Cultural and Language Differences in Voice Quality Perception: A Preliminary Investigation Using Synthesized Signals

Edwin M.-L. Yiu^{a, b} Bruce Murdoch^c Kathryn Hird^d Polly Lau^a
Elaine Mandy Ho^a

^aVoice Research Laboratory, Division of Speech and Hearing Sciences, and ^bCentre for Communication Disorders, The University of Hong Kong, Hong Kong, SAR, China; ^cDivision of Speech Pathology, The University of Queensland, Brisbane, Qld, and ^dSchool of Psychology, Curtin University of Technology, Bentley, W.A., Australia

Key Words

Voice assessment · Perceptual voice · Voice quality · Dysphonia

Abstract

Background: Perceptual voice evaluation is a common clinical tool. However, to date, there is no consensus yet as to which common quality should be measured. Some available evidence shows that voice quality is a language-specific property which may be different across different languages. The familiarity of a language may affect the perception and reliability in rating voice quality. **Aims:** The present study set out to investigate the effects of listeners' cultural and language backgrounds on the perception of voice qualities. **Methods and Procedures:** Forty speech pathology students from Australia and Hong Kong were asked to rate the breathy and rough qualities of synthesized voice signals in Cantonese and English. **Outcome and Results:** Results showed that the English stimulus sets as a whole were rated less severely than the Cantonese stimuli by both groups of listeners. In addition, the male Cantonese and English breathy stimuli were rated differently by the Australian and Hong Kong listeners. **Conclusions and Implications:** These results provided some evidence to support the claim that cultural and language backgrounds of the listeners would affect the

perception for some voice quality types. Thus, the cultural and language backgrounds of judges should be taken into consideration in clinical voice evaluation.

Copyright © 2008 S. Karger AG, Basel

Introduction

A survey of the literature on perceptual voice evaluation shows some agreement as to what specific perceptual vocal qualities should be measured. Studies from Australia [1], Belgium [2], Hong Kong [3–6], Japan [7], Sweden [8], England [9] and North America [10–13] indicated that abnormal roughness and breathiness are the two most commonly studied qualities. Roughness is often perceived as an uneven quality lacking clarity [14, 15], while breathiness is perceived as audible friction noise or air escape during phonation [14]. Listeners generally show a relatively higher intra- and inter-judge reliability in rating rough and breathy qualities than in rating other qualities [16]. The uniformity in the investigation of roughness and breathiness perception facilitates the possibility to share, compare and transfer data across research and clinical centers in different parts of the world. However, cross-centre studies across the world should consider possible cultural and language differences across

data. Neural substrates for voice quality perception have been implicated in a functional magnetic resonance imaging study by Belin et al. [17], which suggests that voice perception is a cognitive process and is likely to be affected by learning experience, such as the cultural and language backgrounds of the judges.

There is evidence to suggest that the production of voice quality is different across languages. Bruyninckx et al. [18] showed that the voice qualities produced by Catalan-Spanish bilinguals, as measured by long-term average spectrum, were significantly different in the two languages spoken by the same person. It has been shown that English speakers with either a British or American cultural backgrounds were more accurate in identifying voices of speakers from their own culture (88% accuracy rate) than those from a different culture (13% accuracy rate) [19]. This suggests that the familiarity of voices allowed native speakers to recognize voices from their own culture more readily. In a cross-language study, Anders et al. [20] found the ratings made by a group of American listeners on a set of German dysphonic stimuli less severe when compared to the ratings made by German and Finnish listeners. The investigators contended that the American listeners were not familiar with German voice qualities and they rated the German voices more conservatively or less severely. However, in their study, the participants had different experience and training in perceptual voice evaluation. The difference in the rating might have been attributed to factors other than the cultural and language familiarity per se.

There is therefore a need to determine how judges from different cultural and language backgrounds would rate voice qualities across different languages. The present study set out to investigate the effect of listeners' cultural and language backgrounds on perceptual voice rating. The possible effect that the language of the voice stimuli might have on perceptual judgment of the listeners is referred to as '*Stimulus*' effect. '*Listener*' effect refers to the cultural, ethnic and language background of the listeners that might have an effect on perceptual judgment. It was hypothesized that listeners would rate the abnormal voice qualities of a language that they were less familiar with by using a less stringent standard when compared to listeners who were native speakers of that language. Synthesized standard references (anchors) and testing stimuli were chosen over natural voice samples as it would be relatively difficult to find natural pathological voice samples that carry only one specific type of quality. Synthesizing signals also have the advantage of having relatively simple acoustic properties than natural patho-

logical voice which facilitates investigation of the relationship between acoustic and perceptual ratings [11, 21–24]. An additional advantage with the use of synthesized signals is that other investigators can easily reproduce the stimuli to replicate or expand the study.

The Sensimetrics' HLSyn software, which is essentially a Klatt synthesizer that uses a parallel or cascade synthesis method [25] was used in the present study. It has been shown that the manipulation of aspiration noise amplitude (AH), spectral tilt, longer open quotient or increased bandwidths of F1 and F2 produced signals simulating breathy quality [5, 26]. Synthesis of rough quality using the Klatt synthesizer is less straight forward than that of breathy quality. Klatt and Klatt [25] suggested increasing the flutter parameter produced a rough quality. However, it has been found that the manipulation of amplitude of voicing (AV) and diplophonia (DI) parameters together produced a quality which better resemble roughness in dysphonic voices [5]. It should be noted however, that the DI parameter synthesizes two glottal pulses in slightly different phases [25] and when the pulse of one glottal pair was attenuated to zero, the fundamental frequency would be halved [5].

Method

Participants

Forty speech pathology students participated as judges in this study. Twenty of them (18 females, 2 males; mean age = 20.85 years, standard deviation = 1.42 years, range = 19–26 years) were Hong Kong Cantonese students from the University of Hong Kong (Hong Kong listener group) and 20 were Australian students (12 females, 8 males; mean age = 19.65 years, standard deviation = 1.42 years, range = 17–23 years) from the University of Queensland (Australian listener group). The higher percentage of female judges represented the gender proportion of the speech pathology profession, and it was expected that the gender of the listeners had no effect on the perceptual abilities. Both groups of participants had no prior experience or exposure to voice patients before participating in this study. They all had normal speech, voice and hearing and were native speakers of either Cantonese (Hong Kong listener group) or English (Australian listener group). All the participants from Hong Kong also spoke English as their second language in addition to Cantonese. All of the Australian subjects were monolingual in English only.

Synthesis of the Prototype Stimuli

Four prototype sentences (female Cantonese, male Cantonese, female English and male English) were synthesized using Sensimetrics' HLSyn Speech Synthesis System (Window PC version). The Cantonese prototype was based on the sentence /papatapɔ/, which literally means 'father hits the ball'. The English prototype was based on the sentence 'a baby boy'. These two sentences were chosen as all the words were made up of simple consonant-vowel

Fig. 1. Spectrograms of synthesized prototype signals.

syllable and nonaspirated (for Cantonese) or voiced (for English) consonants.

The values of the synthesis parameters for these two prototype sentences were determined from analyzing six natural sentences produced by six different native speakers (3 males and 3 females) with normal voices in each language. The analyses were carried out using fast Fourier transform and linear predictive coding available in a Kay Elemetrics' Computerized Speech Lab 4300B system. The average values of the fundamental frequency (f_0), the first four formant frequencies (F1, F2, F3 and F4), and the duration of the vowels for each gender in each language were used to synthesize the corresponding prototype sentence. The values of these synthesis parameters were further fine tuned until a relatively natural sounding voice, as determined by two native speakers for that language, was created. Figure 1 shows the spectrogram for each of the prototype sentences.

Synthesis of Breathy and Rough Qualities

Breathy and rough qualities were added to each prototype sentence using the synthesis parameters determined for the Cantonese sentence in an earlier study [5]. In that study, breathy quality was created by increasing the values of the amplitude of aspiration (AH) synthesis parameter. Rough quality was created by increasing the values of both the AV and DI synthesis parameters. The breathy and rough threshold values for the female voice were AH50 and AV80DI-2.5, respectively, and those for the male voice were AH65 and AV80DI-1.5, respectively [5]. These synthesis values, which were developed for the Cantonese stimulus sets, were also used for the English stimulus sets in the present study so as to allow possible cross-language comparison. For each stimulus set, two signals with synthesis values below the threshold and three signals with synthesis values above the threshold were created. Therefore, each quality set contained five stimulus signals

Table 1. Synthesis values for different stimulus sets

Voice quality	Parameter(s)
<i>Female stimuli</i>	
Breathy	AH5, AH10, AH60, AH70, AH80
Rough	AV65DI-0, AV70DI-0, AV80DI-2.5¹, AV80DI-3, AV80DI-3.5
Normal	Prototype (AH40; AV60)
<i>Male stimuli</i>	
Breathy	AH5, AH10, AH70, AH75, AH80
Rough	AV65DI-0, AV70DI-0, AV80DI-2, AV80DI-2.5, AV80DI-3
Normal	Prototype (AH40; AV60)

Values in bold are above the threshold. Threshold is the minimum value for that particular set of stimuli to be perceived as abnormal.

¹ The threshold of the female rough stimulus set was also included in the stimulus set since the AV + DI parameter allowed only a limited range of synthesizing values for rough quality.

(table 1). A total of eight stimulus sets were synthesized, including two quality sets (breathy and rough) in each gender (female and male) and language (Cantonese and English).

Procedures

The synthesized stimuli, in Microsoft Windows WAV format sampled at 44.1 kHz, were presented using Microsoft Visual Basic program. Participants could choose to listen to each stimulus as frequently as they would like by clicking on the appropriate graphical icon (fig. 2). The stimuli were presented as mono-channel signals in a free field at about 80 dBA, as measured by a sound level meter, through a Pentium III 533 MHz computer with a Creative Sound Blaster Gold sound card and a pair of Sony SRS-PC51 speakers. The speakers were placed 60 cm apart on either side of a monitor and approximately 50 cm in front of the listeners in a sound-treated room. Each participant was required to rate each stimulus for both roughness and breathiness on a visual analogue scale as shown in figure 2. They were asked to mark their ratings by moving the scroll along a 100-mm visual analogue scale, with the left end of the scale representing 'just noticeable' deviation and the right end of the scale representing 'extremely severe' deviation. The distance of each scroll from the left end of the rating scale was measured automatically by the program for each response. Each measurement for the specific quality (breathiness or roughness) was entered into a database to the nearest millimeter.

There were a total of four presentation programs: two gender sets (male and female) for each language (Cantonese and English). Each presentation program consisted of a combination of breathy and rough stimulus sets and a prototype stimulus with normal quality. The breathy and rough stimulus sets were each made up of five stimuli. Each stimulus was presented twice in a randomized order. Each presentation program therefore had 20 dysphonic stimuli (10 breathy and 10 rough) and two prototype stimuli. Each participant took the four presentation programs in a different randomized order in order to eliminate any possible ordering effect.

The participants were given only written definitions of the two qualities (table 2) and no further explanation of the quality types

was given. Five anchors for each of the breathy and rough qualities were given before and during the test. One of the anchors was the normal voice (i.e. the prototype signal). One was threshold breathy signal (female – AH50; male – AH65), and one was threshold rough signal (female – AV80DI-2.5; male – AV80DI-1.5). Both were labeled as 'just noticeable' anchors. The other two were labeled as the 'more obvious' breathy and rough signals, respectively. They used synthesis values which were approximately half of the maximum synthesis values allowed by the synthesizer (female breathy – AH70; female rough – AV80DI-3; male breathy – AH75; male rough – AV80DI-2.5).

The participants were given three trial items to familiarize themselves with the format of the rating procedures and the anchors before attempting each presentation program. The program was designed so that the participants could choose to listen to each stimulus and anchors as many times as they would like but they were not able to re-rate a stimulus once it was already rated.

Data Analysis

The acoustic properties of the synthesized signals were analyzed using Kay Elemetrics' Computerized Speech Lab (CSL 4300B) and Multidimensional Voice Program (MDVP), which has been shown to tolerate to the fluctuation of acoustic properties in connected speech [27]. Each sentence was extracted from the onset of the first word (/pa/ for Cantonese and /a./ for English) to the offset of the last word (/pɔ/ for Cantonese and /boi/ for English). Jitter (relative average frequency perturbation – RAP, and pitch perturbation quotient – PPQ), shimmer (shimmer percent – Shim%, and amplitude perturbation quotient – APQ) and noise to harmonic ratio (NHR) of each signals were determined by the MDVP program.

For the perceptual data, only the ratings of the designated quality for each stimulus were analyzed, i.e. breathiness ratings were analyzed for the AH stimuli and roughness ratings were analyzed for the AV-DI stimuli.

Intra-listener agreement was calculated from the two ratings of the same stimulus to determine the consistency of the listeners in rating breathiness and roughness. Agreement on the two rat-

Fig. 2. A sample page of the presentation program.

Table 2. Definitions of abnormal voice qualities

Rough

Synonymous with 'harshness' or 'hoarseness'

Perceptual correlates

- Irregular quality
- Random fluctuations of glottal pulse
- Lack of clarity
- Uneven quality

Acoustic correlates

- Aperiodic mode of vibration
- Perturbation of the spectrum

Breathy

Synonymous with 'whispery voice' or 'whisperiness'

Perceptual correlates

- Audible sound of expiration
- Audible air escape
- Audible friction noise

Acoustic correlates

- Related to a significant component of noise due to turbulence

ings of the stimulus by the same listener within 10 mm of one another on the rating scale was taken. Pearson's correlation coefficients were also calculated between the listeners' two ratings on the same set of stimuli.

Intraclass correlation coefficient (ICC) was used to determine the inter-listener reliability. The ICC (2,1) model was used as this shows the average agreement between one listener and another [28].

Multivariate analysis of variance (MANOVA) using Pillai's criterion was used to analyze the perceptual data as it is more robust than other multivariate statistics in accommodating possible violations of homogeneity of variance-covariance matrices [29, 30]. The two sets of stimuli (English and Cantonese – 'Stimulus' effect) and the severity levels (five levels of severity – 'Severity' effect) were treated as the within-subject variables and the two listener groups (Australian and Hong Kong – 'Listener' effect) as the between-subject variable.

Results

Acoustic Properties

Breathy Stimulus Sets

The jitter (RAP and PPQ), shimmer (Shim% and APQ) and NHR values of the breathy signals are illustrated in figure 3. Overall, both the jitter (RAP and PPQ) and shimmer (Shim% and APQ) measures showed an increasing trend with increasing AH values in the female and male breathy stimulus sets in both languages. The two shimmer values (PPQ and RAP) showed an increasing trend with increasing AH values in the male and female Cantonese sets and the female English set but the male English set showed a relatively flat contour. The increasing trend in the Shim% was obvious with all four sets of stimuli (fig. 3). The APQ showed similar increas-

Fig. 3. Acoustic properties of breathy stimuli.

ing trends with the male and female Cantonese breathy stimuli but not with the English stimuli. The NHR values, on the other hand, remained relatively similar across the different AH values in all four sets of stimuli (fig. 3).

Rough Stimulus Sets

Figure 4 shows the acoustic properties of the synthesized rough signals. All signals in both gender sets and language sets showed an increasing trend in the jitter values (RAP and PPQ) with increasing AV-DI values. The female Cantonese set showed an increasing trend in the shimmer (Shim% and APQ) and NHR values. However, the shimmer measures (Shim% and APQ) showed a decreasing trend with increasing AV-DI values while the NHR measures showed very little change. This apparent decreasing trend might have been due to the error in the

MDVP program in extracting the shimmer measures as a result of the disappearance of alternate pulses in the signals with increasing DI values.

Reliability in Perceptual Ratings

Table 3 lists the mean intra-listener reliability and the mean agreement values in rating the eight types of stimuli by the Australian and Hong Kong listeners. The intra-listener agreement gave an indication of how well the listeners rated the same stimulus within 10 mm. The probability of two ratings that fall within 10 mm on a 100-mm scale or one unit on an 11-point (10-unit) scale merely by chance is 0.26 $[11 + 2(11-1)/11^2]$ [see Kreiman et al. 28]. The values in rating the breathy stimuli were all above the chance level and ranged from 0.43 to 0.58 (table 3), whereas in rating

Fig. 4. Acoustic properties of rough stimuli.

roughness, they ranged from 0.32 to 0.59. The reliability coefficients in rating the breathy stimuli ranged from 0.73 to 0.94. However, in rating the rough stimuli, lower reliability values were noticed, ranging from 0.41 and 0.84.

The inter-listener reliability was calculated using ICC (table 4). For the female stimulus sets, the values of the ICC varied from 0.38 to 0.60. The male Cantonese breathy ratings showed a negative value (−0.05) while the ICC values of the other male stimulus sets ranged from 0.38 to 0.67.

Perceptual Ratings

Table 5 lists the perceptual ratings made by the two groups of listeners. The MANOVA results are presented in the following sections for each of the gender sets with the breathy and rough stimuli. A Bonferroni adjustment was made on the α -level (0.0125) as four sets of MANOVA were carried out (0.05/4).

Table 3. Mean intra-listener reliability measures (Pearson's correlation coefficients) and agreement (within 10 mm on the rating scale) of perceptual ratings

	Pearson's r		Agreement	
	Australian listeners	Hong Kong listeners	Australian listeners	Hong Kong listeners
<i>Breathy stimuli</i>				
Female Cantonese	0.89	0.89	0.54	0.55
Female English	0.90	0.90	0.55	0.57
Male Cantonese	0.73	0.84	0.43	0.58
Male English	0.93	0.94	0.52	0.61
<i>Rough stimuli</i>				
Female Cantonese	0.41	0.62	0.40	0.44
Female English	0.61	0.49	0.59	0.57
Male Cantonese	0.57	0.84	0.32	0.51
Male English	0.61	0.49	0.59	0.57

All Pearson's r coefficients are significant at 0.01 level.

Breathy Stimulus Sets

Female Set

Figure 5 illustrates the increasing perceptual ratings with increasing AH values by the two groups of listeners. The perceptual ratings also covered a wide range from value 0 to 95 mm. The rating patterns were rather similar between the two groups of listeners across the different language stimulus sets.

- Main effects: No significant main ‘*Stimulus*’ or ‘*Listener*’ effects ($p \geq 0.03$, table 6) were found. Main ‘*Severity*’ effect, however, was significant ($p < 0.0001$, table 6).
- Interaction effects: None of the interaction effects was statistically significant ($p \geq 0.02$, table 6).

Male Set

Figure 6 shows that the severity ratings increased with AH values steadily, covering a range from 0 to 90 mm for three of the four sets of data. The male Cantonese breathy

stimuli rated by the Hong Kong listeners showed an abrupt drop in the severity rating from 80 (for AH75) to 40 mm (for AH80).

- Main effects: The main ‘*Severity*’ effect was statistically significant ($p < 0.0001$, table 6). However, neither the ‘*Stimulus*’ nor the ‘*Listener*’ effect was statistically significant ($p > 0.3$, table 6).

Table 4. Intra-class correlation coefficients

Female Cantonese breathy (AH)	0.60
Female English breathy (AH)	0.38
Male Cantonese breathy (AH)	–0.05
Male English breathy (AH)	0.46
Female Cantonese rough (AV+DI)	0.57
Female English rough (AV+DI)	0.54
Male Cantonese rough (AV+DI)	0.67
Male English rough (AV+DI)	0.59

Table 5. Perceptual ratings by the Australian and Hong Kong listeners

Stimulus types			Listeners	Perceptual ratings				
				AH5	AH10	AH60	AH70	AH80
Female breathy	Cantonese	Hong Kong	0.00	0.08	41.08	78.73	95.38	
		Australian	1.75	1.10	36.08	60.68	88.98	
	English	Hong Kong	1.13	0.43	42.30	83.83	91.25	
		Australian	0.73	2.40	32.15	71.03	86.03	
			AH5	AH10	AH70	AH75	AH80	
Male breathy	Cantonese	Hong Kong	0.00	0.00	43.88	77.95	39.28	
		Australian	5.13	0.25	41.10	58.45	75.68	
	English	Hong Kong	0.03	0.00	50.90	71.08	88.75	
		Australian	0.08	0.03	32.50	68.18	75.20	
			AV65	AV70	AV80DI-2.5	AV80DI-3	AV80DI-3.5	
Female rough	Cantonese	Hong Kong	5.80	16.43	50.88	65.45	63.10	
		Australian	3.95	18.00	49.38	39.48	27.73	
	English	Hong Kong	0.03	0.18	41.38	38.78	39.48	
		Australian	2.30	3.30	40.35	20.48	35.58	
			AV65	AV70	AV80DI-2	AV80DI-2.5	AV80DI-3	
Male rough	Cantonese	Hong Kong	2.43	11.53	67.20	57.00	69.30	
		Australian	3.95	10.33	60.73	53.25	62.30	
	English	Hong Kong	0.08	3.25	36.85	36.33	44.35	
		Australian	4.45	5.43	25.70	21.75	16.00	

Fig. 5. Perceptual ratings for female breathy stimulus sets.

Table 6. MANOVA results of perceptual ratings

Set	Effects	Pillai's trace F	d.f.	p
<i>Breathy stimuli</i>				
Female set	Stimulus	0.28	1	0.59
	Severity	521.46	4	<0.0001*
	Listeners	5.01	1	0.03
	Stimulus × listeners	0.000	1	0.99
	Severity × listeners	1.99	4	0.12
	Stimulus × severity	3.21	4	0.02
	Stimulus × severity × listeners	0.63	4	0.64
Male set	Stimulus	5.09	1	0.03
	Severity	279.97	4	<0.0001*
	Listeners	0.477	1	0.49
	Stimulus × listeners	7.396	1	0.01*
	Severity × listeners	4.41	4	0.005*
	Stimulus × severity	5.92	4	0.001*
	Stimulus × severity × listeners	7.46	4	<0.0001*
<i>Rough stimuli</i>				
Female set	Stimulus	29.82	1	<0.0001*
	Severity	35.14	4	<0.0001*
	Listeners	4.28	1	0.04
	Stimulus × listeners	4.37	1	0.04
	Severity × listeners	3.35	4	0.02
	Stimulus × severity	4.08	4	0.008*
	Stimulus × severity × listeners	2.81	4	0.04
Male set	Stimulus	68.97	1	<0.0001*
	Severity	54.76	4	<0.0001*
	Listeners	3.39	1	0.07
	Stimulus × listeners	1.56	1	0.22
	Severity × listeners	2.37	4	0.07
	Stimulus × severity	15.88	4	<0.0001*
	Stimulus × severity × listeners	1.98	4	0.12

* Significant values at 0.0125 level.

Fig. 6. Perceptual ratings for male breathy stimulus set.

Fig. 7. Perceptual ratings for female rough stimulus sets.

– Interaction effects: All the two-way interaction effects and the three-way interaction effect were statistically significant (table 6). Closer examination of the pooled data showed that the ‘*Listener by Stimulus*’ effect indicated the Hong Kong listeners rated the male Cantonese breathy signals comparatively less severely than the Australian listeners. However, they rated the English breathy stimuli more severely than the Australian listeners. The perceptual ratings of both the Australian and Hong Kong listeners increased with increasing AH values until the value reached AH80, at which point the ratings given by the Hong Kong listeners on Cantonese stimuli decreased. The significant three-way interaction effect (table 6) suggested that the two listener groups performed differently for the different types and severity of stimuli.

Rough Stimulus Sets

Female Sets

Figure 7 shows the severity ratings of the female rough stimulus sets given by the two groups of listeners. The severity ratings increased with increasing AV and DI values. It should be noted that the highest mean perceptual ratings were around 40 mm for the most severe English stimulus in both listener groups. The Cantonese stimuli with a value greater than DI-3 were rated at around 65 mm by the Hong Kong listeners. The Australian listeners, however, perceived the signals with levels above DI-3 as less severe (between 30–40 mm).

– Main effects: The main effect of ‘*Stimulus*’ was statistically significant ($p < 0.0001$, table 6). This indicated that the Cantonese and English stimulus sets were perceived differently by the listeners. It was found that the English stimuli as a whole were rated less severely than the Cantonese stimuli (fig. 7). This increasing trend of

Fig. 8. Perceptual ratings for male rough stimulus sets.

ratings with AV and DI was statistically significant (main 'Severity' effect: $p < 0.0001$, table 6). The main 'Listener' effect, however, was not statistically significant ($p = 0.04$).

- Interaction effects: The only significant interaction effect was the 'Stimulus by Severity' effect ($p = 0.008$, table 6). Closer examination of the data showed that the Cantonese stimuli demonstrated rather steady perceptual ratings as the acoustic values increased. The English stimuli, on the other hand, demonstrated a decreasing trend followed by an increasing trend at around ratings 30–40 mm. Other interaction effects were not statistically significant ($p \geq 0.02$, table 6).

Male Sets

Figure 8 shows the severity ratings of the English and Cantonese male rough stimulus sets given by the two groups of listeners. The severity ratings increased as the acoustic signals increased with AV and DI values. The severity ratings of the Cantonese stimuli by the Hong Kong and Australian listeners ranged from 2 to 69, while the ratings of the English stimuli ranged from 0 to 44 mm.

- Main effects: The main effect of 'Stimulus' was statistically significant ($p < 0.0001$, table 6). Figure 8 shows clearly that the Cantonese stimulus sets were rated more severely than the English stimulus sets by the listeners. The main 'Severity' effect was also significant ($p < 0.0001$, table 6) while the main 'Listener' effect, however, was not statistically significant ($p = 0.07$, table 6)
- Interaction effects: The 'Stimulus by Severity' interaction effect was statistically significant ($p = 0.008$, ta-

ble 6). Figure 8 shows that this interaction effect – with the perceptual ratings of the Cantonese stimulus sets increased more drastically than those of the English stimulus sets. Other interaction effects were not statistically significant ($p \geq 0.02$, table 6).

Discussion

Acoustic Properties

Acoustic analysis was carried out to determine whether the synthesized signals across the two languages were comparable in acoustic properties. The breathy signals demonstrated increasing acoustic perturbation (shimmer and jitter) with increasing synthesis AH values. The increasing trend was comparatively clearer in the Cantonese set than in the English set (fig. 3). The synthesized rough signals demonstrated increasing jitter values with increasing AV-DI values. However, the trend in the shimmer values were unusual as the Shim% and APQ values decreased to around 5 units once DI values were added to the rough signals (fig. 5). The addition of the DI parameter to a signal effectively reduces the number of pulse cycles of the signals to half [5]. Therefore, the shimmer values, which were based on the amplitude of the pulse, might have become inaccurate.

In general, the acoustic properties of the Cantonese and English synthesized signals were broadly similar although the Cantonese showed a slightly clearer increasing trend in some acoustic measures. One would expect that the acoustic properties determine the perceptual qualities of the signals. However, there is no simple relationship between the acoustic and perceptual properties.

The discussion on perceptual data presented in the following section will show this clearly.

Perceptual Ratings

The intra-listener agreement and inter-listener reliability showed no obvious patterns of difficulties in rating a particular stimulus set or by a specific listener group. They were of low to moderate reliability and comparable to those reported in other studies [e.g. 4, 31].

The perceptual ratings of the breathy stimulus sets (AH) covered a wide range. Three of the four AH stimulus sets were perceived as breathy ranging from a mean severity of 0.1–92 on the 100-mm visual analog scale. The male Cantonese AH set showed a narrower range (mean 2.5–57). The rough stimulus sets demonstrated a narrower range, from rating 0 to around 50. These findings suggested that the synthesized rough stimulus sets did not include a sufficient number of signals that were perceived by the listeners as severe enough to represent the upper end of the severity scale. Indeed, this has been a limitation of the Klatt synthesizer in that the model for rough or aperiodic signals is not adequately covered in the Klatt synthesizer [5]. A similar severity effect was also found in the ratings of the four AV-DI set. The ratings increased with increasing AV-DI values but the judges did not seem to use much of the higher (more severe) end of the rating scale. We attributed this to the limitation of the AV-DI parameter in which higher values of DI change the perceptual quality to a more ‘tremor’-like quality rather than merely ‘rough’ [5].

The main ‘Severity’ effect was found in all breathy (AH) and rough (AV-DI) stimulus sets, indicating that the severity ratings increased with increasing AH and AV-DI values. Significant main ‘Stimulus’ effect was found in the female and male rough stimulus sets. The English stimulus set was rated less severely as a whole than the Cantonese stimulus set by the two groups of listeners. Such ‘Stimulus’ effect was not demonstrated in the female and male breathy stimulus sets. One might expect that the acoustic properties of the signals would help explain such difference between the breathy and rough signals. However, the shimmer values (Shim% and APQ) in the Cantonese breathy set were comparatively higher than those in the English set. If the hypothesis based on the acoustic properties is correct, one would expect that the ‘Stimulus’ effect would be demonstrated in the breathy set as well. However, this was not the case. Indeed, there was no simple relationship between the perceptual quality and a specific acoustic property. So if the acoustic properties could not be used to explain the results, the

question of what other factors contributed to the results remains to be answered. It was contended that breathiness is a relatively simpler psycho-perceptual quality than roughness, and therefore, listeners would not be easily affected by the differences in phonetic context between two language sets. However, the psycho-perceptual concept of roughness may not be as distinctive as breathiness, and therefore, the difference in phonetic context between English and Cantonese might have affected the listeners’ judgments differently and more readily.

An objective of this study was to determine whether the listeners rated the stimuli in a language that they were not familiar with more conservatively (i.e. less severely). This could be indicated by the significant ‘Listener by Stimulus’ interaction effect. Only the male breathy stimulus set showed this significant interaction effect. This suggested that these two groups of judges operated differently even though the same anchors were given to assist them in deciding on the severity ratings. However, the results did not support the hypothesis that the listeners used a less stringent criterion for a language that they were not familiar with. The findings showed that the Australian listeners rated the English stimuli less severely than the Cantonese stimuli while the Hong Kong listeners rated the Cantonese stimuli less severely than the English stimuli. Furthermore, it was not easy to explain why only the male breathy set showed this interaction effect. One possible reason was that this was purely a false-positive error. However, a Bonferroni adjustment has been used to set the alpha level at 0.0125, which reduced the chance of committing the type I errors. Another possible reason was that the study did not have enough power to show the effect in the other three sets of stimuli because of the small sample size. Further study with a larger sample size that has enough power would help clarify this finding.

A significant ‘Stimulus by Severity’ interaction effect was found in three sets of stimuli – male breathy, female and male rough. A closer examination of the male breathy stimulus set showed the statistically significant effect was not of much clinical significance as the interaction was caused by a reduction in the perceptual ratings of the Cantonese AH80 signal only. In the male and female rough stimulus sets, the Cantonese stimulus set in each gender set was rated more severely than the corresponding English set across the board. The significant interaction effects could be attributed to the smoother increase in the female Cantonese rough set than in the English set, and the steeper increase in the severity rating of the male Cantonese set than in the English set (fig. 8).

In conclusion, the significant results with some of the stimulus sets indicated that cultural difference plays a role in voice quality perception. The hypothesis that voice quality perception is affected by cultural and language backgrounds of listeners was partially supported by the findings of the present study. Equivocal results might have been attributed to the small sample size which led to an inadequate power.

Acknowledgements

The authors would like to acknowledge that this project was supported in part by grants from the Hong Kong Research Grant Council (HKU7197/98H, HKU72821/04H). We would also like to express our sincere gratitude to the speech pathology students who participated in the study.

References

- Oates J, Russell A: Learning voice analysis using an interactive multi-media package: development and preliminary evaluation. *J Voice* 1998;12:500–512.
- De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C: Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *J Voice* 1997;11:74–80.
- Chan KM-K, Yiu EM-L: Clinical voice assessment I: assessing vocal quality using synthesized anchors. Acquiring knowledge in speech, language and hearing. *Speech Pathol Aust* 2001;3:67–68.
- Chan KM-K, Yiu EM-L: The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res* 2002;45:111–126.
- Yiu EM-L, Murdoch B, Hird K, Lau P: Perception of synthesized abnormal voice quality by Cantonese speakers. *J Acoust Soc Am* 2002;112:1091–1101.
- Yiu EM-L, Ng C-Y: Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clin Linguist Phon* 2004;18:211–229.
- Hirano M, Hibi S, Terasawa R, Fujiu M: Relationship between aerodynamic, vibratory, acoustic and psychoacoustic correlates in dysphonia. *J Phon* 1986;14:445–456.
- Hammarberg B, Fritzell B, Gauffin J, Sundberg J, Wedin L: Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngologica (Stockh)* 1980;90(suppl): 441–451.
- Carding P: Evaluating Voice Therapy: Measuring the Effectiveness of Treatment. London, Whurr Publishers, 2000.
- Kreiman J, Gerratt BR: The perceptual structure of pathologic voice quality. *J Acoust Soc Am* 1996;100:1785–1795.
- Wolfe V, Fitch J, Martin D: Acoustic measures of dysphonic severity across and within voice types. *Folia Phoniatr Logop* 1997;49: 292–299.
- Kreiman J, Gerratt BR, Berke GS: The multidimensional nature of pathologic vocal quality. *J Acoust Soc Am* 1994;96:1291–1302.
- Martin D, Wolfe VI: Effects of perceptual training based upon synthesized voice signals. *Percept Motor Skills* 1996;83:1291–1298.
- Hirano M: Clinical Examination of Voice. Vienna, Springer, 1981.
- Laver J: The Phonetic Description of Voice Quality. Cambridge, Cambridge University Press, 1980.
- Revis J, Giovanni A, Wuyts F, Triglia J: Comparison of different voice samples for perceptual analysis. *Folia Phoniatr Logop* 1999; 51:108–116.
- Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B: Voice-selective areas in human auditory cortex. *Nature* 2000;403:309–312.
- Bruyninckx M, Harmegnies B, Llisterri J, Pocholive D: Language-induced voice quality variability in bilinguals. *J Phon* 1994;22: 19–31.
- Doty ND: The influence of nationality on the accuracy of face and voice recognition. *Am J Psychol* 1998;111:191–214.
- Anders LC, Hollien H, Hurme P, Sonninen A, Wendler J: Perception of hoarseness by several classes of listeners. *Folia Phoniatr* 1988;40:91–100.
- Deal R, Emanuel F: Some waveform and spectral features of vowel roughness. *J Speech Hear Res* 1978;21:250–264.
- Hirano M, Hibi S, Yoshida T, Hirade Y, Kasuya H, Kikuchi Y: Acoustic analysis of pathological voice. *Acta Otolaryngologica (Stockh)* 1988;105:432–438.
- Kreiman J, Gerratt BR, Precoda K: Listener experience and perception of voice quality. *J Speech Hear Res* 1990;33:103–115.
- Martin D, Fitch J, Wolfe V: Pathological voice type and the acoustic prediction of severity. *J Speech Hear Res* 1995;38:765–771.
- Klatt DH, Klatt LC: Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J Acoust Soc Am* 1990;87:820–857.
- Bangayan PT, Long C, Alwan AA, Kreiman J, Gerratt BR: Analysis by synthesis of pathological voices using the Klatt synthesizer. *Speech Commun* 1997;22:343–368.
- Yiu EM-L, Worrall LE, Longland J, Mitchell C: Analysing vocal quality of connected speech using Kay's Computerized Speech Lab: a preliminary finding. *Clin Linguist Phon* 2000;14:295–305.
- Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS: Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res* 1993;36: 21–40.
- Max L, Onghena P: Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. *J Speech Lang Hear Res* 1999;42:261–270.
- Tabachnick BG, Fidell LS: Multivariate analysis of variance and covariance; in Tabachnick BG, Fidell LS (eds): *Using Multivariate Statistics*. Boston, Allyn and Bacon, 2001, pp 375–440.
- Gerratt BR, Kreiman J, Antonanzas-Barroso N, Berke GS: Comparing internal and external standards in voice quality judgments. *J Speech Hear Res* 1993;36:14–20.