Ann. Bull. RILP (1983) No. 17, 187-199

VOICING DISTINCTION IN ESOPHAGEAL SPEECH –PERCEPTUAL, FIBEROPTIC AND ACOUSTIC STUDIES

Hajime Hirose, Masayuki Sawashima and Hirohide Yoshioka

Introduction

Alaryngeal speech is a very specific type of pathological speech in that the patient maintains the normal vocal tract system with an exception of the laryngeal mechanism. Among different types of alaryngeal communication, esophageal speech is perhaps the ideal method, since it is achieved using the potential of the patient's existing structure. Even so, it should be the case that esophageal speakers have difficulty in accomplishing the voiced-voiceless distinction for consonants, as they have no functioning larynx. Surprisingly, however, perceptual studies often reveal that there is a clear voicing distinction resulting in good articulation scores in skilled esophageal speech. As a matter of fact, many esophageal speakers are able to develop intelligible speech for successful oral communication and participate in social activities as they did preoperatively.

The purpose of the present study was to analyze the articulatory ability of skilled esophageal speakers in terms of the voicing distinction in consonants and to explore the physiological background of their articulatory behavior. A perceptual study was first performed to examine the pattern of the confusion that normal listeners would make in comprehending initial and medial consonants spoken by esophageal speakers. Fiberoptic observations of the neoglottis of selected cases of esophageal speakers were then made to explore the underlying physiological mechanism of the voicing distinction in esophageal speech. Finally, spectrographic analyses were performed to investigate the acoustic nature of esophageal speech, with special reference to the voicing distinction.

Procedures

1. Experimental subjects

Eight esophageal speakers (7 males and 1 female) served as the subjects of the present study. They ranged in age from 48 to 77 years, with a mean age of 63 years. The length of the postoperative periods was variable; the shortest was 2 years and one month; while the longest was more than 20 years (Table 1).

A version of this paper is presented at the 19th Congress of International Association of Logopedics and Phoniatrics, Edinburgh, Scotland, August, 1983.

Table 1 *List of length of subjects*

Case	Initial	Age	Sex	Postop. period (ys:ms)
1	TA	56	m	2:1
2	MOf	77	f	2:7
3	YW	60	m	3:3
4	YM	48	m	8:2
5	SH	52	m	9:6
6	MOm	74	m	15:6
7	SO	68	m	18:0
8	SN	65	m	22:4

2. *The perceptual test*

The subjects were required to read the following 20 meaningful Japanese words, each of which was consecutively indicated to the subject in a written form on a card. All the words were uttered in the frame " ____ desu (that is ____)".

/pasu/ /basu/ /teNki/ /deNki/ /keHyu/ /geHyu/
/seNi/ /zeNi/ /sireH/ /zireH/ /kiNpeN/ /kiNbeN/
/iteN/ /ideN/ /koHkeH/ /koHgeH/ /kaisoH/ /kaizoH/
/isi/ /izi/

Recordings were made using a close-talk microphone, and the recorded speech samples were randomized for each subject and used in the subsequent perceptual test.

The randomized utterance samples were presented to 10 judges in an unechoic room. The judges were requested to fill in a blank on an answer sheet corresponding to the underlined mora of each word shown above with one of the Japanese kana based on their perceptual judgment.

3. *Fiberoptic observation*

The movement of the esophageal orifice was observed using a flexible fiberscope inserted through the nose in 4 of the 8 esophageal speakers (Cases 1, 4, 5 and 8). Sixteen mm movies were taken at a frame rate of 24 frames/sec for color film and 50 frames/sec for monochrome film. The latter films were later analyzed using a film analyzer frame-by-frame with reference to the audio-signal recorded simultaneously. As the test words for the fiberoptic filming, the following words were uttered in isolation.

/seHseH/ /teHteH/ /deHdeH/ /zeHzeH/ /seHseHseH/
/teHteHteH/ /deHdeHdeH/ /zeHzeHseH/
/siseH/ /ziseH/ /ziQseH/ /siQseH/
/beHeH/ /eQpeH/ /beHheH/ /peHpeH/
/isi/ /izi/ /eH/ /eHeH/

4. Spectrographic analysis of the recorded utterance samples

The recorded samples obtained for the perceptual tests were analyzed using a sound spectrograph (KAY) with a wide band filter.

Results

1. Results of the perceptual experiment

Figure 1 shows the results of the perceptual experiment in the form of a confusion matrix in which 1,600 data (20 test words × 8 subjects × 10 judges) are included. As a whole, there is a small percentage of voiced-voiceless confusion (168/1,900 = 10.5%), in which mistakes of voiceless consonants for voiced ones are more common (13.5%) than those in the opposite direction (7.5%). There was a considerable individual difference among the subjects in that the lowest confusion ratio was 0.5% (subject SN), while the highest was 18.5% (subject MOf). It was also found that confusion was more dominant in the word initial position (14%) than word medially (7%). The difference in the confusion ratio with reference to place and manner of articulation is presented in Table 2.

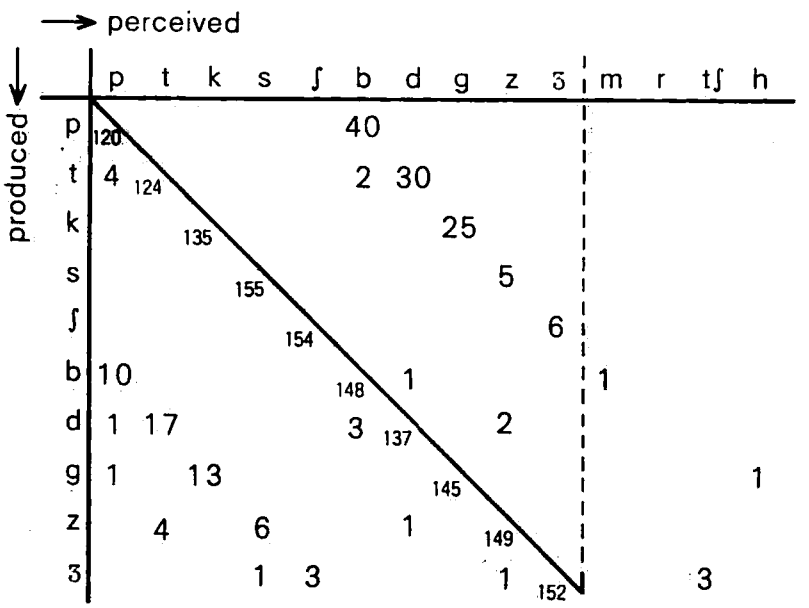
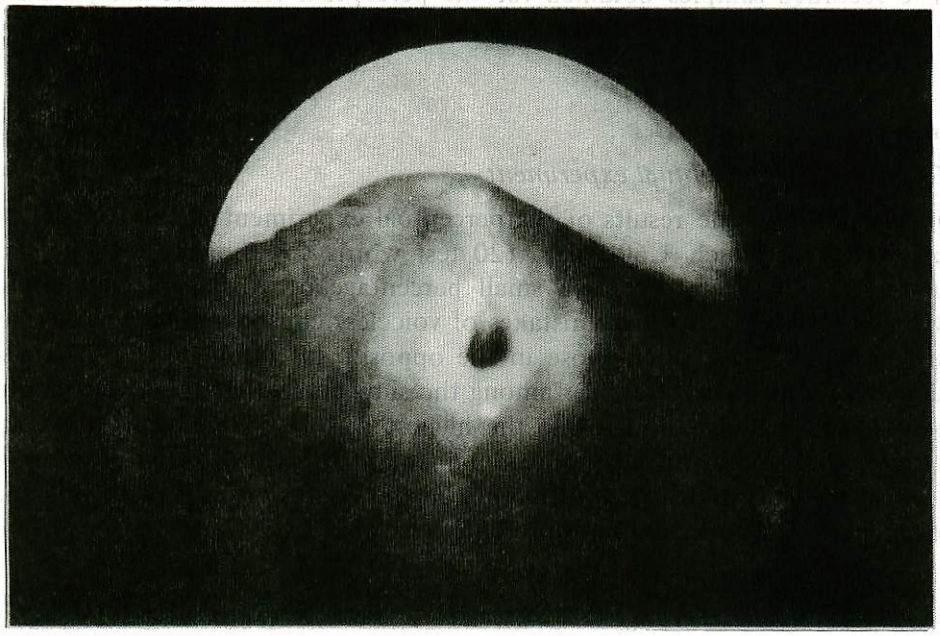


Fig. 1 *Confusion matrix for speech samples obtained from the 8 cases (10 Judges).*

(a)



(b)

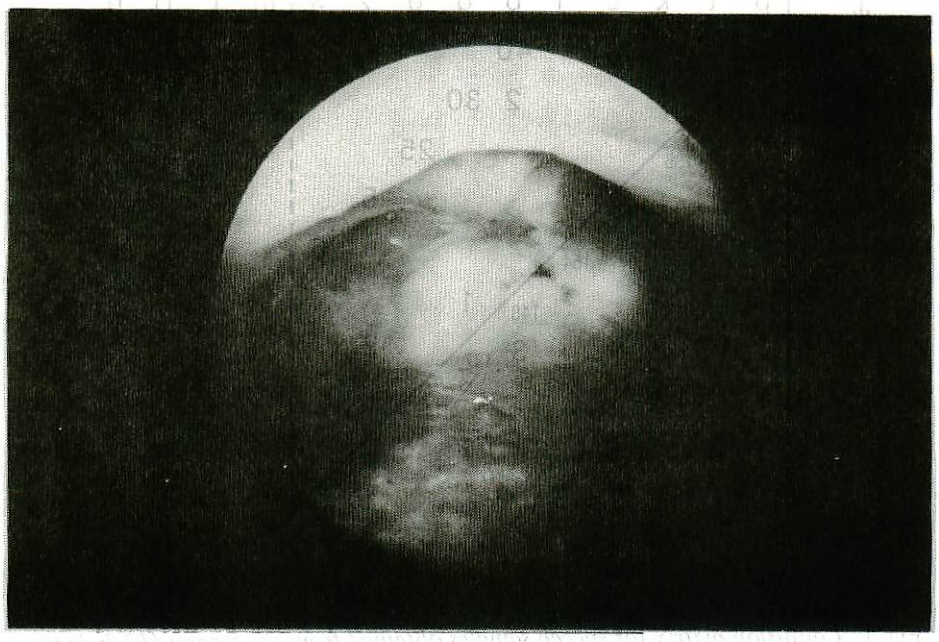


Fig. 2 *Fiberoptic view of the neoglottis taken immediately prior to phonation (prephonatory air-intake) (a); and during a sustained phonation of the vowel /e/ (b).*

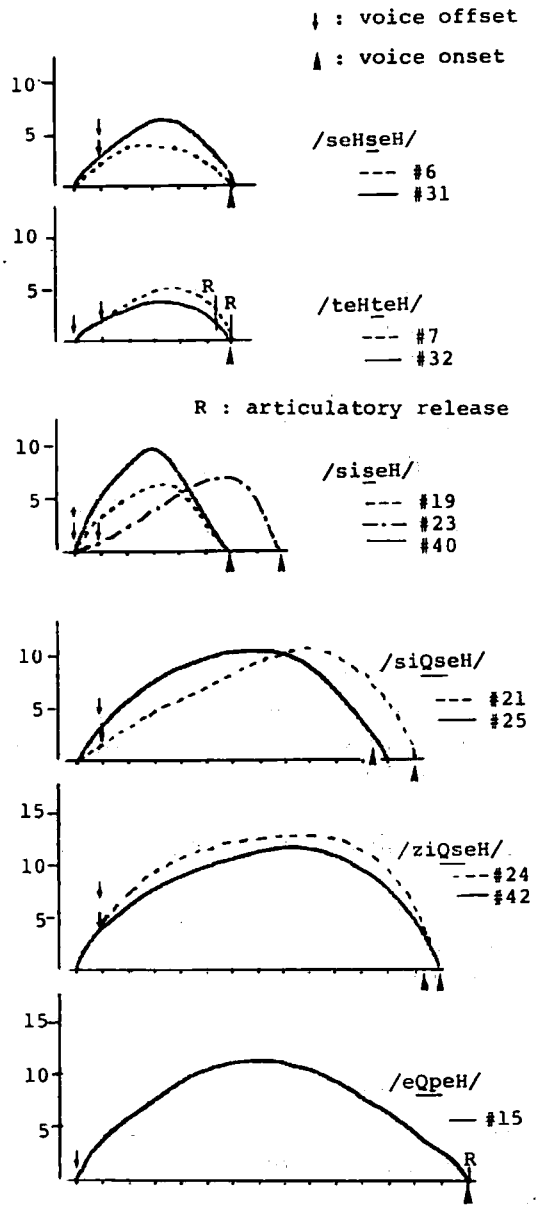


Fig. 3 *Time curves of the opening and closing gestures of the neoglottis for the production of voiceless sounds and sound sequences (subject SN). Curves for the same utterance type are superimposed. Each number above corresponds to the utterance number. The ordinate represents the size of the opening along an artbitrary scale, while the abscissa represents the time course (20 ms. per section).*

Table 2 *Comparison of the rate of confusion with respect to differences in place and manner of articulation*

p	vs.	b	:	15.6%
t	vs.	d	:	15.6%
k	vs.	g	:	12.5%
s	vs.	z	:	4.7%
	vs.		:	4.1%

2. Fiberoptic findings

Immediately prior to phonation, the esophageal orifice was found to open roundly for the intake of air to the esophageal lumen (Fig. 2-a). The closure of the esophageal orifice was then obtained, not by a circular constriction of the entire circumference but by an approximation of the anterior and posterior walls of the orifice. At the very margin of the orifice, in particular, the mucosal surface of the anterior and posterior walls appeared to close towards each other in a fashion similar to bilabial closure and then vibrated (Fig. 2-b). Thus, the mucosal margins of the esophageal orifice seemed to form a neoglottis in most cases.

It was observed that there was a transient opening of the neoglottis for the production of vioceless consonants, but the degree of opening was much less than that for prephonatory air-intake.

Figure 3 shows the time curves of the opening gestures of the neoglottis in Subject SN measured on the antero-posterior axis of the neoglottis with the monochrome film frames for different types of voiceless sounds and sound sequences. In this figure, the time curves are superimposed for the same utterance types. It can be seen that there is a considerable variation in the time course of the opening gesture of the neoglottis even for the same utterance type. Also, both size and duration of the opening for geminates are generally greater than those for single consonants.

After the cessation of each utterance, the esophageal orifice was found to close loosely. Tight constriction of the pharyngeal lumen was observed in none of the subjects during voiceless sound production.

3. Spectrographic analysis

In order to examine the acoustic characteristics of plosive sounds produced by the present subjects, VOT values were measured for plosives in word-initial position, while the closure period was measured for plosives in word-medial position. The presence or absence of voicing during the closure period was also examined. The results are presented in Table 3.

The results shown in Table 3 can be summerized as follows.

1) When the subjects intended to produce voiced plosives,
   a. If a voicing lead in the word-initial position was
      (1) present, then the rate of confusion was 10% (14/140)
      (2) absent, then the ration of confusion was 19% (19/100)
   b. if voicing in the word-medial position was
      (1) present, then the rate of confusion was 0% (0/230)

(2) absent, then the rate of confusion was 60% (6/10)

2) When the subjects intended to produce voiceless plosives
   a. in the word-initial position,
      (1) if VOT was within the range of 0 to 30 msec, then the rate of confusion was 35% (63/180)
      (2) if VOT was over 31 msec, then it was 6.7% (4/60)
   b. in the word-medial position, if voicing was
      (1) present, then the rate of confusion was 38% (19/50)
      (2) absent, then the rate of confusion was 6.8% (13/190)

These results clearly indicate that VOT values in word-initial position and the presence or absence of voicing in word-medial position are quite important cues for the voiced-voiceless distinction. Even so, it is known that there are other cues which are significant for the voicing distinction for plosive consonants. Indeed, close observations of the present utterance samples suggested that the pattern of the F1 transition, and the presence or absence of a silent period after the stop-release could also be taken as important supplementary cues for the voicing distinction.

Representative samples of sound spectrograms are shown in Fig. 4. Fig. 4-a shows sound spectrograms for the utterance samples /pasudesu/ and /basudesu/ produced by Subject SN. There was no confusion for this pair in the perceptual test. In the case of the word-initial /b/, there was a clear voicing lead with a clear F1 transition. For the word-initial /p/, the noise burst of the stop-release followed by a silent period was also clearly seen. These features seem to be important cues for the voicing distinction in this case. It can also be seen that the vowel length of the syllable /ba/ was longer than that of /pa/, and the vibratory pattern was very regular for these vowel segments.

Figure 4-b shows similar utterance samples for Subject SO, for which there was no confusion either. Although there was no voicing lead for /b/, the clearer F1 transition for /b/ than for /p/ seems to have played an important role as a cue for the voicing distinction. Also, the noise burst which was clearly seen for the /p/ release, and the subsequent silent period may also be taken as good cues for the descrimination of /p/ in this case.

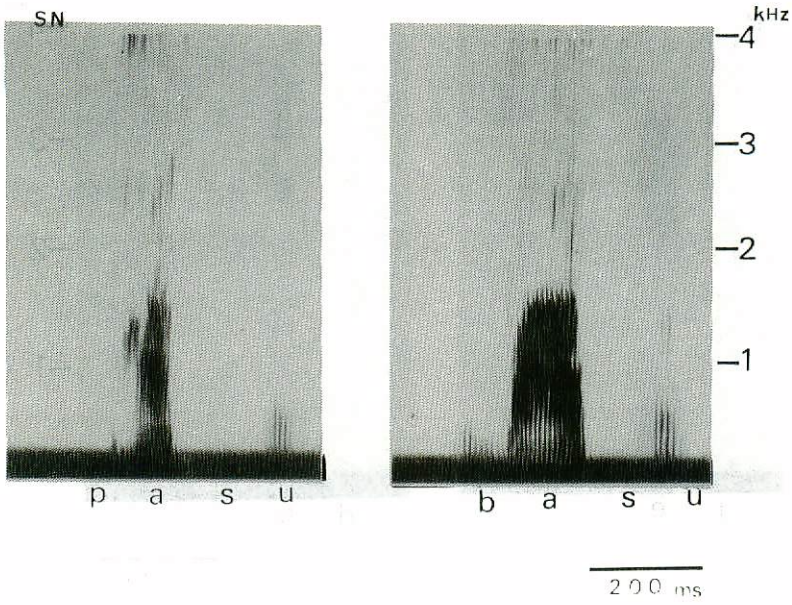
Figure 4-c shows the spectrograms of the utterance samples /teNkidesu/ and /deNkidesu/ produced by Subject YM. There was a confusion of /t/ as /d/ in 6 out of 10 cases: i.e. 6 judges out of 10 thought that they heard a voiced /d/ sound when the subject intended to produce /t/. On the other hand, a confusion of /d/ as /t/ was noted in 8 out of 10 cases. These samples therefore seem to be reversed in terms of the voicing distinction. It can be seen that the F1 transition was unclear for /d/, although there is a voicing lead of very short duration. For /t/, on the other hand, the F1 transition was rather clearer than for /d/, and there was almost no silent period after the /t/ release, the noise burst of which seems slightly clearer than that for /d/.

Table 3 *Results of the measurement of acoustical parameters on the sound spectrograms obtained from each speaker with reference to the pattern of the confusions for voiced-voiceless pairs.*

		Word-Initial				Word-Medial		
CASE	PAIR	voiceless VOT(ms)	conf.	voiced VOT(ms)	conf.	closure period (ms) & voicing voice-less	voiced	confusion
TA	p:b	0	b/p = 5	0	p/b = 1	168 +	104 +	b/p = 6
	t:d	0	d/t = 5	0		64 -	48 +	
	k:g	0	g/k = 6	0		72 -	64 +	g/k = 1
MOf	p:b	0	b/p = 7	-96		336 +	392 +	b/p =4
	t:d	+16	d/t = 2	-72	t/d = 2	128 -	88 +	
	k:g	+48	g/k = 1	-80		216 +	312 -	g/k =3, k/g = 6
YW	p:b	+16	b/p = 3	0	p/b = 3	168 -	96 +	
	t:d	+16		+16	t/d = 5 p/d = 1	88 +	48 +	d/t = 2
	k:g	+16		+16	k/g = 3	120 -	64 +	
YM	p:b	+32		-40	p/b = 1	104 +	136 +	b/p = 4
	t:d	+24	d/t = 6	- 8	t/d = 8	96 -	56 +	
	k:g	+32	g/k = 3	-40	k/g = 3	112 -	64 +	
SH	p:b	0	b/p = 3	0	p/b = 4	96 -	56 +	
	t:d	+16	d/t = 1	0	t/d = 2	104 -	48 +	
	k:g	+32		-64		128 -	80 +	
MOm	p:b	0	b/p = 4	-72		376 -	296 +	b/p = 8
	t:d	0	d/t = 8	-64		136 -	88 +	d/t = 1
	k:g	+16	g/k = 6	0		144 -	96 +	g/k = 3
SO	p:b	+16		0		160 -	200 +	
	t:d	0	d/t = 4 b/t = 1	-56		104 -	104 +	
	k:g	+24	g/k = 1	-96		136 -	120 +	
SN	p:b	+40		-88		160 -	72 +	
	t:d	+16	d/t = 1	-152		96 -	48 +	
	k:g	+48		-120		112 -	120 +	

x/y = N: confusion of y as x in N out of 10

(a)



(b)

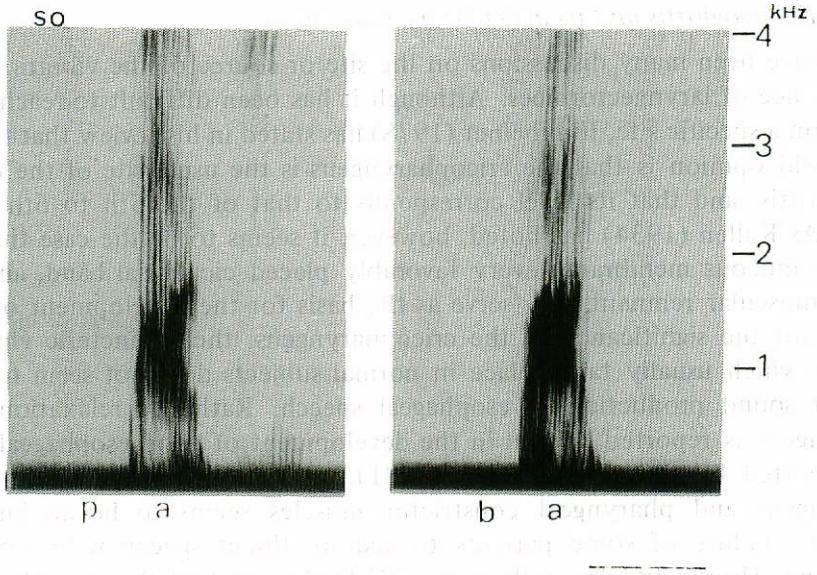
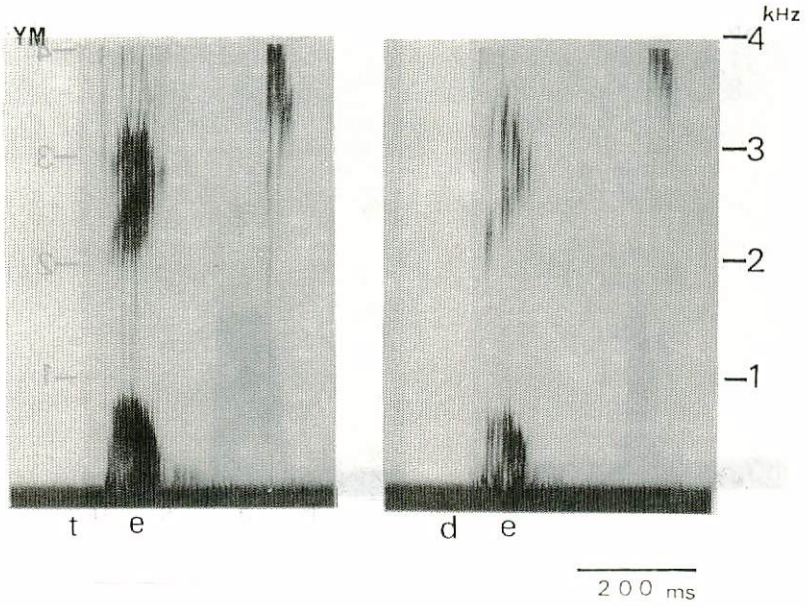


Fig. 4 *Sound spectrograms of speech samples.*
(a) *Comparison of word-initial /p/ vs. /b/ in Subject SN.*
(b) *The above comparison in Subject SO.*
(c) *Comparison of word-initial /t/ vs. /d/ in Subject YM.*

(c)



Comments

1. Site of the neoglottis and its articulatory movement

There have been many discussions on the site or source of the vibration in the substitute voice of laryngectormees. Although it has been difficult to reach general agreement on a specific site, Finkbeiner (1978) has stated in his review that the most generally held opinion is that the cricopharyngeus is the usual site of the neo- (or pseudo-) glottis, and that its level corresponds to that of the 5th to 6th cervical vertebrae. As Kallen (1934) has noted, however, it seems to be the case that every fold of the mucous membrane, every favorably placed cicatricial band, and every muscle or muscular remnant may serve as the basis for the development of a neoglottis. As for the significance of the cricopharyngeus, the sphincteric closure of this muscle which usually takes place in normal subjects does not seem to be important for sound production in esophageal speech. Rather, a relaxation of the cricopharyngeus is reported helpful in the development of good esophageal speech and, as reported by Singer and Blom (1981), an airflow-induced spasm of the cricopharyngeus and pharyngeal constrictor muscles seems to be an important factor in the failure of some patients to acquire fluent speech with esophageal voice training. Hiroto and his colleagues (1974) claimed that the cricopharyngeus should not be tightly closed but rather left unsutured at the time of a laryngectomy so as to avoid postoperative constriction of the esophageal port.

It seems reasonable to conclude that the membranous surface of the pharyngoesophagus is most important for establishing an effective vibration. As early as in 1937, Jackson and Jackson clearly stated that the requirements for substitute voice are closely approximated membranous surfaces and a moving column of air that can

be set in vibration by the membranous surface. The present study reveals that the mucosal margin of the esophageal orifice actively contributes to the voicing distinction and that vibration appears to develop at the highest level of the reconstructed esophageal passage. A similar image can be seen in the ultra high-speed movie frames taken by Rubin and cited by Finkbeiner (1978), in which "labial" opening-closing gestures of the mucosal surface around the esophageal orifice are recognizable. Sawada (1959) also postulated that the neoglottis was located at the entrance of the reconstructed esophagus based on his X-ray cinematographic study.

The mechanism of the production of voiceless sounds in esophageal speech has not been precisely determined. Sawada (1959) carried out an X-ray cinematographic observation and reported that the base of the tongue elevated postero-superiorly together with the hyoid bone during the production of voiceless sounds so as to stop the vibration of the neoglottis by closing the hypopharyngeal lumen. The present study did not confirm these findings.

There seem to be several different strategies for enhancing the perception of voiceless consonants in esophageal speech. For example, as Schwarz (1982) mentioned in his review, esophageal speakers may produce voiceless sounds in isolation through the use of tongue and lip pressure and intra-oral air in syllable-initial position, which can be followed by phonation without undue delay.

The present study indicates that "voicelessness" is achieved by the cessation of the vibration of the mucosal margin which is substantiated by the transient opening of the neoglottis, although the underlying mechanism of the apparent opening gesture is not clean as yet. Certainly, the opening can not be elicited by the normal mechanism of "abduction", but it does not seem to be controlled solely by aerodynamic factors either. There seems to be some tension control mechanism around the region of the neoglottis but further study is needed to explore its nature.

2. *Pattern of the perceptual confusion of esophageal speech and its acoustic characteristics*

There have been many reports on the results of perceptual studies on esophageal speech. Hyman (1955), for example, studied vowel and consonant production in terms of most to least intelligible and claimed that plosives were more intelligible than fricatives. His result does not conincide with that of the present study in which fricatives were more intelligible than plosives. As for the voicing distinction for consonants, Hyman simply stated that voiced-voiceless congnates were often confused.

The present study revealed that the rate of confusion for voiceless consonants was higher than that for voiced consonants. A similar result was reported by Takafuji (1960) for Japanese, and by Nichols (1976) for English.

As the acoustic cues for the voicing distinction for plosives, VOT values in syllable-initial position and the presence or absence of voicing during the closure period in syllable-medial position are generally considered to be most significant. The present study showed that these factors are also quite important for the voicing distinction for plosives in esophageal speech.

Christensen and his colleagues (1978) measured VOT values on phonetically representative syllable-initial plosives in esophageal speech and found that esophageal speakers effected a systematic variation in VOT, and that the general pattern

of the variation paralleled that observed for normal speakers. Conversely, it might be said that speech samples produced by esophageal speakers were regarded as representative when they included plosive consonants with VOT values comparable to those of normal subjects. Incidentally, only 58% of their samples were regarded as representative.

There have been many reports suggesting different perceptual cues effective for the voicing distinction other than VOT (Stevens and Klatt, 1974; Repp, 1979). Regarding the voicing distinction for Japanese plosives, Nakata (1961) reported that the characteristics of the F1 transition can facilitate the voicing distinction as an important perceptual cue. He also indicated the significance of the silent period after the stop release in the voicing distinction. The present study also suggests that these additional acoustic cues play some role in the voicing distinction in esophageal speech.

Based on the results of physiological measurements during obstruent production by hearing-impaired speakers, McGar and Löfqvist (1982) stated that a straightforward relationship between physiology and listener judgement is rather unlikely in such a complex phenomenon as the voiced-voiceless distinction. Their statement should hold true particularly in the case of pathological speech such as esophageal speech. In future studies of esophageal speech we have to take into consideration the possibility that measurements along one single dimension might not always be capable of predicting listener responses. From this viewpoint, a careful combination of physiological, acoustic and perceptual studies seems mandatory for the further analysis of esophageal speech.

Summary

A perceptual study of utterance samples obtained from 8 skilled esophageal speakers revealed that they were able to accomplish the voicing distinction successfully for oral communication. Fiberscopic observation on 4 of the 8 subjects revealed that there was a transient opening gesture for voiceless sound production at the neoglottis which was found in most cases at the highest portion of the reconstructed esophagus. It was assumed that some active tension control mechanism had developed around the neoglottis. Sound spectrographic analysis of the speech samples suggested that voice timing played the most important role in the voicing distinction for stop consonants, although some other acoustical cues should also be taken into consideration.

References

Christensen, J.M., B. Weinberg and P.J. Alfonso (1978); Productive voice onset time characteristics of esophageal speech. J. Speech Hear. Res. 21, 56-62.

Finkbeiner, E.R. (1978): Surgery and speech, the pseudoglottis and respiration in total standard laryngectomy. In Speech Rehabilitation of the Laryngectomized. (ed.) J.C. Snidecor, Charles C Thomas, Springfield, pp. 58-85.

Hiroto, I., T. Morimitsu, S. Komiyama, N. Buma, S. Ryu, H. Watanabe, G. Naruse and O. Kitamura (1974), A fundamental study of vocal rehabilitation of the laryngectomee. Otlogia Fukuoka 20, Suppl. 1 & 2, 362-363.

Hyman, M. (1979): Factors influencing intelligibility of alaryngeal speech. In Laryngectomee Rehabilitation. (ed.) R.L. Kieth and F.L. Darley, College-Hill, Houston, pp. 165-180.

Jackson, C. and C.L. Jackson (1937), The Larynx and Its Diseases. W.B. Sauders, Philadelphia.

Kallen, L.A. (1934): Vicarious vocal mechanism: The anatomy, physiology and development of speech in laryngectomized persons. Arch. Otolaryng. 15, 460-503.

McGarr, N.S. and A. Löfqvist (1982); Obstruent production by hearing-impaired speakers: Inter-articulator timing and acoustics. J. Acoust. Soc. Am. 72, 34-42.

Nakata, K. (1961); A synthetic study of Japanese speech, Quartary Report of Radio Research Laboratories, 4, 525-581.

Repp, B. (1979) Relative amplitide of aspiration noise as a voicing cue for syllable-initial stop consonants. Language and Speech 22, 173-189.

Sawada, K. (1959); An experimental study of alaryngeal speech. Oto-Rhino-Laryng. Clinic (Kyoto), 52, 42-72.

Schwartz, R.J. (1982); Articulatory adjustments in esophageal speech. Paper presented at the III World Congress of Laryngectomees, Tokyo, August, 1982.

Singer, M.I. and E.D. Blom (1981); Selective myotomy for voice restoration after total laryngectomy. Arch. Otolaryng. 107, 670-673.

Stevens, K.N. and D.H. Klatt (1974): Role of formant transitions in the voiced-voiceless distinction. J. Acoust. Soc. Am. 55, 653-659.

Takafuji, T. (1960), Clinical and physiological studies on alaryngeal speech. Proceedings of Symposium, The 12th Annual Meeting of the Japan Bronchoesophagological Society, 1-21.