Perceptual Assessment of Velopharyngeal Dysfunction by Otolaryngology Residents

Sydney C. Butts, MD¹, Alan Truong², Christina Forde², Dimitre G. Stefanov, PhD³, and Eileen Marrinan, MS, MPH⁴

Abstract

Objective. To assess the ability of otolaryngology residents to rate the hypernasal resonance of patients with velopharyngeal dysfunction. We hypothesize that experience (postgraduate year [PGY] level) and training will result in improved ratings of speech samples.

Study Design. Prospective cohort study.

Setting. Otolaryngology training programs at 2 academic medical centers.

Subjects and Methods. Thirty otolaryngology residents (PGY 1-5) were enrolled in the study. All residents rated 30 speech samples 2 separate times. Half the residents completed a training module between the rating exercises, with the other half serving as a control group. Percentage agreement with the expert rating of each speech sample and intrarater reliability were calculated for each resident. Analysis of covariance was used to model accuracy at session 2.

Results. The median percentage agreement at session 1 was 53.3% for all residents. At the second session, the median scores were 53.3% for the control group and 60% for the training group, but this difference was not statistically significant. Intrarater reliability was moderate for both groups. Residents were more accurate in their ratings of normal and severely hypernasal speech. There was no correlation between rating accuracy and PGY level. Score at session 1 positively correlated with score at session 2.

Conclusion. Perceptual training of otolaryngology residents has the potential to improve their ratings of hypernasal speech. Length of time in residency may not be best predictor of perceptual skill. Training modalities incorporating practice with hypernasal speech samples could improve rater skills and should be studied more extensively.

Keywords

hypernasal resonance, cleft palate, resident education, velopharyngeal dysfunction

Otolaryngologists care for patients with wide-ranging communication disorders. Speech disorders are among the most challenging among these. Patients with velopharyngeal dysfunction (VPD) have a speech disorder characterized by hypernasal resonance, auditory nasal emissions, and decreased intraoral pressure.¹-⁵ VPD is used as an umbrella term that describes incomplete closure of the velopharyngeal port from any cause, but most often it is seen in patients with a history of a cleft palate or other neuromuscular disorder resulting in weakness of palatal or pharyngeal muscles.²

During otolaryngology residency training, there may be variable degrees of exposure to patients with VPD. The perceptual speech evaluation is considered the gold standard in the diagnostic hierarchy for the evaluation of VPD and is generally performed by speech language pathologists (SLPs).⁶⁻⁸ Although the evaluation of many speech disorders calls for collaboration between the otolaryngologist and the SLP, otolaryngologists may be the first point of contact for the symptomatic patient. The responsibility for making a preliminary assessment of the speech problem may rest on the otolaryngologist.⁹,¹⁰

Speech disorders are assessed through auditory-perceptual judgments.¹¹,¹² An understanding of the strategies needed to acquire expertise in discriminating features of hypernasal speech has been the focus of a body of work within the speech pathology literature.²,¹¹,¹³⁻¹⁶ Specific guidelines for auditory training in speech disorders are not available for otolaryngologists, nor is there significant literature studying the effectiveness of otolaryngologists or otolaryngology residents as judges of hypernasal speech, although otolaryngologists have been

¹Division of Facial Plastic and Reconstructive Surgery, Department of Otolaryngology, State University of New York Downstate Medical Center, Brooklyn, New York, USA
²College of Medicine, State University of New York Downstate Medical Center, Brooklyn, New York, USA
³Scientific Computing Center, State University of New York Downstate Medical Center, Brooklyn, New York, USA
⁴Central New York Cleft and Craniofacial Center, State University of New York Upstate Medical Center, Syracuse, New York, USA

Corresponding Author:
Sydney C. Butts, MD, Facial Plastic and Reconstructive Surgery, Department of Otolaryngology, State University of New York Downstate Medical Center, 450 Clarkson Avenue, Brooklyn, NY 11203, USA.
Email: sydney.butts@downstate.edu
included in studies of ratings for voice disorders. The diagnostic abilities of otolaryngology residents in assessing speech as it relates to velopharyngeal function may also not be regularly assessed in the clinical setting or on training examinations. Perceptual training has been shown to result in an increase in intrarater reliability and accuracy among SLPs. A similar approach could be applied to the development of training protocols for otolaryngologists.

The goal of this study was to test the ability of otolaryngology residents to accurately grade the severity of hypernasal resonance in a series of speech samples from patients with VPD. We hypothesized that length of otolaryngology residency training is positively correlated with diagnostic accuracy and reliability of speech ratings.

Another goal of this study was to introduce otolaryngology residents to the auditory-perceptual methods used to diagnose speech symptoms associated with VPD. An interactive training tool was developed to test the hypothesis that training in concepts related to the perceptual assessment of patients with VPD would improve otolaryngology residents’ ability to detect resonance differences. Much of the recent emphasis on resident competency assessments and simulated training exercises has focused on operative skills sets. Standardized speech samples could be utilized for resident training exercises, self-assessment, and competency determinations by otolaryngology faculty.

**Methods**

Otolaryngology residents in postgraduate years 1 to 5 (PGY 1-5) were eligible to participate in this study. Resident participation was solicited from our home institutions (State University of New York [SUNY] Downstate and Upstate) and from other training programs in New York City. Recruitment proceeded after approval from program chairs, and participation was on a voluntary basis. Institutional Review Board approval was obtained from both SUNY Downstate and Upstate to conduct the study.

All residents who agreed to participate signed informed consent. Inclusion in the study required normal hearing levels as judged by an initial hearing screening. This included a question regarding symptoms of subjective hearing loss, followed by a tuning fork examination (512 Hz, Weber and Rinne testing). Listener judges (residents) completed a brief survey for assessment of their knowledge of cleft palate repair and level of training (see Appendix A at www.otojournal.org-supplemental). This information was used to describe the level of clinical and didactic exposure among residents, which could have an impact on their experience with speech disorders in patients with cleft palate. Fisher’s exact test was used to compare the prior experience of residents in the treatment and control groups.

Residents were placed into 1 of 2 study arms, determined by alternating group assignment to ensure equal distribution of resident numbers and PGY levels in each group. All residents were instructed during an introductory session by the lead investigator that they would listen to speech samples from patients with VPD, which should be rated per the severity of hypernasal resonance. Both groups rated these speech samples on 2 occasions about 1 to 2 weeks apart. At the second session, the order of the speech samples was changed, but the same set of listening tasks was presented. Residents in the training/treatment group completed an educational module on VPD between the 2 rating sessions. The listening tasks and the training module were all accessible via a web-based program (TalentLMS.com; Epignosis LLC, San Francisco, California). Each listener was provided with headphones and instructed to listen to the speech samples in a quiet environment.

Speech samples from adult and pediatric patients with VPD were obtained from a speech sample database compiled by David Kuehn, PhD, professor emeritus of the Department of Speech and Hearing Science, University of Illinois at Urbana-Champaign. These speech samples have been used for training purposes by the American Cleft Palate–Craniofacial Association, which also granted permission for their use. There were 24 total speech samples: 8 from pediatric patients, 8 from women, and 8 from men. The severity grade rating was determined by Dr Kuehn, and each sample was assigned a value on the severity scale that the residents used. Six of the speech samples were repeated in the series to assess intrarater reliability, for a total of 30 clips, each 15 to 20 seconds in length.

In general, the 3 main parameters that are evaluated in patients suspected of having VPD include resonance; nasal air emissions, turbulence, or rustle; and associated articulation disorders. The choice to focus on the dimension of hypernasal resonance in this study was based on several factors. Resonance is generally the most frequently studied speech dimension by researchers of VPD. In a review of 88 articles studying speech in patients with cleft palate, Lohmander and Olsson showed that hypernasality severity was the most commonly documented dimension in the studies reviewed (measured in 62 of 88 studies). In addition, listeners appear to be most sensitive to the dimension of hypernasal resonance as compared with other dimensions. Brunnegard et al found decreased sensitivity to nasal air emissions as compared with hypernasal resonance, especially for less experienced listeners. The degree of nasal emissions had to be severely problematic for inexperienced listeners to detect it.

In this study, hypernasality was rated on a 4-point scale (normal = 0, mild = 1, moderate = 2, severe = 3), as described by Henningsson and Kuehn. Mild hypernasality is appreciated primarily in certain vowels and may not be socially disturbing to the patient or family; moderate hypernasality is more obvious, as it is audible for most vowels and deemed socially unacceptable; finally, severe hypernasality represents a more pervasive manifestation that would usually prompt a recommendation for intervention by the patient and the clinician. In severe hypernasality, speech intelligibility is significantly diminished.

The educational module consisted of a 40-minute training session. We developed this module and included speech samples demonstrating VPD and anatomic and clinical
Intrarater agreement 66.7 (50.0-83.3) 66.7 (50.0-83.3) .85 66.7 (50.0-83.3) 66.7 (50.0-83.3) .78

Accuracy 53.3 (53.3-60.0) 53.3 (50.0-60.0) .60 60.0 (43.3-66.7) 53.3 (43.3-63.3) .68

Outcome Treatment Control

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Treatment</th>
<th>Control</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>53.3 (53.3-60.0)</td>
<td>53.3 (50.0-60.0)</td>
<td>.60</td>
</tr>
<tr>
<td>Intrarater agreement</td>
<td>66.7 (50.0-83.3)</td>
<td>66.7 (50.0-83.3)</td>
<td>.85</td>
</tr>
</tbody>
</table>

*Data expressed as median (25th-75th percentiles). P values based on Mann-Whitney test.*

correlations. Several points in the training required residents to answer questions about material that they just had covered. Residents completed a brief posttraining survey (Appendix B, available online only) and were given a $50 gift card upon completion of the study.

The primary outcome measured was the percentage agreement with the predetermined severity rating (expert). The Mann-Whitney test was used to compare distributions of scores between the groups. The Wilcoxon signed-rank test was used to compare the scores at session 1 and 2, separately for each group. The Spearman correlation coefficient was used to test for an association between the scores at sessions 1 and 2. Analysis of covariance (ANCOVA) was used to model the accuracy (percentage agreement with expert) at session 2. The covariate (the accuracy at session 1) was included in the model for increased precision in determining the effect of the treatment. The interaction between the covariate and group (control vs treatment) was investigated.

Intrarater reliability was calculated as the percentage agreement among the 6 repeated speech samples for each rater, per session. SAS 9.4 (SAS Institute, Cary, North Carolina) was used for the analysis. The power calculations indicated that a sample size of 34 residents per group would be needed to detect a difference with an effect size of 0.7 (Cohen’s d) between the treatment and the control groups after the training, assuming an alpha of 0.05 and power of 80%.

**Results**

Thirty residents completed the study between October 2013 and March 2015—PGY 1 (n = 8), PGY 2 (n = 6), PGY 3 (n = 6), PGY 4 (n = 6), and PGY 5 (n = 4)—with 15 in the treatment arm and 15 in the control arm. One resident that was screened for inclusion could not participate because of hearing loss. Residents participating in the study had good clinical and didactic exposure to cleft palate and VPD. The majority of residents surveyed had performed palatoplasty (69%) or a pharyngeal flap (37.9%). Many had also attended a multidisciplinary cleft team meeting (41.3%) or had didactic exposure to cleft reconstruction via lectures (82.7%). Review of the responses suggests similar levels of clinical and didactic exposure between the control and training groups at baseline. There was no difference between the surgical exposure or didactic experiences of the control and treatment groups. Among members of the control group, 71.4% and 66.7% of the residents who were assigned to the training arm had participated in palatoplasty cases (P = 1.0); 42.9% and 40% had attended a multidisciplinary cleft team meeting (P = 1.0); and 85.7% and 80% had attended a lecture on the management of cleft palate during residency (P = 1.0), respectively.

The distribution of scores did not differ significantly between the control and treatment groups at session 1 (median score = 53.3% for both groups, P = .60) or session 2 (median score = 53.3% vs 60%, P = .68; Table 1). There was no significant difference in the percentage agreement scores for the treatment (P = .79) or control (P = .84) group at the different time points. There was a significant moderate correlation (r = 0.44, P = .01) between the scores at sessions 1 and 2.

To increase the precision in determining the effect of the treatment on the scores at session 2, we used an ANCOVA model. Model residuals were inspected for outliers; one observation was excluded from the final ANCOVA model based on Cook’s distance statistics (D > 2; therefore, n = 29). The proportion of explained variability in the outcome (R²) was 0.14 and 0.34 for the models with and without the outlier, respectively. Predictors included score at session 1, the group (control vs treatment), and their interaction. Seniority (PGY ≤2 vs PGY ≥3, P = .20) was not found to be significant and not used in the final model. We found that the score at session 1 (P < .001) was a significant predictor, while the group assignment (P = .06) and the interaction between group and session 1 score (P = .055) were not.

The median intrarater agreement score for both groups at session 1 was 66.7%. At the second rating session, the median score was again 66.7% for both groups. The distribution did not differ between the treatment and control groups at session 1 (P = .85) or session 2 (P = .78; Table 1).

We analyzed resident ratings based on the severity of the speech sample. The median percentage agreement for normal speech samples was 100% for both the control and treatment groups at sessions 1 and 2 (Table 2). The median percentage agreement score for severely hypernasal speech samples for both groups at session 1 was 71%. The same median score (71%) was achieved by both groups for these speech samples at session 2. For speech samples falling into the mild to moderate categories, median scores dropped markedly, ranging from 38% to 51% agreement for mildly hypernasal samples and a median score of 42% for both groups in both rating sessions for moderately hypernasal speech (Table 2).
The primary aim of this study was to determine the ability of otolaryngology residents to accurately rate hypernasal resonance in patients with VPD. The speech and voice literature has focused on studying the perceptual accuracy and reliability of SLPs. Less is known about the performance of otolaryngologists as perceptual judges of speech and even less about the performance of residents. Hypernasal resonance is the primary symptom of VPD, and its detection is essential to assessing the success of velopharyngeal surgery. Hypernasal resonance is also a primary finding of an undetected palatal anomaly later in childhood. Preventing further diagnostic delays or, worse yet, ill-advised surgery in these patients (eg, adenoidectomy) requires training in the perceptual features of VPD.

Residency training provides an excellent model for understanding the impact of experience and educational exposure on the development of perceptual judgments of speech. Experience and training have been identified as 2 important factors that can influence the accuracy and reliability of perceptual judgments.

Many studies in the voice and cleft speech literature have tested the hypothesis that experience will result in more accurate and reliable judgments. Resident performance in our study did not correlate with seniority. Laczi et al compared the hypernasality ratings of first-year SLP graduate students and experienced SLPs. The mean scores of hypernasality by experienced judges were significantly lower than those of the inexperienced judges, while the intrarater reliability scores for both groups were similar. Keuning et al found that the reliability ratings for the experienced and inexperienced SLPs were similar. The impact of experience has been studied extensively in the voice literature, with similar findings regarding reliability and severity scores.

Very few studies of cleft palate speech assessment examined the effect of training on the accuracy and/or reliability scores of raters. Lee et al studied naïve listeners to determine how 1 of 3 training scenarios affected hypernasality scores and reliability ratings. They found significant differences in the hypernasality rating scores between (1) the groups receiving training that included practice with and without feedback and (2) the group that had only exposure to speech samples. Groups that had received more comprehensive training generally applied a wider range of scores to the samples they heard. All 3 training groups had comparable intrarater reliability scores. The authors attributed the lack of training benefit on reliability scores to the small sample size and the possibility that training length for some listeners might not have been long enough. The study design did not include a control group or pretraining rating exercises.

The voice literature has examined the role of training programs more extensively. While several studies do show benefits, the training programs described differ significantly; sample sizes may be too small to judge an effect; and some study designs did not include control groups or pretraining rating scores. Our study aimed to address some of these issues by having a designated control group and 2 rating sessions to determine the possible effect of repeat testing alone.

Our prediction model showed that the initial rating score was a significant predictor of performance at the second rating session. Several studies have supported a format that requires trainees to attain a certain level of proficiency to progress to advanced sections of the training module. Our prediction model supports this approach.

Multiple types of training formats have been utilized in prior studies with no consensus on the content, role of feedback, use of anchors/references, or length. Most protocols have employed training speech samples (range: 5-60) as examples of the perceptual feature to be tested in the rating session. The training module used in this study had more didactic content and included fewer training speech samples than other studies.

The results of this study show greater accuracy at the extremes of resonance. All listeners, despite their level of experience or exposure to training, have generally stable judgments of normal speech quality. Mild and moderately hypernasal speech sample scores ranged from 39% to 50%, compared with 96% to 100% accuracy for normal voices. Awan and Lawson found that the use of auditory anchors that listeners could play during the rating task resulted in a significant improvement in the accuracy of severity ratings of voice samples in the normal and mild categories.

The main limitation in this study was the small sample size, with a possibility that the nonsignificant difference between the 2 groups is a false-negative result. Recruitment of residents from other training programs in our region was challenging because of work-hour restrictions that residents face in our current training environment. Study design elements that capture a greater number of residents and an expansion of subjects (fellows and/or faculty) are worth exploration.

### Table 2. Rating Accuracy Compared with Resonance Severity

<table>
<thead>
<tr>
<th>Severity: Session</th>
<th>Control</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>100 (100-100)</td>
<td>100 (100-100)</td>
</tr>
<tr>
<td>2</td>
<td>100 (100-100)</td>
<td>100 (100-100)</td>
</tr>
<tr>
<td>Mild</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>38 (38-63)</td>
<td>38 (38-50)</td>
</tr>
<tr>
<td>2</td>
<td>50 (38-63)</td>
<td>50 (38-75)</td>
</tr>
<tr>
<td>Moderate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>42 (25-50)</td>
<td>42 (33-42)</td>
</tr>
<tr>
<td>2</td>
<td>42 (25-50)</td>
<td>42 (33-50)</td>
</tr>
<tr>
<td>Severe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>71 (71-86)</td>
<td>71 (71-86)</td>
</tr>
<tr>
<td>2</td>
<td>71 (57-86)</td>
<td>71 (57-86)</td>
</tr>
</tbody>
</table>

Data expressed as median (25th-75th percentiles).

### Discussion

Lee et al studied naïve listeners to determine how 1 of 3 training scenarios affected hypernasality scores and reliability ratings. They found significant differences in the hypernasality rating scores between (1) the groups receiving training that included practice with and without feedback and (2) the group that had only exposure to speech samples. Groups that had received more comprehensive training generally applied a wider range of scores to the samples they heard. All 3 training groups had comparable intrarater reliability scores. The authors attributed the lack of training benefit on reliability scores to the small sample size and the possibility that training length for some listeners might not have been long enough. The study design did not include a control group or pretraining rating exercises.

The voice literature has examined the role of training programs more extensively. While several studies do show benefits, the training programs described differ significantly; sample sizes may be too small to judge an effect; and some study designs did not include control groups or pretraining rating scores. Our study aimed to address some of these issues by having a designated control group and 2 rating sessions to determine the possible effect of repeat testing alone.

Our prediction model showed that the initial rating score was a significant predictor of performance at the second rating session. Several studies have supported a format that requires trainees to attain a certain level of proficiency to progress to advanced sections of the training module. Our prediction model supports this approach.

Multiple types of training formats have been utilized in prior studies with no consensus on the content, role of feedback, use of anchors/references, or length. Most protocols have employed training speech samples (range: 5-60) as examples of the perceptual feature to be tested in the rating session. The training module used in this study had more didactic content and included fewer training speech samples than other studies.

The results of this study show greater accuracy at the extremes of resonance. All listeners, despite their level of experience or exposure to training, have generally stable judgments of normal speech quality. Mild and moderately hypernasal speech sample scores ranged from 39% to 50%, compared with 96% to 100% accuracy for normal voices. Awan and Lawson found that the use of auditory anchors that listeners could play during the rating task resulted in a significant improvement in the accuracy of severity ratings of voice samples in the normal and mild categories.

The main limitation in this study was the small sample size, with a possibility that the nonsignificant difference between the 2 groups is a false-negative result. Recruitment of residents from other training programs in our region was challenging because of work-hour restrictions that residents face in our current training environment. Study design elements that capture a greater number of residents and an expansion of subjects (fellows and/or faculty) are worth exploration.
Conclusion and Future Directions

No correlation between years in residency and accuracy of hypernasal resonance ratings was demonstrated in this study. These results suggest that PGY level may not be the best predictor of a resident’s skill level where auditory discrimination is concerned. Improving opportunities for residents to work with SLPs could fine-tune their perceptual skills. Training should incorporate immediate feedback and should be designed to appreciate the greater challenges that mild and moderate hypernasal speech samples require for accurate discrimination. Speech rating exercises could be used to allow residents to assess their readiness to begin a specific training program. Future work could explore abilities of otolaryngologists to discriminate between hyper- and hypoponaslality and to detect other dimensions of VPD, including nasal emissions and articulation errors.

Acknowledgments

We thank Dr Richard Rosenfeld, chair of the Department of Otolaryngology, SUNY Downstate Medical Center, for supporting our recruitment of residents for this study. We also thank Dr Robert Kellman and Dr Amar Suryadevara, chair and program director, respectively, at SUNY Upstate Medical University, for their support of resident recruitment and participation in this study. We further acknowledge David Keulin, PhD, for generously agreeing to grant permission for use of his hypernasal speech samples, as well as the American Cleft Palate–Craniofacial Association leadership who approved of the inclusion of the speech samples in the study. Aaron Cormier in the Department of Biomedical Communications at SUNY Downstate Medical Center provided initial consultation and guidance for developing our online training tool.

Author Contributions

Sydney C. Butts, conception of study design and hypotheses, protocol generation, design of training tool, writing of manuscript, final review of manuscript; Alan Truong, recruitment of subjects, data acquisition, design of training tool, drafting, reviewing final manuscript; Christina Forde, recruitment of subjects, data acquisition, design of training tool, writing critical portions and reviewing final manuscript; Dimitre G. Stefanov, statistical methodology, data analysis, writing/editing methodological section of manuscript, final review; Eileen Marrinan, recruitment of subjects, development of study plan, acquisition of speech samples, drafting and editing of text, final review.

Disclosures

Competing interests: None.
Funding source: CORE health services research grant (2013).

Supplemental Material

Additional supporting information may be found at http://otojournal.org/supplemental.

References


