<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Training perceptual rating of hypernasality with co-existing speech disorders</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Other Contributor(s)</strong></td>
<td>University of Hong Kong.</td>
</tr>
<tr>
<td><strong>Author(s)</strong></td>
<td>Huynh, Yin-sau, Christine</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2007</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/55484">http://hdl.handle.net/10722/55484</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>The author retains all proprietary rights, such as patent rights and the right to use in future works.</td>
</tr>
</tbody>
</table>
Training Perceptual Rating of Hypernasality with Co-existing Speech Disorders

Huynh Yin Sau, Christine

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, June 30, 2007.
Abstract
Perceptual rating of hypernasality is commonly used in clinical settings and research purposes. However, this method is prone to variable reliability and can be influenced by a variety of factors, including the presence of co-existing speech disorders and listeners’ experiences. Researchers have proposed the use of listener training to improve the rater’s reliability. However, few studies have investigated the effectiveness of training programs. Also, no study has systematically controlled the presence of co-existing resonance and speech disorders in the speech samples used. This study investigates and compares the effectiveness of three training programs: simple exposure, practice only and practice with feedback. The training programs aimed to improve the reliability, agreement and accuracy of perceptual rating of hypernasality in speech with co-existing resonance and speech disorders. Connected speech samples from 52 English-speakers with velopharyngeal dysfunction were used. The results showed that the training programs did not lead to a significant improvement in reliability and accuracy. Possible reasons included high variability in perceptual quality of the training stimuli and types of feedback provided. The findings provide insights and directions for future studies on developing hypernasality training programs.
Hypernasality refers to an excessive amount of perceived nasal resonance in speech. (Boone, McFarlane & Von Berg, 2005). It is a resonance disorder which arises primarily from the atypical coupling of the nasal and oral cavities during the production of non-nasal speech sounds (McWilliams, Morris & Shelton, 1990). This disorder is commonly found in individuals with velopharyngeal dysfunction (VPD) (Dworkin, Marunick & Krouse, 2004). VPD can be due to structural anomalies such as cleft palate; neuropathology such as spastic dysarthria; oral surgery such as adenoidectomy; reduced auditory feedback caused by deafness as well as faulty learned behaviors (Dworkin et al., 2004).

Clinically, it is essential to obtain a valid and reliable measure of hypernasality in order to identify the disorder and to provide a severity index to document changes of the disorder over time as well as to evaluate the effectiveness of a surgical repair or treatment program (Whitehill, Lee & Chun, 2002). Both perceptual ratings and instrumental analysis of hypernasality have been used in clinical settings and for research purposes. Instrumental methods including videofluoroscopy, nasendoscopy and nasometry are used to provide objective measurements of aspects related to hypernasality such as velopharyngeal structure and function (Kreiman, Gerrratt, Kempster, Erman & Berks, 1993). However, since hypernasality is a perceptual quality, perceptual rating remains the “gold standard” for evaluating hypernasality in clinical settings and for validating instrumental measures of velopharyngeal dysfunction (Kuehn & Moller, 2000; McWilliams et al., 1990). Instrumental measurements are generally used to supplement perceptual ratings of hypernasality.

In spite of the common use of perceptual ratings for clinical and research purposes, this subjective evaluation method is prone to variable reliability (Counihan & Cullinan, 1970). Perceptual ratings of hypernasality can be influenced by a variety of factors including the types of speech stimuli (Cheung, 2004; Daniel, 1971); intelligibility of the speakers (McWilliams, 1954); co-existing articulation errors (McWilliams, 1954; Star,

Researchers have thus suggested the use of listener training to increase the reliability of perceptual hypernasality judgment (McWilliams et al., 1990). Various studies have provided the subjects with judgment training prior to the experimental judging task. For instance, Bassich & Ledlow (1986), Lintz & Sherman (1961) and Moller and Starr (1984) required the subjects to practice rating hypernasality; Carney & Sherman (1971) and Whitehill et al. (2002) provided the subjects with simple exposure of hypernasal speech samples; Watterson, McFarlane & Wright (1993) provide both exposure and rating training to the subjects. However, only one of these studies has evaluated the effect of the training programs (Bassich & Ledlow, 1986). None of these studies has provided theoretical principle to support the construct of the programs.

Bassich & Ledlow (1986) studied the validity and reliability of using perceptual ratings for assessing 13 dimensions of voice qualities, including nasality, produced by 20 patients with vocal folds nodules of polyps. Four inexperienced listeners were instructed to rate the voice qualities of sustained vowels /a/ and /i/ using equal appearing interval (EAI) scale and the nasality was being rated on a four-point EAI scale. During the training, feedback of each other’s ratings were provided to the subjects and they were required to reach a criterion level of 80% mean inter-rater reliability before proceeding to the experimental task. The subjects had completed sixteen half-hour training sessions. In the post-training experimental, the listeners’ mean inter-rater reliability was 0.63 in rating pathological speech samples and 0.83 in rating normal speech sample. The mean intra-rater agreement was approximated to 0.75. Despite the provision of extensive training, the authors concluded that the reliability scores obtained in post-training experimental task were lower than the data reported by previous studies which used experienced listeners.
(Hammarberg, Fritzell, Gaufini, Sundberg & Wedin, 1980, cited in Bassich & Ledlow, 1986). Moreover, pre-training baseline was not obtained in the study, therefore, the effect of training on improving reliability and agreement of perceptual rating could not be determined.

A recent study by Lee et al. (2005) investigated the effect of three training programs (exposure only, practice only and practice with feedback) on the reliability of perceptual rating of hypernasality by thirty-six inexperienced listeners. Direct magnitude estimation (DME) was chosen over EAI scale in the study. It is because hypernasality was shown to be prothetic i.e., quantitative and addictive in nature and cannot be subdivided in equal interval (Whitehill et al., 2002). Therefore, EAI scaling was shown not to be a valid measure for rating hypernasality (Whitehill et al., 2005). The training paradigm was based on the perceptual learning mechanism suggested by Goldstone (1998). Goldstone (1998) proposed that perceptual learning involved the processes of imprinting and differentiation. Imprinting refers to the development of an internal standard for a specific feature in a stimulus while differentiation refers to the ability to discriminate the signals in a stimulus. It was hypothesized that repeated exposure to a stimulus would facilitate the listener’s development of internal standard for detecting and differentiating the specific perceptual feature. Also, it was hypothesized that provision of feedback would facilitate the development process (Intrator, 1994; Gibson, 1969). Lee et al.’s (2005) findings showed that both practice groups showed statistically significant higher inter-rater reliability ($r = 0.91$ and $0.90$) than the exposure only group ($r = 0.74$) in rating female speakers. The intra-rater reliability was obtained by comparing the number of listeners who showed statistically significant correlation ($p < 0.05$) in test-retest ratings. The results showed that the practice groups, with (10/12) or without (8/12) feedback, had a larger number of listeners obtained significant intra-rater correlation than the exposure group (5/12) for rating male speakers. However, a pre-training baseline was not obtained in the study so the statistical significance of the degree of improvement in reliability of the trained groups could not be identified.
The current study was an extension of previous work by Lee et al., (2005) and aimed to further investigate the effect of practice and feedback on assisting listeners to reliably rate hypernasality.

In addition, it was found that patients with cleft palate, which is a major population of patients with VPD, had a high prevalence of having speech disorders, hyponasality and voice disorders (Keuning, Wienke, Van Wijngaarden & Dejonckere, 2002). Some patients with cleft palate and VPD appeared to develop atypical articulation patterns to compensate for the VPD. The most common compensatory articulation errors include glottal stops and pharyngeal fricatives (Peterson-Falzone, Hardin-Jones & Karnell, 2000). The presence of speech disorders might affect the reliability and accuracy of the judgment of hypernasality (McWilliams, 1954; Star, Moller, Dawson, Graham & Skarr, 1984).

McWilliams (1954) studied the correlation between perceptual ratings of hypernasality, intelligibility and articulation errors. Seven judges rated the hypernasality of 48 speakers with cleft palate using a five-point EAI scale. The speech samples were obtained from passage reading. The results revealed that hypernasality ratings were generally higher in speech samples with higher number of articulation errors and lower intelligibility scores. However, a high intra-rater reliability (r = 0.93) in rating hypernasality was obtained in the study. This suggested that co-occurring speech disorders did not reduce the intra-reliability of hypernasality rating. The accuracy of the ratings was not examined in the study.

Keuning et al. (2002) also studied the correlation between hypernasality and four parameters related to cleft palate speech, i.e., nasal emission, misarticulations, intelligibility and overall severity. Six speech-language pathologists rated the hypernasality and other parameters using five visual analog (VA) scales of 10 cm long. The result revealed that perceived overall severity of cleft palate speech was correlated more with intelligibility than hypernasality rating. The result was consistent with McWilliam’s (1954) suggestion that raters tended to give higher hypernasality ratings when co-existing speech
disorders were present. Inter-rater reliability (0.49) for hypernasality rating was shown to be moderate (0.49) while inter-rater reliability and accuracy of rating were not investigated.

The correlation between rating of nasality and speech disorders were also studied by Starr, Moller, Dawson, Graham & Skaar (1984). The speech samples were obtained from 15 speakers with cleft palate through a passage reading task. There were six groups of raters i.e., speech pathologists experienced in judging cleft palate speech, speech pathologists in general practice, parents of children with cleft palate, parents with normal children, children with cleft palate and hypernasality and normal children. The raters rated the hypernasality and articulation using two eight-point EAI scales. It was found that the correlation between ratings of hypernasality and articulation was $r = 0.45$ for the experienced speech pathologists, $r = 0.77$ for the speech pathologists in general practice, $r = 0.94$ for the parents of children with cleft palate and $r = 0.95$ for the parents with normal children. It was suggested that co-existing speech disorders might correlate with the rating of hypernasality and this correlation was the lowest in the ratings performed by speech pathologists with expertise in judging cleft speech. This could imply that increased exposure or perceptual training, as possessed by experienced speech pathologists, might serve to increase the accuracy in hypernasality training.

The current study was designed to systematically allow participants to progress from rating hypernasality only and to rating hypernasality in speech with one co-existing disorder (hyponasality or speech disorders). It was expected that the training would enable the participants to develop the ability to identify hypernasality clearly and to focus on hypernasality even in the presence of co-existing resonance and speech disorders. Other factors that might also affect hypernasality rating, for example, voice disorders, were not included in the current study in an effort to limit the scope of the study.

In contrast to Lee et al.’s study (2005), a VA scale was used as the rating scale instead of DME.
A recent study by Cheng (2006) demonstrated that DME and VA had similar validity and reliability for perceptual ratings of hypernasality. VA scale was suggested to be a preferable scaling method than DME due to its relative simplicity in rating and subsequent data analysis (Cheng, 2006). The comparison of DME scores rated by different raters requires an equalization procedure. The use of VA scale in the current training program provided the advantage of allowing the participants to get immediate feedback in terms of an expert’s rating once they have rated the stimuli. It was hypothesized that the provision of feedback would help to facilitate the listeners’ to form more consistent internal standards and to improve their ability in differentiating a specific feature i.e., hypernasality in speech stimuli (Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993).

The current study aimed to investigate

i) the effect of practice on intra-rater reliability of perceptual rating of hypernasality in speech with co-existing resonance and speech disorders

ii) the effect of practice on inter-rater reliability of perceptual rating of hypernasality

iii) the effect of provision of feedback during practice on inter-rater and intra-rater reliability of perceptual rating of hypernasality

iv) the effect of practice on accuracy in perceptual rating of hypernasality in naïve listeners.

The result of the study was expected to provide directions on the development of training programs for clinicians to improve the reliability and accuracy of perceptual rating of hypernasality in speech with co-existing resonance and speech disorders.

Method

Participants

Thirty listeners (24 females and 6 males) with a mean age of 20.7 years (SD = 0.64, range = 20-22 years) were recruited to participate in the study on a voluntary basis. All
were undergraduate students from the Division of Speech and Hearing Sciences, The University of Hong Kong. The participants were native Cantonese speakers with normal hearing as defined by a pure-tone audiometric screening at 25 dB HL at octave frequencies from 250Hz to 8000Hz. All of them had limited previous exposure to hypernasal speech samples during coursework but had no clinical experience in judging hypernasal speech. The participants were randomly assigned into three groups: the practice with feedback group (Group PF), the practice group (Group P) and the exposure group (Group E).

*Speech Stimuli*

Two types of stimuli were used in the study: experimental stimuli and training stimuli. All the stimuli were extracted from a database from the University of Wyoming. The database consisted of 4828 sentences produced by 448 English-speaking children and adolescents with VPD. Each speaker produced approximately eleven sentences. The sentences were elicited through reading aloud written sentences or a repetition task, depending on the literacy ability of the speakers.

The experimental and training stimuli used in this study were 76 sentences produced by 52 speakers. The mean age of the speakers was 9.1 years (SD = 3.25, range = 4-20). One to two sentences were selected from each speaker. The two sentences were: “Nick’s grandmother lives in the city” and “We go swimming on a very hot day”.

Connected speech samples were used because it was reported that listeners made more reliable judgment on hypernasality in connected speech samples (Cheung, 2004; Daniel, 1971). The above two sentences were chosen because they contained both nasal and non-nasal consonants, which allowed the judgment of both hypernasality and hyponasality. Although the listener participants were Cantonese, English speech samples were used because an extensive Cantonese database is not yet available. Cheng (2006) found that there was no significant difference in degree of reliability and severity rating of
hypernasality between Cantonese-speaking university students and native English-speaking listeners (Cheng, 2006).

Professor David Jones from the University of Wyoming, who is an expert in perceptual rating of speech in VPD, had previously provided ratings on the severity of hypernasality and articulation disorder on 7-point EAI scales and rating on the severity of hyponasality on a 4-point EAI scales for all the stimuli. Due to the fact that his ratings were based on a whole assessment, which included articulation test, sentence production task and spontaneous speech production, the expert’s ratings might have a discrepancy with the listener’s perceived nasality and articulation ratings of the specific sentences used in this study. For example, speakers with an articulation rating of 5 (1 is normal, 7 is severe) might not have shown an articulation error in producing the specific sentences chosen. Therefore, the samples were also judged by Professor Tara Whitehill of the University of Hong Kong, who had over 10 years of clinical experience in perceptual rating of hypernasality, to select the most representative samples that could best serve the purposes of each part of the training.

Experimental Stimuli

The experimental stimuli were used in the pre- and post- training experimental tasks. The experimental stimuli consisted of 42 sentences produced by 32 speakers (Appendix A). The mean age of the speakers was 9.28 years (SD = 3.20; Range = 4-20). The stimuli included samples with normal resonance and articulation, hypernasality only, hyponasality only, speech disorders only as well as samples with hypernasality with one co-existing disorders i.e., hypernasality plus hyponasality or hypernasality plus speech disorders. Twenty of the 42 experimental stimuli were also used as training stimuli so that comparison of listeners’ reliability of rating trained and untrained stimuli were possible. All the stimuli were repeated once in the experimental tasks for measurement of intra-rater reliability. As a result, there were a total of 84 trials in each experimental task.
Training Stimuli

The training stimuli consisted of 54 English sentences produced by 36 speakers (Appendix A). The mean age of the speakers was 9.03 years (SD = 2.98; Range = 4-17). The stimuli included samples with normal resonance and articulation, hypernasality, hyponasality, speech disorders only as well as samples with hypernasality with one co-occurring disorders i.e. hyponasality or speech disorders. The nasality ratings provided by Professor David Jones was used as a basis for the provision of feedback in Group PF. However, as the original rating was on a 7-point EAI scale but VA scale were used in this study, the ratings were converted to an approximate range of severity by the author (Refer to Appendix A).

Procedures

Participants from all groups first performed a pre-training experimental task. Participants from Group PF and P received a training session immediately after the experimental task. Group E did not receive any training but were given simple exposure to the speech stimuli. Finally, all participants performed a post-training experimental task one week after the training or exposure session. No requirement was imposed for the participants to achieve a specific level of accuracy in the training tasks.

All sessions were administered on an individual basis using specially designed computer program made by E-prime, Microsoft Excel and Visual Basic (Chan & Yiu, 2002). The sessions were conducted in a sound-treated booth and the stimuli were presented binaurally through an Audio-Technica ATH-T2 headphone.

Pre- and post- experimental tasks

The experimental tasks aimed to measure the participants’ baseline performance and performance after training or exposure. All stimuli were presented through a computerized program using Microsoft Excel and Visual Basic. The order of presentation of the stimuli was randomized across the participants. Participants were allowed to listen to each
stimulus a second time if they choose to. Then they were required to provide a hypernasality rating using a 10 cm horizontal scroll bar (VA scale) shown on the computer. The left end of the scroll bar indicated normal resonance and the right end of the scroll bar indicated extremely severe hypernasality. The computer automatically measured and recorded the participants’ responses. The participants were instructed to try to ignore any co-existing resonance and speech disorders and to focus on rating the degrees of hypernasality. Each session took approximately 20 minutes to complete.

Training tasks

The training session was divided into three parts: Part 1- judgment training, Part 2- identification training and Part 3- perceptual rating training. The training paradigm was adopted and modified from Lee et. al. (2005). Parts 1 and 2 were run through an E-prime program (Psychology Software Tools, Pittsburgh, PA). Part 3 was run through a program using Microsoft Excel and Visual Basic.

The training began with a presentation of introductory slides which included the definition of VPD and hypernasality. Other resonance and speech disorders that might affect perceptual rating of hypernasality were also introduced. To let the participants to better understand the concepts, five speech samples representing each of the resonance and speech disorders were included. The presentation materials were adopted and modified with permission from the Microsoft PowerPoint prepared by Lee et al. (2005).

Upon completion of the introductory presentation, Group P and Group PF immediately received the perceptual training while Group E received simple exposure to the speech stimuli.

Part 1 aimed to train the participants to focus attention on a specific disorder and to be able to judge the presence and absence of a disorder. There were three blocks, one for each disorder to be judged i.e. hypernasality, hyponasality or speech errors. Each block contained 10 stimuli, resulting in a total of 30 trials. Five of the stimuli were stimuli with
no abnormality and the other five stimuli had one specific disorder. Each stimulus was presented twice. Then the participants were required to perform a judgment by clicking an “Absence” or a “Presence” button. A model answer was shown on the computer screen after each trial for Group PF while no feedback was given to Group P. The model answer is in sentence form, for example, “This sample is hypernasal” or “This sample is not hypernasal”. Then the participants were required to listen to the stimulus once again before proceeding to the next trial.

Part 2 included 60 trials and aimed to increase the participants’ ability of identifying the co-existing resonance and speech disorders associated with VPD speech. Each stimulus was presented twice. The participants were then required to identify the disorder(s) present in each stimulus. This part was sub-divided into two blocks with increasing difficulty. In the first 20 trials, the participants were required to identify one disorder in each trial. For example, they had to answer by clicking the button of “normal”, “hypernasality”, “hyponasality” or “speech disorders”. In the next 20 trials, they were required to identify any disorder(s) in each stimulus. For example, they had to answer by clicking the button of “normal”, “hypernasality”, “hyponasality”, “speech disorders”, “hypernasality with hyponasality”, “hypernasality with speech disorders. A model answer was shown on the computer screen after each trial for Group PF. The model answer is in sentence form, for example, “This sample is normal”, “This sample is hypernasal”.

Part 3 aimed to allow the participants to practise rating hypernasality using a VA scale. There were 60 stimuli. Participants were allowed to listen to each stimulus twice only in order to control their exposure to the stimuli. In the first 30 trials, the stimuli only had hypernasality, of varying degrees (mild to extremely severe). In the next 30 trials, the stimuli were of hypernasality of varying degrees as well as co-existing resonance and speech disorders (i.e., articulation errors or hyponasality). The participants were required to conduct hypernasality rating using the 10cm scroll bar on the computer. A suggested
rating was shown on the computer screen after each trial for Group PF. The whole training session took approximately one hour to complete.

For Group E, the participants were required to listen to all training stimuli for three times, resulting in a total of 162 trials of exposure. It is to control similar exposure to the stimuli across the three groups. The stimuli were presented through Microsoft PowerPoint on an individual basis. The exposure sessions took approximately half-hour to complete.

Data Analysis

The VA ratings of hypernasality were used to compare the performance of the three groups before and after training. Outcome measures included intra-rater reliability, intra-rater agreement, inter-rater reliability and accuracy of rating.

Pearson’s product-moment correlation was used to determine the intra-rater reliability for each listener. Intra-rater agreement was calculated using each listener’s first and second ratings of an identical stimuli. Ratings that were within one centimeter of one another on the VA scale were considered as agreeing with each other. The percentage agreement was then calculated by dividing the number of agreed rating over the total number of trials (n = 42).

Inter-rater reliability was calculated using intraclass correlation coefficient (ICC type 3,k; equivalent to Cronbach’s alpha) (Kreiman et al. 1993; Shrout & Fleiss, 1979).

For measuring the accuracy of rating, the listener’s rating which is equal to or within the range of suggested rating was considered to be accurate. The percentage accuracy was then obtained by dividing the number of accurate rating over the total number of trials (n = 42).

Repeated two-way analysis of variance (ANOVA) was performed for each of the outcome variables. The two-level variable “session” i.e., pre-training and post-training was treated as the within group factor while the three-level variable “training type” i.e., Group P, PF and E was treated as the between group factor.
Results

Intra-rater reliability

The mean and standard deviation of correlation coefficients of the three groups before and after training are listed in Table 1.

Table 1.
Intra-rater reliability of three groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-training session</th>
<th>Post-training session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Group PF</td>
<td>0.75</td>
<td>0.09</td>
</tr>
<tr>
<td>Group P</td>
<td>0.60</td>
<td>0.17</td>
</tr>
<tr>
<td>Group E</td>
<td>0.64</td>
<td>0.17</td>
</tr>
</tbody>
</table>

p-value < 0.05

A repeated two-way ANOVA was performed to determine if there was any statistical significant difference in mean correlation coefficients among the three groups.

The main effect for session compared the intra-rater reliability before and after the training or exposure session. The results indicated that there was no statistically significant difference in intra-rater reliability between the pre- and post-training session [F (1, 27) = 1.71; df= 27; p> 0.05]. The main effect for training type compared the intra-rater reliability among the three groups i.e. group PF, P and E. The results indicated that there was no statistically significant difference among the three groups session [ F(2, 27) = 0.13; df = 27; p> 0.05]. The results indicated that there was no statistically significant “session” by “training type” interaction effect [F(2, 27) = 1.00; df = 27; p > 0.05].

No post-hoc comparison was performed since the main and interaction effects were not statistically significant.
Intra-rater agreement

The mean and standard deviation of percentage agreement of the three groups before and after training are listed in Table 2.

Table 2

Mean intra-rater agreement of three groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-training session</th>
<th>Post-training session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Group PF</td>
<td>0.49</td>
<td>0.10</td>
</tr>
<tr>
<td>Group P</td>
<td>0.50</td>
<td>0.13</td>
</tr>
<tr>
<td>Group E</td>
<td>0.48</td>
<td>0.10</td>
</tr>
</tbody>
</table>

A repeated two-way ANOVA was performed to determine if there is any significant difference in mean percentage of agreement among the three groups.

The main effect for session compared the intra-rater agreement before and after the training or exposure sessions. The results indicated that there was no statistically significant difference between the intra-rater agreement in the pre- and post- training session \[ F(1, 27) = 1.71; \text{df} = 27; p > 0.05 \]. The main effect for training type compared the intra-rater agreement among the three groups i.e. group PF, P and E. There was a statistically significant difference among the three groups \[ F(2, 27) = 0.13; \text{df} = 27; p < 0.05 \]. There was no statistically significant “session” by “training type” interaction effect \[ F(2, 27) = 1.00; \text{df} = 27; p > 0.05 \].

Since the main effect for training type was found to be statistically significant, a post-hoc comparison using Tukey HSD test was performed to compare the intra-rater agreement among three groups. However, no statistically significant difference was found among the three groups in both the pre- and post- training ratings (p > 0.05 for all).
Inter-rater reliability

Inter-rater reliability of the three groups before and after training is listed in Table 3.

Table 3.

Mean inter-rater reliability of three groups (ICC; 3, k)

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-training session</th>
<th>Post-training session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td></td>
</tr>
<tr>
<td>Group PF</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td>Group P</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>Group E</td>
<td>0.91</td>
<td>0.85</td>
</tr>
</tbody>
</table>

The results indicated that all three groups showed a slight decrease in ICC values after the practice or exposure sessions.

Accuracy of ratings

The mean and standard deviation of percentage accuracy in the three groups before and after training is listed in Table 4.

Table 4

Mean percentage of accuracy of the three groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-training session</th>
<th>Post-training session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Group PF</td>
<td>0.29</td>
<td>0.04</td>
</tr>
<tr>
<td>Group P</td>
<td>0.26</td>
<td>0.06</td>
</tr>
<tr>
<td>Group E</td>
<td>0.26</td>
<td>0.06</td>
</tr>
</tbody>
</table>

A repeated two-way ANOVA was performed to determine if there is any statistical significant difference in mean percentage of accuracy among the three groups.
The main effect for session compared the percentage of accuracy before and after the training or exposure sessions. The results indicated that there was no statistically significant difference in percentage of accuracy between the pre- and post-training session [F (1, 27) = 1.09; df = 27; p > 0.05]. The main effect for training type compared the percentage of accuracy in expert and naïve listeners’ ratings among the three groups i.e. group PF, P and E. The results indicated that there was no statistically significant difference among the three groups [F (1, 27) = 3.30; df = 27; p > 0.05]. The results indicated that there was no statistically significant “session” by “training type” interaction effect [F (2, 27) = 0.01; df = 27; p > 0.05].

No post-hoc comparison was performed since the main and interaction effects were not statistically significant.

Discussion

The first purpose of the study was to investigate the effect of practice on the intra-rater reliability of perceptual rating of hypernasality in speech with co-existing resonance and speech disorders. The effect was evaluated by two measures i.e. comparison of the means of intra-rater reliability as well as means of percentage of agreement of the three groups before and after the practice or exposure sessions. The results revealed that all three groups achieved moderate intra-rater reliability (r = 0.60 - 0.75) and intra-rater agreement (r = 0.48 - 0.53) across the sessions. The absence of statistically significant main effects for session in both measures showed that the intra-rater reliability and agreement of the three groups did not change significantly after training. This suggested that the practice and exposure sessions had no effect on the participants’ intra-rater reliability and agreement. A previous study by Lee et al. (2005) using a similar training paradigm revealed that more listeners in the groups with practice, with or without feedback, had a significant intra-rater correlation than the exposure group in judging male speakers. However, a direct comparison of the
results between Lee et al.’s (2005) and the current study was not possible as no pre-training data were obtained in the first study.

The results of the current study were not consistent with Goldstone’s (1998) principle on perceptual learning mechanism. A possible reason for the present finding is the involvement of abundant exemplars as training stimuli (76 sentences produced by 52 speakers). The original aim to include a large number of stimuli was to avoid the participants to develop an internal standard for a limited individual stimuli instead of the targeted perceptual quality i.e., hypernasality. Goldstone (1998) suggested that imprinting might occur if the specific feature (such as hypernasality) was varied independently and served a functional role for separating two classes of stimuli. The perception of hypernasality has been found to be multidimensional (Zarick et al, 2000) and the judgment might possibly be affected by the presence of other dimensions including voice quality, intelligibility, resonance, loudness and pitch. Therefore, the inclusion of large number of exemplars might thus largely increase the variability among the training stimuli and thus decrease the functional role of the target feature i.e., hypernasality for the distinguish of two stimuli. As a result, the participants might not able to develop a stable internal standard for judgment and rating of hypernasality.

Moreover, Goldstone (1998) suggested that internal standard was formed by repeated exposure to the same stimulus. Thus, an increase in number of exemplars leads to a decrease in the number of exposures of the same stimuli which might negatively affect the participants to develop internal standards of a stimuli or perceptual quality.

In addition, the relationship between the practice and rating tasks might have affected the practice effect. In the one-hour practice program, the judgment and identification tasks accounted for 45% of the trials (100/220) while rating practice accounted for 54% of the trials (120/220). The judgment and identification tasks aimed to develop the participants’ ability in differentiating hypernasality from other commonly co-occurring resonance and
speech disorders. The rating practice was aimed at developing the participants reliability and agreement in performing hypernasality rating and was more directly related to the pre- and post- training task. The number of trials and time used in actual rating practice might not have been sufficient for the participants to develop better ability in judging the different severities of hypernasality.

The second purpose of the study was to investigate the effect of practice on the inter-rater reliability of perceptual rating of hypernasality in speech with co-existing hyponasality and speech disorders. The effect was evaluated by comparing the ICC values for each group before and after the training or exposure sessions. The results revealed that all three groups achieved high inter-rater reliability (r= 0.85-0.96) across the sessions. However, all three groups showed a slight decrease in inter-rater reliability in the post-training rating tasks. This suggested that the practice and exposure sessions were not effective in improving the participants’ inter-rater reliability in hypernasality rating.

The finding might be attributed by a plateau of performance achieved by the participants. All three groups achieved high inter-rater reliability (r= 0.91- 0.96) in the pre-training sessions. The results were consistent with the findings (r= 0.96) of Cheng (2006) which investigated inter-rater reliability of listeners’ with different language background in rating stimuli with hypernasality only. It was possible that the participants had reached plateau in ability in achieving high inter-rater reliability rating, and thus did not demonstrate any significant changes after practice. The slight reductions in ICC values were similar across Group PF (0.04), Group P (0.02) and Group E (0.06) and might have been caused by chance.

The third purpose of the study was to investigate the effects of provision of feedback in practice on improvement of intra- and inter- reliability. The effects were evaluated by comparing the means of intra-rater reliability, means of percentage of agreement and means of inter-rater reliability between the three groups. The absence of a significant main effect
for training type indicated that practice with feedback did not significantly improve the raters’ reliability and agreement.

The finding might be attributed by the lack of validity and specificity of the feedback given. In the current study, the stimuli were collected by an expert who had expertise in hypernasality rating and had well-established judgment on the perceptual quality of the stimuli. Therefore, despite the fact the ratings used a 7-point EAI scale, they were still used as the expert rating, through conversion into an approximate range of severity. However, this conversion method was not validated and it might have led to a discrepancy between the expert’s rating and the suggested rating. The Group PF might thus have been unable to receive the most reliable and valid feedback from the training. This might limit the effect of training in improving the participants’ reliability, agreement and accuracy in rating the stimuli.

Moreover, the expert’s rating was based on his clinical impression throughout the whole session and not specific to a certain sentence stimuli. Therefore, the specific sentences used in this study were judged by another expert to select the stimuli which best served the purpose of the study. Discrepancies on nasality and articulation ratings between the two experts were found. It was hypothesized that the feedback might serve to assist the participated to restructure an internal standard of a specific feature. Thus the reliability and validity of feedback played a significant role in affecting the training effect.

The fourth purpose of the study was to investigate the effect of practice on accuracy. The absence of a significant main effect for session showed that the accuracy of ratings did not change significantly after training for all three groups.

Based on the findings of the current study, several implications for further studies have been generated.

In further studies, researchers might reduce the number of exemplars for the training stimuli. It is hypothesized that the use of less exemplars might reduce the variability among
the stimuli and provide a higher function load of the target feature i.e., hypernasality for distinguishing hypernasality. It might help to reduce the variable internal standards possessed by each listener and thus increase the reliability in perceptual judgment (Kreiman, 1990).

The suggested hypernasality, articulation and hyponasality ratings were obtained from a single expert. The criterion of the articulation rating was not specified. Further studies might include a panel of experts to rate hypernasality using a consensus model (Lee et al., 2005). This might serve to provide a more reliable suggested rating as feedback. Moreover, the experts might also need to provide objective measures for the articulation rating. For example, they might measure the speech intelligibility through calculating the percentage of consonant correct. This would provide a systematic control of one variable that might interact with the judgment of hypernasality.

The absence of effect of practice in the current study may have been due to the unstable internal standard acquired by the trained group. As the post-training rating task was carried out one week after the practice session, it is possible that the trained participants might have failed to maintain the internal standard for imprinting and differentiation ability. Therefore, future studies might include a rating task immediately after training as well as at a delayed time interval. This would allow the researchers to investigate the immediate practice effect as well as the degree of maintenance.

Moreover, it has been suggested that the provision of external standards in the form of anchors could serve improve the reliability in perceptual judgment (Chan & Yiu, 2002; Kreiman, 1992). In this paradigm, listeners can compare the external standard to the stimulus to be rated. This has been shown to decrease variability within and between judges in perceptual rating (Chan & Yiu, 2002). Therefore, future studies might investigate the use of external standard in improving raters’ reliability and agreement.

In conclusion, the current study was one of the few studies which attempted to investigate the effect of practice and feedback in improving naïve listeners’ reliability,
agreement and accuracy in hypernasality rating. The effectiveness of the program might have been affected by factors such as the reliability and validity of the feedback given, selection of training stimuli and the training paradigm. Though the results reflected that the current training program was not effective in improving the perceptual hypernasality rating, it served to provide insights and directions on future studies of hypernasality training.

Acknowledgements

I would like to express my sincere gratitude to Professor Tara Whitehill for her guidance and support throughout the process of the study. I would like to thank Professor David Jones of the University of Wyoming for the provision of speech stimuli. I would also like to thank Dr. Karen Chan for her advice and assistance in the development of the computer programs, Dr. Alice Lee for the provision of the presentation materials used in training and Mr. Dennis Cheng for the provision of edited speech stimuli. Thanks are also extended to Mr. Donald Chan and Mr. Raymond Wu for their technical support, and my fellow classmates for their support and encouragement. Last but not the least, I would like to thank all the participants for their time devoted.
References


Psychology Software Tools, Pittsburgh, PA.


Appendix A

Details of speakers; hypernasality, hyponasality, articulation ratings given by Professor Jones; classification of disorders judged by Professor Whitehill; and suggested range of hypernasality rating given in training task

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Sex</th>
<th>Age</th>
<th>Hyper</th>
<th>Artic</th>
<th>Hypo</th>
<th>Classification</th>
<th>Suggested VA rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>6</td>
<td>M</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>11</td>
<td>F</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Normal</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>Hyper</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>Hyper</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>Hyper</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>16</td>
<td>F</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>Hyper</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>Hyper</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>18</td>
<td>M</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>Hyper</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>11</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>Hyper</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>Speaker</td>
<td>Sex</td>
<td>Age</td>
<td>Hyper</td>
<td>Artic</td>
<td>Hypo</td>
<td>Classification</td>
<td>Suggested VA rating</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>-------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>21</td>
<td>M</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>Hyper</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>22</td>
<td>F</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>Hyper</td>
<td>~ 2-4</td>
</tr>
<tr>
<td>23</td>
<td>M</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>Hyper</td>
<td>~ 2-4</td>
</tr>
<tr>
<td>24</td>
<td>M</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>Hyper</td>
<td>~ 2-4</td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>Hyper</td>
<td>~ 6-8</td>
</tr>
<tr>
<td>26</td>
<td>F</td>
<td>12</td>
<td>5</td>
<td>7</td>
<td>0</td>
<td>Hyper</td>
<td>~ 6-8</td>
</tr>
<tr>
<td>27</td>
<td>M</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Hypo</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>28</td>
<td>F</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Hypo</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>29</td>
<td>F</td>
<td>11</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>Hypo</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>11</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>Hypo</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>31</td>
<td>M</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Hypo</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>32</td>
<td>M</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>Hypo</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>33</td>
<td>M</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>Artic</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>34</td>
<td>M</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>Artic</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>35</td>
<td>M</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>Artic</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>36</td>
<td>M</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>Artic</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>37</td>
<td>M</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>Artic</td>
<td>~ 0-1</td>
</tr>
<tr>
<td>38</td>
<td>F</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>Hyper with Hypo</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>39</td>
<td>F</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>Hyper with Hypo</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>40</td>
<td>M</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>Hyper with Hypo</td>
<td>~ 2-4</td>
</tr>
<tr>
<td>41</td>
<td>F</td>
<td>10</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 1-2</td>
</tr>
<tr>
<td>42</td>
<td>M</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 2-4</td>
</tr>
<tr>
<td>43</td>
<td>F</td>
<td>14</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 2-4</td>
</tr>
<tr>
<td>Speaker</td>
<td>Sex</td>
<td>Age</td>
<td>Hyper</td>
<td>Artic</td>
<td>Hypo</td>
<td>Classification</td>
<td>Suggested VA rating</td>
</tr>
<tr>
<td>---------</td>
<td>-----</td>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>44</td>
<td>M</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 4-6</td>
</tr>
<tr>
<td>45</td>
<td>F</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 4-6</td>
</tr>
<tr>
<td>46</td>
<td>F</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 6-8</td>
</tr>
<tr>
<td>47</td>
<td>F</td>
<td>10</td>
<td>5</td>
<td>7</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 6-8</td>
</tr>
<tr>
<td>48</td>
<td>M</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 6-8</td>
</tr>
<tr>
<td>49</td>
<td>F</td>
<td>11</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 6-8</td>
</tr>
<tr>
<td>50</td>
<td>M</td>
<td>9</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 8-9</td>
</tr>
<tr>
<td>51</td>
<td>M</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 8-9</td>
</tr>
<tr>
<td>52</td>
<td>M</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>Hyper with Artic</td>
<td>~ 8-9</td>
</tr>
</tbody>
</table>

Note: Hyper- Hypernasality; Hypo- Hyponasality; Artic- Articulation disorders.