<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Effect of feedback on the effectiveness of a paired comparison perceptual voice rating training program</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Other</strong></td>
<td><strong>Contributor(s)</strong>: University of Hong Kong.</td>
</tr>
<tr>
<td><strong>Author(s)</strong></td>
<td>Law, Tsz-ying</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/10722/55486">http://hdl.handle.net/10722/55486</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>Creative Commons: Attribution 3.0 Hong Kong License</td>
</tr>
</tbody>
</table>
Effect of Feedback on the Effectiveness of a Paired Comparison Perceptual Voice Rating Training Program

Law Tsz Ying

University of Hong Kong

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

The present study examined the effect of feedback on the effectiveness of a paired-comparison perceptual voice rating training program. Thirty male and 30 female Cantonese speakers without voice disorder and with normal hearing ability were randomly assigned to three gender-balanced groups. One group received training with augmented feedback (Group F), another group received training without augmented feedback (Group NF), and the control group (Group C) received no training. Training effect was assessed throughout three rating sessions—pre-training, post-training (2 days after training), and review session (1 week after training). Generalization was investigated by including non-trained novel stimuli in the rating sessions. Result demonstrated significant improvement in both the feedback and no feedback group. Generalization to novel stimuli was also evidenced. Furthermore, the guidance hypothesis emphasized by Schmidt and Wulf (1997) was not supported in this study. It was found that good retention performance was found in both Group F and Group NF. This result further support Kluger and DeNisi’s (1996) feedback interventions theory (FIT) and Vollmeyer and Rheinbery’s (2005) cognitive-motivational process model that augmented feedback leads learners to deeper procession and develop a more systematic strategy for tackling the task and hence result better retention.
Introduction

Perceptual evaluation of voice quality has been playing an important role in clinical settings in identifying the type and quantifying the severity of the dysphonic quality. Carding, Carlson, Epstein, Mathieson, and Shewell (2000) summarized that the importance of perceptual evaluation of voice quality in both clinical and research settings: a) comparing information obtained from other voice assessments (e.g. acoustic analysis of the speech waveform) with the perceptual voice quality evaluation, b) setting the baseline of the vocal problem for planning therapy and monitoring the progress of therapy, c) being used in voice efficacy literature as a battery for evaluation of voice outcomes, and d) facilitating communication with the patient and other professionals. Nevertheless, the reliability of the traditional perceptual voice evaluation in identifying the type and quantifying the severity of the dysphonic quality by the subjective judgment of listener has been questioned in different studies (e.g. Carding et al., 2000; Gerratt & Kreiman, 2001). In order to improve the reliability of perceptual voice evaluation to the judges, provision of training has been suggested. (e.g. Carding et al., 2000; Gerratt and Kreiman, 2001).

According to Goldstone’s (1998) hypothesis on perceptual learning, two processes are involved in perceptual learning. The first mechanism is stimulus imprinting. During the stimulus imprinting process, internalized detectors or standards are generated owing to the repeated exposure to the stimulus. Based on this principle, Chan and Yiu (2002) further hypothesized that listeners may develop internal standards for the specific voice qualities instead of the individual training stimuli, if listeners are repeatedly exposure to voice samples of the targeted voice qualities with different severity level. The second process of perceptual learning suggested by Goldstone (1998) is stimuli differentiation, which listeners learn to separate different signals during perceptual learning. It has been
shown that higher reliability on evaluating rough and breathy voice was resulted after perceptual voice training was provided to the judges (Chan & Yiu, 2006; Chan & Yiu, 2002). Based on the support that training improved perceptual voice evaluation skills, Chan and Yiu (2006) carried out further investigation to compare the effectiveness of different perceptual voice training approaches.

According to Chan and Yiu (2006), trained listeners showed significant improvement in perceiving breathiness voice after participating in training programs that either used anchoring or paired comparison methods. For the anchoring method, participants were required to match the testing stimulus with one of the six references given. For the paired comparison methods, participants were required to compare a pair of voice stimuli to find out whether the level of breathiness was identical. Chan and Yiu (2006) also found that paired-comparison training showed better maintenance in the perceptual evaluation training. The present study, therefore, focused on the paired-comparison method. In their study, augmented feedback was provided throughout the training session. ‘Augmented feedback is the information provided about the task that is supplemental to, or that augments, inherent feedback’ (Schmidt & Lee, 1999, p.325). Knowledge of results (KR) is one type of augmented feedback frequently used in the studies of perceptual voice evaluation and motor learning (e.g. Chan & Yiu, 2005; Steinhauer & Grayhack, 2000). According to Magill (1998), KR is the information about the outcome of performing a skill or about achieving the goal of the performance.

Augmented feedback has generally been regarded as a strategy to enhance learning. Ammons’s review (1956, cited in Kluger & DeNisi, 1996) reached the two most influential conclusions—augmented feedback improves learning and increase motivation. Kluger & DeNisi (1996) further summarized several theories and researches in the area of
the positive effects of feedback interventions on performance. They concluded that “the
domain of the applications of these theories to augmented feedback is either motivation or
learning” (p.259). Vollmeyer and Rheinbery (2005) further explained this concept by
using the cognitive-motivational process model. They suggested that the implementation
of feedback provides more useful information for learners, triggers deeper processing of
the learning material, and hence develop a more systematic strategy. Learners should
perform better when they had to apply their knowledge and eventually their final
performance would be improved. Salmoni, Schmidt, & Walter (1984) also suggested that
augmented feedback increases cognitive elaboration, helping learners to thinking more
about the task and better retention is resulted due to the deeper processing. This view of
learning is consistent with the preliminary feedback interventions theory (FIT) proposed
by Kluger and DeNisi (1996). They suggested that locus of attention is a probabilistic
process, which attention can be present simultaneously, or with quick alternations among
the 3 levels of hierarchy—with meta-task processes at the top of the hierarchy, followed
by the task-motivation processes, which is the focal task processes, and task-learning
processes at the bottom of the hierarchy. They assumed that attention was generally at the
moderate level of the hierarchy—task motivation processes, and feedback interventions
(FIs) can divert attention either to the higher or lower level.

In general, FIs is compared with the task standards at the task-motivation processes, if the
feedback sign is negative (e.g. ‘Your answer is incorrect!’), more effort will be imputed.
On the other hand, effort will be decreased or maintained if the sign is positive (e.g.
‘You’ve got the correct answer!’). When feedback sign is sustained at the negative level,
additional effort is deemed insufficient for eliminating the feedback-standard discrepancy.
As a result, attention will be diverted to the lower task-learning processes, motivating
learners to focus more on the task, and develop new specific strategies to tackle the task.
Furthermore, Kluger and DeNisi (1996) also suggested that although FIs may help learner focus more on the task and develop new strategies, both salient negative FI (such as failed to identify the voice quality difference of the stimuli consistently) and salient positive FI (such as correctly identify the voice quality differences without any difficulties) may shift attention to self causing competition for cognitive resources, leading to performance loss.

Nevertheless, a number of studies showed that frequent augmented feedback adversely affected the performance (Steinhauer & Grayhack, 2000; Schmidt & Wulf, 1997). Schmidt and Wulf (1997) emphasized the guidance hypothesis for feedback processing. They suggested that knowledge of result (KR) serves as a guidance role, preventing learners from making error. With the presence of KR, learners would heavily rely on it to produce the correction action during practice, keeping good performance during the practice sessions. As a result, the reliance on feedback would shift learners’ attention from acquiring the necessary capability to deal with the intrinsic information on the retention or transfer test, so performance fall when augmented feedback is no longer available. This guidance hypothesis for feedback processing has also been supported by a number of studies (e.g. Kohl & Shea, 1995; Schmidt, Young, Swinnen, & Shapiro, 1989).

In Schmidt, Young, Swinnen, and Shapiro’s study (1989), participants were assigned to four conditions with different summary-KR lengths (1, 5, 10, and 15; summary length of 1 means that KR was given after every trial, summary length of 5 means that KR was given after every 5 trials) in a ballistic-timing task. Result revealed that increased summary length depressed performance during the acquisition state, while inverse relation between the summary length and performance was found in the retention tests. Kohl and Shea (1995) also found that subjects produced less error during acquisition period when auditory feedback was provided, but more error during retention.
Among the studies that investigated the effect of augmented feedback on learning, they mainly focused on the area of motor learning, such as limb motor learning (Schmidt & Wulf, 1997; Swinnen, Lee, Verschueren, Serrien, & Bogaerds, 1997) and voice motor learning (Steinhauer & Grayhack, 2000). This study is therefore motivated to investigate how augmented feedback would affect perceptual voice learning. The research questions are listed as follow:

1) Does the paired comparison training programs on perceptual rating of breathiness help to improve naïve listeners’ ability in detecting perceptual difference of breathiness?

2) Does providing augmented feedback during the paired comparison training programs degrade learner’s final performance when comparing to the no feedback condition?

Based on the result found in the study of Chan and Yiu (2006), it was hypothesized that the paired comparison training program would help improving naïve listeners’ ability on perceptual rating of breathiness. Furthermore, according to the guidance hypothesis emphasized by Schmidt and Wulf (1997), it was hypothesized that the no feedback group would demonstrate better performance than the feedback group in both post-training and review sessions.
Method

Stimuli

Synthesized female voice signals based on a Cantonese sentence /pa pa ta p/ (Father hits the ball) were used. The sentence was based on the prototype developed by Yiu, Murdoch, Hird, and Lau (2002). In order to eliminate the possibility of masking the aspirated noise with the noise from dysphonic voice qualities, all consonants included in the sentence were unaspirated stops. All voice stimuli were synthesized by a Klatt synthesizer, the HLSyn Speech Synthesis System from Sensimetrics (Cambridge, MA). To produce a set of breathy stimuli, the Klatt parameters “amplitude of aspiration” (AH) was adjusted systematically. Yiu et al. (2002) found that listeners consistently rated voice samples as breathy voice with AH manipulated. Therefore, in this study, stimuli that had AH manipulated were labeled as breathy. To contrast with breathiness, the “diplophonia” values (DI) were manipulated to create another dysphonic quality (rough-like quality) in this study. Though no studies have shown that the stimuli that had DI manipulated was labeled as roughness, it did not affect this study as no particular quality were labeled on the DI manipulated stimuli. By manipulating the fundamental frequency (Fo) of the signals, four sets of stimuli with different average fundamental frequency (Fo=200Hz, 220Hz, 240Hz, 260Hz) were produced. Each set included a non-dysphonic signal, five breathy signals, a non-breathy rough-like signal and five breathy rough-like signals. The level of breathiness in each of the four set of stimuli was manipulated by increasing the AH level in steps of 5 dB SPL (from AH 55 to AH 75). The values of the synthesis parameters of these 48 stimuli used in the rating tests and training session are shown in Table 1.
### TABLE 1.

Synthesis Values of the Stimuli Used in Rating Tests and Training Session

<table>
<thead>
<tr>
<th>Synthesis Values</th>
<th>Trained-set</th>
<th>Novel-set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fo 200</td>
<td>Fo 240</td>
</tr>
<tr>
<td>Prototype</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Breathy stimuli

| AH 55                | ✓           | ✓         | ✓         | ✓         |
| AH 60                | ✓           | ✓         | ✓         | ✓         |
| AH 65                | ✓           | ✓         | ✓         | ✓         |
| AH 70                | ✓           | ✓         | ✓         | ✓         |
| AH 75                | ✓           | ✓         | ✓         | ✓         |

Rough-like Stimuli

| DI04 Prototype       | ✓           | ✓         | ✓         | ✓         |
| DI04 AH 55           | ✓           | ✓         | ✓         | ✓         |
| DI04 AH 60           | ✓           | ✓         | ✓         | ✓         |
| DI04 AH 65           | ✓           | ✓         | ✓         | ✓         |
| DI04 AH 70           | ✓           | ✓         | ✓         | ✓         |
| DI04 AH 75           | ✓           | ✓         | ✓         | ✓         |

**Abbreviations:** AH, amplitude of aspiration; DI, diplophonia.

**Note:** Default values for prototype stimulus = DI 0 AH 40

**Participants**

Thirty men and 30 women with the mean age of 21.65 years (SD=1.95; range=19-26) were selected in this study. All participants had met the following criteria: a) native Cantonese speakers, b) had not received any training in perceptual voice evaluation at the time of testing, c) passed the hearing screening test (thresholds at 25dB or lower) and d)
passed the breathiness perception screening test. All participants were recruited on a voluntary basis and they were randomly assigned to three gender-balanced groups: the feedback group (Group F), the no feedback group (Group NF) and the control group (Group C.). For the control group, no training was provided and a listening session was introduced to them.

**Screening tests**

In the hearing screening test, pure tone audiometer were used to conduct a test of threshold at 25dB HL or lower at 250Hz, 500Hz, 1000Hz, 2000Hz, 4000Hz, 6000Hz, and 8000Hz. In the breathiness perception screening test, all participants were required to judge whether the severity of breathiness was identical in 16 pairs of stimuli. Each pair of stimuli either had the same level in AH or had a 10dB SPL difference in AH. The passing criterion was set at 80% accuracy or above.

**Procedures**

Each participant was tested individually over three rating sessions within a 7-day period. All participants took part in the first rating session (pre-training). Group F and Group NF received a training session immediately after the pre-training session. They were tested in the post-training session within two days and the third session (review) was conducted one week after the first session. The control group was tested with the same interval between sessions as the trained groups with the same set of tests being conducted. A listening session was provided for the control group instead of the training session.

All stimuli and responses of the participants were presented and recorded through the E-Prime computer program (Psychology Software Tools, Pittsburgh, PA). The stimuli were presented at a consistent intensity level (approximately 60dB SPL) through a pair of
headphones (Sennheiser, HD-25). A printed version of definition of breathiness was provided to the participants throughout each session. Breathiness was defined as “audible sound of expiration, audible air escape, and audible friction noise. It was found to correlate with the incomplete closure of the vocal folds or glottis during phonation” (Chan & Yiu, 2002, p. 117).

Pre-training session

The participants were required to judge whether the severity of breathiness was identical in each pair of stimuli. All participants were reminded to focus on breathiness only. Each pair of stimuli either had difference of 5dB in AH or had the same level of breathiness. The stimuli will be presented twice to the participants before prompting to answer. The session started with a practice trial for the participants to get familiar with the procedure. This session included 4 blocks, the first two blocks included the trained stimuli which were used in the training session and pre-testing session, and the last two blocks included novel stimuli that were not used in training or in testing. Each block consisted of 30 trials. It took around 40 minutes to complete.

Training session

The training program consisted of 5 blocks with increasing difficulty. In order to proceed to the next block, participants were required to achieve an accuracy of 80% or above in each block. The program would automatically repeat the failed block if the participants failed to reach the criterion. For the feedback group, knowledge of result (correct or wrong) was provided after each trial, while for the no feedback group, no knowledge of result was provided after each trial.
The first block included 15 pairs of stimuli. There were 7 identical pairs (e.g. AH 55 Fo 200 and AH 55 Fo200) and 8 pairs of stimuli with difference only in the severity of breathiness (e.g. AH 55 Fo 200 and AH 60 Fo 200). The second training block included 15 pairs of stimuli. There were 7 stimuli pairs with different fundamental frequency only (e.g. AH 55 Fo 200 and AH 55 Fo 240) and 8 stimuli pairs with different severity of breathiness and fundamental frequency (e.g. AH 55 Fo 200 and AH 60 Fo 240). In the third block, 15 pairs of stimuli were included. There were 7 stimuli pairs only differed in the presence of rough-like quality (e.g. AH 55 Fo200 and DI 04 AH 55 Fo 200) and 8 stimuli pairs differ in the severity of breathiness and the presence of a rough-like quality (e.g. AH55 Fo 200 and DI 04 AH 60 Fo 200). The fourth block also consisted of 15 pairs of stimuli. There will be 7 stimuli pairs differed in the presence of rough-like quality and fundamental frequency only (e.g. AH 55 Fo 200 and DI 04 AH 55 Fo 240), and 9 stimuli pairs differ in all the severity of breathiness, fundamental frequency and presence of rough-like quality (e.g. AH 55 Fo 200 and DI 04 AH 60 Fo 240). The last block included 32 stimuli pairs of stimuli. Eight stimuli pairs in each of the previous four blocks were selected for composing this block. The time for completing the entire training program was around 2 hours.

**Listening session**

For the control group, instead of providing training session, a listening session was introduced in order to provide similar stimuli exposure. Participants were only required to listen to the stimuli passively through a Power point program. Four blocks of stimuli were introduced to group C. The first block contained all breathy stimuli with the average fundamental frequency of 200Hz. The second block included all breathy stimuli with the average fundamental frequency of 240Hz. The third block contained all rough-like stimuli with the average fundamental frequency of 200Hz. The fourth block included all
rough-like stimuli with the average fundamental frequency of 240Hz. Each stimulus was presented six times in each block. The time for completing the entire listening program was around 15 minutes.

Post-training and review rating sessions
All tested stimuli in these two sessions were identical to the pre-training session. Each rating session included 4 blocks, the first two blocks included the trained stimuli which were used in the training session and pre-testing session, and the last two blocks included novel stimuli that were not used in training. Each block consisted of 30 trials. The time for completing each session was around 40 minutes.

Data analysis
The overall percentage of accuracy and the intra-rater exact agreement of the participants’ responses were obtained. Responses that correctly judge whether the AH level of the stimuli pairs were identical were considered as accurate. Furthermore, each test stimulus was repeated twice for the calculation of the intra-rater exact agreement.

Two 2-way repeated MANOVAs were used to analyze the overall accuracy and the intra-rater agreement to determine the effect of training and feedback in facilitating perceptual evaluation of breathiness. The three-levels variable ‘session’ (pre-training, post-training, and review) were treated as the within-group factors and the three-levels variable ‘group’ (Group F, NF and C) were treated as the between group factor. Whenever there are significant main effects found, post hoc comparisons with Bonferronni adjustment was conducted to specify the source of the statistically significant main effect.
Result

The mean accuracy would be reported, followed by the intra-rater agreement. A two-way repeated MANOVA would be conducted for each set of data. In addition, to maintain the overall chance level for each set of statistics at 0.05, the alpha for each test was recalculated so as to avoid possible Type I error.

**Overall Accuracy**

A one-way ANOVA was carried out to compare the overall accuracy across the novel and trained stimuli across each rating session. It was found that there were no significant differenced found between novel and trained stimuli across each rating session \[F(1, 59)=1.51, p=0.23\]. The result showed that the two sets of stimuli (novel and trained) were similar. Therefore, the total overall accuracy by combining the data from novel and trained stimuli would be used for the following data analysis.

The mean accuracy and standard deviation of each rating test are shown in Table 2. A two-way repeated MANOVA was carried out. Two main effects and one two-way interaction effect are reported here.
### TABLE 2.
Mean Accuracy and Standard Deviation Across the Sessions

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-training session (Mean %) (SD)</th>
<th>Post-training session (Mean %) (SD)</th>
<th>Review session (Mean %) (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>69.76 (8.77)</td>
<td>79.23 (7.56)</td>
<td>79.84 (7.84)</td>
</tr>
<tr>
<td>NF</td>
<td>70.26 (8.45)</td>
<td>74.97 (8.09)</td>
<td>76.25 (9.24)</td>
</tr>
<tr>
<td>C</td>
<td>70.12 (6.05)</td>
<td>72.37 (8.98)</td>
<td>71.36 (7.13)</td>
</tr>
</tbody>
</table>

*Abbreviations:* F, feedback group that received feedback in training sessions; NF, no feedback group that received no feedback in training sessions; C, control group that did not receive any training.

**Main ‘session’ effect**

The main ‘session’ effect, comparing the overall accuracy across the three rating sessions (pre-training, post-training and review) was significant [F(2, 56)=23.03, \(p<0.0001\); mean accuracy pooled across the three sessions: pre-training=70.05%, post-training=75.52%, review=75.82%]

**Main ‘group’ effect**

The main ‘group’ effect, comparing the overall accuracy across the three participant groups, was not significant [F(2, 57)=2.51, \(p=0.09\); the mean accuracy pooled across the three group: Group F=76.28%, Group NF=73.83%, Group C=71.28%].

**Two-way ‘session’ by ‘group’ interaction effect**

The ‘session’ by ‘group’ interaction effect was significant [F(4, 114)=4.93; \(p=0.01\)]
Comparison across the sessions

In order to determine how each group of participants improved over the session, separate repeated ANOVAs were carried out for each participant group, with ‘session’ as the within group factor. For group F, ‘training’ effect was significant [F(2,38)=25.46, p<0.0001]. Post hoc comparisons with Bonferroni adjustments showed that the differences between the pre- and post-training sessions and between the pre-training and review session, however, the accuracy in the post-training session was not significantly different from those in the review session (Table 3). For group NF, ‘training’ effect was also significant [F(2, 38)=9.68, p<0.005]. Post hoc comparisons showed that the differences between the pre- and post-training sessions and between the pre-training and review session (p<0.005), however, the accuracy in the post-training session was not significantly different from those in the review session (p<0.005). For group C, no significant ‘training’ effect was found [F(2, 38)=1.05, p=0.36].

TABLE 3.
MANOVAs and Post Hoc Tests Comparing the Accuracy in Each Test Across the Sessions. (Analyzed Separately by Each Participant Groups)

<table>
<thead>
<tr>
<th>Training type</th>
<th>MANOVA</th>
<th>Pre vs. Post</th>
<th>Pre vs. Review</th>
<th>Post vs. Review</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F (2,18)</td>
<td>p</td>
<td>t (19)</td>
<td>p</td>
</tr>
<tr>
<td>Feedback</td>
<td>19.80*</td>
<td>&lt;0.005</td>
<td>36.98*</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>No feedback</td>
<td>9.43*</td>
<td>&lt;0.005</td>
<td>11.24*</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>Control</td>
<td>0.72</td>
<td>0.50</td>
<td>1.52</td>
<td>0.23</td>
</tr>
</tbody>
</table>

*Significant at 0.005 level (nine comparisons were carried out, 0.05/9=0.005).
Comparison across the participant groups

A set of analyses were carried out to compare the accuracy across the participant groups in each rating session. Only the group differences in the review session were significantly different among the three comparisons. \((p=0.007; \text{Table 4})\). Post hoc comparison with Bonferroni adjustments showed that difference between Group F and Group C was significant \((p=0.005)\), while the differences between Group F and Group NF \((p=0.50)\) and Group NF and Group C \((p=0.19)\) were not significant.

**TABLE 4.**

ANOVA Comparing the Accuracy across the participant groups in each rating session.

<table>
<thead>
<tr>
<th>Sessions</th>
<th>F (2,57)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-training</td>
<td>0.02</td>
<td>0.98</td>
</tr>
<tr>
<td>Post-training</td>
<td>3.54</td>
<td>0.04</td>
</tr>
<tr>
<td>Review</td>
<td>5.50*</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

*Significant at 0.017 level (three comparisons were carried out, 0.05/3=0.017)

Intra-rater agreement

A one-way ANOVA was carried out to compare the intra-rater agreement across the novel and trained stimuli across each rating session. It was found that there were no significant differences found between novel and trained stimuli across each rating session \([F(1, 59)=2.43, p=0.13]\). The result showed that the two sets of stimuli (novel and trained) were similar. Therefore, the overall intra-rater agreement by combining the data from novel and trained stimuli would be used for the following data analysis.
The mean intra-rater agreement and standard deviation of each rating test are shown in Table 5. A repeated measure ANOVAs was carried out. Two main effects are reported here.

**TABLE 5.**

**Mean intra-rater agreement and Standard Deviation Across the Sessions**

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre-training session</th>
<th>Post-training session</th>
<th>Review session</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>70.50 (6.78)</td>
<td>74.25 (9.26)</td>
<td>74.42 (9.19)</td>
</tr>
<tr>
<td>NF</td>
<td>70.75 (5.44)</td>
<td>72.59 (6.96)</td>
<td>74.33 (7.79)</td>
</tr>
<tr>
<td>C</td>
<td>72.59 (7.00)</td>
<td>73.92 (6.27)</td>
<td>71.75 (8.16)</td>
</tr>
</tbody>
</table>

*Abbreviations:* F, feedback group that received feedback in training sessions; NF, no feedback group that received no feedback in training sessions; C: control group that did not receive any training.

**Main ‘session’ effect**

The main ‘session’ effect, comparing the intra-rater agreement across the three rating sessions was not significant [F(2, 56) = 2.12, *p* = 0.13; the mean agreement pooled across the three sessions: pre-training = 71.28%, post-training = 73.59%, review = 73.50%].

**Main ‘group’ effect**

The main ‘group’ effect, comparing the intra-rater agreement across the three participant groups was also not significant [F(2, 57) = 0.04, *p* = 0.96; the mean agreement pooled across the three sessions: Group F = 73.06%, Group NF = 72.56%, Group C = 72.75%].
Discussion

The aim of the present study was to explore the effect of augmented feedback on the effectiveness of a paired comparison perceptual voice rating training program. The main ‘session’ effect was significant for both the trained and the novel stimulus sets (Table 3), which indicated that significant improvement on the overall accuracy was noted across the three participant groups after training. The significant two-way ‘session’ by ‘group’ interaction effect suggested that the type of training also affect the rate of improvement across sessions. For the trained groups (Group F and Group NF), improvement across the sessions was significant in rating both the trained and novel stimulus. For the control group, no significant improvement was found across the sessions in rating both the trained and novel stimulus (Table 4). From this result, it can be concluded that trained listeners (for both feedback and no feedback group) demonstrated better improvement across the session than the control listeners. It indicated that the paired comparison perceptual voice training programs (for both with feedback and without feedback) were effective in improving the perceptual skills of the listeners in detecting 5-db differences in synthesized amplitude of aspiration. In addition, no significant ‘session’ and ‘group’ effect was found when comparing the intra-rater agreement across the three participant groups across the three sessions. This indicated that no significant improvement on the intra-rater agreement was noted after the training. The effect of each type of training will be discussed in the following sections.

Feedback vs. no feedback

This study demonstrated comparative accuracy in all participant groups in the pre-training rating session, all of them were around 70%. (Table 2). The ANOVA comparing the accuracy across the participant groups in pre-training rating session showed no significant differences between the three participant groups, this further highlighted the difference of
performance pattern between the training groups and the control group across the three sessions. Therefore, it is suggested that both training with feedback and training without feedback were effective in facilitating voice perceptual ability in detecting subtle differences in breathiness in inexperienced listeners.

However, only Group F showed significantly higher accuracy than Group C in the review sessions. No significantly higher accuracy was found in Group NF than Group C in the review sessions. Though no significantly higher accuracy was found in Group F than Group NF in both the post-training and review sessions, by comparing the performance pattern of Group F and Group NF, Group F demonstrated higher overall accuracy after training (Group F: Post-training = 79%, review = 80%; Group NF: Post-training = 75%, review = 76%). This pattern further suggested that the effectiveness of training without feedback was not better than that of training with feedback.

**Clinical implications**

The result of this study agrees with the result found in Chan and Yiu’s study (2006), which is that paired-comparison training with augmented feedback provided was effective in improving the perceptual ability of naïve listeners in detecting subtle breathiness. This study further proposed that paired-comparison training without feedback provided was also effective in training naïve listeners to perceive breathiness. After a maximum of 2 hours of training, all mean accuracies in both training groups across the post-training and review sessions were over 75% (Table 2). This result further supports Chan and Yiu’s view (2006) that perceptual voice evaluation skills of naïve listeners could be improved through training. Therefore, paired-comparison training is suggested in clinical application for improving clinician’s voice perceptual skill.
Theoretical implications

This study found that both training with augmented feedback and without augmented feedback were effective in training inexperienced listeners to perceive subtle differences of breathiness. Both training groups demonstrated similar generalization to the novel stimuli. However, in comparing the mean accuracy with that of control group, the training with feedback was more effective than that without feedback in both post-training and review sessions.

The good retention performance of the feedback group in this study disagreed with the guidance hypothesis for feedback processing emphasized by Schmidt and Wulf (1997). They suggested that knowledge of result (KR) serves as a guidance role, preventing learners from making error. With the presence of KR, learners would heavily rely on it to produce the correction action during practice, keeping good performance during the practice sessions. As a result, the reliance on feedback would shift learners’ attention from acquiring the necessary capability to deal with the intrinsic information on the retention or transfer test, so performance fall when augmented feedback is no longer available. However, in viewing the result of this study, the feedback group demonstrated good performance in both post-training and review rating tests, which no feedback was provided.

Overall, according to the result, the hypothesis that feedback improves performance was confirmed. This study supported that Kluger and DeNisi’s view (1996) on the shifting of attention in the preliminary feedback interventions theory (FIT). When participants failed to identified the difference of the breathiness level of the paired stimuli consistently under feedback conditions, sustained feedback-standard discrepancy was presence. In this situation, additional effort such as listening to the stimuli repeatedly would be deemed
insufficient for improvement. At this point, attention will be diverted to the lower task-learning processes, motivating learners to focus more on the task, and develop new specific strategies to help discriminating the subtle breathiness level difference. For instance, participants may develop the new strategies of just focusing on the aspirated part of the sound, ignoring the vowels, consonants and tones of the word. This view of learning is consistent with Vollmeyer and Rheinberg’s (2005) cognitive-motivational process model. They suggested that providing feedback to learners help increase the subsequent use of systematic strategies and positive motivation. Moreover, implementing feedback would also provide more knowledge to learners such as giving hints for learners to evaluate the effectiveness of their new developed strategies. As a result, by applying their knowledge, effective strategies would be developed and learners should perform better and eventually their final performance would be improved. These theories can be accounted for the sustained improved performance of group F in the post-training and review sessions in the present study.

Kluger and DeNisi (1996) had also identified different moderators, such as cues, task characteristics and personality, which may affect the effect of feedback on learning. They concluded that by regulating the moderators in certain situations, large and positive effect on performance would be yield by feedback interventions (FIs). In the present study, though group F demonstrated improved performance in this study, no significant differences were found between the performance in Group F and Group NF in the post-training and review sessions, suggesting that moderators were not adjusted to the most favorable situations for yielding great positive effect on performance. They also suggested that both salient negative and salient positive FI may shift attention to self, leading to the shifting of attention from the task diverts cognitive resources to non-task aspects of the meta-task processes such as paying attention to the self. Therefore,
performance loss may result due to the competition for cognitive resources. Throughout the present study, 100% continuous augmented feedback was introduced. With this intensive feedback situation, it is proposed that there would be higher chance that sustained salient positive or salient negative feedbacks would be resulted, and hence decreases the performance. For these reasons, provision of feedback in voice perceptual training is suggested for future clinical application. However, the frequency of feedback should be further adjusted for yielding greater positive effect on performance.

In addition, no significant difference was found when comparing the intra-rater agreement across the three participant groups throughout the three sessions. It was found that naïve listeners demonstrated relative high intra-rater agreement (around 71%) during the pre-training session. Therefore, there would be little room for significant improvement basing on such high starting point. The high intra-rater agreement obtained through the sessions across the three groups is suggested to be an advantage of the method design in the paired-comparison training program. For this reason, paired-comparison training program would be suggested to be the more appropriate training methodology for perceptual voice evaluation under clinical conditions.

**Limitations and further studies**

In the presence study, effect of augmented feedback was found by comparing the 100% concurrent feedback group and the no feedback group only, effects of frequency and distribution of feedback on the effectiveness of perceptual voice training cannot be shown. It has been believed that frequency and distribution of augmented feedback affect performances (Schmidt & Bjork, 1992). Studies showed that frequent augmented feedback adversely affected performances (Winston & Schmidt, 1990; Wulf & Schmidt, 1989). Other studies showed that learning can be facilitated by reducing the presentation
of feedback in a fading way. (Schmidt & Wrisbery, 2000; Schmidt & Lee, 1999). Further research on the effects of distribution and frequency of feedback on the effectiveness of perceptual voice training should be carried out to find out the whole picture of how augmented feedback affect the effectiveness of training in the aspect of perceptual voice evaluation. Moreover, this study only focused on the perceptual evaluation of female stimuli. In reference to Chan and Yiu (2002), different training effects between female stimuli and male stimuli were found. Further studies may also compare the difference of feedback and training effects between male and female stimuli. In addition, breathiness is only one type of different voice qualities. Studies focus on other voice qualities such as roughness would be the future direction of research to find out a clearer picture on the effects of augmented feedback to the effectiveness of perceptual voice training.

Conclusion

Significant improvement was found in naïve listeners after training, with both the paired-comparison trainings with augmented feedback and no feedback. In addition, it was found that both feedback group and no feedback group maintained the improved performance throughout the post-training and review sessions, with feedback group showed better performance. This result disagreed with the guidance hypothesis that feedback group demonstrated poorer retention performance than the no feedback group. On the other hand, this study supported Kluger and DeNisi’s (1996) feedback interventions theory (FIT) and Vollmeyer and Rheinberg’s (2005) cognitive-motivation process model that augmented feedback provided information and shifted learners’ attention to the task so as to help learners to develop a more systematic strategy, resulting better retention of performance.
Reference


**Acknowledgement**

I would like to express my sincere gratitude to Dr. Karen Chan for her guidance and support during the course of this dissertation. I would like to thank Professor Edwin Yiu and Professor Tara Whitehill for their advice in the development of proposal. Thanks also go to Raymond Wu the technician for his help during data collection. Last but not least, I would like to thank all the participants who made this dissertation possible.