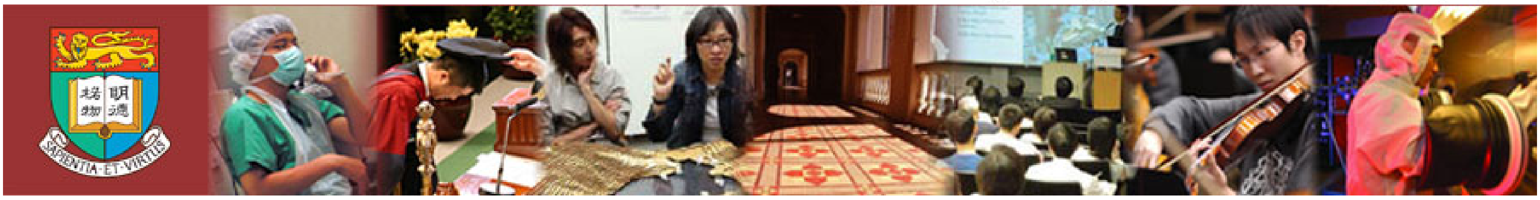


Difference in severity rating between sustained vowels and connected speech in perceptual voice analysis.

Carmen Ma, Ka-man

A dissertation submitted in partial fulfillment of the requirements for the Bachelor of Sciences (Speech and Hearing Sciences), The University of Hong Kong, 7 May 2004.

Abstract

The primary objective of this study was to determine whether there is rating difference in severity between sustained vowels and connected speech for breathiness and roughness in perceptual voice evaluation. The second objective was to investigate whether listeners were more confident in rating sustained vowels than connected speech. Twenty-six undergraduate speech pathology students were recruited to rate the severity of breathiness and roughness of natural voice samples using matching method, and to rate their confidence of their ratings.

The results showed, generally, there was rating difference in severity between sustained vowels and connected speech, while there was no confidence difference in rating these two types of stimuli. These findings suggest that the rating of one type of stimulus cannot represent the rating of another type of stimulus. Therefore, both types of stimuli are suggested to be used in perceptual voice evaluation in order to get a comprehensive analysis of dysphonia.

Introduction

Perceptual voice analysis is most frequently used in clinical voice practice and research. It is also often used to validate acoustic and aerodynamic analyses (Hartl, Hans, Vaissiere & Brasnu, 2003; Kreiman, Gabelman & Gerratt, 2003). However, validity and reliability of perceptual analysis are of great concern because of the subjective nature of its procedure.

Validity refers to measuring what it is supposed to measure. In order to have a valid perceptual voice evaluation, every single component in the procedure must be considered thoroughly (e.g. types of stimuli used, judges' experience and scale used). Among these components, types of stimuli used are controversial. The most commonly used stimuli are sustained vowels and connected speech. Advocates for using sustained vowels argue that perception of dysphonic severity in connected speech can be affected by factors other than voice quality, for example, dialects and articulatory aspects. On the other hand, sustained vowels do not have this limitation and raters can focus more on the dysphonic voice quality (de Krom, 1994; Munoz, Mendoza, Fresneda & Carballo, 2002). On the contrary, advocates for the use of connected speech believe that connected speech is more representative of the conversational voice (de Krom, 1994) while sustained vowels are more like singing voice (de Krom, 1994; Klingholtz, 1990). As the primary aim of voice is speaking, using sustained vowels for perceptual voice evaluation may lack content validity (internal validity), unless the voice quality of sustained vowels is fully representative of the voice quality of connected speech (de Krom, 1994). However, as analysis of perceptual characteristics of phonation types in sustained vowels is not as comprehensive as that in connected speech, dysphonic quality of sustained vowel is not likely to fully represent the dysphonic quality of connected speech (See de Krom, 1994).

As use of sustained vowels or connected speech remained controversial, researchers investigate the correlation between the ratings in connected speech and sustained vowels

(Wolfe, Cornell & Fitch, 1995; Revis, Giovanni, Wuyts & Triglia, 1999) and found a correlation of 0.78 (p<0.01) between them (Wolfe et al, 1995), which is relatively strong. However, the judges they recruited in the study were naïve listeners. Experience in perceptual voice evaluation is an important factor in having reliable judgments (see Revis et al., 1999). Therefore, Revis and his colleagues (1999) replicated the correlational study of clinicians' perceptual voice evaluation in sustained vowels and connected speech. They found the correlations varied from 0.65-0.92 (all with p<0.001) for different rating parameters. However, in their study, no control of the production variability in sustained vowels and connected speech was attempted to make. Therefore, the moderate correlations for some parameters may be due to the different production of the sustained vowels and connected speech, but not solely the different perceptual analysis of the listeners. Therefore, the present study aimed at finding whether there is difference between the severity rating of connected speech and sustained vowels in perceptual voice evaluation.

Moreover, there are no reported Cantonese studies investigating the rating difference between sustained vowels and connected speech. The correlation between sustained vowels and connected speech is studied in English (Wolfe et al., 1995) and French (Revis et al., 1999) only. Cantonese is a very different language from English and French. Cantonese is a tonal language which means words with different lexical tones convey different meanings. In connected speech, the transitions of different tone level (e.g. from tone 1 to tone 3) may give some information to the perception of dysphonic voice quality, while this information is missing in sustained vowels. Secondly, there are no voiced stops in Cantonese. Without voiced stops, the dysphonic voice quality due to vocal fold pathology might not be so pronounced.

As the basis for rating sustained vowels and connected speech is so different, listeners' confidence in rating these two types of stimuli may be different. Therefore, the present study also addressed whether the confidence in rating sustained vowels and that in connected speech are different.

Another parameter receiving much attention in perceptual voice evaluation is reliability. Reliability, which refers to how consistent the measurement would be, is partly associated with validity. If the reliability of a study is poor, the study does not measure what it is supposed to measure, which means, the validity will be poor (Kreiman & Gerratt, 1998). Kreiman, Gerratt, Kempster, Erman and Berke (1993) contended that perceptual analysis is based on the listeners' internal standard (mental representation). As every listener refers to his/her own internal standard, inter-rater reliability could hardly be high and validity will also be threatened as mentioned before, reliability is associated with validity. Moreover, the internal standard is susceptible to be influenced by factors other than the acoustic characteristics of voice. Such factors include listener factors (e.g. sensitivity and experience) and task variables (e.g. scale resolution and scale reality) (Kreiman et al., 1993). In other words, listeners' internal standard may be unstable. Therefore, the intra-judge reliability may also be threatened. For these reasons, anchor (external standard) was developed to eliminate the reference to the unstable internal standards and improve reliability (Kreiman et al., 1993).

Effects of two types of anchors, natural and synthesized, on the reliability of perceptual voice evaluation were studied by Chan and Yiu (2002). They found that the synthesized anchors are more effective than natural anchors in improving reliability in perceptual voice evaluation when training is provided. Yiu and Mok (submitted) also supported the use of synthesized anchors because of three main reasons. First, although the natural anchors could match more easily with the testing stimuli in nature, the individual parameters of natural anchors are difficult to vary systematically, as more than one different voice qualities (e.g. breathiness and roughness) may be present in one voice sample; while the individual parameters of synthesized anchors can be manipulated individually (see Yiu & Mok, submitted). Second, large database is needed for the selection of the natural anchors at different severity levels when natural anchors are used; however, in synthesized anchors, the synthesis parameters can be manipulated systemically to produce a range of severity (see Yiu & Mok, submitted). Third, using synthesized anchors make the replication of studies easier since the anchors can be reproduced when the synthesis parameters are clearly stated (see Yiu & Mok, submitted). Therefore, synthesized anchors will be used in the present study.

More recently, matching method was found to significantly improve reliability in rating normal, mild and moderate stimuli when compared to no provision of anchors (Yiu & Mok, submitted). Matching method means using a set of synthesized anchors as external standards for every rating point. In the study by Yiu & Mok (submitted), eight-point scale was employed, which covers a continuum of severity from normal, mild, moderate to severe. There is a synthetic anchor for every rating point. Therefore, judges can just match the normal or pathological natural voice with the best-matched synthetic anchor. This method is based on the theory that the reference to the unstable internal standard can be avoided altogether and thus should yield a more reliable evaluation. Therefore, the present study used the matching method for increasing the reliability.

Types of voice quality also affect reliability of perceptual voice evaluation. The perceptual parameters of breathiness and roughness are chosen because they are the most widely used parameters in perceptual voice analysis (Yiu & Ho, 1991). Moreover, these two parameters have been shown to yield higher reliability than other parameters (see Yiu & Mok, submitted). In the present study, breathiness is referred to the audible sound of expiration, audible air escape and audible friction noise due to incomplete closure of vocal folds during phonation; while roughness is referred to irregular quality and lack of clarity due to irregularity of vocal folds vibration (Chan & Yiu, 2002).

The present study employed a new scaling method, a knob control, for severity rating. In the study by Wuyts, de Bodt and Van de Heyning (1999), the raters rarely used the extremes of the scale. This may suggest there is a bias in the linear visual scale. Therefore, a knob control is proposed to eliminate this bias. Equal-appearing interval (EAI) scale was used for the confidence rating in order not to confuse the raters with the severity rating scale. EAI scale was used because it demonstrated a better intra-rater agreement than visual analog scale (Yiu & Ng, 2004).

This present study had two objectives.

1) The first objective was to determine whether there was a severity rating difference between sustained vowel and connected speech in the perceptual breathiness and roughness. Results from various studies in the correlation between sustained vowels and connected speech are different and different rating parameters have been employed. Therefore, the present study focused on the two parameter, breathiness and roughness. Moreover, there were no Cantonese studies investigated the rating difference. Therefore, the results of this study should inform clinicians on the choice of stimuli in perceptual voice analysis especially for Cantonese speakers. It was hypothesized that there was a rating difference between the two types of stimuli because listeners have different focus on rating sustained vowels (base solely on the quality) and connected speech (base on a more comprehensive analysis).

2) The second objective was to investigate whether the listeners' confidence levels in evaluating sustained vowels and connected speech were different. Perception of sustained vowels focuses on the voice quality but the perception of connected speech involves a more comprehensive, thus more complicated, analysis. It was hypothesized that listeners would be more confident in rating sustained vowels than connected speech. This would also have a clinical implication on choosing stimuli types.

Pilot Studies

The pilot studies aimed at developing sets of synthesized signals to be used as breathiness and roughness anchors in both sustained vowel and sentence in the main study. The synthesized signals were chosen on the basis of: 1) each of them would comprise a range of severity of roughness and breathiness that would be judged perceptually different from one another by at least 80% of the listeners; 2) the anchors for sentence and vowel would be perceived as correspondent on the same scale point.

Method

Preparation of synthesized signals

Vowel of /a/ and sentence of /pa1pa1ta2pɔ1/ (father hits the ball) were synthesized using HLSyn Speech Synthesis System. The HLSyn is a Klatt synthesizer which was chosen in the present study because it is a widely used, commercially available system and it has a number of synthesis parameters that allow the synthesis of dysphonic voice quality, such as turbulent noise component (Alwan, Bangayan, Gerratt, Kreiman & Long, 2000). The vowel /a/ was selected as it is used frequently in perceptual analysis (Aronson, 1980). The duration of vowel /a/ was 3 seconds as this was the approximate duration of the natural stimuli. The sentence /pa1pa1ta2pɔ1/ was chosen because all the consonants are unaspirated stops which do not have the element of frication noise or aspiration noise and therefore will diminish the chance of masking the breathiness quality (Chan & Yiu, 2002).

Prototypes of vowel /a/ and sentence /pa1pa1ta2pɔ1/ with normal voice quality were synthesized for both genders. The parameters of synthesis of each gender (duration, fundamental frequency and formant frequency) were based on those used by Yiu, Murdoch, Hird and Lau (2002). The parameters associated with abnormal voice qualities (amplitude of aspiration (AH), diplophonia (DI), amplitude of voicing (AV) and spectral tilt (TL)) were based on those used by Yiu and Mok (submitted) to make different levels of severity in both breathiness and roughness. The vowel and sentence share similar adjustment in the synthesis parameters to convey the same degree of quality.

Breathiness

Yiu et al. (2002) and Klatt and Klatt (1990) have illustrated that adding aspiration noise (AH) and spectral tilt (TL) to the stimuli would give the perception of breathiness. For this reason, AH and TL were used to synthesize the breathy stimuli. As suggested by Yiu and Mok (submitted), AH value was increased in 5dB steps (TL value was set at 0dB) until the value reached the ceiling of AH value, i.e. 80dB. Then, TL was increased with 20dB steps (see Table 1).

Roughness

Yiu and Mok (submitted) has found that a fixed 80dB AH value, a fixed 80%AV value and diplophonia (DI) value manipulation would give the perception of primarily roughness with minimum vocal fry quality. The manipulation of DI value is the increase of DI value from 2 to 10 in 4% steps and from 10 to 28 in 6% steps.

All other parameters were kept as the default values as suggested by Klatt and Klatt (1990).

Table 1. *The synthesis values for the synthetic stimuli.*

Female breathy	Female rough	Male breathy	Male rough
PROTOTYPE*	PROTOTYPE	PROTOTYPE	PROTOTYPE
AH60	AV80AH80DI2	AH65	AV80AH80DI2
AH65	AV80AH80DI6	AH70	AV80AH80DI6
AH70	AV80AH80DI10	AH75	AV80AH80DI10
AH75	AV80AH80DI16	AH80	AV80AH80DI16
AH80	AV80AH80DI22	AH80TL20	AV80AH80DI22
AH80TL20	AV80AH80DI28	AH80TL40	AV80AH80DI28

**Note*. Default values for the prototype sentence: AH40, AV60, DI0, TL0

Pilot 1

Participants

Ten naïve listeners were recruited to be the judges (3 males and 7 females, mean age=21.6, SD=1.578, range=19-24). It is believed that if naïve listeners can perceive the differences in the anchors, the anchors would then be perceptually distinguishable to all people. Therefore, the anchor set certainly covered a range of severity. The participants were undergraduate students of University of Hong Kong. The participants are native Cantonese speakers and with no reported hearing or voice problem.

Procedure

The judges were presented sentence stimuli using Microsoft PowerPoint 2000 at a comfortable loudness level in a sound-treated room. The stimuli were presented through a pair of professional-quality headphones (Sennheiser, HD 25) and a Creative Extigy Signal Processing unit. Four sets of stimuli: female breathiness, female roughness, male breathiness and male roughness were presented. The order of presentation of these four sets was randomized. The judges were asked to determine whether two stimuli were the same or not in terms of quality. They could hear the stimuli as many times as they needed. Six practice trials (3 male trials and 3 female trials) were given to the judges to familiarize with the procedures before the actual judgment tasks. The practice items were synthesized using the variation of "flutter" parameter (FL=0, 20, 40, 60, other parameters were kept as the prototype). All listeners completed all the four sets in 40 minutes.

Results for Pilot 1

For the series of breathiness for female, the percentage of judges perceived the stimulus pairs as different regarding the severity level ranged from 50% to 80% for sentence and 60% to 100% for vowel, while for the series of breathiness for male, the percentage ranged from 70% to 100% for both sentence and vowel.

Table 2. *Percentage of listeners that perceived the stimulus pairs as different in severity in the male and female breathiness series.*

Female stimulus pairs	Percentage		Male stimulus pairs	Percentage	
	Sentence	Vowel		Sentence	Vowel
PROTOTYPE*/AH60	80	90	PROTOTYPE/AH65	80	90
AH60/AH65	50	60	AH65/AH70	70	70
AH65/AH70	70	80	AH70/AH75	70	70
AH70/AH75	80	90	AH75/AH80	90	90
AH75/AH80	80	100	AH80/AH80TL20	80	80
AH80/AH80TL20	80	80	AH80TL20/AH80TL40	100	100

**Note*. Default values for the prototype sentence: AH40, AV60, DI0, TL0

For the series of roughness for female, the percentage of judges perceived the stimulus pairs as different regarding the severity level ranged from 60% to 80% for sentence and 60% to 90% for vowel, while for the series of roughness for male, the percentage ranged from 10% to 80% for sentence and 30% to 90% for vowel.

Table 3. *Percentage of listeners that perceived the stimulus pairs as different in severity in the male and female roughness series.*

Female stimulus pairs	Percentage		Male stimulus pairs	Percentage	
	Sentence	Vowel		Sentence	Vowel
PROTOTYPE*/ AV80AH80DI2	80	80	PROTOTYPE/ AV80AH80DI2	80	90
AV80AH80DI2/ AV80AH80DI6	80	90	AV80AH80DI2/ AV80AH80DI6	80	80
AV80AH80DI6/ AV80AH80DI10	60	70	AV80AH80DI6/ AV80AH80DI10	60	70
AV80AH80DI10/ AV80AH80DI16	70	70	AV80AH80DI10/ AV80AH80DI16	60	70
AV80AH80DI16/ AV80AH80DI22	70	60	AV80AH80DI16/ AV80AH80DI22	10	30
AV80AH80DI22/ AV80AH80DI28	60	70	AV80AH80DI22/ AV80AH80DI28	10	60

**Note*. Default values for the prototype sentence: AH40, AV60, DI0, TL0

For sentence, as less than 80% of the judges perceived some of the stimulus pairs in breathiness and roughness series as perceptually different, another pilot study (Pilot 2) was carried out. It aimed to investigate what the adjustment should be made so that the stimulus pairs were perceived as different by at least 80% of the judges.

For vowel, the smallest difference that at least 80% of the judges perceived the stimulus pairs as different was 5dB of AH value for breathiness and 4% of DI value for roughness. This served as a basis for the matching choices in Pilot Study 3.

Pilot 2

Participants

The same ten listeners in Pilot 1 were asked to be the judges in this pilot study.

Procedures

The procedures were the same as Pilot 1. For the breathiness series, the following stimulus pairs, AH60/AH70 and AH80TL20/AH80TL40 (for female), and AH65/AH75 and AH80TL40/AH80TL60 (for male) were presented to the judges. For the roughness series, the incremental values of DI parameters with increase of from 4% to 6% with DI values 6% and from 6% to 8% with DI values of 12% were introduced. The following stimulus pairs, AV80AH80DI6/AV80AH80DI12, AV80AH80DI12/AH80AH80DI20, AV80AH80DI20/AV80AH80DI28 and AV80AH80DI28/AV80AH80DI36 (for both male and female) were presented to the judges. The judges were asked to determine whether the stimulus pairs were the same or not.

Results for Pilot 2

For breathiness series, the percentage of judges that perceived the stimulus pairs as different regarding the severity level was 80% for female, while the percentage ranged from

80% to 100% for male.

Table 4. *Percentage of listeners that perceived the stimulus pairs in the female and male breathiness series as different in severity.*

Female stimulus pairs	Percentage	Male stimulus pairs	Percentage
AH65/AH75	80	AH60/AH70	80
AH80TL20/AH80TL40	80	AH80TL40/AH80TL60	100

For roughness series, the percentage of judges that perceived the stimulus pairs as different regarding the severity level ranged from 80% to 90% for both female and male.

Table 5. *Percentage of listeners that perceived the stimulus pairs as different in severity in the male and female roughness series.*

Female stimulus pairs	Percentage	Male stimulus pairs	Percentage
AV80AH80DI6/ AV80AH80DI12	90	AV80AH80DI6/ AV80AH80DI12	80
AV80AH80DI12/ AV80AH80DI20	80	AV80AH80DI12/ AV80AH80DI20	90
AV80AH80DI20/ AV80AH80DI28	80	AV80AH80DI20/ AV80AH80DI28	80
AV80AH80DI28/ AV80AH80DI36	80	AV80AH80DI28/ AV80AH80DI36	90

Pilot 3

This pilot study aimed at determining what the perceptual corresponding manipulation values of vowel were for the sentence. Even if the values of parameters of vowel and sentence were the same, it was not certain whether they were the same in severity perceptually, i.e. a vowel stimulus of AH60 might not be perceived as severe as a sentence stimulus of AH60. Therefore, matching of the two types of anchors were carried out.

Participants

Twenty-four females and four males (mean age=22.12, SD=0.71, range=21-24) were recruited. They were all native Cantonese-speakers with no reported hearing problem or voice problem. They were speech pathology students of University of Hong Kong. They had attended a 3-hour perceptual voice analysis training session before taking part in this study. All participants had passed hearing screening at 250 Hz, 500 Hz, 1000 Hz, 2000 Hz and 4000Hz (i.e. thresholds at 25dB or lower) in a sound-treated room before the task.

Procedures

The judges were presented the stimuli using Microsoft PowerPoint 2000 at a comfortable loudness level in a sound-treated room. A pair of professional-quality headphones (Sennheiser, HD 25) and a Creative Extigy Signal Processing unit were used. The judges were presented with one sentence stimulus and five vowel stimuli with the same synthetic value of the sentence and plus and minus one and two the smallest difference of value the listeners in pilot 1 could perceive, for example, for sentence stimulus of AH70, vowel stimuli of AH60, AH65, AH70, AH75 and AH80 were presented (see Appendix A and B). The judges were asked to choose which vowel stimulus matches the sentence stimulus the best. The judges could hear the stimuli as many times as they needed. All listeners completed all trials in 15 minutes.

Results for Pilot 3

The manipulation value of the vowel which most of the participants chose to match the sentence anchors was selected. For both female and male breathiness series, the value of the vowel matched with the value of sentence. For female roughness series, only the value of AV80AH80DI20 did not match for both sentence and vowel. Instead, the value of AV80AH80DI16 of the vowel matched with AV80AH80DI20 of the sentence. For male roughness series, only the value of AV80AH80DI2 and AV80AH80DI6 matched for the vowel and sentence, but others form a series with increasing severity (see Table 6). Therefore, series of corresponding rating point in vowel and sentence were formed.

Table 6. *The synthetic value for sentence and vowel that was perceived to be correspondent.*

Female		Male	
Sentence	Vowel	Sentence	Vowel
AV80AH80DI20	AV80AH80DI16	AV80AH80DI12	AV80AH80DI16
--	--	AV80AH80DI20	AV80AH80DI24
--	--	AV80AH80DI28	AV80AH80DI32
--	--	AV80AH80DI36	AV80AH80DI40

Summary and Discussion for Pilot Studies

The aim of the pilot studies was to choose synthesized anchors for the Main Study. The synthesized anchors were chosen based on two criteria: 1) each of the anchors was perceived as different from the next stimulus in the series by at least 80% of the naïve listeners; 2) the sentence and vowel anchors were perceived as correspondent for the same scale point. Seven synthesized anchors were eventually chosen for each type of stimuli (vowel and sentence), each gender (female and male) and each voice quality (breathiness and roughness). The anchors were used for the Main Study (see Table 7 and 8) using an eight-point scale (0-7). The most severe dysphonic quality was not given an anchor as, theoretically, there is no upper limit for severity. A recording of “more severe than anchor six” was used for the last scale point instead.

Table 7. *The synthesized breathiness anchors selected for use in the Main Study.*

Female Breathiness		Male Breathiness	
Sentence	Vowel	Sentence	Vowel
PROTOTYPE*	PROTOTYPE	PROTOTYPE	PROTOTYPE
AH60	AH60	AH65	AH65
AH70	AH70	AH75	AH75
AH75	AH75	AH80	AH80
AH80	AH80	AH80TL20	AH80TL20
AH80TL20	AH80TL20	AH80TL40	AH80TL40
AH80TL40	AH80TL40	AH80TL60	AH80TL60

**Note*. Default values for the prototype sentence: AH40, AV60, DI0, TL0

Table 8. *The synthesized roughness anchors selected for use in the Main Study*

Female Roughness		Male Roughness	
Sentence	Vowel	Sentence	Vowel
PROTOTYPE*	PROTOTYPE	PROTOTYPE	PROTOTYPE
AV80AH80DI2	AV80AH80DI2	AV80AH80DI2	AV80AH80DI2
AV80AH80DI6	AV80AH80DI6	AV80AH80DI6	AV80AH80DI6
AV80AH80DI12	AV80AH80DI12	AV80AH80DI12	AV80AH80DI16
AV80AH80DI20	AV80AH80DI16	AV80AH80DI20	AV80AH80DI24
AV80AH80DI28	AV80AH80DI28	AV80AH80DI28	AV80AH80DI32
AV80AH80DI36	AV80AH80DI36	AV80AH80DI36	AV80AH80DI40

**Note*. Default values for the prototype sentence: AH40, AV60, DI0, TL0

Main Study

The main study aimed at 1) determining if there was a rating difference between sustained vowel and connected speech; 2) finding whether listeners felt more confident in rating sustained vowel.

Method

Preparation of testing natural stimuli

For each type of stimuli (vowel /a/ and sentence /pa1pa1ta2pɔ1/), two gender sets of natural pathological voices (male and female) were chosen from a pool of database collected from Voice Research Lab at the University of Hong Kong as testing stimuli. In order to control the production of the sentence and vowel, three experienced judges in voice evaluation were recruited to choose the stimuli. The judges were asked to rate the vowel and sentence stimuli into three categories (mild, moderate and severe). Only those stimuli that all the judges agree on the same severity level, and the sentence and vowel stimuli produced by the same person were of the same severity level were used. Each gender set included two types of quality (breathy and rough) at three levels of severity (mild, moderate and severe) for two types of stimuli (vowel and sentence). Therefore, there were 12 sets of voices for each gender (breathy/rough X mild/moderate/severe X vowel/sentence). There were two samples from different speakers in each group (therefore, 24 stimuli in each gender set). With addition of two normal voice samples for both sentence and vowel, there were a total of 28 voice samples for each gender set. Half of the voice samples were duplicated, therefore, resulting in a total of 42 (28+14) samples in each gender set.

Participants

The same listeners in Pilot 3 were asked to be the judges in the main study.

Procedures

Each participant was presented the stimuli through a computerized program specifically for this study with a Pentium II 533MHz computer. The participant listened to the stimuli through a Creative Sound Blaster Extigy Signal Processing Unit and a pair of professional-quality headphones (Sennheiser, HD25) in a sound treated room at a comfortable loudness level. Samples of synthesized anchors at three severity levels (mild, moderate and severe) were introduced to the participants using Microsoft PowerPoint 2000 at the beginning of the session. The participants were also presented the whole continuum of the synthetic anchors at the beginning of the session. Written definitions of breathiness and roughness (See Appendix C) were given to the participants throughout the session.

The participants were asked to rate the severity of both breathiness and roughness of each voice stimuli of the four blocks (male sentence, female sentence, male vowel and male sentence) on a computer screen. A knob control with an eight-point scale from 0 (represented normal voice quality) to 7 (represented the most severe dysphonic quality) was used. Each of the scale point from 0 to 6 was represented by an anchor, while no anchor was given for point 7 as, theoretically, there is no upper limit of severity (Chan & Yiu, 2002). The presentation order of the four blocks was randomized and the presentation order of the trials in each block was randomized also. The participants were allowed to listen to the stimuli as frequent as they want. In addition, all the participants were asked to rate their confidence level on a seven-point EAI scale from 1 (indicating wild guess) to 7 (indicating an absolute confidence). Three practice trials were given before each block in order to familiarize the listeners with the rating paradigm.

The whole session took about one and a half hour (See Appendix D).

Results

The ratings of breathiness and roughness of each gender set for both vowel and sentence at different severity level were analyzed. The participants rated both the severity of breathiness and roughness for each stimulus, however, only the relevant ratings were analyzed, i.e. only the ratings of breathiness were analyzed for the assigned breathy stimuli and only the ratings of roughness were analyzed for the assigned rough stimuli.

Rating differences in sustained vowel and sentence

Table 9 shows the mean ratings and the results of Wilcoxon signed-ranks tests for vowel and sentence stimuli. The mean ratings ranged from 1.02 to 5.73. Significantly higher ratings were found in vowel for female mild breathy, male mild breathy, male mild rough and male moderate rough stimuli while significantly higher ratings were found in sentence for female severe rough stimuli and male severe breathy.

Table 9. *Mean ratings and z scores of Wilcoxon signed-ranks tests for vowel and sentence stimuli.*

Stimulus	Vowel			Sentence			*z*	*p*
	Mean	SD	Range	Mean	SD	Range		
Female breathy								
Mild	1.54	0.73	0.50-3.00	1.02	0.61	0.00-2.50	-3.22*	0.00
Moderate	2.35	0.90	1.00-4.00	2.21	0.90	0.50-3.50	-0.98	0.33
Severe	4.33	0.98	2.50-5.50	4.06	1.05	1.50-6.50	-1.26	0.21
Overall	2.74	0.75	1.50-4.00	2.43	0.73	1.00-4.17	-2.60*	0.01
Female rough								
Mild	2.00	0.89	0.50-4.00	2.21	0.76	1.00-4.00	-1.16	0.24
Moderate	2.98	0.98	1.00-5.50	2.75	1.01	1.00-5.00	-1.32	0.19
Severe	5.27	0.84	3.50-7.00	5.73	1.01	4.00-7.00	-2.02*	0.04
Overall	3.42	0.70	2.33-5.00	3.56	0.70	2.17-5.17	-0.70	0.48
Male breathy								
Mild	1.39	0.57	0.00-2.00	1.06	0.61	0.00-2.50	-2.09*	0.04
Moderate	1.65	0.61	0.50-3.00	1.52	0.94	0.00-4.00	-0.98	0.33
Severe	4.70	0.86	3.00-6.00	5.37	0.69	4.00-6.50	-3.30*	0.00
Overall	2.58	0.51	1.17-3.50	2.65	0.60	1.50-4.00	-0.41	0.68
Male rough								
Mild	1.83	0.82	0.50-3.50	1.06	0.61	0.00-3.00	-3.69*	0.00
Moderate	3.69	1.08	1.50-5.50	3.23	1.00	1.50-5.00	-2.31*	0.02
Severe	4.39	1.13	2.50-6.50	4.00	1.28	1.50-7.00	-1.57	0.12
Overall	3.30	0.84	2.00-5.00	2.76	0.76	1.33-4.50	-2.93*	0.00

* Significant level p<0.05

Comparison of correlation and difference between rating sustained vowel and sentence

Correlation of mean ratings between vowel and sentence was calculated using Spearman’s rho. The correlation coefficients varied from 0.17-0.68. The result of correlation was compared to the result of mean difference (see Table 10). It was found that even if there was no difference between the mean ratings and there is significant correlation, the correlation was not high.

Table 10. *Correlation coefficients and z scores of Wilcoxon signed-ranks tests for the ratings of sentence and vowel.*

Stimulus	Spearsman’s rho		Wilcoxon signed test	
	r	*p*	*Z*	*p*
Female breathy				
Mild	0.41*	0.04	-3.22*	0.00
Moderate	0.68*	0.00	-0.98	0.33
Severe	0.46*	0.02	-1.26	0.21
Overall	0.67*	0.00	-2.60*	0.01
Female rough				
Mild	0.34	0.09	-1.16	0.24
Moderate	0.42*	0.03	-1.32	0.19
Severe	0.24	0.24	-2.02*	0.04
Overall	0.63*	0.00	-0.70	0.48
Male breathy				
Mild	0.17	0.40	-2.09*	0.04
Moderate	0.62*	0.00	-0.98	0.33
Severe	0.43*	0.03	-3.30*	0.00
Overall	0.59*	0.00	-0.41	0.68
Male rough				
Mild	0.36	0.07	-3.69*	0.00
Moderate	0.64*	0.00	-2.31*	0.02
Severe	0.46*	0.02	-1.57	0.12
Overall	0.56*	0.00	-2.93*	0.00

* Significant level p<0.05

Confidence in rating sustained vowel and sentence

Table 11 lists the mean confidence ratings and z scores of Wilcoxon signed-ranks tests for vowel and sentence stimuli. The mean confidence ratings ranged from 5.21 to 6.02. Significantly high confidence ratings were found in female moderate rough in rating vowel only.

Table 11. *Mean confidence ratings and z scores of Wilcoxon signed-ranks tests for vowel and sentence stimuli.*

Stimulus		Vowel			Sentence			*z*	*p*
		Mean	SD	Range	Mean	SD	Range		
Female breathy									
	Mild	5.77	0.83	3.00-7.00	5.23	1.25	1.00-7.00	-1.70	0.09
	Moderate	5.48	0.79	3.50-6.50	5.21	0.76	3.50-7.00	-1.63	0.10
	Severe	5.73	0.70	3.50-7.00	5.69	0.78	4.00-7.00	-0.30	0.76
	Overall	5.66	0.54	4.33-6.67	5.38	0.61	4.50-6.67	-1.56	0.12
Female rough									
	Mild	5.48	0.71	4.00-7.00	5.50	0.93	3.50-7.00	-0.58	0.56
	Moderate	5.65	0.72	4.00-7.00	5.35	0.75	3.50-6.50	-2.05*	0.04
	Severe	5.75	0.70	4.00-7.00	5.85	0.63	5.00-7.00	-0.43	0.67
	Overall	5.63	0.54	4.83-6.67	5.56	0.54	4.67-6.50	-0.70	0.48
Male breathy									
	Mild	5.54	0.84	4.00-7.00	5.62	1.04	3.00-7.00	-0.23	0.82
	Moderate	5.25	0.85	3.50-7.00	5.44	0.77	3.00-7.00	-1.06	0.29
	Severe	5.79	0.74	3.50-7.00	6.02	0.78	4.00-7.00	-1.29	0.20
	Overall	5.53	0.53	4.17-6.33	5.69	0.52	4.33-6.50	-1.80	0.07
Male rough									
	Mild	5.42	1.09	1.00-7.00	5.87	0.61	5.00-7.00	-1.89	0.06
	Moderate	5.83	0.79	4.00-7.00	5.66	0.81	4.00-7.00	-0.67	0.50
	Severe	5.85	0.81	4.00-7.00	5.81	0.75	4.00-7.00	-0.39	0.70
	Overall	5.70	0.59	4.50-6.83	5.78	0.52	5.00-6.83	-0.78	0.44

* Significant level p<0.05

Discussion

The first objective of the present study was to determine whether there was a rating difference of breathiness and roughness between sustained vowels and connected speech in perceptual voice evaluation. There were only six out of twelve stimuli sets showing significant rating differences. Among these six sets, four of them showed significantly higher rating in vowel while two of them showed significantly higher rating in connected speech. This implied that there was rating difference between sustained vowel and connected speech. Moreover, although there was significant correlation between the severity ratings of these two types of stimuli, the correlation was not strong. The results supported the hypothesis made: there was difference in severity ratings of sustained vowel and connected speech, i.e. rating of sustained vowels cannot represent the rating of connected speech.

The result of the present study did not support the findings in previous studies (Revis et al., 1999; Wolfe et al., 1995). This was likely to be due to the different participants recruited (naïve listeners were recruited in the study by Wolfe et al., 1995) and control of production of speaker (no control was attempted in both of the previous studies).

It is hypothesized that the perceptions of dysphonic severity in sustained vowels and connected speech are based on different criteria, which contribute to the rating differences between these two types of stimuli. Firstly, perception of vowel is less influenced by factors other than voice quality than that of connected speech. In connected speech, variables other than the voice quality, such as dialect and speech rate, affect the perception of voice quality (de Krom, 1994). Sustained vowels, on the hand, do not have these constraints. Therefore, listeners can focus on the voice quality in evaluation. However, how the variables (e.g. dialect and speech rate) impose the impact on the perception of dysphonic voice was not investigated. Further study may focus on the effect of these variables on perceptual voice evaluation.

Secondly, connected speech comprises different vowels (such as /a/ and /ɔ/ in the sentence used in the present study). Perception of voice quality of different vowels is found to be different (Rees, 1958). This implies that the dysphonic quality may be perceived differently in different segments of in a sentence. This means that listeners need to judge the dysphonic severity in the sentence as a whole, regardless of the fact that there is different perception of dysphonic quality in the sentence.

Thirdly, another difference between sustained vowels and connected speech is the presence of consonants in connected speech. According to Rees (1958), dysphonic quality is perceptually different in different consonant context. In his study, he found that harshness for vowels was perceived less severe in voiceless and plosive consonant context. This is to say, different consonants context in connected speech affects the severity rating in connected speech. This is especially true for Cantonese unaspirated stop. In the sentence used in the present study /pa1pa1ta2pɔ1/, there are unaspirated stops /p/ and /t/, which are not voiced. Therefore, there are transitions from unvoiced (consonants) to voiced (vowels) segments. When compared to the vowels which are voiced, the transitions from unvoiced to voiced segments may contribute to the different rating of sustained vowels and connected speech.

Fourthly, Cantonese is a tonal language, which means words with different tones convey different meanings. Sentence typically includes different tones (such as tone 1 and tone 2 in the sentence of /pa1pa1ta2pɔ1/). Therefore, there are tonal level changes. This means, for example, in the sentence of /pa1pa1ta2pɔ1/, tonal level changes through high-level (pa1), high-level (pa1), high-rising (ta2) to high level (pɔ1). In comparison, there is no tonal change in sustained vowels. Therefore, listeners need to take the dysphonic characteristics present in tonal change into consideration in rating connected speech; but not in rating of sustained vowels.

Last but not least, much valuable information about one’s dysphonic quality is present in connected speech, such as vocal onset, vocal termination and voice breaks (see Wolfe et al. (1995). Therefore, listeners may also consider this information in rating of connected speech.

When investigating the rating difference at each severity level, the severity rating difference was found to be significantly higher in mild breathy vowel stimuli for both female and male. The absence of significant rating difference at the more severe dysphonic voice quality for both female and male stimuli could be interpreted as, in mild breathy stimuli, the variables other than the voice quality may mask the mild breathy dysphonic quality in connected speech. Because of the masking, the breathy quality was perceived as less severe when compared to the breathy quality presented by sustained vowel, which is without other variables affecting the perception of voice quality. However, since no significant difference was found in female mild rough stimuli, the masking for all mild dysphonic stimuli cannot be concluded. It may be possible that there is different masking effect for breathiness and roughness. Further study is needed to investigate how the variables other than voice quality in connected speech affect different severity rating of breathiness and roughness.

It was interesting that more severity rating differences were found for male stimuli (four significant differences out of 12 sets of stimuli) than female stimuli (only two significant differences). It may be due to the different criteria for severity rating for sustained vowels and connected speech in the two genders. However, as in the present study, more female participants were recruited, the result might be interpreted as female listeners base on different criteria for severity rating of the two types of stimuli instead. Therefore, further study on the perception of dysphonic voice quality of different genders is needed before a definite conclusion, which stated the criteria for severity rating in different types of stimuli in different genders were different, could be made.

The second objective of this study was to investigate whether listeners would be more confident in rating sustained vowels than connected speech. The results in the present study did not support this hypothesis. The participants showed a significant higher confidence in vowel only in rating female moderate rough stimuli.

Although rating of sustained vowel is considered as simpler and less influenced by factors other than voice quality, listeners might find it more natural and comprehensive in rating connected speech. The result suggested that people feel more confident in severity rating according to different criteria. This means there is individual difference in confidence levels in rating of sustained vowels and connected speech.

Another hypothesis proposed in this study was that using a knob control may eliminate the visual bias in a linear scale. In the present study, the result shows that the extreme rating points (0 and 7) were not found to be used rarely (see Appendix E). However, the result could not be directly compared to the study by Wuyts et al. (1999), as the rating scales used, parameters studied and participants recruited in the present study were different from that used by Wuyts et al. (1999). Moreover, the present study revealed that there seems to be a bias in using the rating one and a general decrease in the frequency of using more severe rating points. Therefore, it is still doubtful about whether the use of control knob can eliminate the visual bias in a linear scale; or the control knob, also creates a bias in using the less severe rating points than the more severe rating points. Further study on the effect of elimination of visual bias by a knob control is needed.

Limitations of the present study

First, although the production of the natural stimuli of vowel and sentence was controlled by using those stimuli which are of the same level of severity in both vowel and sentence, the productions of vowel and sentence may not be exactly the same. Therefore, the fact that there is rating difference between sustained vowels and connected speech may be due to the productions of the two types of stimuli are different, or a combination of the production and perception difference, but may not be solely the difference in the listeners' perception. Other parameters, such as using the vowels and sentences which are acoustically similar, may be possible to be used in controlling the production factor. Second, clippings were found to be present for the more severe synthesized anchors. Although matching of vowel and sentence on the correspondent scale point was made, the clippings may affect the matching of the natural stimuli to the synthesized anchors. Therefore listeners may find it more difficult to use the anchors as a reference, which may in turn, affects the listeners' performance in severity rating. Reduction in the amplitude of the voice signal may be possible in eliminating the clippings.

Conclusion

The results of this study suggest that the severity rating in sustained vowels differs substantially from that in connected speech. Using one type of stimulus cannot represent the severity of another type of stimulus. As sustained vowels are less influenced by variables other than voice quality, and connected speech represents conversational voice more, it is recommended that both of sustained vowels and connected speech should be included in the perceptual voice evaluation to obtain a more comprehensive analysis of a dysphonic voice.

Confidence levels did not vary between rating of sustained vowels and connected speech. This suggests that there are individual differences in the confidence levels in rating sustained vowels and connected speech. Sustained vowels, being acoustically simpler; connected speech with provision of more perceptual cues or being more natural, make different listeners feel more confident in rating different types of stimuli.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Edwin Yiu, for his support and guidance throughout the development of this study. I am also grateful to Professor Jody Kreiman for her valuable advice to this study. I also thank Ms. Karen Chan, Dr. Estella Ma, Dr. Valter Ciocca and Ms. Brenda Wun for their advices at various stages of this study. I would also like to express my thanks to all speech pathology students who have participated in this study.

References

Alwan, A.A., Bangayan, P.T., Gerratt, B.R., Kreiman, J. & Long, C (2000). Analysis by synthesis of pathological voices using Klatt synthesizer. In R. D. Kent & M J. Ball. *Voice quality measurement.* San Diego: Singular Publishing Group.

Aronson A. (1980). *Clinical voice disorders.* New York: Thieme.

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

de Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragment. *Journal of Speech and Hearing Research, 37*, 985-1000.

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

Hartl, D. M., Hans, S., Vaissiere, J. & Brasnu, D. F. (2003). Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia. *European Archieves of Oto-Rhino Laryngology, 260*(4), 175-182

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

Klingholtz, F. (1990). Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels. *Journal of the Acoustic Society of America*, *87*(5), 2218-2224.

Kreiman, J., Gabelman, B. & Gerratt, B. R. (2003). Perception of vocal tremor. *Journal of Speech, Language and Hearing Research, 46*(1), 203-214.

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

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

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 further research. *Journal of Speech and Hearing Research, 36,* 21-40.

Munoz, J. Mendoza, E, Fresneda, M. D.& Carballo, G.(2002). Perceptual analysis in different voice samples: agreement and reliability. *Perceptual and Motor Skills, 94,* 1187-1195.

Rees, M. (1958). Some variables affecting perceived harshness. *Journal of Speech and Hearing Research, 1,* 155-168.

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

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

Yiu, E. (2001). Clinical voice evaluation protocol- An experimental version. In E. Yiu. *Clinical voice assessment and therapy: A Hong Kong perspective.* INSTEP: The University of Hong Kong.

Yiu, E. M.-L. & Ho, P. S.-P.. (1991). Voice problems in Hong Kong: A preliminary report. *Australian Journal of Human Communication Disorders, 19*(1), 45-58.

Yiu, E. M.-L. & Mok, R. (submitted). Matching method in perceptual voice evaluation. *Submitted to Clinical Linguistics and Phonetics.*

Yiu, E., Murdoch, B., Hird, K. & Lau, P. (2002). Perception of synthesized abnormal voice quality by Cantonese speakers. *Journal of the Acoustical Society of America, 112*(3), 1091-1101.

Yiu, E. M.-L., & Ng, C.-Y. (2004). Reliability of equal appearing interval and visual analogue scales in perceptual voice evaluation. *Clinical Linguistics and Phonetics*, *18*, 1-19.

Appendix A

Matching of sentence and vowel of female anchors in Pilot Study 2

	Sentence	Vowel				
Female	AH 60	1.	2.	3.	4.	5.
Breathy		AH50	AH55	AH 60	AH 65	AH 70
	AH70	1.	2.	3.	4.	5.
		AH60	AH65	AH 70	AH 75	AH 80
	AH75	1.	2.	3.	4.	5.
		AH 65	AH 70	AH 75	AH 80	AH80TL20
	AH80	1.	2.	3.	4.	5.
		AH 70	AH 75	AH 80	AH80tl40	AH80TL60
	AH80TL20	1.	2.	3.	4.	5.
		AH75	AH80	AH80TLl20	AH80TL40	AH80TL60
	AH80TL40	1.	2.	3.	4.	5.
		AH80	AH80TL20	AH80TL40	AH80TL60	AH80TL80
Female	DI 2	1.Prototype	2. DI2	3. DI6	4. DI10	5. DI14
Rough						
	DI 6	1.Prototype	2. DI2	3. DI6	4. DI10	5. DI14
	DI 12	1.DI4	2. DI8	3. DI12	4. DI16	5. Di20
	DI 20	1.DI12	2. DI16	3. DI20	4. DI24	5. DI28
	DI 28	1.DI20	2. DI24	3. DI28	4. DI32	5. DI36
	DI 36	1.DI28	2. DI32	3. DI36	4. DI40	5. DI44

Appendix B

Matching of sentence and vowel of Male anchors in Pilot Study 2

Male	AH 65	1.	2.	3.	4.	5.
Breathy		AH60	AH65	AH70	AH75	AH80
	AH75	1.	2.	3.	4.	5.
		AH65	AH70	AH75	AH80	AH80TL20
	AH78	1.	2.	3.	4.	5.
		AH70	AH75	AH80	AH80TL40	AH80TL60
	AH80TL20	1.	2.	3.	4.	5.
		AH75	AH80	AH80TL20	AH80TL40	AH80TL60
	AH80TL40	1.	2.	3.	4.	5.
		AH80	AH80TL20	AH80TL40	AH80TL60	AH80TL80
	AH80TL60	1.	2.	3.	4.	5.
		AH80	AH80TL20	AH80TL40	AH80TL60	AH80TL80
Male	DI 2	1.prototype	2. DI2	3. DI6	4. DI10	5. DI14
Rough						
	DI 6	1.prototype	2. DI2	3. DI6	4. DI10	5. DI14
	DI 12	1.DI4	2. DI8	3. DI12	4. DI16	5. DI20
	DI 20	1.DI12	2. DI16	3. DI20	4. DI24	5. DI28
	DI 28	1.DI20	2. DI24	3. DI28	4. DI32	5. DI36
	DI 36	1.DI28	2. DI32	3. DI36	4. DI40	5. DI44

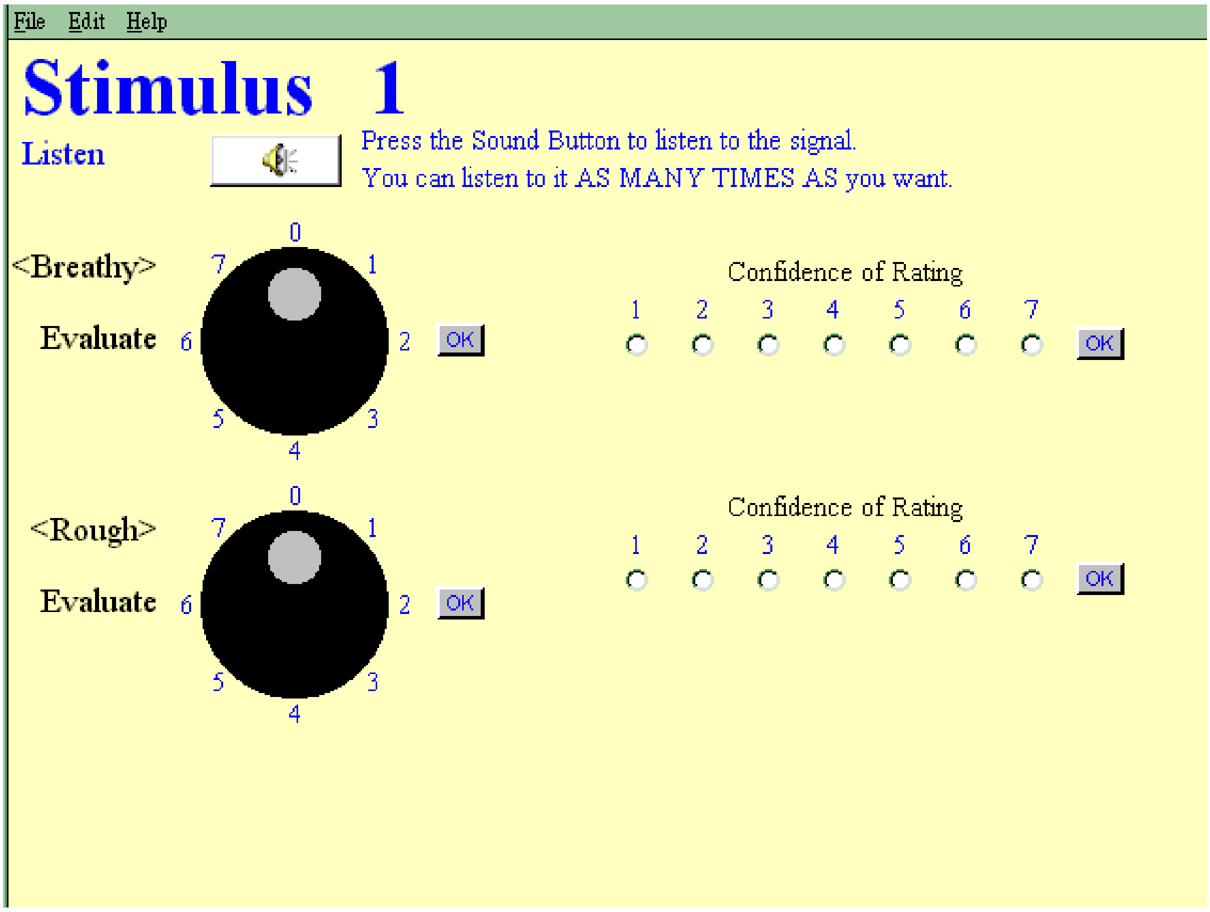
Appendix C

Definitions of breathiness and roughness (Yiu, 2001, p.16)

Voice quality	Perceptual correlates	Physiological correlates
Breathiness	1. Audible sound of expiration 2. Audible sound of 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

Appendix D

Main Study Paradigm



Stimulus 1

Listen

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

<Breathy>

Evaluate

Confidence of Rating

1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ OK

<Rough>

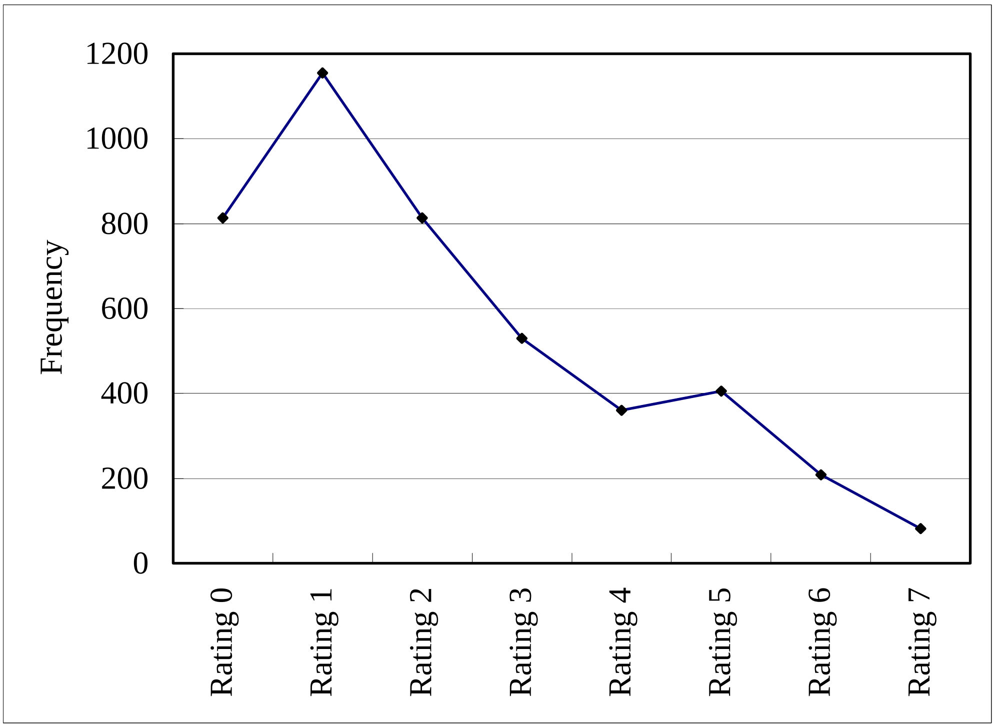
Evaluate

Confidence of Rating

1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ OK

Appendix E

Distribution of frequency of ratings using knob control


Frequency
Rating 0
Rating 1
Rating 2
Rating 3
Rating 4
Rating 5
Rating 6
Rating 7