Title	Effects of listening conditions on perceptual ratings of hypernasal speech
Author(s)	Au, Chi-yeung; 區志漾
Citation	
Issued Date	2010
URL	http://hdl.handle.net/10722/173694
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Effects of listening conditions on perceptual ratings of hypernasal speech

2006223874

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, 2010.

Effects of listening conditions on perceptual ratings of hypernasal speech

Au Chi Yeung

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, 2010.

Abstract

Perceptual speech evaluation is commonly used in clinical settings and research purposes. Nevertheless, criticisms and questions regarding the use of perceptual evaluation exist. Variable reliability and variety of influencing factors, including listeners' experience and listeners' training are concerned by many researchers. Nevertheless, listening condition in perceptual speech evaluation have not been studied since 1984. Updated studies with justifiable experimental procedures and statistic approaches are called. This study investigates and compares the effects of different listening conditions, i.e. high quality headphone condition, regular commercial earphone, and free field speaker condition; on the perceptual rating of hypernasal speech. Outcome measures include the intra- and inter-rater reliability, and intra- and inter-rater agreement. The results showed that the three investigated listening conditions did not pose statistically significant differences in rating hypernasal speech. This study contributes to the construction of standard procedures and provides insights and directions for future studies in perceptual speech evaluation.

Introduction

In cleft palate management, perceptual speech evaluation is one of the crucial procedures used in cleft palate clinics in identifying speech deviation, determining surgical successfulness, and the effectiveness of post-surgery training (Gerratt, Till, Rosenbek, Wertz, & Boysen, 1991; Moller & Starr, 1984; Wyatt, Sell, Russell, Harding, Harland, & Albery, 1996). Instrumental speech evaluation, such as Nasometer, is not preferred as a sole measure mainly because hypernasality is fundamentally perceptual in nature (Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Gerratt et al., 1991; Howard & Heselwood, 2002; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Kuehn & Moller, 2000). Other than clinical uses, perceptual speech evaluation is also extensively used in research. Nevertheless, criticisms and questions regarding the use of perceptual evaluation exist. The main concern is the reliability of perceptual evaluation. Great individual variability in perceptual speech evaluation may influence clinical decision which can cause significant impacts on patients (Moller & Starr, 1984). One way to decrease the variability is to obtain judgments from multiple listeners and use statistical measures of central tendency to describe the variable (Moller & Starr, 1984). This is an effective and valid solution to the problem of single listener. However, clinicians are not always available at the same time. Recording speech samples provides a solution. In the process, not only good audio recording is needed (Sell, John, Harding-Bell, Sweeney, Hegarty, & Freeman, 2009; Stoel-Gammon, 2001); appropriate

listening condition should also be provided to the clinicians. There are many issues which need to be addressed in order to justify the use of most of the procedures. For example, we need to know what are the suitable recording system and listening condition that can provide ratings with the best reliability. There are no clear models and standard procedures for rating speech perceptually (Sell, 2005; Sell, Harding, & Grunwell, 1994; Sell et al., 2009).

For many years, researchers have called for a more detailed description of methods, conditions, procedures, and the demonstration of reliability in rating speech perceptually (Gerratt et al., 1991; Gooch, Hardin-Jones, Chapman, Trost-Cardamone, & Sussman, 2001; Kreiman et al., 1993; Kuehn & Moller, 2000; Lohmander & Olsson, 2004; Moller & Starr, 1984; Sell, 2005). Many studies have investigated factors influencing intrarater and interrater reliability in perceptual speech evaluation. The factors studied have included the type of stimuli (Cheung, 2004), co-existing articulation errors in the stimuli (McWilliams, 1954; Starr, Moller, Dawson, Graham, & Skaar, 1984), the experience of the listeners in judging speech quality perceptually (Kreiman, Gerratt, & Precoda, 1990), the effects of listeners training (Huynh, 2007; Lee, Whitehill, & Ciocca, 2009; Stoeckel, 1980), the influence of individual voice quality (Kreiman, Gerratt, Precoda, & Berke, 1992), the effects of different recording systems and different listening conditions (Moller & Starr, 1984), the scale used in speech rating (Cheng, 2006; Whitehill, Lee, & Chun, 2002; Zraick & Liss, 2000), etc . However, in reviewing fifty seven papers and journals published between 1951 and 1990

extensively, results indicated that both intrarater and interrater reliability fluctuated greatly from studies to studies (Kreiman et al., 1993). Neither intrarater nor interrater reliability varied consistently with any methodological factors studied. No factors can be concluded of affecting reliability or not (Kreiman et al., 1993). Researchers have made a comprehensive discussion to explain the present situation. In summary, firstly, many studies failed in reporting reliability at all, some studies even used the author as the only rater. Secondly, most of the studies did not report or inappropriately reported the estimation of reliability and agreement statistically. Most of them were lack of data of confidence interval. This seriously limited the reliability of the conclusions that could be drawn. Thirdly, the literatures as a whole lacked a clear theoretical approach (Kreiman et al., 1993). The types of scale and statistical approaches used were not based on clearly stated goals or theoretical consideration. Authors had no basis in using different procedures that aimed at obtaining most stable and reliable ratings (Kreiman et al., 1993). Other than literature review done by Kreiman et al., 1993, literature review done by other researchers also found very similar findings (Lohmander & Olsson, 2004; Whitehill, 2002). In order to find out the factors contributing to stable intrarater and interrater reliability, obviously more researches need to be done.

There has been only one published study examining the effect of listening conditions on perceptual speech evaluation (Moller & Starr, 1984). This study investigated the effect of listening condition on several different speech qualities, i.e., speech intelligibility, articulation,

nasality, voice, and overall acceptability (Moller & Starr, 1984). The study investigated the condition effects of three listening conditions including live, audio-visual, and audio listening conditions. Listeners were experienced speech clinicians and speech pathology graduate students from the University of Minnesota Cleft Palate Maxillofacial Clinic. Listeners listened to speech samples of 100 patients with cleft lip palate or associated problems. The age of the patients ranged from 2 to 42 (mean age = 11.1 years; SD = 29.1) with gender ratio not specified. The collected speech samples included conversational speech, reading, reading repeated sentences, counting from 1 to 10, and sustained vowels phonation /i/ and /a/. The number of listeners responsible for rating each speech sample and condition ranged from 3 to 7. Listener trainings were provided to the listeners to decrease the variability among listeners' ratings (Moller & Starr, 1984; Stoeckel, 1980). All the rating scales used in the study were Equal Appearing Interval (EAI) scaling with eight points scale. Under different listening conditions, the speech ratings of the study revealed that similar measurements of intelligibility, resonance, articulation, and overall speech acceptability were obtained. The study concluded that no significant condition effects were posed to resonance and articulation judgments in perceptual speech evaluation (Moller & Starr, 1984).

However, based on the latest studies, some of the procedures used in the Moller & Starr study are now considered invalid. First of all, EAI scaling was used in the study to rate hypernasality. EAI is recently found to be invalid as the rating scale for hypernasality due to

the characteristics of hypernasality as a prothetic continuum (Cheng, 2006; Lee et al., 2009; Zraick & Liss, 2000). Problems exist as raters do not always treat the rating scale as equal intervals when using interval scaling for rating prothetic qualities (Cheng, 2006). DME (Direct Magnitude estimation) and VAS (Visual Analogue Scaling) are suggested as reliable and valid scales in rating nasality (Cheng, 2006; Whitehill et al., 2002; Zraick & Liss, 2000). Secondly, not all the listeners were rating on same set of speech samples. Different speech samples were rated by different listeners. Listener training prior to the rating tasks could not ensure the same internal standard of listeners. Comparisons of speech ratings under this limitation decreased the reliability of the conclusion drawn. Moreover, only three to seven listeners were arranged for each patient and condition. Studies with larger sample size can provide a more reliable statistical analysis. The study was done twenty-six years ago. But nowadays, the recording systems and sound replay system have changed a lot. Furthermore, the research studied the effects of live condition, audio-visual condition and pure audio condition on perceptual speech evaluation. Comparison among pure audio conditions was not administered. Audio recording allows clinicians listening to speech sample anytime and anywhere. The high geographic and temporal mobility of listening to audio recording make it worth for study and further discussion.

According to previous study, rating of other speech qualities such as intelligibility is affected by many external factors (Wyatt, Sell, Russell, Harding, Harland, & Albery, 1996),

and intelligibility of speech is not used as a sole and main speech quality to describe cleft palate speech (Wyatt et al., 1996). In opposite, nasality are very commonly and generally used in describing cleft palate speech (Harding & Grunwell, 1998). In addition, invalid scaling was used for rating hypernasality in the Moller & Starr study. Therefore, the condition effects of perceptual speech evaluation of hypernasality were investigated in the current study.

Recently, researchers have more interest in cross-centre and cross-country studies (Mars, Asher-McDade, Brattstrom, Dahl, McWilliam, Molsted, 1992). Perceptual speech evaluation in different settings and conditions are very commonly involved in these studies. Evidences regarding best listening condition can help the procedures decisions for such cross-centre and cross-country studies.

In summary, the current study aimed at examining the effects of three listening conditions, high quality headphone, regular commercial earphone, and free-field speaker, on perceptual rating of hypernasality. The outcome measures of the study are intrarater and interrater reliability and agreement of listeners' ratings of 23 randomized speech samples. The result of the study was expected to provide evidence-based procedures for rating hypernasal speech perceptually. The result of the study will also contribute to the research purposes, providing direction for further investigation in the field of perceptual speech evaluation.

Methods

Participants Thirty-six listeners (19 females and 17 males) with age ranged from 19 to 25 years (mean age = 21.5 years; SD = 1.52) participated in this study on a voluntary basis. Participants were recruited in a random basis. Twenty two listeners were undergraduate students from the Division of Speech and Hearing Sciences, The University of Hong Kong. The other fourteen listeners were undergraduate students of The University of Hong Kong from other faculties. All the participants were native Cantonese speakers and considered to have no or very limited prior experience in perceptual evaluation of hypernasality so that they lacked specific internal standards for judging hypernasality (Kreiman et al., 1992). All of them had normal hearing as defined by passing a pure-tone audiometric screening at 20 dB HL at octave frequencies from 250Hz to 4000Hz.

Listening conditions Each listener rated the speech samples in all three listening conditions, i.e., 1) high quality headphone (Audio-Technica ATH-T2 headphone), 2) regular commercial earphone (Philips SHE 1360 J PRO earphone), 3) free-field speaker (Harman HK206 speaker). The procedures were administered in a quiet clinic room with sound pressure level under 40dB. The speech stimuli were processed by a Conexant High Definition SmartAudio 221 sound processing unit and were played to listeners binaurally using the high quality headphone, regular commercial earphone and clinical free-field speaker accordingly.

Speech Stimuli Two types of stimuli were used in the study: training stimuli and

experimental stimuli. The training stimuli were extracted from the database of The American Cleft Palate-Craniofacial Association (ACPC) which could be found on the web page <http://www.acpa-cpf.org/educMeetings/speechSamples/index.htm> as speech samples for reference purposes. These stimuli include three sets of speech samples of men, women, and children spanning the severity range in equal intervals from normal to extreme hypernasality (Kuehn et al., 2002). These stimuli were selected due to their clearness of hypernasality and free from co-existing articulation errors. Further information on the sample can be found in Kuehn et al., 2002.

The experimental stimuli were selected from a database provided by Professor David Jones from the University of Wyoming. The database consists of 4828 English sentences produced by 448 English speaking children, adolescents, and adults who were diagnosed with different severity and types of velopharyngeal dysfunction (VPD). All the speech samples in the database were rated on articulation errors, hypernasality, and hyponasality by Professor David Jones, who is an expert in perceptual rating of speech in VPD. Patients with special medical conditions were excluded, e.g., Pierre Robin sequence, and palatal tumor. Speech samples with poor articulation and hyponasality were also excluded. A total of twenty speech samples were finally selected from 11 females and 9 males with age range from 4 years to 13 years (age mean = 8.35; SD = 3.07). Approximate equal numbers of speech samples from different severity of hypernasality were selected. In the speech samples, articulation errors

were unavoidable but articulation errors were not severe compared with hypernasality. The selected patients had a diagnosis of bilateral cleft lip and palate, unilateral cleft lip and palate, cleft lip only, cleft palate only, and soft palate only. Three out of the twenty samples were selected as the repeated samples for measurement of intrarater reliability. The selection was based on the rating of hypernasality, the clearness of speech and the likelihood of memorization effect (Shaughnessy, Zechmeister, & Zechmeister, 2000). Speech samples with least co-existing articulation errors were selected as repeated speech samples. Also, the three repeated speech samples spanned through different hypernasality severity. To minimized memorization effect, speech samples with highly distinctive features were excluded, for examples, unusual high pitch of voice or some laughing or yelling sounds in the stimuli. Detailed information of the selected stimuli and the ratings given by Professor David Jones were listed in Table 1.

Each speech sample in the database consisted of eleven English sentences which were elicited by repetition after examiner or reading aloud. One example of the sentences is “Most boys like to play football”. For each sample, only the best five sentences were selected. Firstly, the listening process would last too long if all the speech samples were to be rated. Effects of fatigue and loss of concentration of listeners may decrease the reliability of the study (Shaughnessy et al., 2000). Secondly, in the case of repeated sentences, the voice of the examiner was included. This may affect the rating of the experimental stimuli. Moreover, in

Table 1

Detailed information and rating given by Professor David Jones on the stimuli selected

Number	Gender	Age	Cleft Type	Articulation*	Hypernasality**	Hyponasality***
1	M	8	BCLP	1	1	0
2	F	8	BCLP	1	1	0
3	M	12	CPO	1	1	0
4	M	7	UCLP	1	2	0
5	M	12	UCLP	1	2	0
6✓	M	12	SPO	1	2	0
7	M	10	SPO	1	3	0
8✓	M	11	UCLP	1	3	0
9	F	13	SPO	1	3	0
10	F	5	UCLP	3	4	0
11	F	5	CLO	4	4	0
12	F	8	CPO	4	4	0
13	M	5	UCLP	4	5	0
14	F	10	UCLP	5	5	0
15✓	F	5	UCLP	5	5	0
16	M	5	UCLP	3	6	0
17	F	4	SPO	4	6	0
18	F	5	SPO	6	6	0
19	F	10	CPO	5	7	0
20	F	12	SPO	5	7	0

Remarks: ✓Repeated stimuli
*For severity of articulation disorder, 1: normal, 7: severe
**For severity of hypernasality: 1: normal resonance, 7: severe hypernasal
***For severity of hyponasality, 0: no hyponasality, 2: severe hyponasal
BCLP – Bilateral Cleft Lip and Palate
UCLP – Unilateral Cleft Lip and Palate
CPO – Cleft Palate Only
CLO – Cleft Lip Only
SPO – Soft Palate Only

many of the speech samples, not all of the eleven sentences were considered of good quality and with clear recording. Therefore, for each sample, the best five sentences were kept and the sentences with poor quality or with the voice of examiner were deleted.

Listener training Listener training was provided to all participants. Listener training has been found to decrease the variability among different listeners in rating nasality (Lee et al., 2009; Moller & Starr, 1984; Stoeckel, 1980). The training procedures were based on those described by Lee et al., (2009). Basically, speech samples with different severity range of hypernasality were played to the listeners. Listeners rated five speech samples first. Since the severity ratings of the training stimuli provided by ACPC were only available in EAI scaling, listeners rated the five speech samples using EAI scaling. As training in this study was primarily for understanding the range of severity of hypernasality, secondarily for getting more exposure and being familiar with hypernasal speech; using EAI scaling in the training was acceptable. Listeners were then told whether their ratings were below or above the ratings given by experienced clinician. They were told to adjust their reference point and rate another five speech samples. This process was repeated until the listeners got all ratings correct in one trial. Other than listening to speech samples, training also included a brief explanation of definition and physiological mechanism of producing hypernasal speech.

Speech rating Nasality was commonly used in describing cleft palate speech (Harding & Grunwell, 1998). As discussed earlier, DME and VAS were suggested as reliable

and valid methods for rating hypernasality (Cheng, 2006; Whitehill et al., 2002; Zraick & Liss, 2000). Hypernasality was rated by VAS in this study as it was considered more straightforward concerning data collection and analysis (Cheng, 2006). VAS referred to the method in which listeners placed an unambiguous mark/stroke in proportion to the perceived hypernasality of each stimulus, along an undifferentiated 10 cm straight line with fixed and predefined extremes of resonance (Cheng, 2006; Eadie & Doyle, 2002). In this study, the left and right endpoints of the 10 cm line were labeled "normal resonance" and "severely hypernasal".

Procedures The 23 speech samples were randomized in three different orders. Listeners were randomly assigned into three groups. Block randomization on the orders of speech samples and listening conditions was applied to ensure that possible memorization effect and fatigue effect were minimized (Shaughnessy et al., 2000). The three repeated samples were evenly distributed in the set of 23 speech samples. Their positions were arranged to ensure that they were separated enough from their original samples.

The procedures were explained to participants and consent forms (Appendix 1) were given and signed by the participants. The sound pressure level of the environment was checked and the hearing ability of the listeners was screened. Training was then given to the listeners for about fifteen minutes. After listener training, listeners listened to speech samples in different conditions according to the pre-set order. The 23 stimuli were presented for each

listening condition including three repeated speech samples for calculating intrarater reliability. Recording sheets were provided. Listeners needed to mark their ratings on the scaling after listening to the whole five sentences of each speech sample. Rating on the severity of hypernasality was emphasized and any articulation errors and distorted voice quality in the stimuli were reminded to be ignored. Listeners could choose to replay each stimulus once in case of loss of concentration. Each session took approximately 45 minutes to complete.

Data Analysis

The VAS ratings of hypernasality were used to compare the rating performance of listeners in different listening conditions. Outcome measures included intrarater reliability, intrarater agreement, interrater reliability, and interrater agreement.

Pearson's product moment correlation was used to calculate the intrarater reliability of each listener (Kreiman et al., 1993; Munro, 2005). The correlation coefficient of each listener in each listening condition was calculated by comparing the repeated stimuli. The average correlation coefficient for each listening condition was calculated by averaging the correlation coefficients of all the listeners in that listening condition.

Confidence interval of intrarater reliability was calculated according to procedures in Munro, 2005. In setting up the confidence interval around a given *r*, *r* must be converted into Fisher's z_r first, using the appendix table provided in the book (Munro, 2005). In calculating

intrarater reliability of *r*, formula of "95% confidence interval = z_r ±(1.96)(standard error)" was used where "standard error = $1/\sqrt{}$(n-3)"(n= number of listeners) (Munro, 2005). The upper limit and the lower limit of z_r were calculated. After converting the two numbers back to *r*, the 95% confidence interval of intrarater reliability was obtained.

Intrarater agreement of each listener in each listening condition was obtained by comparing the first and second ratings of repeatedly rated stimuli of each listener in that listening condition. Ratings that were within one centimeter of one another were considered as agreeing with each other. The percentage of intrarater agreement for each condition was obtained by dividing the agreed rating over the total number of trials.

Intraclass correlation coefficient (2,k) was used to determine the interrater reliability (Kreiman et al., 1993; McGraw & Wong, 1996; Nichols, 1998). ICC (2,k) referred to two-way random average measures (absolute agreement). Two-way random model was used because there was a systematic source of variance associated with rows and also columns (McGraw & Wong, 1996). In this study, the columns represented the speech samples and the rows represented the listener. Average measurement was applied instead of the more commonly used single measurement. This was mainly because ICC in this part was a measure of the reliability of all ratings combined (Ludwig-Mayerhofer, 2006; McGraw & Wong, 1996). Interrater reliability in this study was not finding out "if ratings of one listener are the same as that of the others". Therefore, single measurement was not used in the current

study. Lastly, absolute agreement definition was applied because for consistency measures, column variance was excluded from denominator variance, and for absolute agreement, column variance was not excluded (McGraw & Wong, 1996; Shavelson & Webb, 1991). Column variance was excluded from denominators because the variance was considered as irrelevant in the case of consistency measures(McGraw & Wong, 1996). However, in this study, both the column and row variances were relevant. Therefore, absolute agreement definition was applied. The confidence interval of interrater reliability was obtained as upper bound and lower bound interval from SPSS output.

Interrater agreement was calculated by comparing the "single average" and "group average" of ratings of that listening condition. By averaging the ratings of 20 speech samples of one listener, "single average" of that listener was obtained. By averaging the "single averages" of 36 listeners, a "group average" was obtained. The "single averages" that were within one centimeter from "group average" were considered as agreeing with "group average". The percentage of interrater agreement for each condition was computed by dividing the number of agreed "single averages" over the total number of listeners.

One data point in rating repeated sample in the headphone condition and one in earphone condition were excluded. The two data points were from different listeners. These data were considered outliers as they involved extremely different ratings from all others. It was hypothesized that the listeners have lost concentration or made a mistake at that trial.

Results

Intrarater reliability and agreement

The mean intrarater reliability and agreement of the three listening conditions are listed in Table 1.

Table 1.

Intrarater reliability and agreement of three listening conditions

Listening condition	Reliability (Pearson's *r*)			Percentage of agreement ±1.0 cm
	Mean	SD	95% Confidence Interval	
High quality headphone	0.913	0.163	0.835<*p*<0.955	72.4%
Commercial earphone	0.878	0.213	0.770<*p*<0.940	59.0%
Free field speaker	0.929	0.162	0.870<*p*<0.965	61.1%

A one-way independent group ANOVA was performed to determine if the differences in mean correlation coefficients (*r* value) among the three listening conditions was statistically significant. The results showed that there were no statistically significant differences in intrarater reliability among the three listening conditions [F (2,103) = 0.752; *p*= 0.474]. The 95% confidence interval was also calculated according to the procedures explained previously.

For the percentage of agreement, a one-way independent group ANOVA was performed to determine if the difference in mean percentage among the three listening conditions was

statistically significant. Results indicated that there were no statistically significant differences [F (2,103) = 2.17; p= 0.119].

Interrater reliability and agreement

The interrater reliability and agreement of the three listening conditions are listed in Table 2.

Table 2.

Interrater reliability and agreement of three listening conditions

Listening condition	Reliability ICC (2,k)		Percentage of agreement ±1.0 cm
	Coefficient	95% Confidence Interval	
High quality headphone	0.989	0.976<p<0.994	69.4%
Commercial earphone	0.922	0.982<p<0.995	80.6%
Free field speaker	0.991	0.978<p<0.994	77.8%

The coefficient and the 95% confidence interval were calculated by SPSS. No statistical procedure was available to determine if there was any statistically significant difference among the coefficients of the three listening conditions.

To illustrate the data more clearly and study the interrater agreement, figures plotting the average ratings of each listeners compared with the group average were constructed for each listening conditions (Figures 1-3).

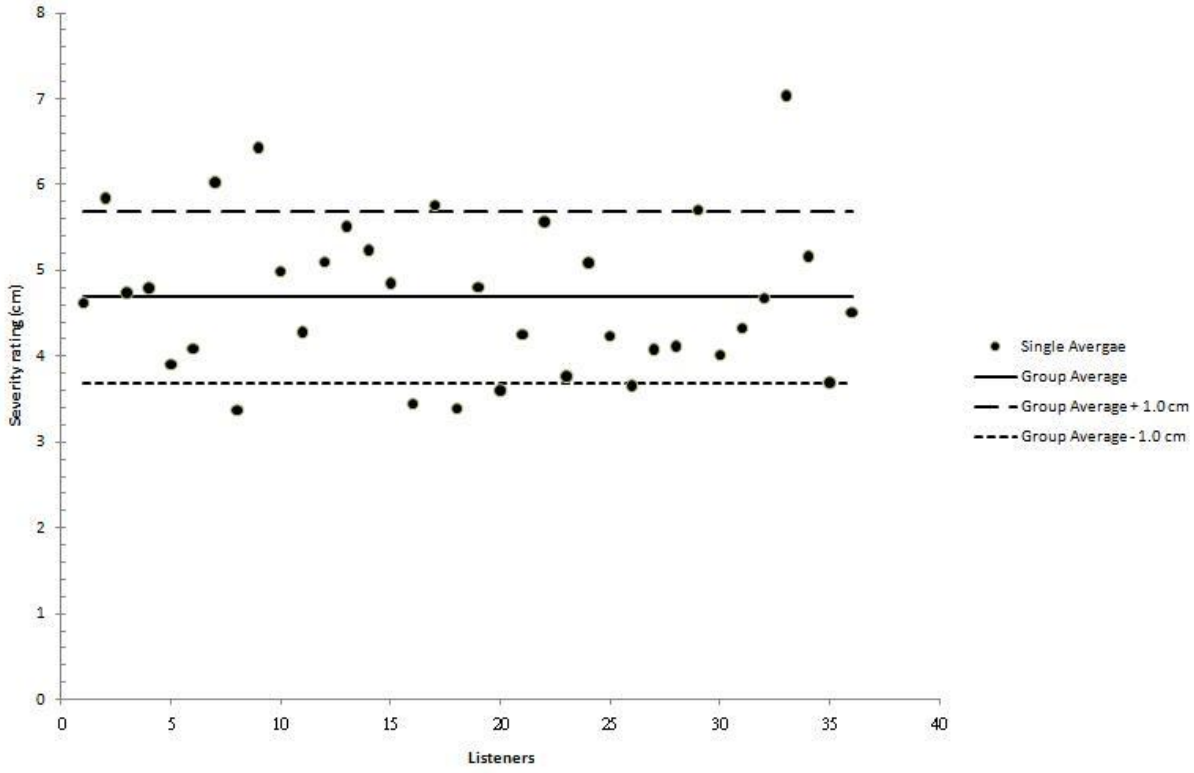

Figure 1. *Average ratings of listeners in headphone condition*

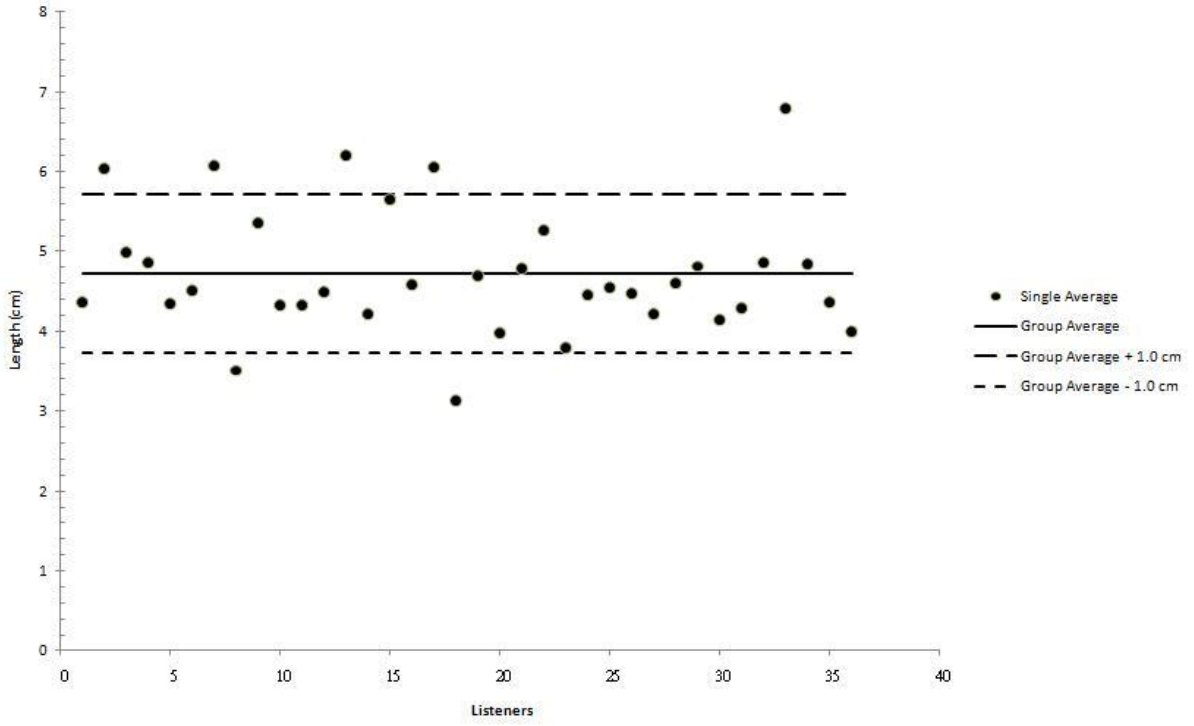

Figure 2. *Average ratings of listeners in earphone condition*

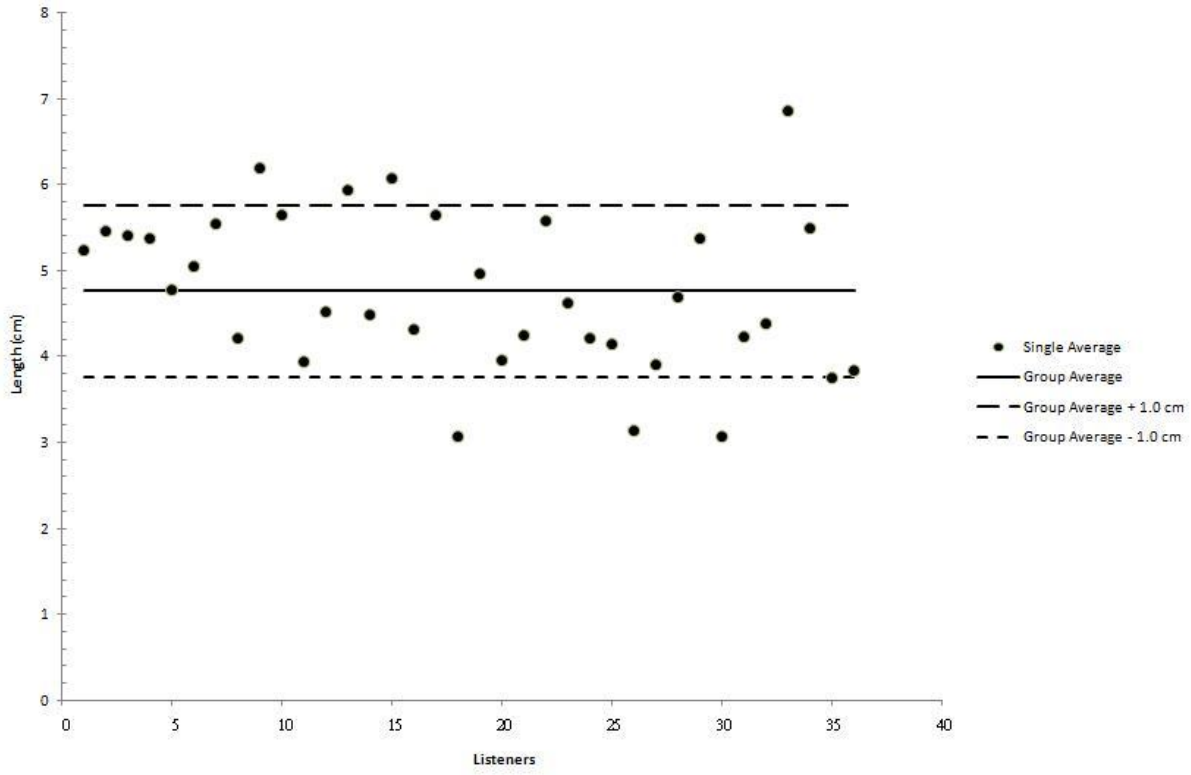

Figure 3. *Average ratings of listeners in free-field condition*

Discussion

Intrarater reliability and agreement

For intrarater reliability, the mean *r* values of the three listening conditions were quite high (headphone condition: $r = 0.91$; earphone condition: $r = 0.88$; free-field speaker condition: $r = 0.93$). The mean *r* value of earphone condition was relatively low compared with the other two. However, the one-way ANOVA indicated that there were no statistically significant differences among the three listening conditions [$F(2,103) = 0.752$; $p = 0.474$]. The percentage of intrarater agreement was varied among the three listening conditions ranging from 59% to 72%. Nevertheless, statistical results of one way ANOVA indicated that there were no statistically significant differences [$F(2,103) = 2.17$; $p = 0.119 > 0.05$].

In term of intrarater reliability and agreement, the three listening conditions showed no effects to listeners in rating hypernasal speech.

Interrater reliability and agreement

The interclass correlation coefficients of the three listening conditions were 0.989 (headphone), 0.922 (earphone), and 0.991 (speaker) (Table 2). All the three coefficients were high. This revealed that difference in listening conditions did not pose any changes in interrater reliability in this study. The percentage of interrater agreement was varied among the three listening conditions ranging from 69% to 81%. Different from the percentage of intrarater agreement, the percentage of agreement here was not an average value; they were obtained from plotting the graphs of average rating of single listener across all listeners in each listening condition (Figures 1, 2, 3). The percentage of agreement was obtained by counting the number of "single average" value that lay within the range of "group average" ± 1.0 cm and dividing the number by the total number of listeners. Therefore, no statistic analysis was computed against the percentage of interrater agreement to justify if there were statistically significant differences. However, by referring to the Figure 1, 2, 3, the data did not show a trend or sign that one condition was different with the other two in interrater agreement.

Concluding the results in table 2, there were no listening conditions that may bring any condition effects influencing the interrater reliability and agreement of listeners in rating

hypernasal speech.

With referring to the Moller & Starr study, results of the current study results were not astonishing but were worth for notice. After modifying the rating scale for hypernasality to VAS, fixing the number of listeners for each speech and condition, and enlarging the samples size; the condition effects of perceptual rating for hypernasality were still statistically insignificant. The current study could ensure that conclusion of the conditions effects was drawn based on valid rating scale for hypernasality and valid procedures.

The current study results may bring great impact to the clinical and research area relating to perceptual rating of hypernasality. As proposed by many different researchers, there were absolute needs to study and find out a methodology in perceptual rating of speech that is repeatable, with theoretical basis in procedures, justifiable statistically, with reliability and confidence interval stated clearly (Kreiman et al., 1993; Lee et al., 2009; Mars et al., 1992; Sell, 2005; Whitehill, 2002; Wyatt et al., 1996). The need was called and the debate continued for years. It was reasonable to understand that studying the whole picture of perceptual speech evaluation in one study is impossible. The current study focused on perceptual rating of hypernasal speech in different listening conditions. The results surely contributed to complete the picture of the whole area. Clinically, the results of the study also contributed to easier and more convenient procedures in rating hypernasality in cleft palate speech. When multiple listeners approach was adopted in the clinic, clinicians could choose

to use either of the instruments in rating the severity. They could choose what was available in their clinic, because high quality headphone, regular earphone, and using free-field speaker gave the same ratings. Or on the other hand, clinicians did not need to buy expensive high quality headphone in rating hypernasal speech.

Limitation of the current study

In the current study, hypernasality was rated to find out the reliability of ratings in different listening conditions. However, hypernasality judgement was influenced by many different factors such as articulation errors, hyponasality and voice quality of the speech (Cheng, 2006). For studies in rating hypernasality, in general, speakers with hypernasality but without co-existing articulation or voice problems were extremely rare. Speakers with co-occurring problems were very common and the rating of hypernasality would really be affected. For example, in the current study, co-existing articulation errors were unavoidable. Either from the database provided by Professor David Jones or other web sources, suitable speech samples with no articulation errors were hard to find. For speech samples with more severe hypernasality, the co-existing articulation errors would be more severe also. Therefore, speech samples number 18, 19, 20 were unavoidably with more severe articulation disorder. This was a limitation for the current study and for most of the studies in rating hypernasality.

In the current study, samples with hyponasality were excluded in the current study. Suitable speech samples with hypernasality but not hyponasality were available. Thus

hypernasality judgement was not influenced by hyponasality. Nevertheless, deviation in voice quality for example breathiness and hoarseness could also affect the hypernasality judgement (Kataoka, Zajac, Mayo, Lutz, & Warren, 2001). Specifically, breathiness was found to raised judgement of slight hypernasality and reduced severe hypernasality (Imatomi, 2005). Same as articulation errors, deviation in voice quality and breathy voice were unavoidable. This was also a limitation of the current study.

Three speech samples were repeatedly listened and rated by each listener in each listening condition. The number of repeated speech samples was limited by the time length of the experiment, and the availability of appropriate speech samples. Memorization effect would be obvious when the proportion of repeated samples in the total number of stimuli was high. However, the total number of stimuli was limited. Loss of concentration and effect of fatigue would occur when too many speech samples were needed to be rated.

Another limitation was that the speech stimuli used in the current study were English sentences, but all the listeners in judging hypernasality in the study was Chinese whose mother tongue was Cantonese. The linguistic background of the listeners and the language of the speech samples were factors that would influence the judgement. One recent publication suggested that non-expert Cantonese and English listeners rated hypernasality in Cantonese speech samples in a similar way(Lee, Brown, & Gibbon, 2008). However, no study concerned, did non-expert Cantonese and English listeners rate hypernasality in English

speech samples in a similar way. The effects of linguistic background of listeners in rating English hypernasal speech were unclear. Moreover, the study also highlighted the needs for further cross-linguistic studies and it was an area that needed to be explored (Lee et al., 2008).

Direction and issues for future research

There are several issues for further research. Firstly, in the future attempts, other languages such as Cantonese and Putonghua which are commonly used in Hong Kong can be involved. For more international purposes and cross countries cross centre collaboration, English and other European languages can be attempted so that the study results can be generalized to many other different areas. Patients under different cultural backgrounds and researchers investigating with different languages can also be benefited from the study results.

Secondly, speech samples can be collected specifically for this research purposes. Speech samples used in the current study were retrieved from the database from Professor David Jones. The data base was not constructed for the current research purposes. In the future attempt, patients with targeted speech characteristics can be recruited so that the voice quality and other variable such as nasality and co-existing articulation errors were well controlled.

Conclusion

In conclusion, listening conditions involving high quality headphone, regular commercial earphone, and clinical free-field speakers did not pose statistically significant differences in rating hypernasal speech in term of intrarater reliability and agreement and interrater reliability and agreement. Therefore, all three listening conditions for rating hypernasal speech were appropriate.

Acknowledgments

I would like to express my sincere gratitude to Professor Tara Whitehill for her guidance and support during the course of this dissertation. I would like to thank Professor David Jones of the University of Wyoming for the provision of speech stimuli. Sincere thanks are also given to all the participants for their time devoted. Besides, I would like to thanks all my family members, my fellow classmates, many brothers and sisters, and all my friends for their support and encouragement, throughout the writing of the dissertation. Last but not least, I would like to give thanks to God, the greatest encourager, who has blessed me with joy and peace through hard times.

References

Cheng, T. H. (2006). *Direct magnitude estimation versus visual analogue scaling in the perceptual rating of hypernasality.* Unpublished BSc dissertation, The University of Hong Kong.

Cheung, S. C. J. (2004). *The effects of stimulus and modulus on perceptual rating of hypernasality.* Unpublished BSc Dissertation, The University of Hong Kong.

Eadie, T., & Doyle, P. (2002). Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *Journal of Speech, Language and Hearing Research, 45*(6), 1088-1096.

Gerratt, B., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research, 36*(1), 14-20.

Gerratt, B. R., Till, J. A., Rosenbek, J. C., Wertz, R. T., & Boysen, A. E. (1991). Use and perceived value of perceptual and instrumental measures in dysarthria management. In C. A. Moore, K. M. Yorkston & D. R. Beukelman (Eds.), *Dysarthria and apraxia of speech* (pp. 77-93). Baltimore, MD: Brookes.

Gooch, J., Hardin-Jones, M., Chapman, K., Trost-Cardamone, J., & Sussman, J. (2001). Reliability of listener transcriptions of compensatory articulations. *The Cleft Palate-Craniofacial Journal, 38*(1), 59-67.

Harding, A., & Grunwell, P. (1998). Active versus passive cleft-type speech characteristics. *International Journal of Language & Communication Disorders, 33*(3). 329-352.

Howard, S., & Heselwood, B. (2002). Learning and teaching phonetic transcription for clinical purposes. *Clinical Linguistics & Phonetics, 16*(5), 371-401.

Huynh, Y. S. C. (2007). *Training perceptual rating of hypernsality with co-existing speech disorders.* Unpublished BSc Dissertation, The University of Hong Kong.

Imatomi, S. (2005). Effects of breathy voice source on ratings of hypernasality. *The Cleft Palate-Craniofacial Journal, 42*(6), 641-648.

Kataoka, R., Zajac, D. J., Mayo, R., Lutz, R. W., & Warren, D. W. (2001). The influence of acoustic and perceptual factors on perceived hypernasality in the vowel /i/: A preliminary study. *Folia Phoniatrica Logopaedica, 53*, 198-212.

Kreiman, J., Gerratt, B., Kempster, G., Erman, A., & Berke, G. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36*(1), 21-40.

Kreiman, J., Gerratt, B., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33*(1), 103-115.

Kreiman, J., Gerratt, B., Precoda, K., & Berke, G. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35*(3), 512-520.

Kuehn, D., Imrey, P., Tomes, L., Jones, D., O'Gara, M., Seaver, E., et al. (2002). Efficacy of

continuous positive airway pressure for treatment of hypernasality. *The Cleft Palate-Craniofacial Journal, 39*(3), 267-276.

Kuehn, D., & Moller, K. (2000). Speech and language issues in the cleft palate population: The state of the art. *The Cleft Palate-Craniofacial Journal, 37*(4), 348-348.

Lee, A., Brown, S., & Gibbon, F. (2008). Effect of listeners' linguistic background on perceptual judgements of hypernasality. *International Journal of Language & Communication Disorders, 43*(5), 487-498.

Lee, A., Whitehill, T., & Ciocca, V. (2009). Effect of listener training on perceptual judgement of hypernasality. *Clinical Linguistics & Phonetics, 23*(5), 319-334.

Lohmander, A., & Olsson, M. (2004). Methodology for perceptual assessment of speech in patients with cleft palate: A critical review of the literature. *The Cleft Palate-Craniofacial Journal, 41*(1), 64-70.

Ludwig-Mayerhofer, W. (2006). Intraclass Correlation. Retrieved from http://www.lrz-muenchen.de/~wlm/wlmsicc.htm

Mars, M., Asher-McDade, C., Brattstrom, V., Dahl, E., McWilliam, J., Molsted, K., et al. (1992). A six-center international study of treatment outcome in patients with clefts of the lip and palate: Part 3. Dental arch relationships. *Cleft Palate Craniofac. J, 29*, 405-408.

McGraw, K., & Wong, S. (1996). Forming inferences about some intraclass correlation

coefficients. *Psychological Methods, 1*(1), 30-46.

McWilliams, B. (1954). Some factors in the intelligibility of cleft-palate speech. *Journal of Speech and Hearing Disorders, 19*(4), 524.

Moller, K. T., & Starr, C. C. (1984). The effects of listening conditions on speech ratings obtained in a clinical setting. *Cleft Palate Journal, 21*(2), 65-69.

Munro, B. (2005). *Statistical methods for health care research* (5th ed.). Philadelphia: Lippincott Williams & Wilkins.

Nichols, D. P. (1998). Choosing an intraclass correlation coefficient. Retrieved from http://support.spss.com/ProductsExt/SPSS/Documentation/Statistics/articles/whichicc.htm

Sell, D. (2005). Issues in perceptual speech analysis in cleft palate and related disorders: A review. *International Journal of Language & Communication Disorders, 40*(2), 103-121.

Sell, D., Harding, A., & Grunwell, P. (1994). A screening assessment of cleft palate speech (Great Ormond Street Speech Assessment). *International Journal of Language & Communication Disorders, 29*(1), 1-15.

Sell, D., John, A., Harding-Bell, A., Sweeney, T., Hegarty, F., & Freeman, J. (2009). Cleft Audit Protocol for Speech (CAPS-A): a comprehensive training package for speech analysis. *International Journal of Language & Communication Disorders, 44*(4),

529-548.

Shaughnessy, J., Zechmeister, E., & Zechmeister, J. (2000). *Research methods in psychology* (5th ed.). Singapore: McGraw-Hill Humanities.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, Calif.: Sage Publications, Inc.

Starr, C., Moller, K., Dawson, W., Graham, J., & Skaar, S. (1984). Speech ratings by speech clinicians, parents and children. *The Cleft Palate Journal, 21*(4), 286-292.

Stoeckel, R. (1980). *Effects of training on nasality ratings.* Unpublished Master of Arts Thesis, University of Minnesota.

Stoel-Gammon, C. (2001). Transcibing the speech of young children. *Topics in Language Disorders, 21*(4), 12-21.

Whitehill, T. (2002). Assessing intelligibility in speakers with cleft palate: A critical review of the literature. *The Cleft Palate-Craniofacial Journal, 39*(1), 50-58.

Whitehill, T., Lee, A., & Chun, J. (2002). Direct magnitude estimation and interval scaling of hypernasality. *Journal of Speech, Language and Hearing Research, 45*(1), 80-88.

Wyatt, R., Sell, D., Russell, J., Harding, A., Harland, K., & Albery, L. (1996). Cleft palate speech dissected: A review of current knowledge and analysis. *British Journal of Plastic Surgery, 49*(3), 143-149.

Zraick, R., & Liss, J. (2000). A comparison of equal-appearing interval scaling and direct

magnitude estimation of nasal voice quality. *Journal of Speech, Language, and Hearing Research, 43*(4), 979-988.

Appendix 1

Informed Consent Form

Effects of listening conditions on perceptual ratings of hypernasality

You are invited to participate in a project research entitled" Effects of listening conditions on perceptual ratings of hypernasality" conducted by Mr. Au Chi Yeung under the supervision of Professor Tara Whitehill of the Division of Speech and Hearing Sciences at the University of Hong Kong.

Purpose of the study The project aims to find out the effects of different listening conditions (high quality headphones, regular commercial earphones, and free-field speakers) on perceptual ratings of hypernasality.

Procedures The research study involves a hearing screening, a brief training on perceptual rating of hypernasality, and rating the severity of hypernasality for 180 speech samples of children with cleft palate. All the procedures will be conducted in the hearing centre and/or clinic rooms at the Division of Speech and Hearing Sciences, The University of Hong Kong. The whole procedure will take about one hour and fifteen minutes.

Potential risks or discomforts There are no potential risks or discomforts.

Potential benefits There are no direct benefits for you. However, the research study can provide valuable information for clinical and research studies of the management of individuals with cleft palate.

Confidentiality Any information obtained in this study will remain strictly confidential, will not be disclosed to any other people, and will be used for research purposes only. Codes, not names, will be used on all research and subject files to protect confidentiality. Participants will not be identified by name in any report of the completed study.

Participation and withdrawal Your participation in this project is voluntary. Withdrawal from this research study at any time, for any reasons, is voluntary and without negative consequences. Part or all of any information obtained from you will be erased upon your request.

Questions and concerns You will be asked to complete and sign the consent form. If

you would like to ask further questions, please contact the investigator Mr. Au Chi Yeung (Email: au532@hkusua.hku.hk) or his supervisor, Professor Tara Whitehill (5/F Prince Philip Dental Hospital, The University of Hong Kong; Tel: 28590599; Email: tara@hku.hk). If you want to know more about the rights as a research participant, please contact the Human Research Ethics Committee for Non-Clinical Faculties, the University of Hong Kong (Tel: 22415267).

We are grateful for all participation in this research study.

Date of preparation: Nov 3, 2009

Informed Consent Form

Effects of listening conditions on perceptual ratings of hypernasality

Code no: _____

I _____ (Name of the participant) have been given the opportunity to ask questions about this study and any questions I raised have been answered to my satisfaction. I understand all the procedures described above and agree to participate in this study.

_____ _____

Name of participant (Block letter) Signature

_____

Date

Date of preparation: Nov 3, 2009