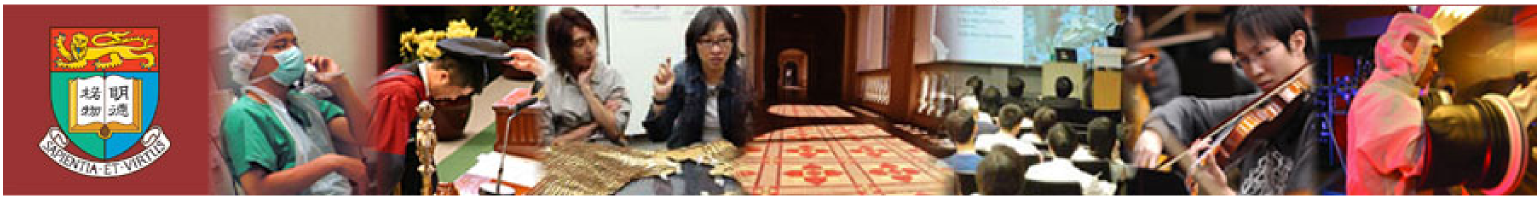

Title	The effectiveness of feedback in training perceptual voice evaluation
Other Contributor(s)	University of Hong Kong.
Author(s)	Li, Wing-sang, Margaret; 李穎生
Citation	
Issued Date	2006
URL	http://hdl.handle.net/10722/50071
Rights	The author retains all proprietary rights, such as patent rights and the right to use in future works.

The Effectiveness of Feedback in Training Perceptual Voice Evaluation

Li Wing Sang Margaret

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Sciences (Speech and Hearing Sciences), The University of Hong Kong, April 30, 2006.

Abstract

The study investigated the effect of training with or without feedback on perceptual voice evaluation. Forty naive listeners randomly assigned to the feedback group or the no feedback group, took part in a training session, a pre-training, a post-training and a review rating sessions involving the reference matching tasks. Feedback group received the correct answer as visual feedback during training. No feedback group received no feedback. Measures of the accuracy and intra-rater agreement were obtained from the rating sessions. The result showed that training with and without feedback had similar effectiveness in improving the reference matching ability of the listeners. The effect of feedback in training perceptual voice evaluation was discussed.

Perceptual voice evaluation is widely used by clinicians to determine the presence and the severity of voice quality impairment and to evaluate the treatment outcome. However, perceptual voice evaluation is a subjective process and the intra and inter rater reliability may be highly variable. Perceptual voice evaluation may subject to variability due to the lack of common understanding of perceptual label of the listeners, the inability to discriminate single perceptual dimensions from complex stimuli and the difficulties for listeners to maintain both the within and across judges consistent judgment (Kent, 1996). The biased and variable nature of perceptual voice evaluation may directly affect the clinical diagnosis and treatment of voice disorder.

The reliability of perceptual voice quality evaluation has been studied extensively in recent studies (e.g., Gerratt & Kreiman, 2001; Carding, Carlson, Epstein, Mathieson, & Shewell, 2000; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Kreiman, Gerratt, Precoda, & Berke, 1992). Methods to improve the reliability of perceptual voice evaluation are also proposed. These include the provision of references or anchors, in which anchors sample are provided for the listeners as external standards that the listener can use to compare with the to-be-rated stimuli (e.g. Chan & Yiu, 2002; Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993), analysis by synthesis techniques that listener are required to vary the speech synthesis parameters to a synthesize signal to match the to-be-rated stimuli (Gerratt & Kreiman, 2001), application of psychometric principles to auditory perceptual voice scaling approaches (Shrivastav, Sapienza, & Nandur, 2005) and listeners' perceptual training (e.g. Martin & Wolfe, 1996, Chan & Yiu, in press).

Provision of anchors as external standards, analysis by synthesis and training are based on the theoretical framework proposed by Kreiman et al. (1993). According to Kreiman et al. (1993), listeners form unstable and idiosyncratic internal standards of voice qualities through exposure to voices with different qualities and these internal standards are stored in the memory (Gerratt, Kreiman, Antonanzas-Barroso & Berke, 1993). During perceptual voice evaluation, those internal standards are retrieved so as to compare the to-be-rated stimuli with these internal standards for judgment. However, Kreiman et al. (1993) suggested that the internal standards are unstable that can be easily affected by acoustic context as well as the listeners' experiences with voices. For example, listeners with different experiences in rating pathological voice may share different internal standards for pathological voice qualities. These lead to the poor reliability of perceptual voice evaluation. In order to improve the reliability of the perceptual voice evaluation, use of fixed external standard to counteract the effects of the unstable internal standards was proposed (Gerratt et al., 1993). External standards are set as anchors or references for the listeners to compare with the to-be-rated stimuli. Studies have shown that external reference might replace the unstable internal standard and thus lead to a relatively more reliable evaluation. In Chan and Yiu (2002), the intra-rater agreement of the listeners to rate natural stimuli improved from 59% without any reference to 76% when a synthesized reference was given.

Another method that has been proposed to improve the reliability of perceptual voice evaluation is to provide training to the rater. Studies suggested that training help to consolidate the internal representation of different pathological voice qualities to make them more stable and thus lead to a more reliable evaluation (e.g. Eadie & Baylor, in press; Chan & Yiu, in press; Chan & Yiu, 2002; Martin & Wolfe, 1996). Martin and Wolfe (1996) trained 28 naive listeners to discriminate pairs of synthesized stimuli in the categories of breathy, rough and hoarse. During training, listeners were required to indicate the sample that was more deviant within a pair of stimuli and the correct answer was given as feedback after every trial. Martin and Wolfe (1996) found that listeners showed an average improvement score of 28 % in the classification of synthesized voice signals after training. However, no control group was included in the study and it was difficult to make conclusion on the effectiveness of the training. In the study by Chan and Yiu (2002) and Chan and Yiu (in press), correct answer was used as feedback in the training program. In these studies, a stimulus-response-feedback-stimulus training paradigm was adopted. The participants were required to listen to a stimulus and then make a response. The correct answer was then given as feedback and the participants were required to listen to the stimulus again to complete a trial (Chan & Yiu, 2002). Listeners were also provided with the definitions of rating dimensions and anchor samples (references) during the training. Chan and Yiu (2002) trained

20 naive listeners to rate roughness and breathiness in a set of natural stimuli with the use of synthesized anchors using visual analog scales. They found that the listeners' intra-rater agreement improved from 63% in pre-training to 73% after training when external references were given. However, no control group was included in the study to document the training effect. In the study by Chan and Yiu (in press), the effectiveness of two training program, the paired comparison training program and the reference matching training program was compared. The pair comparison training program aimed to train the listeners to detect subtle perceptual changes in the aspiration noise. Listeners were required to compare a pair of synthesized stimuli and to judge whether the severity level of breathiness was identical. Reference matching training program was a referenced method to train the listeners to be familiar with a set of synthesized references. Listeners were required to match each training stimulus with one of the set of references provided. Sixty participants were randomly assigned into the paired comparison training group, the reference matching training group and the control group. Participants were required to take part in three rating sessions within a seven-day period. The found that trained listeners had better improvement across the sessions than the control listeners. Listeners could perceive breathiness in synthesized sentences with almost 80% accuracy after two hours of training. This suggested that both the paired comparison and reference matching training programs were effective in improving the ability of naïve listeners to perceive perceptual difference in breathiness.

Within the training programme of the above studies, the correct answer was given as feedback to facilitate perceptual learning of voice qualities or different levels of severity of a voice quality through exposure to the particular voice samples. In the field of motor learning, Sparrow and Summers (1992) and Newell (1991) suggested that feedback enhance and accelerate motor learning of certain skilled movement as it guide the learner toward the target goal. However, numerous studies were against the effectiveness of feedback on different motor learning tasks. For example, some suggested that feedback maybe a redundant information in learning a specific motor tasks as the learners may be able to detect their own errors with the assistance of a reference- of –correctness that enabled them to evaluate their own performance. It is suggested the reference may be established during the experiment through visual and verbal information or it was already available from the previous experiences (Magill, Chamberlin, & Hall, 1991). Schmidt and Wulf (1997) proposed the guidance hypothesis, according to which feedback can guide the learner to the correct response and thus enhances acquisition performance but it can also distract or inhibit learners to pay attention to other information that may important to the retention of the performance such as the information that is required to develop the intrinsic error detection and correction mechanism. In the study by Park, Shea and Wright (2000), participants are required to reproduce a criterion force-production waveform presented on a computer screen. The results indicated the strong guiding effects of feedback illustrated by the boosted performance during acquisition. However, poor performance in retention when feedback was withdrawn also indicated the over reliance on the guidance from feedback that cause detriment in learning.

Voice researchers have also studied the role and the effectiveness of feedback in vocal motor learning (e.g. Yiu, Verdolini & Chow, 2005; Steinhauer & Grayhack, 2000; Ferrand, 1995) Ferrand (1995) studied the effects of practice with or without feedback on phonatory stability on the level of jitter and shimmer during production of vowel /a/ prolongation. Thirty women were randomly assigned into two practice group. Participants were required to take part in a baseline session, two practice sessions which were two days apart for each participants and a final transfer session which was identical to the baseline session seven days after the second practice. The feedback group received visual and verbal feedback during the practice sessions while the no feedback group received no feedback. Visual feedback consisted of watching the waveform of /a/ on a display screen during each prolongation. Verbal feedback involved the discussion of the jitter and shimmer values that was obtained after each prolongation with the investigator. The findings in the study suggested that practice with feedback was effective in increasing vocal motor stability during the practice sessions. However, practice without feedback was more effective in facilitating the carry over effects and retention of the tasks over the longer term. These findings support the guidance hypothesis that strong guidance provided by feedback facilitate immediate performance but degrades learning assessed by the retention tests with no feedback.

The effect of training in perceptual voice evaluation and the effect of feedback on motor learning including the vocal production tasks have been investigated in the previous studies. However, little is known about the role of feedback in facilitating perceptual learning in perceptual voice evaluation. No studies have directly investigated the effectiveness of feedback in training perceptual voice evaluation.

The present study aimed to investigate the effect of training with and without feedback in facilitating the ability of naive listeners to match severity levels of breathiness with the use of references. According the findings by Ferrand (1995), training without feedback was found to be more effective to facilitate learning and retention of the learned skills. Although these findings were based on vocal production tasks, it was adopted in the present study to determine if it may be applicable to the modality of auditory perceptual learning in training perceptual voice evaluation. It was hypothesized that training without feedback would further enhance naive listeners’ ability in matching the severity levels of breathiness with one of the references. The findings would help to provide information about the effectiveness of feedback in training perceptual voice evaluation.

Method

Participants

Twenty male and 20 female with the mean age of 21.9 years (*SD* = 5.3; range = 18-49 years) were recruited in this study. All were native Cantonese speakers and had not received any training in voice disorders or perceptual voice evaluation at the time of testing. The participants were recruited on a voluntary basis within the undergraduate students' community. The participants were randomly assigned to two gender-balanced group: a feedback group (Group F) and no feedback group (Group NF).

Stimuli

The stimuli used in this study were based on the stimuli used in the study by Chan and Yiu (2002). All stimuli were synthesized female voice signals. They were based on a Cantonese sentence /pa pa ta p□/ ("father hits the ball"), which were the prototype developed by Yiu, Murdoch, Hird and Lau (2002). The voice stimuli were synthesized using a Klatt synthesizer, the *HLSyn Speech Synthesis System* from Sensimetrics (Cambridge, MA) (Klatt & Klatt, 1990). A set of breathy and rough like signals were produced by adjusting the Klatt parameters, "amplitude of aspiration" (AH) and "diplophonia" (DI), respectively.

The value of the synthesis parameters of the stimuli used in the rating tests and training session are shown in Appendix A. The fundamental frequency (F_0) of the signals was manipulated to produce four sets of stimuli with different average fundamental frequency (F_0 = 200Hz, 220Hz, 240Hz, 260Hz). Each set included a non-dysphonic signal and five breathy signals. The 200-Hz and 260-Hz sets also included a non-breathy rough –like signal and five breathy rough-like signals. The rough-like signals were included to create another dysphonic quality to contrast with breathiness. If only breathiness quality exist, listeners may focus on the perceptual feature other than breathiness, such as loudness in the rating. The inclusion of the rough quality to contrast with breathiness helps listeners to focus on the perceptual feature of breathiness. 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) resulting in 36 stimuli.

Procedure

All participants were required to pass a discrimination screening test and a hearing screening test before they took part in i) three rating sessions and ii) a training session within a seven-day period. All screening tests, rating and training sessions were carried out in a sound-treated booth.

Screening procedure

In the discrimination screening test, participants were required to judge whether the severity of breathiness was identical in 16 pairs of stimuli. The stimuli were identical to some of the stimuli used in training. Each pair of stimuli were either had 10 dB SPL difference in AH (e.g., F_0 260 AH55 were paired with F_0 260 AH65) or had the same level in AH (e.g. F_0 240 AH75 were paired with F_0 240 AH75). A 10 dB SPL difference was selected as the level of breathiness was manipulated by increasing the AH level in steps of 5 dB SPL (from AH 55 to AH 75) in the experiment. Participant should at least be able to discriminate 10 dB SPL difference before they were trained to detect the 5 dB SPL differences in breathiness throughout the experiment. The passing criterion was accuracy of 80% or above. This was to ensure a similar minimum ability to perceive differences in breathiness by all participants. In the hearing screening test, test of threshold at 25 dB or lower at 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, 6000 Hz, and 8000 Hz were conducted using a pure tone audiometer. This was to ensure the participants had normal hearing.

Rating and training sessions

All participants were then took part in first rating session (pre-training) as a baseline measurement. They were given a training session immediately after the pre-training rating session. The participants were tested two days after the first session (post-training) and one week after the first session (review).

During the rating and training sessions, participants were asked to match a target stimulus with one of the six references given in a reference matching program that was adopted from the study by Chan and Yiu (in press). A non-dysphonic stimulus and five breathy stimuli were included as the six references. The breathy references were increased in steps of 5 dB SPL in AH (from AH 55 to AH 75). Each target stimulus and the references were presented as a graphic icon on a page of the program. The reference stimuli were labeled as stimulus *0* to *5* with increasing level of breathiness. Participants were required to listen to the target stimulus and the references before selecting the reference that match with the target stimulus by clicking on the appropriate icon as response. During rating sessions, the target stimuli were not always identical to one of the references. They may differ in terms of the fundamental frequency or in the presence of the rough-like quality. In the training session, the target stimuli were always identical to one of the references. The participants could listen to the target stimulus and the breathy references as many times as they wish in each trial by clicking on the appropriate icon. In addition, the participants were required to listen to all references in every four trials in order to encourage the participants to judge based on the references but not based on memory.

The stimulus presentation and response collection were controlled by a specifically designed stand-alone computer programs based on Microsoft excel through a Genie-IV Intel Pentium III 533MHz computer. The stimuli were presented through a pair of headphones (Sennheiser, Wedemark, Germany; HD-25) at a consistent intensity level. Participants were provided with a printed version of the definition of breathiness in English at the beginning of each session. This information was available throughout the sessions. Breathiness was defined as audible sound of expiration, audible air escape, and audible friction noise. Its physiological correlation is incomplete closure of vocal folds or glottis during phonation. (Chan & Yiu, 2002)

Training procedure

The aim of the training program was to train the participant to become familiar with the breathy references. For Group F, a stimulus-response-feedback-stimulus paradigm that was adopted from the study of Chan and Yiu (in press) was used. The participants were required to listen to a training stimulus and then match each training stimulus with one of the six references. The correct answer was shown as on the computer screen as visual feedback by showing the labeled number of the reference that match the training stimuli. The training stimuli were always identical to one the six references. The participants were required to listen to the stimulus once again to finish the trial. The training program for Group NF was identical to Group F except no feedback was given.

The training program was divided into three blocks of 36 stimuli each. The presentation order of the training blocks and the stimuli in each block was randomized for each participant. Two blocks consisted of one non-dysphonic stimulus and five breathy stimuli with fundamental frequency of 240 Hz and 260 Hz, respectively. Another block consisted of one non-breathy rough-like stimulus and five breathy rough-like stimuli with fundamental frequency of 260 Hz. Each reference was repeated six times in each training block resulting in 36 stimuli. The participants were required to reach 80% accuracy in each block in order to move to another block or the program would repeat the same failed training block automatically. The accuracy of response in percentage was shown at the end of each training block on the computer screen. Four participants had to repeat and two of them had to repeat one of the training blocks once and the other two had to repeat one of the training blocks for twice. The participants were encouraged to take a break if necessary within the training session. It took approximately one hour for each participant to complete the training program.

Rating procedure

The pre-training rating session aimed to measure the baseline performance of the participants. Listeners were informed to start with a practice trial at the beginning of the test .They were told to be familiarized with the rating procedure through the practice trail at the beginning followed by 36 test trials using the training stimuli in Appendix. Eighteen target stimuli were repeated twice to obtain the intra-rater agreement for each participant resulting in 36 test trials. This session took about 15 minutes to complete.

The purpose of the post-training and review sessions were to measure how well the participants learn from the training program and how well they maintain the performance. These two rating sessions were identical. Each of them made up of two blocks. One block was identical to the pre-training session. The other block consisted of stimuli and a set of references that were not used in training (ie, novel stimuli). Each block consisted of 36 trials. Each session lasted for approximately 30 minutes.

The presentation order of all stimuli and test blocks were randomized across participants and across the three rating sessions to counterbalance any possible memory and learning effects from the order of presentations.

Data analysis

The accuracy of the response of the participants was calculated in each rating test to determine the ability of the participants to match the severity level of breathiness of the to-be-rated stimulus with one the references provided. Response that can match the level of AH of the stimulus with the corresponding reference will be considered as an accurate response.

As each test stimulus was repeated twice in each set of the testing stimuli, the percentage of intra-rater agreement of the participants was also calculated to determine the agreement in rating two identical stimuli of each participant in each rating tests. Two rating that are identical to each other were considered to agree with each other. The intra-rater agreement was taken as a measure of the consistency and thus the reliability of each participant's performance in perceptual voice evaluation.

As there is no rating of novel stimuli in the pre-training session, a three-way analysis of variance (ANOVA) cannot be performed. Since the main focus of this study was not to investigate the participants' difference in matching the trained and novel stimuli, the data obtained in the trained and novel stimuli would be averaged if participants show similar performance towards these two stimuli types. Separate match sample *t* tests would be performed on the accuracy of response and the percentage of intra-rater agreement for each participant group in each rating sessions.

Two-way analysis of variance (ANOVA) with repeated measure was used to analyze the data on the accuracy of response and the percentage of intra-rater agreement to determine the effectiveness of feedback in facilitating the ability of naïve listeners to match the severity level of breathiness of the to-be-rated stimulus with one the references provided.

The two-level variable “group” (Group F or Group NF) was treated as between group factor. The three-level variable “session” (pre-training, post-training and review) were treated as within group factor. Post hoc comparisons with Bonferonni adjustment was conducted when the main effects was significant to specify the source of the statistically significant main effect. Because three comparisons were carried out, the alpha level of each test was recalculated and set at .0167 (0.05/3).

Result

Trained vs. novel stimuli

No significant differences were found between the trained and novel stimuli in the accuracy of response and the percentage of intra-rater agreement for any participants groups in any rating sessions (p > .05 for all matched sample t tests).

Participants showed similar performance on the accuracy of response and the percentage of intra-rater agreement towards the trained and novel stimuli in the rating tests. Therefore, data on the trained and novel stimuli were averaged for each participant in each rating sessions in the following analysis. Two two-way ANOVAs with repeated measure were carried out with one for the accuracy of response and one for the percentage of intra-rater agreement.

Accuracy of response

The mean accuracy and standard deviation is shown in Table 1.

Table 1.

Mean accuracy of response and standard deviation across the sessions

	Mean Accuracy (%) (*SD*)		
Group	Pre-training	Post-training	Review
Group F	72.79(14.98)	86.11(7.48)	82.01(10.90)
Group NF	72.22(13.78)	84.79(8.29)	84.23(11.34)

The main session effect, which compared the accuracy across the three sessions, was significant, *F*(2, 37) = 15.80, *p* = < .0001. Post hoc comparisons with Bonferonni adjustment showed that significant improvement occurred between the pre-training and post-training session, *t*(39) = -5.39, *p* = <.0001, and between the pre-training and review session, *t*(39) = -4.09, *p* = <.0001. The highest accuracy attained was in the post-training session (See Table 1). However, the drop in performance between the post-training and review session was not significant, *t*(39) = 2.33, *p* = .025.

The main between group effect compared the overall accuracy between the two groups of participants. The main effect was not significant, *F*(1, 38) = 0.34, *p* > .05. The result showed that there was no significant difference between the performance of Group F and Group NF across the sessions.

The session by group interaction effect was not significant, *F*(2, 37) = 2.37, *p* >.05. The learning pattern for the two participant groups across the sessions was similar.

Intra-rater agreement

The mean percentage and standard deviation is shown in Table 2.

Table 2.

Mean percentage of intra-rater agreement and standard deviation across the sessions

	Mean Percentage (%) (*SD*)		
Group	Pre-training	Post-training	Review
Group F	72.51(10.41)	81.67(10.48)	75.84(12.74)
Group NF	74.44(11.58)	82.22(9.24)	78.75(14.13)

The main session effect, which compared the percentage of intra-rater agreement across the three sessions, was significant, *F*(2, 37) = 15.38, *p* < .0001. Post hoc comparisons with Bonferonni adjustment showed that significant improvement occurred only between the pre-training and post-training session, *t*(39) = -5.17, *p* = <.0001. The improvement occurred between the pre-training and review session was not significant, *t*(39) = -1.84, *p* = .074.

The main between group effect compared the overall accuracy across the two groups of participants. The main effect was not significant, *F*(1, 38) = 0.36, *p* > .05. The result showed that there was no significant difference in the percentage of the intra-rater agreement between Group F and Group NF across the sessions.

The session by group interaction effect was not significant, *F*(2, 37) = 0.33, *p* > .05. The pattern of the differences for the two participant groups across the sessions was similar.

Discussion

In this study, the effect of training with and without feedback in facilitating the ability of naive listeners to match severity levels of breathiness with the use of references was investigated. The correct answer of each matching trial was given as feedback in the stimulus-response-feedback- stimulus training paradigm for the feedback group while no feedback was given for the no feedback group. The significant main session effect of the accuracy of response suggested that naive listeners in both training condition improved significantly after training (See Table 1). This demonstrated the training effect in facilitating the ability of naïve listeners in matching severity levels of breathiness with the use of references. However, the improvement for the percentage of intra-rater agreement was only significant between the pre-training and post-training session but not between the pre-training and review session (See Table 2). This implied that training might be less effective to enhance the consistency and thus reliability of the participants' performance in this study. On the other hand, the failure to obtain the significant session by group interaction effect and the insignificant main group effect for both the accuracy of response and the percentage of intra-rater agreement suggested that training with and without feedback had similar effectiveness in facilitating the matching ability of naive listeners in this study.

Consistent with the findings of Chan and Yiu (in press), the significant improvement in the overall accuracy of response across the feedback group and the no feedback group after training showed that the reference matching training program was effective in improving naive listeners' ability to match severity levels of breathiness of with the use of references. These studies support training to improve the perceptual voice evaluation skills of naive listeners. Although only breathiness was focused in these studies, it is also possible that the reference matching training method would be effective in training the perceptual voice evaluation of other voice qualities.

Furthermore, the insignificant improvement in the overall percentage of intra-rater agreement between the pre-training and the review session suggested that the training program used in this study was less effective to enhance the consistency and thus the reliability of the participants in perceptual voice evaluation. This also indicates that more training or different training and feedback protocol may be needed to strengthen the consistency of individual listeners' judgment which is also essential in a reliable perceptual voice evaluation.

However, the similar learning pattern for both the feedback and no feedback group in this study appeared to be inconsistent with the findings reported for vocal production tasks. Ferrand (1995) suggested that practice without feedback would facilitate or enhance the carryover effects of the learned skills compared with practice with feedback. The following possible explanations can be conceived for this finding. Feedback (provision of correct answers) might be considered as redundant information in the reference matching training program in this study. According to Magill, Chamberlin and Hall (1991), provision of feedback has been found to be redundant information when intrinsic information feedback is readily available in a anticipation timing skill. In this study, participants might be able to process the significant information source that is important to facilitate learning through the repeated exposure to the labeled reference and stimuli when implementing the response in the matching tasks throughout the training session. Listeners might also learn the level of severity of the labeled references according to the numbering labels of the reference through the repeated listening to the same set of references throughout the training session. Therefore, listeners might replace their relatively unstable internal standards with these references as their internal representations of breathiness at different severity level through the repeated listening to the labeled references (Chan & Yiu, in press). These relatively stable internal standards might then be retrieved from the memory to compare with the to-be-rated stimuli with these internal standards for judgment in post training and review session. In this way, perceptual learning of the participants might mainly occur through the matching process and the repeated exposure to the labeled response during training. Participants in the feedback group might less rely on the guidance of feedback. Therefore, over reliance of the guidance from feedback that cause detriments in learning might reduced. This might explain the similar learning pattern for both the training groups. As training with or without feedback had similar effectiveness in facilitating the matching ability of the naive listeners in this study, feedback (provision of correct answer) may be considered to be withdrawn in the reference matching training program.

Although the findings in this study failed to demonstrate that training with no feedback would further facilitate the matching ability of naive listeners compared with training with feedback, several limitations must be noted. Firstly, 36 rating stimuli used in the rating tests could have been too small to elicit the group difference. It might be possible that increasing the number of stimuli would be able to elicit the predicted group differences. Secondly, only one type of feedback and one training program was used the current study. The findings from this study might only applicable for this particular type of feedback (provision of correct answer) and the reference matching training program. It is difficult to conclude that feedback might not be used in training perceptual voice evaluation. There might be a possibility that training with no feedback in another training program (e.g. the paired comparison training program) or another type of feedback (e.g. discussion of the answer with the trainer) would further enhance perceptual learning compared with training that with feedback. Finally, no control group was used in the current study. It is difficult to make conclusion that the improvement of the participants after training is due to the training instead of the repeated exposure to the stimuli.

These data suggest that additional studies using a larger number of stimuli in the rating testes should be perform to investigate if the differences between groups could be elicited with an increased number of stimuli. However, the number of stimuli should be carefully determined to balance the fatigue effect that may affect the validity of the study. In addition, studies using other training program or other feedback protocol should be performed to investigate the effects of different feedback in different perceptual voice evaluation training program. This would help to explore more on the effectiveness of feedback and whether feedback should be used in training perceptual voice evaluation. This might also provide information on whether the views in motor learning (e.g. guidance hypothesis) are applicable to perceptual learning in training perceptual voice evaluation. Finally, a control group that do not receive any training but only exposure to the stimuli used in the training program should be included in future studies to document the training effect.

Conclusion

This study found that reference matching training program was effective in improving the ability of naive listeners in matching severity levels of breathiness with the use of references. However, it was found that training with or without feedback had similar effectiveness in facilitating the matching ability of the naive listeners. The findings were inconsistent with that of the vocal motor learning tasks. Further studies incorporating different training methods and different training protocols could provide more information on the effectiveness of feedback in perceptual learning in training perceptual voice evaluation. Also, future research to examine whether the views in motor learning could be applied to the auditory perceptual learning in training perceptual voice evaluation is warranted. This study proposed that feedback might be withdrawn in the reference matching training program.

Acknowledgements

The author is grateful to all the subjects for their participation in this study.

Deepest thanks are devoted to Dr. Karen Chan and Dr. Valter Ciocca for their kind supervision, valuable comments and guidance. I would also like to express my gratitude to my fellow classmates, other staff in our department, friends and family for their support.

References

Chan, K. M. K., & Yiu, E. M. L. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research, 45,* 111-126.

Chan, K. M. K., & Yiu, E. M. L. (in press). A comparison of two perceptual voice evaluation training programs for naive listeners. *Journal of Voice*.

Carding, P., Carlson, E., Epstein, R., Mathieson, L., & Shewell, C. (2000). Formal perceptual evaluation of voice quality in the United Kingdom. *Logopedics, Phoniatrics, Vocology,* 25, 133-138.

Eadie, T. L. & Baylor, C. R. (in press). The effects of perceptual training on inexperienced listeners' judgment of dysphonic voice. *Journal of Voice*.

Ferrand, C. T. (1995). Effects of practice with and without knowledge of results on jitter and shimmer levels in normally speaking women. *Journal of Voice,* 9, 409-423.

Gerratt, B. R., & Kreiman, J. (2001). Measuring vocal quality with speech synthesis. *Journal of the Acoustical Society of the America,* 110, 2560-2566.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Breke, G.. (1993). Comparing internal and external standards in voice quality judgements. *Journal of Speech and Hearing Research*, 36, 14-20.

Janelle, C. M., Kim, J. & Singer, R. N. (1995). Subject-controlled performance feedback and learning of a closed motor skill. *Perceptual and Motor Skills,* 81, 627-634.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of the America,* 87, 820-857.

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

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research*, 35, 512-520.

Magill, R. A., Chamberlin, C. J. & Hall, K.G.. (1991). Verbal knowledge of results as redundant information for learning an anticipation timing skill. *Human Movement Science*, 10, 485-507.

Martin, D. P., & Wolfe, V. I. (1996). Effects of perceptual training based upon synthesized voice signals. *Perceptual and Motor Skills*, 83, 1291-1298.

Newell, K. M. (1991) Motor skills acquisition. *Annual Review of Psychology*, 42, 213-137.

Park, J. H., Shea, C. H. & Wright, D.L. (2000) Reduced- frequency concurrent and terminal feedback: A test of the guidance hypothesis, *Journal of Motor Behavior,* 32, 3, 287-296.

Schmidt, R. A. & Wulf, G. (1997). Continuous concurrent feedback degrades skill learning: implications for training and simulation. *Human Factors*, 39, 509-525.

Shrivastav, R., Sapienza, C. M. & Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research*, 48, 323-335.

Sparrow, W. A., Summers. J. J. (1992). Performance on trials without knowledge of result (KR) in reduced relative frequency presentations of KR. *Journal of Motor Behavior*, 24, 197-209.

Steinhauer, K. & Grayhack, J. P. (2000). The role of knowledge of results in performance and learning of a voice motor task. *Journal of Voice*, 14, 2, 137-145.

Swinnen, S. P. (1996). Information feedback for motor skill learning: A review. In H. N. Zelazink (Ed.), *Advances in Motor Learning and Control* (pp. 37-66). Champaign, IL: Human Kinetics

Yiu, E. M. L., Murdoch, B., Hird, K., & Lau, P. (2002). Perception of synthesized voice quality in connected speech by Cantonese speakers. *Journal of the Acoustical Society of the America,* 87, 820-857.

Yiu, E. M. L., Verdolini, K. & Chow, L. P. Y. (2005). Electromyographic study of motor learning for a voice production task, *Journal of Speech, Language, and Hearing Research*, 48, 1254-1268.

Appendix A

Synthesis Values of the stimuli used in rating tests and training session

Synthesis Values	**Training**		**Novel testing**	
	F_0 240	**F_0 260**	**F_0 200**	**F_0 220**
Prototype	✓*	✓	✓	✓*
		Breathy stimuli		
AH55	✓*	✓	✓	✓*
AH60	✓*	✓	✓	✓*
AH65	✓*	✓	✓	✓*
AH70	✓*	✓	✓	✓*
AH75	✓*	✓	✓	✓*
		“ Rough-like” stimuli		
DI04 Prototype		✓	✓	
DI04 AH55		✓	✓	
DI04 AH60		✓	✓	
DI04 AH65		✓	✓	
DI04 AH70		✓	✓	
DI04 AH75		✓	✓	

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

Note: Default values for prototype stimulus = D10 AH40.

* stimuli that were used as the reference in the reference matching tests

Appendix B

Consent Form

CONSENT TO PARTICIPATE IN RESEARCH

Perceptual voice evaluation

You are invited to participate in an undergraduate dissertation conducted by Li Wing Sang Margaret, a final year student from the Division of Speech and Hearing Sciences at the University of Hong Kong.

PURPOSE OF THE STUDY

This study explores the effectiveness of feedback in the training perceptual voice evaluation

PROCEDURES

You will be invited to do the following things:

1. You will be invited to participate a first rating session to rate the synthesized voice signal (about 20 minutes).
2. You will then receive a training session. In the training session, you will be asked to rate the synthesized voice signal similar to the first rating session (about 60 minutes).
3. Two days and one week after the first rating session, you will be invited back for a second and third ratings sessions similar to the first test respectively (about 30 minutes).

POTENTIAL RISKS AND DISCOMFORTS

While you may feel frustrated or tired during the rating tests, such discomfort will be kept minimal. To minimize fatigue, the test will be mostly self-paced with several short breaks.

POTENTIAL BENEFITS

This project can provide useful information to the effectiveness of feedback in the training perceptual voice evaluation.

CONFIDENTIALITY

Any information obtained in this study will remain confidential.

PARTICIPATION AND WITHDRAWAL

If you volunteer to be in this study, you may withdraw at any time without any consequences.

QUESTIONS AND CONCERNS

If you have any questions or concerns about the research, please feel free to contact Li Wing Sang Margaret at 93152768 or h0201574@hksua.hku.hk.

SIGNATURE

I ___________________________________ (Name of Participant) understand the procedures described above and agree to participate in this study.

____ ____________________________

Signature of Participant Date