The Intelligibility of Tracheoesophageal Speech: First Results

P. Jongmans[1-2], F.J.M. Hilgers [1-2], L.C.W. Pols[1], C.J. van As-Brooks[2]

[1)] Institute of Phonetic Sciences, ACLC, University of Amsterdam; [2)] Netherlands Cancer Institute

p.jongmans@nki.nl

Abstract

A total laryngectomy changes the anatomy and physiology of the vocal tract, with a most noticeable effect on speech. By applying a voice prosthesis, enabling the patient to use tracheoesophageal (TE) speech, speech is of better quality than with esophageal or electrolarynx speech, but still very deviant from laryngeal speech. Most studies on TE speech have focused on voice quality. For the understanding of the physiology of the neoglottis and for improving the results of rehabilitation, it is important to study intelligibility as well. This paper will discuss the first results of a study on TE speech intelligibility.

1. Introduction

One of the most obvious consequences of total laryngectomy is the loss of natural voice. Over the years, different methods of voice rehabilitation have been developed, the two earliest being the use of esophageal speech or an electrolarynx. From 1980 onwards voice prostheses became available, enabling the patients to use tracheoesophageal (TE) speech. The obvious advantage of TE speech is that, like normal laryngeal voicing, it is pulmonary driven, i.e. air from the lungs is used to set the tissues of the pharyngo-esophageal segment (PE segment; also called neoglottis or pseudo glottis) into vibration. This results in longer phonation time and a higher intelligibility rate compared to esophageal or electrolarynx speech [1, 2, 3, 4].

For Dutch, three pilot studies on TE speech intelligibility have been performed [5, 6, 7]. These studies rendered interesting, but mostly general results. Therefore a study has been initiated that concerns investigating intelligibility of TE speech in more detail. As we only have a marginal idea of what to expect, a pilot study has been carried out first. The purpose of this pilot study is to gain more insight in the speech production/intelligibility of tracheoesophageal (TE) speakers. As an initial step, perceptual phoneme confusions of TE speech, as judged by naïve listeners will be studied.

2. Patients and methods

2.1 Subjects

Subjects were 11 laryngectomized patients. All of them underwent a standard total laryngectomy. They were all using TE speech by means of an indwelling (Provox®) voice prosthesis. Five speakers used hands free speech, the other six used digital occlusion of the stoma. All speakers were male, with a mean age of 66.9 years (age range 44 to 78 years). Post-operation times ranged from 2;2 to 17;5 years (mean time 9;4 years). Ten patients had received irradiation. Subjects were obtained from the records of the Netherlands Cancer Institute.

2.2 Recordings

Recordings were made in a sound treated room with a Marantz CDR 770 audio cd recorder. A Sennheiser microphone was placed at a microphone-to-mouth distance of 30 cm. Before the actual recording was made, the sound level was optimally adjusted for that particular subject and a calibration signal was recorded onto cd.

2.3 Speech material

The speech material consisted of syllables, consonant clusters, nonsense sentences, intonation sentences and spontaneous speech. Only the syllables will be discussed in this paper. They had the following structures: C(onsonant)V(owel) (n=54), VCV (n=63), VC (n=33) and h-V-t (15). The syllables consist of all possible consonants and vowels in Dutch (except loan phonemes). The consonants are combined with the vowels /i/, /u/ and /a:/. The vowels (including diphthongs) are preceded by an /h/ and followed by a /t/ to minimize articulation effects.

2.4 Listening Task

The listeners consisted of 10 naïve listeners, with no prior experience with TE speech. The experiment was performed online, so that listeners could participate from their own home. The whole experiment consisted of 8 parts. For the syllables listeners listened to the speech samples and for each syllable they typed what they perceived in normal spelling, which is unambiguous in Dutch. This automatically rendered response-files that were used to investigate which phonemes are difficult to produce and which confusions are made between phonemes.

3. Results

In this section consonants and vowels will be discussed. Due to limited space we cannot give a full overview of all the results and their implications.

There are 330 responses for each consonant (3 vowel contexts x 11 speakers x 10 listeners) and 110 for vowels (1 vowel x 11 speakers x 10 listeners)

3.1 Consonants

Inter-rater reliability was measured using Cronbach's alpha for each consonant position individually. For initial position a value of .992 was found, for medial and final position a value of .996 and .998. These values indicate that inter-rater reliability is sufficiently high and that averages over the listeners can be used later on.

Table 1: Overview of the consonants used for each consonant position

Manner	initial	medial	final
plosive	p b t d k	p b t d k t j	p t k
fricative	f v s z h x ʃ	f v s z h x ʃ	f s x
nasal	m n	m n ɲ	m n
approximant	ʋ j l r	ʋ j l r	l r

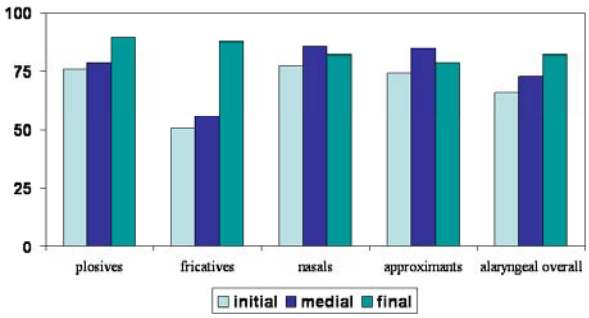

Figure 1: percentage correct score for manner of articulation in initial, medial and final position. The last cluster shows the overall total percentage correct score for each consonant position.

The overall score of 72% is markedly lower than the percentage correct score of 94% as found by Pols [8] (93.7 , 91.6 and 93.8 % for initial, medial and final position), who investigated consonant intelligibility in Dutch in nonsense CVCVC words for normal, clean, masked and reverberant speech. In our material, consistently a large range was found for each score, indicating large differences in performance between individual speakers.

Significant differences were found between manners of articulation for all consonant positions ($F(3,40)=7.350$, $p<.01$; $F(3,40)=20.297$, $p<.01$; $F(3,40)=12.303$, $p<.01$, for initial, medial and final position). A Post Hoc test (Tukey HSD) shows that the scores for fricatives are significantly lower than for plosives, nasals and approximants in initial and medial position ($p<.01$) and that nasals score lowest in final position ($p<.01$).

In our research we will particularly focus on the voiced-voiceless distinction. Confusions between these two are found frequently in every study on TE speech intelligibility that has been performed so far, especially within the plosive and fricative group [9, 10, 11, 12]. Besides this, the (in)ability to produce this contrast may teach us more about the functioning of the neoglottis and whether TE speech is myoelastic-aerodynamic or only aerodynamic.

Table 2: Overall confusion between voiced and voiceless sounds

Initial position				
stimulus	response different	response same	Missing values	Total
voiceless	623	1822	195	2640
voiced	370	2821	109	3300
Total	993	4643	304	5940
Medial position				
stimulus	response different	response same	Missing values	Total
voiceless	566	2289	115	2970
voiced	430	3134	66	3630
Total	996	5423	181	6600
Final position				
stimulus	response different	response same	Missing values	Total
voiceless	78	1861	41	1980
voiced	39	1169	112	1320
Total	117	3030	153	3300

It can be seen that in initial and medial position, voiceless sounds become voiced more often than voiced sounds become voiceless (response different). (Chi-square, $p < .01$). For final position, no significant effect was found. Note that missing values consist of deletions and sounds that were incomprehensible for the listener.

Confusion matrices were devised to be able to look more closely at the kind of confusions that were made.

	p	b	t	d	f	v	s	z	other
p	**70**	*23*	0.3						…
b	*3.9*	**78.2**	0.3	2.1					…
t	2.4	1.2	**67.3**	*22.4*					…
d		1.5	*3.6*	**87.8**					…
f					**55.8**	*33.9*	0.3		…
v					*37.9*	**48.2**	1.2	0.6	…
s						0.3	**56.1**	*31.2*	…
z					0.9	0.9	*24.2*	**60.9**	…

Figure 2: Confusion matrix for initial position. Numbers in percentages.

The confusion matrices were only made for the plosives and the fricatives that have a voiced (or voiceless) counterpart as we are especially interested in confusions occurring in this group.

In initial position, we see that for the plosives /p/ is perceived as /b/ more often than the other way round. The same is true for /t/ and /d/. For the fricatives, /s/ and /z/ follow the same pattern as the plosives, but /f/ and /v/ deviate from this pattern: /v/ is perceived more often as /f/ than /f/ as /v/. What is also apparent from the matrix is that plosives are most often confused with other plosives, and fricatives with other fricatives. This means that manner of articulation is mostly perceived correctly.

	p	b	t	d	f	v	s	z	other
p	**87**	*11.2*							…
b	*7*	**76**		1.2					…
t	3	0.3	**84**	*8.5*					…
d		1.5	*2.4*	**84.5**					…
f	1.5	0.3		0.3	**59.4**	*34.5*			…
v	1.2	0.3	0.3		*44.8*	**47**			…
s					0.9	0.3	**62.7**	*20.6*	…
z					0.6	1.5	45.2	***41.2***	…

Figure 3: Confusion matrix for medial position. Numbers in percentages

For the medial plosives we see the same pattern here as for initial position, with /p t/ being perceived as their voiced counterpart more often than the other way round. With the fricatives, we see the opposite of this, with more confusion from voiced to voiceless. Just as in initial position, we see that the responses are clustered within the same manner of articulation, but that there is slightly more variation in the responses. Though not visible in this matrix, especially /b/ and /d/ show greater variation in their responses with confusions that belong to the nasal and approximant category.

	p	t	f	s	b	d	v	z	other
p	**93**		0.3		*3.3*				…
t	5.5	**82.7**		0.3		*3.6*			…
f	0.9	0.6	**91.2**				*3*		…
s			1.2	**89.4**				*1.2*	…

Figure 3: Confusion matrix for final position. Numbers are in percentages

Even though we do not expect to find any confusions from voiceless to voiced for plosives and fricatives in final position, because the latter do not exist in Dutch, we can see in fig. 3 that they do occur in final position. As in the initial and medial position, we again find that plosives are mainly perceived as plosives, and fricatives as fricatives, but this would be clearer had the entire matrix been shown.

3.2 Vowels

Inter-rater reliability was measured using Cronbach's alpha. A value of .988 was found allowing us to use averages again.

Table :3 percentage correct score for vowels

Stimulus	Percentage correct score
Short vowels (ɪ ɛ ɑ ɔ ʏ)	80.5
Long vowels (i eː aː oː u y øː)	64
Diphthongs (ɛi œy ʌu)	87.3
Overall	74

The overall score of 74% is lower than the 84.3% score found in literature for normal speakers [13]. There is large variation in the scores between individual vowels, as can also be seen in the confusion matrix (fig. 5). Also, long vowels are misperceived significantly more often than short vowels and diphthongs ($F(2,30)=14.452$, $p<.01$; Tukey HSD, $p<.01$). Overall scores of individual speakers show a large variation in performance (66-87% correct) as well.

From the matrix, some very interesting confusions can be observed. Sometimes only length is involved as in /aː/ and /ɑ/. However, in general, a lowering of the vowels is visible, so moving to a more open articulation (when you look at the highest number of confusions for each individual vowel)
Three vowels in particular deserve more attention because of their diphthongization: /oː/ /eː/ and /øː/. /oː/ and /øː/ are also the two phonemes that have a very low intelligibility score. These three vowels are frequently perceived as /ʌu/, ɛi/ and /œy/ respectively.

4. Discussion

In this paper we focused on the perception experiment and the first results emerging from this experiment. Remarkable are the high values for inter-rater reliability. Even though TE speech sounds very different, listeners still show high agreement. Another important observation is that the overall correct scores are much lower than the scores found for normal laryngeal speakers. Even though alaryngeal speech has improved greatly with the introduction of the voice prosthesis, enabling TE speech, the intelligibility rates are still rather low and deserve detailed investigation.

We have also seen that intelligibility rates are lowest in initial position and highest in final position. This division can also be found in the literature [5, 9, 10]. The reason that final position shows the highest score may be found in the fact that there are fewer consonants possible in final position than in initial or medial position. Also, for fricatives and plosives it is mainly the feature voice that is misperceived often. As in

	aː	ɑ	ɛ	ɪ	eː	ɛi	i	ɔ	oː	ʌu	u	ʏ	øː	œy	y	ai
aː	**95.5**	4.5														
ɑ	*24.6*	**68.2**	1.8					3.6		0.9		0.9				
ɛ			**90.1**	1.8								3.6		3.6	0.9	
ɪ			8.1	**76.4**	7.3	0.9	5.5							0.9	0.9	
eː			0.9	6.4	**64.6**	*24.5*	0.9							2.7		
ɛi		0.9	0.9		2.7	**84.6**								9.1		1.8
i				4.5	*14.6*		**80**							0.9		
ɔ		*13.7*						**80**	0.9	2.7	0.9	1.8				
oː								5.5	**32.6**	*55.5*	5.5			0.9		
ʌu		0.9			0.9	*8.2*		0.9		**89.1**						
u									*9.1*	1.8	**89.1**					
ʏ			0.9	0.9	0.9							**88.2**	0.9	*5.5*	2.7	
øː					0.9							17.3	**20.9**	*54.5*	6.4	
œy	2.7	1.8	0.9			0.9				2.7		1.8	0.9	**88.3**		
y			0.9	0.9	0.9							*16.4*	6.4	9.1	**65.4**	

Figure 5: Confusion matrix for vowels. Numbers in percentages

Dutch we do not have voiced plosives and fricatives in final position, the listener will probably be less inclined to perceive a plosive or fricative as voiced. In initial and medial position, the fricatives were most often misperceived. In final position, the nasals were misperceived most. Boon-Kamma [6] showed the same results in Dutch. In American-English [10] and Spanish [11] fricatives also caused problems, just as plosives did. Problems with the feature voice are discussed in many papers on TE speech intelligibility [5, 6, 9, 10, 11, 12] and were also clearly present in this research. Most studies found confusions from voiceless to voiced, but Nord et al. [12], for example, found more confusions from voiced to voiceless. In the present research, we found a mix. The plosives in initial and medial position mainly show confusions from voiceless to voiced, but the fricatives showed a different pattern: in initial position /s/ was perceived as /z/ more than the other way round and in medial position both /s/ and /f/ were perceived as their voiced counterpart more often than the other way round. What was also apparent from the confusion matrices is that plosives are mostly perceived as plosives and fricatives as fricatives. Apparently manner of articulation for these two groups is easy to perceive, it is mostly the feature voice or in some cases the feature place of articulation that causes difficulties. Similar results are found in research performed by Pols [8] who looked at the intelligibility of consonant features and found the same clustered responses for plosives and fricatives. He also found that in the confusions for fricatives especially the feature voice played a role.

The vowels rendered some interesting results. For Dutch, Oubrie [7] has looked at the intelligibility of vowels and their formant values for TE speakers. She found an overall score correct of 49.2%. Even though our overall score of 74% is much higher, the rest of the data are very similar. She found the same types of confusions and the same diphthongization for the vowels /oː/ /eː/ and /øː/. Where the lowering of sounds is concerned, perception might be influenced by higher formant values for TE speakers. Another factor is the low F0, that we did not measure in these syllables, but we know exists in TE speech. A large difference between an F0 and a F1 will also influence the judgment of a listener in such a way that he will perceive an even higher F1 than what is actually present in the signal. These findings will need to be investigated in more detail. Oubrie [7] also gives a possible explanation for the dihthongization. /oː/ /eː/ and /øː/ lie near each other in the vowel diagram at almost the same height and also near the short vowels /ɔ/, /ɪ/ and /ʏ/. In normal speech, we see that formants change during the articulation of the vowel (so actually also slight diphthongization), which we call colouring of the formants. It might be that for TE speakers this colouring is too large, moving the formant values too strongly into the direction of /u/, /iː/ and /y/, resulting in the diphthongs /ʌu/, /ɛi/ and /œy/. An explanation for this strong formant colouring might be that TE speakers have little control over the dorsal part of their tongue.

At this point, it is still difficult to give a thorough explanation for all confusions found. The explanation might come from combining these data with the results of the planned acoustic analyses and imaging studies of the neoglottis and vocal tract, and we hope that these further investigations will reveal some useful information about the underlying causes for the intelligibility problems. The knowledge thus gained will form the basis for possible further improvement of the results of speech rehabilitation through adapted surgical and/or clinical intervention approaches.

5. Bibliography

[1]Williams, S.E. & Watson, J.B. (1987): "Speaking proficiency variations according to method of alaryngeal voicing", *Laryngoscope* 97, 737-739.

[2]Debruyne, F., Delaere, P., Wouters, J. & Uwents, P. (1994): "Acoustic analysis of tracheo-esophageal versus esophageal speech", *Journal of Laryngology and Otology* 108, 325-328.

[3]Bertino, G., Bellomo, A., Miani, C., Ferrero, F., & Staffieri, A. (1996): "Spectrographic differences between tracheo-esophageal and esophageal voice", *Folia Phoniatrica et Logopaedica* 48, 255-261.

[4]Max, L., Steurs, W. & De Bruyn, W. (1996): "Vocal capacities in esophageal and tracheoesophageal speakers", *Laryngoscope* 106, 93-96..

[5]Roeleven, S. & Polak, R. (1999): *De Verstaanbaarheid van consonanten in de Nederlandse trachea-esophageale spraak*. Bachelor's thesis, Rotterdam College of Logopedics.

[6]Boon-Kamma, B. (2001): *Verstaanbaarheid na totale laryngectomie. De verstaanbaarheid van sprekers met een Provox® 2 stemprothese*. Master's thesis, University of Amsterdam. 70 pp.

[7]Oubrie, A. (1999): *Formantwaarden en verstaanbaarheid van Nederlandse klinkers bij tracheo-esophageale sprekers*, Master's thesis, University of Nijmegen, 50 pp.

[8] Pols, L.C.W. (1983): "Three-mode principal components analysis of confusion matrices, based on the identification of Dutch consonants, under various conditions of noise and reverberation", *Speech Communication* 2, 275-293

[9]Hammarberg, B., Lundström, E. & Nord, L. (1990): "Consonant intelligibility in esophageal and tracheoesophageal speech. A progress report", *Phoniatric and Logopedic Progress Report. Department of Logopedics and Phoniatrics, Huddinge University Hospital & Karolinska Institute*, Stockholm, Sweden. Report no. 7.

[10]Doyle, P.C., Danhauer, J.L. & Reed, C.G. (1988): "Listeners' perceptions of consonants produced by esophageal and tracheoesophageal talkers", *Journal of Speech and Hearing Disorders* 53, 400-407.

[11] Miralles, J.L. & Cervera, T. (1995): "Voice intelligibility in patients who have undergone laryngectomies", *Journal of Speech and Hearing Research* 38, 564-571.

[12]Nord, L., Hammarberg, B. & Lundström, E. (1992): "Voiced-voiceless distinction in alaryngeal speech-acoustic and articulatory observations", *Phoniatric and Logopedic Progress Report. Department of Logopedics and Phoniatrics, Huddinge University Hospital & Karolinska Institute*, Stockholm, Sweden, report 2-3, 19-22.

[13]Koopmans-van Beinum, F.J. (1980): *Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions*, PhD. Thesis, University of Amsterdam.