Title	Development and validation of a comprehensive assessment questionnaire for Cantonese alaryngeal speakers' speech performance
Author(s)	Yip, Chui-yan; 葉翠茵
Citation	
Issued Date	2011
URL	http://hdl.handle.net/10722/192911
Rights	The author retains all proprietary rights, (such as patent rights) and the right to use in future works.

Development and validation of a comprehensive assessment questionnaire for Cantonese alaryngeal speakers' speech performance

Yip Chui Yan

Abstract

The study devised and validated the perceptual assessment questionnaire for evaluating the speech performance of Cantonese alaryngeal speakers. Forty-eight male alaryngeal speakers participated in the study: 10 electrolaryngeal, 10 esophageal, 9 tracheoesophageal, 9 pneumatic artificial and 10 normal laryngeal speakers. Five speech therapists also participated in the perceptual rating procedures. Results indicated moderate to strong inter-rater reliability in all parameters that involve only auditory judgment except that of rating electrolarynx noise. Assessment parameters that require both auditory and visual judgment might require further modification. For tone perception, moderate to strong inter-rater reliability was also noted. High intra-rater reliability of the assessment questionnaire was also found. In addition, the parameters adopted were reported to have significant correlation with the acoustic correlates except that for pitch rating. The assessment questionnaire suggested appeared to be valid for evaluating auditory dependent speech characteristics of the four types of alaryngeal speech.

Introduction

Total laryngectomy is a surgical procedure of removing the entire larynx, sometimes including the hyoid bone, epiglottis, thyroid and cricoid cartilages, and the first 2-3 tracheal rings (Doyle, 2005). The procedure is usually carried out in patients with late stage laryngeal cancer. After the surgery, a permanent opening which connects trachea to the anterior lower neck, known as the tracheostoma, is created for breathing purpose (Balm, 2007). Due to the amputation of laryngeal structures, laryngectomees suffer from total loss of phonation post-operatively. Therefore, learning to adopt an alternative phonation method to speak again is crucial in post-surgical rehabilitation (Doyle, 2005).

In Hong Kong, four types of post-laryngectomy phonation methods are adopted by the laryngectomees, including electrolaryngeal (EL), pneumatic artificial (PA), esophageal (SE) and tracheoesophageal (TE) speech. These alaryngeal phonation methods utilize the same resonating structures of the oral, nasal, and pharyngeal cavities, and various articulators in the vocal tract as normal laryngeal speakers. Nevertheless, they differ from each other by the air supply mechanism and the alternative vibrating sound source. During SE and TE phonation, the pharyngoesophageal (PE) segment (known as the neoglottis) is set into vibration to function as a new sound source (van As-Brooks & Fuller, 2007). In EL and PA phonation, sound is generated by an external device: electrolarynx for EL speech and pneumatic artificial larynx for PA speech (Salmon, 1999). The sound generated by an external sound source is transmitted to the vocal tract for articulation. Regarding the use of air reservoir, EL phonation is independent of air source; PA and TE phonations are driven by pulmonary air, whereas SE phonation is driven by the air expelled from the upper esophagus. It should be noted that the vital capacity of human lungs is about 3,000 c.c. (Seikel, King, & Drumright, 2005), while the upper esophagus only stores up to 40 - 80 c.c. of air (Searl & Reeves, 2007). The difference in the phonatory mechanism renders unique characteristics

associated with different alaryngeal speech, which results in their different perceptual and acoustical characteristics. Previous research reported significant differences in perceptual speech performance of different alaryngeal Cantonese speakers including differences in the aspect of voice quality, amount of noise present and pitch variability (e.g., Ng, Kwok, & Chow, 1997). Differences in the acoustical characteristics among different types of alaryngeal speech were also reported. For example, Ng, Gilbert and Lerman (2001) reported that fundamental frequency (F0) characteristics, F0, vowel duration, and intensity were significantly different among SE, NL and EL speech. Although it has been found that alaryngeal speech was associated with a reduced intelligibility when compared to normal laryngeal (NL) speech (Yiu, van Hasselt, Williams, & Woo, 1994), no significant difference in intelligibility was found among the four types of superior alaryngeal speakers (Ng et al., 1997). In other words, it is assumed that with sufficient and effective training, intelligibility of alaryngeal speech can be regained/retained.

In post-laryngectomy speech rehabilitation, speech therapists provide systematic training to alaryngeal speakers. Different alaryngeal speakers receive training by enrolling in different therapeutic regimens with different focuses and targets. In order for such training to be effective and efficient, attending speech therapists need to determine which aspects hinder the overall speech performance of alaryngeal speech. However, as discussed above, alaryngeal speech phonation mechanism differs from that of laryngeal speech. For instance, the additional involvement of PE segment and the complex aspects of its coordination that is believed by some researchers as an aerodynamic-myoelastic event which make SE and TE speech very unique in nature (Lundstrom, Hammarberg, Munck-Wikland, & Edsborg , 2008). A robust and valid assessment tool which can highlight the specific aspects of alaryngeal speech is needed. Currently, speech therapists only rely on personal subjective perception to judge the speech performance of different alaryngeal speakers. This is apparently unreliable

and not evidence-based. An objective assessment protocol is urgently needed in order to reliably reveal speech performance of Cantonese alaryngeal speakers.

According to recent statistics from the Hong Kong Cancer Society, there are over 200 laryngeal cancer new cases being diagnosed per year in Hong Kong (Hong Kong Cancer Registry, 2007). Currently, there are over 3,000 laryngectomee members in the New Voice Club of Hong Kong, an organization which promotes self-help and mutual-help among laryngectomees in Hong Kong. Despite the large number of laryngectomees, such a comprehensive assessment tool is lacking. Previous studies have investigated different assessment protocols for alaryngeal speech. Schuster et al. (2006) and Maier et al. (2009) suggested an objective assessment to quantify the quality alaryngeal speech by using an automatic speech recognition (ASR) system that could recognize speech by means of acoustic analysis. Significant correlation was found between subjective rating of intelligibility by voice professionals and the data obtained from the ASR system. Yet, noise has been suggested to have an effect on the ASR system. Thus, EL, SE and PA speech that associated with stoma or electrolarynx noise might not yield valid result from such assessment. In addition, participants recruited in the studies were only German TE speakers. Results from the study thus might not be applicable to alaryngeal speakers of a tonal language such as Mandarin and Cantonese, given the marked differences between tonal and non-tonal languages. In a tonal language such as Cantonese, semantic word distinction depends mainly on the F0 contour characteristics of the associated word (Ng, Gilbert, & Lerman, 2001). The same syllable may have different meanings when produced at different lexical tones. This contributes to the unique characteristics in acoustical and perceptual measurements of Cantonese alaryngeal speech.

For Cantonese alaryngeal speech assessment, Wong, Cheung, Yuen, Ho, and Wei (1997) suggested an assessment scale for evaluating TE speech by using both subjective

(5-point scale) and objective parameters. However, this tool failed to address the evaluation of other alaryngeal speech apart from TE speech of the Cantonese population. Therefore, a robust assessment tool that can be used for different kinds of alaryngeal speech of Cantonese is currently unavailable.

The present study attempts to develop a comprehensive assessment questionnaire that can be used to evaluate proficiency of Cantonese alaryngeal speech. It is hypothesized that by establishing and validating specific parameters that can reflect the alaryngeal speech intelligibility and listeners' acceptability (Balm, 2007), speech therapists will be able to evaluate the proficiency of alaryngeal speech and promote rehabilitation efficiency. These parameters are not assumed to have equal importance in constituting intelligible and acceptable speech, as a combination of parameters are expected to have influence on the speech perception. The assessment questionnaire proposed in the current study adopted perceptual ratings. Previous researches have suggested significant inter-rater reliability for the perceptual ratings for TE speech (van As, Hilgers, Verdonck-de Leeuw, & Koopmans-van Beinum, 1998). Also, different perceptual parameters have been adopted for TE and SE speech evaluation (Most, Tobin, & Mimran, 2000; van As-Brooks, Hilgers, Koopmans-van Beinum, & Pols, 2005), PA speech (Xu, Chen, Lu, & Qiao, 2009), and EL speech (Liu, Wan, Wang, & Niu, 2004). As statistically significant results were only found in the specific types of alaryngeal speech and different perceptual parameters (differ in both wordings and the aspects being investigated) were being adopted, ineffective communication among professionals might be resulted. After reviewing the perceptual and acoustic properties of different alaryngeal speech and gathering clinical experience from expert speech therapists, the following aspects of parameters were suggested in the questionnaire (see Appendix A):

A. Voice quality

Variability in voice quality of laryngeal speech is related to the quality of the vibrating

source (Colton, Casper, & Leonard, 2006). In alaryngeal speech, it is assumed that SE and TE speeches' quality, which is related to the PE segment vibration, can be trained. Therefore, this array of parameter was only adopted for SE and TE speeches' evaluation and evaluating the voice quality might reflect the proficiency in control of PE segment.

B. Pitch and Loudness

Research showed that F0 is significantly correlated with intelligibility and acceptability in SE and TE speech (Most et al., 2000). EL speech is characterized by the lack of F0 variation, which is considered as its main acoustic deficit that affects intelligibility (Meltzner & Hillman, 2005; Liu et al., 2006). Therefore, F0 values and F0 variation that are perceived as pitch level and pitch variation respectively affect the intelligibility of speech and served as an important parameter. On the other hand, loudness is determined by the driving air pressure and the medial compression of the vibrating source (Colton, Casper, & Leonard, 2006). As discussed, different alaryngeal speakers utilize different air reservoir; placement of coupling device may also influence the energy transmission. Therefore, assessing loudness reveals the performance of vibratory behavior of the new sound source.

C. Fluency

Rating of speech rate, maximum phonation time (MPT) and syllables per breath group were included. Syllables per breath group and MPT are determined by the air source and adduction of the vibrating apparatus (Colton, Casper, & Leonard, 2006). For inefficient alaryngeal speakers, shorter syllables per breath group may be observed due to the insufficient air source, poor coordination between sound source and articulation, and/or incomplete adduction of PE segment. Inappropriate pauses might contribute to poor intelligibility. In English, an average MPT of 2 - 3 s and 5 - 9 syllables per breath group are necessary for producing useable SE speech (Duguary, 1999). Research showed that more intelligible SE speakers exhibited significantly more syllables per second than those of less

intelligibility (Most et al., 2000). Therefore, fluency served as a parameter in the study.

D. Audible distraction

EL speakers might exhibit radiated noise which is not filtered by the vocal tract during EL speech production. Research revealed that noise is one of the acoustic deficits for EL speech (Liu & Ng, 2007). PA, SE and TE speakers might also generate stoma noise that negatively affects intelligibility (Globek, Stajner-Katusic, Musura, Horga, & Liker, 2004). Therefore, degree of audible noise served as a parameter for assessing speech performance.

E. Articulation

Both articulation and tone production proficiency were included for assessment. Similar to laryngeal speech, imprecise articulation influences alaryngeal speech intelligibility. With the assumption of source-filter theory, the articulatory and laryngeal systems are independent. Amputation of laryngeal structure should have no impact on the articulatory (filter) system (Kent & Read, 1992). Therefore, imprecision of articulation might suggest inefficient alaryngeal speech. For instance, Salmon (1999), and Searl and Reeves (2007) reported that EL and PA speakers show difficulty in producing frication and aspiration contrasts due to the lack of intraoral pressure; SE and TE speakers might show denasalization with no manner of articulation errors (Searl, Carpenter, & Banta, 2001). Articulation proficiency thus differs in different alaryngeal speech and needed to be evaluated.

Cantonese is a tonal language. Cantonese speakers use six lexical tones to convey different semantic meanings. Therefore, proficiency in tone production can help determine speaker’s effectiveness in transmitting linguistic information. Tonal contrast depends on F0 contour which is dependent on the source and vibration segments (Ng et al., 2001). Research showed that Cantonese SE, TE and PA speakers can convey lexical tone information while EL speakers did not reach significant level for total informational transmission (Yiu et al., 1994). It should thus be expected that, except for EL speech, efficient alaryngeal speakers are

able to convey tonal contrast.

F. Coordination and placement of coupling device

Previous clinical experience in alaryngeal rehabilitation reveals that different alaryngeal speech requires different coordination for effective phonation. However, subtle coordination parameters and the most efficient placement of the coupling devices require more than acoustic signals to assist perceptual judgment. For instance, in EL and PA speech, a "sweet spot" must be located by the coupling device so that radiation can be transmitted most efficiently (Wong, 2009; Kwok, 2009; Searl & Reeves, 2007; Salmon, 1999; van As-Brooks & Fuller, 2007; Balm, 2007; Doyle, 2005). Therefore, video recording, providing both visual and audio input was adopted for this study. The validity of such assessment will be evaluated.

Method

Speakers

Forty-eight male alaryngeal speakers, including 10 EL (Servox type), 10 SE, 9 TE and 9 PA speakers and 10 normal laryngeal (NL) speakers participated in the study. All participants were native Cantonese speakers with ages ranging from 41-81years (mean = 62.8 years). They had no known history of speech/voice problems, except that associated with laryngectomy for alaryngeal speakers. Alaryngeal participants were reported with history of total laryngectomy and were at least one year post-operation. Their average experience in using alaryngeal speech is summarized in Table 1. To ensure that the speakers were utilizing their most efficient phonation, all selected speakers had received primitive speech therapy for the alaryngeal speech being used; SE and TE speakers were suggested not to have food intake immediately before assessment as swallowing might affect the neoglottis vibration (van As-Brooks & Fuller, 2007). Participants were randomly selected from the New Voice Club of Hong Kong,

Table 1.

The participants' average experience (year) in using alaryngeal speech

	Types of speech			
	TE	SE	PA	EL
Experience (years)	8	12	11	7

Rater

Five practicing speech therapists with at least one year work experience were invited to participate in the rating experiment. They were native Cantonese speakers and had no known history of hearing problem.

Procedures

The study was divided into two stages. The first stage aimed at increasing the content validity of the assessment tool by gathering expert opinion on the preliminary assessment questionnaire. Two practicing speech therapists who were experienced in alaryngeal speech rehabilitation were invited to evaluate the questionnaire in the following areas: (1) The choice of parameters for clinical assessment, treatment planning and measuring change in intervention. (2) The feasibility and comprehensibility of the questionnaire. The proposed questionnaire was amended according to the comments collected. The second stage of the study targeted to determine the convergent validity and reliability of the revised questionnaire. This validation stage included recording and rating procedures which are described below.

Speech materials

Four types of speech materials were included: (1) a short passage, (2) monosyllables, (3) vowel /a/ prolongation, and (4) loudness glide. All speech materials were printed on a card for easy reading. The 134-word standard Chinese passage "The North Wind and the Sun" (see Appendix B) was adopted for continuous speech production task. In addition, CV monosyllabic words of /ji/ and /si/ produced at six contrastive tones were presented in randomized order to avoid order effect. (see Appendix C)To increase naturalness

of productions, each word was embedded in a carrier phrase /ŋɔ23 jiu33 tʊk3_/, meaning “I want to read __”. The carrier phrase was constructed so that the target monosyllable was preceded and followed by a voiceless plosive which allowed easy identification during acoustic analysis.

Recording and listening procedures

The recording procedure took place in a quiet room at the New Voice Club of Hong Kong. Acoustic signals were collected using a high quality microphone (SM58, Shure) via a preamplification unit (PreMobile USB, M-Audio). Signals were digitized at a sampling rate of 20 kHz and a quantization rate of 16 bits/sample by using PRAAT. To measure the actual loudness, calibration was carried out prior to recording. Three calibration (pure tone) signals of 60 dB SPL, 70 dB SPL and 80 dB SPL at 1,000 Hz were generated and recorded using the same instrumental setup. The signals were used to generate a regression equation using which the actual intensity levels were calculated. To control loudness, a constant microphone-to-mouth distance of 10 cm was maintained during throughout the recording.

The speakers were instructed to read the speech materials once in the following order: (1) passage, (2) vowel /a/ prolongation, (3) monosyllables, and (4) loudness glide. They were also instructed to read the passage with comfortable pitch and loudness level. Sufficient practice time was allowed before recording so as to familiarize themselves with the reading materials and the environment. Voice and video recordings were obtained simultaneously during the recording procedure.

To carry out the perceptual experiment, tentative notes were provided to ensure the raters' understanding on the proposed parameters and brief introduction of the mechanism of the different alaryngeal speech was also provided to familiarize them with alaryngeal speech. The rating task was a take-home exercise and the recorded speech samples were randomized according to groups and presented to the listeners via speakers and a video monitor.

Table 2.

Types of tasks and the corresponding parameters being rated in perceptual rating experiment.

		Types of speech/ video samples				
Speech tasks	**Perceptual parameters**	**NL**	**SE**	**TE**	**PA**	**EL**
Vowel /a/ prolongation	Hoarseness	√	√	√	-	-
	Breathiness	√	√	√	-	-
	Effortfulness	√	√	√	-	-
	Phonation breaks	√	√	√	-	-
	MPT	√	√	√	√	-
Passage	Average Pitch	√	√	√	√	√
	Pitch variation	√	√	√	√	-
	Loudness	√	√	√	√	√
	Speech rate	√	√	√	√	√
	Phrase length	√	√	√	√	√
	Stoma noise	√	√	√	√	√
	Electrolarynx noise	-	-	-	-	√
	Articulation Proficiency	√	√	√	√	√
	Intelligibility	√	√	√	√	√
Loudness glide	Loudness variation	√	√	√	√	√
Video	Coordination of placement of coupling device with articulation	-	-	-	√	√
	Coupling device Placement accuracy and consistency	-	-	-	√	√
	Coordination of stoma occlusion and phonation	-	-	√	√	-
	Coordination of respiration and phonation	-	√	√	√	-

Throughout the rating, loudness control was kept constant to maintain the actual loudness difference between speakers. Based on the speech samples, they were instructed to rate the performance by means of a questionnaire containing items in 7 equal-interval Likert scales. Different arrays of parameters corresponding to the types of speech or video samples presented were included in the questionnaire for rating. The parameters used for rating

different speech samples are listed in Table 2. For tone perception, CV monosyllables were isolated from the carrier phrase and presented to the listeners. To eliminate order effect, the presentation order was randomized. The raters were instructed to identify corresponding words upon listening to the samples. In addition, the raters were advised to take rest between blocks to avoid fatigue effect.

Acoustic analysis

Three types of speech samples were analyzed acoustically using PRAAT. For all speech samples, the acoustic pulses within pitch range of 75 – 300 Hz and intensity range of 50 - 150 dB were analyzed. Only voiced portions were selected for analysis with all the natural and unnatural pauses eliminated. For MPT calculation, mean MPT was obtained by averaging the MPT values of the three trials. The second trial of vowel prolongation was selected for acoustic analysis.

Results

In order to validate the assessment questionnaire, validity and reliability were evaluated. The average perceptual ratings on different parameters were presented in Figure 1-3 (see Appendix D) and the acoustic data of the different speaker groups are presented in Table 3.

Table 3

Various acoustical measurement associated with different speaker groups

	Mean (SD)				
	Intensity (dB)	Average F0 (Hz)	F0 variation	Jitter (RAP)(%)	MPT (s)
NL	85.3 (2.7)	121.9 (12.5)	25.9	0.2	18.3
SE	79.0 (3.4)	163.6 (35.2)	55.8	3.1	2.0
TE	84.5 (4.1)	116.9 (21.1)	25.6	1.9	4.2
PA	82.2 (7.7)	133.4 (17.6)	15.3	NA	7.4
EL	83.9 (6.1)	93.0 (30.8)	11.0	NA	NA

Inter-rater reliability of the speech parameters

Intra-class correlation coefficient (ICC) with 95% confidence interval (CI) was calculated for each parameter to assess the inter-rater reliability for the five raters. The variability of different rater's ratings and the total variance across all ratings were compared. In the present study, a two-way random effect ICC model was adopted based on the assumption that the raters were selected from the population. In addition, ICCs for absolute agreement were adopted with the rater variance taken into account. Therefore, the reliability reported is generalizable to the speech therapist community and reflects the flexibility of raters. The ICCs for the speech parameters between the five raters are shown in Table 4.The interpretation of ICC value followed the criteria of : 0-0.2 indicates poor agreement; 0.3-0.4 reflects fair agreement ; 0.5-0.6 indicates moderate agreement; 0.7-0.8 indicates strong agreement; and >0.8 indicates almost perfect agreement ("Intraclass correlation for parametric data introduction and explanation," n.d.).

Taking all parameters into account, the results showed overall moderate inter-rater reliability (66%). Strong agreements was found between raters for rating: MPT (73.2%), hoarseness (76.1%), breathiness (76.4%), phrase length (87.6%) and intelligibility (76.4%). Moderate reliability for rating: effortfulness (69.6%), phonation breaks (59.3%), pitch (63.6%), pitch variation (69.3%), loudness (60.2), loudness variation (57.5%), speech rate (50.8%), stoma noise (52.9%), articulation proficiency (61.0%), tone production proficiency (61.2%) and coordination between respiration and phonation (51.4%). The degree of reliability between raters was fair for rating coordination of coupling device placement with articulation (31.8%) and coupling device placement accuracy (29.3%). However, poor agreement between raters was noted for rating electrolarynx noise (15.7%), coordination of stoma occlusion and phonation (20%), and coupling device placement consistency (9.5%).

Table 4.

ICC Percentages and 95% CI of each speech parameters for the five raters

Speech Parameters	Measures obtained from the Five Raters ICC% (95% CI)	Cronbach's Alpha
Voice quality		
MPT	73.2*	0.947
Hoarseness	76.1*	0.946
Breathiness	76.4*	0.948
Effortful	69.6*	0.926
Phonation Breaks	59.3*	0.913
Pitch and Loudness		
Pitch	63.6*	0.920
Pitch Variation	69.3*	0.937
Loudness	60.2*	0.907
Loudness Variation	57.5*	0.891
Fluency		
Speech Rate	50.8*	0.885
Phrase Length	87.6*	0.905
Audible Distraction		
Stoma Noise	52.9*	0.879
Electrolarynx noise	15.7	0.604
Articulation		
Articulation Proficiency	61.0*	0.913
Tone Production Proficiency	61.2*	0.923
Coordination		
Coordination of coupling device placement with articulation	31.8*	0.774
Coordination of stoma occlusion and phonation	20*	0.699
Coordination of respiration and phonation	51.4*	0.854
Placement of coupling device		
Coupling device placement accuracy	29.3*	0.779
Coupling device placement consistency	9.5	0.591
Overall Proficiency		
Intelligibility	76.4*	0.952
Overall (All parameters)	**66***	**0.907**

*Statistically significant ICC was noted ($p < 0.01$)

In these poorly agreed parameters, ratings of coordination of stoma occlusion and phonation in TE speakers varied more (SD = 1.069) between raters than that of PA speech (SD = 0.892); while similar standard deviation values were noticed for rating placement accuracy of coupling device between PA (SD = 1.452) and EL speakers (SD = 1.601).

Inter-rater reliability of tone perception

The raters were asked to identify the tones associated with the monosyllables /si/ and /ji/ produced at the six Cantonese tones. Among them, 64% of the tone identification was randomly selected for analysis. The tone of the Cantonese monosyllables regardless of the initial consonant was identified for statistical analysis. Inter-rater reliability of tone identification between raters was assessed by using ICCs and the results are shown in the Table 5. ICC results indicated strong reliability (72.6%) between five raters for identifying alaryngeal speech samples. Agreement of different types of speakers was also evaluated and results are listed in Table 6. The greatest agreement was found in raters judging the tones of NL speech samples (82.7%), while the lowest was found in perceiving EL tones (54.5%).

Intra-rater reliability

The reliability of the raters was evaluated using the Pearson product-moment correlation test. To assess the consistency and reliability of ratings, 20% of samples were judged twice by the raters. Percent agreement of the first and second ratings was calculated to indicate the average intra-rater reliability. The results are listed in Table 7. Significant and high positive correlation of $r = 0.884$ ($p < 0.01$) was found.

Table 5.

ICC Percentages and 95% CI of tone perception for the five raters

	Measures obtained from the five Raters	
Tone	**ICC% (95% CI)**	**Cronbach's Alpha**
Overall (all tones)	72.6	0.742

Statistically significant ICC was noted ($p < 0.01$) in the listed data

Table 6.

Agreement of tone perception between the five speech therapist raters

Types of speech	**Percentage of agreement (%) among five raters**
NL	82.7
SE	61.2
TE	66.6
PA	63.9
EL	54.5
Overall (all tones)	**65.8**

Table 7.

Pearson product-moment correlation of intra-rater ratings across five raters

	Raters					
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	All Raters
r	0.868	0.910	0.900	0.858	0.903	0.884

Significant correlation with $p < 0.01$ was noted in all the result listed.

Table 8.

Pearson product-moment correlation coefficient values indicating correlation between average acoustic value and the corresponding perceptual ratings.

Speaker Group	**Pearson product-moment correlation coefficient (p value)**				
	Average intensity and loudness rating	Average F0 and pitch rating	Frequency variation and pitch variation rating	Jitter and Hoarseness rating	MPT and maximum phonation rating
NL	0.390 ($p = 0.265$)	0.287 ($p = 0.421$)	0.540 ($p = 0.107$)	0.429 ($p = 0.216$)	0.898* ($p < 0.01$)
SE	0.789* ($p = 0.007$)	-0.192 ($p = 0.595$)	-0.726 ($p = 0.017$)	-0.489 ($p = 0.151$)	0.878* ($p = 0.001$)
TE	0.538 ($p = 0.135$)	-0.538 ($p = 0.135$)	-0.375 ($p = 0.320$)	-0.612 ($p = 0.080$)	0.777 ($p = 0.014$)
PA	0.935* ($p < 0.01$)	0.278 ($p = 0.469$)	0.091 ($p = 0.815$)	-	0.891* ($p = 0.001$)
EL	0.835* ($p = 0.003$)	0.018 (p= 0.960)	-	-	-
ALL	**0.640* ($p < 0.01$)**	**0.016 ($p = 0.916$)**	**-0.302 ($p = 0.065$)**	**-0.679* ($p < 0.01$)**	**0.879* ($p < 0.01$)**

*Statistically significant correlation was noted ($p < 0.01$)

Convergent validity

To examine validity of the parameters included in the questionnaire, convergent validity was evaluated by calculating the agreement between acoustical (objective measures) and corresponding perceptual parameters. Pearson product-moment correlation was used to determine the convergent validity and the result is shown in Table 8. Significant and moderate to high correlation was found in all of the objective and perceptual measurement pairs except for pitch rating and F0 ($r = 0.016$, $p = 0.916$); pitch variation rating and frequency variation ($r = -0.302$, $p = 0.065$). These correlations were not significant among all speaker groups. In addition, in NL group, no significant correlation was found in all the perceptual and acoustical pairs except that in MPT .

Discussion

A robust and reliable assessment tool is essential for objectively evaluating speech performance for Hong Kong laryngectomees and designing intervention regimen in alaryngeal speech rehabilitation. The present study attempted to develop such assessment tool by evaluating perceptual ratings rated by five practicing speech therapists through statistical validation. It is hypothesized that with high convergent, inter and intra-rater reliability on the suggested parameters, the questionnaire can be validated and adopted for evaluating male Cantonese alaryngeal population.

Inter-rater reliability of speech parameters rating

Inter-rater reliability reflects the degree of agreement of speech therapists' perception on the related parameters. The present results indicated that our speech therapists had a moderate-to-strong agreement in ratings of overall proficiency, aspects of voice quality, pitch, loudness, and fluency, fair-to-poor agreement for rating coordination and placement of coupling devices and for rating electrolarynx noise. Speech parameters judged based only on acoustics appeared to have a better inter-rater reliability. For the parameters involving

coordination and placement of coupling device both visually and audibly, fair-to-poor inter-rater reliability was noted.

Generally, no perfect agreement between raters was found. This might be attributed to a number of error sources. First, questionnaire of ordinal rating scale was adopted in the study. Sources of errors, which generally influence ordinal ratings of pathological voices, were documented. According to Kreiman and Gerratt (2000), it is difficult to isolate single dimension from complex stimuli. For instance, incompetent coordination and placement accuracy of coupling devices together contribute to the poor speech performance. Low inter-rater reliability might be ascribed to the difficulty in isolating single dimension of the speech characteristics. Secondly, identification of the source of speech disturbance might require more than acoustic cues. Video recording which provided less visual spatial information than real live assessment might not have provided sufficient information for accurate perceptual rating. Both auditory and visual perceptual ratings are crucial as intelligibility does not solely depend on acoustical parameters and might be compromised by articulatory adjustment and contextual information. In addition, rating of subtle adjustments, including the coordination and the placement of coupling devices, depended mainly on visual judgment might be affected. Thus, environmental factor might also have contributed to the low reliability (Kreiman & Gerratt, 2000) Thirdly, rater experience (with or without alaryngeal speech rehabilitation experience), rating environment and rater methods (Kreiman & Gerratt, 2000) also contributed to the source of inconsistency in rating between raters.

Further examination of the poor inter-rater reliability in judging coordination of stoma occlusion and phonation also revealed that the variation occurred more in TE speech than PA speech. TE and PA speech require occlusion by finger and pneumatic device, respectively. Digital occlusion involves fine motor control which might be more subtle to be identified via video recording. This conformed to the previous hypothesis on the possible error induced by

video assessment. On the other hand, rating of EL noise is also reported to have low inter-rater reliability. EL noise is the radiated buzzing noise leaked out from the poorly coupled device that is not filtered by vocal tract. Increased amplitude of EL noise might be influenced by the pressure against the neck and placement accuracy of the device (van As-Brooks & Fuller, 2007).Therefore, rating of EL noise required both visual and auditory perception which might induce more variance in perceptual rating. Furthermore, EL noise is inevitable when an external vibrating source is utilized. Clinical experience might be important in determining the acceptability degree of EL noise. Poor inter-rater reliability might therefore be attributed to the difference in relevant clinical experience.

Inter-rater reliability of tone perception

For tone perception, inter-rater reliability analysis indicated strong agreement in ratings provided by different raters, implying that speech therapists were reliable in rating tones produced by alaryngeal speakers. Tone production proficiency of different alaryngeal speakers has been reported in the literatures. Yiu et al.(1994) suggested SE, TE and PA speakers were able to convey lexical tone information while TE speakers were more proficient in tone production. Ching, Williams, and van Hasselt (1994), however, reported that SE and PA speech were significantly more proficient than TE speech. Ng, Gilbert, and Lerman (2001) suggested SE speakers exhibited F0 contours that were similar to NL speakers, and SE tones were perceptually similar to NL tones. Although it seems difficult to compare tone proficiency between TE, SE and PA speakers as it might be hindered by other speech parameters, the literature unanimously reported that EL speech lacked pitch and F0 contour variation and thus led to inefficient tone production. As tone is a very important part of Cantonese alaryngeal speech performance, the current study investigated the inter-rater agreement in judging tone production by different alaryngeal speakers. Intuitively, more proficient tone production should result in higher inter-rater agreement. The present results

conformed to the findings reported previously that perceptually SE, TE and PA speakers were more superior in producing tones and they yielded high inter-rater agreement than EL speech, while TE speech was slightly more proficient than PA which in turns better than SE speech. None of the alaryngeal speech reached the level of inter-rater agreement as that in NL speech. This contradicted to the previous literature (Ng, Gilbert, & Lerman, 2001). The discrepancy of not reaching full potential of tone production might be due to the ageing of PE segment which lowers the efficiency of vibration and adduction and thus F0 contour adjustment. Participants' experience of using SE and TE speech in this study was reported in Table 1, 8 - 12 years of using alaryngeal speech has been reported. Aging of PE segment can possibly contribute to the reduced proficiency in tone production as compared with NL speech. Further research on the ageing effect on tone production might be needed.

The lowest inter-rater agreement (54.5%) was noted for EL speech. It was noted that within EL speakers, different tones were consistently perceived as the same level tones(i.e., the high level, mid level and low level tones). In addition, different raters perceived the tones produced by the same speaker differently (i.e., with a low consistency). All of the above might have contributed to the low inter-rater agreement. This finding on EL speech conforms to that reported by Ng, Gilbert, and Lerman (2001). EL speech lacked pitch variation and thus hindered correct tone perception as only level tones were identified. Further study on how F0 variation affects tone perception should be conducted which may provide implications on how electrolarynx design can help EL speakers produce different tones.

Intra-rater reliability

Strong intra-rater reliability was found. In addition, the intra-rater reliability($r = 0.884$) was considerably stronger than that of inter-rater reliability ($r = 0.66$), suggesting that the rating scale and the parameters adopted in the questionnaire allowed little variance in rating within raters. Same rater should be responsible to rate all sessions in order to minimize

measurement error when implementing the assessment.

Convergent validity

Convergent validity was evaluated to determine if the items established truly reflected the targeted aspects of speech. Statistically significant and moderate-to-strong correlation was found between the three pairs of perceptual and acoustic correlates. Overall moderate correlation was reported between average intensity and the average loudness level. Loudness is related to the air pressure variation during speech production and it closely corresponds to intensity measure (defined as the average energy per unit time per unit area) (Placka & Carlyon, 1995). Given the parameters proposed in the questionnaire, the perceptual rating on loudness appeared to be valid. Although an overall moderate correlation was reported, no significant correlation was found for TE speech. Lundstrom et al., (2008) recently examined the PE segment performance and their results suggested that a change of vocal intensity was related to the increased subglottic pressure and the lateral surface of the PE segment. The unique feature in utilizing PE segment in TE speech might contribute to this insignificant correlation.

Convergent validity of the parameters, hoarseness and MPT were evaluated. Jitter, shimmer and noise-to-harmonic ratio values have been reported to contribute to hoarseness perception (Gorham-Rowan & Laures-Gore, 2006). Perceived severity of hoarseness has been found to significantly correlate with jitter measurement (Jones, Trabold, Plante, Cheetham, & Earis, 2000) and with jitter relative average perturbation (RAP) (van As et al., 1998). Jitter (RAP) was thus included in the present study for correlating with hoarseness quality in alaryngeal speech. Hoarseness is mainly caused by the aperiodic vibration or incomplete adduction of the vibrating source, which is the PE segment in TE and SE speech. The mechanical vibratory source associated with EL and PA speech generates a periodic sound, and thus hoarseness is minimal. Results revealed a significant negative correlation

between jitter and hoarseness rating. On the other hand, appropriateness of perceiving a single trial of maximum phonation time was significantly correlated with the actual MPT (in second) measured from averaging three trials as analyzed in PRAAT. Such results imply a high convergent validity for the perceptual parameter of hoarseness rating and perceptual rating on the appropriateness of MPT in a single trial. Rating of MPT also suggested the validity of judging the phonation onset and offset of the alaryngeal speech.

As discussed above, fundamental frequency (F0) as an acoustic correlate of pitch was adopted in the assessment questionnaire. Validity of average pitch and pitch variation rating were investigated and results showed insignificant correlation between both parameters. In fact, similar finding (low correlation between pitch and F0) has been reported in previous research. In TE speech, it was reported that no significant correlation was found between F0 and perceptual evaluation of pitch and voice quality, and it has been hypothesized that the presence of hoarseness might affect pitch perception (van As et al., 1998). In the current study, considerable amount of perceived hoarseness was reported. For SE and TE speech, average perceptual ratings of 3.1 and 2.5 respectively were reported. It follows that the hoarseness perceived in SE and TE voices may also play a part in the pitch rating. In addition to hoarseness, for SE and TE speech, significant correlation between high F0 and the perception of hyperfunction voice quality and breathiness were accounted (Lundstrom et al., 2008). Similarly, breathiness rating scores of 4 and 3 in SE and TE speech; effortfulness ratings of 2.6 and 3.8 in SE and TE speech were reported in the current study. Other perceptual qualities that co-existed with pitch variation might affect the pitch perception and thus contributed to the insignificant correlation as revealed. Furthermore, unlike laryngeal phonation, SE and TE speech make use of the PE segment as the vibrating source. Correlation between pitch and F0 may be different from laryngeal speech. Lundstrom et al. (2008) investigated the relationship between physiological data of PE segment, acoustic and perceptual measurement of SE and

TE speech. They reported that when pitch changes, the position of PE segment also varies. Instead of changing the vocal fold tension and length for frequency or pitch modification in laryngeal phonation (Seikel et al., 2005), higher F0 is correlated to a higher PE segment at cervical region (approximately at the level of the fourth and/or fifth cervical vertebra, C4/5), yielding a shortened vocal tract. In addition, the adjustments of PE segment for phonation are not as consistent as that of vocal folds; a large individual difference has been observed (Lundstrom et al., 2008). Moreover, in the present study, SE and TE speakers had SE/TE speech experience of 8 –12 years. Individual differences such as ageing effect of PE segment might have contributed to the insignificant correlation between perceived pitch and measured F0.

In fact, for NL speech, acoustical findings and perceptual ratings were not significantly correlated. This was contradictory to other findings reported previously (e.g., Kent & Read, 1992). In this study, speech therapists were instructed to rate laryngeal and alaryngeal speech at the same time. Raters might have made relative judgment between laryngeal and alaryngeal voice samples, which led to high rating on the NL speech and thus the insignificant correlation.

Although the correlations found between some of the acoustical and perceptual measurements were insignificant, great perceptual differences between alaryngeal and control speakers were noted. This suggests that perceptual parameters can be of interest for clinical assessment. As there are insignificant validity and reliability concluded in some of the parameters, those parameters may need further revisions. Further studies will be warranted before revising these parameters in the proposed assessment questionnaire.

Possible modifications

Three issues were of concern in this project. First, as only male speakers were recruited in the study, results may be biased. Further investigation on the female population

and normative data will be needed before the tool can be conclusive. Five raters were recruited in this study and their experience in alaryngeal rehabilitation varied. As clinical experience is important in the validity and reliability of rating the speech parameters suggested, this variable might contribute negatively to the findings. Second, concerning the rating procedure, as the rating was conducted in the speech therapists' workplace instead of a sound-proof room, background noise might have imposed a negative effect on loudness rating. In addition, lengthy (4 - 6 hours) rating sessions were required during the rating procedure. Although breaks were suggested between blocks, possible fatigue effect might have affected listeners' perceptual judgment. Third, concerning the reading stimuli, all speakers were asked to read a standard passage "the North Wind and the Sun". Upon listening to a known passage, presence of contextual cues might induce the possibility of over-estimating the rating of intelligibility.

Conclusion

The current study is considered as an initial attempt to develop and validate an assessment questionnaire for evaluating the speech performance of Cantonese male alaryngeal speakers. To conclude, overall moderate inter-rater reliability and high significant correlation in intra-rater reliability were shown in rating the parameters adopted in the preliminary assessment questionnaire. Moderate-to-strong inter-rater reliability was found in rating the aspects of voice quality, pitch and loudness, fluency, articulation and the overall proficiency, while fair-to-poor reliability was revealed in aspects of rating audible distraction, coordination and placement of coupling devices. Concerning the rating of tone production, strong inter-rater reliability was found. Finally, moderate-to-high convergent validity was found in the perceptual parameters and the corresponding acoustic pairs except in the pairs involving F0. The insignificant validity, however, might be contributed by influence of other

perceptual qualities. Future modifications on validating the perceptual judgment on parameters which require more than acoustical information are warranted. As no comprehensive assessment tool has been developed for Cantonese alaryngeal speech assessment, validation of the questionnaire is a major step forward to assist the assessment and thus treatment efficacy for this patient population.

References

Balm, A. J. M. (2007). Laryngeal and hypopharyngeal cancer: intervention approaches. In E. C. Ward & C. J. As-Brooks (Eds.), *Head and neck cancer: treatment, rehabilitation, and outcomes* (pp. 123-137). San Diego, CA: Plural Publishing Group.

Ching, T. Y. C., & Williams, R. (1994). Communication of lexical tones in Cantonese alaryngeal speech. *Journal of Speech and Hearing Research, 37*, 557-563.

Colton, R. H., Casper, J. K., & Leonard, R. (2006). *Understanding voice problems: A physiological perspective for diagnosis and treatment* (6th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Doyle, P. C. (2005). Clinical procedures for training use of the electronic artificial larynx. In P.C. Doyle & R.L. Keith (Eds.), *Contemporary considerations in the treatment and rehabilitation of head and neck cancer: voice, speech, and swallowing* (pp. 545-570). Austin, TX: Pro-Ed.

Duguary, M. J. (1999). Esophageal speech training: The initial phase. In S. J. Salmon (Ed.). *Alaryngeal speech rehabilitation: For clinicians by clinicians* (pp. 165-201). Austin, TX: Pro-Ed.

Globek, D., Stajner-Katusic, S., Musura, M., Horga, D., & Liker, M. (2004). Comparison of alaryngeal voice and speech. *Logopedics Phoniatrics Vocology, 29,* 87-91.

Gorham-Rowan, M. M., & Laures-Gore, J. (2006) Acoustic-perceptual correlates of voice quality in elderly men and women. *Journal of Communication Disorders, 39*, 171–184.

Hospital Authority: Hong Kong Cancer Registry (2007). Retrieved October, 2010, from www3.ha.org.hk/cancereg/e_stat.asp

Jones, T. M., Trabold, M., Plante, F., Cheetham, B. M. G., & Earis, J. E. (2000). Objective assessment of hoarseness by measuring jitter. *Otorhinolaryngology, 26*, 29-32.

Kent, R. D., & Read, C. (1992). *The Acoustic Analysis of Speech.* San Diego, CA: Singular Publishing Group.

Kreiman, J., & Gerratt, B. R., (2000). Sources of listener disagreement in voice quality assessment. *Journal of the Acoustical Society of America, 108*, 1867-1876.

Kwok, I. C. L. (2009, November). Post-laryngectomy speech rehabilitation [PowerPoint Slides]. Paper presented at The Division of Speech & Hearing Sciences, The University of Hong Kong, Hong Kong.

Liu, H., & Ng, M. (2007). Electrolarynx in voice rehabilitation. *Auris Nasus Larynx, 34*, 327-332.

Liu. H., Wan. M., Wang, S., & Niu, H. (2004). Aerodynamic characteristics of laryngectomees breathing quietly and speaking with the electrolarynx. *Journal of Voice, 18*, 567-577.

Liu, H., Wan, M., Ng, M. L., Wang, S., & Lu, C. (2006). Tonal perceptions in normal laryngeal, esophageal, and electrolaryngeal speech of mandarin. *Folia Phoniatrica et Logopaedica, 58*, 340-352.

Lundstrom, E., Hammarberg, U., Munck-Wikland, E., & Edsborg, N. (2008). The pharyngoesophageal segment in laryngectomees - videoradiographic, acoustic, and voice quality perceptual data. *Logopedics Phoniatrics Vocology, 33*, 115-125.

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., & Noth, E. (2009). PEAKS - A system for the automatic evaluation of voice and speech disorders. *Speech Communication, 51*, 425-437.

Meltzner, G. S., & Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language, and Hearing Research, 48,* 766-779.

Most, T., Tobin, Y., & Mimran, R. C. (2000). Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *Journal of Communication Disorders, 33,* 165-181.

Ng, M. L., Gilbert, H. R., & Lerman, J. W. (2001). Fundamental frequency, intensity, and vowel duration characteristics related to Cantonese alaryngeal speech. *Folia Phoniatrica et Logopaedica, 53,* 36-47.

Ng, M. L., Kwok, C. L. I., & Chow S. F. W. (1997). Speech performance of adult Cantonese-speaking laryngectomees using different types of alaryngeal phonation. *Journal of Voice, 11,* 338-344.

Placka, C. J., & Carlyon, R. P. (1995). Loudness perception and intensity coding. In B. C. J. Moore (Ed). *Hearing* (pp. 123-160). San Diego, CA: Academic Press.

Salmon, S. J. (1999). Artificial larynx devices and their use. In S.J. Salmon (Ed.*). Alaryngeal speech rehabilitation for clinicians by clinicians* (pp. 79-104). Austin, TX: Pro-Ed.

Schuster, M., Haderlein, T., Noth, E., Lohscheller, J., Eysholdt, U., & Rosanowski, F. (2006). Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *European Archives of Oto-Rhino-Laryngology, 263,* 188-193.

Searl, J. F., Carpenter, M. A., & Banta, C. L. (2001). Intelligibility of stops and fricatives in tracheoesophageal speech. *Journal of Communication Disorders, 34,* 305-321.

Searl, J. P., & Reeves, I. S. (2007). Nonsurgical voice restoration following total laryngectomy. In E. C. Ward & C. J. van As-Brooks (Eds.), *Head and neck cancer: treatment, rehabilitation, and outcomes* (pp. 123-137). San Diego, CA: Plural Publishing Group.

Seikel, J. A., King, D. W., & Drumright, D. G. (2005). *Anatomy and physiology for speech, language, and hearing*.Australia: Thomson Delmar Learning.

van As, C. J. V., Hilgers, F. J. M., Verdonck-de Leeuw, I. M., Koopmans-van Beinum, F. J. (1998). Acoustic analysis and perceptual evaluation of tracheoesophageal prosthetic voice. *Journal of Voice, 12*, 239-248.

van As-Brooks, C. J. V., Hilgers, F. J. M., Koopmans-van Beinum, F. J., & Pols, C. W. L. (2005). Anatomical and functional correlates of voice quality in tracheoesophageal speech. *Journal of Voice, 19*, 360-372.

Van As-Brooks, C. J. V., & Fuller, D. P. (2007). Prosthetic tracheoesophageal voice restoration following total laryngectomy. In E. C. Ward & C. J. van As-Brooks (Eds.), *Head and neck cancer: treatment, rehabilitation, and outcomes* (pp. 123-137). San Diego, CA: Plural Publishing Group.

Wong, K. R. (2009, March). 復聲導師講座 [PowerPoint slides]. Paper presented at The New Voice Club of Hong Kong, Hong Kong.

Wong, S. H. W., Cheung, C. C. H., Yuen, A. P. W., Ho, W. K., & Wei, I. (1997). Assessment of tracheoesophageal speech in a tonal language. *Archives of Otolaryngology - Head &Neck Surgery, 123*, 88-92.

Xu, J. J., Chen, X., Lu, M. P., & Qiao, M.Z. (2009). Perceptual evaluation and acoustic analysis of pneumatic artificial larynx. *Otolaryngology - Head and Neck Surgery, 141*, 776-780.

Yiu, E. M., van Hasselt, C. A., Williams, S. R., & Woo, J. K. S. (1994). Speech intelligibility in tone language (Chinese) laryngectomy speakers. *European Journal of Disorders of Communication, 29*, 339-347.

Intraclass correlation for parametric data introduction and explanation. (n.d.). Retrieved March 30, 2011, from http://amchang.net/StatTools/ICC_Exp.php#Data%20entry

Acknowledgement

Sincere appreciation is given to my dissertation supervisor, Dr. Manwa Lawrence Ng, Associate professor, The University of Hong Kong, who has driven me all the way with his expert opinion and comments. Special thanks are extended to the participants and staff at the New Voice Club of Hong Kong for their full support; Mr. Raymond Fong and Mr. David Chow for their expert clinical opinions on the questionnaire evaluation; speech therapists for their contribution in rating the speech samples. Appreciation also is expressed to friends, including Grace Yip, Ivy Mok, Melody Yu and Stephen Li, classmates, Eric Tong, May Mak and Morris Poon and family for their full support.

Appendix A: Assessment questionnaire

Preliminary Assessment Questionnaire for

Cantonese Alaryngeal Speakers' Speech Performance(For ST use)

Client's Name: _____ Gender/Age: _____ Date of assessment: _____

History of hearing loss: Y/N Visual impairment: Y/N Manual Dexterity: Good/ Fair/ poor

Type of alaryngeal speech spoken: EL(☐ Neck ☐ Oral) /TE/ SE/ PA

Objective Parameters

Maximum phonation time ***(measured by stop watch)*** **[For SE TE and PA Only]**

Production of /a/:

Trial1:_____Trial2:_____Trial3:_____

Perceptual Parameters

[Task: Reading standard passage "The North Wind and theSun"]

A. Voice Quality **[*For SE and TE Only]**

	Severe			**Moderate**			**Appropriate**
*Hoarseness	1	2	3	4	5	6	7
*Breathiness	1	2	3	4	5	6	7
*Effortful	1	2	3	4	5	6	7
*Phonation breaks	1	2	3	4	5	6	7

B. Pitch and Loudness

	Mono/ Too high/ Too low			**Moderate**			**Appropriate**
Average pitch	1	2	3	4	5	6	7
*Pitch variation **(For SE, TE & PA)**	1	2	3	4	5	6	7
Average Loudness	1	2	3	4	5	6	7
*Loudness Variation **(For SE, TE & PA)**	1	2	3	4	5	6	7

C. Fluency

	Too Fast/slow			**Fair**			**Appropriate**
Speech Rate	1	2	3	4	5	6	7
	Too Short			**Fair**			**Appropriate**
Phrase length	1	2	3	4	5	6	7

D. Audible Distraction

	Noisy			**Fair**			**Acceptable**
Stoma noise	1	2	3	4	5	6	7
*Electrolarynx noise **(For EL only)**	1	2	3	4	5	6	7

E. Articulation [Speech task: CV monosyllabic words of /ji/ and /si/ produced at six tones]

	Poor			**Fair**			**Acceptable**
Articulation proficiency	1	2	3	4	5	6	7
Tone production proficiency	1	2	3	4	5	6	7

F. Coordination

[*For EL and PA speakers]

	Poor			**Fair**			**Good**
*Coordination: placement and articulation	1	2	3	4	5	6	7

[*For TE and PA speakers]

	Poor			**Fair**			**Good**
*Coordination: stoma occlusion and phonation	1	2	3	4	5	6	7

[*For SE, TE and PA speakers]

	Poor			**Fair**			**Good**
*Coordination:respiration and phonation	1	2	3	4	5	6	7

G. Placement of coupling device **[*For EL and PA speakers]**

Accuracy = Correct location at the sweet spot

	Poor			**Fair**			**Good**
*Accuracy	1	2	3	4	5	6	7
*Consistency	1	2	3	4	5	6	7

H. Overall proficiency

	Poor			**Fair**			**Accurate**
Intelligibility	1	2	3	4	5	6	7

Appendix B: Reading materials: The North Wind and the Sun

北風和太陽

有一次，北風同太陽喺度拗緊邊個叻啲。佢哋啱啱睇到外面有個人行過，哩個人著住件大褸同恤衫。佢哋就話嘞，邊個可以整到哩個人除咗件褸呢，就算邊個叻啲。於是，北風就拼命咁吹。點知，佢越吹得犀利，嗰個人就越係會執實件褸。最後，北風冇晒符，唯有放棄。跟住，太陽出嚟晒咗一陣，嗰個人就即刻除咗件褸嘞。於是，北風唯有認輸啦。

Appendix C: Reading Materials : Monosyllabic words for tonal contrast

史　　（嘗）試

時　　意

二　　市

以　　移

詩　　衣

椅

是

Appendix D: Figures

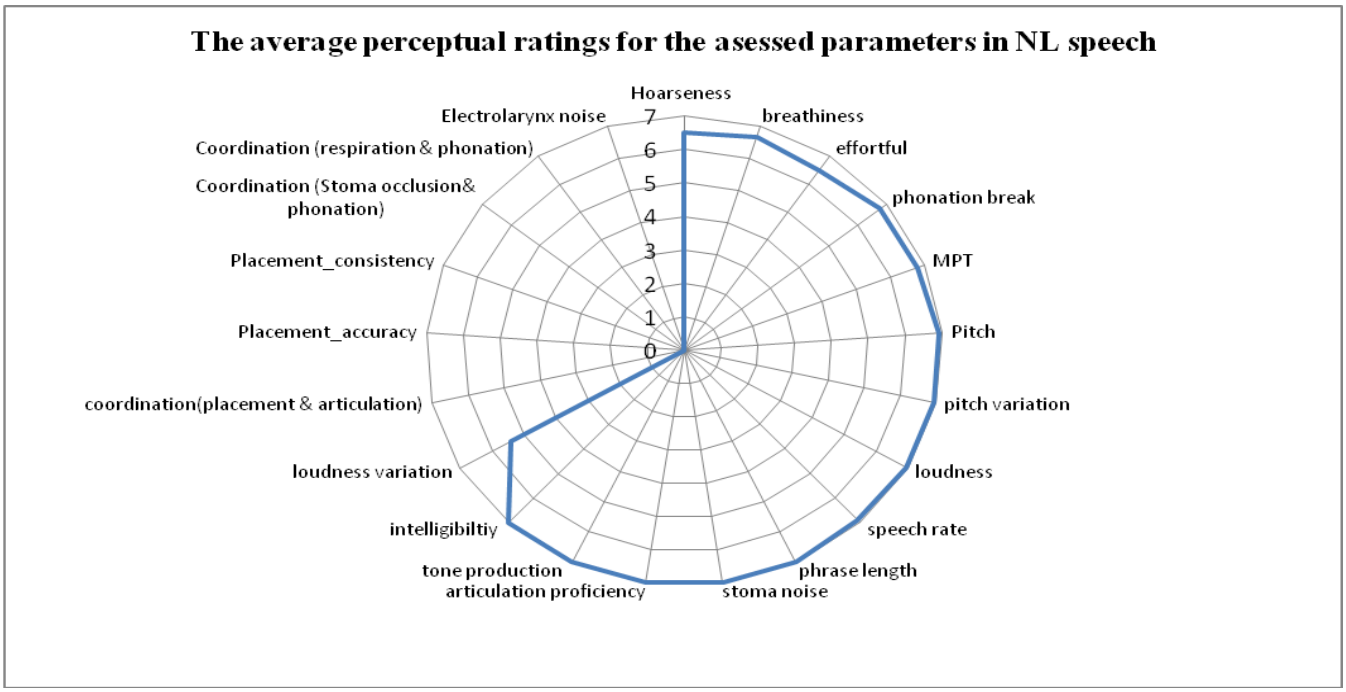

Figure 1. Average perceptual ratings on assessed parameters of NL speech

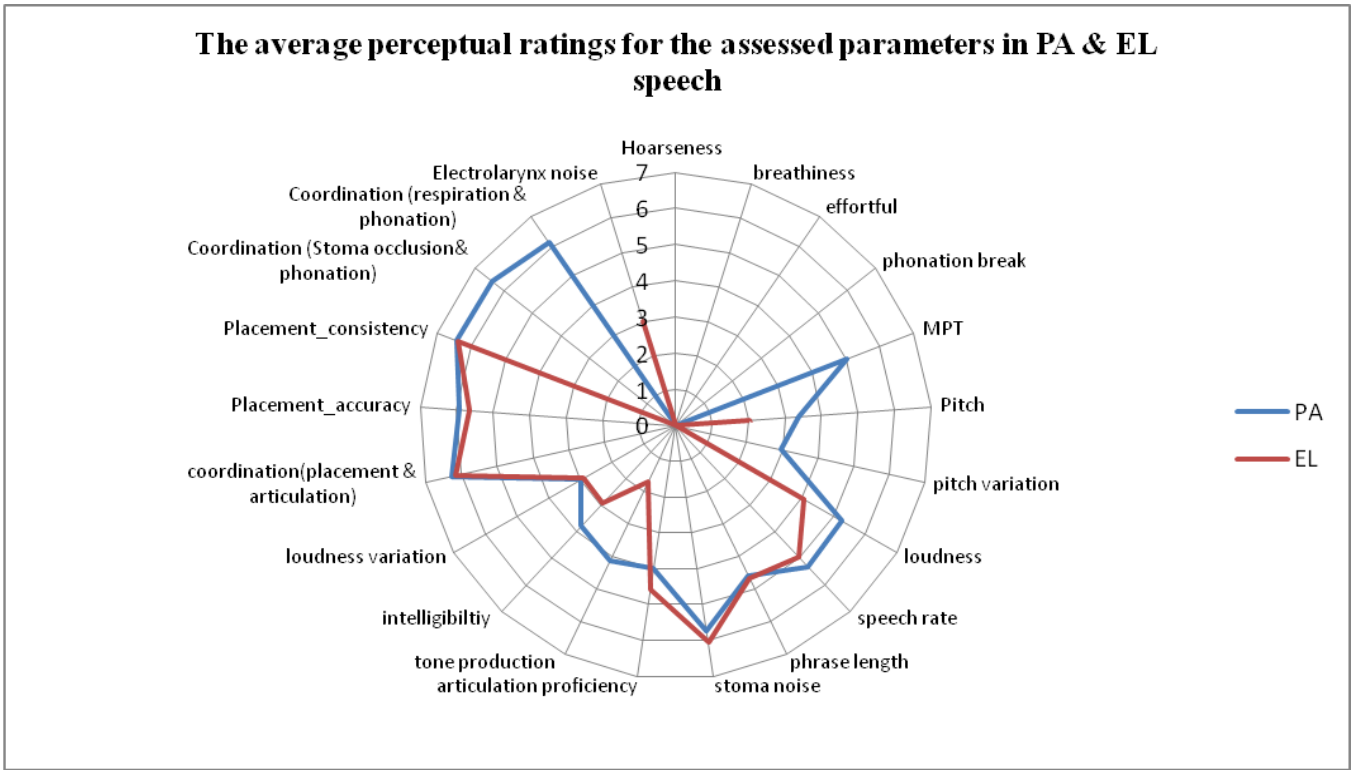

Figure 2. Average perceptual ratings on assessed parameters of SE and TE speech

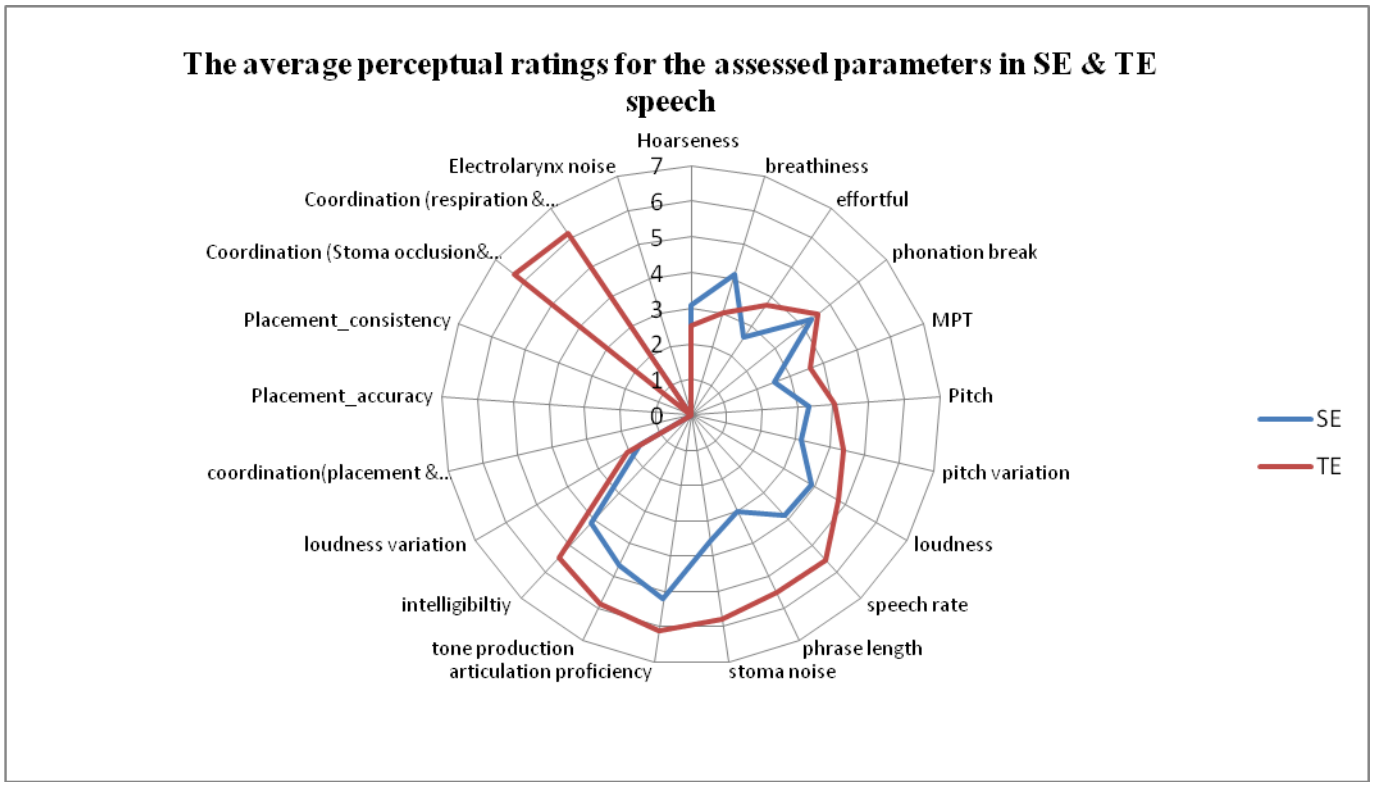

Figure3. Average perceptual ratings on assessed parameters of PA and EL speech