Clinical Linguistics & Phonetics, March 2008; 22(3): 215–231

Prosodic boundaries in alaryngeal speech

M. A. VAN ROSSUM[1], H. QUENÉ[2], & S. G. NOOTEBOOM[2]

[1]*Leiden University Medical Centre, Leiden, the Netherlands, and* [2]*Utrecht institute of Linguistics, Utrecht University, Utrecht, the Netherlands*

(Received 24 September 2007; accepted 5 December 2007)

Abstract
Alaryngeal speakers (speakers in whom the larynx has been removed) have inconsistent control over acoustic parameters such as F_0 and duration. This study investigated whether proficient tracheoesophageal and oesophageal speakers consistently convey phrase boundaries. It was further investigated if these alaryngeal speakers used the same hierarchy of acoustic boundary cues that is found in normal speakers. A perception experiment revealed that listeners identified prosodic boundaries less accurately in oesophageal speakers. Acoustic analyses showed that laryngeal speakers used pre-boundary lengthening and pitch movements at phrase boundaries, as expected. Tracheoesophageal speakers used pre-boundary-lengthening and pauses and oesophageal speakers used pauses to convey phrase boundaries. Two oesophageal speakers also paused inappropriately, within phrases. Although these two speakers differentiated between air-injection and prosodic pauses, listeners were unable to tell the two types of pauses apart. Alaryngeal speakers might benefit from therapy that specifically teaches them how to optimize their prosodic abilities.

Keywords: *Prosodic boundaries, alaryngeal speech, perception, acoustic analyses*

Introduction

The present study investigated how alaryngeal speakers convey prosodic boundaries. Alaryngeal speakers are speakers in whom the larynx has been surgically removed, usually due to laryngeal cancer. In the resulting alaryngeal speech, the muscle and mucosa at the entrance to the oesophagus function as an alternative sound source (neo-glottis). In the Netherlands, the most widely used modes of alaryngeal voicing are tracheoesophageal (TE) and oesophageal (Es). In TE voicing, pulmonary air causes the neo-glottis to vibrate. The air is shunted into the oesophagus by means of a one-way silicone-prosthesis inserted in a surgically constructed tracheoesophageal puncture. In Es voicing, air is injected from the oral cavity into the oesophageal lumen. When this air is then ejected, it causes the neo-glottis to vibrate. Although the voicing source is the same in TE and Es speech, the driving-force is different. In contrast to normal laryngeal and TE speakers, who can have a pulmonary air supply of approximately 3 litres, the air supply available to Es speakers is

Correspondence: Maya van Rossum, present address: Leiden University Medical Centre (Department of ENT and Head and Neck Surgery), PO Box 9600, 2300 RC Leiden, the Netherlands. Tel: +31 71 526 2405. Fax: +31 71 526 6910. E-mail: M.A.van_Rossum@LUMC.nl

ISSN 0269-9206 print/ISSN 1464-5076 online
DOI: 10.1080/02699200701847160

limited to small volumes of approximately 80ml (Van den Berg & Moolenaar-Bijl, 1959; Casper & Colton, 1993). The alaryngeal voicing source can be said to be a grossly controlled structure when compared to the fine-tuning capabilities of the larynx. In comparison to normal laryngeal speech, both TE and Es speech is noisy and less intelligible (Christensen & Dwyer, 1990; Miralles & Cervera, 1995), and the ability of alaryngeal speakers to realize prosodic cues such as pitch movements, or intensity and durational variations, has been described as unpredictable and inconsistent (McHenry, Reich, & Minifie, 1982; Gandour & Weinberg, 1985; Van Rossum, De Krom, Nooteboom, & Quené, 2002). Durational and pitch variations are prosodic cues especially necessary to convey prosodic boundaries.

Prosodic boundaries mark consecutive units of speech, such as phrases, clauses or sentences (Streeter, 1978; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991). The importance of prosodic boundaries is clearly illustrated when potentially ambiguous sentences need to be disambiguated (Lehiste, 1973; Scott, 1982; Lehiste, 1983; Price et al., 1991). For example in the sentence: ''John and Mary or Jim might come'', it could be that John and Mary might come, or Jim. Or John might come with either Mary or Jim. With the first option, a break after Mary is essential: ''(John and Mary) or Jim'', whereas with the second option, the break would need to be directly after John: ''(John) and Mary or Jim''. If these boundaries are not positioned appropriately, or more than one boundary exists, listeners will find it difficult to interpret the meaning of the sentence.

The acoustic cues associated with prosodic boundaries form a hierarchy (e.g. Wightman, Shattuck-Hufnagel, & Price, 1992; Shattuck-Hufnagel & Turk, 1996; Gussenhoven & Rietveld, 1992). The most ''basic'' cue seems to be pre-boundary lengthening: speakers slow down towards the end of a unit of speech. The higher the perceptual strength of a boundary, the more cues are associated with the prosodic boundary (De Pijper & Sanderman, 1994). Phrases are generally demarcated by final lengthening and boundary-marking pitch movements. In clauses or sentences, the degree of lengthening increases and pausing may be added to the boundary marking pitch movements. A number of studies have investigated the acoustic cues associated with prosodic boundaries. De Rooij (1979) found that pre-boundary lengthening alone was sufficient for the perception of a prosodic boundary, whereas F_0 on its own is not sufficient to signal the location of a boundary (De Rooij, 1979; Terken & Collier, 1992). Also, insertion of pauses in the absence of final lengthening is perceived as a dysfluency (De Rooij, 1979) and whereas boundary-marking pitch movements occur frequently without pauses, pauses are normally accompanied by pitch movements (Klatt, 1975; Price et al., 1991; Blaauw, 1994; De Pijper & Sanderman, 1994).

The question arises to what extent alaryngeal speakers are able to adhere to the hierarchy outlined above. First, voice modulation (variation of the fundamental frequency) in both types of alaryngeal speech is generally erratic, and voice range is more restricted, when compared to normal laryngeal voicing (i.e. Robbins, Fisher, Blom, & Singer, 1984; Moon & Weinberg, 1987; Gandour, Weinberg, Petty & Dardarananda, 1988; Qi & Weinberg, 1995). This may affect adequate realization of boundary marking pitch movements in both TE and Es speakers.

Second, Es speakers produce approximately seven syllables per air charge compared to approximately 19 syllables in normal speakers or TE speakers (Moolenaar-Bijl, 1951; Snidecor & Curry, 1959; Max, Steurs & De Bruyn, 1996). Es speakers might therefore be forced to pause more often because of their small supply of air. Inappropriately positioned pauses have a negative effect on speech recognition (Scharpff & Van Heuven, 1988; Nooteboom, Scharpff, & Van Heuven, 1990; Sanderman & Collier, 1997). The study by

Nooteboom et al. (1990) further illustrated that appropriate pausing is especially beneficial in speech that is of lesser quality or less intelligible than normal speech. Thus, Es speakers, whose speech quality and intelligibility might already be compromised when compared to normal speech, might not be able to position pauses appropriately and might be forced to pause within phrases (when phrases consist of a larger number of syllables). Third, Es speakers might also be limited in the amount of pre-boundary lengthening, as this further depletes the limited air supply.

Thus, TE and even more so Es speakers might not be able to adhere to the expected prosodic hierarchy because of physical limitations. However, these speakers might rely on one specific cue, or a different combination of cues than normal speakers, to convey the presence of prosodic boundaries consistently.

The research questions are as follows:

1. Can listeners identify the intended prosodic boundaries in TE and Es speech?
2. Does the length of phrases have an effect on how accurately prosodic boundaries are conveyed in Es speech?
3. Which prosodic cues do TE and Es speakers (consistently) manipulate to convey prosodic boundaries?

It is expected that the TE speaker group will use final lengthening and boundary marking pitch tunes to convey the presence of a phrase-boundary, although this group might not be able to use pitch as consistently as normal laryngeal speakers. We further expect that the Es group might not be able to convey boundaries accurately, especially when the phrases in an utterance are long. The limited air supply might influence consistent use of final lengthening, and might cause inappropriate within-phrase pauses.

Method

Speakers

Nine speakers participated. Three laryngeal speakers produced the stimulus sentences in voiced (LV) as well as whispered mode (LW). In this way, we included a condition in which speakers could consistently manipulate duration (final lengthening) and pitch (boundary-marking pitch movements), and a condition in which speakers could consistently manipulate duration (final lengthening). The laryngeal speakers functioned as controls.

Three tracheoesophageal speakers and three oesophageal speakers participated. All speakers were male. Alaryngeal speakers were proficient speakers. This judgement was based on an evaluation procedure developed by Bors, Wicherlink, Schutte, and Mahieu (1986), which includes voice quality, articulation, voice modulation and dynamics, speaking rate and fluency. For oesophageal speakers, these criteria also specify that the production of more than five syllables per air injection is considered ''good''. Participating oesophageal speakers all used the air injection technique. In the Netherlands, oesophageal speakers are generally taught to use the injection technique, which combines the glosso-pharyngeal press method with the production of plosives (Moolenaar-Bijl, 1951, 1953; Van den Berg & Moolenaar-Bijl, 1959). With this method, oesophageal speakers rely on plosives to re-inflate the oesophageal reservoir with air: ''…air, which is subsequently used for phonation, is brought into the oesophagus by means of the articulation of certain consonants'' (Molenaar-Bijl, 1951, p. 21). The advantage of this method is its supposed unobtrusiveness, because re-inflating the oesophagus coincides (blends) with the articulation of plosives. The assumption is that this results in fluent phrasing. Since

plosives can be found at regular intervals in speech, the air supply can be ''recharged'' without disrupting the natural flow of speech. However, in utterances with no plosives, oesophageal speakers' phrasing might be jeopardized, because the injections become separate phonetic events instead of being integrated in a speech sound (Moolenaar-Bijl, 1951, 1953; Van den Berg & Moolenaar-Bijl, 1959), and might be perceived by a listener as a pause or phrase boundary. The training strategy through which oesophageal speakers are taught the injection method is to gradually increase the complexity of words. To start with, monosyllabic words containing a voiceless plosive in the initial position (e.g. pit), then polysyllabic words and phrases containing plosives (e.g. paperclip), and eventually production of polysyllabic words and phrases containing no plosives, is acquired (e.g. Miami).

Table I gives relevant information per speaker.

Stimulus material

There were nine sentences. The stimulus sentences were the same as those used by Lehiste (1983), but with different proper nouns (which will be explained below). These sentences were chosen, firstly because of their potential syntactic ambiguity which rendered them suitable to study prosodic effects (Beach, 1991), and secondly because speakers and listeners are known to disambiguate these type of sentences easily and precisely (Streeter, 1978; Scott, 1982; Lehiste, 1983).

The stimulus sentences consisted of two or more alternative groupings of noun phrases within the main noun phrase, depending on the conjunction that was used. The conjunctions *of* (''or'') and *en* (''and'') occurred. Sentences containing ''or'' had two possible groupings per sentence (see 1a and 1b or 2a and 2b), and sentences containing ''and'' had three possible groupings per sentence (see 3a, 3b, 3c). N stands for the first name:

1a. Ik zou (N1 en N2), of N3 uitnodigen;
I would invite (N1 and N2), or N3. *or*

Table I. Relevant information of speakers participating in this study. Speaker group abbreviations: L=laryngeal; TE=tracheoesophageal; Es=oesophageal. n.a.=not applicable

group	speaker	age at recording	time lapsed since operation (yr;mth)	type of surgery	radiation: primary or post-op	average number of syllables per injection
L	1	64	n.a.	n.a.	n.a.	n.a.
	2	61	n.a.	n.a.	n.a.	n.a.
	3	57	n.a.	n.a.	n.a.	n.a.
TE	1	58	3;7	total laryngectomy	primary	n.a.
	2	55	5;1	total laryngectomy+unilateral neck dissection	post-op	n.a.
	3	54	4;6	total laryngectomy	primary	n.a.
Es	1	67	9	total laryngectomy+unilateral neck dissection	primary	6
	2	59	6;4	total laryngectomy+bilateral neck dissection	primary	7
	3	56	5;11	total laryngectomy+unilateral neck dissection	primary	9

1b. Ik zou N1, en (N2 of N3) uitnodigen;
I would invite N1, and (N2 or N3).
2a. Ik zou N1, of (N2 en N3) uitnodigen;
I would invite N1, or (N2 and N3). *or*
2b. Ik zou (N1 of N2) en N3 uitnodigen;
I would invite (N1 or N2) and N3.
3a. Ik zou N1, en **(**N2 en N3**)**, en **(**N4 en N5**)** uitnodigen;
I would invite N1, and (N2 and N3), and (N4 and N5). *or*
3b. Ik zou (N1 en N2), en N3, en **(**N4 en N5**)** uitnodigen;
I would invite (N1 and N2), and N3, and (N4 and N5) *or*
3c. Ik zou (N1 en N2), en (N3 en N4), en N5 uitnodigen;
I would invite (N1 and N2), and (N3 and N4), and N5.

The sentences containing ''or'' had a version in which ''or'' was positioned after the first name and a version in which ''or'' was positioned after the second name. Two versions of ''or'' sentences, as well as ''and'' sentences (in which five proper nouns occurred) were included to increase the number of items that occurred in phrase-initial as well as pre-boundary position, thus increasing the number of items that could be included in the acoustic analyses (see below).

Fifteen different first names were included. These names varied in complexity (including or excluding plosives) and length (monosyllabic or polysyllabic). It is surmised that oesophageal speakers using the injection method will experience no difficulties when phrases contain plosives (the length of the phrases should not be an issue), but, as was mentioned above, they might find it difficult to convey proper phrasing when phrases do not contain plosives. Thus, in this study, the oesophageal speakers were tested to the utmost of their ability, because the stimulus sentences in the present study reflected this increasing degree of complexity:

1. Sentences with monosyllabic first names, containing plosives (e.g. Dutch first/proper names, *Kees*, *Toos*), which resulted in three syllables per phrase.
2. Sentences with polysyllabic names but still containing plosives (e.g. *Patricia*, *Catharina*), which resulted in eight or nine syllables per phrase, depending on the names.
3. Sentences with polysyllabic names but not containing any plosives (e.g. *Annemarie*, *Josefien*), which resulted in seven or eight syllables per phrase, but without the advantage of plosives.

In total, there were 21 stimulus sentences:

((2 ''of'' versions $\times$ 2 bracketings)+(1 ''en'' version $\times$ 3 bracketings)) $\times$ 3 levels of complexity.

Recording procedure

The audio recordings were made in a quiet room. Speakers were first instructed to read the sentences quietly. The sentences were presented to the speakers on paper. Brackets illustrated the different structural versions of the sentences (see sentences 1a, 1b, 2a, 2b, 3a, 3b, 3c). Speakers were made aware of the ambiguity of the sentences, because speakers only make active use of prosodic cues when they are aware of the different possible interpretations of a sentence (Lehiste, 1973). It was explained that the speaker's aim should be to disambiguate the different versions of the sentences, using whichever cues the speakers regarded as necessary. *Without* these instructions, it would not have been possible

to conclude whether speakers could not, or would not disambiguate the sentences, and the purpose of this study was to determine the speakers' *ability* to convey prosodic boundaries. Speakers practiced the sentences before actual recording. When speakers substituted the order of names or mispronounced names, they were asked to repeat the sentence.

Perception experiment

Listeners. Twenty-seven native speakers of Dutch between the ages of 18 to 27 participated. No-one had any known hearing deficiencies. None were familiar with alaryngeal speech. Listeners were paid for their participation.

Procedure. A binary forced-choice identification task was used in this perception experiment, because it was felt that this more closely resembled real-life conversation: a listener will attempt to interpret the speaker's intention. Thus, when a listener is confronted with these ambiguous utterances, a listener will have to choose between two possibilities. For each spoken utterance, a choice was given between two written response possibilities. The ''or'' utterances could be matched with one of two differently bracketed sentences (see sentences 1a, 1b and 2a, 2b). However, the ''and'' utterances could be matched with three differently bracketed sentences (see sentence 3a, b, c). Therefore, each ''and'' utterance was presented twice, so that the listener had the opportunity to choose among all three written possibilities. For example, spoken utterance (3b) had written sentence versions (3a) and (3b) as response possibilities in trial one, and written sentence versions (3b) and (3c) as response possibilities in a trial two, etc. The computer program used to present the stimulus material, allowed the trials to be presented in random order. This resulted in two listener judgements for each ''and'' utterance, and one listener judgement for each ''or'' utterance.

Listeners were seated in a sound-treated booth. A speaker's *spoken utterance* (e.g. 1a, spoken, see above) was presented over headphones, and two versions of the corresponding *written sentence* (e.g. 1b and 1a, see above) were represented as text buttons on the computer screen in front of them. The listeners were asked to identify the way in which the speaker combined the names in the sentence into pairs. In other words, listeners chose which of the two written versions (1b or 1a, see above) was heard, by matching the utterance (audio) with one of two differently bracketed versions of the sentence (visual orthography). The chosen version was selected by clicking the appropriate text button. The correct answer was not represented by one specific text button; sometimes the utterance would match one text button, sometimes the other text button. Listeners were instructed to guess when uncertain.

In total, listeners had to judge 360 utterances {([3TE+3Es+3 LV+3LW] $\times$ 12 ''or'' utterances)+([3TE+3Es+3 LV+3LW] $\times$ (2 $\times$ 9 ''en'' utterances)}. Listeners needed, on average, 1 hour 40 minutes to complete the experiment.

Acoustic analyses

Stimulus material. As a result of the bracketing, names were positioned phrase-initially in one version and phrase-finally in another version of an utterance. Per speaker, 15 names occurred in phrase-initial and phrase-final position. Measurements were made on these phrase-initial and phrase-final names, and then compared. In total, 360 names were analysed.

Pre-boundary lengthening. The domain of pre-boundary lengthening is the final, pre-boundary syllable, but if the final syllable of a test name contained a schwa, the syllable preceding the final syllable was included in the measurement (Cambier-Langeveld, Nespor, & Van Heuven 1997). Durations of the test names' final syllable were measured in milliseconds. Segmentation was based on combined audio-visual (oscillographic and spectrographic) information, according to criteria given by Van Zanten, Damen, and Van Houten (1991).

Fundamental frequency. F_0 was determined using the auto-correlation pitch-detection algorithm (Boersma, 1993). Subsequent F_0-contours were re-synthesized by means of the PSOLA-analysis by synthesis technique (Moulines & Laroche, 1995), with the sole purpose of quickly tracing and manually correcting faults introduced by the pitch detection algorithm: a comparison of the pitch contour, with the (lack of) periodicity observed in the oscillogram, revealed any faults introduced by the pitch detection algorithm. In the present study, perceptual and visual inspection of the speech signal revealed that speakers mostly used rising tunes, and occasionally falling tunes, but never level tunes. F_0-excursions were therefore measured within the final syllable of the test names (or, if the final syllable contained a schwa, the syllable preceding the final syllable was included in the measurement). The distance (measured in Hz) between the F_0-maximum and F_0-minimum in the final syllable was expressed in semitones.

Pauses. Two types of pauses were differentiated: appropriate and inappropriate pauses. For example, in the utterance: ''Ik zou N1 en **(**N2 en N3**)**, en **(**N4 en N5**)** uitnodigen'', a pause might follow the phrase-final test names N3 or N5. Similarly, a pause might precede the phrase-initial test names N2 or N4. In contrast, pauses within an intended phrase (preceding N3 and N5 or following N2 and N4), were deemed inappropriate (because they would be positioned within a phrase). The durations of silent intervals (absence of amplitude in oscillogram) between words following or preceding the test names, and the test names themselves were measured in milliseconds (e.g. between the end of N5 and the start of ''uitnodigen''). Perceptual and auditory judgements of the utterances revealed that Es speakers actually used the prosodically motivated pause to inject air. Thus, at the points where prosodic pauses occurred, air injections also occurred. Speakers might therefore differentiate between prosodically motivated pauses and air injection pauses by controlling the duration of the silent interval. In other words, the difference between a prosodic pause and an injection pause might not be the presence or absence of an injection, but the duration of the silent interval.

In summary, pre-boundary lengthening, F_0-excursions and pausing were measured, using Praat (Boersma & Weenink, 1998).

Results

Perception experiment

The first research question was: ''Can listeners identify the intended prosodic boundaries in tracheoesophageal and oesophageal speech?'' The speakers' utterances were presented in a perception experiment, and listeners were asked to identify how the utterances were phrased.

Task consequences

As mentioned above (Method, Task) there were two trials for the ''and'' utterances. There was no significant difference between the responses of the listeners on the two ''and'' trials (t (214)=.246, p=.806). We therefore only included the first ''and'' trial in further analyses, so that the number of listener judgements for the ''or'' and ''and'' utterances was equal.

Results per speaker group

Table II gives, per speaker group, the average percentage of utterances that listeners identified correctly.

In the laryngeal (voiced and whispered) as well as the TE speaker groups, listeners accurately identified how the utterances were phrased (99%, 97% and 96% respectively). The average percentage correctly identified utterances for the oesophageal group, although lower (80%), indicates that listeners still identified the intended phrasing quite often and better than chance. However, the variation in the oesophageal group was much larger than in the other speaker groups. Thus, the oesophageal group seems to differ from the other groups and phrase length might have played a role, as stated in the second research question: ''Does the length of a phrase affect prosodic boundaries in Es speech?'' The results are therefore broken down by complexity and given in Figure 1 for each speaker group.

As expected, only the oesophageal group conveyed the intended phrasing less accurately. All phrases containing polysyllabic names were identified less accurately, regardless of whether plosives were present in the phrases. In contrast to findings of previous studies on injection-oesophageal speech (Moolenaar-Bijl, 1951, 1953; Van den Berg & Moolenaar-Bijl, 1959) the presence of plosives apparently did not aid fluent phrasing in the Es speakers' utterances. It seemed that the length of the phrases (in number of syllables) had a negative effect on how accurately phrasing was conveyed by these speakers.

Because the results were in the form of percentages, they were transformed using the arcsine transformation (Studebaker, 1985). Percentage of correctly identified phrasing was then entered into analyses of variance on ''speakers'' and ''sentences''. ''Speaker groups'' (LV, LW, TE, Es) and ''level of complexity'' (monosyllabic containing plosives, polysyllabic containing plosives, polysyllabic without plosives; nested, under ''sentences'') were fixed factors; ''sentences'' and ''speakers'' (nested under ''groups'') were random factors.

The effect of ''speaker group'' did not reach significance ($F_{1\ (3,6)}=26.6$, $p<.001$; $F_{2\ (3,8)}=2.99$, $p=.096$; min $F'_{\ (3,5)}=2{,}68$, $p>.05$), possibly because there were too few sentences or too few speakers per group. The interaction between ''level of complexity'' and ''speaker group'' was significant ($F_{2\ (6,16)}=4.4$, $p=.008$), as illustrated in Figure 1. Furthermore, the main effect of ''level of complexity'' was significant ($F_{1\ (6,216)}=2.72$,

Table II. Average percentage of correctly identified phrasing, broken down per speaker group. Averaged over 3 speakers per group, $\times 21$ utterances per speaker, $\times 27$ listener judgements.

Speaker group	n	accuracy in % (standard error)
Laryngeal Voiced	1701	99 ± 2.8
Laryngeal Whispered	1701	97 ± 5.7
Tracheoesophageal	1701	96 ± 15.4
Oesophageal	1701	80 ± 29

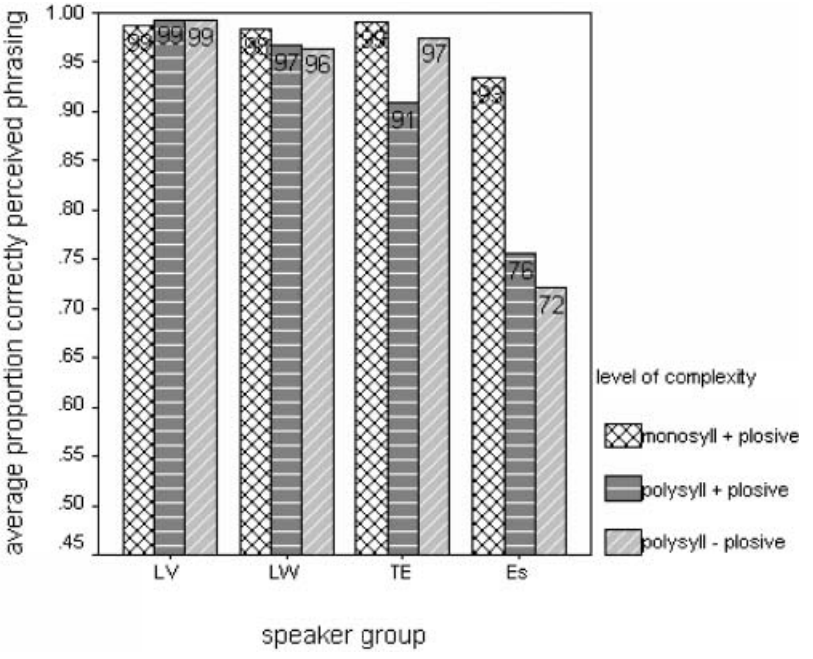

Figure 1. Phrasing identification: average proportion of correctly identified utterances given for each speaker group and for each level of complexity, pooled over sentences (7) and listeners (27).

$p=.015$; $F_{2\ (2,16)}=8.41$, $p=.003$). Post-hoc analysis confirmed that the utterances with monosyllabic names containing plosives differed significantly from both the utterances with polysyllabic names containing plosives and the utterances with polysyllabic names without plosives (Tukey's HSD, $p<.05$), as illustrated in Figure 1.

The effect of "speaker within speaker group" was also significant ($F_{2\ (8,216)}=16.4$, $p<.001$). This was somewhat unexpected, because only proficient speakers had been selected. Table II showed that there was considerable variation within the oesophageal speaker group, and although this variation was ascribed to the different levels of complexity, it might additionally have been caused by individual differences among the oesophageal speakers.

Because the effect of level of complexity is associated with the oesophageal group and we suspect differences among speakers in this group, the levels of complexity are presented for each oesophageal speaker in Figure 2.

Figure 2 confirmed that there were indeed differences among the oesophageal speakers. Es1 accurately conveyed the intended phrasing, regardless of the number of syllables per phrase or the absence of plosives. In fact, Es1's results were comparable to those for speakers in the other groups. The pattern for the other oesophageal speakers more closely mirrored the expectations as given in the Introduction: Es2 and Es3 could not disambiguate utterances that contained longer phrases caused by the polysyllabic names.

Acoustic analysis

The third research question was: "Which prosodic cues do tracheoesophageal and oesophageal speakers consistently manipulate to convey prosodic boundaries?" The

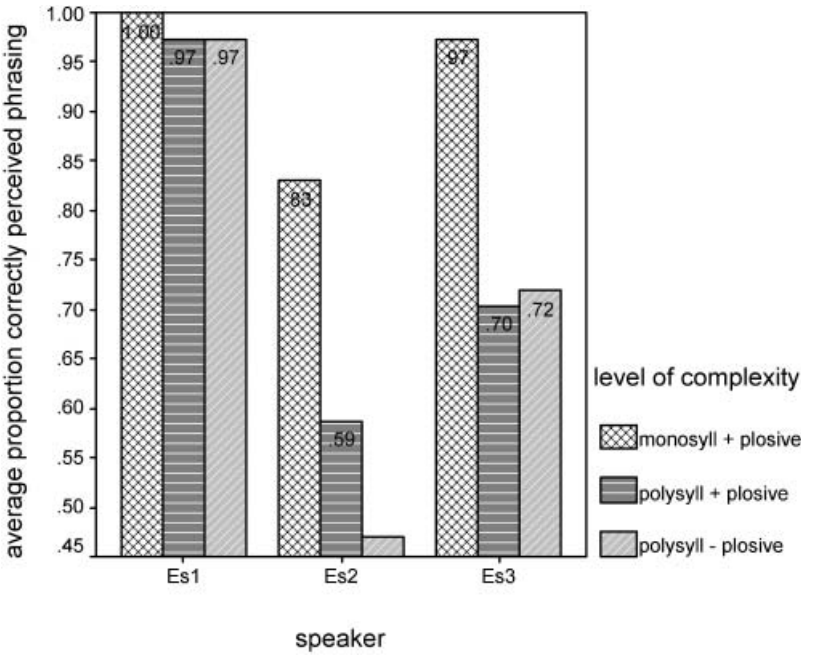

Figure 2. Phrasing identification: average proportion of correctly identified utterances, per oesophageal speaker, for the different levels of complexity.

speakers' utterances were analysed and the consistency with which acoustic cues were used to convey prosodic boundaries was determined.

Determining consistency

As mentioned in the Introduction, alaryngeal speakers' ability to produce prosodic cues tends to be unpredictable and inconsistent. Hence, calculating and comparing the averages of the two conditions (pre-boundary values versus phrase-initial values), does not necessarily indicate if a speaker uses a cue consistently (for the interested reader, the average values are also given per group, in appendix A). The consistency with which speaker groups produced acoustic cues was therefore investigated.

First, the difference between pre-boundary values and phrase-initial values was calculated for each individual name (resulting in a total of 15 ''difference'' values per speaker, one for each name). Second, the differences had to be perceptually meaningful (in other words, the differences should be large enough for listeners to perceive). The criteria according to which a difference was deemed perceptually meaningful were as follows. For final lengthening to count as a difference, a just noticeable difference of 10% was used (Klatt, 1976). Thus, the pre-boundary syllable had to be at least 10% longer in duration than its phrase-initial counterpart. For a boundary-marking pitch tune to count as a difference, the F_0-excursion in the phrase-final syllable had to be at least 1.5 semitones larger than the F_0-excursion in the phrase-initial syllable (Rietveld & Gussenhoven, 1985). No restriction was placed on the duration of pauses, because the Es speakers' injections might have been preceded by a short pause.

Third, based on these criteria, we totalled the number of times that a cue was perceptually meaningful. Thus, the consistency with which speakers differentiated between pre-boundary names and their phrase-initial counterparts could be determined. The Binomial Test was used to determine if the number of perceptually meaningful differences was significant. If the Binomial Test was significant, the occurrence of a cue was taken to be consistent; the higher the level of significance the greater the consistency.

Differences between speaker groups

Results for the groups are presented in Table III.

Final lengthening of the pre-boundary syllable was the most consistently used cue in the laryngeal group when voicing and when whispering. The laryngeal group, when voicing, also consistently produced boundary marking pitch tunes. The TE group consistently realized pre-boundary lengthening and post-phrase pausing (after the phrase-final name).

As a group, the Es speakers seemed to be especially consistent in their use of pausing. Visual and auditory inspection of the speech signal revealed that all Es speakers also, always, injected air when pausing. Whenever pauses were situated at prosodic boundaries, oesophageal speakers used these boundary pauses to inject air. However, they also paused consistently after the first name *within* a phrase (second column from the right, Table III; see also explanation under Method, Acoustic Analyses, Pauses).

Differences among Es speakers

The Perception experiment revealed that there were differences among the oesophageal speakers. Results for individual Es speakers are given in Table IV.

For Es1, post-phrase pausing was used the most consistently, followed by pre-boundary pitch tunes. Post-phrase pausing was also the most consistently used cue for Es2, followed by pre-phrase pausing. However, this speaker also paused consistently within phrases, although pauses after the first name within a phrase occurred more consistently than pauses before the last name within a phrase (two right-most columns of table 4: inappropriate pauses). Es3 used both pre-phrase and post-phrase pausing consistently to convey the prosodic boundaries. This speaker also paused within phrases, but not as consistently as Es2. Overall, especially Es2 and Es3 added inappropriately positioned injection pauses.

Table III. Per group, consistency with which different acoustic cues were used to convey prosodic boundaries (according to Binomial Test: the greater the significance, the greater the consistency).

group	final lengthening	F_0 excursion	pausing: pre-phrase	pausing: post-phrase	inappropriate within-phrase pausing 1	inappropriate within-phrase pausing 2
LV	***	**	ns	ns	0	0
LW	***	n.a.	ns	ns	0	0
TE	***	ns	ns	***	0	0
Es	ns	ns	***	***	***	ns

LV=laryngeal voiced; LW=l aryngeal whispered; TE=tracheoesophageal; Es=oesophageal. 1=pause after first name within phrase; 2=pause before last name within phrase; *=p<.05; **=p<.01; ***=p<.001: significant according to Binomial Test; ns=not significant; n.a.=not applicable; 0=did not occur.

Table IV. Per speaker (Es speaker group), consistency with which different acoustic cues were used to convey prosodic boundaries (according to Binomial Test: the greater the significance, the greater the consistency).

speaker	final lengthening	F_0 excursion	pausing: pre-phrase	pausing: post-phrase	inappropriate within-phrase pausing 1	inappropriate within-phrase pausing 2
Es1	ns	*	ns	***	ns	ns
Es2	ns	ns	*	***	***	*
Es3	ns	ns	***	***	**	ns

Es=oesophageal; 1=pause after first name within phrase; 2=pause before last name within phrase; *=$p<.05$; **=$p<.01$; ***=$p\leqslant.001$: significant according to Binomial Test; ns=not significant.

Effect of phrase length in oesophageal speech

The results of the Perception experiment indicated that the Es speakers conveyed phrasing less accurately when the phrases were longer (containing polysyllabic names). This might indicate that the production of inappropriate within-phrase pauses occurred more frequently in longer phrases. To investigate if this is indeed true, the results of the Es group for the various complexity levels are given in Table V.

The results for the different levels of complexity, when pooled over the Es speakers, showed that the greater the level of complexity, the more consistently pauses were used. In the utterances with monosyllabic names containing plosives (thus, the shortest phrases), only post-phrase pauses occurred consistently. In the utterances with polysyllabic names containing plosives (thus, longer phrases, but with the opportunity to inject air unobtrusively), post-phrase as well as, to a lesser extent, pre-phrase pausing was used to convey prosodic boundaries. In the utterances with polysyllabic names not containing plosives (thus, longer phrases, but without the opportunity to inject air unobtrusively), post-phrase and pre-phrase pausing occurred equally consistently. With regard to inappropriate within-phrase pausing, the column second from the right in Table V reveals that the consistency with which inappropriate pauses occurred also increased as the level of complexity increased. This result was in line with the idea that oesophageal speakers would find it more difficult to inject air unobtrusively when no plosives were available.

Table V. Per level of complexity (pooled over Es speakers), consistency with which different acoustic cues were used to convey prosodic boundaries (according to Binomial Test: the greater the significance, the greater the consistency).

level of complexity	final lengthening	F_0 excursion	pausing: pre-phrase	pausing: post-phrase	inappropriate within-phrase pausing 1	inappropriate within-phrase pausing 2
mono. +plos.	ns	ns	ns	***	ns	ns
poly +plos	ns	ns	**	***	*	ns
poly −plos	ns	ns	***	***	***	ns

Es=oesophageal; mono=monosyllabic; poly=polysyllabic; +plos=phrases with names containing plosives; −plos=phrases with names NOT containing plosives; 1=pause after first name within phrase; 2=pause before last name within phrase; *=$p<.05$; **=$p<.01$; ***=$p\leqslant.001$: significant according to Binomial Test; ns=not significant.

Differentiating between ''injection'' and ''prosodic'' pauses

Because of the consistent use of inappropriate pauses (especially ES2 and Es3), Es speakers might differentiate between (appropriate) prosodic pauses and pauses necessary for air injection. Thus, the prosodic pauses were expected to be significantly longer in duration than the air injection pauses. This was indeed the case (Wilcoxon $z=-1.989$, $p=.047$): prosodic pauses were 443ms on average (standard error=22) compared to 286 milliseconds on average (standard error=17) for air injection pauses. However, this difference was much more marked in short phrases with plosives, where prosodic pauses were on average 150% longer in duration than injection pauses, but note that the number of inappropriate pauses was much lower (not reaching significance, as can be seen in Table V). In longer phrases containing plosives, prosodic pauses were 43% longer than injection pauses and in longer phrases without plosives prosodic pauses were 28% longer.

General discussion

This study investigated whether TE and Es speakers are able to convey prosodic boundaries. The Perception experiment generally showed that the TE group conveyed the intended boundaries very consistently, whereas, as a group, the Es speakers conveyed the intended boundaries less consistently when the phrases became longer. However, one Es speaker managed to convey the prosodic boundaries as consistently as the LV, LW and TE speakers. Before attempting to explain this phenomenon, we shall discuss a question which was also posed in the Introduction, namely, to what extent the different speaker groups adhered to the hierarchy of acoustic cues outlined in the Introduction. An hierarchical prosodic organization implies that speakers and listeners rely on the same cues in the speech signal to determine the strength of a boundary. The stronger the boundary is the greater the number and strength of the cues will be (e.g. Klatt, 1975; Price et al., 1991; De Pijper & Sanderman, 1994; Shattuck-Hufnagel & Turk, 1996). In the present study, laryngeal speakers, when voicing, used final lengthening as well as pre-boundary pitch movements, as would be expected at noun phrase boundaries (e.g. Blaauw, 1994). Thus, the hierarchical prosodic organization was adhered to. When whispering, the laryngeal group relied on pre-boundary lengthening only and did not compensate for the lack of voicing by adding an alternative cue. Final lengthening on its own was sufficient to convey the presence of a prosodic boundary, a result that is in agreement with the findings of De Rooij (1979) and Lehiste (1983).

TE speakers consistently used final lengthening, but they did not consistently manipulate the pitch cue normally associated with this level of prosodic boundary. Instead, these speakers consistently used pausing, which is more often associated with a sentence or clause boundary. It was mentioned in the Introduction that, although pre-boundary pitch movements frequently occur without pauses, pauses are normally accompanied by pre-boundary pitch movements (Wightman et al., 1992; De Pijper & Sanderman, 1994). The TE speakers' strategy therefore seems at variance with the accepted prosodic hierarchy as explained above. One possible explanation is that TE speakers might have used pausing as a prosodic cue to compensate for their loss of ''normal'' speech quality. This would correspond to Nooteboom (1985) who suggested that the introduction of grammatical pauses helps to maintain intelligibility.

The Es speakers as a group relied on pausing to convey prosodic boundaries. Final lengthening seems to have been substituted with pausing. This, again, deviates from the cues normally associated with phrase boundaries and might have been a compensatory

strategy. The limited air-supply associated with oesophageal speech might have prohibited consistent lengthening of the pre-boundary syllable, especially in longer phrases. The inability to use final lengthening is unfortunate, because final lengthening allows a gradual slowing down without which speech sounds jerky and unnatural (De Rooij, 1979). Furthermore, the absence of pre-boundary lengthening in this speaker group violates the expected temporal pattern, and might therefore have an adverse effect on listener's speech processing (i.e. Nooteboom, 1973; Nakatani & Schaffer, 1978).

The alaryngeal speakers in the present study were all proficient speakers. We therefore conclude that, whereas normal laryngeal speakers apply different prosodic cues in an hierarchically ordered fashion, a majority of alaryngeal speakers might not (be able to) use prosody to indicate whether a boundary is meant to be major or minor. Thus, loss of the normal voicing source also leads to loss of prosodic output, although the effect seems greater for Es speakers than for TE speakers.

The differences within the Es group seem to be quite large. It was mentioned above that one Es speaker conveyed prosodic boundaries as consistently as laryngeal and TE speakers. This Es speaker consistently combined pre-boundary pitch movements and pauses, but hardly ever paused inappropriately within a phrase. The other two Es speakers used only pausing consistently, but also consistently paused *inappropriately* within-phrases. This might indicate that the opportunities available (through the presence of plosives and boundary pauses) were insufficient to recharge the oesophageal air supply and that additional ''injection pauses'' were needed—at inappropriate places within a phrase—to recharge the oesophageal air supply. However, the average number of syllables per injection during conversational speech, as given in Table I, is slightly lower for the first Es speaker than for the other two speakers. Moreover, the average number of syllables per injection for the other two Es speakers as given in Table I suggests that only the longest phrases, consisting of eight to nine syllables, should have caused inappropriate within-phrase pauses. It therefore seems that different strategies were used. The first speaker might have compressed his speech, thus ''forcing'' out complete phrases on one injection. Although this seemed to benefit the listeners when having to identify prosodic boundaries, this strategy might have compromised intelligibility (too many syllables produced on too little air). The other two speakers seemed to pause at regular intervals, regardless of the required phrasing. It has been shown that normal speakers sometimes position pauses so that the number of words between pauses is balanced (Klatt, 1976; Gee & Grosjean, 1983). Two of the Es speakers in the present study might have adopted a similar strategy, which might have resulted in the high number of inappropriate within-phrase pauses. This strategy might seem odd, because speakers were explicitly instructed to concentrate on proper phrasing. However, these speakers did differentiate between pauses at prosodic boundaries and injection pauses by increasing the silent interval of the prosodic pauses. Unfortunately, the Perception experiment showed that listeners were not able to differentiate between the two types of pauses, suggesting that this compensatory strategy was not effective.

Given these findings, we conclude that the TE and Es speakers participating in this study attempted to compensate for their physical limitations. However, these compensatory strategies were not necessarily optimal. The ineffective pausing strategy of two Es speakers illustrated that speakers may not have been aware of what listeners needed. Furthermore, speakers' attempts might not have been as optimal as physically possible, as illustrated by the fact that none of the TE speakers varied pitch *consistently*, whereas one Es speaker did. Evidently, one can not assume that physical limitations dictate the speech ability, but neither can one expect that optimal compensation will take place automatically. This

underscores the importance of speech rehabilitation for alaryngeal speakers. Speakers need to be made aware of listeners' needs and should be explicitly taught how to optimize their prosodic and other speech abilities. Speech language pathologists need to be aware of which cues are available for boundary-signalling, and the possibility of cue-trading for intelligibility. Furthermore, knowledge of what constitutes appropriate and inappropriate pausing, and the effect of inappropriate pausing in the absence of final lengthening should be considered. Es speakers need to be actively taught how phrasing is conveyed within the context of a longer utterance, and not only how to convey short phrases in isolation. The sentences used in the present paper might therefore be suitable as training material.

To conclude, this study investigated prosodic boundaries in alaryngeal speech, thus shedding some light on the effect physical limitations may have on language use and communication, and giving insight into the effectiveness of alaryngeal speakers' compensatory strategies.

Acknowledgements

We thank Marika Voerman (Leiden University Medical Center) for assisting us in securing subjects to participate in this study. We further thank Vincent van Heuven and a number of anonymous reviewers for their valuable comments, as well as Frans Hilgers for his comments on the final draft.

References

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: evidence for cue trading relations. *Journal of Memory and Language*, *30*, 644–663.

Blaauw, E. (1994). The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication*, *14*, 359–375.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings*, *17*, 97–110.

Boersma, P., & Weenink, D. (1998). Praat: a system for doing phonetics by computer (version 3.7). Institute of Phonetic Sciences, University of Amsterdam, *Report*, *132*, available at: http://www.praat.org.

Bors, E. F. M., Wicherlink, W. H., Schutte, H. K., & Mahieu, H. F. (1986). Evaluatie Esophagusstem. *Logopedie en Foniatrie*, *58*, 230–234.

Cambier-Langeveld, G. M., Nespor, M., & Van Heuven, V. (1997). The domain of final lengthening in production and perception in Dutch. ESCA. *Eurospeech Proceedings* 931–935.

Casper, J. K., & Colton, R. H. (1993). *Clinical manual for laryngectomy and head and neck cancer rehabilitation*. San Diego, CA: Singular.

Christensen, M. J., & Dwyer, P. E. (1990). Improving alaryngeal speech intelligibility. *Journal of Communication Disorders*, *23*, 445–451.

De Pijper J., R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, *96*(4), 2037–2047.

De Rooij J., J. (1979). *Speech punctuation. An acoustic and perceptual study of some aspects of speech prosody in Dutch*, Doctoral Thesis. Utrecht: University of Utrecht.

Gandour, J., & Weinberg, B. (1985). Production of speech melody and contrastive stress in oesophageal and tracheoesophageal speech. *Journal of Phonetics*, *13*, 83–85.

Gandour, J., Weinberg, B., Petty, S. H., & Dardarananda, R. (1988). Tone in Thai alaryngeal speech. *Journal of Speech and Hearing Disorders*, *53*, 23–29.

Gee, J. P., & Grosjean, F. (1983). Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, *15*, 411–458.

Gussenhoven, C., & Rietveld, A. C. M. (1992). Transcription of Dutch intonation-courseware available at: http://lands.let.kun.nl/todi.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*, 129–140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*, 1208–1222.

Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, *7*, 107–122.

Lehiste, I. (1983). Signalling of syntactic structure in whispered speech. *Folia Linguistica*, *17*, 239–245.

Max, L., Steurs, W., & De Bruyn, W. (1996). Vocal capacities in oesophageal and tracheoesophageal speakers. *Laryngoscope*, *106*, 93–96.

McHenry, M., Reich, A., & Minifie, F. (1982). Acoustical characteristics of intended syllabic stress in excellent oesophageal speakers. *Journal of Speech and Hearing Research*, *25*, 564–573.

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

Moolenaar-Bijl, A. (1951). Some data on speech without larynx. *Folia Phoniatrica et Logopaedica*, *3*, 20–24.

Moolenaar-Bijl, A. (1953). The importance of certain consonants in oesophageal voice after laryngectomy. *Annals of Otology, Rhinology & Laryngology*, *62*, 979–989.

Moon, J. B., & Weinberg, B. (1987). Aerodynamic and myoelastic contributions to tracheoesophageal voice production. *Journal of Speech and Hearing Research*, *30*, 387–395.

Moulines, E., & Laroche, J. (1995). Non-parametric techniques for pitch scale and time-scale modification of speech. *Speech Communication*, *16*(2), 175–205.

Nakatani, L. H., & Schaffer, J. A. (1978). Hearing ''words'' without words. *Journal of the Acoustical Society of America*, *63*, 234–245.

Nooteboom, S. G. (1973). The perceptual reality of some prosodic durations. *Journal of Phonetics*, *1*, 25–45.

Nooteboom, S. G. (1985). A functional view of prosodic timing. In J. A. Michon, & J. L. Jackson (Eds.), *Time, mind and behavior*. New York: Springer-Verlag.

Nooteboom, S. G., Scharpff, P., & Van Heuven, V. J. (1990). Effects of several pause strategies on the recognizability of words in synthetic speech. *International Conference on spoken language processing*, Kobe, Japan, Vol. 1 385–387.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, *90*(6), 2956–2970.

Qi, Y., & Weinberg, B. (1995). Characteristics of voicing source waveforms produced by oesophageal and tracheoesophageal speakers. *Journal of Speech and Hearing Research*, *38*, 536–548.

Robbins, J., Fisher, H. B., Blom, E. C., & Singer, M. I. (1984). A comparitive acoustic study of normal, oesophageal and tracheoesophageal speech production. *Journal of Speech and Hearing Disorders*, *49*, 202–210.

Rietveld, A. C. M., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics*, *13*, 299–308.

Sanderman, A. A., & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech*, *40*, 391–409.

Scharpff, P., & Van Heuven, V. J. (1988). Effects of pause insertion on the intelligibility of low quality speech. *Proceedings of the 7th FASE Symposium*, Edinburgh 261–268.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America*, *71*(4), 996–1007.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*(2), 193–247.

Snidecor, J. C., & Curry, E. T. (1959). Temporal and pitch aspects of superior oesophageal speech. *Annals of Otology, Rhinology and Laryngology*, *68*, 1–14.

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, *64*(6), 1582–1592.

Studebaker, G. A. (1985). A rationalized arcsine transform. *Journal of Speech and Hearing Research*, *28*, 455–462.

Terken, J. M. B., & Collier, R. (1992). Syntactic influences on prosody. In Y. Tokhura, E. Vatikiotis-Bateson, & Y. Sagisaki (Eds.), *Speech Perception, Production and Linguistic Structure*. Amsterdam: IOS.

Van den Berg, J., & Moolenaar-Bijl, A. J. (1959). Crico-pharyngeal sphincter, pitch, intensity and fluency in oesophageal speech. *Pract. Oto-rhino-laryngology*, *21*, 298–315.

Van Rossum, M. A., De Krom, G., Nooteboom, S. G., & Quené, H. (2002). ''Pitch'' accent in alaryngeal speech. *Journal of Speech, Language and Hearing Research*, *45*, 1106–1118.

Van Zanten, E., Damen, L., & van Houten, E. (1991). The ASSP speech database. *SPIN/ASSP-report*, *41*, Speech Technology foundation, Utrecht, the Netherlands.

Wightman, C. W., Shattuck-Hufnagel, S., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *91*(3), 1707–1717.

Appendix

Average values and standard error (in parenthesis) of different prosodic cues, for the pre-boundary condition and the phrase-initial condition; given per speaker group (LV=laryngeal voiced; LW=laryngeal whispered; TE=tracheoesophageal; Es=oesophageal); n.a.=not applicable; ms=milliseconds; st=semitones.

Speaker Group	Lengthening (ms)		F0-excursion (st)		Pausing (ms)	
	Pre-boundary	Phrase-initial	Pre-boundary	Phrase-initial	Appropriate	Inappropriate
LV	390±13	244±7	6±.5	1±.3	288±24	0
LW	406±11	267±8	n.a.	n.a.	345±34	0
TE	417±11	309±12	3±.5	1±.2	470±26	0
Es	394±13	364±13	4±.6	2±.4	443±22	286±17