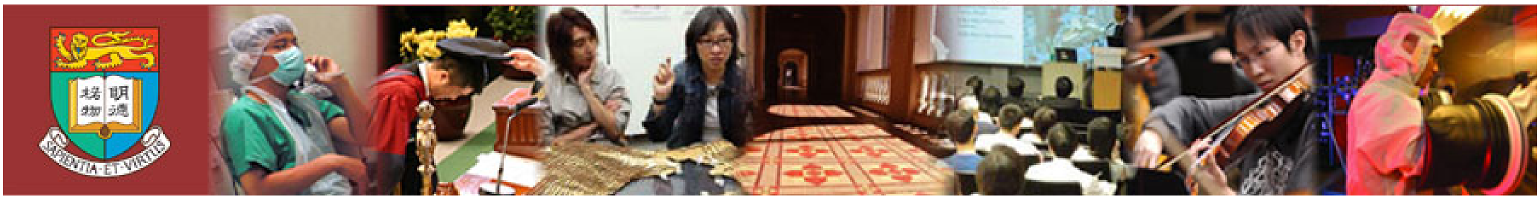


Title	Effects of cultural and linguistic backgrounds on perceptual voice quality rating
Other Contributor(s)	University of Hong Kong
Author(s)	Ho, Elaine Mandy
Citation	
Issued Date	2005
URL	http://hdl.handle.net/10722/56211
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Effects of Cultural and Linguistic Backgrounds on Perceptual Voice Quality Rating

Ho, Elaine Mandy

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Speech and Hearing Sciences), The University of Hong Kong, May, 6th, 2005

Abstract

Perceptual voice judgment is a standard procedure in voice quality evaluation. However, the cultural and linguistic backgrounds of the judges often affect its reliability. Use of anchors has been shown to improve the reliability of the procedure. Therefore, this study aimed to develop female synthesized anchors in Cantonese, English, and Putonghua and investigate the reliability of perceptual voice rating across speakers of three native languages using an anchor-matching method. Ten native Cantonese speakers, 10 native English speakers and 10 native Putonghua speakers were recruited to rate the severity of breathiness and roughness of pathological voice samples in Cantonese, English and Putonghua. Results showed high intra-rater agreement (average 87.23%) and inter-rater reliability (average 0.94) across listener groups and stimuli sets. The Putonghua listeners in this study demonstrated significantly lower intra-rater agreement than the Cantonese listeners, suggesting possible cultural and linguistic differences in perceptual voice evaluation between Putonghua and Cantonese listeners.

Introduction

Perceptual voice judgment is considered as a standard procedure in voice quality evaluation. It is widely used in research and clinical voice assessment for rating the severity of voice quality impairments. The procedure is often used as a standard for validating acoustic and aerodynamic measurements (e.g., De Krom, 1995; Yumoto, Sasaki, & Okamura, 1984). However, the reliability of the procedure has been a major issue. Both inter- and intra-rater reliability measures were reported to vary greatly. These measures range from 17% to 98% due to various reasons (see review by Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). First, the subjective nature of the procedure inevitably leads to individuals' variability. Second, the lack of widely accepted definitions for specific voice qualities also results in the quality variability as listeners may attend to a quality not precisely specified by the definition. Third, the multidimensional nature of pathological voices may lead to the perception of different features of voice qualities by different listeners (e.g., Chan & Yiu, 2002; Gerratt & Kreiman, 2001; Hartelius, Theodoros, Cahill, & Lillvik, 2003; Yiu, Chan, & Mok, submitted). Fourth, differences in the use of rating scales and rating samples, together with the experiences or backgrounds of the listeners may also contribute to low reliability (Gerratt & Kreiman, 2001; Kreiman et al., 1993; Kreiman, Gerratt, & Precoda, 1990).

Recently, the use of explicit external anchors has been shown to achieve the reliability of perceptual voice rating in native Cantonese and English speakers up to 75% (e.g., Chan & Yiu, 2002; Gerratt & Kreiman, 2001; Kreiman et al., 1993; Yiu et al., submitted). It has been proposed that the source of listener variability caused by the use of unstable internal standards would be reduced with the provision of anchors.

The types of samples to be rated can also contribute to low reliability (see review by Kreiman et al., 1993). Sustained vowels have been used in most studies. However, it has been argued that connected speech is more representative of daily voice use (e.g. Chan & Yiu, 2002;

Hammerberg, Fritzell, Gauffin, Sundberg, & Wedin, 1980; Wolfe, Cornell, Fitch, 1995; Yiu et al., submitted; Yiu, Worrall, Longland, & Mitchell, 2000). Connected speech carries important transitory information, such as coarticulation, voice onset and termination, voice break, and supraglottic articulatory dynamics that are important to voice quality perception (Revis, Giovanni, Wuyts, & Triglia, 1999; Wolfe et al., 1995). Yet, these elements, which may be vulnerable to cultural or linguistic effects, are not present in sustained vowels. With the relative scarcity of cross-language studies, it is yet unknown if language itself, as well as difference in the cultural and linguistic backgrounds between judges, may have different effects on the reliability.

With over half of the world's population being bilingual (Hartelius et al., 2003), and specifically in Hong Kong, with an increasing number of Putonghua speakers emigrating from Mainland China to Hong Kong, it is important to investigate the effects of familiarity of language and cultural backgrounds on the reliability of perceptual voice evaluation. Studies on how native judges would rate Cantonese and Putonghua pathological voice samples differently or similarly would be of research and clinical significance as it is necessary to develop a cross-culturally valid research and clinical tool. It would be useful to investigate how reliable listeners would rate pathological voice quality in both native and non-native languages utilizing an anchor-matching paradigm. So far, there has been no standard voice assessment protocol available for Putonghua. With this lack of research and clinical tool on voice quality rating in Putonghua, there is a need to develop an assessment protocol and more research for the Putonghua speakers, who constitute the largest population in the world.

This study aimed to investigate the effects of cultural and linguistic backgrounds on perceptual voice quality rating. In this study, voice quality rating performances of judges with different cultural and linguistic backgrounds, including native Cantonese, native Putonghua and native English judges were compared.

Contradictory results concerning the roles of cultural and linguistic backgrounds on voice quality perception and production have been documented. Different voice quality production across cultural or languages groups has been reported. According to Esling (2000), voice quality is socially acquired and is relatively idiosyncratic. For example, results from the study by Majewski, Hollien and Zalewski (1972) showed that Polish male speakers have a higher speaking fundamental frequency than American male speakers by 8Hz. However, it has been argued that the effects of cultural and linguistic differences might have been confounded by the difference in physical size of the subjects. Bruyninckx, Harmegnies, Llisterri and Poch-Olivé (1994) also found significant difference in the long-term average spectrum in the two languages spoken by Catalan-Spanish bilinguals. Additionally, falsetto and harshness was noticed in African-English when compared to White American English (Esling, 2000) and higher-pitch was noted in Australian English when compared to Swedish (Hartelius et al., 2003).

The above review indicated that both cultural and linguistic backgrounds do have a role on voice quality production. Nonetheless, it is inconclusive of the effects of cultural and linguistic backgrounds on voice quality perception as studies in this area have shown equivocal results. Anders, Hollien, Hurme, Sonninen and Wendler (1988) compared American, Finnish and German listeners in rating the severity of dysphonic voice qualities in German. No statistically significant differences were found. However, the American listeners were milder in their judgments than the Finnish and German listeners. Nonetheless, the difference in experience and training between listener groups in this study was a confounding variable which warranted caution on the interpretation of data. Hartelius et al. (2003) investigated the effects of familiarity of languages on severity rating of various speech and voice qualities. They compared Australian and Swedish listeners on assessing speech characteristics and voice quality of both Australian and Swedish patients with multiple sclerosis. Both listener groups showed high reliability. Nonetheless, the Australian judges showed higher mean agreement, although not statistically significant, and were less critical than the Swedish judges, despite their lack of knowledge of the Swedish language. However the small sample size in both listener groups and the differences in their professional backgrounds and linguistic commands warranted caution over generalization of the results.

Comparable results were also found in a study by Yamaguchi, Shrivastav, Andrews and Niimi (2003). They compared voice quality ratings by American and Japanese listeners using the GRBAS scale. Similar ratings for both groups of listeners were documented and the authors concluded that neither cultural nor linguistic factors have much effect on perceptual ratings of voice qualities. Yet, a number of methodological problems were observed in this study. First, the levels of experience in both groups of listeners were different. Second, the rating procedures were different between the two groups. Third, a four-point equal-appearing interval scale was used, which was considered to be relatively insensitive (Kreiman et al., 1993) with a respective agreement by chance as high as 38% (Wuyts et al., 1999). Lastly, sustained vowels were used, as mentioned earlier, that this type of rating sample carries relatively few cultural or linguistic cues. With these methodological problems, it is still not conclusive as to how cultural and linguistic factors affect perceptual ratings.

In addition to the types of rating sample, the type of quality to be rated is also a significant factor that affects the reliability of perceptual voice evaluation. Breathiness and roughness are chosen in the present study for investigation as they have been widely investigated and used in clinical procedures. They also demonstrated higher reliability than other perceptual qualities in perceptual judgment tasks (Gerratt et al., 1993; Hammerberg et al., 1980; Wolfe et al., 1997) and described the physiological consequences of a wide range of voice pathologies (Chan & Yiu, 2002; Kreiman et al., 1993; Martin & Wolfe, 1996; Yiu et al., submitted). Breathiness is referred to as audible air escape and frication noise caused by incomplete closure of the vocal folds during voice production (Hirano, 1981). Roughness is perceived as irregular quality and lack of clarity resulted from aperiodic vocal fold vibrations (Chan & Yiu, 2002).

Studies have shown that the provision of external anchors in perceptual voice evaluation helps to improve listeners' reliability (e.g., Chan & Yiu, 1992; Gerratt, Kreiman, Antonanazas-Barroso, & Berke, 1993; Kreiman & Gerratt, 2000; Yiu et al., submitted). The present study used synthesized voice anchors over natural anchors for the following reasons. First, synthesized signals can be manipulated individually and systematically to represent different types of quality and with different degree of severity (e.g., Kreiman & Gerratt, 2000; Yiu et al., submitted; Yiu, Murdoch, Hird, & Lau, 2001). On the other hand, a large database of natural voice samples must be available to allow the selection of appropriate anchors, which is sometimes not feasible. Second, synthesized signals may be easily created to exhibit only a specific voice quality by adjusting appropriate parameters. Yet natural dysphonic samples are often multidimensional in nature. It is not easy to find natural dysphonic samples that differ only in one dimension with different degrees of severity. For example, it is not easy to find one sample which is 'twice' as severe as another sample in the breathiness rating but not differ in other quality ratings. Lastly, synthesized signals are relatively less acoustically complex when compared to natural stimuli (e.g., Kreiman et al., 1990; Wolfe, Fitch, & Martic, 1997; Yiu et al., submitted), which enhance investigations on the relationships of acoustic measurements and individual voice quality perception. The relative simplicity and ease of reproducibility of the stimuli also allows replication of studies for further investigation.

This study had two objectives. First, it aimed to develop appropriate sets of female Cantonese, English, and Putonghua synthesized anchors of various degrees of severity in breathiness and roughness. Second, it aimed to investigate the reliability of perceptual voice evaluation across speakers of three different native languages and their respective stimuli. An anchor-matching paradigm was used and listeners were asked to select the synthesized anchor that best matched with the breathy and rough quality of the target voice sample in corresponding languages. It was hypothesized that similar and high intra- and inter-rater reliability would be shown across listeners and stimuli and that there would be minimal effects of cultural and linguistic backgrounds on perceptual voice evaluation when a anchor matching paradigm was used. Thus, it was expected that the use of anchor-matching paradigm would be a reliable mean of perceptual judgment for both research and clinical use in Cantonese, English and Putonghua.

Pilot Study

A pilot study was first conducted to develop sets of female synthesized breathy and rough anchors in Cantonese, English, and Putonghua. The anchors were created to cover a range of severity levels, from normal, mild, moderate to severe, and each of them was perceptually different in terms of severity from the other anchors in the same breathy or rough continua.

Subjects

Three listeners (two female and one male) were asked to be judges. They were all native Cantonese speakers fluent in all three languages. They all have two or more years of clinical experience in rating voice qualities.

Preparation of Synthesized Stimuli

Female prototype sentences with normal voice quality were first synthesized. The Cantonese series used the sentence /pa pa ta k—k—/ (meaning father hits brother), the English series used the phrase ‘a baby girl’, and the Putonghua series used the sentence /pa pa ta k□ k□/ (meaning father hits brother). These sentences were chosen as all the consonants were unaspirated plosives which avoided any possible turbulent noise caused by fricatives or aspirated consonants that might mask the breathy voice quality (Chan & Yiu, 2002; Yiu et al., submitted). The prototype sentences were synthesized at a sampling rate of 11,025 Hz using the HLSyn Speech Synthesis System (version 2.2; Sensimetrics; Microsoft Window version). The

HLSyn synthesizer is a commercially available Klatt synthesizer that has been used to create signals with specific voice qualities in different degree of severity (e.g. Bangayan, Long, Alwan, Kreiman, & Gerratt, 1997; Chan & Yiu, 2002; Klatt and Klatt, 1990; Yiu et al., 2002; Yiu et al., submitted). Formant information from natural speech produced by native speakers were extracted using LPC (Kay Elemetrics ASL Model 5104) and used in the synthesis. Fine adjustments of the formant values were manipulated to make the synthesized signals to sound as natural as possible.

The synthesis parameters (amplitude of aspiration [AH], diplophonia [DI] and amplitude of voicing [AV]) that were used in the study by Yiu et al. (submitted) were employed in the present study to create continua of breathy and rough qualities.

Breathiness. The manipulation of AH values have been shown to result in the perception of breathiness (Klatt & Klatt, 1990; Yiu et al., 2002; Yiu et al., submitted). Yiu et al. (2002) also showed that breathiness was perceived when AH was increased from 40 dB to 50 dB. Therefore, AH 50 was set as the starting level for creating breathy stimuli in this study. The AH value was increased in steps of 5dB until it reached its maximum value of 80dB. A total of seven breathy signals were synthesized (Table 1). By including the prototype signal, there were a total of eight signals in the breathy continuum. Amplitude clippings in the waveforms of the signals were noted when AH reached 80 dB level. Therefore, the default values of gain control of voicing (GV), aspiration (GH) and frication (GF) were reduced to 52dB for all eight signals in order to avoid any possible amplitude clippings (the original default level was 60 dB for each of these three parameters).

Roughness. The manipulation of DI values has been shown to result in the perception of roughness. Previous studies have shown that signals with AH and AV values both set to 80dB and DI set to 2% were perceived to be primarily rough with a slight vocal fry (Yiu et al., 2002; Yiu et al., submitted). The values of GV, GH and GF were also set to 52dB for all signals to avoid amplitude clippings. Following the suggestion by Yiu et al. (submitted), the DI value was increased in 4% steps between 2% and 10% and in 6% steps between 10% and 28%, resulting in a total of six synthesized rough signals (Table 1). By including the prototype signal, there were a total of seven signals in the rough continuum.

Table 1. Values of synthesis parameters for the first dysphonic stimulus continua

Breathiness	Roughness [AV 80 AH 80]
Prototype	Prototype
AH 50	DI 2
AH 55	DI 6
AH 60	DI 10
AH 65	DI 16
AH 70	DI 22
AH 75	DI 28
AH 80	

** Prototype – (AV 60 AH 40 DI 0)*

Procedures

The stimuli were presented through a USB digital sound processor (Creative Extigy Signal Processing unit) and a pair of professional-quality headphones (Senheiser, HD 25) via a Microsoft Power-Point 2000 program (Pentium IV 2.26GHz computer) at a comfortable loudness level adjusted by the judges themselves in a quiet room. The stimuli were presented in six blocks, in terms of three languages and two voice quality sets. Synthesized stimuli were presented in pairs. Each pairs consisted of either two identical stimuli or a combination of two consecutive stimuli in the same breathy or rough continuum. As the prototype stimulus was added to the breathy and rough continua, a total of 15 breathy and 13 rough stimulus pairs for each language were formed. The judges were asked to decide whether each stimulus pair was the same or different in severity level. The presentation order of the stimulus pairs was randomized across the subjects.

Stimulus pairs that were perceived to be different by at least two out of the three judges were selected to be used in the Main Study. If stimulus pairs were not perceived to be different by at least two judges, meaning that the two consecutive stimuli in the continuum were not synthetically contrastive enough to be perceived as different, other synthesis parameters would be further manipulated and adjusted.

Results and Further Synthesis

Four breathy signals (AH 55 – AH 70) together with the prototype stimulus (AV 60 AH 40 DI 0) in all three languages were perceptually distinguishable by at least two of the three judges. They were therefore selected as the anchors for the Main Study. However, synthesized signals at AH 75 and AH 80 in all three languages were not perceived to be distinguishable by at least two judges. For the rough signals, only the first two rough signals in the continuum (DI 2 and DI 6) in all three languages were perceptually distinguishable by at least two judges.

Breathy signals were further modified by reducing the AV values in steps of 10 dB when the AH value increased beyond 70 dB. By reducing the amplitude of voicing (AV), the breathy quality would increase in proportion as the intensity of the signal is reduced. For the rough signals, the judges reported that they detected the presence of breathy quality in the rough signals, which had masked the rough quality perceptually. Thus, the two rough signals that were perceptually distinguishable to the judges (DI 2 and DI 6) were not used. The rough continuum was further designed with DI level increased in steps of 5% and the AH value of the rough signals was further modified. The AH value was increased in steps of 10 dB from AH 40 to AH 60 in the first three signals (DI 0 – DI 10) of the rough continuum, and then held at fixed value (AH 60) for the next four signals (DI 15 – DI 30). The synthesis values for these two further sets of dysphonic stimulus continua are given in Table 2. The synthesized stimuli were presented in pairs with either an identical stimulus or with the consecutive stimulus next in the same breathy or rough continuum. Five pairs of breathy stimuli and 13 pairs of rough stimuli were prepared and presented to the three judges again using the procedure described above.

Table 2. Values of synthesis parameters for the second dysphonic stimulus continua

Breathiness	Roughness
AV 60 AH 70	Prototype (AV 60 AH 40 DI 0)
AV 50 AH 70	AH 50 DI 5
AV 40 AH 70	AH 60 DI 10
	AH 60 DI 15
	AH 60 DI 20
	AH 60 DI 25
	AH 60 DI 30

For the breathy series, all stimulus pairs in all three languages were perceived to be different by at least two of the three judges. Therefore, together with the three stimuli determined earlier to be appropriate, a total of six breathy anchors were adopted for use in the Main Study. For the rough series, only the first four stimuli (from AH40 DI 0 to AH60 DI 20) in all three languages were perceived to be different from the next stimulus in the continuum.

Additional rough stimuli above DI 20 were therefore synthesized with the DI values increased in steps of 10% (Table 3a) and 15% (Table 3b). Stimuli pairs were prepared and presented as described earlier, each pairs consisted of either two identical stimuli or a combination of two consecutive stimuli in the rough continuum. This made up a total of 10 stimulus pairs.

Table 3. Values of synthesis parameters for the third dysphonic stimulus continua

(a) Roughness	(b) Roughness
AH 60 DI 20	AH 60 DI 20
AH 60 DI 30	AH 60 DI 35
AH 60 DI 40	AH 60 DI 40

Only the stimulus pairs with a difference of 15% in DI were judged to be perceptually different by at least two judges in all three languages.

Summary and Discussion for Pilot Study

The objective of the pilot study was to develop and select six sets of synthesized signals (three languages by two quality types) to be used as anchors in the Main Study. The anchors chosen had to be perceptually distinguishable in severity from the next stimulus in each of the breathiness and roughness continua by at least two experienced judges. Six synthesized stimuli for each voice quality (breathiness and roughness) and language (Cantonese, English and Putonghua) were selected together with the prototype stimulus (see Table 4).

Table 4. Synthesized anchors selected for the Main Study

Breathiness	Roughness
PROTYPE (AV 60 AH 40 DI 0)	PROTYPE (AV 60 AH 40 DI 0)
AV 60 AH 55	AH 50 DI 5
AV 60 AH 60	AH 60 DI 10
AV 60 AH 65	AH 60 DI 15
AV 60 AH 70	AH 60 DI 20
AV 50 AH 70	AH 60 DI 35
AV 40 AH 70	AH 60 DI 40

Acoustic Properties of the Selected Synthesized Signals

Kay’s Computerized Speech Lab 4400 and Multidimensional Voice Program (MDVP) were used to conduct acoustic analysis on the selected synthesized signals. MDVP was chosen as it has been shown to tolerate acoustic variations in connected speech (Yiu et al., 2002). Acoustic measurements including fundamental frequency (F_0), relative average perturbation (RAP), pitch perturbation quotient (PPQ), shimmer percent (Shim %), amplitude perturbation quotient (APQ) and noise to harmonic ratio (NHR) were carried out. These measures were used as they are commonly reported in the literature and were believed to be correlated with breathy and rough qualities (Eskenazi, Childers, & Hicks, 1990).

For the breathy signals in Cantonese, English and Putonghua, the fundamental frequency ranged from 203- 249Hz. An increasing trend was noted in all five acoustic measurements for all the breathy signals (Appendix A). The most severe English breathy signals AV50 AH70 and AV40 AH70, as well as the Putonghua breathy signal AV40 AH70 were not analyzable acoustically as the program reported insufficient voice data to conduct accurate perturbation analysis.

For the rough signals, the fundament frequency across Cantonese, English and Putonghua signals ranged from 102 – 153 Hz. The fundamental frequency was halved as synthesis of roughness was based on manipulation of the DI parameter, which uses two glottal pulses in slightly different phases (Klatt & Klatt, 1990). When the pulse of one glottal pair was attenuated to zero, the fundamental frequency would be halved (Yiu et al., 2002). The roughness scale showed a mild increasing trend in all five acoustic measurements (Appendix A). The acoustic data in general also supported the increasing severity of the synthesized stimuli.

Method

Preparation of Natural Testing Stimuli

Three sets of natural female voice samples, in Cantonese, English and Putonghua, were selected from the database of recorded voice samples collected at the Voice Research Laboratory, the University of Hong Kong. Only female voice stimuli were used, as they represent the major clinical population (Yiu & Ho, 1991) and the task would have been too long for the subjects if stimuli from both genders were used. All testing stimuli were recorded in a sound-treated room using Kay's Computerized Speech Lab 4400 at 50 kHz. They were downsampled to 44.1 kHz and equalized in amplitude using Cool Edit (2000; Syntrillium) for compatibility with the sound processing system of the computer. The voice samples were recorded from 30 female speakers (20 dysphonic, five with normal voice, and five with normal voice simulating dysphonic voices). The mean age of speakers was 29.83 years (SD = 9.60, range = 20-34).

Each language set covered two types of quality (breathiness and roughness) at three severity levels (mild, moderate and severe), resulting in six categories in each set. There were two voice samples from different speakers in each category. Together with the two normal voice samples, there were 14 voice samples in each language set. Each voice sample was duplicated, resulting in 28 testing stimuli in each set. The dysphonic quality and severity level were judged to represent appropriately by two experienced judges, each with over four years of experience in perceptual voice evaluation.

Subjects

Ten native Cantonese speakers, 10 native English speakers and 10 native Putonghua speakers with a mean age of 26.30 years (*SD* = 5.29, range = 19 – 38) participated as judges in this study. All of them were reported to have normal voice and health conditions and had not received any training in voice disorders or perceptual voice evaluation. All of the subjects received a hearing screening before the session in a sound-treated room, included test of thresholds at 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz and 8000 Hz at 25dB. All of them had normal hearing.

Procedures

The rating test was presented using a specifically designed computer program through a Pentium II 533 Hz computer in a sound-treated room. The stimuli were presented through an external sound card (Creative Extigy Signal Processing unit) and a pair of professional-quality headphones (Senheiser, HD 25) at a comfortable listening level. Printed definitions of breathiness and roughness (see Table 5) was shown and explained to the participants at the beginning of the test. The rating test was divided into three blocks - Cantonese, English and Putonghua stimulus block. The participants were required to rate the breathiness and roughness of the testing stimuli in all three languages. A computerized program with a knob control with a seven-point scale from 0 (representing normal voice quality) to 6 (most severe dysphonic quality) was used and each scale point was represented with a synthesized voice anchor (Appendix B). The participants were asked to choose the breathy and rough anchor that best-matched with each of the natural testing stimuli. They were allowed to listen to both the synthesized anchors and the natural stimuli as many times as they would like to. Also, they were asked to rate their confidence in rating each testing stimulus using a seven-point equal-appearing interval scale from 1 (representing wild guess) to 7 (absolute confidence). The order of presentation of language blocks and test stimuli within each block were randomized across the listeners. The average duration of the test taken by the participants was 45 minutes.

Table 5. Definition of breathiness and roughness (Yiu, 2001, p. 16)

Voice quality	Perceptual correlates	Physiological correlates
Breathiness	1. Audible sound of expiration 2. Audible sound escape 3. Audible friction noise	Incomplete closure of vocal folds during phonation
Roughness	1. Irregular quality 2. Random fluctuations of glottal pulse 3. Lack of clarity	(Believed to be) due to irregular vibration of the vocal folds

Data Analysis

The perceptual ratings of breathiness and roughness were analyzed to determine the agreement and reliability of each listener group across each stimulus set. Only ratings of relevant voice quality were analyzed for the designated voice sample, i.e. ratings of breathiness were analyzed for the designated breathy stimuli and ratings of roughness were analyzed for the designated rough stimuli.

The mean ratings of breathiness and roughness for each listener group across each stimulus set were calculated. Intra-rater agreement of both breathiness and roughness ratings and inter-rater reliability of both voice qualities were calculated separately and analyzed using repeated ANOVA to compare data between listener groups (between-subject factor) and stimulus sets (within-subject factor). Pillai’s Trace ANOVA was chosen as it was considered to be more robust for violation in assumptions of ANOVA calculations (Tabachnick & Fidell, 2001).

Results

The mean ratings of breathiness and roughness were calculated for each voice quality, listener group and language set (Appendix C).

Main (Listener, Stimulus) and Interaction (Listener x Stimulus) Effects

Intra-rater Agreement. Intra-rater agreement was calculated for each judge as each testing stimulus was rated twice. The mean percentage of exact agreement and mean percentage agreement within one scale point were calculated. Agreement ratings at each severity level were not calculated as there were only two voice samples available per level. The chance probability is 14.3% (1 out 7 possible ratings) when listeners responded to any rating and 42.9% when judges responded within one scale point, based on the formula $[n+2(n-1)]/n^2$ where n equals to the number of points in the scale (Kreiman et al., 1993).

The overall mean exact agreement was above chance level, varying between 28.75% and 62.75% (Table 8), and the values for mean agreement within one scale point varied between 78.75% and 96.25%, which was also well above the chance level. Pillai’s Trace ANOVA was used to determine if there were any significant differences between the listener groups, stimulus sets and language sets for ratings on agreement within one scale point. ANOVA results (Table 9) showed significant main listener and main stimulus effects for the roughness ratings. Post-hoc comparison (using Tukey’s HSD test) and planned contrast was calculated with adjusted p-level (0.0167) as the test was repeated three times each (0.05/3). Results showed that significant higher agreement in Cantonese listeners than in Putonghua listeners and also significant higher agreement in English stimulus than in Cantonese stimulus (Table 10).

Table 8. Mean percentage of intra-rater agreement

	Mean exact agreement		**Mean agreement ± one scale point**	
Listeners	**Breathy (SD)**	**Rough (SD)**	**Breathy (SD)**	**Rough (SD)**
Cantonese stimulus				
Cantonese	41.25 (10.29)	55.00 (32.29)	87.50 (10.21)	90.00 (14.19)
English	46.25 (27.04)	41.25 (17.73)	81.25 (17.48)	81.25 (12.15)
Putonghua	60.00 (19.36)	28.75 (10.29)	85.00 (19.76)	78.75 (10.29)
English Stimulus				
Cantonese	62.50 (27.00)	62.75 (12.43)	91.25 (15.65)	96.25 (6.04)
English	43.75 (23.75)	55.00 (14.67)	78.75 (24.33)	92.50 (8.74)
Putonghua	57.50 (19.72)	46.25 (15.65)	88.75 (16.08)	82.50 (12.08)
Putonghua Stimulus				
Cantonese	51.25 (25.31)	55.00 (25.82)	91.25 (10.29)	96.25 (8.44)
English	37.50 (25.00)	52.50 (20.24)	86.25 (14.97)	90.00 (12.91)
Putonghua	46.25 (17.73)	41.25 (16.72)	90.00 (12.91)	82.50 (12.08)

Table 9. ANOVA results of intra-rater agreement

		Agreement ± 1 scale point	
Effects	**Quality**	***F (2, 26)***	***p***
Listener	Breathy	1.17	0.33
	Rough	*7.23**	*0.003*
Stimulus	Breathy	26.00	0.40
	Rough	*26.00**	*0.02*
Listener x	Breathy	54.00	0.91
Stimulus	Rough	54.00	0.77

** Significant level p < 0.05*

Table 10. Post-Hoc analysis and planned contrast on the intra-rater agreement on rough signals

Mean Intra-rating agreement rankings	**Agreement ± 1 scale point**	
	F (2, 26)	***p***
Listener effects - post – hoc comparison		
Cantonese > Putonghua	*12.92**	*0.002*
Cantonese > English	6.25	0.18
English > Putonghua	6.67	0.14
Stimulus effects - planned contrast		
Putonghua > Cantonese	5.23	0.03
English > Cantonese	*9.46**	*0.005*
English > Putonghua	0.11	0.75

** Significant level p < 0.016*

Inter-rater Reliability. For inter-rater reliability, both intra-class correlation coefficient (ICC) (Shrout & Fleiss, 1979) and a variability score were reported. ICC was used to measure the variability of perceptual ratings of breathiness and roughness in each listener group across each stimuli set. A variability score (see Chan & Yiu, 2002) was also reported as ICC only reveals an overall pattern of the entire group of listener and may not reflect the pattern for each stimulus (Kreiman & Geratt, 1998). The variability score was calculated by averaging the sum of the squared differences between each listener's rating and the mean rating of all the listeners for that stimulus. Perceptual ratings that were close to the mean would result in a low variability score, while ratings differed largely from the mean would result in a high score (Chan & Yiu, 2002).

Intraclass Correlation Coefficient. ICC was calculated using a two-way random effects modal (ICC[2, 10]) ((Shrout & Fleiss, 1979). The model was chosen as the listeners were selected randomly from a larger population of listeners and they rated all the testing stimuli randomly selected from a database of voice stimuli. Violation of homogeneity of variance due to small sample size did not permit calculation of ICC for each severity level. The ICC values for breathiness and roughness at each level varied between 0.84 and 0.98 (Table 11). All the values were within the 95% confidence interval of each other and no significant difference (p> 0.05) was found.

Table 11. Average measure overall intra-class correlation coefficient (ICC)

	Cantonese listeners		**English listeners**		**Putonghua listeners**	
Stimulus	**ICC**	**95% Confidence**	**ICC**	**95% Confidence**	**ICC**	**95% Confidence**
Breathy						
Cantonese	0.92	$0.84 < p < 0.97$	0.93	$0.86 < p < 0.97$	0.84	$0.69 < p < 0.94$
English	0.97	$0.93 < p < 0.99$	0.97	$0.95 < p < 0.99$	0.95	$0.91 < p < 0.98$
Putonghua	0.98	$0.96 < p < 0.99$	0.98	$0.97 < p < 0.99$	0.95	$0.91 < p < 0.98$
Rough						
Cantonese	0.95	$0.91 < p < 0.98$	0.96	$0.92 < p < 0.98$	0.88	$0.77 < p < 0.95$
English	0.96	$0.93 < p < 0.99$	0.96	$0.92 < p < 0.98$	0.91	$0.82 < p < 0.96$
Putonghua	0.95	$0.91 < p < 0.98$	0.96	$0.92 < p < 0.98$	0.93	$0.87 < p < 0.97$

Inter-rater Variability Score. The variability score obtained varied between 0.21 and 2.49 across each dysphonic level (Table 12). Pillai’s Trace ANOVAs were carried out (Table 13) with adjusted p-level (0.0125) as the test was repeated four times (0.05/4) for different severity levels. Significant stimulus effects were noted for moderately and severely breathy signals. Tukey’s HSD post-hoc comparisons were used with adjusted p-level (0.0167) as the test was repeated three times (0.05/3) for comparison within the stimulus sets. Results showed that for moderately breathy signals, significant lower variability score was found for Putonghua stimulus when compared to Cantonese stimulus, and for the severely breathy signals, significant lower variability scores was found for Putonghua stimulus when compared to English stimulus.

Table 12. Mean and standard deviation (SD) of inter-rater variability scores

	Cantonese stimulus			**English stimulus**			**Putonghua stimulus**		
	mean (SD)			**mean (SD)**			**mean (SD)**		
Listeners	**Cant**	**Eng**	**PTH**	**Cant**	**Eng**	**PTH**	**Cant**	**Eng**	**PTH**
Breathy									
Normal	0.58	0.59	2.28	0.53	0.81	0.99	0.73	0.85	0.92
	(0.32)	(0.74)	(2.07)	(0.59)	(0.40)	(0.70)	(0.52)	(0.59)	(0.40)
Mild	0.53	1.45	2.49	0.68	2.08	1.66	0.92	1.33	1.61
	(0.25)	(0.56)	(0.90)	(0.14)	(0.60)	(0.50)	(0.38)	(0.45)	(1.11)
Moderate	0.92	2.01	2.12	0.85	2.28	1.96	0.87	0.96	1.41
	(0.29)	(0.58)	(0.41)	(0.21)	(0.47)	(0.25)	(0.74)	(0.25)	(0.51)
Severe	1.04	1.47	1.64	1.30	0.77	1.75	0.37	0.45	1.48
	(0.41)	(0.34)	(0.33)	(0.98)	(0.46)	(0.78)	(0.19)	(0.01)	(0.79)
Rough									
Normal	0.21	0.52	1.43	0.52	1.85	0.99	0.49	1.26	1.21
	(0.16)	(0.31)	(0.51)	(0.37)	(0.51)	(0.46)	(0.33)	(0.64)	(0.60)
Mild	0.42	1.01	1.94	0.42	1.05	1.16	1.09	0.33	1.49
	(0.20)	(0.30)	(0.42)	(0.02)	(0.94)	(0.44)	(0.58)	(0.19)	(0.73)
Moderate	1.15	1.30	1.64	0.87	1.84	1.75	1.23	0.82	1.38
	(0.33)	(0.88)	(0.52)	(0.10)	(0.31)	(0.37)	(0.88)	(0.82)	(0.74)
Severe	1.39	1.16	1.66	1.06	1.22	1.71	1.21	0.71	0.92
	(0.27)	(0.27)	(0.66)	(0.63)	(0.51)	(0.48)	(0.70)	(0.48)	(0.21)

Cant = Cantonese Eng = English PTH = Putonghua

Table 13. ANOVA results for inter-rater variability scores

	Breathy		**Rough**	
Effects	***F (2, 26)***	***p***	***F (2, 26)***	***p***
Normal				
Listener	4.63	0.02	2.61	0.09
Stimulus	0.86	0.43	1.39	0.27
Stimulus x Listener	2.01	0.11	0.85	0.13
Mild				
Listener	3.40	0.50	2.91	0.07
Stimulus	0.48	0.63	0.90	0.42
Stimulus x Listener	1.40	0.25	2.91	0.07
Moderate				
Listener	3.33	0.05	1.84	0.18
Stimulus	*6.69**	*0.01*	0.21	0.81
Stimulus x Listener	1.85	0.13	0.97	0.43
Severe				
Listener	3.53	0.04	1.18	0.32
Stimulus	*5.87**	*0.01*	3.84	0.03
Stimulus x Listener	1.18	0.33	1.28	0.29

** Significant level p < 0.0125*

Table 14. Post-Hoc analysis on stimulus effect of inter-rater variability scores

		F (2, 26)	***P***
Moderate breathy	Cantonese > Putonghua	*6.61**	*0.01*
	English > Cantonese	0.00	0.97
	English > Putonghua	4.27	0.05
Severe breathy	Cantonese > Putonghua	5.20	0.03
	Cantonese > English	0.14	0.71
	English > Putonghua	*8.99**	*0.01*

** Significant level p < 0.0167*

Confidence in Rating Breathiness and Roughness

Mean confidence rating varied between 4.76 and 5.91 (Table 15). Pillai’s Trace ANOVAs were calculated and no significant difference was found (p > 0.1) (Table 16).

Table 15. Mean confidence ratings

	Cantonese stimulus		**English stimulus**		**Putonghua stimulus**	
Listeners	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Breathy						
Cantonese	5.91	0.86	5.83	0.90	5.83	0.79
English	5.63	1.02	5.63	1.02	5.73	0.91
Putonghua	4.76	1.55	5.25	1.16	5.30	1.09
Rough						
Cantonese	5.80	0.89	5.78	0.85	5.83	0.79
English	5.72	1.01	5.73	0.87	5.72	1.02
Putonghua	5.23	1.10	5.19	1.17	5.22	1.30

Table 16. ANOVA results for confidence ratings

	Breathy		**Rough**	
Effects	***F (2, 26)***	***p***	***F (2, 26)***	***p***
Listener	3.70	0.38	2.24	0.13
Stimulus	0.73	0.49	0.37	0.96
Listener x Stimulus	0.70	0.60	0.41	1.00

** Significant level p < 0.05*

Discussion

The first objective of this study was to develop appropriate sets of female synthesized signals that represent perceptually a range of mildly to severely breathy and rough signals in Cantonese, English, and Putonghua. The breathy signals were synthesized by manipulating the parameters amplitude of aspiration (AH) and amplitude of voicing (AV), while the rough signals were synthesized by manipulating AH and diplophonia (DI). The pilot study showed that the expert listeners perceived a 5 dB difference in the synthesized amplitude of aspiration for the breathy signals until it reaches its maximum value of 70 dB, then a minimum of 10 dB difference was required in detecting differences in the synthesized amplitude of voicing. For the rough signals, the listeners were able to perceive a 5% and 15% difference in the synthesized diplophonia at DI 0 to DI 20 and DI 20 to DI 50 respectively. For the acoustic measures, the jitter, shimmer and noise measures for the breathy and rough signals in Cantonese, English and Putonghua were gradually increasing as the severity of breathiness and roughness increased. The perceptual and acoustic measurements both showed that the synthesized signals represented a range of breathiness and roughness.

The second objective was to determine if there were any differences in the agreement and reliability of perceptual voice evaluation across native speakers of three different languages. The present study hypothesized that similar and high intra- and inter-rater reliability would be shown across listeners and stimuli using an anchor matching method. The findings supported this hypothesis. The average overall mean agreements within one scale point for breathy and rough ratings were 86.67% and 87.78% respectively (Table 8). These values were comparable to that obtained in the study by Yiu et al. (submitted) (92.5%) and slightly lower than that in the study by Gerratt et al. (1993) (99.6%). Such difference might have been attributed to the variation in the listeners' experiences with pathological voices, which is often a major source of variability in perceptual voice evaluation (Kreiman et al., 1993). Naïve listeners were recruited in the current study, while listeners in the study by Yiu et al. (submitted) were given training in perceptual voice rating and experienced listeners were used in the study by Gerratt et al. (1993). Nonetheless, the agreement measurements demonstrated by the naïve listeners in the present study were well above the 80% level, indicating that the use of the anchor-matching method is a reliable alternative means of perceptual judgment for use in Cantonese, English and Putonghua.

It was also hypothesized that there would be minimal effects of cultural and linguistic backgrounds on perceptual voice evaluation when an anchor matching paradigm was used. The findings only partially supported this hypothesis.

Listener Effects

The Cantonese listeners showed significantly higher mean intra-rater agreement within one scale point than Putonghua listeners when rating rough signals. One possible reason was that the Putonghua judges might not have grasped the concept of roughness accurately. The concepts of breathiness and roughness were introduced to the judges using definitions in English since there was no direct translation of the definition of roughness in Chinese (Cantonese and Putonghua). On the other hand, the Cantonese listeners, who were all knowledgeable in all three languages, might have a better grasp of this concept being described in English. Kreiman and Gerratt (1994) found that listeners differed considerably when judging vocal roughness, as they varied in their attention to different dimension of roughness, compared to rating breathiness, leading to disagreement and variability in their perceptual ratings. It was hypothesized that the Putonghua listeners in this study were more vulnerable to such differential attention among and within themselves when rating vocal roughness due to their poor conceptualization of this voice quality, resulting in the significantly poorer intra-rater agreement obtained when compared to the Cantonese listeners. Indeed, the Putonghua listeners showed slightly lower mean intra-rater agreement than the Cantonese and English listeners in all the ratings of the rough stimuli and also showed high mean inter-rater variability scores except when rating severely rough Putonghua signals, although no significant differences were found. Future studies with more Putonghua subjects will be needed to confirm this.

However, comparable findings were not found in the inter-rater reliability measurements. One possible reason was that the two inter-rater reliability measurements were average measures. Intra-class correlation coefficient (ICC) is calculated by summing all the ratings to give a single measure and reveals only the overall agreement pattern of the entire group of listeners. Kreiman and her colleagues (1993) criticized it to be too coarse for measuring reliability. The variability score was also calculated by averaging the sum of the squared differences between each listener's rating and the mean rating of all the listeners for that stimulus. Both measurements were calculated by averaging across all the ratings, where variability amongst listeners might have cancelled each other out. Nonetheless, relatively higher mean inter-rater variability scores were found in the Putonghua listeners than in the Cantonese listeners in all the ratings of rough stimuli except for the severely rough continuum, indicating lower inter-rater reliability within this listener group.

Additionally, the familiarity of listeners with the three languages of the stimuli tested might have affected the results. As indicated above, no significant difference was found between the English and Cantonese and between the English and Putonghua listeners. One possible reason was that half of the English listeners recruited in this study have been living in Hong Kong for more than five years. They might have some exposure to all three languages of the stimuli. Furthermore, the Cantonese listeners were all knowledgeable of the three languages tested. However, majority (80%) of the Putonghua listeners have only lived in Hong Kong for less than a year, presumably with limited exposure to Cantonese and English. Differences in their familiarity with the three languages tested might have been a confounding variable.

Stimulus Effects

Significant stimulus effects were also found in this study. First, results showed listeners rated the English rough stimuli with significantly higher agreement when compared to the Cantonese series. This might be due to the reduced naturalness in the Cantonese synthesized anchors when compared to the English series. Cantonese is a tonal language with variations in nine tones. It was possible that the manipulation of the Klatt parameters in creating natural-sounding synthesized signals was more difficult for signals in tonal languages. Thus the significant difference might have been caused by the difference in the naturalness of the synthesized anchors. Future research may be carried out on the manipulation of Klatt parameters in creating natural-sounding signals.

Results also showed significant difference between the Putonghua and Cantonese series when listeners rated the moderately breathy continuum and between Putonghua and English series in the severely breathy continuum. This might have been caused by the disproportional ratio in the use of simulated and non-simulated natural female voice testing stimuli across the Cantonese, English and Putonghua series. Simulated natural dysphonic stimuli were used as there were not enough dysphonic voice samples from the database to cover the full range of severity level in both the breathy and rough continua. Both the natural voice stimuli used for moderately and severely breathy series in the Putonghua continuum were simulated, while few simulated stimuli were used in the Cantonese and English breathy continua. Simulated natural dysphonic stimuli might have been easier for the raters as only one dysphonic quality was simulated, whereas the real dysphonic voice signals might have multidimensional qualities, resulting in the significantly lower inter-rater variability score with the moderately and severely breathy Putonghua signals when compared to the Cantonese and English ones respectively. Hence, future studies should avoid the mixed used of simulated and non-simulated natural dysphonic voice samples.

Limitations of the Present Study and Future Studies

There were a number of limitations that warrant further studies. First, the listeners recruited in this study were different in their familiarity with the three languages being tested. Further study recruiting only monolingual listeners would allow further examination of effects of specific culture on perceptual voice evaluation. Second, judges of both genders were mixed in the listener groups and only female voice samples were employed. Future studies with the use of gender-balanced listener groups and voice samples of both genders should be carried out to investigate on possible gender effects in perceptual voice evaluation in listeners with different native languages. Last, only two natural testing stimuli were included in each severity level in both voice qualities in each language set. Calculations of intra-rater agreement and ICC at each dysphonic level were not permitted due to the small sample size of stimuli. Future studies with the use of more natural testing stimuli would allow ensure reliability at each severity level.

Conclusion

Results from this study supported that the anchor-matching paradigm could be an alternative perceptual voice evaluation method as demonstrated by its acceptable intra-rater agreement and inter-rater reliability in all three listener groups across the different stimuli sets. The findings also suggested that there are possible cultural and linguistic differences in perceptual voice evaluation between Putonghua and Cantonese listeners. Further studies with the recruitment of monolingual listeners in gender-balanced groups, use of real natural stimuli in both genders, improvement in the naturalness of the synthesized anchors, and inclusion of more natural testing stimuli at each severity level are recommended for future research.

Acknowledgment

I would like to express my sincere to Dr. Edwin Yiu for his guidance and support throughout this study. I would like to thank all the volunteers for their participation. I am also grateful for Dr. Karen Chan, Mr. Donald Chan, Ms. Elaine Kwong and Mr. Raymond Wu and for their advice and help in this study.

References

Anders, L. Ch., Hollien, H., Hurme, P., Sonninen, A., & Wendler, J. (1988). Perception of hoarseness by several classes of listeners. *Folia Phoniatrica, 40*, 91-100.

Bangayan, P., Long, C., Alwan, A. A., Kreiman, J., & Gerratt, B. R. (1997). Analysis by synthesis of pathological voices using the Klatt synthesizer. *Speech Communication, 22,* 343-368.

Bruyninckx, M., Harmegnies, B., Llisterri, J., & Poch-Olive, D. (1994). Language-induced voice quality variability in bilinguals. *Journal of Phonetics, 22*, 19-31.

Chan, K. M. K., & Yiu, E. M.-L. (2002). The effects of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language and Hearing Research, 45*(1), 111-126.

De Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research, 38*, 794-811.

Eskenazi, L., Childers, D. G., & Hicks, D. M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33,* 298-306.

Esling, J. H. (2000). Crosslinguistic aspects of voice quality. In Kent, R. D., & Ball, M. J. (Eds.). *Voice Quality Measurement.* San Diego: Singular Publishing Group.

Gerratt, B. R., & Kreiman, J. (2001). Measuring vocal quality with speech synthesis. *Journal of Acoustical Society of America, 110*(5), 2560-2566.

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

Hammerberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngology Supplement (Stockholm), 90*, 441-451.

Hartelius, L., Theodoros, D., Cahill, L., & Lillvik, M. (2003). Comparability of perceptual analysis of speech characteristics in Australian and Swedish Speakers with multiple sclerosis. *Folia Phoniatrica et Logopaedica, 55*(4), 177-188.

Hirano, M. (1981). *Clinical Examination of Voice.* Vienna: Springer Verlag.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87*(2), 820-857.

Kreiman, J., & Gerratt, B. R. (2000). Sources of listener disagreement in voice quality assessment. *Journal of the Acoustical Society of America, 108*(4), 10867-1876.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *Journal of Acoustical Society of America, 104*(3), 1598-1608.

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.

Martin, D. P., & Wolfe, V. I. (1996). Effects of perceptual training based upon synthesized voice signals. *Perceptual and Motor Skills, 83*(3 (part2)), 1291-1298.

Majewski, W., Hollien, H., & Zalewski, J. (1972). Speaking fundamental frequency of Polish adult males. *Phonetica, 25*, 119-125.

Revis, J., Giovanni, A., Wuyts, F., & Triglia, J.-M. (1999). Comparison of different voice samples for perceptual analysis. *Folia Phoniatrica et Logopaedica, 51*, 108-116.

Wolfe, V., Cornell, R., Fitch, J. (1995). Sentence/vowel correlation in the evaluation of dysphonia. *Journal of Voice, 9*(3), 297-303.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Tabachnick, B. G.., & Fidell, L. S. (2001). *Using multivariate statistics* (4th Ed.). Boston: Allyn and Bacon

Wolfe, V., Fitch, J., & Martin, D. (1997). Acoustic measures of dysphonic severity across and within voice types. *Folia Phoniatrica et Logopaedica, 49*, 292-299.

Wuyts, F. L., de Bodt, M. S., & Van de Heyning, P. H. (1999). Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *Journal of Voice, 13*(4), 508-517.

Yamaguchi, H., Shrivastav, R., Andrews, M. L., & Niimi, S. (2003). A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. *Folia Phoniatrica et Logopaedica, 55*(3), 147-157.

Yiu, E. M.-L., Chan, K. M. L., & Mok, S. M. R. (submitted). Reliability and confidence in using a pair comparison paradigm in perceptual voice quality evaluation. *Submitted to Journal of Speech and Hearing Research.*

Yiu, E. M. L., & Ho, P. S. P. (1991). Voice problems in Hong Kong: A preliminary report. *Australia Journal of Human Communication Disorders*, 19, 45-58.

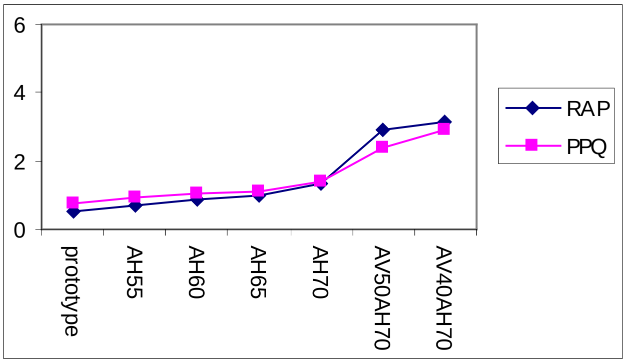
Yiu, E. M.-L., Murdoch, B., Hird, L., & Lau, P. (2001). Perception of synthesized voice quality in connected speech by Cantonese Speakers. *Journal of the Acoustical Society of America, 112*(3), 1091-1101.

Yiu, E., Worrall, L. E., Longland, J., & Mitchell, C. (2000). Analyzing vocal quality of connected speech and using Kay's Computerized Speech Lab: A preliminary finding. *Clinical Linguistics and Phonetics, 14*(4), 295-305.

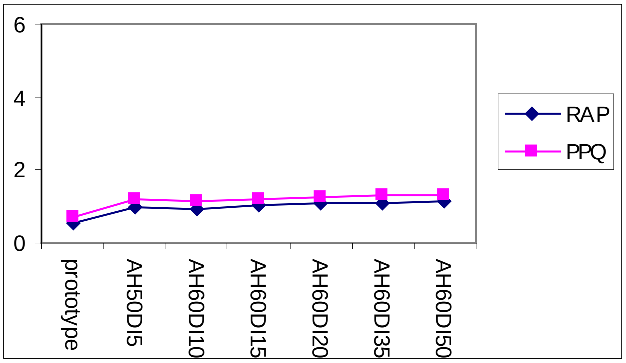
Yumoto, E., Sasaki, Y., & Okamura, H. (1984). Harmonics-to-noise ratio and psychophysical measurement of degree of hoarseness. *Journal of Speech and Hearing Research, 27*, 2-6.

APPENDIX A

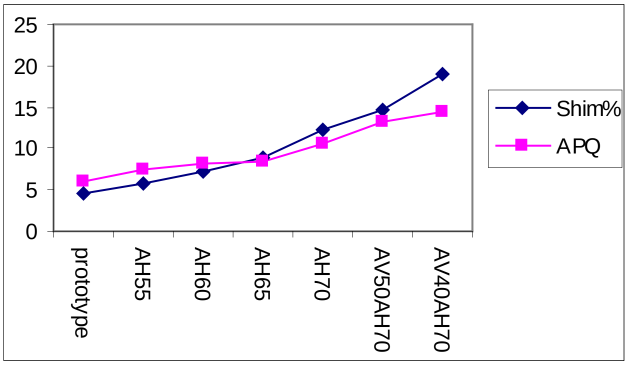
Acoustic properties of the synthesized anchors.



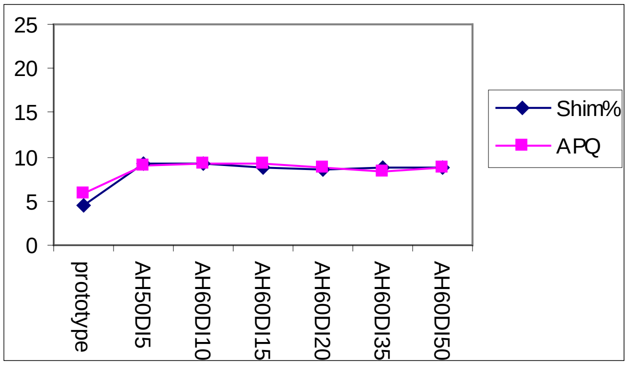
(a) Jitter measures of Cantonese breathy signals



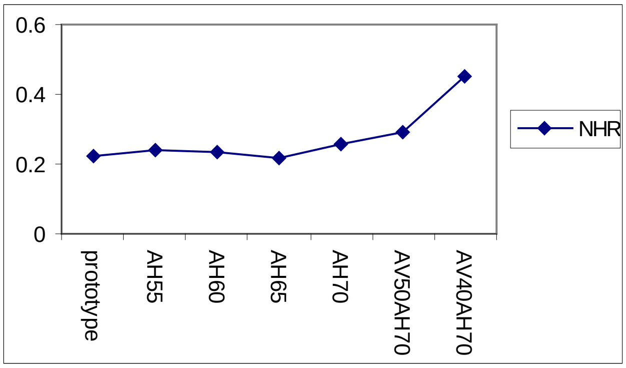
(j) Jitter measures of Cantonese rough signals



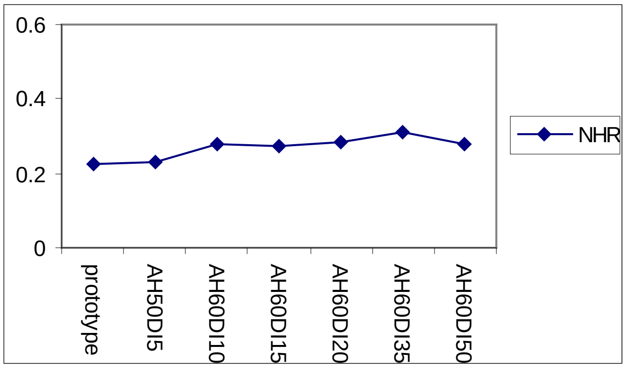
(b) Shimmer measures of Cantonese breathy signals



(k) Shimmer measures of Cantonese rough signals

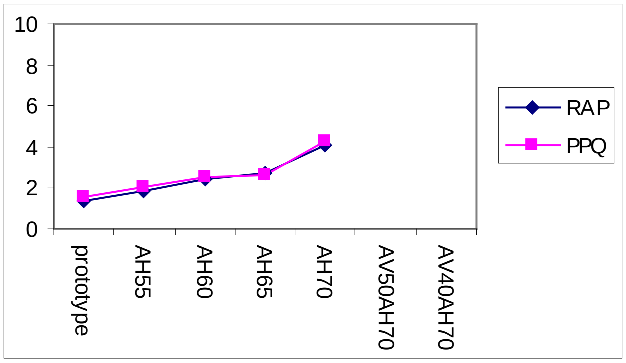


(c) Noise measures of Cantonese breathy signals

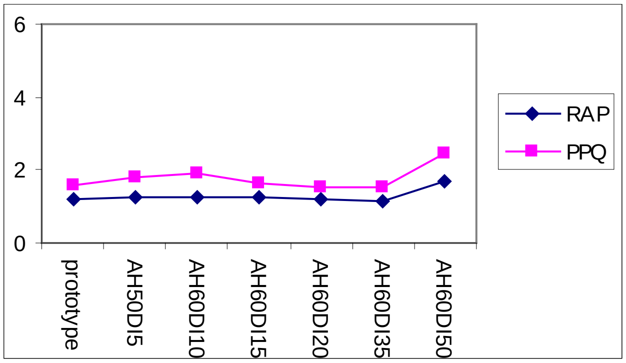


(l) Noise measures of Cantonese rough signals

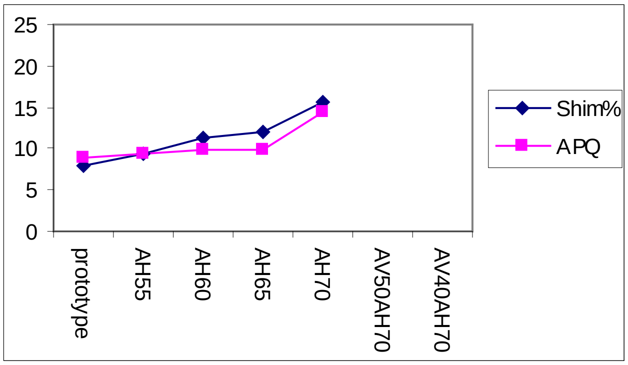
**Note:* F_0 - fundamental frequency; RAP - relative average perturbation; PPQ - pitch perturbation quotient, Shim% - shimmer percent, APQ - amplitude perturbation quotient; NHR - noise to harmonic ratio (NHR)



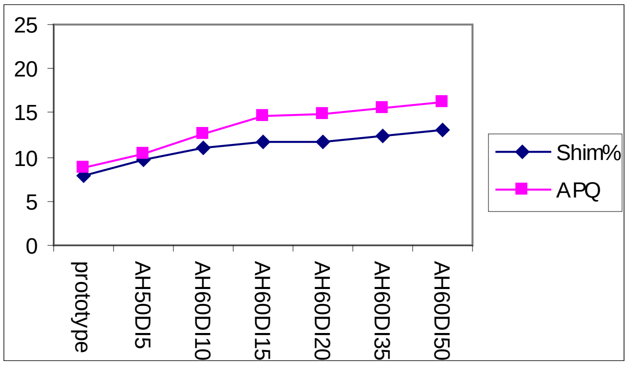
(d) Jitter measures of English breathy signals



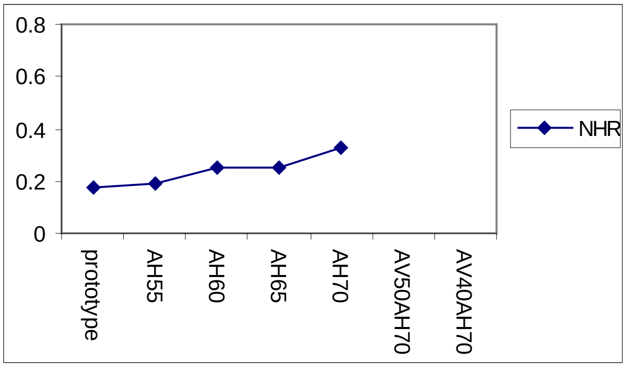
(m) Jitter measures of English rough signals



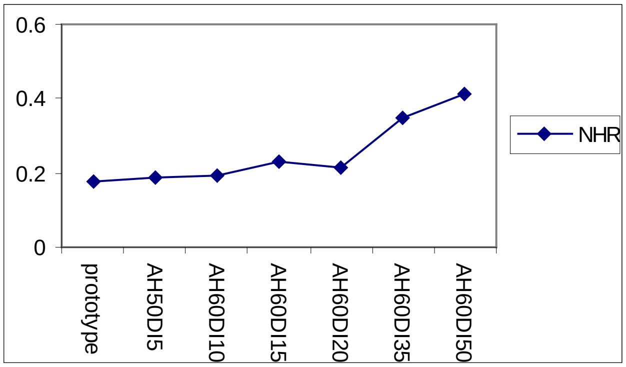
(e) Shimmer measures of English breathy signals



(n) Shimmer measures of English rough signals

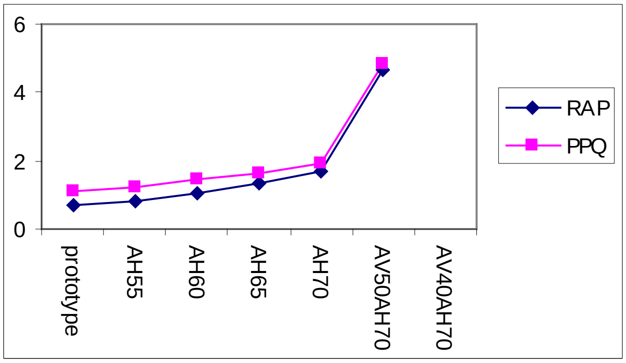


(f) Noise measures of English breathy signals

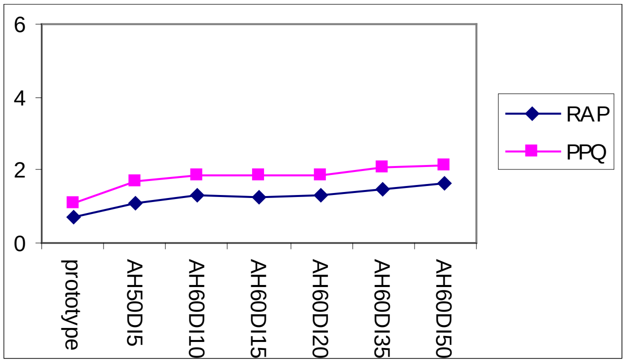


(o) Noise measures of English rough signals

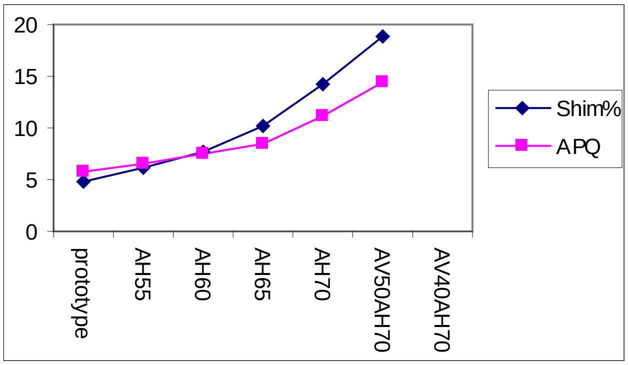
**Note:* F_0 - fundamental frequency; RAP - relative average perturbation; PPQ - pitch perturbation quotient, Shim% - shimmer percent, APQ - amplitude perturbation quotient; NHR - noise to harmonic ratio (NHR)



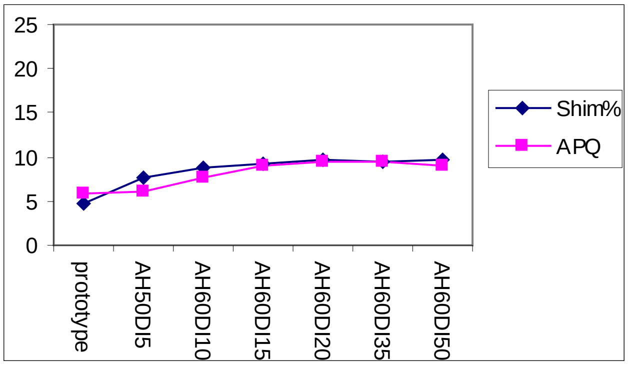
(g) Jitter measures of Putonghua breathy signals



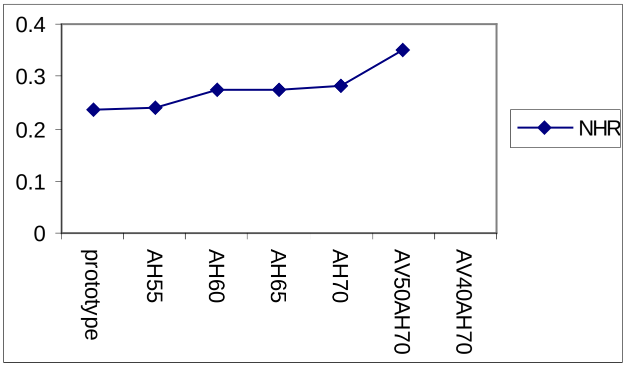
(p) Jitter measures of Putonghua rough signals



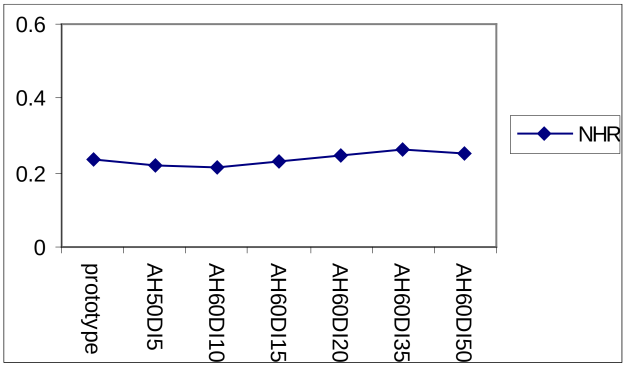
(h) Shimmer measures of Putonghua breathy signals



(q) Shimmer measures of Putonghua rough signals



(i) Noise measures of Putonghua breathy signals

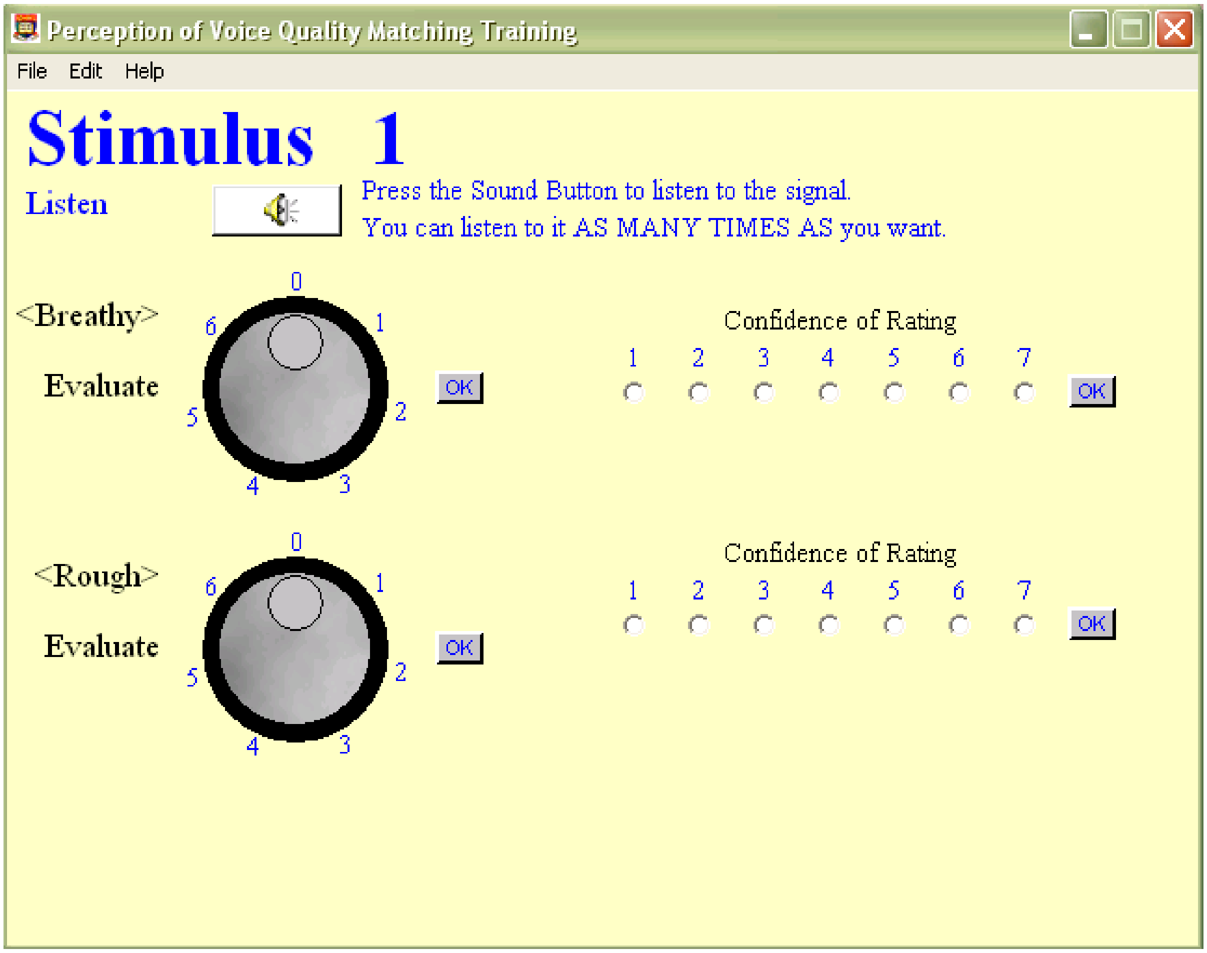


(r) Noise measures of Putonghua rough signals

*Note: F$_0$ - fundamental frequency; RAP - relative average perturbation; PPQ - pitch perturbation quotient, Shim% - shimmer percent, APQ - amplitude perturbation quotient; NHR - noise to harmonic ratio (NHR)

APPENDIX B

Sample page of the rating program



Stimulus 1

Press the Sound Button to listen to the signal.
You can listen to it AS MANY TIMES AS you want.

APPENDIX C

Mean breathiness and roughness ratings

	Cantonese listeners			English listeners			Putonghua listeners		
Stimulus	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**	**Mean**	**SD**	**Range**
Cantonese Breathy									
Normal	0.90	0.81	0 – 3	0.55	0.85	0 – 4	1.15	1.37	0 – 5
Mild	1.10	0.78	0 – 2	1.15	1.23	0 – 5	2.03	1.39	0 – 5
Moderate	2.05	1.08	0 – 4	1.95	1.45	0 – 5	3.03	1.46	0 – 5
Severe	2.93	1.07	1 - 5	3.33	1.37	0 – 6	3.58	1.34	1 – 6
Cantonese Rough									
Normal	0.23	0.48	0 – 2	0.40	0.67	0 – 3	1.18	1.26	0 – 4
Mild	0.80	0.69	0 – 3	1.58	1.03	0 – 4	1.98	1.48	0 – 5
Moderate	1.93	1.10	0 – 5	2.70	1.36	1 – 6	2.63	1.44	0 – 5
Severe	3.00	1.22	1 – 6	4.00	1.24	2 – 6	3.7	1.30	1 – 6
English Breathy									
Normal	0.48	0.75	0 – 4	0.58	0.93	0 – 3	0.58	1.03	0 – 4
Mild	1.33	0.92	0 – 3	1.78	1.53	0 – 5	1.98	1.49	0 – 5
Moderate	1.65	0.98	0 – 4	2.43	1.58	0 – 5	2.30	1.42	0 – 5
Severe	4.38	1.29	0 – 6	5.03	1.00	2 – 6	4.13	1.47	1 – 6
English Rough									
Normal	0.45	0.75	0 – 3	1.40	0.39	0 – 4	0.70	1.02	0 – 3
Mild	0.53	0.68	0 – 2	1.18	1.11	0 – 5	1.68	1.10	0 – 4
Moderate	1.20	0.97	0 – 4	1.95	1.57	0 – 5	2.13	1.38	0 – 5
Severe	3.43	1.34	2 – 6	4.70	1.16	2 – 6	3.45	1.43	0 – 6
Putonghua Breathy									
Normal	0.85	0.92	0 – 4	0.65	0.95	0 – 4	1.08	1.00	0 – 4
Mild	1.93	1.27	0 – 5	2.50	1.60	0 – 5	2.83	1.50	0 – 6
Moderate	3.93	1.02	1 – 5	4.15	1.03	2 – 6	3.98	1.23	1 – 6
Severe	5.15	0.62	4 – 6	5.45	0.68	4 – 6	4.65	1.29	2 – 6
Putonghua Rough									
Normal	0.53	0.75	0 – 3	1.00	1.15	0 – 5	1.25	1.17	0 – 5
Mild	1.40	1.06	0 – 5	1.83	1.20	0 – 5	1.35	1.39	0 – 5
Moderate	1.65	1.29	0 – 6	2.40	1.57	0 – 6	3.10	1.22	1 – 6
Severe	3.63	1.43	1 – 6	4.55	1.45	2 – 6	3.98	1.21	1 – 6