ACOUSTIC CORRELATES OF THE VOICED-VOICELESS DISTINCTION IN DUTCH NORMAL AND TRACHEOESOPHAGEAL SPEAKERS

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

[1] Institute of Phonetic Sciences/ACLC, Universiteit van Amsterdam/ [2] Netherlands Cancer Institute
p.jongmans@nki.nl

ABSTRACT

Confusions between voiced and voiceless plosives and fricatives are the most common confusions in Dutch tracheoesophageal (TE) speech. The problem is attributed to the working of the new voice source: the pharyngo-esophageal segment, or neoglottis. In order to learn how these speakers convey the voiced-voiceless distinction, detailed analyses are necessary. 15 acoustic correlates (and a subset of 6 for the fricatives) were selected and analyzed. Statistical analyses were then used to determine which correlates are used to distinguish between voiced and voiceless sounds. The data show that TE speakers do not differ much from normal laryngeal speakers, except where voicing is concerned.

Keywords: voiced-voiceless distinction, acoustic analyses, pathological voices.

1. INTRODUCTION

Total laryngectomy is often necessary in people with laryngeal cancer. With this operation the entire larynx is removed, including the vocal folds. The upper and lower airways become disconnected, the digestive tract is re-established and the patient has to breathe through a permanent tracheostoma at the base of the neck. This operation has profound influences on the voice quality and speech intelligibility. Next to esophageal and electrolarynx speech, voice prostheses became available from 1980 onwards, enabling the patients to use prosthetic tracheoesophageal (TE) speech [9]. The advantage of TE speech over the other methods of voice rehabilitation is that like normal laryngeal (NL) voicing it is pulmonary driven, i.e. air from the lungs is used to set the tissues of the pharyngoesophageal segment into vibration allowing for longer phonation times and a higher intelligibility rate. However, when we compare this intelligibility rate with that of NL speakers, there is still a substantial difference that warrants further investigation. Several TE speech intelligibility studies exist and one of the most common confusions found was between voiced and voiceless sounds, also for Dutch [e.g. 1,3,6]. It is argued that the production of voicing is difficult for a TE speaker as it is assumed that the neoglottis is less pliable than the vocal folds and cannot be easily adjusted by will. Perceptual data seem to confirm this assumption. It is an important area to study in more detail because of its importance for word intelligibility. Knowledge about the (in) ability to produce this contrast consistently, may also teach us more about the amount of control TE speakers have over their neoglottis. To look at the voiced-voiceless distinction in more detail, acoustic analyses were performed. Several other studies [3-5,7,8] have done so by using a variety of acoustic correlates. Our study has combined these acoustic correlates and has complemented them with other acoustic correlates described in literature on normal laryngeal voicing [10]. The question we wish to answer is which acoustic correlates are used for the production of a *correct* voiced-voiceless distinction and whether TE speakers differ from NL speakers in the use or values of the correlates.

We expect that TE speakers will exaggerate certain correlates, especially segmental ones, to convey a correct voiced-voiceless distinction and that TE speakers will show more problems with actual voicing (pitch) than NL speakers.

2. ACOUSTIC ANALYSES

2.1. Patients and methods

2.1.1. Subjects

Subjects were 11 male Dutch TE speakers, all with a standard total laryngectomy and an indwelling (Provox®) voice prosthesis [2]. Mean age was 66.9 years (age range 44-78). Mean post-operation time was 9;4 years (range 2;2-17;5). Ten subjects had received irradiation. Subjects were obtained from the records of the Netherlands Cancer Insti-

tute. The study was approved by the Protocol Review Board of the Netherlands Cancer Institute.

Five male control subjects were also included, with a mean age of 56 (age range 45;9-72;3).

2.1.2. Recordings

For the TE speakers, 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. Beforehand, the sound level was optimally adjusted for each subject individually and a calibration signal was recorded onto CD. Recordings for normal speakers were made in a recording studio with a Pioneer PDR-555 RW CD recorder and a preamp Sennheiser MKH 105T microphone.

2.1.3. Speech material

The stimuli consisted of [p b t d f v s z] in medial position (VCV) with V being /i/, /u/ or /aː/. This amounts to 33 stimuli (11 speakers * 3 vowels) per consonant for the TE speakers and 15 (5 speakers * 3 vowels) for the NL speakers.

Ten naïve listeners with no prior experience with TE speech participated in the listening experiment. They typed in what they perceived in normal spelling, which is unambiguous in Dutch.

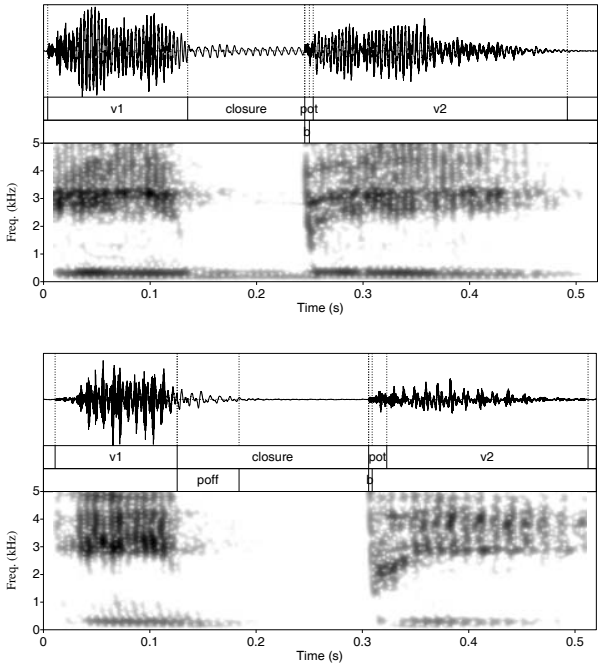
2.2. Measurements

The acoustic correlates were investigated by manually segmenting the acoustic speech signal using the program Praat [www.praat.org]. TE speech often contains noise and often lacks periodicity. This makes the segmenting task rather complicated. For that reason, both the correlates and the segmentation criteria were defined carefully (to be published elsewhere). A Praat script was used to calculate the durations of the segments measured and the various correlates. In fig. 1 examples are given of segmented TE speech signals, showing all segments investigated in the study.

V1 is the first vowel, *closure* is the closure phase of the plosive (*cd*), *pot* is the phonation onset time and *V2* is the following vowel. Phonation onset time (*pot)* is a parameter which is similar to VOT: it seems to be a first attempt at producing the following vowel, but in our case is mostly aperiodic.

On the second tier in the lower part of fig. 1, *poff* is the phonation offset after the first vowel. It

Fig 1: Examples of segmented speech signal. /ibi/ above, /ipi/ below, both correctly perceived, and produced by a TE speaker. cf. audio_file_1/2 (.wav/ .collection)



is possible that the length of *poff* may be a cue to perceive a sound as voiced or voiceless [4,5,8]. The plosive *burst* is also marked on this tier. Based on these segments, other correlates were calculated:

1. Phonation offset as percentage of the closure duration (*relpoff*)
2. Percentage of voiced frames in the closure duration (*percvoicedframes*);
3. HNR of the voiced frames in the closure (*hnrvoicedframes*);
4. HNR of the closure (*hnrcd*);
5. Relative burst intensity (*relbint*);
6. Closure + burst (*cd+durb*)

The segmentation criteria were based on literature [5,7,10], except for *poff*, where intensity was used rather than glottal pulses [8] to mark the end of phonation, and *pot* as defined here was introduced specifically for this study. Where pitch measurements were required, default parameter values in Praat were adjusted to accommodate for the deviant voices of TE speakers.

2.3. Results

2.3.1. Plosives

For the TE speakers, 126 out of a possible 132 plosives were segmented. Six realizations were omitted due to the low quality of the signal. For the NL speakers, all 60 plosives were segmented.

Our first interest in this study is a better understanding of the importance of specific acoustic features once utterances are properly understood. We therefore *only* took into account the stimuli that were perceived more than 80 percent correct by all listeners (N=52 for voiced, N=47 for voiceless plosives for TE speakers; N=30 for voiced, N=30 for voiceless plosives for NL speakers). Even though it is interesting to investigate acoustic differences between accurately and inaccurately perceived TE samples as well, too few samples that were unanimously inaccurately perceived exist, making it impossible with the current data set to investigate this issue.

A 'two-level conditional hierarchical linear model' was used with the following factors (Proc Mixed, SAS 9.1.3.):

1. Speaker type (NL vs. TE speakers)
2. Place of articulation (alveolar vs. labial)
3. Vowel type (a, i, u)
4. Voicing (voiced vs. voiceless)

Effects were found for vowel type and place of articulation, but not consistently for all correlates. Mean scores were corrected for these effects.

2.3.2. Segmental durations

In table 1 the mean lengths of the segments are given for voiced and voiceless plosives and for NL and TE speakers. These values are raw values, meaning they have not been corrected yet for vowel or place of articulation effects and are used here mainly for illustration of the results.

Except for *pot*, no significant difference was found for the speaker groups. Therefore, mean values of the two groups together were used for further analysis. For all segment durations a significant difference was found between voiced and voiceless plosives.

For *pot*, a difference between TE speakers and NL speakers was found for the voiceless plosives (longer duration for TE speakers, p<.01).

All findings were as could have been expected from the literature: *V1* and *V2* were longer for the voiced plosives, whereas the *cd*, *cd+durb*, *durb* and *pot* were longer for voiceless plosives.

Table 1: Mean durations of the voiced (V) and unvoiced (UV) segmental correlates for TE and NL speakers. Significance given per correlate and per speaker group. * indicates that a significant difference exists between TE and NL speakers for that particular duration.

Correlate	Voice	TE (sd)	Sign.	NL (sd)	Sign.
V1-ms	V UV	206 (56) 165 (47)	p<.01	189 (58) 160 (51)	p<.01
cd-ms	V UV	87 (45) 135 (59)	p<.01	99 (39) 143 (43)	p<.01
cd+durb-ms	V UV	92 (45) 149 (58)	p<.01	103 (40) 153 (44)	p<.01
durb-ms	V UV	5 (6) 14 (13)	p<.01	5 (5) 10 (8)	p<.01
pot-ms	V UV	13 (13) 33* (17)	p<.01	8 (6) 19* (9)	p<.01
V2-ms	V UV	249 (52) 219 (51)	p<.01	243 (46) 240 (55)	p<.01

Phonation offset time *poff* and *relpoff* were measured as well, but appear to be so complicated that they are left out of consideration for now.

2.3.3. Other correlates

In table 2 again the 'raw' means are given for illustration.

Significant effects for all correlates were found for the speaker groups except for *relbint*. *Relbint* is also the only correlate with an only moderately significant difference between voiced and voiceless plosives. Both TE and NL speakers showed a strong significant difference between voiced and voiceless for the other three correlates. When speaker groups are compared, it can be seen that the groups differ significantly for *percvoicframes* in voiced plosives, for *hnrvoicedframes* for voiced and voiceless plosives and for *hnrcd* for voiced and voiceless plosives, with TE speakers showing lower scores. All differences found between voiced

Table 2: Mean values of the voiced (V) and unvoiced (UV) correlates for TE and NL speakers. Significance given per correlate and per speaker group. * indicates that a significant difference exists between TE and NL speakers for that particular correlate.

Correlate	Voice	TE (sd)	Sign.	NL (sd)	Sign.
relbint-ratio	V UV	.22 (.37) .43 (.48)	p= .0105	.17 (.16) .15 (.07)	p= .0105
percvoiced frames-%	V UV	77* (32) 40 (35)	p<.01	98* (2) 43 (19)	p<.01
hnrvoiced frames-dB	V UV	7* (4) 3* (3)	p<.01	17* (6) 9* (3)	p<.01
hnrcd-dB	V UV	7* (5) 3* (4)	p<.01	17* (6) 8* (3)	p<.01

and voiceless confirm the literature: a higher burst intensity for the voiceless plosives, a higher percentage of voiced frames for the voiced plosives, a higher HNR value for the voiced frames in voiced plosives and a higher HNR in the closure for voiced plosives.

2.3.4. Fricatives

For the fricatives, *V1*, *cd* (for fricatives *consonant* duration), *V2*, *percvoicedframes*, *hnrvoicedframes* and *hnrcd* were analyzed using the same methods as for the plosives. Due to limited space, only main findings are discussed here.

Only very few fricatives were perceived more than 80 percent correct (TE: a total of 23; NL: a total of 56), which means that one has to be careful when interpreting the present results.

Significant differences between voiced and voiceless were found for *V1* (longer for voiced), *cd* (longer for voiceless) and for *percvoicedframes* (higher percentage for voiced) for both speaker groups. A significant effect for speaker group was found only for *percvoicedframes*. For *hnrcd* and *hnrvoicedframes* a difference between voiced and voiceless was found only for NL speakers.

3. DISCUSSION & CONCLUSION

In the introduction, we hypothesized that TE speakers would exaggerate (segmental) correlates to convey a voiced-voiceless contrast and that this would distinguish them from NL speakers. From the results so far, we cannot support this hypothesis as TE speakers only differed from NL speakers on phonation onset time. A likely explanation for this specific difference is that it takes TE speakers longer to start up a vowel than NL speakers due to the changed anatomy and physiology of the neoglottis and vocal tract. However, our expectation that voicing would be a problem was confirmed: TE speakers show a lower percentage voiced frames in voiced plosives than NL speakers (the high percentage of voiced frames in voiceless sounds for NL speakers (42.7%) was caused by two outliers). The correlate *hnrvoicedframes* says something about the quality of the voicing in the voiced frames. For this correlate, the speaker groups differ significantly as well: TE speakers show a lower HNR both for the voiced and voiceless plosives. Related to the voiced frames in the closure is the *hnrcd*. Voiced sounds have a better HNR, which was also found for both speaker groups, but the TE speakers perform worse than the NL speakers. These results suggest that TE speakers have more difficulty employing actual voicing (pitch) as a distinguishing correlate than NL speakers. Also the quality of the voicing is poorer than that of NL speakers. It does not mean, however, that TE speakers do not make use of these correlates to make a correct voiced/voiceless contrast. Based on the results TE speakers seem capable, to a greater or lesser extent, to employ voicing at appropriate times, at least for the plosives.

Summarizing, only the hypothesis that TE speakers have problems producing voicing could be confirmed. Contrary to expectations, TE speakers showed significant differences between voiced and voiceless plosives for all acoustic correlates and in that do not differ much from NL speakers. Further statistical analyses (e.g. CART) will be used to determine which of the acoustic correlates best predict class membership of voiced or voiceless plosives and fricatives, both for TE and NL speakers.

4. ACKNOWLEDGEMENTS

We greatly acknowledge the research grant of the "Stichting Breuning ten Cate" which makes this research possible, and Rob van Son and Ton Wempe for their technical support.

5. REFERENCES

[1] Doyle, P., Danhauer, J., Reed, C. 1988. Listeners' perceptions of consonants produced by esophageal and tracheoesophageal talkers. *J Speech Hear Disord* 53, 400-407.

[2] Hilgers, F., Schouwenburg, P. 1990. A new low-resistance, self retaining prosthesis (Provox®) for voice rehabilitation after total laryngectomy. *Laryngoscope* 100, 1202-1207.

[3] Jongmans, P., Hilgers, F., Pols, L., van As-Brooks, C. 2006. The intelligibility of tracheoesophageal speech, with an emphasis on the voiced-voiceless distinction. *Logoped. Phoniatr. Vocol*. 31, 172-181.

[4] Robbins, J., Christensen, J., Kempster, G. 1986. Characteristics of speech production after tracheoesophageal puncture: voice onset time and vowel duration. *J Speech Hear Res* 29, 499-504.

[5] Saito, M., Kinishi, M., Amatsu, M. 2000. Acoustic analyses clarify voiced-voiceless distinction in tracheoesophageal speech. *Acta Otolaryngol* 120, 771-777.

[6] Searl, J., Carpenter, M., Banta, C. 2001. Intelligibility of stops and fricatives in tracheoesophageal speech. *J Commun Disord* 34, 305-321.

[7] Searl, J., Carpenter, M. 2002. Acoustic cues to the voicing feature in tracheoesophageal speech. *J Speech Lang Hear Res* 45, 282-294.

[8] Searl, J., Ousley, T. 2004. Phonation offset in tracheoesophageal speech. *J Commun Disord* 37, 371-387.

[9] Singer, M.I., Blom, E.D. 1980. An endoscopic technique for restoration of voice after laryngectomy. *Ann Otol Rhinol Laryngol* 89, 529-533.

[10] Slis, I.H., Cohen, A. 1969. On the complex regulating the voiced-voiceless distinction I. *Language and Speech* 12, 80-102.