INTER-JUDGE VARIABILITY IN PERCEPTION OF VOICE QUALITIES

F. D. Minifie, J. Green, J. Smith, and D. Z. Huang
Speech and Hearing Sciences, University of Washington, Seattle, WA 98195

ABSTRACT

Perceptual ratings by 10 speech-language pathologists of breathiness, harshness and hoarseness in disordered, and synthetic vowels were compared. The synthetic vowels were varied in jitter, shimmer and glottal noise. Rather wide variations in inter-judge ratings were observed.

INTRODUCTION

Perceptual judgements of voice quality are one of the key indicators in the clinical diagnosis of voice disorders, yet little is known about the sources of inter-judge variability in classifications of breathiness, harshness and hoarseness. The purpose of the present study was to compare individual differences in the ratings of breathiness, harshness and hoarseness by using 10 trained speech-language pathologists as judges. By using a repeated measures experimental design, it was assumed that rater variability within and among judges can be specified for the corpus of disordered vowel stimuli, thereby providing a general indication of the stability of measurement within and across judges. However, it was further assumed that the sources of perceptual variability could be studied in greater detail via the use of synthetic stimuli wherein a single acoustic variable at a time can be manipulated.

METHODS AND PROCEDURES

Digitally recorded tokens of sustained vowel sounds produced by talkers with verified laryngeal pathologies and computer synthesized sustained vowels generated to simulate varying levels of acoustic perturbations assumed to result from vocal pathology were used as stimuli in this perceptual (speech-language pathologists) were used as judges and asked to perceptually rate the sustained vowel tokens.

Disordered Vowel Stimuli

The disordered voice samples used in the present investigation were selected from among those collected by Hirano at Kurume University in 1987. Two hundred and eight recordings of the vowel /ae/ produced by talkers presenting disordered voice samples were used as the original corpus of disordered vowel tokens for this study. It should be mentioned that not all of the vowel tokens were perceived as the vowel /ae/ by listeners raised in the United States. The perceptual variations in vowel quality are probably related to differences in vowel space for Japanese versus American English listeners. Nevertheless, these vowel stimuli were acoustically analyzed to obtain measures of jitter, shimmer and normalized glottal noise energy (NNE), acoustic perturbations of the vocal sound source assumed to be related to perceptual judgements of breathiness, harshness and hoarseness. The durations of the sustained vowels produced by the voice disordered Japanese talkers varied widely - from 200 ms to 3000 ms due largely to the extent of laryngeal pathology and consequent difficulty in phonation. Some of the disordered vowel tokens were unanalyzable, due to unacceptable levels of noise from nonvocalic sources (environmental noises), or from the presence of silence gaps within the sustained vowel. The presence of the silence gaps in the disordered voice samples served to reduce the number of measurable contiguous pitch periods to an unacceptably small number. Hence, unreliable perturbation data on these voice samples rendered them unusuable for the present study. The remaining 158 sustained vowel tokens from the Kurume recordings served as the disordered vowel stimuli for the present study.

Synthetic Vowel Stimuli

A series of 462 synthetic vowel stimuli was created in which only one acoustic perturbation dimension at a time was varied. These stimuli were used to determine the specific perceptual consequences of jitter, shimmer and NNE on judgements of breathiness, harshness and hoarseness.

Three-formant synthesized vowels were generated using the *Dr. Speech Science for Windows* software developed by Tiger Electronics. The synthesized vowels generated were /i, I, E, ae, a, U, u/ . To account for gender differences, synthesized vowels were created using both male and female fundamental frequencies and formant frequency data [1]. The vowels averaged 400 msec in duration. Half of the vowels were produced using a flat intonation and half were produced using a rise-fall contour.

Each vowel was synthesized with a specified amount of jitter, shimmer or NNE. The amount of the acoustic variable imposed on a vowel was predetermined to fall within one of the five acceptable levels (ranges) for that acoustic dimension. Table 1 shows the values of jitter, shimmer and NNE at each of the five levels of variation used in this experiment. Level one represented minimal amounts of that acoustic dimension (a normal vowel production level), whereas level five represented a maximum amount of that acoustic variable (comparable to the level of variation observed in severely disordered vowel productions).

In order to isolate the perceptual effects of each acoustic variable, only one acoustic change was imposed on an interation of each vowel. Given that jitter, shimmer and NNE are not mutually exclusive, phenomena, the manipulation of one variable inadvertently has some effect on other variables. For instance, when synthesizing a vowel with extensive jitter, trace levels of NNE or Shimmer would also be detected. Although the contaminations were small, they will be addressed in subsequent papers.

Table 1: Acceptable Ranges of Jitter, Shimmer and NNE

	% Jitter
Level	Acceptable Range
1	(0.00-0.00)
2	(0.68-0.83)
3	(1.35-1.65)
4	(2.03-2.50)
5	(2.70-3.30)
	% Shimmer
Level	Acceptable Range
1	(0.00-0.00)
2	(3.60-4.40)
3	(7.20-8.80)
4	(10.80-13.20)
5	(14.40-17.60)
	% NNE
Level	Acceptable Range
1	(-18)-(-20)
2	(-14)-(-16)
3	(-10)-(-12)
4	(-6)-(-8)
5	(-2)-(-4)

Twenty percent of the vowel tokens rated in this experiment were duplicated and randomly inserted into the sequence of tokens to be rated. Repeated judgements on these tokens provided the data for evaluations of the intra-judge and inter-judge reliability in the perception of vowel quality.

Perceptual Judgements

Perceptual judgements of were made by 10 graduate students in speech-language pathology. They rated

breathiness, harshness and hoarseness on each sustained vowel token produced by the patients with laryngeal pathology, the disordered vowel samples, and for each synthetic vowel stimulus. The listening task took approximately 5 hours (3 listening sessions of about 100 minutes each). Subjects were required to complete the listening task in a sound proofed booth (IIC 1200 series). All stimuli were presented through loudspeakers at t comfortable loudness level.

Judges were required to judge only one voice quality at a time. For example, they listened to 208 disordered voice samples and 462 samples of synthetic vowels in approximately 100 minutes during which they rated only one of the voice qualities (breathiness, harshness or hoarseness). During that time the judges were provided a 1-minute break after every 52 samples, a 5 minute break after hearing the disordered voice samples, and a 5-minute break after 231 tokens of the synthetic vowels. These rest breaks were inserted to resist fatigue during the listening task.

Listener Training

All listeners in the perceptual judging task received training at the beginning of the first listening session. In addition, judges were provided with a definition of the voice quality to be rated at the beginning of each listening session. The definitions were adapted from Bassich and Ludlow [2]. Perceptual training included the presentation and rating of representative vowel tokens from the real voice samples and from the corpus of synthetic vowels. The tokens used during the training session were selected by an experienced voice clinician to be representative of the range of severity for each of the voice qualities (breathiness, harshness and hoarseness) to be rated. During training the students were familiarized with the rating form.

Perceptual Ratings

Vertically arranged continuous rating scales of 10 cm in length were used to obtain ratings of breathiness, harshness and hoarseness. The polar positions on the scales were identified with dimensional adjectives (e.g. "no breathiness" to "extremely breathy"). After the brief training period during which judges practiced rating 8 sample vowel tokens for the voice quality to be rated in that session, the blocks of 52 experimental vowel tokens were presented. Each token was presented two times in succession, separated by a 500 msec interstimulus interval, followed by a three second judging interval. Judges were instructed to indicate the severity of the vowel quality being rated by marking the rating sheet immediately following the second presentation of each vowel token. The rating sheets contained a separate vertical line for each vowel token. The top of the vertical scale was "most severe," and the bottom end of the scale was "normal" vowel quality. When judging tokens the subjects were instructed to make a horizontal mark across the vertical line at a point most representative of the severity of a specified voice quality.

RESULTS AND DISCUSSION

Three major findings of this study were obtained.

1. Listeners perceived greater changes in breathiness, harshness and hoarseness when rating the disordered voice samples collected from patients with pathological vocal mechanisms, than when rating the computer generated synthetic voice samples. This trend was particularly evident for the breathiness and hoarseness conditions, and less so for the harshness ratings. This trend may result from the relatively different levels of complexity of the stimulus from live vowels to synthetic vowel tokens.

2. Inter-judge reliability varied as a function of vowel and the voice quality being rated. As shown in Table 2, judges were more reliable in repeated estimates of breathiness than they were fore hoarseness and harshness. The perception of harshness appeared to be the most difficult for judges, based on the low reliability values reported.

Table 2
Reliability - Mean Perceptual Ratings

Token	N	Breathy	Harsh	Hoarse
/ae/	50	0.82	0.46	0.81
/a/	40	0.78	0.64	0.61
/e/	10	0.71	0.48	-0.24
/i/	50	0.70	0.67	0.62
/I/	10	0.36	0.31	0.00
/U/	10	0.92	0.38	0.70
/u/	40	0.72	0.69	0.59
All vowels		0.79	0.59	0.69

3. When the perceptual ratings on repeated syllables were compared within judges, there was considerable variation in the patterns of data obtained. Clearly, some of the judges were reliable when rating all of the voice qualities identified in this experiment. Other judges appeared to be considerably more reliable when rating a particular voice quality and not so reliable on another. (See Table 3).

Table 3
Intra-judge Reliability in the Perception of Voice Quality

Judge	Breathy	Harsh	Hoarse	All
1	0.33	0.91	-0.10	0.77
2	0.83			0.83
3	0.73	0.83	0.44	0.66
4	0.81	0.61	0.60	0.72
5	0.91	0.72	0.71	0.76
6	0.80	0.48	0.79	0.74
7	0.70	0.48		0.57
8	0.93	-0.05	0.69	0.55
9	0.87	0.74	0.77	0.79
10	0.34	0.50	0.76	0.51

Comparison of these data with transformed data from obtained three normalization procedures will be compared and the implications for reporting perceptual ratings of voice qualities will be discussed.

CONCLUSIONS

Based on the foregoing analysis of data obtained from perceptual ratings of voice quality, it would seem prudent to proceed cautiously when reporting group data or data from an individual judge when reporting perceptual ratings of vowel qualities. Individual judges appear to be more reliable when rating some voice qualities than others. These data are particularly troubling when attempting to rationalize inconsistencies in clinical ratings of voice qualities. Perhaps objective measurements of jitter, shimmer and glottal noise combined with perceptual judgements will provide increased stability of measurement.

REFERENCES

[1] Peterson, G. and Barney, H.L. (1952), "Control Methods used in the study of vowels," *J. Speech Hear. Res.*, vol. 9, 68-99.

[2] Bassich, J. and Ludlow, C. (1986), "The use of perceptual methods by new clinicians for assessing voice quality," *J. Speech Hear. Dis.*, vol. 51, 133.