal information encoded in voice (Murray & Arnott, 1993), and in how personal quality can influence spoken word recognition (Bradlow & Pisoni, 1999). Engineers seek to develop compression and transmission algorithms that preserve voice quality (Reynolds,

Address correspondence to B. R. Gerratt, 31-24 Rehab Center, Los Angeles, CA 90095-1794, U.S.A. E-mail: bgerratt@ucla.edu

0095–4470/00/030335 + 08 $35.00/0

Zissman, Quatieri, O'Leary & Carlson, 1995). Finally, speech pathologists and laryngologists are interested in how pathological voice quality conveys clinical information (Kent & Ball, 2000). Although much could be learned from a unified, cross-disciplinary effort to describe and understand these different facets of voice quality perception, the literature demonstrates little interdisciplinary fertilization of ideas. The one element that does link these diverse disciplines is the common use of verbal category labels like "breathy", "rough", and "hoarse" to describe various vocal attributes.

This special issue examines how theories in communication disorders are developed and how results from research on clinical populations may contribute to broad theoretical frameworks for the description of speech and language. In the case of voice, researchers have generally appropriated scales and categories derived from those traditionally used in the study of normal vocal quality and applied them to pathological voices. However, the development and use of verbal categories for normal voice quality has received little substantial theoretical discussion, resulting in an impoverished framework for their understanding. This problem is unfortunately replicated in the understanding of pathological voice quality.

In this paper, we describe some theoretical and methodological problems that may underlie this research area, primarily taking an inverse approach relative to the charge of this issue. In fact, we present a cautionary tale of how methodology first developed to characterize normal behavior and then applied to clinical phenomena without much questioning of validity may poorly illuminate both. We will also describe a theoretical foundation for the experimental study of voice quality and alternative approaches to examining the stimulus and response relationships in voice quality perception. Finally, we will discuss how methodology and information derived from the study of pathological voice quality may contribute to a basic understanding of perception of the voice quality of normal speakers.

The perception of pathological voice quality is of primary importance in the evaluation and treatment of patients with voice disorders. These patients usually seek clinical care because of their own perception of a voice quality deviation, and most often they judge the success of treatment for the voice problem by improvement in their voice quality. A clinician may judge success by documenting changes in laryngeal anatomy or physiology, but in general, patients are more concerned with how their voices sound after treatment. Because of this great interest, most research on voice quality has been concerned with pathological, rather than normal, voices.

The overall quality of a sound is defined as everything in the acoustic signal that is not pitch or loudness, including the spectral envelope and its changes over time, any noise components, fluctuations in F0, and other factors (ANSI, 1960; Plomp, 1976). Thus, quality is by definition multidimensional. In addition, pathological voices are characterized by great acoustic variability, within and across both utterances and speakers. Thus, we may expect that their quality will be associated with particularly complex psychophysical functions that will probably be difficult to specify. Nevertheless, investigations of voice quality have almost always assessed quality with small sets of unidimensional scales, without examining how the importance of a scale may vary depending on the total perceptual context provided by overall vocal quality (see, e.g., Kreiman & Gerratt, 2000, for review). Whether individual rating scales are treated as individual features of voices that may be studied independently, or as subordinate aspects of some other superordinate quality (for example, Fairbanks's (1940) view of breathiness and roughness as components of hoarseness), the use of such independent scales to measure complex, varying multidimensional structures raises significant issues of rating instrument validity.

As a preliminary investigation of the validity of such rating scale approaches as models of overall vocal quality, we asked listeners to judge the dissimilarity of a very large set of pairs of pathological voices (Kreiman & Gerratt, 1996). We analyzed listeners' judgments with multidimensional scaling (e.g., Schiffman, Reynolds & Young, 1981), a statistical technique that let us define the dimensions that underlie perceived vocal quality, without requiring *a priori* assumptions about what those dimensions might be. Analyses showed that voices did not disperse in a perceptual space along continuous scale-like dimensions, but instead clustered together along each dimension to form groups that lacked subjective unifying percepts (Fig. 1). Neither clusters nor dimensions corresponded to traditional qualities like breathiness or roughness, suggesting that such scales are not good indices of what listeners hear, and that linear scales in general may not adequately model listeners' perceptual processes. Further, different voices clustered together for each individual listener's data, so that two voices were never consistently treated as similar in the perceptual spaces. These results led us to the further interpretation that listeners do not share a common set of perceptual features for pathological voices. They also provided some indirect evidence against the validity of traditional voice quality rating scales, because we could find no evidence that listeners judged vocal similarity along dimensions that corresponded (even approximately) to these scales.

Establishing or disproving the validity of perceptual measures is always challenging. It is particularly difficult to find previously-validated standard measurements to which a perceptual measure can be compared. This is certainly true in the study of voice quality. At least in languages like English, voice qualities do not contrast in the way that phonemes do, so there is no basis for defining what the "correct' quality judgment should be for a given stimulus. Thus, accuracy of judgments cannot serve as a criterion for

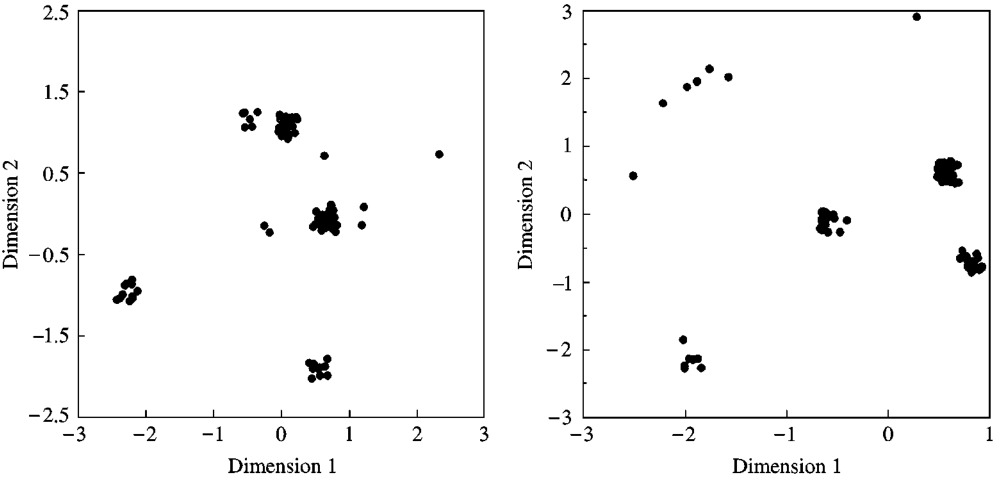

**Figure 1.** Sample perceptual spaces for judgments of the dissimilarity of pairs of the voices of 80 male speakers with vocal pathology. Data have been jittered slightly to show overlapping points. Within a space, voices that are close together were perceived as similar by that listener; but across listeners different voices clustered together, so no two voices consistently emerged as "similar". Previously unpublished data from Kreiman & Gerratt (1996).

establishing scale validity. However, patterns of agreement and disagreement among listeners can provide such evidence, because by definition, unreliable scales cannot be valid (e.g., Suen & Ary, 1989). To determine whether traditional rating scales for voice were reliable enough to be potentially valid, we examined the likelihood that any two raters in a listening task would agree in their ratings of a single voice (Kreiman & Gerratt, 1998). Results demonstrated that listeners agreed very poorly in the midrange of the scales for breathiness and roughness (Fig. 2). For voices with mean ratings between 2.5 and 5.5 on a 7-point scale, the likelihood that two raters would agree exactly averaged 0.21, while the likelihood of agreement within 1 scale value averaged 0.57. Although these values exceed chance, they are quite low; in fact, these disagreements are so large that we question whether mean ratings in the midrange of a scale represent the extent to which a voice possesses a quality, rather than simply indicating that listeners disagreed. Further, in all our perceptual data we did not find any voices that listeners agreed were moderately deviant in quality. These data provided further evidence that traditional unidimensional scales for voice quality are not adequate measures of what listeners hear when they listen to pathological voices.

Although listener agreement is quite low in the midrange of traditional scales for voice quality, statistics measuring overall reliability are actually quite high for these data. Application of statistics like Cronbach's alpha (Cronbach, 1951) and intraclass correlation coefficients (ICCs; e.g., Shrout & Fleiss, 1979) to data like these may result in serious misinterpretations and acceptance of essentially unreliable data. For example, Cronbach's alpha and the ICC both equal 0.99 for the data set shown in Fig. 2, despite the low overall likelihood of agreement among pairs of listeners. Traditionally used reliability statistics thus appear to mask important differences in levels of listener agreement for different voices, and may lead to incorrect conclusions about the reliability of the data. These problems may derive from generalization of statistics that were originally designed

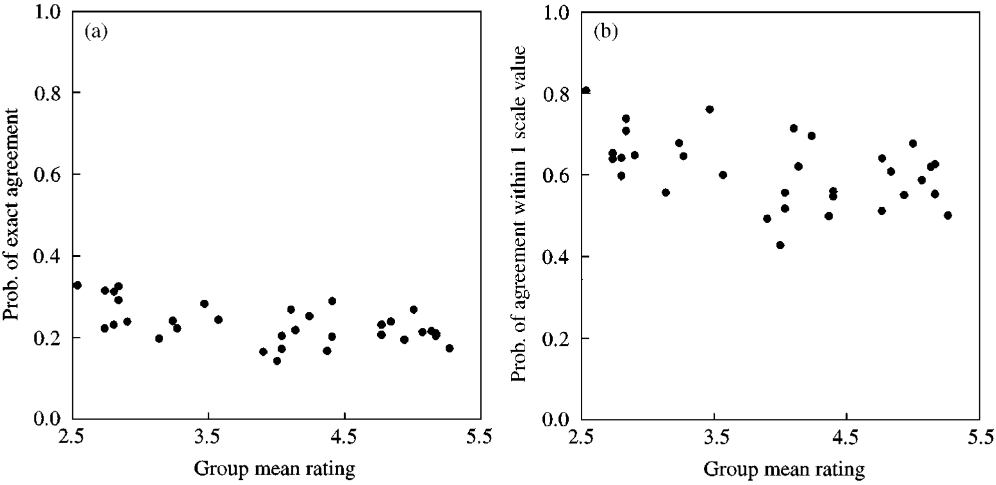

**Figure 2.** For each voice in a set, the probability that two raters agreed in their ratings of that voice, *vs.* the overall mean rating for that voice. Raters judged the roughness of the stimuli on 7-point equal-appearing interval scales; only voices with mean ratings between 2.5 and 5.5 are shown. (a) The probability that two raters would agree exactly in their ratings of a given voice. (b) The probability that two raters would agree within one scale value in their ratings. Data from Kreiman *et al.* (1993); figure adapted from Kreiman & Gerratt (1998).

for use in construction of psychological tests. In adapting this statistical framework, researchers substituted listeners for test items, and replaced test subjects with voices. Reliability as measured by Cronbach's alpha or the ICC implies that another sample of listeners would produce the same mean ratings for the same test voices, but does not necessarily inform us how the subjects would agree in their ratings of a new set of voices. However, a primary interest in using a perceptual measurement is knowing the extent of agreement across individuals in their responses to a stimulus. Large disagreement among listeners probably indicates that subjects use the measurement system in different ways from each other. The reliability statistics that are derived from psychological test construction do not answer important questions such as this, because they cannot represent patterns of agreement among raters (like those shown in Fig. 2), and they cannot indicate overall agreement for specific voice samples.

An additional research goal borrowed from studies of the perception of nonpathologic signals is the search for acoustic correlates of vocal quality. Many researchers have pursued acoustic measures in an apparent attempt to find objective measures as substitutes for subjective perceptual judgments of voice quality. For example, since 1961 more than 200 papers have appeared describing and applying measures of vocal perturbation alone (see Buder, 2000, for review). Possibly, this search for objective correlates of voice quality stems from successful studies of the relatively straightforward relationship between frequency and judgments of pitch, or intensity and loudness. However, as discussed above, the study of vocal quality requires multidimensional stimuli, rather than the unidimensional ones often used in pitch and loudness experiments (e.g., Moore, 1998; Scharf, 1998). Further, because voice quality is a concept describing the perceptual response to a particular acoustic stimulus, we cannot know the perceptual importance of particular aspects of the acoustic signal without valid measures of that perceptual response. Thus, acoustic measures that purport to quantify vocal quality must derive their validity as measures of voice quality from their relationship to perceptual measures. We have shown in our work that traditional rating protocols for assessing vocal quality consistently fail to adequately represent listeners' perceptions, in part because listeners are unable to selectively attend to individual elements or dimensions of quality in such multidimensional stimuli (Kreiman & Gerratt, 2000). Thus, because acoustic measures derive their validity from their correlation to perceptual scales of questionable validity, we argue that the validity of these acoustic measures as measures of vocal quality must also be suspect.

Another difficulty with the search for acoustic correlates of vocal quality derives from the statistical truth that correlation does not imply causality. Classic psychophysical studies of pitch and loudness were primarily inferential, in that acoustic signals were systematically manipulated and changes in perception were then assessed (see, e.g., Yost & Nielsen, 1977, for review). In contrast, most studies of pathologic voice quality perception examine the correlation of an acoustic measure with a rating scale variable. However, simply knowing the association of an acoustic variable with a perceptual one does not necessarily illuminate its contribution to perceived quality. Even if an acoustic variable were important to a listener's judgment of vocal quality, the nature of that contribution would not be revealed by a correlation coefficient. Further, given the great variability in perceptual strategies and habits that individual listeners demonstrate in their use of traditional rating scales, the overall correlation between acoustic and perceptual variables, averaged across samples of listeners and voices, fails to provide useful insight into the perceptual process.

To recapitulate, classification systems for pathological voices have been derived from venerable studies of normal phonation and quality. Unidimensional verbal rating scales that were inherited from this tradition typically have been applied to study pathological voice quality, but questions of rating scale validity have not been adequately addressed, and recent studies suggest poor scale validity. Statistical models have been imported from other disciplines, and answer questions about mean ratings rather than about the likelihood of raters agreeing in their judgments of a given voice. The correlational approach of modeling voice quality using acoustic measures may ultimately derive from studies of stimuli that vary in only one dimension, such as frequency or intensity. This historical approach to the study of pathological voice quality, derived from other disciplines or from the study of normal functions, has failed to propel our knowledge very far, and may represent a theoretical dead-end.

Clearly, other approaches are needed to understand how listeners exploit the many acoustic elements of the voice signal in their perception of voice quality. An alternative to traditional methods is to develop theories of stimulus-response relationships in voice quality perception, which are then systematically, experimentally evaluated using speech synthesis. In this approach, synthesis parameters are first selected to adequately represent voice features. To determine these synthesis variables, candidate parameters are used to synthesize vowels that match (as closely as possible) natural vowels produced by speakers with vocal pathology. (Selection of parameters having physiological or acoustic relevance will enhance the theoretical appeal of this experimental approach.) The synthesized tokens are then presented to listeners who use similarity judgments to compare vowels constructed with different synthesis parameters. The results of such studies provide confirmatory information regarding the perceptual importance of the acoustic variables that were used in the synthesis, in contrast to the traditional correlational approach. This comparative method not only can help determine a parsimonious list of synthesis parameters necessary for adequate modeling of pathological voice samples, but also has the potential to reveal the relative importance of the stimulus attributes that actually underlie listeners' voice quality perceptions.

Following this determination of relevant synthesis parameters is the use of analysis resynthesis, in which listeners themselves are asked to change speech synthesizer parameters to create an acceptable auditory match to a pathological voice stimulus. In this method, which integrates acoustic and perceptual approaches to voice quality, quality can be modeled as a whole without the need to isolate specific aspects or dimensions. By measuring quality as the set of synthesis parameters that will generate a given percept, issues of the validity and utility of acoustic measurements are simultaneously addressed.

Preliminary efforts to apply such an approach are promising. In a recent study (Gerratt & Kreiman, 2000), synthetic copies were made of 12 natural samples of the vowel /a/, spoken by subjects with different vocal pathologies. Listeners were asked to adjust the signal-to-noise ratio for the synthetic voices, so that they matched the perceived levels of spectral noise present in the natural target voice samples. Listeners were also asked to rate the perceived noisiness of the natural stimuli, using a traditional voice rating scale. For nine of the 12 voices, variance in the synthesizer settings the listeners chose was significantly (and substantially) less than that in the traditional scalar ratings of noisiness, although both scales were physically identical and ranged from 0 to 100. Variances in responses for the two tasks did not differ significantly for the remaining three voices. However, subsequent analyses indicated that the majority of apparent "listener disagreements" in the analysis–resynthesis task in fact resulted from the perceptually-arbitrary calibration of the synthesizer scale, and did not reflect actual listener disagreement or unreliability.

These results indicate that listeners are able to agree in their perceptual evaluations of pathological voice stimuli when they are given a suitable tool for reporting their perceptions. In particular, listeners agreed perfectly in midrange of the synthesizer scale. This *never* occurs for traditional rating protocols. Consequently, disagreements observed in the rating scale task in this study and in past studies likely reflected task-related difficulties, rather than actual differences in voice quality perception.

Finally, we note in closing that more systematic, theoretically motivated studies of pathological voice quality may prove to be fertile source of inspiration (and data) for studies of normal processes in voice production and perception, including speaker recognition, phonation type contrasts occurring in various languages, prosody, recognition of emotional states, and other segmental and nonsegmental aspects of speech. Pathological voice qualities encompass the extremes of human phonatory possibilities. Methods and models that are successful in describing, understanding and predicting such a rich spectrum of qualities will provide a solid foundation for inquiry into normally occurring vocal qualities. Interestingly, qualities that are often considered pathological (for example, vocal fry, diplophonia, bifurcations, breathiness) also occur in phonation by normal speakers, but intellectual tradition treats pathology as separate from normal processes. Consequently, models of normal processes may not reflect the full range of phonatory possibilities; and separate study of clinical and normal voice quality may lead to different acoustic and/or perceptual terms and models that account for what are essentially the same phenomena. This appears to be the case in the study of supraperiodic phonatory modes such as diplophonia and bifurcated (subharmonic series) phonation (Gerratt & Kreiman, in press), in which apparently similar phonatory phenomena are labeled differently by different academic disciplines. The advantages of developing a unified theoretical approach to describing acoustically and perceptually similar phenomena are obvious.

Developing a unified approach to the assessment of voice quality will provide a significant challenge to theory and technology in the future. It is not easy—either intellectually or practically—to abandon concepts, models and methods that have roots far back in western culture, but it appears that our knowledge of voice quality will not proceed much farther in the present track.

The authors acknowledge the support of the National Institute on Deafness and Other Communication Disorders by grant DC01797.

## References

ANSI (1960) *USA standard: acoustical terminology* (*S*1.1). New York: American National Standards Institute, Inc.

Bradlow, A. R. & Pisoni, D. B. (1999) Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors, *Journal of the Acoustical Society of America*, **106,** 2074–2085.

Buder, E. H. (2000) Acoustic analysis of voice quality: a tabulation of algorithms 1902–1990. In *Voice quality measurement* (R. D. Kent & M. J. Ball, editors), pp. 119–244. San Diego: Singular Publishing Group.

Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests, *Psychometrika*, **16,** 297–334.

Fairbanks, G. (1940). *Voice and articulation drillbook*. New York: Harper and Brothers.

Gerratt, B. R. & Kreiman, J. (2000) *Measuring vocal quality with speech synthesis*. Paper presented at the 139th meeting of the Acoustical Society of America, Atlanta, GA.

Gerratt, B. R. & Kreiman, J. (in press) Toward a taxonomy of nonmodal phonation. *Journal of Phonetics* (to appear).

Gray, G. W. (1943) The "voice qualities" in the history of elocution, *Quarterly Journal of Speech*, **29,** 475–480.

Hammersley, R. & Read, J. D. (1996) Voice identification by humans and computers. In *Psychological issues in eyewitness identification* (S. L. Sporer, R. S. Malpass & G. Koehnken, editor), pp. 117–152. Hillsdale, NJ: Lawrence Erlbaum.

Kent, R. D. & Ball, M. J. (2000) *Voice quality measurement*. San Diego, CA: Singular.

Kreiman, J. & Gerratt, B. R. (1996) The perceptual structure of pathologic voice quality, *Journal of the Acoustical Society of America*, **100,** 1787–1795.

Kreiman, J. & Gerratt, B. R. (1998) Validity of rating scale measures of voice quality, *Journal of the Acoustical Society of America* **104,** 1598–1608.

Kreiman, J. & Gerratt, B. R. (2000) Sources of listener disagreement in voice quality assessment, *Journal of the Acoustical Society of America*, **108,** 1867–1877.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A. & Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research*, **36,** 21–40.

Ladefoged P., Maddieson, I. & Jackson, M. (1988) Investigating phonation types in different languages. In *Vocal fold physiology*: *voice production*, *mechanisms and functions* (O. Fujimura, editor), pp. 297–317. New York: Raven Press.

Laver, J. (1981) The analysis of vocal quality: from the classical period to the 20th century. In *Toward a history of phonetics* (R. Asher & E. Henderson, editor), pp. 79–99. Edinburgh: Edinburgh University Press.

Moore, B. C. J. (1998) Frequency analysis and pitch perception. In *Handbook of acoustics* (M. J. Crocker, editor), pp. 1167–1180. New York: John Wiley and Sons.

Murray, I. R. & Arnott, J. L. (1993) Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *Journal of the Acoustical Society of America*, **93,** 1097–1108.

Plomp, R. (1976) *Aspects of tone sensation*. London: Academic Press.

Reynolds, D. A., Zissman, M. A., Quatieri, T. F., O'Leary, G. C. & Carlson, B. A. (1995) The effects of telephone transmission degradations on speaker recognition performance. *Proceedings of ICASSP*-95, pp. 329–332.

Scharf, B. (1998) Loudness. In *Handbook of acoustics* (M. J. Crocker, editor), pp. 1181–1196. New York: John Wiley and Sons.

Scherer, K. R. (1995) Expression of emotion in voice and music, *Journal of Voice*, **9,** 235–248.

Schiffman, S., Reynolds, M. & Young, F. (1981) *Introduction to multidimensional scaling*: *theory*, *method*, *and applications*. New York: Academic.

Shrout, P. & Fleiss, J. (1979) Intraclass correlations: uses in assessing rater reliability *Psychological Bulletin*, **86,** 420–428.

Suen, H. K. & Ary, D. (1989) *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Yost, W. A. & Nielsen, D. W. (1977) *Fundamentals of hearing*. New York: Holt, Rinehart, and Winston.