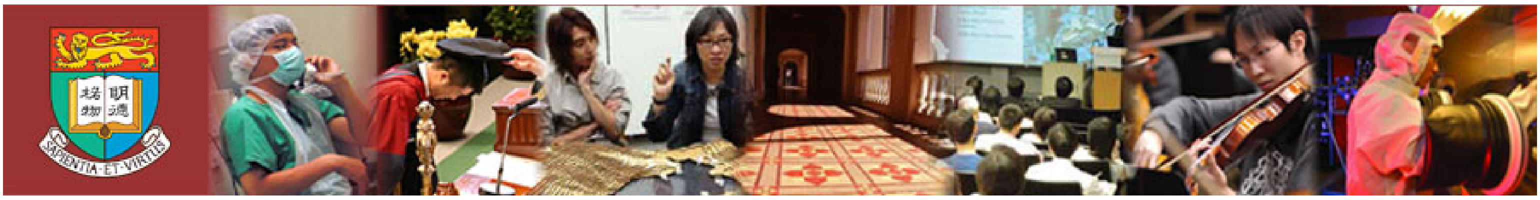


Reliability of rating synthesized hypernasal speech signals in connected speech and vowels

Wong Chun Ho Eddy

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

Reliability of rating synthesized hypernasal speech signals in connected speech and vowels

Wong Chun Ho Eddy

Abstract

The study investigated whether valid hypernasal stimuli (i.e., with higher inter-rater and intra-rater reliability) could be synthesized by using a Klatt synthesizer (Klatt & Klatt, 1990). Two sets of synthesis parameters (i.e., the high-level [HL] synthesis parameter and the low-level [LL] synthesis parameters) in the synthesizer were used to create five sets of Cantonese stimuli (i.e., the HL sentence, vowel /i/ and /ɔ/ and the LL vowel /i/ and /ɔ/). Eleven Cantonese speaking speech therapists were asked to listen to the stimuli and rated whether they detected hypernasality. The result showed that all /i/ stimuli from low-level synthesis parameters and 70% connected speech stimuli from the high-level synthesis parameter were rated as hypernasal. Paired comparison task and perceptual rating task were administered in order to measure the severity of the synthesized hypernasal signals. In addition to the high intra-rater agreement (ranged from 68.60% to 88.64%) and the ICC values (ranged from 0.662 to 0.948), the current study suggested that the low-level synthesis parameters were said to be a set of preferable parameters for synthesizing hypernasal stimuli and both connected speech and vowel /i/ were preferable than vowel /ɔ/ in perceptual evaluation of hypernasality. Research implications are discussed.

Reliability of rating synthesized hypernasal speech signals in connected speech and vowels

Introduction

Hypernasality is defined as an excessive amount of perceived nasal resonance which results from coupling of nasal and oral cavity due to insufficient velopharyngeal closure during the production of normally non-nasal sounds (Boone, McFarlane & Von Berg, 2005). This resonance disorder may be found in individuals with structural anomalies (e.g., cleft palate), with neuropathology (e.g., dysarthria), hearing impairment, or a consequence of oral surgeries (Whitehill, Lee & Chun, 2002; Dworkin, Marunick & Krouse, 2004). In order to evaluate nasality, perceptual evaluation is used (Kuehn & Moller, 2000). Kreiman, Gerratt, Kempster, Erman, and Berke (1993) suggested that perceptual voice evaluation highly depends on listener's experience because internal standards for different voice qualities were developed through exposure to different voices. Based on this theoretical framework, Chan and Yiu (2002) synthesized anchors of disordered voice qualities (i.e. breathiness and roughness) by using the Klatt synthesizer (Klatt & Klatt, 1990) in order to provide judges with signals as external standard. These external representations built up the internal representations of specific voice quality of listeners and resulted in increasing the reliability of perceptually evaluating these different voice qualities. However, no hypernasal stimuli were synthesized in their study. Zrack and Liss (2002) created synthesized hypernasal sounds in order to compare two different scaling methods (i.e., Equal-Appearing Interval Scaling and Direct Magnitude Estimation) in rating hypernasality. However, no process of validation (i.e. determining whether the signals were hypernasal) on the synthesized hypernasal signals was conducted. The reliability and validity of using these signals as anchors in the process of perceptual training was being questioned by author in the current study. In order to develop an external standard of hypernasal stimuli and provide basis for future studies on synthesized hypernasal stimuli, the main objective of this study was to determine whether valid hypernasal stimuli (i.e., with higher inter-rater and intra-rater reliability) could be synthesized by using synthesizers.

Yiu, Murdoch, Hird and Lau (2002) opened up the possibility of using synthesized signals as external standards. These synthesized stimuli provided the basis for the use of anchors (Chan & Yiu, 2002) and perceptual training (Chan & Yiu, 2002) which were supported by researchers (Gerratt, Kreiman, Antonaszas-Barroso & Berke, 1993; Martin & Wolfe, 1996). Chan and Yiu (2002) found that intra-rater agreement of listeners in rating synthesized stimuli improved from 49.11% in pre-training sessions to 75.89% in post-training sessions. This showed that use of anchors and perceptual trainings aimed at using external standards to build up the internal representations of specific voice quality of listeners and resulted in increasing the reliability of perceptually evaluating these different voice qualities. However, only breathiness and roughness were created as standards and developed as anchors. No synthesized hypernasal stimuli were developed.

Zraick and Liss (2000) created hypernasal synthesized signals by using the Klatt synthesizer (Klatt & Klatt, 1990). Seven low-level (LL) synthesis parameters (i.e., F1, first formant frequency; B1, first formant bandwidth; F2, second formant frequency; FNP, frequency of the nasal pole; FNZ, frequency of the nasal zero; BNP, bandwidth of the nasal pole; BNZ, bandwidth of the nasal zero) were varied in order to produce four different levels of severity in hypernasal vowel /i/. According to Kent and Read (2002), nasal sounds (i.e. vowels and consonants) are produced with the opening of velopharyngeal port so that the nasal tract and oral tract coupling together to allow sound energy passes through. Hypernasal sounds are characterized by reduced overall energy of the acoustic signal, increased formant bandwidths so that formant energy appears broader in spectrograms, a slight increase of the F1 and a slight lowering of the F2 and F3, and the presence of one or more anti-formants which can further reduce overall energy of the signals (Kent & Read, 2002). The most prominent feature of nasal sounds is that poles (spectral speak) and zero (deep valley) are involved in transfer function (Kent & Read, 2002). Therefore, the seven parameters were varied in the synthesizer in order to create hypernasal stimuli.

Apart from the low-level synthesis parameters, in the Klatt synthesizer, there is a set of high-level (HL) parameters for synthesizing stimuli. According to Sensimetrics' HLSyn Speech Synthesis System (1997), nasality is synthesized by manipulating AN (the cross-sectional area of the opening of the nasal cavity). In order to synthesize different severity of hypernasal sounds, AN should be varied in order to change the volume of coupled oral and nasal cavity results in changes of resonance frequency. However, no previous study investigated the reliability of rating the signals synthesized by this HL parameter. In the current study, two synthesis parameters were used to synthesize hypernasal signals. The first objective of the study was to determine whether both sets of parameters could be used for the synthesis of valid hypernasal stimuli.

The effect of using sustained vowel versus connected speech for perceptual evaluation has been discussed for a long time. Some researchers supported the use of sustained vowels because they are controllable, easily elicited, easily standardized and have less effect across cultures (Zraick, Wendel, & Smith-Olinde, 2004). However, according to Zriack and Liss (2000), connected speech was closer to daily speech behaviors and allowed for more detailed description of the characteristics. Yiu et al. (2002) also supported the use of connected speech stimuli because they were more representative to the daily speech tasks. Some other studies also involved the use of connected speech (i.e., sentences with a variety of vowels) and reading passage (Lewis, Watterson and Honghton, 2003; Sherman and Hall, 1978). Cheung (2004) found that using non-nasal sentences would lead to significantly higher intra-listeners and inter-listeners reliability in the perceptual rating of hypernasality than isolated vowels and monosyllabic words. She used only real voice samples for her study. There was no direct comparison about the effect of synthesized sustained vowels and connected speech on perceptual evaluation. Therefore, the second objective of the current study was to determine which type of stimuli (i.e., synthesized hypernasal sentences or vowels) was preferable for perceptually evaluating synthesized signals.

Sustained vowel /i/ and /ɔ/ were synthesized and used in the current study. Zraick and Liss (2002) created vowel /i/ as the stimuli since it was a high-front vowel in which less nasal coupling was required to elicit nasal percepts (Abramson, Nye, Henderson & Marshall, 1981). Apart from vowel /I, the use of this vowel /ɔ/ was also suggested by Zraick and Liss (2002) as it was a low-back vowel. Exploration on the use of different vowels was highly recommended. Vowel /ɔ/ was also included in the sentence type stimuli in the current study. This would facilitate the comparison between sustained vowels and sentence.

Valid and reliable synthesized hypernasal signals would provide basis for further development of anchors which could be used for perceptual training in rating hypernasality. However, before developing anchors, some other information should be obtained. The first one would be the Just Noticeable Differences (JND). “JND is the smallest difference in a specified modality of sensory input that is detectable by a human being or other animal” (Wikipedia, the free encyclopedia, 2007). The general idea on the minimum differences that could be detected by judges on the synthesized hypernasal stimuli (i.e., created by both the high-level and low-level synthesis parameters) was obtained in this study. The results of this study would be used as a reference for determining the JND of the synthesized hypernasal stimuli. Further, hypernasal stimuli with different severity of hypernasality could be synthesized and developed as anchors.

Indeed, severity of hypernasality is a continuum. In order to develop a universal set of anchors which represent different levels of severity in the continuum, it was important, in the current study, to determine how the synthesized signals represented the level of severity of hypernasality in the continuum. Visual analogue (VA) scaling, one of the valid and reliable scaling methods, supported by Cheng (2006), for perceptual judgement of hypernasality, would be used to evaluate the severity of hypernasality in the synthesized signals created in the current study.

In summary, the current study investigated whether valid hypernasal sounds (i.e., with higher inter-rater and intra-rater reliability) could be synthesized by using a Klatt synthesizer (Klatt & Klatt, 1990). It could be achieved through investigating, first, which type of synthesis parameters (i.e., the HL or the LL) was more preferable; and second, which type of stimuli (i.e., sentence or vowel, high front vowel or low back vowel) was preferable for evaluating synthesized hypernasal signals.

Method

Preparation of the stimuli

Male voice synthesized signals were created by using Sensimetrics' HLSyn Speech Synthesis System (1997) in a Microsoft Window platform with fundamental frequency between 100Hz and 150Hz. There were two types of stimuli, i.e., the high-level (HL) and the low-level (LL) stimuli. Two sets of stimuli, i.e., sustained vowel /i/ and /ɔ/, were synthesized as the LL sets and three sets of stimuli, i.e., sustained vowel /i/ and /ɔ/ and sentence /pa1 pa1 ta2 kɔ1 kɔ1/ ("father hits the elder brother"), were synthesized as the HL sets. In total, there were five sets of stimuli, i.e., "HL vowel /i/", "HL vowel /ɔ/", "HL sentence", "LL vowel /i/", and "LL vowel /ɔ/".

The sustained vowels were 3 seconds long, which was 1.5 seconds longer than the stimuli created by Zraick and Liss (2002). The extension aimed to provide sufficient time for the listeners to detect hypernasality (Whalen & Beddor, 1989).

The non-nasal sustained vowels of both the high-level and the low-level synthesis parameters were created by following the user manual of Sensimetrics' HLSyn Speech Synthesis System (1997). The non-nasal sentence was based on the prototype stimuli of Yiu et al. (2002). The synthesis parameters were varied slightly from the originals to achieve natural sounding prototype stimuli, as determined by two native Cantonese speakers.

The nasality of the HL sets was synthesized by varying the AN parameter. Each stimulus was differed by 5 AN value with the range from 0 to 100 AN value (e.g., AN0, AN5, AN10, AN15, etc). Twenty one stimuli were prepared for each of the three HL sets. The other HL synthesis parameter (i.e., “AG”, the average values of glottal opening) were slightly modified in order to achieve natural sounding stimuli. For example, AG was increased by one unit at the beginning of the entire vowel /ɔ/ stimuli and the insertion of AG was delayed by 5ms at the beginning of the third syllable (/ta/) in the sentence set.

The seven LL synthesis parameters (i.e., F1, B1, F2, FNP, FNZ, BNP and BNZ) which were used by Zraick and Liss (2002) were varied independently in the current study to synthesize seven levels of hypernasality for vowel /i/ and /ɔ/ (Table 1 and 2). For the /i/ stimuli set, four of the levels were exactly the same as those synthesized by Zraick and Liss (2002) and three intermediate levels were added in between the original levels. The F1, B1, F2, BNP and BNZ of non-nasal vowel /i/ were increased (40 for F1, 20 for B1, 790 for F2, 110 for both BNP and BNZ) when compared with the oral stimuli synthesized by Zraick and Liss (2002) in order to achieve a natural sounding stimulus. For the /ɔ/ stimuli set, the non-nasal stimuli was synthesized automatically by the Klatt synthesizer. The seven levels of the LL synthesis parameters were similarly varied as vowel /i/ set. The range of F1 and F2 values in the /ɔ/ stimuli set were narrowed and the values of B1, FNP, FNZ, BNP and BNZ were slightly reduced when compared with the /i/ stimuli set.

Table 1: *Values of the low-level synthesis parameters for /i/.*

		Levels						
	Non-nasal	1*	2	3*	4	5*	6	7*
F1	310	262	256	250	244	238	232	225
B1	80	300	275	250	225	200	175	150
F2	2290	2325	2362	2400	2438	2475	2512	2550
FNP	500	700	700	800	850	900	900	1000
FNZ	500	1200	1250	1300	1350	1400	1400	1500
BNP	200	150	150	150	150	150	150	150
BNZ	200	250	250	250	250	250	250	250

* Same as Zraick and Liss (2002)

Note: F1- first formant frequency, B1- first formant bandwidth, F2- second formant frequency, FNP- frequency of the nasal pole, FNZ- frequency of the nasal zero, BNP- bandwidth of the nasal.

Table 2: *Values of low-level synthesis parameters for /ɔ/.*

		Levels						
	Non-nasal	1	2	3	4	5	6	7
F1	506	493	490	488	485	482	480	477
B1	200	200	175	150	125	100	75	50
F2	840	864	876	888	901	914	926	938
FNP	570	520	570	620	670	720	700	820
FNZ	500	671	721	771	821	871	921	971
BNP	80	100	100	100	100	100	100	100
BNZ	200	200	200	200	200	200	200	200

Note: F1- first formant frequency, B1- first formant bandwidth, F2- second formant frequency, FNP- frequency of the nasal pole, FNZ- frequency of the nasal zero, BNP- bandwidth of the nasal.

Judges

Eleven expert listeners (one male and ten females), native Cantonese speaking speech therapist with at least three years of experience in clinical practices and had the experience of assessing or treating patients with hypernasality, were asked to serve as judges.

Procedure

Four tasks were included in this study. There were a Yes/No task, two paired comparison tasks and a perceptual rating task. These four tasks were presented in the same order. The first three tasks were presented by E-prime (Psychology Software Tools, Pittsburgh, PA) and the last task was presented by using a specifically designed computer program based in Microsoft Excel. All stimuli were presented through headphone (Sennheiser, HD-25). The hardware system included an external sounds card (Aardvark, direct mix usb[3]) and a notebook computer (IBM, thinkpad R32 with Pentium 4). All tasks were carried in a quite room at the working places of the judges.

Yes/No task

There were five blocks in this task. Each block had one stimuli set. Total 79 stimuli in this task were individually and automatically presented to the judges who were asked to determine whether they detected hypernasality in each trial. Each stimulus was rated twice (totally 158 trials) with 0.5s time interval in order to determine the intra-rater reliability. A binary choice question was visually presented to the judges who were asked to rate whether they detected hypernasality in the stimuli. The judges were expected to respond by typing either “Y” (for yes) or “N” (for no) on the keyboard. The presentation order of the blocks and stimuli in each block were randomized.

Paired comparison task 1

There were three blocks in this task. Each block had one HL stimuli set. In each trial stimuli were paired up according to their AN values (they were differed by 15 AN values, e.g., AN0 and AN15, AN15 and AN30, etc). One-third of the pairs in each block (i.e., three pairs) were paired by the same stimulus. Totally nine pairs were presented to the judges in each block (total 27 pairs in this task). Each pair of stimuli (with 0.5s time interval) was judged twice (totally 54 trials) with 0.5s time interval in order to determine the intra-rater reliability. Binary choice question was visually presented to the judges who were asked to rate whether each pair was “same” or “different” in terms of hypernasality by pressing either “F” (for same) or “J” (for different) on the keyboard. The presentation order of the blocks and paired stimuli in each block were randomized.

Paired comparison task 2

There were five blocks (i.e., three for the HL stimuli sets and two for the LL stimuli sets) in this task. In each block of the HL stimuli set, stimuli were paired up according to their AN values (each pair of stimuli was differed by 10 AN values, e.g., AN0 and AN10, AN10 and AN20, etc). One-third of the pairs in each block (i.e., five pairs) were paired by the same stimulus. Totally 15 pairs were presented to the judges in each block of the HL stimuli sets. Judges were asked to rate whether the paired stimuli were “same” or “different” in terms of hypernasality. Each pair of stimuli (with 0.5s time interval) were judged twice (totally 90 trials) with 0.5s time interval in order to determine the intra-rater reliability. The responding method was same as that in paired comparison task 1 and this method was used in all blocks of this task. The presentation order of the block and the stimuli in each block were randomized. Accuracy was automatically calculated by the end of each block. If 80% or above accuracy was achieved in each block, an extra block of that particular stimuli type would be presented. In the extra block, stimuli were paired with 5 AN values difference (e.g., AN0 and AN5, AN10 and AN15, etc). One-third of the pairs in this block (i.e., five pairs) were paired with the same stimulus. Totally 15 pairs were presented in each extra block and judges were asked to rate whether the paired stimuli were “same” or “different”. Each pair of stimuli was judged twice in order to determine the intra-rater reliability. Same responding method was used. The presentation order of the stimuli was randomized. If lower than 80% accuracy was obtained, no extra block would be presented.

In each block of the LL stimuli sets, stimuli were paired up with every another level (e.g., 1 and 3, 2 and 4, etc). Three pairs of stimuli in each block were paired up by the same stimulus. Each pair of stimuli was rated twice in order to determine the intra-rater reliability of the judges. Totally 16 paired stimuli in each block were presented to the judges who were asked to rate whether the paired stimuli were “same” or “different” in terms of hypernasality. The presentation order of the blocks and stimuli in each block were randomized. Accuracy was automatically calculated by the end of each block. If 80% or above accuracy was achieved in each block, an extra block of that particular stimuli type would be presented. In the extra block, stimuli were paired up with the next level (e.g., 1 and 2, 2 and 3, 3 and 4, etc). Four pairs in this extra block were paired with the same stimulus. Each pair of stimuli was judged twice in order to determine the intra-rater reliability. Totally 18 trials of rating were obtained in each extra block. The presentation order of the stimuli was randomized. If lower than 80% accuracy was obtained, no extra block would be presented.

Perceptual rating task

Judges were asked to rate the severity of hypernasality of each stimulus by using a visual analogue (VA) scale which was found to be a valid and reliable scaling method for perceptual judgment of hypernasality (Cheng, 2006). The instruction of using VA scales was verbally explained to the listeners and was visually shown in the computer program. Judges could listen to the stimuli as many times as they would like.

Data analysis

The intra-rater agreement for hypernasality rating in Yes/No task and paired comparison task 1 and 2 were calculated. As stimulus was presented twice in these tasks, the percentage agreements were calculated by dividing the number of same rating judgment by the total number of stimuli in each task. The value was then multiplied by one hundred percent. A one-way ANOVA was done to compare intra-rater agreement of 11 judges in rating the hypernasality of the HL and the LL vowels sets in the Yes/No task in order to compare and evaluate the two sets of parameters. Since sentence type stimuli was only synthesized by the HL synthesis parameter, sentence type stimuli was not involved in this comparison. For the perceptual rating tasks, the inter-rater reliability of the hypernasality rating was calculated by using intraclass correlation coefficient (ICC) (3, 11).

Results

Yes/No task

Eleven judges were asked to rate whether they detected hypernasality in all stimuli. Each stimulus was rated twice, resulting in 22 judgments for each type of stimuli.

Figure 1: Number of judgments that detected hypernasality in the HL sets.

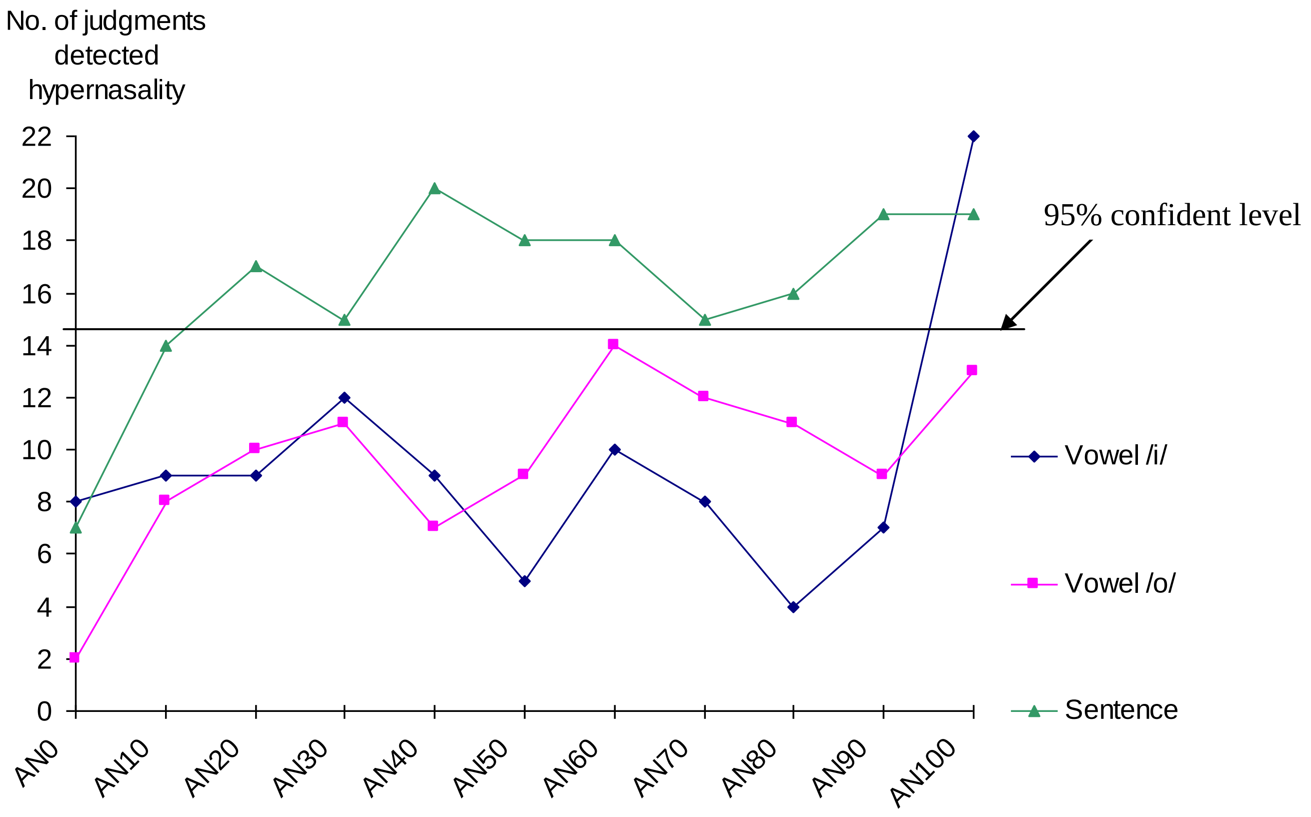
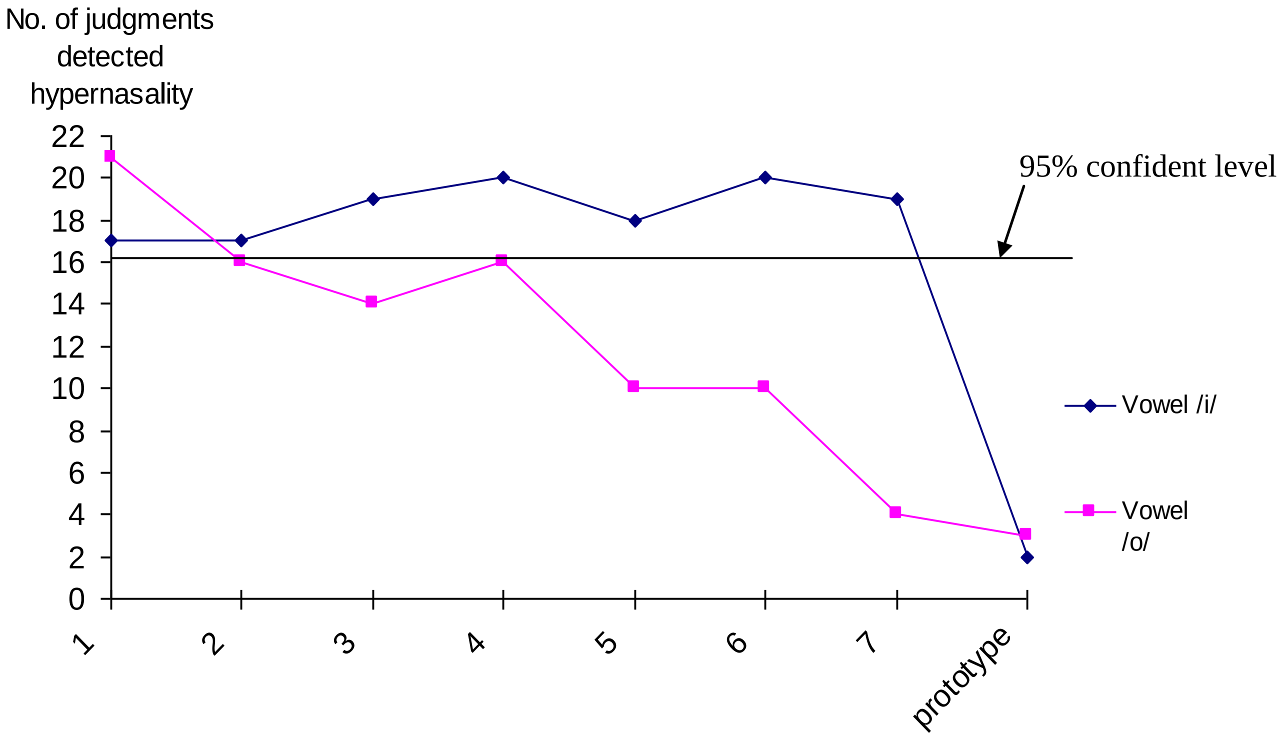


Figure 2: Number of judgments that detected hypernasality in the LL sets.



With a total of 22 judgments for each stimulus, we need at least 16 judgments that rated the stimulus as hypernasal in order to reach the 95% confidence level (Shaughnessy, Zechmeister & Zechmeister, 2003) which was same as the confidence level used by Yiu et al. (2002).

For the HL sets (figure 1), the results showed that the higher the AN values, the higher the number of judges who perceived the stimuli as hypernasal. The number of hypernasal judgments of all vowel /ɔ/ stimuli did not reach the 95% confidence level. In the vowel /i/ stimuli set, only AN100 reached this 95% confidence level. Seven out of 11 sentence stimuli (i.e., AN20, AN40, AN50, AN60, AN80, AN90 and AN100) reached the 95% confidence level.

Figure 2 showed the number of judgments that detected hypernasality in the LL sets. The confidence level was drawn on the graph. Three vowel /ɔ/ stimuli (i.e., 1, 2 and 4) and seven out of eight vowel /i/ stimuli (i.e., from 1 to 7) reached the 95% confidence level.

Paired comparison task

Paired up stimuli were given to the judges who were asked to discriminate whether they were the “same” or “different” in terms of hypernasality. Since only sentence stimuli from the HL set and vowel /i/ stimuli in the LL set reached the 95% confidence level in the Yes/No task, only these two types of stimuli would be presented here.

Figure 3: Number of correct rating of the HL sentences set with 15 AN values differences.

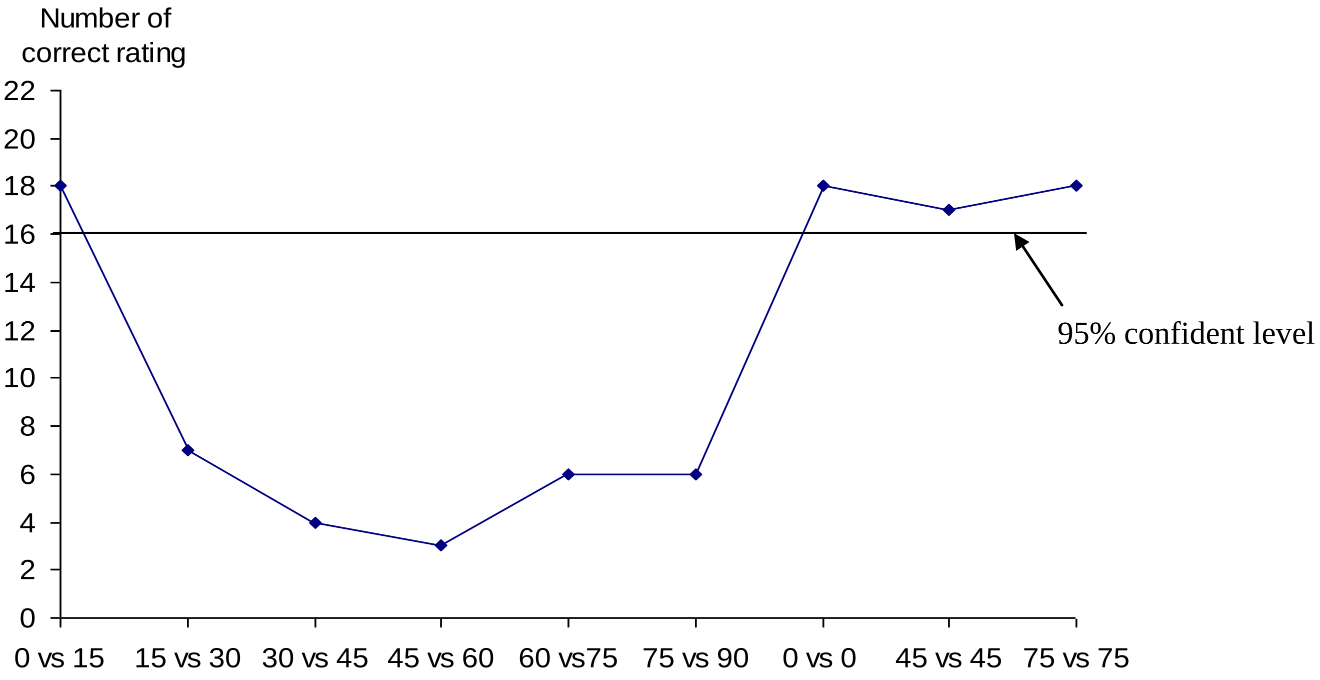


Figure 4: Number of correct rating of the LL vowel /i/ paired with every another level.

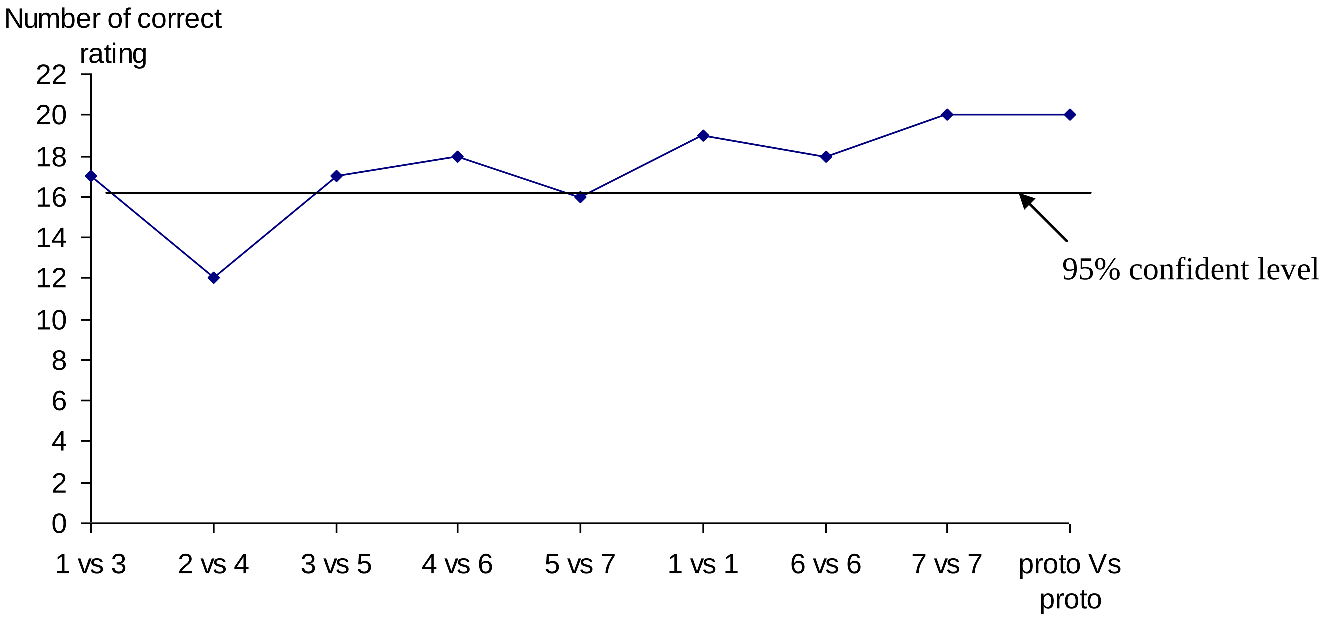
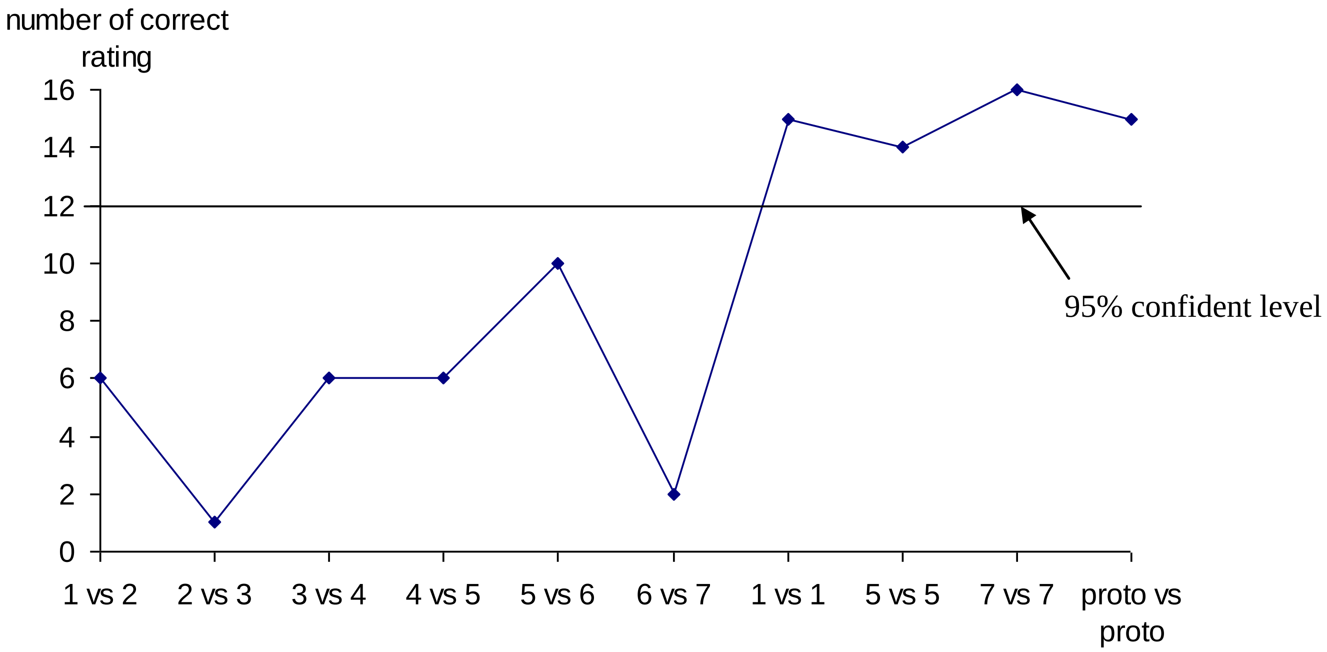


Figure 5: Number of correct rating of the LL vowel /ɔ/ paired with next level.



For the LL vowel /i/ stimuli set (figure 4), the 95% confidence level was drawn (Figure 5). It showed that 7 out of 8 pairs reached the 95% confidence level. However, for the LL vowel /ɔ/ stimuli set, only 4 out of 10 pairs reached the 95% confidence level and these four pairs were paired with same stimuli. This showed that judges were able to determine the differences of vowel/i/ stimuli at every another level (except the pair Level 2 vs Level 4) but not for the pairs with the next level.

Perceptual rating task

Judges were asked to rate the severity of hypernasality of all 79 stimuli on a 10 cm continuous line. In the yes/no task, only the HL sentence set and the LL vowel /i/ set reached the 95% confidence level (i.e., had more than 16 out of 22 judgements of detecting hypernasality), therefore, only these two stimuli sets would be rated in the perceptual rating task. Table 2 listed the results of the mean visual analogue (VA) rating of 11 judges.

Table 3: *Means of visual analogue (VA) rating of the HL sentence and the LL vowel /i/ sets.*

The HL sentence	Mean	SD	Level of the LL vowel /i/	Mean	SD
AN 0	1.58	2.70	1	6.01	2.64
AN 5	1.76	2.59	2	4.53	3.57
AN10	2.35	2.35	3	4.90	3.06
AN15	2.92	2.80	4	4.99	3.26
AN20	2.75	2.55	5	5.34	2.48
AN25	3.10	2.97	6	5.12	3.36
AN30	3.19	2.64	7	5.64	2.74
AN35	3.56	2.47	Prototype	1.72	2.26
AN40	2.93	2.60			
AN45	3.01	2.66			
AN50	4.15	2.48			
AN55	3.62	2.73			
AN60	3.35	2.71			
AN65	3.19	2.47			
AN70	3.85	2.34			
AN75	3.84	2.75			
AN80	3.03	2.56			
AN85	3.72	2.88			
AN90	3.84	2.81			
AN95	3.55	2.95			
AN100	3.89	2.33			

Note: AN (the cross-sectional area of the opening of the nasal cavity) – variable of the high-level synthesis parameters in Sensimetrics' HLSyn Speech Synthesis System (1997).

The mean VA rating of all the HL sentences stimuli ranged from 1.58 to 4.15 and from 1.72 to 6.01 for vowel /i/ stimuli.

Intra-rater agreement

All synthesized signals were rated twice by judges in each task except the perceptual rating task. Percentage agreement was calculated in order to determine the intra-rater reliability. Table 3 summarized the mean, standard deviation and range of percentage agreement of the judges in each of the tasks (i.e., Yes/No task and two paired comparison tasks).

Table 4: *Percentage agreement (%) of judges in Yes/No task and two paired comparison tasks.*

	Yes/No task		Paired comparison task 1	Paired comparison task 2		
	HL sentence	LL vowel /i/	HL sentence (different by 15 AN values)	HL sentence (different by 10 AN values)	LL vowel /i/ (with every another level)	LL vowel /i/ (with next level)
	(N=11)	(N=11)	(N=11)	(N=11)	(N=11)	(N=8)
Mean	68.60	88.64	73.74	75.15	76.77	81.25
Standard deviation	23.83	15.26	14.29	17.91	11.61	11.26
Range	27-100	50-100	56-100	40-100	55-100	60-90

The mean percentage agreement ranged from 68.60% (in the HL sentence) to 88.64% (in the LL vowel /i/) in the three tasks. A one-way ANOVA was done to compare intra-rater agreement of 11 judges in rating the hypernasality of vowels in two types of synthesis parameters (i.e., the HL and the LL) in the Yes/No task in order to compare and evaluate the two sets of parameters. Since sentence set was only synthesized by the HL synthesis parameter, it was not involved in this comparison. Results showed that there was a significant difference between the judges' rating on the HL and the LL sets ($F[3,40] = 4.405$, $p<0.05$).

Inter-rater reliability

Intraclass correlation coefficient (ICC) (3, 11) was calculated from perceptual rating task in order to determine how closely judges agreed with each other on the hypernasality ratings.

Table 5: *ICC (3, 11) values of the HL sentence and the LL vowel /i/ in perceptual rating task.*

	HL set			LL set	
Stimuli	Vowel /i/	Vowel /ɔ/	sentence	vowel /i/	Vowel /ɔ/
ICC (3, k)	0.948	0.696	0.622	0.865	0.788

Discussion

The main purpose of the current study was to determine whether reliable hypernasal stimuli could be synthesized by using a Klatt synthesizer. It could be achieved through investigating two objectives, i.e., comparing two different sets of parameter (the high-level and the low-level synthesis parameters) in synthesizing hypernasal stimuli and determining the most preferable type of stimuli among sentence and sustained vowel /i/ and /ɔ/.

High-level versus low-level synthesis parameters

The first objective of the current study was to compare which type of parameters (i.e., the high-level and the low-level synthesis parameters) would be preferable for creating synthesized hypernasal stimuli.

<u>*Yes/No task*</u>

Vowel /i/ and /ɔ/ were synthesized by both the high-level and the low-level parameters and were presented in the yes/no task. It was noted that only the LL vowel /i/ set was rated as hypernasal with 95% confidence level. Three out of eight stimuli in the LL vowel /ɔ/ set were rated as hypernasal with the same confidence level. All stimuli from the HL vowel /i/ set and the HL vowel /ɔ/ set failed to reach the 95% confidence level. The results suggested that the LL vowel /i/ and the LL vowel /ɔ/ were more representative of hypernasality when compared with the HL vowel /i/ and the HL vowel /ɔ/. This meant that the LL synthesis parameters from the Klatt synthesizer (Klatt & Klatt, 1990) were more preferable than the HL synthesis parameters from the same synthesizer. Researchers developing synthesizing hypernasal speech stimuli should explore different ways of synthesis other than the high-level synthesis parameters from Sensimetrics' HLSyn Speech Synthesis System (1997).

In the low-level vowels, seven synthesis parameters (i.e., F1, B1, F2, FNP, FNZ, BNP and BNZ) were selected and varied independently according to the way proposed by Zraick and Liss (2002). The result of the current study provided support to their works. However, it should be noted that five parameters (i.e., F1, B1, F2, BNP and BNZ) of non-nasal vowel /i/ were slightly increased in the values when compared with the oral stimuli synthesized by Zraick and Liss (2002) in order to achieve a natural sounding stimulus. Besides, the range of F1 and F2 in /ɔ/ stimuli were narrowed and the values of B1, FNP, FNZ, BNP and BNZ were slightly reduced when compared with /i/ stimuli in order to have a natural sounding vowel /ɔ/. This suggested that natural sounding hypernasal stimuli could be created by synthesizer and could be used to develop as anchors for perceptual training on rating hypernasality which was supported by Chan and Yiu (2002).

Paired comparison task

In order to apply the synthesized signals into a perceptual training program, it is important to first determine the degree of differences judges could differentiate among stimuli. In the paired comparison task, for the HL sentence set, judges were unable to determine the differences between stimuli with 15 AN values difference, except for the AN0 and AN15 pair. The results revealed that 15 AN values difference was not noticeable enough for judges to differentiate differences between stimuli in terms of hypernasality. However, the nasality difference between prototype sentence and AN15 stimulus was detected by the judges. This showed that the high-level synthesis parameter (i.e., AN value) was responsible for synthesizing hypernasality. But 15 AN values difference was a small, not representative and unnoticeable increment. This poor noticeable difference of the HL sentence set further supported that the high-level synthesis parameter was not a preferable synthesis parameter for creating hypernasal speech signals.

For the LL vowel /i/ set, the results showed that judges were able to detect the differences between stimuli when the pair differed by two levels (e.g.., 1 and 3) but not when the pairs included consecutive levels. The results revealed that the original sets of synthesis parameters proposed by Zraick and Liss (2002) had enough noticeable differences which could be detected by judges. When intermediate levels were added, listeners failed to detect the differences.

Perceptual rating task

Visual Analogue (VA) scales were used to determine the severity of hypernasality of synthesized stimuli. Table 3 showed that the severity of the HL sentence set was generally increased (except AN 50 which was rated as the highest level of severity) when AN values increased. This meant that AN100 had the highest degree of hypernasality. The exceptional rating on AN50 could be explained by the order of presentation which was presented immediately after AN0. The nasality difference between the non-nasal stimuli (AN0) and the AN50 might confuse the judges and as a result, higher severity rating on AN50 obtained. Secondly, the severity of the LL vowel /i/ set was generally decreased from 1 to 7 with exceptions which might due to the rigid order of presentation of the stimuli. The results showed that stimulus 1 was the most severe stimulus in terms of hypernasality and the stimulus 7 was the least severe. More specifically, severity of hypernasality in the low-level synthesis parameters decreased when F1 and B1 decreased and F2, FNP and FNZ increased. Thirdly, the mean of the VA ratings of the LL vowel /i/ set (i.e., 4.78) were higher than the

HL sentence set (i.e., 3.20). The range of the means of VA rating of the LL vowel /i/ set (i.e., 4.29) was larger than that of the HL sentence set (i.e., 2.57). This showed that LL vowel /i/ set represented a larger continuum on the severity of hypernasality than the HL sentence set. These results supported the finding in the previous parts, in which the low-level synthesis parameters were more preferable than the high-level synthesis parameters. Lastly, it was noted that the HL sentence set were rated between 1.58 and 4.15 and the LL vowel /i/ set ranged between 4.53 and 6.01. This suggested that the HL sentence set were synthesized as mild to moderate severity in terms of hypernasality in the continuum and the LL vowel /i/ set were synthesized as moderate severity. Both sets of stimuli did not represent the entire range of severity of hypernasality. Researchers should explore other synthesizer in order to synthesize representative stimuli all over the severity continuum.

Intra-rater reliability

The intra-rater reliability was calculated by percentage agreement in Yes/No task and two paired comparison tasks (i.e., the HL sentences with 10 and 15 AN values differences and the LL vowel /i/ paired with the every another level and with the next level) in order to determine how judges' rating agreed themselves. The percentage agreement ranged from 68.60% for the HL sentence set in Yes/No task to 81.25% for the LL vowel /i/ set paired with the next level. It was a relatively higher percentage agreement when compared with 71% intra-rater percentage agreement of experienced listeners of mode rating in rating nonnasal sentence by using four point EAI scale done by Laczi, Sussman, Stathopoulos and Huber (2005) and 58% percentage agreement of experienced listeners of mode rating in rating four types of sentences by using a 5-point scale done by Karling, Larson, Leanderson, Galyas, & Serpa-Leitao (1993). This relatively higher intra-rater reliability indicated that the judges recruited in the current study had their own consistent rating in rating synthesized hypernasal stimuli.

Intra-rater agreement analysis showed that the judges' ratings in the HL sets were significantly lower (68.60%) than the LL sets (88.64%) from the Yes/No task. It could be concluded that judges had more consistent nasality ratings when judging the LL stimuli than the HL stimuli. This also supported that the LL synthesis parameters was more preferable than the HL synthesis parameters for synthesizing nasality.

Inter-rater reliability

The inter-rater reliability was calculated by intra-class correlation (ICC) (3, 11) in the perceptual rating task in order to learn how well the judges' ratings agreed with each other. The results (Table 5) showed that the ICC values of the LL sets ranged from 0.788 to 0.865 and that of the HL sets ranged from 0.622 to 0.948. The relatively narrow range of the ICC values in the LL sets revealed that judges showed similar agreement with each other across different sets of stimuli in the LL sets than in the HL sets. This also supported that the LL synthesis parameters were more preferable than the HL synthesis parameters.

Vowel versus sentence

The second objective of the current study was to determine which type of stimuli (i.e., sentence or vowels) was preferable for perceptually evaluating hypernasality.

Yes/No task

In the yes/no task, five sets of stimuli were judged by the listeners. The results showed that the HL sentence set and the LL vowel /i/ set reached 95% confidence level.

For the LL sets, sustained vowel /i/ and /ɔ/ were synthesized as stimuli in order to compare the effect of vowel context in perceptual evaluation of hypernasality. It was noted that the LL vowel /i/ set had more stimuli reached 95% confidence level than that of the LL vowel /ɔ/ set. This showed that, in a preferable method of synthesizing hypernasal signals, vowel /i/, a high-front vowel, was more preferred than low-back vowel (i.e., vowel /ɔ/) in perceptual evaluating synthesized hypernasal speech stimuli.

For the HL sets, the result revealed that only the sentence stimuli were synthesized as hypernasal. This suggested that sentence was a preferable type of stimuli for perceptual evaluation of hypernasality. Concurrent with the findings from Cheung (2004), in which, for real voices, sentences without nasal sounds were more preferable for perceptual evaluation of hypernasality than isolated vowels and monosyllabic words, both real and synthesized hypernasal sentences were preferable in perceptually evaluating hypernasality. Since connected speech was more natural and more representative of "voice" used in daily speech activities (Yiu et al, 2002), sentence stimuli should be used in future study for developing anchors for perceptual training in evaluating hyperanasality. However, it should be noted that, from the current study, the sentence stimuli were only synthesized by the high-level synthesis parameter which was not a preferable way of synthesizing hypernasality, as shown in the current study. Further exploration on other synthesis parameters (e.g. the LL synthesis parameters from Klatt synthesizer (Klatt & Klatt, 1990)) in synthesizing sentence stimuli was recommended.

Inter-rater reliability

Table 5 showed the ICC values in perceptual rating task. By comparing the results with Cheung (2004), in which the ICC values for sentences ranged between 0.92 and 0.96 and ranged between 0.40 and 0.76 for isolated vowel /a/ and /i/, the ICC values of the HL sentence set (0.622) was relatively low and that of sustained vowels (ranged from 0.696 to 0.948) was relatively high. The relatively low ICC values in the HL sentence set suggested that judges showed poor agreements with each other in rating the severity of hypernasality. The results were different from Cheung (2004). It could be explained by the use of the non-preferable way of synthesizing hypernasality in the HL sentence set. Further exploration on other synthesis parameters (e.g. the LL synthesis parameters from Klatt synthesizer (Klatt & Klatt, 1990)) in synthesizing sentence stimuli was recommended. The relatively high ICC values in the sustained vowels suggested that judges agreed with each other in rating the severity of hypernasality in vowels. The results also showed that both the LL and the HL vowel /i/ sets (0.865 and 0.948 respectively) had a higher ICC values than that of vowel /ɔ/ sets (0.788 and 0.696). This revealed that judges’ ratings agreed with each other more in the vowel /i/ sets than in the vowel /ɔ/ sets and also sentence set. This supported that vowel /i/ was also a preferable stimuli in evaluating synthesized hypernasality.

In summary, the current study found that the low-level synthesis parameters were more preferable than the high-level synthesis parameter in synthesizing hypernasal speech signals. Besides, among connected speech and the vowels, sentence and vowel /i/ was more preferable than vowel /ɔ/. It was suggested that in synthesizing hypernasal speech signals, either connected speech or vowel /i/ could be included. High-front vowels rather than low-back vowels were also recommended. Low-level synthesis parameters or other available synthesizers should be explored in order to further compare which stimuli (sentence or vowel /i/) was the most preferable.

Limitation of current study

Firstly, in the processing of synthesizing signals by using low-level synthesis parameters, it was noted that the parameters used by Zraick and Liss (2002) involved difficult manipulation. For example, F1 and F2 were different in different vowels. Manipulating these two parameters would lead to a noticeable vowel distortion on the stimuli (i.e., /ɔ/ → [a]). The borderline values between vowel /ɔ/ and /a/ were used in the current study in order to prepare seven levels of stimuli. Therefore, some of the low-synthesis signals were perceived as /a/ by listeners rather than /ɔ/ even the F1 and F2 values were acoustically within vowel /ɔ/. If the low-level synthesis parameters are used to synthesize hypernasal signals, perceptual judgment on the vowel context should also be considered in order to achieve a higher agreement and then minimize the confusion of vowels. Secondly, no LL sentence stimuli were synthesized in the current study. The study failed to directly compare the LL sentence set with the LL vowels sets and therefore was unable to determine which type of stimuli were the most preferable. LL sentence set was highly recommended to be synthesized in the future research studies. Thirdly, each stimulus was presented once in VA scaling in the current study. No intra-rater reliability was calculated. This could not evaluate the listeners' agreement in this task. It was recommended that the stimuli should be presented twice in order to evaluate the intra-rater reliability. Lastly, in the VA scaling task, the HL sets were always presented before the LL sets. There may be order effect. It is possible that listeners showed facilitation or fatigue after exploring to the HL sets and then perform better or worse in rating LL sets, resulted in an unreliable result. It was suggested that precautions for order effect should be taken when similar study is carried out in the future.

Conclusion

The current study showed that the low-level synthesis parameters from Klatt synthesizer (Klatt & Klatt, 1990) were reliable and valid parameters for synthesizing hypernasal signals. Second, it was found that connected speech and vowel /i/ were better than vowel /ɔ/ for synthesizing and perceptually evaluating hypernasal signals.

Acknowledgement

I would like to express my sincere gratitude to Dr. Karen Chan for her nice guidance and supervision during the course of this dissertation. I would like to also thank Dr. Tara Whitehill for her temporary supervision during the leave of Dr. Karen Chan. Thanks are also extended to all the subjects who participated in this dissertation for their selfless devotion of time. Lastly, I would like to thank all participants who made this dissertation possible.

References

Abramson, A.S., Nye, P.W., Henderson, J., & Marshall, C.W. (1981). Vowel height and the perception of consonantal nasality. *Journal of the Acoustical Society of America, 70,* 329-339.

Boone, McFarlane & Von Berg (2005). *The Voice and Voice therapy (7th ed.),* Boston, Pearson Education, Inc.

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 (2006). A Comparison of Two Perceptual Voice Evaluation Training Programs for Naïve Listeners. *Journal of Voice, 20(2)*, 229-241.

Cheng, T.H.D. (2006). *Direct Magnitude Estimation Versus Visual Analogue Scaling in The Perceptual Rating of Hypernasality*. Unpublished BSc dissertation. Hong Kong. The University of Hong Kong.

Cheung, S.C.J. (2004). *The Effects of Stimulus and Modulus on Perceptual Rating of Hypernasality.* Unpublished BSc dissertation. Hong Kong. The University of Hong Kong.

Dworkin, J.P. Marunick, M.T. & Krouse, J.H. (2004). Velopharyngeal dysfunction: Speech characteristics, variable etiologies, evaluation techniques, and differential treatments. *Language, Speech and Hearing Services in Schools, 35,* 333-352.

Sensimetrics's HLSyn Speech Synthesis System (1997). *HLsyn High-Level Parameter Speech Synthesis System Version 2.2. User Interface Manual.* Sensimetrics Corporation: Cambridge.

Karling, J., Larson, O., Leanderson, R.,Galyas, K., & Serpa-Leitao, A. (1993). NORAM – and instrument used in the assessment of hypernasality: a clinical investigation. *Cleft Palate Craniofac Journal. 30*. 135-140.

Kataoka, R., Zajac, D.J., Mayo, R., Lutz, R.W., and Warren, D.W. (2001). The Influence of Acoustic and Perceptual Factors on Perceived Hypernasality in the Vowel [i]: A Preliminary Study. *Folia Phoniartica et Logopaedica, 53*, 198-212.

Kent, R.D. & Read, C. (2002). *Acoustic analysis of speech, 2nd edition*. Singular: Thomson Learning.

Klatt, D.H. & Klatt, L.C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal Acoustical Society of America, 87, (2)*, 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 for future research. *Journal of Speech and Hearing Research, 36*, 21-40.

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

Laczi, E., Sussman, J.E., Stathopoulos, E.T., & Huber, J. (2005). Perceptual Evaluation of Hypernasality Compared to HONC Measures: The Role of Experience. *The Cleft Palate – Craniofacial Journal. 42, 2*. 202-211.

Lewis, K.E. Watterson, T.L. & Honghton, S.M. (2003). The influence of listener experience and academic training on rating of nasality. *Journal of Communication Disorders, 36*, 49-58.

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

Shaughnessy, J.J., Zechmeister, E.B., & Zechmeister, J.S. (2003). *Research Methods in Psychology, sixth edition*. New York: McGraw-Hill.

Sherman, D. & Hall, P.K. (1978). Nasality and precision of articulation. *Perceptual and Motor Skills, 46,* 115-118.

Whalen, D.H. & Beddor, P.S. (1989). Connections between nasality and vowel duration and height. *Language, 65*, 457-486.

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

Wikipedia contributors. (2007). Just noticeable difference. *Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Just_noticeable_difference&oldid=116195686.

Yiu, E. M.L., Murdoch, B., Hird, K., & Lau, P. (2002). Perception of synthesized voice quality in connected speech by Cantonese speakers. *Journal Acoustical Society of America, 112,* 1091-1101.

Zraick, R.I. & Liss, M.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.

Zraick, R.I., Liss, M.J. & Beals, S.P. (2000). Multidimensional scaling of nasal voice quality. *Journal of Speech, Language, and Hearing Research, 43,* 989-996.

Zraick, R.I., Wendel, K., & Smith-Olinde, L. (2004). The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice, 19(4),* 574-581.