말소리와 음성과학 제1권 제4호 (2009)
pp. 165~170

Multiple Average Ratings of Auditory Perceptual Analysis for Dysphonia

Choi, Seong-Hee[1], Choi, Hong-Shik[2]

ABSTRACT

This study was to investigate for comparison between single rating and average ratings from multiple presentations of the same stimulus for measuring the voice quality of dysphonia using 7-point equal-appearing interval (EAI) rating scale. Overall severity of voice quality for 46 /a/ vowel stimuli (23 stimuli from dysphonia, 23 stimuli from control) was rated by 3 experienced speech-language pathologists (averaged 19 years; range = 7 to 40 years). For average ratings, each stimulus was rated five times in random order and averaged from two to five times. Although higher inter-rater reliability was found in average ratings than in single rating, there were no significant differences in rating scores between single and multiple average ratings judged by experienced listeners, suggesting that auditory perceptual ratings judged by well-trained listeners have relatively good agreement with the same stimulus across the judgment. Larger variations in perceptual ratings were observed for moderate voices than for mild or severe voices, even in the average ratings.

Keywords : multiple average ratings, voice quality, auditory perceptual evaluation, dysphonia

1. Introduction

Perceptual analysis of voice is by definition subjective, guided by the rater's cumulative experience with each particular voice disorders. Auditory perceptual evaluation has been considered as the "gold standard" of voice assessment, and used in conjunction with objective measures for voice quality (de Krom, 1995; Kreiman & Garnett, 2000). Previous studies using rating scales to evaluate voice quality have varied widely in methodology. For examples, studies using perceptual rating scales have included categorical ratings requiring assignment of voice samples to discrete and unordered categories, equal-appearing interval (EAI) scales, visual analog scales, and direct magnitude estimation (DME) that assign a specific magnitude to an anchor voice characteristic (Eadie & Doyle, 2002; Wuyts et al., 1999; Yiu et al., 2004; 2007).

Clinically, Grade, Roughness, Breathiness, Aesthenia, Strain (GRBAS) perceptual 4-point equal appearing interval rating scale has been commonly used (Hirano, 1981). In addition, recently, new developed consensus auditory-perceptual evaluation of voice (CAPE-V)(Kempster et al., 2008) has been recommended to document more voice quality features than GRBAS across more speech tasks in the clinical practice by American Speech-Language-Hearing Association. Visual analog scale is used to assess each of the 6 voice quality features –overall severity, roughness, breathiness, strain, pitch, loudness- and allows flexibility to add other perceptual features of interest. It also includes the ordinal ratings of mild, moderate, and severe. Although these perceptual ratings provide clinical standard for auditory-perceptual evaluation and have been recommended for the clinical assessment of voice quality, methodogical weakness and unsolved problems remain. The most concerns for auditory perceptual evaluation of voice quality are intra- and inter- judge variability and reliability related to sources of listeners disagreement, differences among raters in perceptual strategy, and potential errors on perceptual reponses (Kreiman et al., 1993 Kreiman, Gerratt, Berke, 1994; Kreiman & Gerratt, 2000). A

[1] University of Wisconsin - Madison, Department of Surgery - Otolaryngology, choi@surgery.wisc.edu
[2] Yonsei University College of Medicine, Dept. of Otorhinolaryngology, Institute of Logopedics and Phoniatrics, hschoi@yuhs.ac

Received: July 28, 2009
Revision received: August 29, 2009
Accepted: September 11, 2009

listener's response to a stimulus may be influenced by a number of factors other than the physical characteristics of the stimulus. Two kinds of errors are frequently identified when using a rating scale for a psychophysical task (Shrivastav et al., 2005). The first kind, which is called 'random errors' such as rapid changes in attention, listeners fatigue, chance and lead to random variability in a listeners' response. The other kind of error is 'criterion errors' or 'response bias', resulting from the use of different criteria by different listeners. To explore the source of listener disagreement, Kreiman & Gerratt (2000) used synthesized vowel samples of disordered voices to identify and quantify sources of listener disagreement. They found that 84% of the variance in the extent to which listeners do or do not agree could be accounted for by four factors - instability of internal memory standards for levels of a perceptual dimension, ability to isolate single dimension in a complex context, scale resolution, and absolute magnitude of the attribute being measured. This approach, however, provided new insight into the source disagreement, synthesizing the disordered voices being assessed is still challenging.

In an alternative approach, Shrivastav et al. (2005) have recently shown that listener variability can be minimized by applying psychometric principles. Such principles include averaging the ratings from multiple presentations of the same voice samples, with investigators showing that a minimum of five repetitions provide the best results for judging voice quality in sustained vowels. Hence, the present study was attempted to see if multiple average rating methods with increasing number of averaged ratings judged by experienced listeners can improve listener agreement and reliability than single rating for measuring the voice quality.

2. Methods

Stimuli

Voice samples from the Voice Disorders database Version 1.03 (Kay Elemetrics, Inc. Boston, MA) were used for both dysphonia groups and controls. 23 /a/ vowels for control and 23 /a/ vowels for dysphonia were selected for auditory perceptual analysis.

Auditory-Perceptual ratings

Perceptual ratings were performed in a sound-treated room. TDH-39 headphones, not computer speaker at a comfortable listening level, approximately 70dB SPL were used for all ratings. 3 listeners rated overall severity of voice quality for 46 /a/ vowel stimuli using a 7- point equal-appearing interval (EAI) rating scale (1= normal voice, 7 = the largest deviation from normalcy).

All raters were trained and experienced speech-language pathologists. They averaged 19 years' postgraduate experience evaluating voice disorders (SD = 18.2 years; range = 7 to 40 years). None of the raters reported any history of hearing, speech, voice, or language difficulties. Each stimulus included one presentation of each stimulus but it was presented to the listeners 5 times in random order. For this, five lists, with 46 stimuli each (23 stimuli from each patient, 23 stimuli from control) were made with randomized order and saved stimuli in form of wave file in DVD. Tasks were then given with the fixed same presentation order across listeners to avoid effects of different task orders for within-task variability. Therefore, all listeners heard the same experimental DVD, but performed the tasks independently. Once a judgment was made, raters have then moved to the next sample, but were not allowed to go back and revise a rating. All ratings should be completed in a single setting. A 5 - 10 min break was allowed between each of the five lists to optimize listeners' attention and minimize fatigue.

Auditory-perceptual analysis

Each listener completed ratings from 5 randomized lists of stimuli. To achieve greater agreement and reliability the multiple ratings of each stimulus from each listener were averaged (Shrivastav, Sapienza & Nandur, 2005). First, discrete ratings were achieved from the absolute ratings for each of 5 lists of stimuli. Second, the multiple ratings from two to five times ratings were averaged across listeners.

Statistical analysis

All statistics were performed using SPSS version 11.5 for windows (SPSS, Cary, NC). Pearson correlations were computed for inter-rater reliability of all three judges. One-way repeated measures analysis of variance (ANOVA) was used to compare rating scores across 5 lists for within single rating and average ratings. A one-way ANOVA with post hoc Tukey tests was used to determine if rating scores with averaging successive ratings were different with scores of single rating in normal and dysphonia group. A one-way ANOVA was used to investigate variability depending on the severity of voice quality. Statistical p-values less than 0.05 were considered a significant difference.

3. Results

3.1. Single ratings vs. multiple average ratings

The rating scores for single rating and average rating made by three experienced listeners were summarized in Table1 and Table 2. One-way repeated measures analysis of variance (ANOVA) that compared rating scores across the 5 lists showed that there were significant differences among raters in all single rating with 5 lists (p<.01) and multiple average ratings (p<.01) whereas no significant differences within raters across five set of ratings (p>.05).

Table 1. Rating scores obtained from single ratings of each five lists among raters.

single ratings	rater1	rater2	rater3	p-value

List 1	M=3.57
	SD=2.21
		M=2.65
		SD=1.83
			M=2.15
			SD=1.89
				p<.01*

List 2	M=3.72
	SD=2.01
		M=2.67
		SD=1.85
			M=2.43
			SD=2.07
				p<.01*

List 3	M=3.65
	SD=2.05
		M=2.50
		SD=1.91
			M=2.15
			SD=1.92
				p<.01*

List 4	M=3.60
	SD=2.04
		M=2.65
		SD=1.89
			M=2.17
			SD=1.90
				p<.01*

List 5	M=3.65
	SD=2.04
		M=2.57
		SD=1.93
			M=2.04
			SD=1.86
				p<.01*

Table 2. Rating scores obtained from averaging each list from twice to five times among raters.

Average ratings	rater1	rater2	rater3	p-value

A2	M=3.64
	SD=2.10
		M=2.66
		SD=1.98
			M=2.29
			SD=1.98
				p<.01*

A3	M=3.64
	SD=2.09
		M=2.49
		SD=1.96
			M=2.24
			SD=1.96
				p<.01*

A4	M=3.64
	SD=2.07
		M=2.41
		SD=1.94
			M=2.23
			SD=1.94
				p<.01*

A5	M=3.64
	SD=2.07
		M=2.34
		SD=1.93
			M=2.19
			SD=1.93
				p<.01*

Each listener's average rating for each stimulus was obtained by averaging the ratings for that stimulus on two or more lists (A2, A3, A4, A5). A2 refers to average list 1 and 2. A3 refers to average list 1, and 2, and 3. A4 refers to average list 1, and 2, and 3, and 4. A5 refers to average list 1, and 2, and 3, and 4, and 5 so on.

Although there was a small decrease in standard deviation between A2 and A3 and subsequent comparisons, SD was decreased when a greater number of ratings were averaged, implying variations within raters were decreased as shown in Table 2.

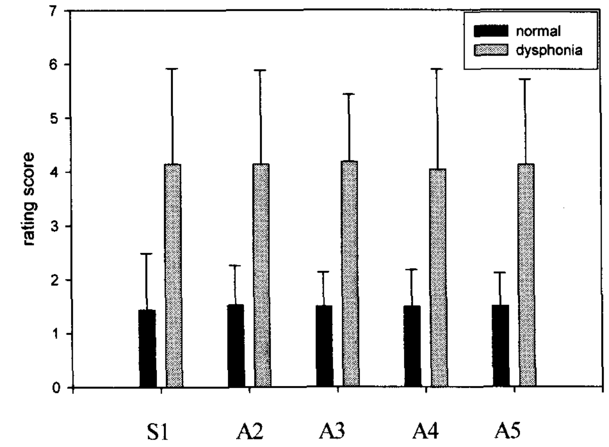

Figure 1. Rating scores obtained from single rating of 1st list and average ratings for each stimulus obtained by averaging the ratings for that stimulus on two or more lists (A2, A3, A4, A5) in normal and dysphonia group.

Rating scores of stimuli were calculated by single ratings from 1st list (Normal: M=1.43, SD=1.05, Dysphonia: M=4.14, SD=1.77) and average ratings from A2 (Normal: M=1.52, SD=0.73, Dysphonia: M=4.14, SD=1.74), which averaged 1st and 2nd list. Likewise, A3 (Normal : M=1.49, SD=0.63, Dysphonia: M=4.18, SD=1.25) averaged from 1st to 3rd list, A4(Normal: M=1.49, SD=0.68, Dysphohia: M=4.03, SD=1.86) averaged from 1st to 4th list, and A5 (Normal: M=1.50, SD=0.61, Dysphonia: M=4.12, SD=1.58) averaged from 1st to 5th list in normal voice and dysphonia. All of the rating scores of dysphonia were significantly higher than normal voice (p<.001).

A one-way ANOVA with post hoc Tukey tests was used to determine if rating scores with averaging successive ratings were different when compared to scores of single rating in normal and dysphonia group. Tukey post hoc comparison showed that the rating score of the first single rating was not significantly different from A2, A3, A4, A5 in both normal voice and dysphonia as shown in Figure 1.

3.2. Inter-rater reliability of single rating and multiple average ratings

Pearson's r was calculated to measure the inter-rater reliability for single ratings and five time average ratings between raters.

Pearson's r for 1 st set of single rating and five times average ratings are shown in Table 3 and Table 4, respectively. Pearson's r in single rating was found from .684 to .935. Five time average ratings, however, showed that inter-rater reliability was improved ranged from .762 to .960 than single rating.

Table 3. Correlation matrix for perceptual ratings of overall severity for each listener of 1st set of single rating and the group mean ratings (p<.001).

Perceptual	Perceptual rater			group
rater	rater1	rater2	rater3	mean
rater1	1.000	-	-	-
rater2	.815**	1.000	-	-
rater3	.684**	.789**	1.000	-
group mean	.914**	.935**	.894**	1.000

Table 4. Correlation matrix for perceptual ratings of overall severity for each listener of five times average ratings and the group mean ratings (p<.001).

Perceptual	Perceptual rater			group
rater	rater1	rater2	rater3	mean
rater1	1.000	-	-	-
rater2	.856**	1.000	-	-
rater3	.762**	.853**	1.000	-
group mean	.932**	.960**	.926**	1.000

3.3. Average ratings in terms of the severity of dysphonia

To investigate variability in terms of the severity of voice quality including normal and dysphonia, stimuli were classified by severity based on the rating scale - mild (1-2), moderate (3, 4, 5), severe (6-7). Although average rating scales have higher inter-reliability than single rating, one-way ANOVA showed that SD between-rater in moderate severity voices is significant higher than in mild or severe (p<.001) as shown in Figure 2, reflecting moderate dysphonia has higher variability in the auditory perceptual ratings between raters.

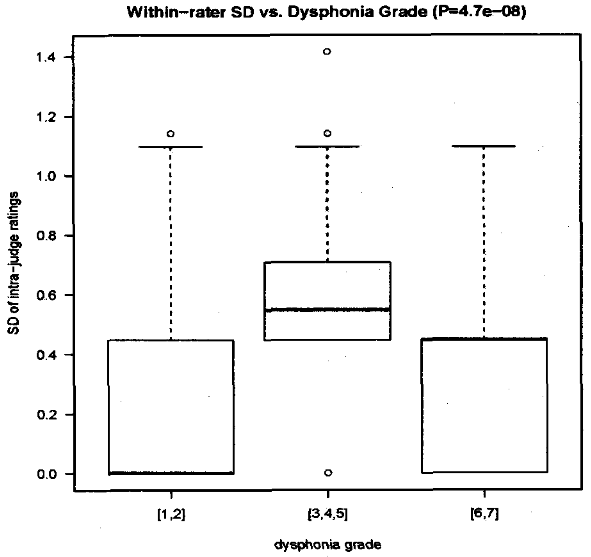

Figure 2. Inter-rater standard deviation in multiple average rating in terms of severity of dysphonia.

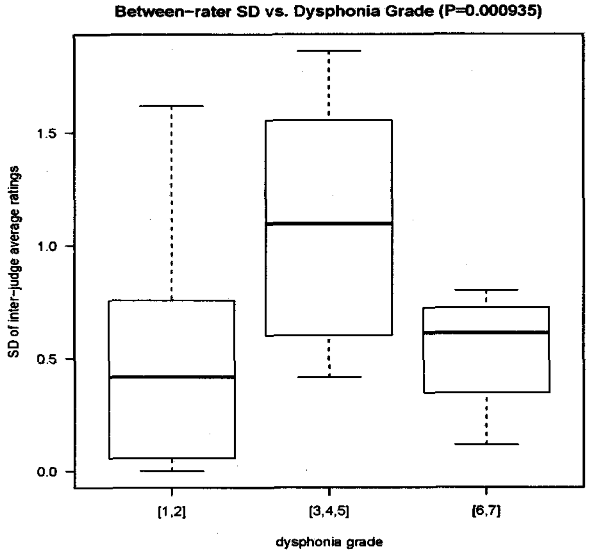

Figure 3. Intra-rater standard deviation in multiple average rating in terms of severity of dysphonia.

In addition, one-way ANOVA was used to compare SD within-rater in terms of severity of voice quality. Results demonstrated that significant higher SD of intra-judge average ratings was observed as SD of intra-judge average ratings between-rater.

4. Discussion

Auditory perceptual ratings has been widely used for diagnosis and treatment efficacy for dysphonia. Although rating scales may be capable of providing useful information about voice quality perception, they need to be used with a good understanding on how listeners perceive voice quality. Furthermore, a listener"s response to a stimulus may be influenced by a number of factors such as momentary changes in attention, fatigue, memory of previously presented stimuli, training, and past experiences with the stimuli/task, as well as other chance factors (Poulton, 1989).

Large body of literatures and research have shown that level of interater reliability and agreement vary across scales, listener groups, and voice sets. Yumoto, Sasaki, and Okamura (1984) found correlations ranging from .51 to .79 when eight laryngologists rated the hoarseness of 87 voices on a 4-point equal-appearing interval scale. Bassich and Ludlow (1986) had four inexperienced but intensively trained listeners rate 10 pathological voices on 13 seven-point scales. They reported a mean intraclass correlation of .71, with a range of .19 to .96 across the 13 scales.

Some previous research has shown the variability using probability of exact agreement and suggested that individual listeners, even when experienced, differed greatly in their judgments (Kreiman & Gerratt, 2000).

Recently, multiple average ratings were proposed to minimize listeners' variability and improve interjudge reliability in perceptual judgments (Shrivastav et al., 2005). They found that agreement and reliability were observed to improve when multiple ratings of each stimulus from each listener were averaged. Moreover, when listeners rated across the 10 presentations, averaging a minimum of five ratings was needed to obtain a significant improvement in agreement for the stimuli. In the present experiment, listeners were presented with each of 46 vowel stimuli 5 times in random order and were asked to rate overall severity using a 7-point rating scale. Hence, based on previous study, five presentations for auditory perceptual ratings were used in our experiments.

For evidence-based clinical practice, we also need to see if multiple average ratings will be required to evaluate auditory perceptual ratings in the clinical setting to improve agreement and reliability. Thus, the present study was attempted to compare the single rating with multiple average ratings rated by three experienced speech-language pathologists. Since time is limited in the clinical evaluation to measure the voice quality, we questioned if application of average ratings may result in more reliable perceptual rating outcomes. Our experiments results showed that source disagreement was found for same stimuli among listeners for overall severity voice quality even though the raters are well-trained and experienced speech-language pathologists. In this study, we didn't use any anchors to evaluate overall severity for normal and dysphonic voices before perceptual ratings and it may imply that listeners may have different strategies and different cues to evaluate perceived overall severity.

For reliability and agreement of multiple ratings of auditory-perceptual evaluation, our hypothesis is that experienced clinicians may show higher reliability in both single and multiple ratings. Because they may have learned more to calibrate judgments in normal voice and dysphonic voices than inexperienced listeners although we didn't include inexperienced listeners in this study. Our results showed that higher correlations were observed for inter-rater reliability in five time average ratings when compared to single rating. However, there was no significant differences between single rating and more than two average ratings up to five although SD tended to decrease with increasing number of averaged ratings. In the present study, listeners had averaged 19 years' postgraduate experience evaluating voice disorders ranging from 7 to 40 years. Shrivastav et al., (2005) found that reliability and agreement for perceptual ratings were improved with average multiple ratings. In contrasts to present study, all listeners were graduate students in speech-language pathology and had taken at least one class on voice disorders and thus they did have no much experience for evaluating voice quality. Regarding listener experiences, Kreiman et al. (1990) found that Clinicians and naive listeners attended to different cues when judging the same voices, for both normal and pathological voices. In addition, Gerratt et al. (1992) demonstrated individual differences in voice quality perception between clinicians and naive listeners. Clinicians apparently develop idiosyncratic approaches to rating pathological voice qualities in the course of their clinical training and practice. Extensive experience with pathologic voices may provide richer auditory strategies for subtle judgments regarding pathologic voice quality than that of naive listeners.

We speculate another possibility for what dimension of voice quality was measured. It is still unclear when judging the quality of a given voice, experienced listeners can consistently show the good reliability when judging the different voice quality to determine the difference between single rating and multiple average ratings. Because we measured only overall severity, not specific voice quality parameters such as breathiness and/or roughness in this study.

There are additional considerations for auditory perceptual analysis using rating scales. Generally, mild and severe voice quality has more good agreement in figure 2 and figure 3. However, the greater variability of listeners in the midpoints of a rating scale occurred even when we used average ratings, because stimuli at the scale midpoints are perceptually more similar than those at the scale end points.

5. Conclusion

In the present study, multiple average ratings were applied to measure the voice quality and compared to single rating. Although inter-rater reliability was increased with multiple rating compared to single rating, there was no evidence of differences between single rating and average ratings judged by experienced listeners. In the clinical view, well-trained listeners might do auditory perceptual analysis consistently when the same stimuli were presented repeatedly, suggesting that more reliable results for auditory-perceptual analysis may gain from the experienced listeners. Further study might be needed to confirm the present results by using different types of listeners such as inexperienced listeners, graduate students and different types of voice quality dimensions such as breathiness, strained-strangled, roughness, which may be differently affected in terms of listener's experience.

Acknowledgements : Authors are thankful to Diane M. Bless and Nathan V. Welham for auditory perceptual ratings.

References

Bassich, C.J., Ludlow, C.L.(1986). "The use of perceptual methods by new clinicians for assessing voice quality", *J Speech Hear Dis* Vol.51, 125-133.

De Keom, G. (1995). "Some spectral correlates of pathologic breathy and rough voice quality for different types of vowel fragments", *J Speech Hear Res,* Vol. *38,* pp.794-811.

Eadie, T.L., Doyle, P.C. (2002). "Direct magnitude, estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers". J Acoust Soc Am Vol.112, No.6, pp.3014-21.

Gerratt B.R, Kreiman J, Antonanzas-Barroso N, Berke GS. (1993). "Comparing internal and external standards in voice quality judgments", *J Speech Hear Res.* Vol.36, No.1, pp. 14-20.

Gerratt, B.R,. Kreiman. J. (2001). "Measuring vocal quality with speech synthesis", *J Acoust Soc Am,* Vol.110, pp.2560-2566.

Hirano, M. (1981). *Clinical examination of voice,* New York: Springer Verlag.

Kempster, G.B., Gerratt, B.R., Verdolini, A.K., Barkmeier - Kraemer, J., Hillman, R.E. (2008). "Consensus auditory - perceptual evaluation of voice: Development of a standardized clinical protocol", *Am J Speech Lang Pathol* Vol.18, No.2, pp. 124-32.

Kreiman, J., Gerratt, B.R. (2000). "Sources of listener disagreement in voice quality assessment", *J Acoust Soc Am* Vol.108, No.4, pp.1867-1876.

Kreiman J., Gerratt B.R., Precoda K. (1990) "Listener experience and perceptio of voice quality", *J Speech Hear Res* Vol.33, No.1, pp. 103-15.

Kreiman, J., Gerratt, B.R., Precoda, K., Berke, G.S. (1992). "Individual differences in voice quality perception", *J Speech Hear Res.* Vol.35, No.3, pp. 512-20.

Kreiman, J., Gerratt. B.R.(1998). "Validity of rating scale measures of voice quality assessment", *J Acoust Soc Am,* Vol. 104, pp. 1598-1608.

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., & Berke, G.S. (1993). "Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research", *J Speech Hear Res* Vol.36, pp.21-40.

Kreiman, J., Gerratt, B.R., & Berke, G.S. (1994). "The multidimensional nature of pathological voice quality", *J Acoust Soc Am* Vol.96, pp.1291-1302.

Kreiman, J., Gerratt, B.R. (2000). "Sources of listener disagreement in voice quality assessment", *J Acoust Soc Am* Vol. 108, No.4, pp.1867-1876.

Poulton, E. C. (1989). *Bias in quantifying judgments.* Hove, United Kingdom: Erlbaum.

Shrivstav, R., Sapienza, C.M., Nandur, V. (2005). "Application of psychometric theory to the measurement of voice quality using rating scales", *J Speech Hear Res* Vol. 48, No.2, pp.323-335.

Yiu, E.M., Ng, C.Y. (2004). "Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness", *Clin Linuist Phon* Vol.18, No.3, pp.211-29.

Yiu, E.M., Chan, K.M., Mok, R.S. (2007). "Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation", *Clin Linguist Phon* Vol.21, No.2, pp.129-45.

Yumoto, E., Sasaki, Y., Okamura, H. (1984). "Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness", *J Speech Hear Res* Vol.27. pp.2-6.

Wuyts, F.L., DeBodt, M.S. (1999). "Is the reliability of a visual analogue scale higher than an ordinal scale? An experiment with GRBAS scale for the perceptual evaluation of dysphonia". *J Voice* Vol.13, No.4, pp.508-17.

• **Choi, Seong Hee**
University of Wisconsin-Madison, Dept. of Surgery,
Otolaryngology. 5036 Wisconsin Institutes Medical Research,
1111 Highland Ave, Madison, WI 53705, USA
Tel : 1-608- 265-2268
Email: choi@surgery.wisc.edu

• **Choi, Hong-Shik**
Kangnam Severance Hospital, Yonsei University College of
Medicine. Institute of Logopedics and Phoniatrics
612 Enjuro Kangnamgu Seoul, Korea
Tel: 02-2019-3460 Fax: 02-3463-4750
Email: hschoi@yuhs.ac.