

*Prosody in Alaryngeal
Speech*

Maya van Rossum

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6006
fax: +31 30 253 6000
e-mail: lot@let.uu.nl
<http://www.lot.let.uu.nl/>

This dissertation was co-sponsored by ATOS-medical

Cover illustration: Photo of Protea, Northern Drakensberg, South Africa,
taken by D. van Rossum.

ISBN 90-76864-74-8
NUR 632

Copyright © 2005: Maya van Rossum. All rights reserved.

Prosody in Alaryngeal Speech

Prosodie in alaryngeale spraak

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de Rector Magnificus, Prof. Dr. W.H. Gispen
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op woensdag 15 juni 2005
des ochtends te 10.30 uur

door

Maartje Adriana van Rossum
Geboren op 29 oktober 1965 te Maassluis

Promotor: Prof. Dr. S.G. Nootboom
Copromotor: Dr. H. Quené

CONTENTS

ACKNOWLEDGEMENTS	v
1. GENERAL INTRODUCTION	1
1.1 <i>Speech Communication</i>	2
1.2 <i>Laryngeal and Alaryngeal Voice Production</i>	3
1.2.1 Laryngeal Voice Production	3
1.2.2 Alaryngeal Voice Production	5
1.3 <i>Alaryngeal Speech Quality and Intelligibility</i>	7
1.3.1 Speech Quality	8
1.3.2 Intelligibility	9
1.4 <i>Functions of Prosody</i>	10
1.4.1 Focus	11
1.4.2 Phrasing	13
1.5 <i>Prosody in Alaryngeal Speech</i>	15
1.6 <i>General Research Hypotheses</i>	17
1.7 <i>Outline of the Present Dissertation</i>	19
2. PITCH ACCENT	21
2.1 <i>Introduction</i>	22
2.2 <i>General Method</i>	24
2.2.1 Speakers	24
2.2.2 Stimulus material	25
2.2.3 Procedure	26
2.3 <i>Accent Perception Experiment</i>	26
2.3.1 Method	26
2.3.1.1 Stimulus material	26
2.3.1.2 Listeners	26
2.3.1.3 Procedure	27
2.3.2 Results	27
2.4 <i>Acoustic Attributes of Accent</i>	30
2.4.1 Method	30
2.4.1.1 F0 movement	30
2.4.1.2 Peak intensity (in dB)	31
2.4.1.3 Spectral tilt	31
2.4.1.4 Word duration	32
2.4.1.5 Pauses	32
2.4.1.6 Consistency with which cues are used	32
2.4.2 Results	33
2.4.2.1 F0 measurability	33
2.4.2.2 Consistency with which cues were manipulated	35
2.4.2.3 Pauses	37
2.4.2.4 Trade-off between F0 and other cues	38

2.5 <i>Perception of Alternative Pitch</i>	40
2.5.1 Method	40
2.5.1.1 Stimulus material	40
2.5.1.2 Listeners	40
2.5.1.3 Procedure	41
2.5.1.4 Analysis	41
2.5.2 Results	41
2.6 <i>General Discussion</i>	42
3. PERCEPTION OF SPEECH MELODY IN SPEECH WITHOUT F0	47
3.1 <i>Introduction</i>	48
3.2 <i>Eliciting non-F0 Speech Melodies to be used in Three Perception Experiments</i>	51
3.2.1 Method	51
3.2.1.1 Speakers	51
3.2.1.2 Validity of speaker selection	52
3.2.1.3 Stimulus sentences	54
3.2.1.4 Construction of reference utterances	55
3.2.1.4.1 Production of Reference Utterances	55
3.2.1.4.2 Evaluation of Reference Utterances	55
3.2.1.5 Feasibility of imitation task	56
3.2.1.6 Procedure of imitation task	57
3.3 <i>Perception of non-F0 Speech Melodies</i>	57
3.3.1 Method	58
3.3.1.1 Stimulus material	58
3.3.1.2 Listeners	58
3.3.1.3 Procedure	58
3.3.2 Results	59
3.4 <i>Comparing non-F0 and F0 Speech Melody</i>	61
3.4.1 Method	62
3.4.1.1 Stimulus material	62
3.4.1.2 Procedure	63
3.4.2 Results	63
3.5 <i>Transcription of Speech Melody</i>	65
3.5.1 Method	65
3.5.1.1 Stimulus material	65
3.5.1.2 Transcribers	65
3.5.1.3 Transcription task	65
3.5.1.4 Order of transcriptions	66
3.5.2 Results	66
3.5.2.1 Transcriber agreement	67
3.5.2.2 Confusion matrices	67
3.6 <i>General Discussion</i>	69

4. IN SEARCH OF NON-F0 PITCH	73
4.1 Introduction	74
4.2 Imitation Experiment	76
4.2.1 Method	77
4.2.1.1 Speakers	77
4.2.1.2 Stimulus material	77
4.2.1.3 Participants	79
4.2.1.4 Procedure	79
4.2.1.5 Transcription of imitated utterances	79
4.2.2 Results	80
4.3 Perception Experiment	85
4.3.1 Method	86
4.3.1.1 Stimulus material	86
4.3.1.1.1 Excised Stretches of Speech	86
4.3.1.1.2 Construction of Band-filtered Sounds	86
4.3.1.2 Experimental design	86
4.3.1.3 Listeners	87
4.3.1.4 Procedure	87
4.3.2 Results of Unfiltered Speech Fragments	87
4.3.3 Results of Band-filtered Sounds	90
4.4 Acoustic Analyses	93
4.4.1 Method	94
4.4.1.1 Material	94
4.4.1.2 Analyses	94
4.4.1.3 F0 excursion	94
4.4.1.4 High frequency intensity	95
4.4.1.5 Spectral tilt	96
4.4.2 Results	96
4.4.2.1 Fundamental frequency	96
4.4.2.2 High frequency intensity	100
4.4.2.3 Spectral tilt	100
4.5 General Discussion	102
5. PROSODIC BOUNDARIES	105
5.1 Introduction	106
5.2 General Method	109
5.2.1 Stimulus material	109
5.2.2 Speakers	111
5.2.3 Recording procedure	112
5.3 Perception Experiment	112
5.3.1 Method	113
5.3.1.1 Stimulus material	113
5.3.1.2 Listeners	113
5.3.1.3 Procedure	113
5.3.1.4 Design	113
5.3.2 Results	114

<i>5.4 Acoustic Analyses</i>	<i>117</i>
5.4.1 Method	118
5.4.1.1 Stimulus material	118
5.4.1.2 Analyses	119
5.4.1.2.1 Final Lengthening	119
5.4.1.2.2 Pauses	119
5.4.1.2.3 Fundamental Frequency	119
5.4.1.3 Comparison between pre-boundary syllable and its phrase-initial counterpart	120
5.4.2 Results	121
<i>5.5 General Discussion</i>	<i>124</i>
6. FINAL DISCUSSION	129
<i>6.1 Introduction</i>	<i>130</i>
<i>6.2 Main Findings and Conclusions</i>	<i>130</i>
<i>6.3 Linguistic Implications</i>	<i>133</i>
<i>6.4 Clinical Implications</i>	<i>135</i>
<i>6.5 Limitations and Suggestions for Further Research</i>	<i>137</i>
REFERENCES	141
APPENDICES	151
1. Stimulus material used in Chapter 2	151
2. Average values of cues to accent	153
3. Stimulus material used in Chapter 3	154
4. Control experiment: Intonation bias	156
5. Stimulus material used in Chapter 5	158
6. Average values of boundary cues	159
7. Suggestions that might improve alaryngeal speakers' prosodic ability	160
SAMENVATTING (SUMMARY IN DUTCH)	161
CURRICULUM VITAE	165

ACKNOWLEDGEMENTS

That this dissertation actually exists is due to following people:

Sieb Nooteboom, Hugo Quené
Guus de Krom

Participating Speakers and Listeners

Professors Hilgers, Van Heuven,
Schutte, Dejonckere
Doctors Langeveld, Maassen

Peter Pabon, Theo Veenker, Bert Schouten, Gerrit Bloothoof
Johanneke Caspers, Marc Swerts

Marika Voerman, Corina van As

Esther Janse, Brigit van der Pas,
Elise de Bree, Carien Wilsenach

THANK YOU!!

1

General Introduction

ABSTRACT

This research project investigates to what extent production of prosody in alaryngeal speakers is similar to that in normal, laryngeal speakers, and if listeners perceive prosody in alaryngeal speech as accurately as they perceive prosody in normal, laryngeal speech. This chapter provides background information on laryngeal and alaryngeal voicing. It also provides an overview of the literature on relevant topics such as alaryngeal speech quality and intelligibility, and prosody in laryngeal and alaryngeal speech. This information forms the basis of the general research hypotheses that are given towards the end of this chapter.

1.1 SPEECH COMMUNICATION

This dissertation investigates an important aspect of speech communication, viz. prosody in alaryngeal speakers.

Speech communication involves a speaker and a listener. Generally, it is assumed that the purpose of the speaker is to convey a message to the listener. The speaker has an idea that he would like to communicate, and uses speech to do so. The process of speaking can be said to consist of three layers: a plan that determines what will be said (the concept), a programme that generates how it will be said (the linguistic structure) and the performance, which is the actual execution of the programme (cf. Cohen, 1968; Levelt, 1989).

This dissertation concentrates on the performance layer, which will be described here in slightly more detail. A speaker's performance is the product of a number of factors. Performance is directed by the concept to be communicated. Phonological, morphological, syntactic and prosodic rules determine how this concept will be presented in an utterance. Speakers further tune their performance according to communicative and situational demands: according to Lindblom (1990), speakers are expected to vary their speech along a continuum of hyper- and hypospeech, and the speaker's adaptations along this continuum reflect his awareness of the listener's needs. In other words, the speaker anticipates what a listener might need, and provides the necessary information (Nooteboom, 1983; Lindblom, 1990). This is achieved by estimating the redundancy of a message, and the auditory quality of an utterance (Nooteboom, 1985). Context may determine (lack of) redundancy, whereas auditory quality is usually related to external factors, such as the distance between speaker and listener, or environmental noise, or a possible hearing deficit in the listener.

The emphasis from here onwards will be on the role of prosody within the performance layer. The main carriers of prosody are duration, loudness and pitch. Prosody forms one of the organizational layers that structure an utterance, and prosodic characteristics are further adapted to compensate for 'interfering' factors. When introducing a new concept to a listener, a speaker highlights the important words by varying pitch. When communication takes place in a noisy environment, the speaker will produce greater pitch variations, speak more loudly and slowly, and pause more often.

In normal speech communication, performance, and prosody in particular, is generally accomplished through a well-controlled, well-functioning larynx.

The focus of this dissertation, however is on speakers who do not have a larynx. The aim of this dissertation is to investigate to what extent production of prosody in alaryngeal speakers is similar to that in normal speakers, and if listeners perceive prosody in alaryngeal speech as accurately as they perceive prosody in normal speech.

In comparison to the larynx, the control that alaryngeal speakers have over the *alaryngeal* voice source is less predictable. Furthermore, the alaryngeal voice quality is known to be poor. However, there are more similarities than differences between laryngeal and alaryngeal speakers. Both laryngeal and alaryngeal speakers have the same communicative purpose: to successfully convey a message. Both have access to, and are guided by the same linguistic rules. Both have a vocal tract to shape speech sounds. Both have a voice source that may be controlled to a greater or lesser degree.

The main topic of this dissertation will be clarified in more detail before specific research hypotheses will be formulated towards the end of this chapter. In section 1.2, laryngeal and alaryngeal voice production are explained. In section 1.3, the effects that alaryngeal voice production has on speech quality and intelligibility are described. The role and functions of prosody in speech communication are reviewed in more detail in section 1.4 to explain why control over prosodic features may be especially important in alaryngeal speech. In section 1.5, a review is given of available studies on prosody in alaryngeal speech. In section 1.6, the general hypotheses for this dissertation are formulated, and the outline of this thesis is described in section 1.7.

1.2 LARYNGEAL AND ALARYNGEAL VOICE PRODUCTION

1.2.1 LARYNGEAL VOICE PRODUCTION

Normal, laryngeal voice production is extensively described in many excellent texts (e.g., Titze, 1994; Sataloff, 1997; Prater & Swift, 1984; Lieberman, 1977). The present section only includes the most essential information that is necessary to understand the difference between laryngeal and alaryngeal voice production, and it is therefore a summary of information selected from the academic texts mentioned above.

The larynx extends from the trachea (windpipe) inferiorly, to the root of the tongue superiorly, and is made up of cartilages, joints, ligaments and membranes, muscles and nerves. Although all of these are essential, only the most important structures directly involved in voice production will be mentioned.

The hyoid bone (superior) is a point of attachment for a number of laryngeal muscles and ligaments and is essential for all laryngeal functions. The thyroid cartilage is the largest laryngeal cartilage and the thyroarytenoid muscles, which form the body of the vocal folds, run from the thyroid at the front to the arytenoid cartilages at the rear. The space between the vocal folds is known as the glottis. Voice production is dependent on the finely balanced relationship between the forces exerted by the intrinsic muscles of the larynx and the force exerted by the airflow as it passes through the glottis. The intrinsic adductor muscles cause the vocal folds to be appropriately approximated, while the outflowing air accelerates as it moves through the increasingly narrow glottis. Because the velocity of the air between the folds increases, there is a concomitant decrease in the air pressure between the folds. This phenomenon is known as the Bernoulli effect, and this causes the vocal folds to be drawn together. After the vocal folds have completely blocked the airway, subglottal pressure increases again until it is sufficient to blow open the vocal folds. The lateral, outward movement of the vocal folds continues until the natural elasticity of the tissue pulls the vocal folds back inwards to their original, closed position. Then, the cycle begins again. The properties of the larynx in fact allow a quasi-periodic vibration of the vocal folds. Speakers have active control over the elastic properties of the vocal folds: by regulating the activity of different laryngeal muscles the length and stiffness of the vocal folds is altered. Thus, the frequency of vibration (fundamental frequency, hence F_0) can be changed volitionally. Voice production is therefore a combination of aerodynamic forces (lung pressure and Bernoulli effect) and the elastic properties of the vocal folds, and this combination is also known as the myoelastic-aerodynamic model of phonation (Van den Berg, 1958). This model has since been expanded to depict the wave-like motion of the vocal folds, which actually move from bottom to top, with the bottom edge leading the way. Models of phonation are increasingly becoming more realistic, such as the 16-mass model of Titze (1994), but these will not be discussed here. It is sufficient to realize that the tension and length of the vocal folds are

continually altered to produce the intended F0 during speech and that this variation in F0 is perceived as pitch variation.

The intensity of phonation can also be varied. Raising the pressure of the air supply effectively increases the amount of air that is pushed past the vocal folds in each phonation cycle. This, combined with an increased resistance to the airflow caused by the degree and duration of vocal fold closure, leads to some passive increase in pitch, but primarily it leads to an increased vocal intensity and a flatter spectral tilt. This is perceived as increased loudness (e.g. Moore, 1989).

Duration of words or speech sounds is dependent on an air supply, in that sounds can only be lengthened if there is a sufficient amount of air to allow sound production of a (prolonged) word or speech sound.

This brief description reveals that voice production is complex. The fine-tuning capabilities of the voice are most clearly demonstrated in singers, who can volitionally produce the intended tone at the desired pitch, loudness and duration.

1.2.2 ALARYNGEAL VOICE PRODUCTION

Laryngectomy – usually necessitated by laryngeal cancer – involves the surgical removal of the entire larynx, including the first two or three tracheal rings (inferiorly) and the hyoid bone (superiorly). This also causes the respiratory tract to be separated from the vocal tract, including the oral and nasal cavities. Consequently, breathing now occurs via the tracheostoma, an opening that is created by attaching the trachea to the skin in the neck.

The alaryngeal speakers participating in the present research project were tracheoesophageal (TE) and esophageal (Es) speakers. Therefore, only the voice and speech characteristics associated with these speakers will be discussed.

The new voice source (neo-glottis) is situated at the entrance to the esophagus (foodpipe): it is formed by the same structures as the upper esophageal sphincter. Thus, the source of vibration is composed of mucosa and musculature that is normally present in this area, such as the cricopharyngeal muscle and the constrictor pharyngeus muscles (Van Weissenbruch, 1996).

Similar to normal voicing, vibration of the neo-glottis in *tracheoesophageal* voicing relies on the airflow from the lungs. The air from the lungs is shunted into the esophagus by means of a silicone-prosthesis

inserted in a surgically constructed tracheoesophageal puncture and causes the neo-glottis to vibrate.

In *esophageal* voicing, air is injected from the oral cavity into the esophagus, thus insufflating the esophagus beneath the neo-glottis. This injected air is then released, and causes the neo-glottis to vibrate. The injection technique is described as an uninterrupted movement, merging the injection with the initial phoneme of the following word (Damsté, 1958). In contrast to normal laryngeal and tracheoesophageal speakers who can have an air supply of approximately 3 liters, the air supply available to esophageal speakers is limited to small volumes of approximately 80 milliliters (Van den Berg & Moolenaar-Bijl, 1959; Casper & Colton, 1993).

Alaryngeal voice production is, in principle, comparable to laryngeal voice production, because both rely on the combination of a driving force and vibrating tissue. Van As (2001) extensively investigated voice source factors that influence vibration and voice quality. Van As found substantial variability in the anatomical and morphological characteristics of the neo-glottis. Vibration could be in an anterior-posterior direction, but could also be left-to-right, or include more sides of the neo-glottis. The location of the vibration, as well as the tonicity of the neo-glottis influenced voice quality. A circular or side-to-side shaped neo-glottis, displaying optimal closure (not too tight or too slack) resulted in the best voice quality. A wave-like motion similar to that of the vocal folds could be seen in approximately half of the participating tracheoesophageal speakers. Given the variability of the neo-glottis described above, it is not surprising that highly variable voice source waveforms were found in both tracheoesophageal and esophageal speakers (Qi & Weinberg, 1995), in contrast to laryngeal waveforms that tend to be homogeneous (Hirano & Bless, 1993).

It is not entirely clear whether the rate of vibration of the neo-glottis is a result only of the aerodynamic forces, or whether there is a myoelastic component as well. Moon and Weinberg (1987) found that some tracheoesophageal speakers volitionally adjusted their voice source to modulate F_0 . However, this volitional adjustment was not consistent, as it is in the laryngeal voice source. Moon and Weinberg therefore concluded that F_0 modulation was consistently mediated on an aerodynamic basis, whereas the contribution of myoelastic adjustment of the neo-glottis greatly varied between and within speakers. Periodicity was indeed observed more often in tracheoesophageal than in esophageal voicing, because the more efficient driving force of the pulmonic air supply results in a more regular and stable

vibration of the neo-glottis (e.g., Debruyne, Delaere, Wouters & Uwents, 1994; Bertino, Bellomo, Miani, Ferrero & Staffieri, 1996). In general, F0 is however less consistent in esophageal *and* tracheoesophageal speakers than in normal laryngeal speakers (e.g., Gandour & Weinberg, 1985; Robbins, Fisher, Blom & Singer, 1984; Qi & Weinberg, 1995).

In comparison to normal laryngeal phonation, higher pressure is also needed to initiate and sustain vibrations in the neo-glottis, as measured in higher sub neo-glottic pressures and higher resistance values in the neo-glottis (Weinberg, Horii, Blom & Singer, 1982; Moon & Weinberg, 1987).

Alaryngeal speakers increase intensity by increasing the airflow and air pressure (Robbins, et al., 1984; Moon & Weinberg, 1987). In comparison to normal laryngeal speech, the average intensity in tracheoesophageal speech tends to be higher, and in esophageal speech lower (e.g., Robbins, et al., 1984).

Esophageal speakers further differ from tracheoesophageal and laryngeal speakers with regard to timing. Not only is the maximum phonation time shorter, but also the number of syllables produced per phrase is far less in esophageal speakers (e.g., Robbins, 1984).

Compared to the larynx, the alaryngeal voice source can at best be described as a grossly controlled structure. The voice sound that is produced, although it may contain the same features as the laryngeal voice, also has different acoustic characteristics, which influence speech quality and intelligibility, among other properties. Because both speech quality and intelligibility have an effect on speech communication, these will be discussed in the next section.

1.3 ALARYNGEAL SPEECH QUALITY AND INTELLIGIBILITY

In section 1.2 it was shown that the control of the alaryngeal voice source may not be as consistent as control of the larynx. It therefore seems reasonable that the auditory quality of alaryngeal speech, will be negatively affected. Prosody plays a more important role when the auditory quality of speech, and intelligibility of speech is less than normal, as will be explained in section 1.4. It is therefore important to know if alaryngeal speech quality is indeed poorer than normal speech quality. In this section an overview will be given of some studies that investigated quality and intelligibility in alaryngeal speech. Alaryngeal speech quality will be discussed first, and thereafter alaryngeal speech intelligibility.

1.3.1 SPEECH QUALITY

Many studies have investigated quality of alaryngeal speech, by evaluating acoustic and perceptual characteristics, and comparing these to normal speech. First, some studies investigating acoustic properties are mentioned, second, perceptual characteristics are described, and third we will look briefly at a study that related perceptual characteristics to acoustic properties in tracheoesophageal speech.

Robbins et al. (1984), found that variability of F0 and intensity was greatest in esophageal speakers, while both esophageal and tracheoesophageal groups had greater variability than normal speakers. The esophageal speakers also obtained the most deviant perturbation measures such as harmonics-to-noise ratio, and percentage of shimmer and jitter, but perturbation measures in both alaryngeal groups were again more deviant than those of normal speakers. Esophageal speakers further had the slowest speaking rate. Pindzola and Cain (1989) similarly found that the esophageal group produced the least number of words, and perturbation measures were significantly poorer for esophageal speakers than for normal speakers. F0, intensity and duration measures were more tightly related in esophageal speakers than in tracheoesophageal speakers (Max, Steurs & De Bruyn, 1996). In comparison to normal speakers, tracheoesophageal and esophageal speakers showed flatter long term average spectra with a relatively higher level of energy above 4 kHz (Debruyne, et al., 1994), which was related to a greater amount of turbulent noise. Weinberg, Horii and Smith (1980), and Qi and Weinberg (1991) also found flattened spectra in alaryngeal speech, although the spectra were not as flattened as in whispered speech.

Thus, acoustically, alaryngeal speech is generally noisier than normal speech, and features such as F0 and intensity are less stable. Esophageal speakers further seem to be at a greater disadvantage than tracheoesophageal speakers. Presumably, the acoustic characteristics will be mirrored in how listeners judge alaryngeal quality of speech.

Naïve listeners rated tracheoesophageal speech quality as highly acceptable (Tardy-Mitzell, Andrews & Bowman, 1985). However, in a different perceptual study by Williams and Watson (1987), tracheoesophageal and esophageal speech were rated more poorly than normal speech, on parameters such as quality and noise, intelligibility, pitch and speaking rate. Nieboer, De Graaf and Schutte (1988) found that

tracheoesophageal speakers were judged more favourably than esophageal speakers on perceptual characteristics such as “smoothness”, “briskness”, “intelligibility”, “quickness” and “expressiveness”.

Van As (2001) investigated perceptual *and* acoustic characteristics, but only in tracheoesophageal speech. The perceptual ratings as well as acoustic measures revealed great variability among speakers. Tracheoesophageal speech in general was judged to be deviant. Tracheoesophageal speakers could be divided into three different voice quality subgroups (good, moderate and poor), and these voice quality subgroups could be related to a number of acoustic features: the F0 standard deviation, the proportion-of-voicing, the amount of high frequency energy, and harmonics-to-noise all showed more favourable values in the better voice quality group. The acoustic and perceptual variability among speakers was related to the large anatomical and morphological variations of the neo-glottis (this was also shown by Damsté, 1958). For example, some tracheoesophageal speakers with a morphologically favourable neo-glottis were judged as having good speech quality, and also displayed stable F0 and little perturbation. In contrast, some tracheoesophageal speakers that displayed a more variable speech signal, or were judged as poor speakers, also showed a morphologically less favourable neo-glottis.

Generally, these studies illustrate that alaryngeal speech quality may indeed deviate from normal speech quality as expected, but they also show the variability that exists among speakers.

1.3.2 INTELLIGIBILITY

Intelligibility in alaryngeal speech has been investigated in different ways, and the results of a few key studies are summarized below.

Tracheoesophageal speakers were rated as being more intelligible than esophageal speakers. Doyle, Danhauer and Reed (1988) found that intelligibility scores for tracheoesophageal speakers were higher (65%) than for esophageal speakers (56%). Miralles and Cervera (1995) found that both tracheoesophageal and esophageal speakers had difficulty conveying the voicing distinction, and suggested that the alaryngeal speaker is less capable of controlling the onset and offset of vibration of the neo-glottis. Velar phonemes were not conveyed regularly, and this was attributed to the changed morphological characteristics of the vocal tract after laryngectomy. Apart from errors caused by the voicing distinction, fricatives were often not

conveyed. Approximants were conveyed most accurately, followed by nasals and affricates, but error patterns were not identical for esophageal and tracheoesophageal speakers, indicating that some distinctions between these speakers might have been because of the difference in driving force (pulmonal versus injected air). Intelligibility scores for esophageal speakers in adverse noise conditions were especially poor for liquids, glides and nasals (Horii & Weinberg, 1975).

The intelligibility of alaryngeal speech is generally poorer than the intelligibility of normal speech. Esophageal speakers are further affected differently when compared to tracheoesophageal speakers, apparently because of the difference in the driving force.

In summary, the alaryngeal voice source can be unstable, and a source of noise, affecting speech quality and intelligibility. When the quality of speech is affected, prosody becomes more important in the process of speech communication. This will be explained in the next section, when functions of prosody are reviewed.

1.4 FUNCTIONS OF PROSODY

Prosody helps to structure an utterance. For example, it divides the speech into meaningful “chunks” and highlights important information. This structure is achieved by varying features such as pitch. In this sense, prosodic functions (what prosody does) can be distinguished from prosodic features (what prosody is). The latter were introduced in section 1.1 as “carriers” of prosody: F0 or pitch, intensity or loudness, and duration or length.

Prosody has an extra-linguistic function as it gives information on gender or age: for example, the average pitch height in females is higher than in males (cf. Lieberman, 1977). Prosody also has a paralinguistic function, for example, greater loudness and pitch height can signal excitement or anger (e.g., Ladd, 1996; Gussenhoven, 2004). Prosody further has linguistic functions, which will be the topic of this section.

Linguistic functions ascribed to prosody are varied, and depend on the language in question. This dissertation concentrates on the sentence level as it concerns Dutch. Therefore only those functions relevant for this study, namely focus marking and phrasing, will be discussed.

Prosody may fulfil a function related to syntax, or semantics. The (morpho-) syntactic or semantic structure influences what the prosodic

structure will be in an utterance, although the relation between prosody and syntax or semantics is not one-to-one. This means that alternative prosodic structures, which are equally likely or acceptable, may be used to express the syntactic or semantic structure in question (e.g., Cutler, Dahan & Van Donselaar, 1997). In practice, syntactic or semantic requirements determine which prosodic structures are beneficial to the listener's processing of the message, prosodic-phonological rules determine which prosodic structures are allowed, and the speaker decides – depending on background noise, for example – how he will convey the prosodic structure.

An explanation of the prosodic functions *focus marking* and *phrasing* is given next, as well as how they are conveyed. Also, a few studies will be mentioned to illustrate in what way a prosodic function may benefit speech processing.

1.4.1 FOCUS

Focus, somewhat simplified, has a highlighting function associated with important, or new information in a phrase or sentence (e.g., Bolinger, 1958; Ladd, 1996). In other words, the listeners' attention is directed to the semantically significant elements of the message.

For example, the answer to the question “*When are you going?*” can be rather emphatic, focusing on important information, as in “*I've **TOLD** you when I'll be going!*” or the answer can merely provide new information, as in “*I'm not sure, but I may go **TONIGHT***”.

The perception of focus is triggered by the presence of a pitch accent (Bolinger, 1958; Van Donzel, 1999). It has been firmly established that pitch is the most important cue to accent perception for Dutch (Cohen & 't Hart, 1967; Sluijter, 1995): the speaker uses an accent-cueing F0 event to focus the listener's attention on the important information.

On the one hand, a speaker has “prosodic freedom”. To a certain degree, he is free to add other, secondary accents apart from the primary accents we just described. For example, “*going*” in the first answer may also carry an accent; or “*may*” and “*sure*” in the second answer, etcetera. The speaker may also choose the type of accent (e.g., a rise-and-fall), although his choice will have consequences for the type and position of other accents in an utterance. It is also up to the speaker to decide how prominent he wants the accent to be. This last factor may be associated with the speaker's emotional

state, or interference of external noise, but also with the presence of other accents.

On the other hand, the speaker's prosody is also "fixed". The context directs which word(s) need to be highlighted so that a specific meaning is communicated. Depending on the semantic or pragmatic context, different prosodic structures are therefore appropriate (Cutler, et al., 1997). Accent distribution is also a product of rules, because not every word in the sentence may carry an accent (Ladd, 1996; Gussenhoven 2004). Rules further disallow the adjacent occurrence of certain accents ('t Hart, et al., 1990; Gussenhoven, 2004). These last two factors prevent uniformity of structure which would otherwise result when, for example, all the words are accented, or all are de-accented, or when the same accents are positioned next to each other. For this reason there is also a perceptual difference between more prominent and less prominent accents (Rietveld & Gussenhoven, 1985; Rump & Collier, 1996). These rules governing accentuation imply that a speaker should be able to control the acoustic features associated with accent quite accurately, so that the exact position and degree of prominence may be conveyed consistently, during speech communication.

As mentioned above, accent is *primarily* realized through a pitch event on the stressed syllable of a word (e.g., Bolinger, 1958; 't Hart, Collier & Cohen, 1990; Sluijter, 1995), although the accented word may also be louder and longer (Nooteboom, 1973; Eefting, 1991). These cues form a hierarchy, with F0-change as the most important acoustic cue to accent, followed by intensity and duration.

Numerous studies have investigated the effect of accent on processing in listeners. Some of these studies will be mentioned here. Words that were in focus were retained in memory longer (Birch & Garnsey, 1995). Appropriately placed accents facilitated sentence comprehension, in that listeners responded faster when new information was accented and information that had already been mentioned was de-accented. In contrast, inappropriate accentuation on words that were repeated slowed down response times (Nooteboom & Terken, 1982). The degree of speech intelligibility influences listeners' reliance on prosody. When the segmental quality of speech was impaired, listeners relied more strongly on the available prosody to identify new or important words (Van Donselaar & Lentz, 1994).

Given the poorer alaryngeal speech quality and intelligibility described in section 1.3, listeners would benefit from accenting in alaryngeal speech, but

most probably only if alaryngeal speakers are able to convey accents accurately: through F0 (pitch) movements that are appropriately positioned, with the exact degree of prominence.

1.4.2 PHRASING

Basically, phrasing divides speech into meaningful chunks of information. The prosodic boundaries in an utterance mark consecutive phrases (Streeter, 1978; Price, Ostendorf, Shattuck-Hufnagel & Fong, 1991) and listeners can accurately locate major syntactic boundaries from the prosodic information that signals prosodic boundaries. Intuitively, phrasing seems to be intimately related to the syntactic structure of a sentence, but the syntactic structure does not automatically predict how a speaker will phrase an utterance. As mentioned in 1.4.1, there are often a variety of prosodic phrasing possibilities, and some do not even line up with the syntactic structure (Shattuck-Hufnagel & Turk, 1996), because other factors also influence the phrasing of an utterance, such as accent patterns, speaking rate, and rhythm. The latter is illustrated by the role of symmetry: a speaker prefers to partition his speech into stretches of approximately equal length (Klatt, 1976; Gee & Grosjean, 1983). Similar to accentuation, the speaker is therefore free to convey or omit prosodic boundaries, as long as the relevant boundaries are available to the listener.

The acoustic cues associated with prosodic boundaries form a hierarchy, with final lengthening as the most “basic” cue: pre-boundary lengthening is a sufficient cue for the perception of a prosodic boundary (e.g., De Rooij, 1979). Listeners even “hear” pauses when only final lengthening is available (Scott, 1982), whereas insertion of pauses in the absence of final lengthening is perceived as a disfluency (De Rooij, 1979). Although F0 as a cue is often present, F0 on its own is not sufficient to signal the location of a boundary (Terken & Collier, 1992, De Rooij, 1979), and is also not necessary when durational cues are available (Lehiste, 1983). In practice, F0 and pauses are hierarchically added onto final lengthening. As the syntactic boundary becomes more important, the degree of lengthening tends to increase, F0-excursions are added and then pausing: pitch movements occur frequently without pauses, but pauses are normally accompanied by boundary marking pitch movements (Blaauw, 1994; De Pijper & Sanderman, 1994; Klatt, 1975; Price, et al., 1991; Wightman, et al., 1992; Shattuck-Hufnagel & Turk, 1996; Gussenhoven & Rietveld, 1992).

The effect of prosodic boundaries on speech processing is nicely illustrated by studies that manipulated prosodic cues so that they did not fit the expected syntactic structure: Sanderman and Collier (1997) found, for example, that listeners' processing time slowed down if prosodic boundaries were not appropriately realized and appropriately positioned. Similarly, syntactic parsing was adversely affected when prosodic boundaries conflicted with the syntactic expectation (e.g., Speer, Kjelgaard & Dobroth, 1996). The importance of prosodic boundaries is further confirmed when potentially ambiguous sentences need to be disambiguated (Lehiste, 1973; Lehiste, 1983; Scott, 1982; 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. Further, pauses that are positioned within phrases instead of at syntactically motivated phrase boundaries have a negative effect on speech recognition (Scharpff & Van Heuven, 1988; Nootboom, Scharpff & Van Heuven, 1990; Sanderman & Collier, 1997). Parts between the boundaries should preferably be produced on one breath without pausing (De Rooij, 1979). More importantly, the study by Nootboom, et al., (1990) illustrated that well-formed phrasing is especially beneficial in speech that is of lesser quality or less intelligible than normal speech.

In summary, listeners need an actual difference in prosodic realization if they are to distinguish between the absence and presence of a prosodic boundary, or if they are to identify to what extent a specific boundary is more important than another boundary. To convey this contrast, a speaker must be able to manipulate durational cues with some consistency.

Given the poorer speech quality and intelligibility in alaryngeal speakers, proper use of focus, as well as appropriate phrasing might be of great benefit during speech communication. Conversely, the alaryngeal voice source might actually prevent alaryngeal speakers from consistently manipulating the necessary prosodic features. For accenting, consistent control over F0 is the most important, and this might well be problematic for some alaryngeal speakers, as revealed in sections 1.2.2 and 1.3. For proper phrasing,

durational cues are essential, but realizing the appropriate durational contrasts might prove difficult for esophageal speakers, because of their limited air supply. In the next section an overview will be given of the available literature on prosody in alaryngeal speech, to get an idea of how well or how poorly alaryngeal speakers convey prosodic functions.

1.5 PROSODY IN ALARYNGEAL SPEECH

A number of studies dealing with prosody in alaryngeal speech were found. Nearly all were part of a project that investigated how proficient alaryngeal speakers of American English realize prosodic contrasts (Gandour & Weinberg, 1982; 1983; 1985; Gandour, Weinberg & Garzione, 1983; Gandour, Weinberg & Kosowsky, 1982; McHenry, Reich & Minifie, 1982). The prosodic functions that were investigated concerned perception and production of noun-verb contrasts (“*Object*” versus “*obJECT*”), minimally distinguished noun compounds and noun phrases (“*BLACKboard*” versus “*black BOARD*”), as well as contrasts on sentence level (questions versus statements, and contrastive focus). The latter contrasts concerned sentences such as “*Bev loves Bob*”, which could be produced as interrogative or declarative, and in which either name could be contrasted (“*BEV*” versus “*Bev*”, and “*BOB*” versus “*Bob*”). A two-interval-forced-choice task was used in all the perception experiments: listeners had to indicate the order in which a spoken contrast was presented (e.g., first the question and then the statement, or the other way around). The results of these perception experiments revealed that listeners were able to identify the intended prosodic contrasts with high accuracy, both in the tracheoesophageal and the esophageal speakers. Results for the alaryngeal speaker groups did also not differ significantly from the normal laryngeal speaker group.

However, listeners were presented with both contrasting utterances, so that they could make a comparison, whereas in normal speech communication a listener has to identify the intention without the benefit of directly comparing one intention with the other. Also, the participating alaryngeal speakers were highly proficient: only speakers who could fulfill daily speaking activities in a “problem-free fashion”, who produced fluent, continuous discourse without distracting extraneous noises or vocal roughness, and who spoke with a high level of intelligibility were included. Speakers were further allowed to repeat the speaking task until they were

satisfied that their utterances were as intended. The stimulus material was chosen so that individual phrases were short, words were bisyllabic at most, and only voiced consonants were included. Thus, possible physical limitations in the alaryngeal speakers were accommodated as much as possible. Because of this stringent selection of speakers and materials, there does not seem to be much scope to evaluate variability between groups or between speakers, let alone within speakers. Because speakers could select which utterances were “as intended”, it is unclear if they could *consistently* convey the intended contrast. The number of contrasts that had to be discarded (because they were not as intended), was not mentioned in the studies by Gandour and Weinberg, but McHenry et al. revealed that 14% of the esophageal speakers’ utterances had to be discarded, compared to none of the normal speakers’ utterances.

For the acoustic analyses, only those speakers and utterances were selected in which the prosodic intent had been signalled successfully, as investigated in the perception experiments. To contrast compound nouns and noun phrases, the alaryngeal speaker groups displayed systematic changes similar to those measured in the normal speaker group: an increase in the average F0 peak, intensity peak and vowel duration. To investigate the intonational contrasts (statement versus question), F0 contours were obtained for individual speakers. The authors found that esophageal speakers’ F0 contours “only represent noisy approximations” of normal F0 contours (Gandour & Weinberg, 1985: 86). In only two of the four tracheoesophageal speakers could an F0 contour be determined. The F0 contours that could be measured in esophageal and tracheoesophageal speakers showed great variability over time, with many disruptions and breaks, so that the contrasts that are generally clearly discernible in normal F0 contours, were not at all that distinct in the alaryngeal speakers. Yet, these rough, degraded F0 contours were sufficient to signal the intended prosodic contrasts, as the perceptual results revealed. It would seem from the results that listeners are well able to discriminate between prosodic contrasts in alaryngeal speech as long as the appropriate cue provides a perceptually salient difference that roughly resembles the difference found in normal speech. In those instances when alaryngeal speakers produced the *opposite* effect of what was expected listeners’ responses were no longer accurate, for example, a lower F0 peak and shorter duration of the syllable that was in focus than the syllable that was not in focus (Gandour & Weinberg, 1985).

In section 1.4, it was mentioned that consistent control over F0 is important for accenting and that the realization of appropriate durational cues is important for proper phrasing. The studies reported on in this section indicate that under the most favorable conditions, excellent alaryngeal speakers' best efforts are sufficient to convey certain prosodic contrasts. This was achieved by manipulating the appropriate cues in a manner resembling normal speakers.

The knowledge gained, as well as the limitations of the studies reviewed in this section, combined with the information that was given in the previous sections, will form the basis of the general hypotheses that this dissertation wishes to address, and these are given in the next section.

1.6 GENERAL RESEARCH HYPOTHESES

In section 1.1 it was mentioned that the aim of this dissertation is to investigate to what extent production of prosody in alaryngeal speakers is similar to that in laryngeal speakers and perception of prosody in alaryngeal speech is similar to perception of prosody in normal, laryngeal speech. Section 1.5 showed that alaryngeal speakers produced the same cues as normal speakers, albeit as rough approximations, and listeners accurately perceived the intended prosodic contrast. Thus, production of prosody by alaryngeal speakers and listeners' perception of prosody in alaryngeal speakers can be said to be similar to normal speakers. However, this can only be said of excellent speakers performing under speaker-friendly conditions. Sections 1.2.2 and 1.3 indicate that the alaryngeal population does not only include excellent speakers, and that control over essential acoustic variables such as F0 and timing, may be less consistent than the studies in section 1.5 revealed. If the sample is representative of the alaryngeal population as a whole, it is likely that some speakers might not be able to manipulate the essential prosodic features (F0 and timing) described in section 1.4, or might not be able to do so consistently. Also, the stimulus material should test, not accommodate, the possible limitations of the alaryngeal speakers. In this way, a more representative picture of alaryngeal speakers' prosodic abilities may emerge. In contrast to the studies mentioned in section 1.5, the research project reported on in this dissertation sets out to test the limitations of prosodic abilities in alaryngeal speakers, by including more demanding stimulus material and less proficient speakers. In the

present research project, the effect of an unreliable F0 is evaluated by including less proficient speakers, and the effect that a limited air supply might have on proper phrasing is investigated by including esophageal speakers.

From the information provided in the previous sections, the following general hypothesis might be formed:

H1. *All* alaryngeal speakers, regardless of proficiency will conform to the same rules in communication as normal speakers, and will therefore strive to convey necessary prosodic contrasts accurately, although the “hierarchy” of acoustic cues that is used may be dissimilar to the hierarchy found in normal speakers.

Thus, speakers will compensate.

The next hypothesis addresses the possible effect of this expected compensation.

H2. Because the phonetic realization of prosodic structure in alaryngeal speakers is dissimilar when compared to the phonetic realization of prosodic structure in normal speakers, listeners will not perceive the intended prosodic structure in alaryngeal speakers as accurately as the intended prosodic structure in normal speakers.

This hypothesis implies that the difference between laryngeal and alaryngeal speakers will also be reflected in how the linguistic structure is communicated: because the phonetic realization differs, alaryngeal speakers will not convey the prosodic structure accurately, and the syntactic or semantic structure will also no longer be communicated correctly.

The null hypothesis of the second hypothesis would therefore be:

H2Ø. Although the phonetic realization of prosodic structure in alaryngeal speakers is dissimilar when compared to the phonetic realization of prosodic structure in normal speakers, listeners will perceive the intended prosodic structure in alaryngeal speakers as accurately as the intended prosodic structure in normal speakers.

This hypothesis implies that the difference between laryngeal and alaryngeal speakers will not also lead to a difference in how accurately the linguistic structure is communicated: even if the phonetic realization differs, the prosodic structure will be conveyed and therefore syntactic or semantic structure will be communicated.

To test the general hypotheses, the experiments reported on in the different chapters concentrate on the production and perception of pitch and timing. The next section gives an overview of the different chapters.

1.7 OUTLINE OF THE PRESENT DISSERTATION

Each chapter was written as an independent study, addressing its own specific research questions. Some repetition was therefore unavoidable, especially in the Introductions.

In **chapter two** the perception and production of (contrastive) pitch accent is investigated. Because tracheoesophageal and esophageal speakers as well as normal control speakers participated, differences between the groups can be determined. It is also investigated which acoustic cues the alaryngeal groups use, compared to laryngeal speakers, and if esophageal speakers exploit pauses to compensate for their limited air supply. More importantly, because F0 was absent in a number of the participating, alaryngeal speakers we can investigate to what extent this influenced accurate perception of pitch accent, and how these speakers compensated for the absence of F0. Based on the acoustic analyses and another perception experiment, the notion of an alternative, non-F0 pitch-like property is tentatively proposed.

In **chapter three**, the perception of intonation in speech without F0 is investigated. The perception of pitch accent is dependent on the presence of a pitch event (or pitch-like event). Perception of speech melody implies the ability to convey different types of pitch movements, some more subtle than others: speech melody consists of a sequence of pitch events that include prominence-cueing events, *as well as* different boundary tunes, and tunes that signal sentence type. If ‘non-F0’ speakers indeed manipulate an alternative, pitch-like property, this non-F0 pitch may be used to convey speech melody. Three tracheoesophageal and two whispering speakers (all were considered non-F0 speakers) participated in the experiments of this chapter.

Chapter four addresses the next step in this search for an alternative to F0. First, we investigate how accurately the direction of non-F0 pitch is perceived, compared to the direction of F0 pitch. If an alternative non-F0 pitch-like property is to be counted as a true substitute of F0, it should be able to fulfill the same role as F0. Pitch direction is important as it may, for example, signal sentence type: the difference between a statement and a

question may be signaled through a falling or rising pitch movement. In this chapter, we also search for the acoustic properties that constitute this non-F0 pitch: which specific cues listeners interpret as the intended pitch direction, and how non-F0 speakers produce this non-F0 pitch direction. Because both non-F0 tracheoesophageal and laryngeal whispering speakers participated, we can further compare if control of non-F0 pitch is similar for both type of speakers.

In **chapter five** production and perception of phrasing is investigated. In this chapter, alaryngeal speakers were matched in terms of proficiency, so that the distinguishing factor would be the difference in air supply, and not variables such as intelligibility or speech quality. Laryngeal speakers also participated as controls, in voiced as well as whispered mode. The stimulus material consisted of potentially ambiguous sentences that can be easily disambiguated using prosodic cues. The sentences were constructed to test the limitations of esophageal speakers. It could therefore be determined if, and how esophageal speakers compensate to overcome the limitations of the air supply available to them.

In **chapter six**, the main findings of the research project are discussed. Once we have established to what extent the speakers conveyed the prosodic structures, we might also infer to what extent alaryngeal speakers communicate syntactic and semantic structures. In other words, it is investigated to what extent linguistic structures are achieved, given the possible constraints of the alaryngeal voice source.

*Pitch Accent*¹

ABSTRACT

The present study investigated whether alaryngeal speakers, if they have not been selected on grounds of proficiency, are able to convey pitch accent. The participating speakers (10 tracheoesophageal, 9 esophageal and 10 laryngeal) produced sentences in which accent was cued by the preceding context. For each utterance, a group of listeners identified which word conveyed accent. Results show that the accuracy with which speakers conveyed accent, varied. Acoustic analyses showed that some alaryngeal speakers had little or no control over fundamental frequency. These “non-F0” speakers did not compensate by using non-melodic cues whereas speakers using F0 did use non-melodic cues. Thus, temporal and intensity cues seem to be concomitant with the use of F0 (if F0 is affected, these non-melodic cues will be as well). A pitch perception experiment showed that listeners did however perceive pitch movements in non-F0 speakers’ utterances. Non-F0 speakers apparently rely on an alternative pitch system to convey accents and other pitch movements.

¹ An earlier version of this chapter was published as “Pitch” accent in alaryngeal speech, *Journal of Speech Language and Hearing Research*, 45, 1106-1118

2.1 INTRODUCTION

Suprasegmental features such as timing, loudness and pitch are generally referred to as “prosodic”. A speaker uses prosody to focus a listener’s attention on new or important information (Nooiteboom & Terken, 1982): the essential word or phrase is presented as in focus. The perception of focus is triggered by a pitch accent (Bolinger, 1958; Van Donzel, 1999). In Dutch, a pitch accent is realized through a prominence-lending pitch movement: a pitch rise, pitch fall, or combination of a rise-and-fall (‘t Hart, Collier & Cohen, 1990). Accented words are generally louder and longer, but for Dutch, it has been firmly established that pitch movement is the most important cue to accent perception (Cohen & ‘t Hart, 1967; ‘t Hart & Cohen, 1973; Van Katwijk, 1974; Sluijter, 1995).

Pitch is the perceptual correlate of F₀, the fundamental or repetition frequency of a sound. A normal speaker voluntarily alters the tension of the laryngeal musculature to produce F₀ changes typically associated with pitch accent.

This chapter investigates whether tracheoesophageal (TE) and esophageal (Es) speakers are able to convey pitch accent. As the alaryngeal voicing source does not have the same fine-tuning capabilities as the larynx, one might expect prosody to be compromised in alaryngeal speech.

As mentioned in chapter 1 (section 1.5), several studies have investigated prosody in alaryngeal speakers. A short review will be given here as well. Gandour and Weinberg (1982; 1983) looked at the perception of contrastive stress (“*BEV loves Bob*”; “*Bev loves BOB*”), and intonational contrasts (statement; question). Gandour, Weinberg and Garziona (1983) looked at perception of lexical stress (“*OBject*”; “*obJECT*”). These studies indicated that Es and TE speakers successfully conveyed prosodic intent. However, their stimulus materials were adapted to suit the capabilities of alaryngeal speakers, and only the speakers’ best attempts were used in the actual experiments. This makes it difficult to evaluate the consistency with which speakers conveyed these effects. Furthermore, these studies concentrated on highly proficient speakers (fluent, highly intelligible, and without distracting extraneous noises). The prosodic abilities of the speakers in these studies might not be representative of the alaryngeal population as a whole. If one allows whichever speakers to participate regardless of their proficiency, the speaker group might be more diverse in terms of proficiency, and the results

might give a more accurate picture of the prosodic abilities of the alaryngeal population in general. The first question in this study therefore was:

1. Are alaryngeal speakers, who have not been selected on the basis of proficiency, able to convey accent?

In their study on the production of intonational contrasts and contrastive stress, Gandour and Weinberg (1985) concluded that proficient Es and TE speakers manipulated the same acoustic cues as laryngeal speakers. Similar results were found for Es speakers producing syllabic stress: “*BLACKboard*” versus “*black BOARD*” (McHenry, Reich & Minifie, 1982). There was however, considerable variation among individual speakers, and especially Es speakers did not manipulate acoustical cues consistently (McHenry et al., 1982). Because the present study also included less proficient speakers, the second research question was:

2. Do *all* the alaryngeal speakers in the present study, regardless of proficiency, consistently use the same cues as laryngeal speakers to convey accent?

Speech intelligibility of Es speakers is compromised when compared to TE speakers: plosives, fricatives and liquid-glides were found to be significantly more intelligible in TE speech (Doyle, Danhauer & Reed, 1988). When segmental speech quality is impaired, listeners rely more heavily on accent to identify new information (Van Donselaar & Lentz, 1994). The recognition of speech that is segmentally unclear is also improved by inserting pauses at appropriate positions in sentences (Nooteboom, Scharpff and Van Heuven, 1990). Es speakers are known to insert more pauses than TE speakers, because the air supply that these speakers rely on is limited (Robbins, Fisher, Blom & Singer, 1984). As mentioned in the previous chapter, they have to pause regularly to inject air from the mouth into the esophagus. It is conceivable that Es speakers might exploit these injection pauses to signal accent by positioning the injection pause before or after an accented word, but not before or after an unaccented word. The third question was therefore as follows:

3. Do esophageal speakers use pauses as an extra acoustic cue to signal accent?

Moon and Weinberg (1987) looked at the relationship between F0 variation and transsource airflows: although the participating TE speakers could adjust the voicing source to influence the rate of vibration, this active adjustment was not used consistently to vary F0. The consistency with which F0 was manipulated furthermore varied within a speaker as well as between

speakers. Gandour and Weinberg (1985) found in their study that F0 could not be measured in one TE speaker and that one Es speaker did not vary F0 effectively. Aperiodicity and resulting absence of any harmonic structure in alaryngeal speakers was also mentioned by Van As, Hilgers, Koopmans-van Beinum & Ackerstaff (1998). Some speakers in the present chapter might not be able to manipulate F0 consistently. Given that F0-movement is the most important cue to accent in Dutch, the fourth question was:

4. Do speakers with little control over F0 convey accent as accurately as speakers with good control over F0?

Individual proficient alaryngeal speakers manipulated acoustic cues differently (Gandour & Weinberg, 1985). This could point to a trading effect in that a speaker with limited control over F0 might rely more heavily on intensity or duration (Slavin & Ferrand, 1995). Such effects might in fact be more evident in fair or poor speakers than in proficient speakers (McHenry et al., 1982). The fifth question therefore was:

5. Is there a trade-off effect between F0 and other acoustic cues in speakers without F0?

2.2 GENERAL METHOD

2.2.1 Speakers

Laryngeal, TE and Es speakers participated in this study. Speakers were approximately matched for age (on average four years difference was allowed: \pm four years), but no suitable laryngeal speaker was found to match the two oldest alaryngeal speakers. Apart from age, there were no selection criteria: all speakers who were available at the time recordings took place, were included in the study. All speakers were volunteers. Table 2.1 gives general information per speaker group.

Table 2.1. Information regarding speaker groups: median age: years; months (range); time since operation: yrs; mnths (range); number of speakers per speech proficiency category; gender distribution.

Speaker group	age	Time post-op	good	fair	poor	(male / female)
L	53;6 (40;9-63;5)	n.a.	10	-	-	10 (8/2)
TE	56;9 (46;1-77;8)	6;4 (1;1-11)	4	3	3	10 (8/2)
Es	57;6 (44;6-77;3)	5;5 (1;8-6;9)	2	3	4	9 (7/2)

The author informally rated speakers' overall proficiency as good, fair or poor, depending on a speaker's general intelligibility, beauty/clarity of voice and pitch. This was a routine, clinical impression based on the author's extensive experience with alaryngeal speech. Although the TE speaker group contained more "good" speakers, both alaryngeal groups had similar numbers of "fair" and "poor" speakers. The laryngeal speakers were all native Dutch speakers with no language, speech or voice problems (as judged by the author). All alaryngeal speakers had received post-operative radiotherapy. As explained in the previous chapter, in TE speakers the air from the lungs is shunted into the esophagus by means of a silicone-prosthesis. All TE speakers used the Provox prosthesis with an HME filter (Hilgers & Schouwenburg, 1990; Hilgers, Ackerstaff, Balm & Gregor, 1996).

2.2.2 Stimulus material

The material consisted of 10 different items. The second part of each item was taken from the Speech Reception Threshold sentences (Plomp & Mimpen, 1979). The first part was a precursor phrase, providing a semantic context for the second (test) part. This ensured that the desired accent pattern was cued. Speakers read the entire sentence (the underlined test part plus the preceding context) and emphasized the word printed in capitals. Two examples are given below, the complete list of sentences can be found in Appendix 1.

De schoen vloog niet over de schutting, de BAL vloog over de schutting
(The shoe did not fly over the fence; the BALL flew over the fence)

De bal vloog niet over de muur, de bal vloog over de SCHUTTING
(The ball did not fly over the wall; the ball flew over the FENCE)

Each contrastive pair of test utterances produced by the speakers yielded two accented words (e.g., "SCHUTTING" and "BAL") plus their unaccented versions, resulting in an utterance with accent towards the beginning of the test sentence (early) and another utterance with accent towards its end (late). Thus, the utterance with early accent included the unaccented counterpart of the utterance with late accent, and vice versa.

2.2.3 Procedure

Audio recordings were made in a quiet environment, using a condenser microphone (Sennheiser electret model ME 40) at a mouth-to-microphone distance of about 30 cm. The speech signals were recorded on a portable DAT recorder (AIWA HHB 1PRO, sample frequency 48 kHz).

Speakers read all 10 utterances twice, in random order. Five utterances were added at the beginning and the end of the list, which were used as training items in the perceptual experiment. Each speaker read 50 utterances: 10 utterances x 2 accent positions x 2 repetition + 10 additional utterances. From the two realizations of an utterance, the one with the fewest mistakes (repetitions, substitutions or omissions of words) was chosen for evaluation. The test utterances were downsampled to 22.05 kHz, yielding 580 utterances (29 speakers x 10 sentences x 2 accent positions). On average 20 minutes were needed to complete the task. Some speakers read all sentences without pausing, while other speakers paused more than once and found the task quite tiring.

2.3 ACCENT PERCEPTION EXPERIMENT

The first question in the Introduction was: “are alaryngeal speakers, who have not been selected on the basis of proficiency, able to convey accent?” A perception experiment was used to establish how well listeners could identify accent in alaryngeal speech.

2.3.1 METHOD

2.3.1.1 Stimulus material

Each stimulus used in the accent perception experiment consisted of a combination of a recorded utterance (the sound part of the stimulus) and two corresponding questions (the text part). The entire stimulus list comprised 580 items.

2.3.1.2 Listeners

Twenty-two native speakers of Dutch between the ages of 19 and 30 participated as listeners. All reported normal hearing. Listeners were not

informed about the purpose of the experiment. During everyday communication, alaryngeal speakers are confronted with listeners who have not been exposed to alaryngeal speech. Thus, to eliminate any effect of experience, participating listeners were unfamiliar with alaryngeal speech and inexperienced in speech evaluation. The listeners were paid for their participation.

2.3.1.3 Procedure

Listeners were seated in a sound-treated booth. A speaker's *spoken utterance* was presented over headphones, and two *written questions* were represented as text buttons on the computer screen in front of them. For example:

Question 1: Wat vloog over de schutting? (What flew over the fence?)

Question 2: Waar vloog de bal overheen? (What did the ball fly over?)

The listeners were asked to match the spoken utterance with one of the two questions. For example, if in the test utterance "*de bal vloog over de schutting*" (the ball flew over the fence) an accent was perceived on "*bal*", question 1 would have been selected. If an accent was perceived on "*schutting*", question 2 would have been selected. In other words, listeners chose which of the two *written* versions was answered by the *spoken* utterance. The chosen version was selected by clicking the appropriate text button. Listeners were instructed to guess when uncertain. Each utterance was presented once to each listener, yielding 22 judgments per utterance. After a response, the next stimulus was presented. There was a delay of three seconds between the appearance of the questions and the presentation of the test utterance over the headphones. All 580 stimuli were presented in random order.

2.3.2 RESULTS

Table 2.2 gives, per speaker group, the average percentage of utterances in which accent was perceived correctly, as well as the ranges (and N and SD).

Table 2.2 Percentage of utterances in which accent was correctly perceived, broken down per speaker group. (SD = standard deviation). N = total number of utterances multiplied by the number of judgments. Range (across utterances)

Speaker Group	Average Percentage (SD)	(N)	Range
Laryngeal (n=10)	95% (21%)	4400	91%-99%
Tracheoesophageal (n=10)	91% (29%)	4400	72%-98%
Esophageal (n=9)	82% (39%)	3960	74%-91%

Alaryngeal speakers, without prior selection based on proficiency, are generally able to convey accent, but there was more variation in the alaryngeal groups than in the laryngeal group. Some TE and Es speakers conveyed accent as accurately as laryngeal speakers, but not all.

The (arcsine transformed) percentages of correctly identified accents were entered into univariate analyses of variance on Speakers and Sentences. Speaker Groups (L, TE, Es) and Accent Position (early or late, nested under Sentences) were fixed factors; Sentences and Speakers (nested under Speaker Groups) were random factors.

The effect of Speaker Group was significant ($F_1(2,24)=14.02, p < .001$, $F_2(2,18)=90.37, p < .001$; $\min F'(2,31)=12.137, p < .001$). Post hoc analysis revealed that the Es group differed from the laryngeal group (Tukey's HSD, $p < .005$). Although not expected, the effect of Sentence was significant ($F(9,18)=6.79, p < .001$). Results for sentence 5 were significantly lower than results for sentences 1, 4, 6 and 7 (Tukey's HSD, $p < .005$). The reason for this is unclear as sentence 5 was not structurally different from other sentences. Results concerning the effects of Accent Position were inconclusive: they were significant in the first ANOVA, but not in the second. The effect of Speakers within speaker group was also significant ($F_2(24,486)=4.25, p < .001$). This is not surprising, as could be seen in Table 2.2. Furthermore, as explained in the Introduction, variation between alaryngeal speakers is known to exist.

Although the variation in the alaryngeal groups was greater than in the laryngeal group, we saw that only the Es group differed significantly from the laryngeal group. The differences and similarities between the groups are illustrated in Figure 2.1.

Figure 2.1 shows that the laryngeal speakers are grouped closely together, as are the TE speakers (except one TE speaker: a number of this speaker's utterances had more than one accented word, which might have confused listeners). For six TE speakers, the percentage correctly identified accent was in fact higher than the lowest percentage for the L speakers. For only

two Es speakers, the percentage correctly identified accent was similar to the laryngeal speakers. This explains why the Es group differed significantly from the L group: most Es speakers conveyed accent less accurately than the L speakers and – apart from one TE speaker – from the TE speakers.

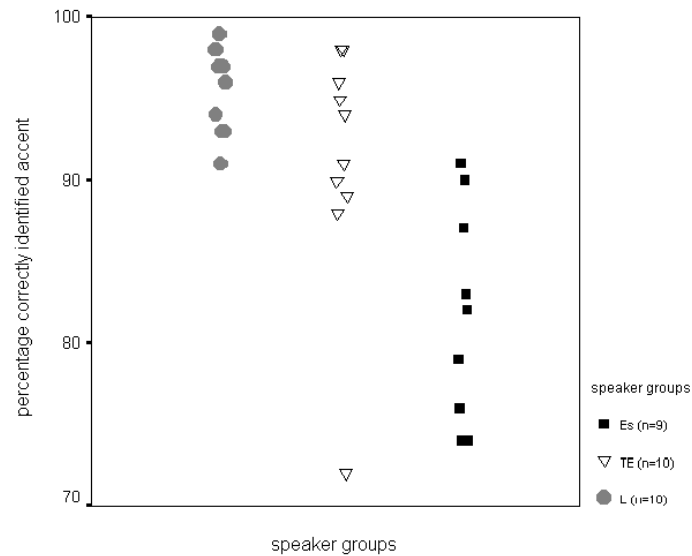


Figure 2.1. Percentages correctly identified accent, given per speaker, per group. If the 22 listeners had correctly identified accent in all of a speaker's utterances, this speaker would have scored 100% (L=laryngeal; TE=tracheoesophageal; Es=esophageal).

The fact that a number of TE and Es speakers conveyed the presence of accent as accurately as laryngeal speakers indicates that production of accent is not necessarily affected by the physiological limitations of the speakers. Some alaryngeal speakers may have sufficient control over their voicing source to effect the subtle changes necessary to signal pitch movements, or speakers have become sufficiently adept at using non-melodic cues when conveying the presence of an accent. To investigate if alaryngeal speakers indeed manipulated F₀, or other cues associated with accent consistently, the speakers' utterances were analysed. These acoustic analyses are described in the next section.

2.4 ACOUSTIC ATTRIBUTES OF ACCENT

Acoustic analyses were done to find an answer to the second question: “do the alaryngeal speakers in the present study consistently use the same cues as laryngeal speakers to convey accent?”

In Dutch, the most important cue for accent is a prominence-lending F0-movement in the accented syllable (Sluijter & Van Heuven, 1996). In accented syllables, overall intensity is higher, and spectral tilt flatter, than in unaccented ones (Sluijter & Van Heuven, 1996). Further, two studies of Dutch suggest that the domain of lengthening is the entire word (Eefting, 1991, Sluijter & Van Heuven, 1995). Pausing was also investigated, because Es speakers might use pausing as an alternative cue to accent.

2.4.1 METHOD

The acoustic cues (F0, intensity, spectral tilt, word duration and pausing) were investigated, using the *Praat* speech analysis program (Boersma & Weenink, 1996).

2.4.1.1 F0 movement

F0 was determined using sub-harmonic summation (Hermes, 1988). Subsequent F0-contours were re-synthesized by means of the PSOLA-analysis by synthesis technique (Moulines & Laroche, 1995). These synthetic contours were close-copy stylized (de Pijper, 1983). A close-copy stylisation is a synthetic approximation of the natural course of pitch, and it has to meet two criteria:

(1) The copy must be perceptually indistinguishable from the original. This is judged by the experimenter (in this instance the author) through analytical listening. This method has been validated in listening experiments by De Pijper (1983), ‘t Hart & Cohen, (1973) ‘t Hart, Cohen & Collier (1990)).

(2) The copy must contain the smallest number of line segments with which this perceptual equality can be obtained (in the time-log F0 domain).

Thus, the close-copy only contains F0-fluctuations that are relevant for the perception of intonation and excludes minor local fluctuations that result from co-intrinsic properties of adjacent segments, which have been shown

not to contribute to the perception of intonation (de Pijper, 1983). Figure 2.2 gives an example of an F0-contour and its close-copy stylization.

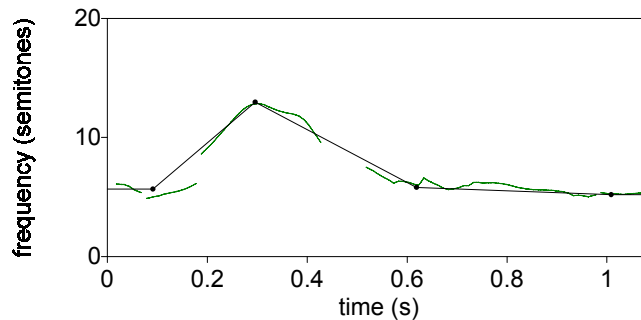


Figure 2.2. Original F0-contour (discontinuous line) & close-copy stylization (continuous line) of an utterance.

The discontinuous line represents the original F0-contour and the continuous line the close-copy stylization. The close-copy stylization consists of manually inserted pivotal points and connecting lines. As mentioned above, local, non-intonational F0-fluctuations are present in the original contour but absent in the close-copy. Using the stylized contours, the excursions between the F0-peak (F0-maximum in Hz) within the (un) accented syllable and the previous or following pivotal point (F0-minimum in Hz) were measured manually, and the distance in hertz between these two points was then converted to semitones (cf. Sluijter, 1995).

2.4.1.2 Peak intensity (in dB)

RMS intensity levels were calculated over a pitch-synchronous time window of 3.2 pitch periods (Boersma & Weenink, 1996).

2.4.1.3 Spectral tilt

Increased speaking effort associated with accent translates into more energy in the mid- and high frequencies (typically above 500 Hz) as well as an upward shift of formant frequencies. Both factors lead to a reduced spectral tilt (Sluijter, 1995). Thus, the intensity of two frequency bands: 25 to 500 Hz and 500 Hz to 4000 Hz was determined over the whole

(un)accented syllable and the difference in dB between the low and the high band was calculated.

2.4.1.4 Word duration

Durations of the complete (un)accented words were measured and compared. Word boundaries were determined on the basis of combined audio-visual (oscillographic and spectrographic) information, using the criteria described by Van Zanten, Damen and Van Houten (1991). Because the pre-release silence of a voiceless plosive was also part of the pre-accent pause, all words starting with a plosive were measured from the moment of the release. All Es speakers directly injected air into the esophagus except one, who used the inhalation technique to insufflate the esophagus. The injection technique is described as an uninterrupted movement, merging the injection with the initial phoneme of the following word (Damsté, 1958). In this study, some of the speakers' injections were auditorily and visually separate, independent entities (inspection of the oscillogram and the spectrogram revealed easily identifiable, short, vowel-like bursts of energy). However, excluding the injection would have meant that silent pauses consisted of silence plus the injection sound. Word durations were therefore measured by including the injection.

2.4.1.5 Pauses

The durations of silent intervals (absence of amplitude in the oscillogram) between words preceding or following the accented and unaccented words were measured. However, if a word started with a plosive, the pause consisted of everything up to but not including the release of the plosive (see word duration). As the silent interval before a voiceless plosive can approximate 100 ms (Slis & Cohen, 1969), pauses were defined as silences exceeding 100ms.

2.4.1.6 Consistency with which cues are used

As mentioned in the Introduction, alaryngeal speakers' ability to realize prosodic cues tends to be unpredictable and inconsistent. Hence, calculating and comparing the averages of the two conditions (accented values versus unaccented values), does not prove that a speaker uses a cue consistently.

We therefore chose to investigate the consistency with which speakers produced acoustic cues. For the interested reader, the average values are also given per group in Appendix 2.

To determine consistency, we proceeded as follows:

First, the difference between accented values and unaccented values was calculated for each (un)accented word (resulting in a total of 20 “difference” values per speaker: one for each (un)accented word. Late and early positions of accent were pooled, because the Perception experiment revealed that position of accent had little effect).

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 an F0 pitch movement to count as a difference, the F0-excursion in the accented word had to be at least 1.5 semitones larger than the F0-excursion in the unaccented word (Rietveld & Gussenhoven, 1985).

For to count as a difference, a difference of at least 1dB was adhered to (Moore, 1989): the intensity in accented syllable had to be at least 1dB more than its unaccented counterpart.

For word duration to count as a difference, a JND of 10% was adhered to (Klatt, 1976). Thus, the accented word had to be at least 10% longer in duration than its unaccented counterpart.

For a pause to count, its duration had to be at least 100 ms (see 2.4.1.5).

If the spectral tilt in the accented version was flatter than the spectral tilt in the unaccented version, this was taken as a difference.

Third, based on the criteria above, we totaled, for each speaker, the number of times that a cue was perceptually meaningful.

Finally, the Binomial Test (alpha is 0.05) was used to determine if the number of times that a cue was used (see 2.4.1.6) was indeed significant. If the Binomial Test was significant, the occurrence of a cue was taken to be consistent.

2.4.2 RESULTS

2.4.2.1 F0 measurability

In some alaryngeal speakers F0 could not be measured consistently. This failure could be a result of aperiodic excitations (complete absence of

harmonic structure), or to the inability of the F0 detection algorithm to detect periodicity in speech signals with relatively weak periodicity. Oscillograms and spectrograms of the utterances in question were visually examined. Glottal pulses were marked manually. In many instances, there was a striking lack of periodicity (gross irregularities in period duration, but also in peak amplitude), even within vowels. In the corresponding spectrograms no harmonics could be found, or, at the most, one or two very erratic harmonics.

Figure 2.3 shows that, whereas the word spoken by a laryngeal speaker has clear periodicity (repetition of same wave), no such regularity can be seen in the same utterance spoken by a TE speaker.

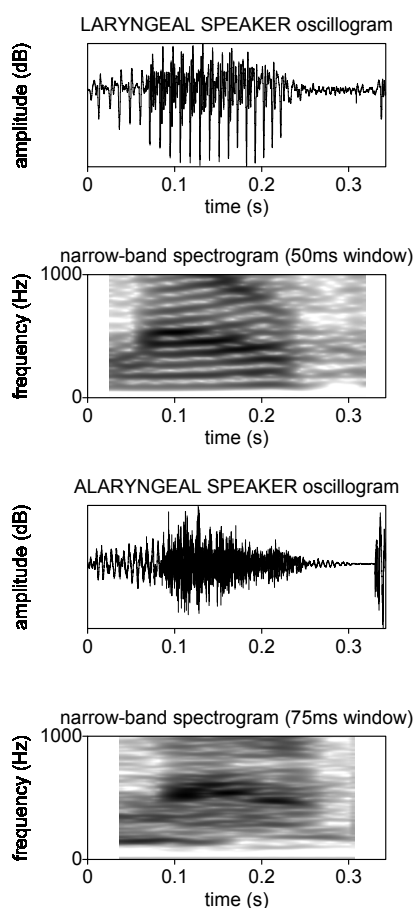


Figure 2.3. Oscillograms and narrow band spectrograms of the Dutch word 'loop' spoken by a laryngeal and an alaryngeal speaker. In comparison to the laryngeal speaker, the alaryngeal speaker's oscillogram shows an unstable signal and hardly any harmonics can be seen in the spectrogram

The narrow-band spectrograms reveal clear harmonic structure in the laryngeal utterance and the absence of harmonics in the TE utterance. An informal perception test was carried out to verify that the alaryngeal utterances, in which no F0 could be measured, did not contain F0. This test was based on the work of Cohen and 't Hart (1967): if an utterance contains F0, it is possible to obtain the whole F0-contour (pattern of rises, falls and level stretches) by asking listeners to match successive 30 ms portions of the utterance. This procedure will result in the same contour as found in the complete utterance. In the present study a number of alaryngeal speakers' utterances that did not seem to contain consistent F0, were divided into successive 30 ms portions. Two listeners listened to each portion and indicated if it was higher, lower or equal to the previous portion. The "pitch-height-rating" of the successive portions of these utterances did however not result in contours, but in a random collection of unrelated "movements". This is in contrast to results of speech containing F0 where whole and portioned utterances yield the same contours. It therefore seemed that F0 was indeed absent, or at least so inconsistent in these utterances, that listeners could not perceive F0 related pitch contours.

In this study, speakers could therefore be divided into a group with consistently measurable F0-movements (F0-speakers) and a group without consistent F0 (non-F0 speakers), demonstrating the variability that marks alaryngeal speakers as a group. In the non F0-speaker group, harmonics were completely absent in some speakers, whereas other speakers had erratic F0 (first harmonic) that could sporadically be measured. This lack of consistency prevented meaningful assessment of F0. Six out of eight F0-users conveyed accent more accurately than the non-F0 users. It is questionable whether an inconsistent F0 is a reliable cue to convey accent (this latter question will be taken up in Chapter 4 of this dissertation).

2.4.2.2 Consistency with which cues were manipulated

The second research question was: "do the alaryngeal speakers in the present study consistently use the same cues as laryngeal speakers to convey accent?" The results presented above indicated that some speakers manipulated F0, whereas others did not. Similarly, some speakers might, for example, manipulate duration whereas others might not. Because of this expected variability, it was determined how many speakers in each speaker

group consistently manipulated a specific cue (see section 2.4.1.6 for full explanation).

Figure 2.4 gives, for each group, the number of speakers that manipulated a cue consistently.

In general, the alaryngeal group did not consistently use the same cues as the laryngeal group to signal accent, with the intensity cue as the only exception used by all groups. The most striking difference between the laryngeal and alaryngeal groups was the (in)consistent use of F0. Five of the 10 TE speakers and only three out of the nine Es speakers had consistently measurable F0. In the other alaryngeal speakers, F0 was either absent, or so sporadic that no measurements could be made. The alaryngeal speakers could be subdivided in a F0 group (5 TE + 3 Es speakers) and a non-F0 group (5 TE + 6 Es speakers). It is conceivable that the non-F0 group relied on non-melodic cues to signal accent. This is investigated in 2.4.2.4.

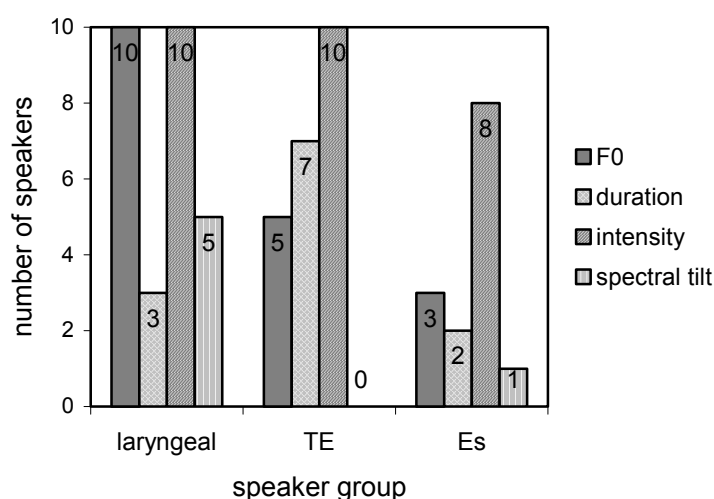


Figure 2.4. For each group, number of speakers that manipulated a cue consistently

In answer to the second question in the Introduction, the following can be said: the laryngeal group made the most consistent use of F0 and intensity, followed by spectral tilt and then duration. The TE group made the most consistent use of intensity, followed by duration and then F0. The Es group also made the most consistent use of intensity, followed by F0, however, the number of speakers that manipulated various cues was lower in the Es group

than in the TE group. It might be that the Es speakers relied on pausing as an extra cue, which is investigated in the following section.

2.4.2.3 Pauses

The third question was: “do esophageal speakers use pauses as an alternative cue to signal accent?”

No laryngeal speakers used pauses. Only one TE speaker used pauses consistently to mark accent. The pause strategy in the Es speakers is different to the other groups. Results for the Es group are given in Figure 2.5. Average durations of the pauses in the accented and non-accented versions are shown, pooled over all the Es speakers. Values for pre-and post-word pauses are given separately.

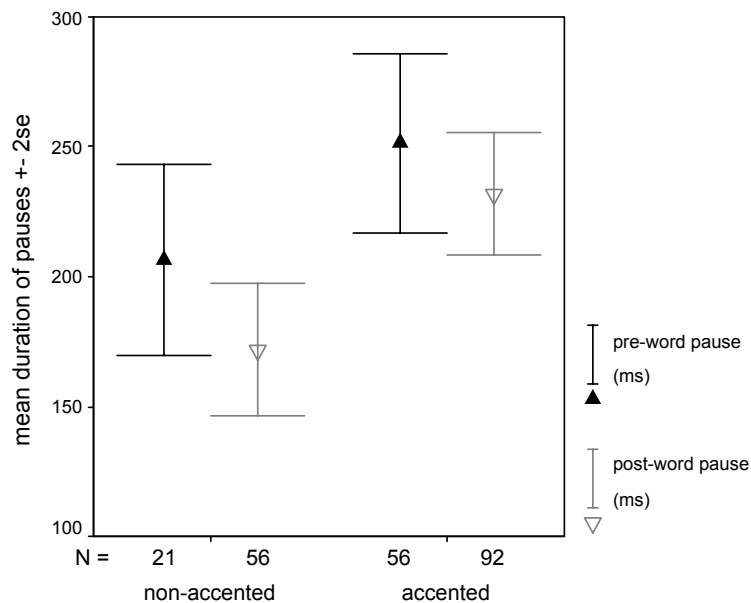


Figure 2.5. Average duration of pauses (and twice the standard error) in accented and non-accented versions, pooled over all Es speakers, given separately for pre-word and post-word positions

On first inspection, there does not seem to be much difference between the pre-word and post-word pauses. In both positions, the average duration of pauses is longer in the accented condition than in the unaccented

condition. Further, for both the pre-word and post-word positions, the number of pauses is higher in the accented version (56 and 92 respectively) than in the non-accented version (21 and 56 respectively). On closer inspection, there is a difference between the pauses in the pre-word position and pauses in the post-word position. There is considerable overlap in the pre-word position: pause durations in the non-accented version are often similar to pause durations in the accented version. This is not true for the post-word position, where speakers apparently made a clearer distinction between pause durations in the accented version versus pause durations in the non-accented version. The number of pauses in the post-word position is also higher than in the pre-word position.

A more detailed inspection of the pause distribution data further revealed the following: in the accented version, simultaneous pre-and post- word pausing occurred in 13 cases, compared to two in the unaccented version. Of the total number of accent-related pauses (righthand columns in Figure 2.5), pauses *also* occurred in the unaccented versions in 62 cases (44 were in the post-word position and 17 in the pre-word position). Further, of these 62, the duration of 18 pauses in the *unaccented* version were longer than in the accented version. These numbers indicate that the use of pauses might be less systematic than Figure 2.5 suggests. With so much variation in pause duration and the presence of pauses in the non-accented versions, it is unclear if listeners could effectively distinguish between “accent” pauses and “non-accent” pauses.

2.4.2.4 Trade-off between F0 and other cues

Although it is clear from the results of the acoustic analyses that speakers may be divided in a F0 group and a non-F0 group, it has not yet been determined if non-F0 speakers convey accent less accurately than F0 users. It is also unclear if non-F0 speakers rely more heavily on non-melodic cues than F0 speakers.

The fourth question was: “do speakers with little control over F0 convey accent as accurately as speakers with good control over F0?”

The average correctly identified accent for the group of F0 users was 90% (s.d. 29%) and the average correctly identified accent for the non-F0 user group was 83% (s.d. 37%). Although the difference between these groups was significant (Kolmogorov-Smirnov, $Z = 2.779$, $p < 0.001$), the

average percentages mentioned above and the results in Table 2.2 indicate that all speakers were generally often able to convey accent.

This could mean that non F0-speakers compensate by using non-melodic cues associated with accent (using intensity or duration instead of F0). The fifth question was: “is there a trade-off between F0 and other acoustic cues in speakers with limited control over F0?” (e.g., Slavin & Ferrand 1995).

Apart from F0, three non-melodic cues were measured (overall intensity, spectral tilt and duration; pausing will not be considered, because it seemed not to have been used as a cue to signal accent). Figure 2.6 compares how many of these other (non-F0) cues are used by F0-speakers and by non F0-speakers. The number of cues given (x-axis) does not include F0.

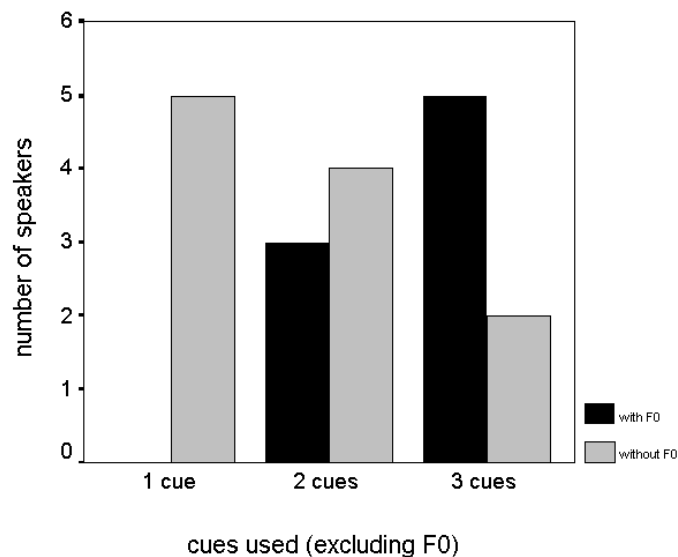


Figure 2.6. Number of cues used by different speakers. Speakers are divided in groups according to presence or absence of F0

It is clear from Figure 2.6 that F0-speakers exploited non-melodic cues more than non F0-speakers. Contrary to expectation, non F0-speakers did not rely more heavily on alternative cues to signal accent. Approximately half of this group only used one cue, whereas more than half of the F0-speakers also relied on three other cues.

Thus, there seems to be no trade-off effect in speakers who have little control over F0. In other words, non-F0 speakers do not seem to compensate for the absence of F0 by using non-melodic cues, at least not the cues that

were investigated in this chapter. However, listeners often perceived the accents that these speakers conveyed. Apart from the acoustic cues investigated in the previous section, speakers might manipulate other cues so that pitch accent is perceived. The next experiment looks more closely at the existence of a possible “alternative” (non-F0) pitch.

2.5 PERCEPTION OF ALTERNATIVE PITCH

Studies have shown that listeners perceive something pitch-like in whispered speech and that this “pitch” perception may be related to formants (e.g., (Higashikawa, Nakai, Sakakura & Takahashi, 1996). Similarly, listeners might rely on another feature to perceive pitch in the non-F0 users’ utterances. If this is true, one would expect listeners to perceive the same type of pitch movements on the same (accented) word in an utterance. If there is no, or inconsistent “pitch” information, different listeners would most probably perceive different types of pitch movements; or the perceived pitch movements would not be on the same word. To determine whether listeners could perceive pitch movements in utterances that did not seem to contain F0, another small experiment was carried out.

2.5.1 METHOD

2.5.1.1 Stimulus material

The stimulus material consisted of a random selection of non-F0 speakers’ utterances (see explanation following Figure 3). In total, 40 utterances spoken by four different male TE speakers were used. In this experiment Es speakers were not included so that listeners would not be distracted by the audible injections. Early- and late-accent positions were equally represented.

2.5.1.2 Listeners

Three experienced phoneticians participated in this experiment.

2.5.1.3 Procedure

Listeners were seated in a sound-treated booth, wearing headphones and with a microphone in front of them. They listened to each test utterance and concentrated on the intonation pattern of each utterance. The utterances were also available in writing, so that listeners could read the sentences intended by the speakers. Accented words were not marked in the written version. After listening (repeatedly) to an utterance, listeners imitated the utterance. These imitations were recorded and stored on disc.

2.5.1.4 Analysis

For each of the imitated utterances (produced by the listeners), the F0-contour was processed as described in 2.4.1.1. Per listener, the words and phrases on which pitch movements had been produced were listed, as well as the type of movement that was produced (rise, fall, rise-and-fall, or none).

Per word, it was determined how many of the listeners produced a pitch movement. If all three listeners produced a pitch movement on a word, and at least two out of the three listeners produced the *same* pitch-movement, that word was counted as a word carrying pitch movement. Thus, if one out of the three listeners did not produce a pitch movement on a word, that word was not considered to carry a convincing pitch movement.

2.5.2 RESULTS

Results for the ‘accented’ words (those words in which the desired accent pattern had been cued as described in 2.2) were as follows:

Presumably, the non-F0 users intended these words to carry a “pitch” accent, although this “pitch” would then have been conveyed by some other means than F0. With 40 utterances, there were also 40 accented words (one for every utterance). In 26 of these (65%), the three listeners produced the same movement, and in 13 (32%), two of the three listeners produced the same movement whereas one listener produced a different movement (e.g., two produced a rise whereas one perceived a rise-and-fall). Thus, in 39 (97%) of accented words, three listeners perceived pitch movement, although it was not always the same movement. These accents were intended by the speaker, and it seems from these results that the listeners in fact perceived the speakers’ intended “pitch” movements reasonably accurately.

Because one specific word in each utterance was cued so that it carried accent, the rest of the words were not expected to carry accent.

Results for the 'non-accent words' were as follows:

There were 197 non-accent words in total. In 49 (25%) of these non-accent words, all three listeners produced the same pitch movement. In another 26 non-accent words, two listeners produced the same pitch movement whereas one listener produced a different movement. Thus, in 75 (38%) of the words not cued to carry accent, three listeners apparently perceived pitch movement. This excludes declination, which was produced in many, but not all, utterances (as declination was not produced in some instances this indicates that these listeners did not automatically use declination, but imitated what they perceived). Although 38% seems low, it has to be remembered that the sentences consisted on average of six words. Since the main accent was cued in every sentence, one would not expect any other significant pitch movements (clear rises and/or falls). The fact that listeners perceived pitch movement in 75 words which were not cued to receive accent, might indicate that these words possess (as yet unknown) acoustic information. This was apparently perceived as pitch movement, unless there was a response bias which resulted in the perception of similar 'pitch movements' on certain words.

The notion of perceived pitch-movement in the face of unmeasurable F0 is not new, as the example of whispered speech mentioned above, illustrated. Furthermore, two TE speakers produced "largely aperiodic excitations" (unmeasurable F0), even though these speakers were able to signal the *intended* intonation successfully (Gandour & Weinberg, 1985). These authors concluded that the perceived pitch in alaryngeal speech might not correspond to F0 in a one-to-one fashion.

It might be that some of the alaryngeal speakers participating in the present study convey pitch movements, not through F0-movement, but through an alternative pitch system.

2.6 GENERAL DISCUSSION

Similar to Gandour and Weinberg (1982), this study looked at how well alaryngeal speakers convey accent. Both English and Dutch are stress-accent languages: stress is a structural, linguistic property that specifies which syllable in a word is strong; accent is used to focus important information. Further, American English and Dutch do not differ greatly in terms of

acoustic cues used to signal accent (Sluijter, 1995). In this sense, the present study is comparable to the work done by Gandour and Weinberg. There were however, a number of differences. Gandour and Weinberg used short, all-voiced phrases that only included monosyllabic words. In the present study, length of sentences and accented words varied. Unlike the present study, Gandour and Weinberg used the speakers' best attempts and all their speakers were judged as highly proficient. For three out of four Es speakers and three out of four TE speakers results of 93% or higher were calculated. In the present study, such high percentages were calculated for five out of 10 TE speakers and none of the Es speakers. In the present study, there was more variation among alaryngeal speakers, but it could still be concluded that all the alaryngeal speakers in the present study often conveyed accent, and some conveyed accent as accurately as normal speakers.

McHenry, et al. (1982), found that Es speakers used the same cues as laryngeal speakers to convey syllabic contrasts, but the Es group did not manipulate these cues as consistently as laryngeal speakers. Gandour and Weinberg (1985) also found that individual speakers differed in how they used each acoustic cue. Some alaryngeal speakers participating in the present study only used peak intensity consistently. Peak intensity is however very susceptible to environmental background noise and the chance that it has any communicative significance is small. Less than half of the alaryngeal speakers used F0 consistently and not many speakers used any of the other cues consistently. Apart from the cues usually associated with accent, the present study also investigated pausing, especially in esophageal speakers. The way pauses were distributed over accented and unaccented words suggests that pausing in Es speakers is determined more by a physiological need to replenish the limited air supply available for speech production, than by linguistic needs. As a result, the rhythmical pattern of Es speech is most probably compromised. If the rhythm of speech is unnatural, listeners will have difficulty using durational information as a cue to accent. This might explain why in the perception experiment, results for the majority of Es speakers were worse than the results for the TE speakers. Apart from the effect this apparently had on the perception of accent, this unnatural rhythm might also have consequences for the perception of prosodic boundaries. The effect of pausing in Es speakers on the perception of prosodic boundaries is investigated in Chapter 5.

The large inconsistency with which alaryngeal speakers manipulated F0 further highlights inter-speaker variability. F0 was measured consistently in

only five out of 10 TE speakers and three out of nine Es speakers. In the other speakers, F0 was either absent or its presence was so erratic, that it could not be measured. Speakers with little control over F0 also did not convey accent as accurately as F0-users. If individual speakers are ranked according to the results of the Perception experiment, six out of eight F0-users conveyed accent more accurately than the non F0-users. This substantiates the importance of F0 as a cue to perceive accent.

Speakers with little control over F0 did not rely more heavily on non-melodic cues to convey accent. This finding therefore contrasts with the idea of cue trading described by Slavin and Ferrand (1995). A possible explanation might be that acoustic cues such as intensity and duration are strongly associated with F0. In a prosodic function such as accent, they do not exist independently or separately from F0, but are concomitant. Further, physiologically, the more control one has over the voicing source, the more one is probably able to manipulate other acoustic cues. Es and TE speakers rely on the neoglottis as the voicing source (cf. Diedrich, 1968; Moon & Weinberg, 1987). Voice quality differences can be explained by the physiological characteristics of the neoglottis (Dworkin, et al., 1998). As the anatomical and morphological characteristics are quite variable (Van As, 2001), it is not surprising that interspeaker variability is the defining characteristic of TE and Es voicing. From the results in the present chapter, it is concluded that voicing source variability might also have influenced the consistency with which F0 and timing were manipulated.

All of the F0-speakers were subjectively rated as good speakers in terms of proficiency, except one Es speaker who also received a lower percentage in the Perception experiment. One Es speaker who was subjectively judged as a good speaker achieved a lower percentage than some fair or poor speakers. Some poor speakers also achieved higher percentages than a number of 'fairly proficient' speakers. The term "proficiency" as used in this study was comparable to "proficiency" as used by Gandour & Weinberg. Two of the speakers in the Gandour & Weinberg study (1982) achieved much lower percentages than the other speakers (74% versus 93% or higher) although these speakers were judged as highly proficient. A subjective rating of "proficiency", apparently does not predict prosodic ability.

The second and third experiment showed that some speakers might manipulate acoustic properties other than the ones investigated to convey pitch. Sisty and Weinberg (1972), studying formant frequencies in Es speakers, observed that in addition to "appropriate" formant frequencies,

unexpected concentrations of energy were found in 12% (20 out of 190) of vowel spectra examined. These energy concentrations were “formant like” in appearance, in frequency regions where formants are not expected, and present throughout the major portion of the vowel. These extra resonances were seen in 10 out of the 27 participating Es speakers. Coleman (1971) demonstrated that the location of vocal tract resonances in the frequency spectra also plays a role in the perception of vocal pitch. Research on whispered speech has shown that there is a relation between perceived pitch and formants (Higashikawa, Nakai, Sakakura & Takahashi, 1996). Giet (1956) claimed that pitch movements can be heard in whispered utterances such as “*Nein.*” versus “*Nein?*” Meyer-Eppler (1957) investigated acoustic features in whispered German speech and concluded that two substitutes exist for periodic pitch movement: in some vowels, gaps in the higher frequencies are filled with noisy components. In other vowels, formants shift upwards. Thus, in laryngeal speakers, pitch movements might be produced independently of F0. Some alaryngeal speakers possibly manipulate spectral properties in a similar fashion to convey pitch.

If an alternative non-F0 pitch system exists, non-F0 speakers should be able to produce meaningful pitch contours, and listeners should be able to perceive the pitch movements as they were intended by the speaker. The next chapter investigates the latter claim: whether listeners indeed perceive the speech melodies as intended by non-F0 speakers.

Perception of Speech Melody in Speech without F0

ABSTRACT

An intonation language, when spoken, contains pitch events that signal semantic or syntactic information. These pitch events, which are strongly related to the fundamental frequency, form components of speech melody. This study investigates if listeners perceive speech melody in alaryngeal speech, in which F0 is absent. Three tracheoesophageal and two whispering laryngeal speakers participated. A perception experiment revealed that listeners perceive speech melody in the absence of F0. A rating experiment revealed that non-F0 speakers' *intended* speech melodies were partly recognized. Non-F0 speech melodies were also transcribed. Confusion matrices showed that gradual pitch changes are perceived less accurately than abrupt changes. Pitch *direction* (rising vs falling) is hardly ever confused. Non-F0 speech apparently contains something pitch-like that simulates F0 pitch.

3.1 INTRODUCTION

Pitch can be defined as “that attribute of auditory sensation in terms of which sound may be ordered on a musical scale” (American Standards Association, 1960). In other words, variations in pitch give rise to a sense of melody (Moore, 1989). Spoken language also contains variations in pitch, which can be regarded as sequences of pitch events (Ladd, 1996). Para-linguistically these pitch events signal information about gender, age or emotional state. Linguistically pitch events signal different lexical meanings in tone languages such as Chinese, whereas in intonation languages such as English or Dutch, they convey semantic or syntactic features of the message. Tonal languages are characterized by more pitch fluctuations (as a function of time) than intonation languages (Eady, 1982), but acoustic and physical correlates of pitch are the same for both tone and intonation languages (Ladd, 1996).

Perceived pitch is strongly related to changes in the fundamental frequency (F0). Control of F0 in speech is accomplished through a combination of aerodynamic forces and volitional laryngeal adjustment. A speaker varies the length and tension of the vocal folds, thus controlling rate of vibration, to produce communicatively relevant F0 changes (‘t Hart, Collier & Cohen, 1990).

Since F0 constitutes an important cue to the abovementioned linguistic functions, these functions might be hampered when F0 is absent. The question arises if speech, in which F0 is absent – henceforth “non-F0 speech” – conveys something pitch-like that functions as a substitute for F0-based pitch: can speakers without F0 compensate for the absence of F0?

For example, some of the speakers participating in the previous chapter did not seem to be able to manipulate F0 consistently. In fact, the presence of F0 seemed to be very erratic in some speakers, and completely absent in others. This group of ‘non-F0’ speakers did not convey accent as accurately as speakers with F0, but still conveyed the presence of accent quite often. These speakers might therefore not be able to convey the linguistic functions of pitch that are normally conveyed by F0, unless their speech contains another pitch-like phenomenon.

However, do alaryngeal speakers in general convey linguistic correlates of pitch? Studies that have investigated linguistic correlates in alaryngeal speech yielded diverging results. Most of these studies dealt with tonal

languages. In a study on the perception of tones in Cantonese tracheoesophageal speakers, 52% of the tones were correctly perceived (Ching & Williams, 1994). In another group of Cantonese alaryngeal speakers, only 26 out of 44 could imitate five out of the six word tones correctly (Wong, et al., 1997). In another study, monosyllabic words produced by Chinese alaryngeal speakers were presented to listeners for identification. For (tracheo)-esophageal speakers, half of the tones were correctly identified (Yiu, Hasselt, Williams & Woo, 1994). Unfortunately, these three studies did not differentiate between speakers with F0 and speakers without F0. A study on tone in Thai alaryngeal speech showed that tones produced by two alaryngeal speakers were correctly identified in 70% and 39% respectively. Reliable F0 contours could not be extracted for the latter speaker (Gandour, Weinberg, Petty & Dardarananda, 1988).

In two studies that dealt with an intonation language, Gandour & Weinberg (1983, 1985) found that alaryngeal speakers conveyed intonational contrasts (question versus statement), even though F0 could not be measured reliably in some speakers.

It is unclear from these results to what extent the absence of F0 might affect how accurately pitch events are perceived. It is however unclear what the difference is between alaryngeal speakers with, or without F0. It seems conceivable that F0 speakers conveyed pitch events much more accurately than non-F0 speakers (Gandour et al., 1988). The question arises if alaryngeal speech without F0 contains something pitch-like that allows listeners to perceive pitch events at least to some extent.

The present study therefore investigates to what extent listeners perceive pitch in alaryngeal speech *without* F0.

In addition to alaryngeal speakers, we included whispering laryngeal speakers in the present study as controls, because F0 is typically absent in whispered speech. It is conceivable that alaryngeal speakers without F0 will convey a similar pitch-like phenomenon that is found in whispered speech. Fortunately, this phenomenon has been investigated in whispered speech. For example, according to Panconcelli-Calzia (1955), a missionary in China claimed that he had no difficulty hearing and understanding confessions that were whispered in Chinese, but Panconcelli-Calzia argued that that could not be true, as the necessary information is not present in the signal. In response, the same missionary asserted that “auch beim Flüstern spricht und hört der Chinese die ‘Töne’” (Giet, 1956:376). He further suggested that the presence of ‘pitch’ is self-evident when one whispers the word “*nein*” (“no”), first as

a question then as a statement. Jensen (1958) tested the recognition of whispered minimal word pairs in four tonal languages. He concluded that tones in whispered speech were perceived, but more accurately, for example, in Mandarin than in Slovenian. Miller (1961) replicated Jensen's experiments in whispered Vietnamese, but concluded instead that very little tone was transmitted: a confusion matrix of one word with six different tones revealed that there was confusion across the board.

We found one published study related to whisper in an intonation language. Meyer-Eppler (1957) made sound spectrograms of a sentence, first whispered as a statement and then as a question. He concluded that formants and spectral tilt were modulated to replace F0 pitch. Unfortunately, the perception of these F0 pitch substitutes was not investigated. An unpublished study on intonation in whispered Dutch showed that the second formant might be related to the perception of certain intonation contours (Heeren, 2001).

In summary, the results on whispered speech seem rather ambiguous: certain pitch events can be perceived, but it seemed to depend on the listener.

Based on the literature reviewed above, we expect that something pitch-like might be perceived to some extent, in the non-F0 speakers participating in the present study.

Since Dutch is an intonation language, we elicited specific speech melodies through an imitation task, so that the investigators knew exactly what the speakers' intended pitch events were. Thus, we ensured that the stimulus utterances were suitable to answer the present study's research questions:

1. Do naïve listeners perceive speech melody *at all*, in non-F0 speech?
2. Do naïve listeners perceive the *intended* speech melodies in non-F0 speech?
3. Do listeners perceive certain non-F0 pitch events more accurately than other non-F0 pitch events?

The participating speakers' utterances were presented to listeners in perception experiments, and each perception experiment was designed to answer one of the three research questions.

3.2 ELICITING NON-F0 SPEECH MELODIES TO BE USED IN THREE PERCEPTION EXPERIMENTS

The main goal of this study is to investigate if listeners perceive something pitch-like in non-F0 speech melodies produced by non-F0 speakers. However, not only did we want to know if speech melody was perceived at all in non-F0 speech, but also if listeners correctly perceived the speakers' *intended* speech melody.

To achieve this, the selected non-F0 speakers imitated reference utterances. The reference utterances contained melodic 'recipes' (known sequences of pitch events).

This section describes how speakers were selected. It further describes how the reference utterances containing the melodic recipes, and the non-F0 utterances containing imitations of the melodic recipes, were obtained.

3.2.1 METHOD

3.2.1.1 Speakers

One normal-speaking laryngeal female speaker (the author, functioning as reference speaker), who also whispered (whispering laryngeal female: WLF), a whispering laryngeal male speaker (WLM) and three tracheoesophageal speakers (TE1, TE2, TE3) participated.

Since the goal of the present study was to investigate speech melody in *non-F0* speakers, TE speakers were selected, based on visual inspection of their acoustic signal, as explained below. All TE speakers were male. Standard laryngectomies had been performed on TE2 and TE3. For TE1, the extent of the surgery had been as follows: a total laryngectomy with a partial pharyngectomy, reconstructed with a myocutaneous pectoralis major flap. The TE speakers had all received post-operative radiotherapy. All TE speakers used the Provox prosthesis (Hilgers & Schouwenburg, 1990). The speakers were not matched for age. Age at time of recording was 55 for TE1, 77 for TE2 and 48 for TE3. The time of the recording was more than four years post-laryngectomy for each speaker.

3.2.1.2 Validity of speaker selection

Tracheoesophageal speakers can be divided into subgroups, based on the visual appearance of the acoustic signal (Van As, 2001). This acoustic signal typing differentiates between speakers by considering the presence and stability of F0 as well as the harmonic strength of the speech signal. The classification introduced by Van As is primarily based on the visual inspection of narrow-band spectrogram and spectrum (presence of harmonics), as well as visual inspection of the oscillogram (periodicity), and on the result of pitch extraction (percentage ‘voiced’ found for the calculation of F0). The TE speakers participating in the present study could all be classed as type IV: a highly unstable signal with no, or at most, fleeting traces of periodicity. To illustrate, Figure 3.1 compares the periodicity in the reference speaker (R) and TE1, as seen in the oscillogram of a stable part of a vowel.

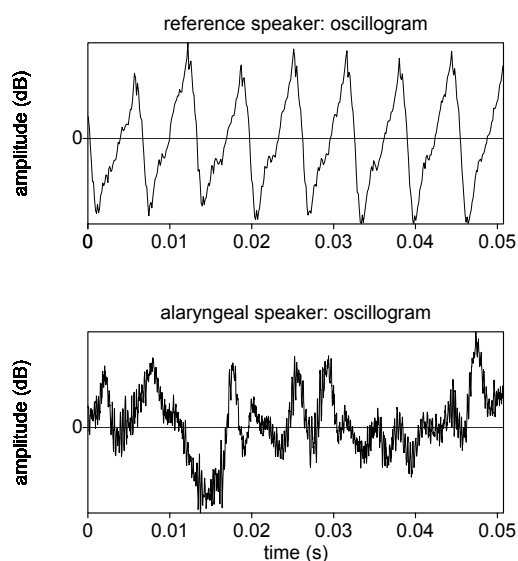


Figure 3.1. Typical oscillograms of a vowel produced by a laryngeal speaker (A) and alaryngeal speaker (bottom).

Whereas the A oscillogram shows clear periodicity, the bottom oscillogram is clearly aperiodic. The difference between the reference speaker and TE1 is further illustrated in the narrow-band spectrograms in Figure 3.2.

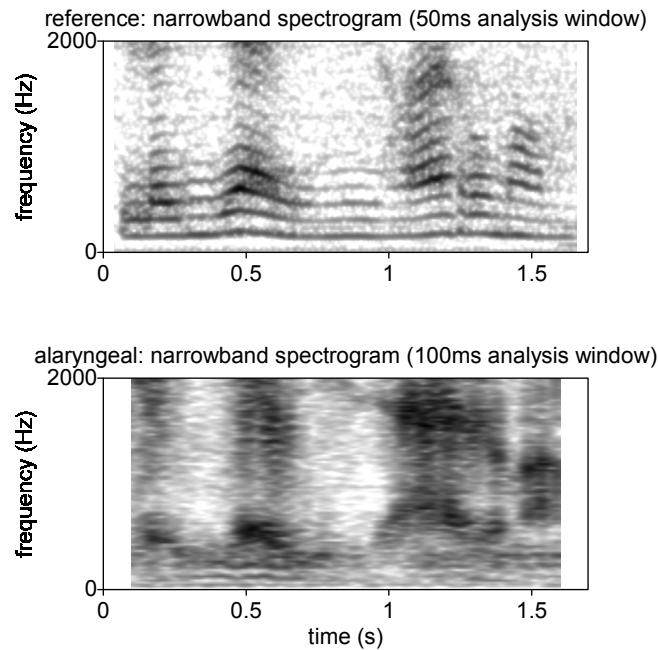


Figure 3.2. *Narrowband spectrograms of laryngeal speaker (top) and alaryngeal speaker (bottom) of the sentence: ‘leen jij nu haar roman’ (borrow you now her novel).*

In the top spectrogram, the harmonic structure and F0 contour is clearly visible, indicating the positions of a number of rising and falling pitch movements. In the bottom spectrogram no harmonic structure or F0 contour can be seen, although one can see that the frequency of the energy peaks in the spectrum seem to mimic the F0 contour in the top spectrogram. TE1 seems to manipulate or filter some aspects of his speech to produce a pitch-like effect. The examples of the oscillogram and narrow-band spectrogram are representative of all the TE speakers. Based on inspection of the speech signal, all TE speakers were considered non-F0 speakers.

The whispering speakers were included because one can be certain that their utterances would not contain F0. We expected any pitch-like phenomena found in alaryngeal speech to be similar to pitch-like phenomena in whispered speech.

The normal-speaking laryngeal speaker’s utterances were the only utterances that contained consistent, regular fundamental frequency. These were reference utterances that contained realizations of the melodic recipes.

3.2.1.3 Stimulus sentences

There were 11 sentences. Words containing mostly sonorants were used in the sentences, to maximize the number of phonologically voiced segments. Each sentence was supplied with several speech melodies: ‘melodic recipes’ on paper, that consisted of known sequences of different pitch events. Eight sentences were provided with two different melodies and three sentences with three different melodies.

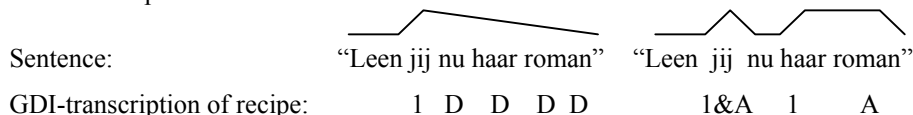
The melodic recipes were based on the IPO-Grammar of Dutch Intonation (GDI), as described in ‘t Hart, et al. (1990)¹. The Grammar of Dutch Intonation is an experimental-phonetic, bottom-up approach to speech melody that has resulted in a melodic description of all possible pitch contours of Dutch speech melody. In this approach, pitch contours are seen as melodic entities that tend to coincide with clauses or complete utterances. The pitch contour is viewed as a sequence of configurations. Configurations are described as smaller structural units that consist of one or more perceptually relevant pitch events (e.g., boundary tones or accents). These perceptually relevant pitch events are taken to be F0 variations intentionally produced by the speaker (in contrast to involuntary F0 variations, inherent to the segmental structure of speech, which do not contribute to the perceived speech melody, according the Grammar of Dutch Intonation). The GDI differentiates 10 different pitch events as well as two types of declination. Each event is represented by a specific symbol. We used 10 different pitch events in the present study.

Each sentence was provided with GDI-pitch configurations (in symbols and as drawn patterns), which constituted the melodic recipes. An example of a sentence and its accompanying melodies is given below:

The sentences, with the different melodic recipes, were used to produce reference utterances.

¹ The GDI is not the only transcription system designed to capture pitch events of Dutch intonation. ToDI (Gussenhoven, Terken & Rietveld, 1999) also focuses on the transcription of phonological contrasts of Dutch intonation.

Melodic recipe:



3.2.1.4 Construction of reference utterances

3.2.1.4.1 Production of Reference Utterances

The author read the sentences, which were supplied with the melodic recipes as illustrated above. Recordings were made while seated in a sound treated booth, using a Sennheiser MKH 50 P48 microphone with a mouth-to-microphone distance of approximately 20 cm. The author repeated each sentence until she was confident that the melodic recipe accompanying the sentence, had been adequately produced. The author’s ‘adequately produced’ realizations (as judged perceptually by the author) were stored on disk.

Unfortunately, approximately one quarter of these realizations deviated from the melodic recipe. However, pitch events in the Grammar of Dutch Intonation are specified in terms of timing (in the syllable), direction and excursion size, making it possible to construct a synthetic F₀-contour and replace the deviations in the F₀-contour. This was done as follows:

For each of the author’s realizations, F₀ was determined using sub-harmonic summation (Hermes, 1988). The F₀ contours were replaced with artificial ones using the PSOLA analysis-by-synthesis technique (Moulines & Laroche, 1995). These synthetic contours were close-copy stylized, as described in the previous chapter, section 2.4.1.1 (De Pijper, 1983).

The PSOLA synthetic contour was manipulated to correct the deviations in the author’s F₀-contour, so that the pitch events matched the pitch events stipulated by the melodic recipe. In this way, we obtained reference utterances containing the melodic recipes.

3.2.1.4.2 Evaluation of Reference Utterances

Since the participating speakers were expected to imitate the reference utterances accurately, the (manipulated) reference utterances had to sound *natural* (in terms of the speech melody), and *acceptable* (in terms of the overall quality and intelligibility).

To verify these properties, two listeners, L1 and L2, judged the reference utterances. The utterances were presented over headphones in a quiet

environment. L1 and L2 were not familiar with the GDI, as the naturalness of the reference utterances was more important than absolute correctness in terms of the GDI. For each utterance, L1 and L2 were asked to rate the two properties mentioned above (0 being poor and 10 being excellent). The reference utterances that received poor ratings (receiving a rating of 5 or less for either property) were discarded. Utterances that were judged to be of insufficient quality or not completely natural or intelligible (receiving a rating of 6 or 7) were re-recorded and again subjected to the same process.

This process (recording of spoken sentences, modeling of F0-contour, perceptual rating and subsequent modeling of utterances) was repeated twice. Of the original 25, 11 utterances were discarded, because of synthetic, and sometimes unnatural sounding speech melodies. This left 14 reference utterances. There were five sentences with two different melodic realizations, one sentence with three different melodic realizations, and one sentence with one melodic realization.

A different set of 5 reference utterances was constructed in a similar procedure, to be used as practice material.

3.2.1.5 Feasibility of imitation task

Imitation is a natural skill that humans use, for example, in language acquisition (Sloate & Voyat, 1983). Through imitation, a person attempts to accurately reproduce what is perceived.

The reference utterances that remained after the evaluation should contain melodies that are easy to perceive and imitate. To verify that the speech melodies were not too complex or too subtle and to ensure that it was possible to imitate the reference utterances, L1 and L2 were also asked to listen to and imitate the reference utterances.

L1 and L2's imitations were recorded and stored on disk. The F0 contours in their utterances were determined, using the pitch detection algorithm mentioned above, and compare to the original reference utterances' F0 contours. Visual and auditory inspection of the speech signal showed that L1 and L2's imitations corresponded to the reference utterances in terms of number, type and position of pitch events. This indicated that the reference utterances were suitable for the imitation task for which they had been designed.

3.2.1.6 Procedure of imitation task

The TE speakers indicated that they had been able to imitate simple tunes before their laryngectomy. WLM and WLF were trained phoneticians who were familiar with performing speech imitation tasks.

The reference utterances (practice and test sets) were presented to the non-F0 speakers over headphones. Each reference utterance was presented as often as a speaker wished. Speakers were instructed to give special attention to the speech melody. They then imitated the reference utterance as accurately as possible. Speakers repeated their utterances until they were satisfied with the result.

The recordings of WLM and WLF were made in a sound-treated booth. A Sennheiser MKH 50 P48 microphone was used with a Sennheiser MZA 14 P48 amplifier and Beyerdynamic DT 250 DAT recorder. Recordings of TE1, TE2 and TE3 were made in a quiet environment, using a condenser microphone (Sennheiser electret reference ME 40) at a mouth-to-microphone distance of about 30 cm. The speech signals were recorded on a portable DAT recorder (AIWA HHB 1PRO, sample frequency 48 kHz).

The speakers indicated which of their utterances was the most accurate imitation of a reference utterance and these were stored on disk.

There were 70 non-F0 utterances (5 non-F0 speakers x 14 utterances per speaker) and five practice utterances per speaker. These utterances were used in three different perception experiments. Each perception experiment was designed to answer one of the research questions in the Introduction. The first perception experiment is described in the next section.

3.3 PERCEPTION OF NON-F0 SPEECH MELODIES

Speech melody consists of a sequence of pitch events that include prominence-cueing events, as well as pitch tunes that signal, for example, sentence type (Ladd, 1996). In the third, small-scale perception experiment presented in the previous chapter (section 2.5), we saw that experienced phoneticians seemed to perceive something pitch-like, also in words that were not cued to carry accent. To confirm this finding, the first research question in the Introduction of this chapter was: “Do *naïve* listeners perceive speech melody *at all*, in whispered and non-F0 alaryngeal speech?” The non-F0 utterances were presented to listeners in a perception experiment and

listeners were asked to judge *if* speech melody was present in each of these utterances.

We expected that listeners would perceive speech melody to a similar extent in both whisperers and alaryngeal speakers. The alaryngeal speakers were selected non-F0 speakers and it is conceivable that alaryngeal speakers without F0 will convey similar pitch-like phenomena as perceived in whispered speech. We also expected that utterances containing a large number of pitch events might more often be judged as containing speech melody than utterances containing only one pitch event.

3.3.1 METHOD

3.3.1.1 Stimulus material

All the speakers' test utterances, described in the previous section, were used as test items in this experiment. There were 70 test items (5 non-F0 speakers x 14 utterances per speaker).

3.3.1.2 Listeners

Eighteen listeners between the ages of 18 and 35 participated. All listeners were native speakers of Dutch and all reported normal hearing. Listeners were not informed about the purpose of the experiment. They were unfamiliar with tracheoesophageal speech, and inexperienced in speech evaluation. The listeners were paid for their participation.

3.3.1.3 Procedure

There were two tasks for each utterance. In the first task listeners judged if speech melody was present (this experiment). In the second task, listeners rated the similarity between non-F0 speech melodies and the reference utterances' speech melodies (next experiment). Only the first task will be described in this section.

Listeners were seated in a sound-treated room with a computer screen in front of them.

At the top of the screen were 'play' buttons. One button had the text "original". Only this button was used in this experiment, the second 'play' button was used in the next experiment. By clicking on the "original" button, an utterance produced by one of the five speakers was presented over

headphones. There were two questions on the computer screen. Only the first was relevant for this experiment. The text of question 1 was the Dutch equivalent of: “Do you hear speech melody in the ‘original’ utterance?” and was positioned underneath the play buttons. Next to this question, a “yes” and “no” button were positioned. The listeners could listen as often as they wanted to the “original” and then click on “yes” if they thought the utterance contained speech melody or click on “no” if they could not perceive any melody. Once the listeners had completed this task for one test item, they could continue with the experimental task of Experiment 2, described in the next section.

There were 25 practice items (5 speakers x 5 practice utterances per speaker), which were presented separately before the actual experiment started to familiarize listeners with the task. After the practice session, listeners could clarify any uncertainties they might have had.

3.3.2 RESULTS

The answer to the first research question, “do naïve listeners perceive speech melody *at all*, in whispered and non-F0 alaryngeal speech?” is positive: Speech melody was perceived in 91% of the non-F0 speakers’ utterances. We conclude that listeners perceived speech melody even when the most important cue to speech melody, F0, was absent. Table 3.1 gives the results. Table 3.1 shows that TE1 achieved the highest score and TE3 the lowest. TE3 also had the largest range, but the minimum value that indicates that half of the listeners did not perceive speech melody in TE3, concerned only one of this speaker’s utterances. TE1 and WLF seemed to be the most alike. There does not seem to be much difference among speakers.

Table 3.1 Average percentage of utterances in which speech melody was perceived; Standard error (in parentheses), and range (across utterances) are also given. Given per speaker, pooled over utterances (14) and listeners (18).

Speaker	Mean (s.e.)	Range
TE1	96% (1%)	83% – 100%
TE2	88% (2%)	72% – 94%
TE3	85% (2%)	56% – 100%
WLM	90% (2%)	72% – 100%
WLF	95% (1%)	83% - 100%

Logistic regression (Kleinbaum, 1992; Hosmer & Lemeshow, 2000) was used to determine the differences between speakers and the possible

influence of the number of pitch events on the perception of speech melody. The dependent variable (perceived speech melody) was dichotomous: hit (1) if a listener perceived speech melody, and miss (0) otherwise. The listener responses were aggregated into proportions of hits. Logistic regression transforms the dependent variable, a proportion, into a logit variable (the natural log of the odds ratio of the hits and misses). The logistic regression model was used to assess which factors influenced the occurrence of ‘yes’ responses. The individual speakers and the number of pitch events (one to four) formed independent categorical variables or predictors. WLF was the same speaker as the Reference speaker. Thus, she was the only speaker who was familiar with the melodic recipes. We therefore used WLF as the reference or baseline category for Speakers. The means for the other speakers were expressed as deviances (in logits) from the logit average of WLF. The category ‘four pitch events’ (the maximum) was used as reference category for the effect of number of pitch events. The means of the other numbers of pitch events were expressed as deviances (in logits) from the logit average of the ‘four pitch events’ condition. If an independent variable’s coefficient was positive, this meant that this variable scored higher on ‘yes, speech melody was perceived’ than the reference category, if negative, then lower. A significant logistic regression coefficient means that the predictor is significantly different from the reference category. Logistic regression results of the best fitting model are presented in Table 3.2.

Table 3.2. Logistic regression coefficients (with standard error), for the effects of speaker and number of pitch events, on the proportion of hits (the dependent variable was: ‘yes, speech melody was perceived’). The reference category was WLF for speakers and ‘four pitch events’ for number of pitch events

variable	coefficient (S.E.)
intercept	3.155** (0.337)
TE1	0.176 (0.421)
TE2	-0.956** (0.345)
TE3	-1.195** (0.336)
WLM	-0.665 (0.358)
One pitch event	-0.480 (0.279)
Two pitch events	-0.552* (0.259)
Three pitch events	0.326 (0.335)
-2 Log likelihood	741.331
Nagelkerke R ²	0.064

** indicates that the coefficient is statistically significant at the 0.01 level, and * indicates that the coefficient is significant at the 0.05 level.

The log likelihood and R^2 suggest that this logistic model explains the data reasonably well. The intercept means that the overall proportion of hits for WLF's four pitch events can be calculated as $\log(P/(1-P)) = 3.155$. This corresponds to $P = 0.959$.

TE2 and TE3 differed significantly from WLF, but TE1 and WLM did not. Thus, although there were individual differences between speakers, these differences were not related to a specific group. We therefore chose to combine whispering and alaryngeal speakers into one non-F0 speaker group for the remainder of this study.

Table 3.2 further shows that, although utterances with fewer pitch movements scored fewer hits (lower proportion of 'speech melody was perceived' responses), only the utterances with two pitch events differed significantly from the utterances containing four pitch movements, but utterances with one pitch event did not.

Obviously, listeners still perceived speech melody in the absence of F0. This might indicate that variation of acoustic cues other than F0 were sufficient to convey the presence of speech melody. However, the reader has to bear in mind that the perception of speech melody does not necessarily imply that listeners perceived the same speech melody as specified by the melodic recipes and as intended by the speaker. Perceived speech melodies *could* have been fictitious (in the mind of the listener), or listeners might have interpreted acoustic cues not related to the speakers' intention as speech melody. The next experiment investigated to what extent listeners perceived the *intended* speech melodies.

3.4 COMPARING NON-F0 AND F0 SPEECH MELODY

The second research question was: "do naïve listeners perceive the *intended* speech melodies in non-F0 speech?" Listeners compared the speech melody in the non-F0 speakers' utterances with the melodic realization in the reference utterances and rated the degree of similarity.

When a speech melody is judged as an accurate imitation, it indicates that the pitch events are perceived as categorically the same in both utterances ('t Hart, et al., 1990). We therefore surmise that if listeners perceive all the pitch events in a non-F0 utterance to be the same as the pitch events in the reference utterance, the degree of similarity will be judged as very high. If none of the pitch events are the same, the degree of similarity will be judged

as very low. Listeners may, of course, only perceive some, but not all pitch events in a non-F0 utterance to be the same as in a reference utterance, so that the similarity is neither very good nor very poor.

The non-F0 utterances were paired with the reference utterances to create test and control pairs. One member of a pair was therefore always a reference utterance, and the other member a non-F0 utterance. In both the test, and the control pairs the non-F0 speaker's utterance *always* matched the reference utterance *segmentally*.

As already discussed above, non-F0 speakers indicated which of their utterances were the most accurate imitations of the reference utterance. Thus, each speaker produced, for example, utterance 1, which corresponded to reference utterance 1. A test pair consisted of these matching utterances (non-F0 utterance 1 and reference utterance 1, etc.). This meant that in the test pairs, the number and type of intended pitch events in the non-F0 speaker's utterance closely resembled those in the reference utterance.

In contrast to the test pairs, control pairs were non-matching in terms of number and type of pitch events. The compilation of control pairs was possible since multiple speech melody versions existed of most of the sentences (see section 3.2.1.3, and Appendix 3). Thus, we combined a non-F0 speaker's utterance 1 with, for example, reference utterance 2. These combinations were random; the only criterion being that the utterances did not match in terms of speech melody.

If listeners recognised the non-F0 speakers' intended melody, one would expect higher similarity ratings for the matching test pairs as compared to the similarity ratings for the non-matching control pairs.

3.4.1 METHOD

3.4.1.1 Stimulus material

As was mentioned above, there were two types of stimuli: test pairs and control pairs. There were 70 test pairs (5 speakers x 14 utterances). A test pair consisted of a reference utterance plus the speaker's imitation of that reference utterance. There were 25 control pairs (5 speakers x 5 utterances). Control pairs consisted of a speaker's utterance plus a reference utterance, which was identical except for the speech melody pattern: the non-F0 speaker's utterance did not match (the speech melody pattern of) the reference utterance.

3.4.1.2 Procedure

The 18 listeners, still seated in the sound-treated booths, compared the reference utterances with the non-F0 speakers' utterances. Listeners concentrated on the speech melody and ignored (poor) intelligibility. In a try-out it was found that listeners preferred the reference utterance to be presented as imitation of the non-F0 utterance, because that made it easier to focus on the speech melody and disregard voice quality and intelligibility. In the perception experiment the *non-F0 speakers'* utterances were therefore presented as if they were "original" and the *reference* utterances were presented as if they were imitations of the non-F0 users' utterances.

At the top of the screen were two play buttons. One button had the text "original"; the other button had the text "imitation". By clicking on the "original" button, an utterance produced by one of the five non-F0 speakers was presented over headphones (see also previous experimental task, described in section 3.3.1.3). By clicking on the "imitation" button, a reference utterance was presented over headphones. The Dutch version of the text of the second question, "How good is the 'imitation' when compared to the 'original'?" was positioned towards the bottom of the screen. Underneath this question was a sliding scale marked "very poor" on the left and "very good" on the right. Listeners compared the "original" with the "imitation" as often as they needed and then moved the button on the sliding scale. The subsequent rating could be any value between and including 0 (very poor) and 99 (very good).

Once the listeners had completed both experimental tasks (previous experiment and this experiment, see 3.3.1.3), a button with the text "next" appeared. If the listener clicked on this button, the next set of utterances could be judged.

3.4.2 RESULTS

As mentioned above, the ratings of the matching test pairs were expected to be higher than the ratings of the non-matching control pairs. We therefore expected a larger percentage of 'very good' ratings for the test pairs, and a larger percentage of 'very poor' ratings for the control pairs (0 = very poor, 99 = very good). However, if listeners did not perceive the non-F0 speakers' intended melody, there should not be a difference in the way test and control items were rated. Figure 3.3 shows the percentage of listener judgments, as distributed over the rating scale. In this figure, the rating scale was divided

into five categories, to illustrate more clearly the differences between the two conditions.

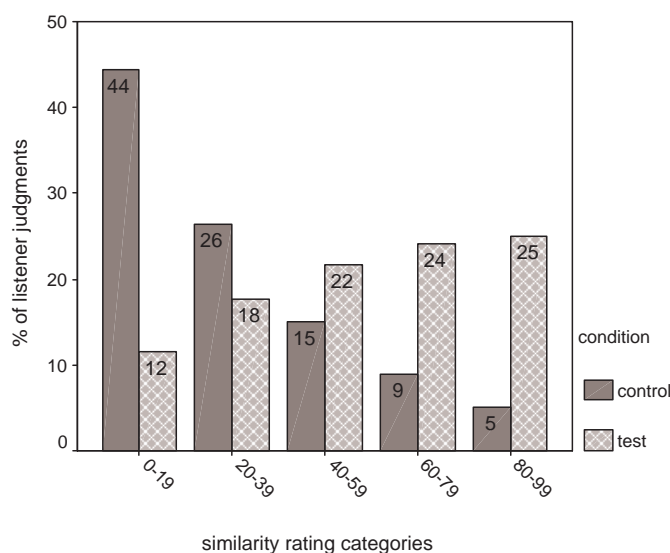


Figure 3.3. Similarity ratings: distribution of listener judgments over the rating scale (divided into categories), given for matching test and non-matching control conditions.

Figure 3.3 shows that for most of the control pairs, the similarity ratings are poor: the lowest category (similarity rating of 0-19) contains the highest percentage of judgments. The higher the rating category, the lower the percentage of listener judgments. The opposite is true for the test condition, although judgments were distributed more evenly over the different categories. The test condition differed significantly from the control condition, as calculated over the original distributions (Kolmogorov-Smirnov, $z = 7.605$, $p < 0.001$). Thus, listeners could generally hear whether the speakers' utterances did match, or did not match the reference utterances. In the matching test condition, many of the non-F0 speech melodies were judged neither as very similar, nor as very dissimilar, when compared to the reference speech melodies (Figure 3.3, middle three rating categories). As explained above, we surmise that the degree of similarity is related to the number of pitch events that are perceived as being the same as in the reference utterance ('t Hart, et al., 1990; see explanation in section 3.4). Based on this supposition it seems that many non-F0 utterances contained

pitch events that were perceived as being different from the intended pitch events: many of the speaker's utterances were not rated as very similar to the reference utterances, in the test condition. In the next section, we therefore investigate which specific pitch events were correctly perceived and which were not.

3.5 TRANSCRIPTION OF SPEECH MELODY

The third research question was: “are certain non-F0 pitch events perceived more accurately than other pitch events?” We determined that naïve listeners perceived speech melody in non-F0 speech, and recognised parts of the intended speech melodies, but do not know which parts. Accurate perception of certain pitch events might be related, for example, to the degree of prominence cued by the pitch event.

Two expert listeners transcribed the pitch events that were perceived in the reference utterances and the non-F0 speakers' utterances. Subsequently intra- and inter transcriber agreement was calculated, as well as the agreement between transcriptions (of F0 utterances and non-F0 utterances) and the melodic recipes.

3.5.1 METHOD

3.5.1.1 Stimulus material

The stimulus material consisted of the utterances produced by the five non-F0 speakers and the reference speaker. In total, 84 utterances (14 sentences x 6 speakers) were presented.

3.5.1.2 Transcribers

Two phoneticians with expert knowledge of the IPO-GDI were asked to transcribe the test utterances, using the GDI symbols.

3.5.1.3 Transcription task

The transcribers (TR1 & TR2) transcribed the pitch events they perceived. Separate transcription forms were provided, one for each speaker. A transcription form consisted of the written test sentences plus space underneath to add the transcription symbols. There was no time limit and the

transcribers were allowed to listen to each utterance or part of an utterance as often as they wished. They could also alter any transcriptions until they were satisfied.

3.5.1.4 Order of transcriptions

The 84 utterances (70 non-F0 and 14 reference) were transcribed twice. The non-F0 speakers' utterances were transcribed with a minimum of two months between the first and second transcription. After the transcriptions of the non-F0 speakers' utterances were completed, the reference utterances, each time presented in a different order, were transcribed.

3.5.2 RESULTS

This experiment was set up to investigate which non-F0 pitch events were perceived. We first determined transcriber agreement as well as the accuracy with which the transcribers transcribed the reference utterances. If these agreements were good, the transcription of pitch events in non-F0 utterances was also taken to be reliable. Agreements were calculated using Cohen's kappa (κ) (Landis & Koch, 1977):

0.00 – 0.20:	Slight agreement
0.21 – 0.40:	Fair agreement
0.41 – 0.60:	Moderate agreement
0.61 – 0.80:	Substantial agreement
0.81 – 1.00:	Almost perfect agreement

When perceptually analysing a pitch contour in the IPO-GDI tradition, each syllable of the utterance is assigned one or more pitch events, depending on which events occur on a syllable ('t Hart, Collier and Cohen, 1990). Thus, agreement between transcriptions was calculated for configurations within syllables: successive pitch events within a syllable had to agree.

In the non-F0 utterances, the transcribers did not use some of the possible transcription categories. To determine agreements, we needed an equal number of categories for each transcription of each speaker. The number of categories were therefore slightly simplified to seven transcription categories, based on the degree of prominence as described by 't Hart and

Cohen (1973), 't Hart and Collier (1975) and 't Hart, Collier and Cohen (1990). The modified categories are given in Table 3.3.

Table 3.3. Description of transcription categories

Transcription Category	GDI symbols	Clarification
1	1	full rise = prominence
2	2	final rise
3	4	gradual rise
4	1A, 5A	rise-and-fall = prominence
5	A	full fall = prominence
6	D, E	gradual fall / half fall
0		no pitch change

3.5.2.1 Transcriber agreement

For each transcriber, the agreement between the first and second transcription was calculated, for the reference utterances as well as for the non-F0 utterances. The intra-transcriber agreement for the reference speaker's utterances was good: for TR1 $\kappa = 0.85$ and for TR2 $\kappa = 0.92$. The intra-transcriber agreement for the non-F0 speakers' utterances was substantial: for TR1 $\kappa = 0.71$ and for TR2 $\kappa = 0.77$. The agreement between the TR1 and TR2's first transcriptions, as well as between their second transcriptions was calculated for the reference utterances as well as for the non-F0 utterances. Inter-transcriber agreement was also substantial: for the reference utterances $\kappa = 0.70$ for the first transcriptions and 0.76 for the second transcriptions. For the non-F0 utterances $\kappa = 0.63$ for the first transcriptions and 0.64 for the second transcriptions. We conclude that the transcriptions were sufficiently reliable for our purposes.

3.5.2.2 Confusion matrices

To investigate which pitch events were perceived accurately, confusion matrices were compiled. One confusion matrix compared the transcriptions of the F0 utterances with the melodic recipes, and one confusion matrix compared the transcriptions of the non-F0 utterances with the melodic recipes.

The agreement between *each* transcription (2 transcribers x 2 transcriptions x 2 utterance types (F0 and non-F0) and the intended pitch events (as prescribed by the *melodic recipes* on paper) was also calculated to determine the accuracy of the transcribed pitch events.

The confusion matrices give information on the pattern of confusions: which pitch events are confused, which pitch events are not perceived, and if pitch events are perceived when no pitch event was intended.

Table 3.4 *Confusion matrix for F0 and non-F0 utterances' pitch categories compared to the intended pitch categories. See Table 3.3 for clarification of transcription categories*

A

Intended pitch category	Transcribed pitch category (F0 utterances)							Total
	0	1	2	3	4	5	6	
0	177 (96%)	2	1	0	0	0	4	184 (100%)
1	0	47 (98%)	0	0	1	0	0	48 (100%)
2	0	0	4 (100%)	0	0	0	0	4 (100%)
3	6	3	0	35 (79%)	0	0	0	44 (100%)
4	0	1	0	0	21 (75%)	6	0	28 (100%)
5	0	0	0	0	1	27 (96%)	0	28 (100%)
6	17	0	0	0	2	7	18 (41%)	44 (100%)
Total	200	53	5	35	25	40	22	380

B

Intended pitch category	Transcribed pitch category (non-F0 utterances)							Total
	0	1	2	3	4	5	6	
0	816 (87%)	30	13	14	23	4	20	920 (100%)
1	30	133 (55%)	0	2	63	8	4	240 (100%)
2	0	0	17 (85%)	0	0	3	0	20 (100%)
3	145	47	0	26 (12%)	2	0	0	220 (100%)
4	4	10	0	0	70 (50%)	56	0	140 (100%)
5	19	7	4	0	15	86 (61%)	9	140 (100%)
6	148	10	1	0	22	29	10 (5%)	220 (100%)
Total	1162	237	35	42	195	186	43	1900

We can also investigate if the pattern of confusions, omissions and additions is the same for F0 and non-F0 utterances.

Overall, transcriptions of the reference utterances (Table 3.4 A) were very accurate when compared to the melodic recipes: 329 out of the intended 380 pitch events were perceived correctly (87%). For TR1 $\kappa = 0.72$ for the first transcription, and 0.86 for the second transcription. For TR2 $\kappa = 0.82$ for the first, and 0.80 for the second transcription, indicating that for both transcribers the agreement between the transcriptions and the melodic recipes was good.

Table 3.4B gives the confusion matrix for the non-F0 utterances. Overall, 1158 out of the 1900 pitch events (61%) were perceived correctly in the non-F0 utterances; this is much lower than for the F0 utterances. For TR1 $\kappa = 0.36$ for the first transcription and 0.46 for the second. For TR2 $\kappa = 0.45$ for the first, and 0.40 for the second transcription, indicating that for both transcribers the agreement between the non-F0 utterances and the melodic recipes was fair to moderate.

The answer to the third research question, whether certain non-F0 pitch events are perceived more accurately than others, is positive. Abrupt changes (full fall and - rise and rise-and-fall) were transcribed much more accurately than gradual pitch changes (categories 3 and 6). Table 3.4 (B) further shows that full falls were transcribed more accurately than full rises and rise-and-fall. This might be explained by the finding that falls are more conspicuous than either rise or rise-and-fall (Hermes & Rump, 1994). Rise-and-fall was also confused with fall, which is possibly a recency-effect caused by the more conspicuous fall. Note however that although rise-and-fall is often confused with fall, rise is not. Neither is fall confused with rise.

We conclude that non-F0 speech contains something pitch-like that simulates F0 pitch to some extent.

3.6 GENERAL DISCUSSION

Pitch perception is strongly related to the F0-variations in an utterance, but this study shows that listeners also perceived something pitch-like in the absence of F0. It is unclear to what extent this pitch-like something fits the ASA definition of pitch given in the Introduction: the ordering of sound on a musical scale, but we do not really imagine that the non-F0 alaryngeal speakers could sing clearly audible musical tunes.

The first perception experiment showed that ‘pitch’ was perceived to the same extent in both the whisperers and the alaryngeal speakers. However, the second and third perception experiment revealed that not all the pitch-like events were related to *intended* pitch events. In the second experiment, listeners rated only some of the non-F0 utterances as good imitations (according to ‘t Hart, et al., 1990, utterances are good imitations when the types of pitch events are perceived as categorically alike). Thus, the utterances that did not receive a good rating contained pitch events that were perceived as different from the intended pitch events. This was confirmed by the third perception experiment: agreement between the non-F0 utterances and the melodic recipes was moderate at best, and more importantly, the confusion matrices showed that more errors occurred in the non-F0 utterances than in the F0 reference utterances.

Comparison with results of alaryngeal speakers found in literature is not straightforward, because most studies concerned tone languages. However, since so little information is available on pitch in alaryngeal speech, a comparison is still worthwhile. In contrast to the present study, results were averaged over F0 and non-F0 speakers. Generally, it seems that even when F0 alaryngeal speakers were included, results for alaryngeal speakers in tone languages were poorer than the results in the present study on an intonation language. For instance, accurate perception of up to 6 tones in Cantonese tracheoesophageal speakers resulted in an average of 52% (Ching & Williams, 1994). Average percentages given by Wong, et al. (1997), and Yiu, et al. (1994), likewise on perception of Cantonese tones in alaryngeal speakers, were 59% and 50% respectively. In the present study, the perception of the (6 possible) intonational pitch events was 61%. Results by Gandour et al. (1988) indicated that correct perception of 5 tones was only 39% in a non-F0 alaryngeal Thai speaker, compared to 70% in an alaryngeal speaker with F0. The absence of F0, and the type and number of tones might affect correct perception of pitch. This seems to be confirmed by a study on intonation: Gandour and Weinberg (1983, 1985) found that non-F0 alaryngeal speakers conveyed simple questions and statements correctly in 81% (for F0 alaryngeal speakers this was above 95%). The non-F0 speakers in the present study succeeded in conveying final rises and full falls (comparable to question versus statement), but slightly less accurately (73%). However, in the present study the utterances were longer and contained more pitch events of different types. The task of the listeners in Gandour and Weinberg’s study was also a simple choice (question or

statement), whereas the transcribers in the present could choose among 10 pitch events. Overall, these results indicate that an alaryngeal speaker's performance will be most affected when F0 is *absent* and/or the language concerned is a *tone* language.

The explanation that Gandour et al. (1988) gave for the discrepancy between a tone and intonation language in alaryngeal speech, was unfortunately based on the results of the F0-alaryngeal speaker, who could not produce a fast rate of change in F0. They surmise that the faster rate of change needed to convey a tone, may exceed limitations of the voicing source. We noticed in this study that *gradual* pitch changes were generally not conveyed in the non-F0 utterances: in the absence of F0, the perceptual reality of gradual non-F0 pitch changes seemed doubtful, whereas *abrupt* pitch changes seemed to be perceptually more robust. This suggests that the strategy used by alaryngeal non-F0 speakers to convey something pitch-like differs from the (limited) voicing source variation used to convey F0-pitch.

Given that listeners in the present study perceived something pitch-like, the question arises what the acoustic correlates of this perceived 'pitch' might be. A number of studies have shown that pitch perception in whispered speech is related to formant frequencies, especially F2 (Thomas, 1969; McGlone & Manning, 1979; Higashikawa et al., 1996). Whisperers also increased high frequency energy when producing a rise and decreased high frequency energy when producing a fall (Krull, 2001; Meyer-Eppler, 1957). Alaryngeal speakers without F0 might have adopted a similar 'pitch' strategy to whisperers.

Understanding which strategy non-F0 speakers use is important, because it might also have implications for rehabilitation: training non-F0 speakers, who clearly have limited capabilities, to convey musical scales, would be unproductive. However, non-F0 speakers might benefit from purposeful training of perceptually robust rising and falling pitch.

The next chapter investigates if the non-F0 speakers in the present chapter use similar strategies as whisperers to convey pitch.

4

In Search of non-F0 Pitch

ABSTRACT

This chapter investigates if listeners perceive the intended pitch direction in tracheoesophageal and whispered speech, in which F0 is absent. First, naïve listeners imitated the perceived speech melody in stimuli from 3 alaryngeal speakers and 2 laryngeal control speakers (1 whispering, 1 whispering and phonating normally). Listeners were able to imitate rising and falling pitch on target stretches of speech, but this was probably due to the influence of an intonation bias.

Second, the excised stretches of speech (speech fragments) were filtered into 5 frequency bands. Listeners identified the direction of pitch, in the excised speech fragments and their filtered bands.

Third, the speech fragments were analysed acoustically for possible correlates of non-F0 pitch. Results suggest that one alaryngeal speaker produced a semblance of periodicity in the lowest band, which listeners interpreted as pitch; speech fragments of the other 2 alaryngeals contained no consistent perceptual cues to pitch; and the 2 whisperers modified spectral tilt, which was correlated to listeners' perceived direction of pitch.

4.1 INTRODUCTION

As already mentioned in the previous chapters, pitch is important in speech communication, because it facilitates the perception of linguistic functions in a spoken utterance. In Dutch, an intonation language, pitch signals linguistic functions such as sentence accents and type of sentence (question or statement). The pitch events that signal these linguistic functions are components of the utterance's speech melody. A speech melody consists of successive pitch rises and pitch falls, and listeners rely on this consecutive pattern of rising and falling pitch to recognise the speech melody ('t Hart, Collier & Cohen, 1990).

A *speaker* conveys rising and falling pitch by increasing or decreasing the fundamental frequency (F0). Increasing and decreasing F0 can therefore be regarded as the acoustic correlate of the rising and falling pitch that is perceived in speech melody.

This chapter investigates if an alternative to F0 pitch can be found in the utterances produced by the non-F0 speakers from the previous chapter: can speakers TE1, TE2, TE3, WLM and WLF, in whose speech F0 appears to be absent, convey rising or falling pitch? As was explained in the Introduction to chapter three, previous studies on pitch in alaryngeal speech did not distinguish between non-F0 and F0 speakers, but the previous chapter on the perception of speech melody in whispered and non-F0 alaryngeal speech indicated that several of the intended rising and falling pitch events were often perceived.

Since linguistic functions of pitch are associated with changing pitch over time, an alternative (non-F0) pitch should be able to operate in a similar fashion, if it is to be a useful substitute of F0. The existence of an alternative to F0 pitch has been a topic of investigation for nearly fifty years. Studies in whispered speech that have looked at time-varying non-F0 pitch showed, for instance, that "tones" could be perceived in whispered Mandarin (Jensen, 1958). Yet, these results could not be replicated for whispered Vietnamese (Miller, 1961).

Krull (2001) looked at the perception of Estonian word prosody in whispered speech. In her study, listeners could accurately differentiate between falling and rising versions of the stimulus word. The whisperers participating in Krull's study increased high frequency energy when producing a rise, and decreased high frequency energy when producing a fall. However, these whisperers also varied duration, which might have been

an important cue to pitch perception. Meyer-Eppler (1957) looked at short whispered utterances, produced first as statement and then as question. Spectrographic analyses of these utterances showed two substitutes for F0: intensity in the higher frequencies was increased and an extra formant appeared when a pitch rise was realized. Unfortunately, it was not investigated if listeners actually perceived these F0 substitutes.

From the above studies it seems that listeners did perceive rising and falling pitch in non-F0 speech to a certain extent, but it is uncertain which acoustic cues caused this perceived pitch.

A systematic search for an acoustic correlate of non-F0 pitch has focussed mostly on vowels, and as far as we know, mostly in whispered speech, not in alaryngeal speech. A number of studies have shown that pitch perception in whispered vowels is related to formant frequencies, especially F_2 (Harbold, 1958; Thomas, 1969; McGlone & Manning, 1979). Higashikawa, et al. (1996), showed that listeners perceived the difference in height between, for example, an /a/, that was whispered at “high, “normal” or “low” pitch levels and that there was a relation with the frequency height of F_1 and F_2 . A study on synthetically generated ‘whisper’ vowels further showed that simultaneous increase or decrease of F_1 and F_2 had a greater effect on the perception of pitch (Higashikawa & Minifie, 1999). Unfortunately, these studies were limited to sustained vowels in which the formant tracks were fixed over time, whereas listeners should also perceive the *variations* in the non-F0 pitch contour, if it is to be a true alternative to F0 pitch. Remez & Rubin (1993) investigated intonation of sinusoidal sentences in which the fundamental was excluded, but which contained four time-varying sinusoids representing formants. Listeners consistently selected the tonal contour representing F_1 as matching their impression of intonation. Remez & Rubin relate this effect to the dominance region: the auditory system is roughly keyed to detect pitch from excitation in the range of 0.4 – 1 kHz, which is also roughly the range of F_1 . None of these studies investigated perception and production of rising and falling pitch, as intended by the speaker. It is therefore still unclear if an alternative pitch exists that can fulfil a similar linguistic role to F0.

The aim of the present chapter is therefore to investigate the existence of non-F0 pitch, but we looked at rising and falling pitch, as intended by the non-F0 speakers mentioned above, and produced within the context of a normal utterance’s speech melody. When non-F0 pitch exists under these circumstances, it can be thought of as an alternative to F0 pitch.

The utterances used in the previous chapter contained stretches of speech that differed only in the intended pitch direction: rising versus falling pitch. These utterances are used as stimulus material in the present chapter, precisely because non-F0 speakers produced them.

Normal laryngeal phonation is the result of myoelastic-aerodynamic properties that cause the vocal folds to vibrate (Van den Berg, 1958). The frequency of vibration (F0) is determined primarily by adjusting the length and tension of the vocal folds (e.g., Hirano & Bless, 1993). During whispering there is no periodic vibration of the sound source. In the previous chapter, the participating tracheoesophageal speakers produced an acoustic signal that was highly unstable and in which periodicity was predominantly absent. Thus, both the whispering laryngeal speakers and the tracheoesophageal speakers participating in the previous chapter on speech melody were classified as non-F0 speakers.

The sound source that these non-F0 speakers rely on could be described as a noise source: in the absence of any consistently vibrating structures that could generate F0, the airstream becomes turbulent and generates noise when it moves through a constriction. By systematically modulating the constrictions and thus, the noise, non-F0 speakers could have conveyed an utterance's speech melody and thus the direction of pitch in the stretches of speech mentioned above. If listeners can perceive the intended direction of pitch, it would mean that an alternative pitch exists that may fulfill a communicative function similar to F0. The research questions addressed in this chapter are as follows:

1. Can naïve listeners perceive rising and falling pitch in non-F0 speech?
2. Which acoustic information do listeners use to perceive rising and falling pitch?
3. Which acoustic information do speakers use to convey rising and falling pitch?

4.2 IMITATION EXPERIMENT

The aim of this chapter was to determine whether an alternative pitch (unrelated to F0) exists that has a similar communicative function as F0: an alternative pitch that conveys the intended rising or falling pitch.

An imitation task was chosen to answer the first research question: “can naïve participants perceive rising or falling non-F0 pitch?”

As was mentioned in chapter three, imitation is a natural skill found in language acquisition (Sloate & Voyat, 1983). Through imitation, a person attempts to accurately reproduce what is perceived. If an alternative to F0 pitch exists, imitators might be able to reproduce the intended rising and falling pitch contained in the non-F0 utterances' speech melody, similar to how rising and falling pitch is reproduced in F0 utterances.

4.2.1 METHOD

4.2.1.1 Speakers

Utterances produced by the five speakers described in the previous chapter were used as stimulus material (see 3.2.1.1): the normal-speaking female laryngeal speaker (R: reference), who also whispered the utterances (WLF: whispering laryngeal female); a whispering laryngeal male speaker (WLM); and three tracheoesophageal speakers (TE1, TE2 & TE3).

The normal-speaking laryngeal speaker's utterances were the only utterances that contained consistent, regular fundamental frequency; they were used as reference and control utterances. The whispered utterances were included because we could be certain that they would not contain harmonics. With the TE speakers we aimed at including non-F0 speakers, based on the acoustic signal typing developed for tracheoesophageal speech (Van As, 2001, described in the previous chapter).

4.2.1.2 Stimulus material

The stimulus material in the present chapter was a selection of a larger set of utterances from the previous chapter on perception of speech melody (see 3.2.1.3).

As explained in the Introduction, these utterances contained stretches of speech that differed in the intended pitch direction.

The original *sentences* contained nine potential contrasting stretches. Four of the nine contrasts were not included in the experiments of this chapter. In three of these contrasts, the Reference speaker's most acceptable and natural spoken renditions deviated too much from the intended pitch events as designed on paper. Furthermore, for another contrast which was meant to be monosyllabic ('warm'), three of the non-F0 speakers produced the speech fragment as bisyllabic /wɑrəm/ in one version and monosyllabic

/warm/ in the other version, hindering proper comparison. This resulted in a very small set of stimulus materials.

The utterances containing the remaining five contrasts are given below with the contrasting stretches printed in italics, bold and underlined:

- 1a (rise) Marianne ***en Willem*** doen allebei raar.
 1b (fall) Marianne ***en Willem*** doen allebei raar.
 (Marian and William act both weirdly)
 2a (rise) Ja wij willen v ***ooral wi*** nnen.
 2b (fall) Ja wij willen v ***ooral wi*** nnen.
 (Yes we want especially to win)
 3a (rise) Wil hij wel weer ***mee?***
 (Will he again come along?)
 3b (fall) Hij nam haar ***mee!***
 (He took her along!)
 4a (rise) Hij wil een ***meloe*** n
 4b (fall) Hij wil een ***meloe*** n
 (He wants a melon)
 5a (rise) Vrij warm maar wel ***mooi wa*** ndelweer.
 5b (fall) Vrij warm maar wel ***mooi wa*** ndelweer.
 (Quite warm, but certainly nice walking weather)

Thus, the utterances constituted the test items, but we especially investigate the bold, underlined, italicised stretches of speech, in this chapter. These stretches of speech were not defined phonologically, but were defined phonetically. They started at the lowest point of an F0 excursion and ended at the highest point of an F0 excursion, or vice versa: beginning at the highest point and ending at the lowest point of an F0 excursion. Thus, speakers' stretches of speech were based on F0-movements as measured in the Reference speaker's stretches of speech, regardless of the boundaries of a syllable (or word). For example, if an excursion originated in the syllable preceding the stressed syllable (or word) in question, this syllable was judged to be part of the stimulus item, or if the excursion ended before the end of a syllable, the remaining part was not judged to be part of the stimulus item (cf. Collier, 1970; Hasegawa & Hata, 1992; Hermes, 1997).

For each speaker, both versions of the *complete* utterances were presented in the Imitation experiment so that the stretches of speech were still 'embedded' in the natural context in which they had been produced. As

explained above, imitation of the complete utterances is an ecologically valid task. Further, the versions of the sentences were designed to be segmentally the same (except ‘mee’), to exclude the risk of a sentential context bias: a stretch of speech held the same position in a sentence, and was surrounded by the same context regardless of the version (rise or fall).

In total 60 utterances were included: 5 utterances x 2 versions per utterance (fall and rise) x 5 speakers (TE1-TE3, WLM & WLF) and the Reference speaker’s utterances (hence: R).

The signal strength of speakers TE1, TE2, TE3, WLM and WLFs’ utterances was substantially lower than those found for the normal speaker. The signal strength for these speakers was therefore amplified, so that all speakers’ utterances were approximately equal (average of 60 dB SPL).

4.2.1.3 Participants

Participants who imitated the utterances (hence: imitators) were 18 native Dutch graduate students. None reported hearing deficiencies. All were paid for their participation.

4.2.1.4 Procedure

Imitators were seated in a quiet environment, wearing headphones and with a microphone placed in front of them. The speakers’ utterances were presented over headphones. The order, in which the non-F0 speakers were presented, was random, but R’s utterances were always presented last. The imitators were asked to concentrate on the speech melody. The sentences were also provided in written form, to which no punctuation was added. After listening (repeatedly) to an utterance, the imitators imitated the utterance.

4.2.1.5 Transcription of imitated utterances

To determine if imitators had identified the speakers’ pitch direction accurately, the author transcribed the relevant stretches of speech, embedded in the imitators’ imitations. Four transcription categories were used: a test item could be transcribed as rise, fall, rise-and-fall or no pitch change. To ensure that the imitators’ pitch events were transcribed accurately, transcriptions were based on perceived pitch, but also on visual inspection of

the F0-contour: pitch events that could not be transcribed with confidence, and could also not be measured in the F0-contour were transcribed as ‘no change’. Vice versa, F0-excursions that could be seen in the F0-contour but were not perceived as a pitch event were also transcribed as ‘no change’. For each of the imitated utterances (produced by the imitators), the F0-contour was obtained using an auto-correlation pitch-detection algorithm (Boersma, 1993). All the stretches of speech were transcribed twice. There was an interval of approximately two months between the first and the second transcription. The agreement between the two transcriptions was 93%, which was considered sufficient for further analysis. The results of the last transcription were used for further analysis, yielding 60 x 18 transcriptions.

4.2.2 RESULTS

If non-F0 pitch exists, we expect imitators to perceive and reproduce the rising and falling pitch, as intended by non-F0 speakers. Thus, the imitators’ pitch direction should correspond to the speakers’ intended pitch direction: they should reproduce rising pitch in the rise version of a test item and falling pitch in the fall version. If non-F0 pitch does not exist, the imitators would not have perceived the intended pitch direction, and then the imitators’ responses over the four categories should be independent of the intended pitch direction: the imitators’ responses are then not expected to be clustered in the rise category (for the intended rise version) or clustered in the fall category (for the intended fall version).

Figure 4.1 gives the results for the rise versions and for the fall versions. The results for the Reference speaker (hence R) on the left in each panel, indicates that imitators were capable of the task. They accurately reproduced the intended pitch direction when F0 was present in the signal: imitators responded with more than 85% rise responses in the rise version and more than 90% fall responses in the fall version. If an effective alternative pitch exists, the non-F0 speakers should display a comparable pattern to R: mostly rises in the rise version and falls in the fall version. Figure 4.1 shows that this is indeed the case, but to a lesser degree than R. Hence, a χ^2 test was performed, in which the non-F0 speakers’ results were compared with R’s results (a χ^2 was chosen, because the imitation task resulted in a number of response categories). The *expected* values were *derived* from the results displayed by R, but to calculate χ^2 , each of R’s categories had to contain at least 5 observations. Therefore R’s categories were adjusted: observations

were transferred from the category containing the largest number of responses to the categories containing too few responses. Separate tests were performed for the rise version and the fall version.

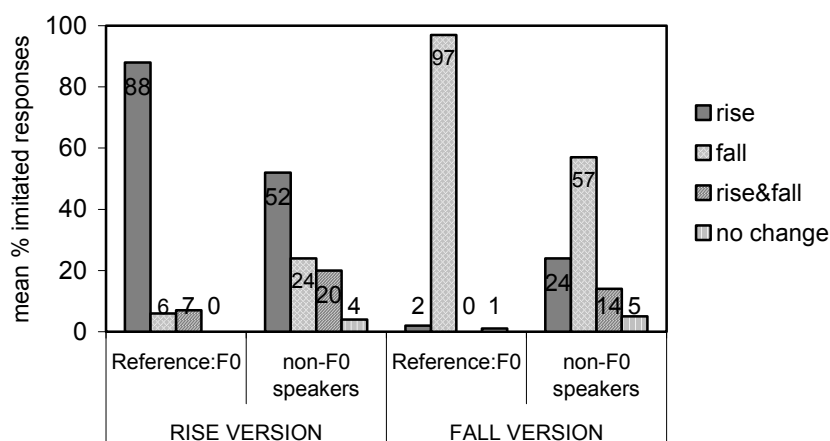


Figure 4.1. For Reference speaker and non-F0 speakers (x-axis): Percentage of responses (y-axis) produced by imitators per transcription category. Left: speakers' intention: rise; right: speakers' intention: fall. Pooled over imitators (18), stretches of speech (5) and speaker(s): 1 F0 speaker; 5 non-F0 speakers.

Distribution of imitators' responses for non-F0 speakers differed significantly from the distribution of imitators' responses for R, both in the rise version ($\chi^2 = 588.946$; $p < 0.001$; $df = 3$) and in the fall version ($\chi^2 = 2092.718$; $p < 0.001$; $df = 3$), confirming the results seen in Figure 4.1: direction of intended non-F0 pitch was not as consistently imitated as R's intended F0 pitch.

If imitators did not perceive the intended rising or falling pitch, their responses would not be dependent on the intended pitch direction, and there should be no difference between the response distribution of the rise version and the response distribution of the fall version, as we controlled for sentential context. However, the results in Figure 1 do reveal a difference between the rise and fall versions; a χ^2 test confirmed that responses in the non-F0 speakers' rise version were distributed significantly differently from those in the non-F0 speakers' fall version ($\chi^2 = 114$; $p < 0.001$; $df = 3$), as predicted.

The results of the non-F0 speakers were not as convincing as for R. Imitators perceived intended rising and falling pitch to a certain extent, but not always. It might be that certain speakers, for example the whisperers, conveyed pitch direction more accurately. For this reason, we present the results of the individual non-F0 speakers in Figure 4.2. Results for rise (left) and fall (right) versions are again given separately.

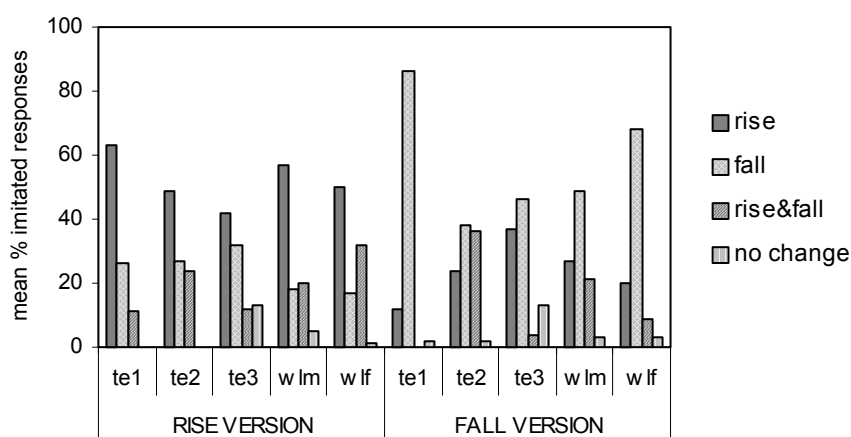


Figure 4.2. For non-F0 speakers (x-axis): Percentage of responses (y-axis) produced by imitators per transcription category. Left panel: speakers' intention: rise; right panel: speakers' intention: fall. Pooled over imitators (18) and stretches of speech (5).

The response distributions varied among different non-F0 speakers, for example TE1 achieved higher percentages than the whisperers and the other TE speakers. It would seem that this speaker conveys pitch direction more accurately than the other speakers. The whisperers achieved higher percentages than TE2 and TE3. However, in *all* the non-F0 speakers, the rise category received the highest percentage of responses when the intended version was a rise, and the fall category received the highest percentage of responses when the intended version was a fall. Conversely, we can also state that a considerable percentage of the imitators' responses did *not* match the intended pitch direction, and this was true for all the non-F0 speakers (except TE1's fall version).

Thus, except for speaker TE1, imitators did not reproduce the intended non-F0 pitch direction as consistently as they reproduced F0 pitch direction in R's speech. One explanation might be that none of the non-F0 speakers consistently produced the intended pitch direction, but this seems unlikely: at least WLF should have consistently produced the intended pitch direction, since she was also the Reference speaker and therefore aware of which pitch movements were required, and when. Alternatively, in the absence of a dominant F0 cue and confronted with unnatural or pathological speech, imitators might have been prejudiced by their internal knowledge of Dutch intonation rules (e.g., a fall is more likely at the end of the sentence), so that some of the test items attracted a 'preferred' pitch event, regardless of the speakers' intent.

This would constitute an intonation bias: the rules of intonation influenced imitators in their choice of the most likely pitch event.

An intonation bias is easily revealed: if we look at a certain stretch of speech (e.g., 'mooiwa'), one response category should contain significantly more responses than any other category, regardless of the speakers' intended pitch direction. If intonation rules did not bias imitators in their response, however, and non-F0 does exist, then we expect that the proportion of rise responses will be the greatest in the rise version and the proportion of fall responses the greatest in the fall version.

For each stretch of speech we compared the proportion of responses among the four categories, separately for the rise and fall versions. Thus, we were forced to work with very small numbers, which made formal testing impossible. The results are presented in Figure 4.3.

As can be seen in Figure 4.3, there was a strong bias in two stretches of speech: for 'enwi' the rise category contained most of the responses and for 'mooiwa' the fall category contained most of the responses, regardless whether the intention was rise or fall. For 'mee' (interrogative) the rise category contained more responses and for 'mee' (declarative) the fall category contained more responses. This can either indicate a sentential bias, or imitators indeed perceived the intended pitch direction. It is unclear to what extent intonation biased the imitators for 'meloe': imitators responded mostly with rise-and-fall in the rise version, but with fall in the fall version, both rise-and-fall and fall can occur at the end of a sentence. For 'oralwi' imitators responded mostly with rise in the rise version and with fall in the fall version, indicating that imitators perceived the intended pitch direction.

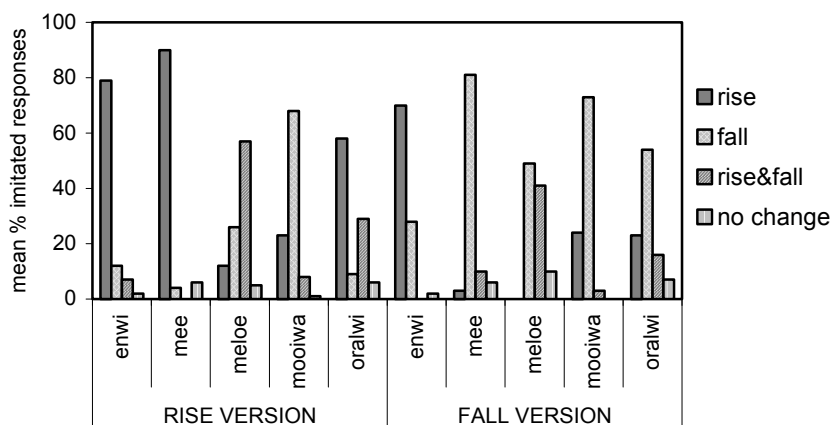


Figure 4.3. For stretches of speech: Percentage of responses (y-axis) produced by imitators per transcription category. Left panel: speakers' intention: rise; right panel: speakers' intention: fall. Pooled over imitators (18) and non-F0 speakers (5).

When we consider 'oralwi', and possibly 'meloe' and 'mee', responses in the Imitation experiment seemed to be, at least partly, steered by non-F0 pitch, although the "correct" responses in 'mee' might have been caused by a sentential bias. However, an intonation bias also influenced the imitators' responses, for example in 'enwi' and 'mooiwa', and possibly in 'meloe'. Although we now know that non-F0 pitch is perceived to some degree, we do not know to what extent imitators' responses were controlled by non-F0 pitch, and to what extent their responses were controlled by the intonation bias. We therefore need to differentiate between the influence of the intonation bias and the influence of non-F0 pitch.

In a control experiment the existence of an intonational bias in these stretches of speech was investigated. This experiment, which is presented in Appendix 4 for the interested reader, indeed confirmed the existence of an intonation bias.

The perception experiment in the next section was designed to eliminate the influence of an intonation bias, so that we could determine the true effect of non-F0 pitch.

4.3 PERCEPTION EXPERIMENT

Although the Imitation experiment revealed that imitators perceived non-F0 pitch direction to some extent, we do not know to what extent the intonation bias overshadowed the intended non-F0 pitch. The goal of this Perception experiment is therefore to investigate the perception of non-F0 pitch direction when the possibility of an intonation bias is eliminated. In this Perception experiment listeners had to identify the pitch direction of the *excised* stretches of speech (speech fragments), thus eliminating intonational expectations, and obliging listeners to rely on the acoustic information present in these *excised* stretches of speech (speech fragments).

We also want to know *which* information listeners use to identify the direction of non-F0 pitch, as stated in the second research question: “which features do listeners use to perceive pitch direction in non-F0 speech?” The speech fragments were therefore filtered into several frequency bands, so that potential “pitch” cues could be isolated. As we already explained in the Introduction, the literature suggests a number of alternative acoustic features that are associated with pitch perception. A brief summary will be given for the convenience of the reader:

Studies on whispered speech have shown that pitch perception was related to *formant frequencies*, especially F_2 (i.e., Thomas, 1969; McGlone & Manning, 1979).

A study by Krull (2001) on whispered word prosody indicated that listeners might interpret the same word, with an increase or decrease in energy above approximately 1.7 kHz as a rise or a fall. Comparing the same (un)-stressed syllables, the perception of stress was related to an increase in high frequency energy (Sluijter, 1995; Grant & Walden, 1996). Whispering and alaryngeal speakers might therefore manipulate *spectral tilt* to convey a rising or falling pitch change.

Some *traces of periodicity* that could not be measured through regular means (pitch detection algorithms or visual inspection of the speech signal) might still be found when speech is filtered (Gauffin & Sundberg, 1989; Grant & Walden, 1996).

Thus, each frequency band included one of the acoustic properties mentioned above.

Both the complete, unfiltered speech fragments and separate band-filtered sounds were presented to the listeners, who were asked to identify the direction of the pitch movement (‘rise’ or ‘fall’).

4.3.1 METHOD

4.3.1.1 Stimulus material

4.3.1.1.1 Excised Stretches of Speech (speech fragments)

The test items were excised from the speakers' utterances at the time points described in 4.2.1.2.

Final consonants were excluded, as the F0 maximum (or minimum) was located in the vowel preceding the final consonant.

As in the previous experiment, the total number of speech fragments was 60 (6 speakers, including R, x 5 speech fragments x 2 versions (rise and fall) of each speech fragment).

4.3.1.1.2 Construction of Band-filtered Sounds

An FFT of each speech fragment was computed from which the power spectrum was obtained. These spectra were decomposed into consecutive frequency bands, using a Hanning-window with 100 Hz smoothing (Boersma & Weenink, 1996). The lowest band was chosen so that it included fundamental frequency (Gauffin & Sundberg, 1989). The second, third and fourth band were roughly chosen so that these bands included F₁, F₂, and F₃ respectively (taken into consideration that formants, especially F₁, are higher in whispering and alaryngeal speakers cf. Ecklund & Traunmüller, 1997; Sisty & Weinberg, 1972). The highest frequency band was chosen so that it did not contain information on formants important for speech.

Therefore, Band 1 = 0.05-0.5 kHz; Band 2 = 0.5-1 kHz; Band 3 = 1-2kHz; Band 4 = 2-4 kHz; Band 5 = 4-8 kHz.

Each band-filtered sound was stored on disk as a separate sound file.

Three hundred and sixty test items were presented in the perception experiment (60 unfiltered speech fragments + (5 x 60) band-filtered sounds).

4.3.1.2 Experimental design

The perception experiment was divided into two sessions split over two days. A session lasted approximately 25 minutes. Per session, a listener would judge three blocks of band-filtered sounds or two blocks of band-filtered sounds plus the block of unfiltered speech fragments. There was also

a short pause between two blocks of stimuli. The band-filtered sounds were blocked by frequency band (thus, all the Band 1 sounds were presented together in a block, etc.). Items within each block were presented randomly. The order in which the different blocks were presented also differed per listener. Each listener judged each test item once.

4.3.1.3 Listeners

Eighteen listeners between the ages of 19 and 30 participated. All reported normal hearing. All were native Dutch speakers. The majority of the listeners had not participated in the Imitation or Control experiment.

4.3.1.4 Procedure

Listeners were seated in a sound-treated booth and listened to each test item (presented over headphones). There were two buttons on the computer screen in front of the listener: one button labeled as “*stijging*” (‘rise’) and one button labeled as “*daling*” (‘fall’). Depending on whether a listener perceived a rise or a fall, the corresponding button on the computer screen was selected by a mouseclick. Thus if a listener perceived a rise in a given test item, he/she would select “*stijging*”, and vice versa. Listeners were instructed to guess when uncertain.

4.3.2 RESULTS OF UNFILTERED SPEECH FRAGMENTS

The Imitation experiment had revealed a possible bias for a number of the speech fragments. In the Perception experiment the speech fragments were therefore presented in isolation. Listeners in the perception experiment were only given a choice between rise and fall.

From the question: “can naïve listeners perceive the pitch direction in non-F0 speech?” three possibilities follow: first, if listeners perceive intended pitch direction *accurately*, we expected mostly rise responses in the rise version and mostly fall responses in the fall version. Second, if listeners perceive the intended pitch direction to some extent, but not accurately, we expect the proportion of rise responses in the rise version or fall responses in the fall version to be above chance. Third, if listeners do not perceive the intended pitch direction, listeners’ responses should be at chance level. Figure 4.4 gives the results per speaker, separately for rise and fall.

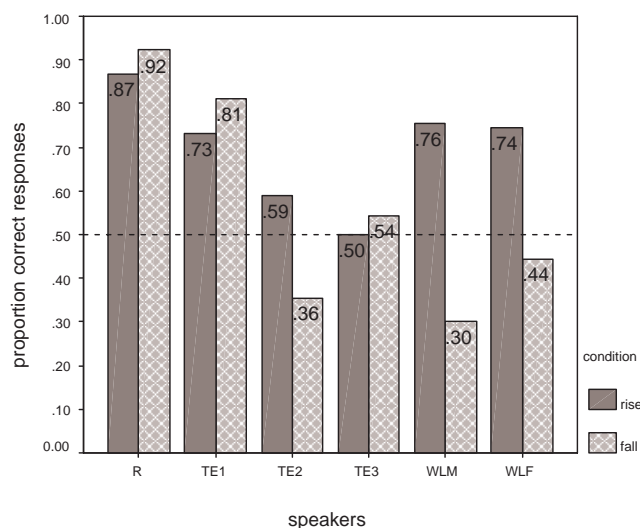


Figure 4.4. Accuracy of listeners' responses. Given per speaker, separated for rise and fall.

Listeners perceived R's intended pitch directions, as well as TE1's intended pitch. For TE3, the rise and fall versions were close to chance, indicating that listeners did not perceive TE3's intended pitch direction. For TE2, WLM and WLF the proportion of correct responses was high in the rise version but poor in the fall version. For these three speakers, two questions arise: does the proportion of correct responses for the rise version differ significantly from chance, and is the seeming bias toward 'rise' responses significant.

Logistic regression was used to answer these questions (Kleinbaum, 1992; Hosmer & Lemeshow, 2000), see also section 3.3.2. The dependent variable (correct responses) was dichotomous: hit (1) if a listener responded with the same pitch direction as the intended pitch direction, and miss (0) otherwise. Listener responses were aggregated into proportions of hits (1). The logistic regression model was used to investigate which factors influence the occurrence of 'correct responses.' The individual speakers and the condition (rise or fall) formed independent categorical variables. As TE3's rise version is at chance level, any of the other speakers' versions that differed from TE3's rise version would differ from chance. Thus, TE3's rise version was the reference category. The means for other speakers were

expressed as differences in logits from the logit average of TE3's 'rise' condition. If an independent variable's coefficient was positive, it yielded more correct responses than the reference category (TE3), if negative, then lower. A significant coefficient means that the independent variable is significantly different from the reference category on 'correct response'.

Logistic regression results of the best fitting model are presented in Table 4.1.

Table 4.1. Results of the logistic regression. The dependent variable was 'correct response' and the reference category was TE3's rise condition. ** indicates that the coefficient is statistically significant at the 0.01 level or beyond. * indicates that the coefficient is significant at the 0.05 level.

variable	coefficient (S.E.)
intercept	0.000 (1)
R	1.872** (0.375)
TE1	1.012** (0.318)
TE2	0.359 (0.301)
WLM	1.128** (0.323)
WLF	1.069** (0.321)
Condition (TE3 rise vs fall)	0.178 (0.299)
Cond*R	0.423 (0.583)
Cond* TE1	0.267 (0.468)
Cond* TE2	-1.132** (0.429)
Cond*WLM	-2.154** (0.450)
Cond*WLF	-1.154** (0.439)
-2 Log likelihood	1235.386
Nagelkerke R ²	0.217

Speakers and condition were the independent variables. The log likelihood and R² in Table 4.1 suggest that this logistic model explains the data reasonably well. All the speakers, except TE2, differed significantly from TE3 in the rise condition. Thus, listeners' perception of rising pitch was above chance for speakers R, TE1, WLM and WLF. For R and TE1 the proportion of correct responses is slightly higher in the fall version than in the rise version (see Figure 4.4). Listeners' perception of falling pitch is therefore also above chance for these two speakers. For the interaction between Speaker and Condition, the reference was the difference between TE3's fall and rise conditions. The results of this interaction show that R and TE1 did not differ significantly from TE3 in terms of the distribution of correct responses over the two conditions. TE2, WLM and WLF had a significantly smaller proportion of correct fall than rise responses, when compared to TE3's distribution, which indicates that for these three

speakers, there was a significant bias toward rise responses. Based on the results in Table 4.1 and Figure 4.4, we conclude that R and TE1 conveyed both rising and falling pitch. We further conclude that WLM and WLF mostly conveyed rising pitch, resulting in a bias toward rise. Although TE2 had a similar pattern to WLM and WLF, we conclude that TE2 and TE3 did not convey the intended pitch direction at all.

This Perception experiment was designed to exclude the intonation bias found in the Imitation experiment. Figures 4.1, 4.2 and 4.4 suggest that the results for R and TE1 are fairly similar in both experiments (in terms of the degree to which responses matched intended pitch direction). For the other speakers it seems that responses in the Imitation experiment were indeed influenced by the imitators' intonational expectations, and, in the absence of an intonational context, listeners identified the intended pitch direction less accurately.

Based on the results in Table 4.1 (illustrated in Figure 4.4), certain speakers seem to be more similar, which might indicate that they have something in common when it comes to conveying pitch direction. We therefore divided the data from stimuli spoken by the participating speakers into subgroups: first the Reference speaker, then TE1, then TE2 and TE3, and lastly WLM and WLF (the whisperers).

4.3.3 RESULTS OF BAND-FILTERED SOUNDS

As explained in section 4.3.1.1, the speech fragments were filtered into adjacent frequency bands to isolate different cues that might contribute to the perception of non-F0 pitch direction.

It was expected that listeners used different acoustic cues for the whisperers than for R. We base our expectations on the results in Table 4.1 (illustrated in Figure 4.4) and on the literature mentioned in the introduction to the Perception Experiment.

Since R used F0, it was expected that the changes in the inter-harmonic difference conveyed the intended pitch direction. The frequency band containing the third to fifth harmonics tends to dominate the pitch sensation (Ritsma, 1967). Thus, for R, the second frequency band is expected to contain the highest proportion correct responses. Depending on the strength of the harmonics, higher frequency bands might contain equal, or less information on 'correct responses'.

We further expect that TE1 either manipulated alternative F0 cues (F_2 or F_3 or spectral tilt) more effectively than TE2 and TE3 or the whisperers, or that some type of periodicity still existed, to be found in the lower frequency bands.

We did not expect TE2 and TE3 to use an alternative cue, and therefore none of the frequency bands is expected to yield a lot of information on ‘correct responses’.

We expect the whisperers to use an alternative cue to F0, which could consist of manipulation of formants or manipulation of spectral tilt. We expect the third or fourth frequency band to yield more information on ‘correct responses’ if listeners relied on F_2 or F_3 , and we expect the fifth frequency band to contain information on ‘correct responses’ if listeners relied on spectral tilt to perceive the intended pitch direction.

Figure 4.5 gives the results separately for the different subgroups.

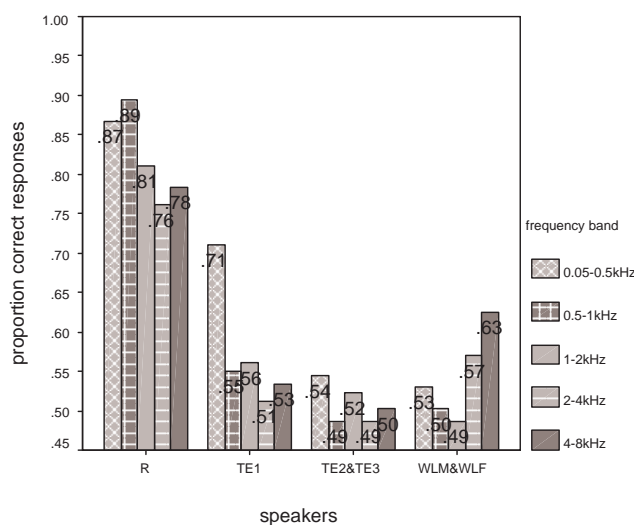


Figure 4.5. Proportion correct responses broken down by subgroup and frequency band. Pooled over rise and fall.

Figure 4.5 shows that although band two contained the highest proportion of correct responses for R, the other frequency bands still contained a higher proportion of correct responses than any of the other subgroups’ frequency bands. For TE1, band one was the most important for the perception of intended pitch direction. For TE2 and TE3, the proportion of correct responses was poor for all the frequency bands. For the whisperers, the proportion of correct responses was poor for all the frequency bands, but was

best for band five. To confirm the results in Figure 4.5, we again used logistic regression, with (the logit of the proportion of) correct responses as dependent variable, and frequency bands as categorical independent variables. Band two was the reference category here. Based on the literature mentioned above, it was expected that band two was the least likely to contain any information aiding perception of *intended* pitch direction in the non-F0 speakers. Also, this band was at chance level for the whisperers. We expected that for whisperers band five might differ significantly from band two, and thus from chance. For TE2 and TE3 we did not expect any frequency bands to differ from band two. For TE1 band one might differ significantly from band two, which was also close to chance level. For R, higher frequency bands might, or might not differ significantly from band two. Because the predictions for the different subgroups differed, logistic regression was done separately for each subgroup. The results of these logistic regressions are given in Table 4.2. Four regression models are presented, one for each subgroup.

Table 4.2. Results for the logistic regression. The dependent variable was ‘correct response’ and the reference category was band 2. Model 1: reference speaker, Model 2: TE1, Model 3: TE2 and TE3, Model 4: whisperers. Regression coefficients (with se in parenthesis) indicates that the coefficient is statistically significant at the 0.01 level. * indicates that the coefficient is significant at the 0.05 level.**

	Reference speaker	TE1	TE2 & TE3	whisperers
Constant (band 2)	2.137** (0.243)	0.044 (0.149)	-0.056 (0.105)	0.011 (0.105)
band 1	-0.256 (0.327)	0.856 (0.222)**	0.234 (0.149)	0.111 (0.149)
band 3	-0.680 (0.308)*	0.201 (0.212)	0.145 (0.149)	-0.067 (0.149)
band 4	-0.978 (0.299)**	0.156 (0.211)	0.000 (0.149)	0.268 (0.150)
band 5	-0.852 (0.303)**	0.089 (0.211)	0.067 (0.149)	0.500** (0.152)
Model χ^2 (df)	16.074 (4)**	19.076 (4)**	3.629 (4)	18.150 (4) **
-2 Log likelihood	823.268	1209.159	2491.209	2463.988

The log likelihood “goodness of fit” gives the scaled deviance that reflects error associated with the model; thus the smaller the number, the better the fit. Table 4.2 shows that the regression model for the Reference speaker has a better overall model fit than the regression models for TE1, TE2 and TE3, or the whisperers. The overall model fit for TE2 and TE3 and for the whisperers is poorer, indicating that there is a larger amount of unexplained variance in their correct responses.

The pattern for R differed from the other subgroups. Band one did not differ significantly from band two, but the size and direction of the coefficients show that the higher frequency bands contained significantly less information on intended pitch direction than band two: apparently the presence of harmonics was more salient in the two lower frequency bands than in the higher frequency bands.

The results for TE1 show that only band one differed significantly from band two which indicates that information on TE1's intended pitch direction was mostly found in band one.

For TE2 and TE3 none of the frequency bands differed significantly from band two.

For the whisperers the proportion correct responses in band five differed significantly from the reference band (band two, which happened to yield accuracy scores at chance level). This contrasts with the results of the other subgroups and indicates that information on intended pitch direction was found in band five for these speakers.

In answer to the second research question: "which acoustic information do listeners use to perceive intended pitch direction?" we conclude that for the Reference speaker, listeners mostly relied on information in the lower frequency bands to identify the intended pitch direction. We therefore expect that F0 (as well as higher harmonics) cued perception of pitch direction. For TE1, listeners relied on information in the lowest frequency band to identify intended pitch direction. We expect that the presence of (quasi)-periodicity might have cued perception of pitch direction. For TE2 and TE3 listeners apparently did not rely on any of the frequency bands to identify intended pitch direction, and it is not expected that these speakers manipulated any acoustic cues consistently. For the whisperers, listeners relied only on information in the highest frequency band to identify intended pitch direction. We expect that changes in spectral tilt might have cued perception of pitch direction.

Acoustic analyses were carried out to confirm these expectations.

4.4 ACOUSTIC ANALYSES

The aim of the Acoustic Analyses was to investigate which acoustic features speakers manipulated to convey pitch direction. The Perception experiment showed that frequency band one (for TE1), and five (for whisperers) contributed to correctly identified pitch direction. For the

Reference speaker all the frequency bands contained information on pitch direction, but mostly bands one and two. As explained in the Method section of the Perception experiment, the choice of the different frequency bands was based on literature. Band one was therefore expected to contain traces of periodicity and band five was expected to reflect the effects of spectral tilt.

Acoustic features were examined using *Praat* (Boersma & Weenink, 1996).

4.4.1 METHOD

4.4.1.1 Material

The speech fragments (test items as defined in 4.2.1.2 and excised for the Perception experiment) were analysed. In total there were 60 items (5 speech fragments x 2 versions x 6 speakers).

4.4.1.2 Analyses

As was explained in section 4.2.1.2, the speech fragments were excised according to the F0-movement observed in the Reference utterances: the F0 excursion started at the beginning of the speech fragment and finished at the end of the speech fragment. If speakers modeled the F0-excursions, the largest contrast would also be between the first part of the speech fragment and the last part of the speech fragment. Measurements were therefore made over the first 50 ms of the speech fragments and the last 50 ms of the speech fragments. The difference was then calculated between the initial and final parts of a speech fragment.

4.4.1.3 F0 excursion

After a try-out with different algorithms the auto-correlation pitch-detection algorithm (Boersma, 1993) was chosen to measure the fundamental frequency. When it came to measuring periodicity in the filtered sound bands, we had the impression that there were fewer pitch detection faults with the AC algorithm than with the SHS pitch detection algorithm used previously.

From previous analyses it was known that F0 was completely absent, or so sporadic in the other speakers' utterances, that F0-excursions could not be

calculated. It was hoped that, if ‘periodicity’ did exist in these speakers, even if sporadic, this might be more apparent and measurable in a low-pass band: if higher frequency perturbations have been filtered out, the presence of some kind of periodicity might be found. Thus, apart from measuring F0 in the complete speech fragments, F0 was measured in the lowest frequency band with the same pitch detection algorithm used for the R’s speech fragments. Any resulting F0 contours were subsequently re-synthesised by means of the PSOLA-analysis by synthesis technique (Moulines & Laroche, 1995) the sole purpose being that faults introduced by the pitch detection algorithm could quickly be traced and corrected manually, using band one’s oscillogram.

No F0 could be measured for TE2 and TE3 or the whisperers, neither in the complete speech fragments nor in band one.

F0 could be measured for R both in the complete speech fragments as well as in bands one to three, but less accurately in bands four and five.

For TE1 previous inspection of narrow-band spectrograms revealed very erratic short-term harmonics. F0 could only be measured sporadically. To illustrate (for the selected all-voiced speech fragments in this chapter): the average percentage voicing for R was 99% (range: 95 – 100%). After manually correcting pitch detection errors, the average percent voicing for speaker TE1 was 34% (range: 0 – 74%). One speech fragment contained sufficient information from which F0 could be calculated. For two speech fragments band one contained sufficient periodicity to calculate F0. In four items F0 could only be calculated after manually correcting the contour, using band one’s oscillogram. Three items contained insufficient periodicity to calculate F0 either in the speech fragment or in band one.

Using the (manually corrected) F0-contours, the mean F0 values were calculated for the 50 ms sections of the initial and final parts of the speech fragment. The difference was then determined in semitones.

4.4.1.4 High frequency intensity

The intensity of the fifth band (4 kHz – 8 kHz) was determined as follows:

The initial and final 50 ms sections (4.4.1.2) were windowed and extracted. Of each extracted section an FFT was computed from which the power spectrum was obtained. Using a Hanning-shaped window, the two

filter bands were constructed and the intensity in dB SPL was calculated over each extracted filtered section.

Another, more common way to measure spectral tilt was also performed:

4.4.1.5 Spectral tilt

The power spectra (see above) of the extracted sections were used to determine the spectral tilt. For each section the spectral tilt was calculated. A low frequency band (0.05 – 4 kHz) was contrasted with band five (4 – 8 kHz). Thus, the band that, at least for the whisperers, seemed to have contained information on the direction of pitch (as revealed in the perception experiment) was contrasted with the rest of the frequency domain. The mean intensity of each sectioned frequency band was measured as described in 4.4.1.4 and the difference between the low and the high band was calculated.

4.4.2 RESULTS

Based on the results of the Perception Experiment and on the literature mentioned in the Method section of the Perception Experiment, we expect the Reference speaker and TE1 to convey pitch direction using periodicity. We further expect WLF and WLM to manipulate spectral tilt. We did not expect TE2 and TE3 to manipulate any cues.

The Mann-Whitney U Test was used to determine if differences between the ‘rise’ and ‘fall’ conditions were significant.

4.4.2.1 Fundamental frequency

Although it was expected that R and TE1 manipulated F0, it was described in the Method section how difficult it was to find periodicity in TE1. Periodicity in this speaker could only be measured in one complete speech fragment. In the other items, if periodicity could be measured, it could only be found in band one. Figure 4.6 illustrates periodicity in TE1 as seen in two different oscillograms. The oscillogram at the top is from a section of a complete, unfiltered vowel and the oscillogram at the bottom is its bandfiltered equivalent (band one). Compared to the top oscillogram (complete unfiltered speech fragment), periodicity is much more noticeable in the bottom oscillogram (band 1).

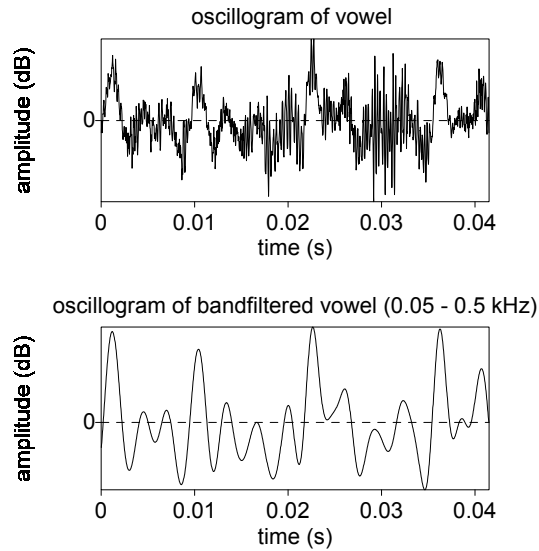


Figure 4.6. Speaker TE1. Periodicity in unfiltered part of a vowel (top) and in the lowest frequency band (bottom).

The periodicity displayed in speaker TE1's speech signal also differed from the periodicity seen in R's speech signal. In Figure 4.7 the quality of the periodicity in three different speakers is shown. The oscillograms were all taken from a bandfiltered (band 1; see above) section of a vowel. The top oscillogram illustrates clear periodicity in the Reference speaker, the middle oscillogram illustrates some periodicity in TE1 and the bottom oscillogram is completely aperiodic (TE2).

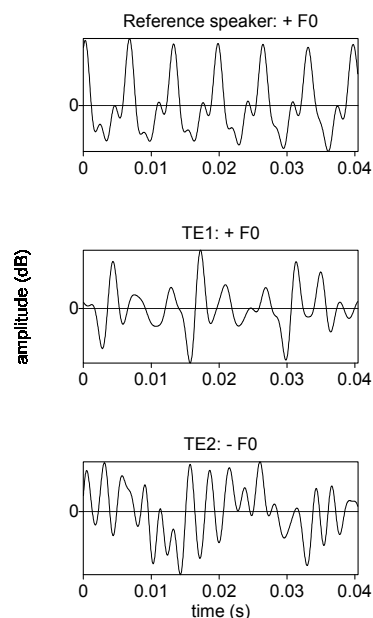


Figure 4.7. Comparison of periodicity in band 1, for different speakers: top = R; middle = TE1; bottom = TE2.

Another difference between R and speaker TE1 concerned the harmonic structure. Figure 4.8 shows the difference in harmonic structure between these two speakers. The spectra were taken from a section of the same vowel (40ms).

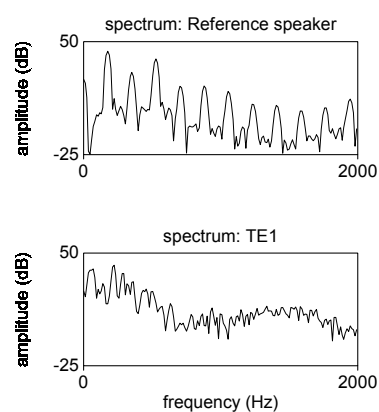


Figure 4.8 Difference in harmonic structure in part of a steady-state vowel. Top: spectrum of R's vowel; bottom: spectrum of TE1's vowel.

Whereas R's spectrum (top) shows a clear harmonic structure at least up to 2 kHz, TE1's spectrum (bottom) shows two peaks on the left, which could probably be interpreted as harmonics. This confirms the results of the Perception Experiment given in Figure 4.5, where listeners perceived the Reference speakers' intended pitch direction also in the higher frequency bands, whereas intended pitch direction was perceived primarily in TE1's lowest frequency band.

The type of periodicity shown in Figures 4.6 and 4.7 indicates that the sound source in TE1 functions very differently with regard the regularity of the vibratory cycle, and the limited harmonic structure in TE1 (Figure 4.8) seems to indicate that the speed of closure is also different. The morphological structure and the tonicity of the sound source apparently influenced TE1's ability to produce consistent periodicity.

Based on the results of the Perception Experiment, it was expected that the difference (in terms of F0) between rise and fall would be significant. There should be an increase when rising pitch is conveyed, and a decrease when falling pitch is conveyed.

Table 4.3 gives, in semitones, the average rising pitch and falling pitch. The Mann-Whitney U Test was used to determine the difference between the conditions.

Table 4.3. F0 difference-values measured in band 1, in semitones, for the two conditions (rise and fall). Averaged over 5 items, given separately per speaker. Standard error of the mean is also given in brackets, as well as the Z value and the P value (Mann-Whitney U Test)

speaker	RISE Mean in semitones (s.e.)	FALL Mean in semitones (s.e.)	Z value	P value
R	+ 5.4 (1.2)	- 3.8 (0.7)	-2.627	0.009
TE1	+ 5.0 (1.9)	- 4.0 (1.2)	-2.121	0.034

The difference between the two conditions, rise and fall, was significant for both speakers. The Reference speaker and TE1 both increased F0 when rising pitch was conveyed and decreased F0 when falling pitch was conveyed. We saw that TE1 displayed limited harmonic structure, and we measured periodicity in the lowest frequency band. Listeners also perceived pitch direction in TE1's lowest frequency band. We therefore conclude that TE1 produced a semblance of periodicity, which speakers perceived as the intended pitch.

The Perception experiment further showed that listeners relied on band five to identify pitch direction in the whisperers. It was therefore expected that the whisperers manipulated spectral tilt to convey pitch direction, and that the perceptual effect of spectral tilt was especially noticeable in band five. R and TE1 manipulated F0, and might not have needed to manipulate spectral tilt to convey pitch direction. We also did not expect TE2 and TE3 to manipulate spectral tilt. The results of the high frequency intensity in band 5, and spectral tilt are given for R, TE1, TE2 and TE3, and the whisperers.

4.4.2.2 High frequency intensity

If speakers manipulated intensity cues, the intensity would increase in the rise version and decrease in the fall version. Thus, the pattern seen in Table 4.3 should also be found for high frequency intensity, but measured in dB. For each group, Table 4.4 gives, in dB, the average rising pitch and falling pitch.

Table 4.4. Intensity in Band 5 (4 – 8 kHz). Intensity difference-values, in dB, for the two conditions (rise and fall). Given separately for the four subgroups. Standard error of the mean is also given in brackets, as well as the Z value and the P value (Mann-Whitney U Test).

group	RISE Mean in dB (s.e.)	FALL Mean in dB (s.e.)	Z value	P value
R	+8.2 (5.2)	+3.8 (3.4)	+0.674	0.690
TE1	+2.0 (3.3)	+1.8 (2.7)	-0.210	0.833
TE2 & TE3	+5.0 (2.0)	0.1 (3.0)	-1.859	0.063
whisperers	+11.7 (1.6)	+1.4 (1.7)	-3.293	0.001

Table 4.4 shows that neither R, nor TE1 manipulated intensity in band 5 to convey pitch direction. Although the difference in averages for TE2 and TE3 seems meaningful, it is much smaller and the variation much greater than in the whisperers. The difference between rise and fall was significant for the whisperers. Although whisperers increased intensity to convey rising pitch, they did not decrease intensity effectively to convey falling pitch.

4.4.2.3 Spectral tilt

If speakers manipulated energy in the high frequencies to convey pitch direction, one would expect the energy distribution in the rise condition to be the opposite of the energy distribution in the fall condition: the spectral tilt in

the rise condition would become flatter, because speakers would enhance the high frequencies to convey a rising pitch change. In contrast, the spectral tilt in the fall condition would become steeper. The spectral tilt was used as an alternative way to measure high frequency intensity because it is a more common method; the difference being that the intensity increase or decrease in band five is measured against the intensity below 4 kHz. It was expected that the results would be similar to the results of the High Frequency Intensity. Table 4.5 gives the results.

Table 4.5. Spectral tilt (0.05 – 4 kHz versus 4 – 8 kHz). Intensity difference-values, in dB, for the two conditions (rise and fall). Given separately for the four subgroups. Standard error of the mean is also given in brackets, as well as the Z value and the p value (Mann-Whitney U Test).

group	RISE Mean in dB (s.e.)	FALL Mean in dB (s.e.)	Z value	P value
R	+5.4 (5.1)	+7.4 (3.9)	-.313	0.754
TE1	+2.4 (3.0)	+2.6 (4.5)	-.525	0.699
TE2 & TE3	+1.1 (2.4)	-0.9 (1.7)	-0.757	0.481
whisperers	+6.8 (1.3)	+0.9 (1.0)	-2.985	0.003

The results in Table 4.5 confirm the results in Table 4.4. As expected, R and TE1 did not manipulate spectral tilt. The average difference between rise and fall is again much smaller, and the variation much larger for TE2 and TE3, when compared to the whisperers. The pattern for the whisperers is the same as in Table 4.4. Spectral tilt became flatter when rising pitch was conveyed, but not steeper when falling pitch was conveyed.

We conclude that whisperers increased intensity in the high frequencies to convey rising pitch, but did not effectively decrease high frequency intensity to convey falling pitch.

The results of the Acoustic analyses confirm the results of the Perception experiment:

Listeners perceived rising pitch, but not falling pitch in the whisperers, and relied on information in band 5 to identify the intended pitch direction, and whisperers indeed manipulated spectral tilt to convey the intended pitch direction.

Listeners did not perceive the pitch direction in TE 2 and TE3, whether in the unfiltered speech fragments or in the different frequency bands, and TE2 and TE3 also did not manipulate any cues consistently enough to convey the intended pitch direction.

Listeners perceived both rising and falling pitch in R and TE1, and relied on information in the lowest frequency band(s) to identify the intended pitch direction, and R and TE1 also manipulated periodicity to convey the intended pitch direction.

4.5 GENERAL DISCUSSION

The aim of this chapter was to search for an alternative, or non-F0 pitch. In normal speech, variations in F0 signal sentence type (question or statement), boundary tones and type of accent. Listeners associate F0 pitch changes with these linguistic functions.

In both perception experiments reported above, the results for the normal phonating reference speaker confirm the unique function of F0. The importance of F0 was also demonstrated in the Imitation experiment. If F0 is absent, imitators switch to a top-down listening strategy; they insert expected pitch events based on their internal knowledge of speech melody, regardless of the speakers' intended pitch direction. This chapter showed that there was no alternative pitch that could *effectively* convey changing pitch direction generally associated with linguistic functions.

Before discussing the alaryngeal speakers, we shall focus on the whisperers, because the assumptions on whispered pitch form the basis to understanding what might have occurred in alaryngeal speech. In the Perception experiment, listeners perceived the whisperers' rising pitch and acoustic analyses showed that whisperers systematically flattened spectral tilt to produce rising pitch. This result is similar to Meyer-Eppler's (1957), who compared whisperers' statements to questions. Although his examples of questions clearly revealed an increase in formants or high frequency energy, there was no clear evidence of a decrease in formants or high frequency energy in his examples of statements.

We can explain the results from both studies in terms of changes in vocal effort. In *voiced* speech, formants have been reported to increase as a result of increasing vocal effort (e.g., Lienard & Di Benedetto, 1999; Eriksson & Traunmüller, 2002; Traunmüller & Eriksson, 2000). Furthermore, in voiced speech spectral tilt and vocal effort are also strongly related (e.g., Glave & Rietveld, 1975; Klatt, 1980; Sluijter, 1995), as are spectral tilt and stress (Sluijter & van Heuven, 1996). Vocal effort and spectral tilt in voiced speech, and stress in *both* voiced and whispered speech, have been related to an increase in glottal tension, in subglottal pressure and in mouth opening

(Vilkman, et al., 1987). Thus, the same physiological process is fundamental to all these effects in voiced, as well as whispered speech. We propose that the perceived increase in pitch is due to an increase in vocal effort.

However, whisperers did not convey falling pitch accurately. Instead, there was a bias towards rising pitch in the Perception experiment and acoustic analyses showed that spectral tilt was slightly *decreased* (flattened) in the fall version. This indicates that vocal effort was slightly *increased* by whisperers when attempting to convey falling pitch. Whisperers did apparently not mimic a falling F0 contour as might be expected: starting with an increased vocal effort and ending with a decrease in vocal effort. Vocal effort was actually manipulated to a much greater degree when a rise was produced than when a fall was produced. Most of the pitch changes in this chapter were prominence cueing. It might be that the non-F0 speakers preferred to mark the stretch of speech that carried an accent-cueing F0, by increasing vocal effort during the course of that stretch of speech. In other words, speakers preferred the production of rises to falls when attempting to convey prominence. This preference might have caused the results seen above: prominence-cueing rises that were clearly marked by an increase in vocal effort and prominence-cueing falls that were not clearly marked by a change in vocal effort.

Alaryngeal speakers differed from the whisperers. Unlike the whisperers, the alaryngeal speakers did not seem to vary spectral tilt consistently to convey pitch direction. However, from the literature it would seem that the underlying mechanism to vary spectral tilt is similar to whispered speech: in alaryngeal speech, an increase in the level of the first formant has been associated with an increased vocal effort (Nord, Hammarberg & Lundstrom, 1995). Further, an increase in vocal effort has also been related to increased tracheal (sub neo-glottic) pressure and increased transsource rate of airflow, the latter indicating increased tension of the neoglottis (Moon & Weinberg, 1987). Thus, if TE speakers increase vocal effort, one might expect listeners to perceive and interpret this increase as prominence. Indeed, the work reported on in chapter two showed that non-F0 alaryngeal speakers did convey prominence, although not as accurately as F0 alaryngeal speakers. The reason why TE2 and TE3 did not vary spectral tilt as consistently as the whisperers in the present study might be as follows: to convey rising and falling pitch, an increase in vocal effort should be produced *gradually over time*. Whisperers achieve this because they still control the tension and resistance of the glottis with great precision, whereas alaryngeal speakers'

control over the neo-glottis is unpredictable and inconsistent (e.g., Moon & Weinberg, 1987). This limited ability might have prevented the alaryngeal speakers from consistently conveying a gradual change in vocal effort, and thus pitch direction.

TE2 and TE3 not only differed from the whisperers, but also from the other alaryngeal speaker, TE1, who produced a semblance of periodicity. Unlike the reference speaker, TE1's periodicity lacked consistency, which complicated F0 measurements. TE1 was classified as a Type IV speaker (Van As, 2001). Van As mentioned that one of her Type IV speakers revealed some voicing (18% of the total voice sample). In this chapter, some of speaker TE1's speech fragments contained even less voicing, but after these speech fragments were low-pass filtered, some F0 could be calculated. In terms of linguistic functions associated with pitch, it seems that alaryngeal speakers with a limited periodicity will have an advantage over alaryngeal speakers without any periodicity.

Training of gradual pitch changes might be worthwhile if alaryngeal speakers have some semblance of periodicity, but would not be effective in non-F0 alaryngeal speakers. However, training non-F0 alaryngeal speakers to convey prominence (and therefore the *perception* of sentence accent) might greatly improve their communicative ability.

In conclusion, this chapter shows that, although inconsistent and unpredictable, a semblance of F0 pitch could still be found in a classified non-F0 alaryngeal speaker. When F0 pitch is completely absent in alaryngeal speakers, listeners do not perceive the intended rising and falling pitch, probably because these alaryngeal speakers are not able to gradually change their vocal effort. Whisperers did gradually increase vocal effort, and listeners perceived this as rising pitch when given the choice between rise, or fall.

Chapters two, three and four have predominantly focussed on the role of pitch in alaryngeal speakers. In the next chapter, the role of timing will be investigated in greater detail.

Prosodic Boundaries

ABSTRACT

Prosodic cues convey the boundaries of (syntactically motivated) phrases. Realization of prosodic cues in alaryngeal speakers can be unpredictable and inconsistent. The present chapter investigated prosodic boundaries in proficient tracheoesophageal and esophageal speakers. Laryngeal speakers (voiced and whispered mode) functioned as controls. Listeners perceived intended phrasing more accurately in laryngeal and tracheoesophageal speakers than in esophageal speakers. However, one esophageal speaker achieved similar results to laryngeal speakers, regardless of number of syllables per phrase. Acoustic analyses showed that different speaker groups used different combinations of cues. One esophageal speaker adapted his speaking style to minimize wrongly positioned within-phrase pausing. The other esophageal speakers differentiated between type of pause (syntactically-motivated versus air-injection), but this was less effective, perceptually.

5.1 INTRODUCTION

In chapter one, the General Introduction, phrasing was also mentioned as a function of prosody. Simply stated, prosody helps a listener interpret the speaker's message, because the prosodic structure groups words into phrases that are strongly related to the syntactic structure (i.e., Wightman, Shattuck-Hufnagel, Ostendorf & Price, 1991; Scott, 1982). In other words, the presence of prosodic information cues the location of prosodic boundaries, and these are not randomly distributed, but often located at syntactically motivated boundaries, for example, at the end of a noun phrase (De Rooij, 1979; Cutler, Dahan & Van Donselaar, 1997). This implies that speakers deliberately manipulate prosodic cues to signal where boundaries are located.

The present chapter investigates to what extent prosodic boundaries are conveyed by alaryngeal speakers.

In comparison to normal laryngeal speech, both TE and Es speech are noisy and less intelligible (Christensen & Dwyer, 1990; Miralles & Cervera, 1995). Both intelligibility and voice quality determine speech quality and because prosody becomes more important when speech quality is less than normal (Nooteboom, 1985), alaryngeal speakers might be more dependent on prosody to convey a message accurately.

Conversely, because the alaryngeal voicing source is a grossly controlled structure when compared to the fine-tuning capabilities of the larynx, especially prosodic cues might be conveyed less accurately.

To understand if, and how alaryngeal speakers might be restricted when attempting to produce prosodic boundaries, we need to know which prosodic elements cue prosodic boundaries. These were described in chapter one, but will be explained here in slightly more detail. The presence of preboundary lengthening alone is a sufficient cue for the perception of a prosodic boundary (De Rooij, 1979; Lehiste, 1983). Nevertheless, minor syntactic boundaries such as a verb, or noun phrase are also often accompanied by a boundary-marking pitch movement, whereas major syntactic boundaries such as sentences and clauses, are also often accompanied by an increase in pre-boundary lengthening, a greater boundary-marking pitch movement and a pause (e.g., Blaauw, 1994; Terken & Collier, 1992; Klatt, 1975). This increase of prosodic information at higher syntactic boundaries might be explained by the finding that listeners perceive boundaries as stronger when they contain more cues (De Pijper & Sanderman, 1994). Yet, different

speakers use prosodic cues differently: sometimes they occur together, sometimes separately, and sometimes not at all. Likewise, a review of various perception studies seems to indicate that listeners often fail to exploit available prosodic information (Cutler, Dahan & Van Donselaar, 1997).

The above might give the impression that, although prosodic boundaries are helpful, they are not essential for the process of speech comprehension. It turns out, however, that when the prosodic cues do not fit the expected syntactic structure, then processing is impeded. For instance, poorly phrased utterances in which boundary-marking pitch changes and pauses did not match the expected syntactic structure, slowed down listeners' processing time (Sanderman & Collier, 1997). Similarly, syntactic parsing was adversely affected when boundary-marking pitch changes and lengthening conflicted with the syntactic expectation (Speer, Kjelgaard and Dobroth, 1996). Scott (1982) found that listener judgments shifted away from the original interpretation of a potentially ambiguous sentence when she inserted a pause, lengthening, or a pause plus lengthening at an alternative boundary. This effect was greater when pauses were accompanied by lengthening. Pauses located at prosodically motivated positions in a sentence improve speech recognition, whereas pauses in other positions have a negative effect on speech recognition (Nooteboom, Scharpff and Van Heuven 1990).

In summary, it would seem that adequately realized prosodic boundaries help speech processing, but more importantly, inadequately realized prosodic boundaries hamper speech processing, either by slowing down, or confusing the listener.

This may happen in alaryngeal speech, because prosodic boundaries in tracheoesophageal and esophageal speech might not be realized adequately for a number of reasons. For example, voice modulation (modulation of the fundamental frequency) in both types of alaryngeal speech is generally erratic and voice range more restricted, when compared to normal laryngeal speech (e.g., Moon & Weinberg, 1987; Robbins, Fisher, Blom, Singer 1984; Gandour, Weinberg, Petty & Dardarananda, 1988; Qi & Weinberg, 1995). This may affect adequate realization of boundary marking pitch movements in both TE and Es speakers.

Although the voicing source is the same in TE and Es speech, the driving-force differs, so that we may also expect differences between TE and Es speech.

In contrast to normal laryngeal and TE speakers who can have an air supply of approximately 3 liters, the air supply available to ES speakers is

limited to small volumes of approximately 80ml (Van den Berg & Moolenaar-Bijl, 1959; Casper & Colton, 1993). Es speakers may produce seven syllables per air charge (Snidecor & Curry, 1959; Moolenaar-Bijl, 1951; Max, Steurs & De Bruijn, 1996, Gandour Weinberg, Petty & Dardarananda, 1986), although an experiment by Gandour et al. (1986) on *phrasing* revealed that a proficient esophageal speaker tended to produce three syllables per phrase, compared to 19 syllables per phrase by normal laryngeal speakers. Since the small supply of air in Es speakers limits the ability to produce longer phrases, Es speakers are forced to pause more often, and pauses are often located at non-phrasal boundaries. This might have an adverse affect on the realization of prosodic boundaries, unless Es speakers successfully differentiate between prosodic pauses and air injection pauses. The results on pausing in chapter two did unfortunately not clarify if Es speakers *effectively* differentiate between prosodic pauses and air injection pauses.

Es speakers might also be limited in the amount of pre-boundary lengthening, as this further depletes the limited air supply. Gandour et al. (1986), found pre-boundary lengthening in their Es speaker, but a phrase in their study was defined as “the portion of a waveform between two pauses”, and did not necessarily coincide with syntactic phrases. Unfortunately, we could not find a study on alaryngeal speech in which pre-boundary lengthening associated with syntactically motivated phrasing was investigated, but the possible absence of this cue in esophageal speakers might also affect the realization of prosodic boundaries.

To summarize, certain prosodic cues might be deficient, either lacking or being inconsistent, both in tracheoesophageal and esophageal speakers. Based on the literature reviewed above, we expect that perception of prosodic boundaries will be affected most in alaryngeal speakers who are unable to convey pre-boundary lengthening and / or well-positioned pauses. In other words, we expect differences between TE speakers and Es speakers. Where TE speakers are expected to convey at least pre-boundary lengthening and might add pausing to compensate for poor speech quality, Es speakers might not be able to convey either pre-boundary lengthening or proper boundary-marking pauses.

The effect of inadequate prosodic boundaries is most clearly illustrated in utterances that are ambiguous by virtue of having more than one underlying syntactic structure, as both speakers and listeners are forced to rely on the prosodic boundaries to disambiguate the utterance. This is particularly true

for bracketing ambiguities, for example, “*(John or Mary) and Sam*” versus “*John or (Mary and Sam)*” (Lehiste, 1983; Streeter, 1978; Scott, 1982). This class of sentences was therefore used in the present study as stimulus material.

Thus, the present chapter specifically investigates listeners’ perception of prosodic boundaries in alaryngeal speech, and alaryngeal speakers’ production of prosodic boundaries, using syntactically ambiguous sentences.

The research questions are as follows:

1. Can listeners perceive prosodic boundaries in TE and Es speech?
2. Which prosodic cues do TE and Es speakers (consistently) manipulate to convey prosodic boundaries?

5.2 GENERAL METHOD

5.2.1 Stimulus material

There were 9 sentences. As mentioned above, sentences were chosen because of their potential syntactic ambiguity, and were therefore suitable to study prosodic effects (Beach, 1991). However, not all ambiguous sentences can be disambiguated (cf. Price, Ostendorf, Shattuck-Hufnagel & Fong, 1991), therefore sentences in the present study are of a class that speakers and listeners are known to disambiguate easily and precisely (Lehiste, 1983; Streeter, 1978; Scott, 1982).

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 “*of*” had two possible groupings per sentence, and sentences containing “*en*” had three possible groupings per sentence.

The different bracketings had equally probable alternative meanings, causing truly “practical” ambiguity (Streeter, 1978). For example:

- 1a. Ik zou (N1 en N2), of N3 uitnodigen;
I would invite (N1 and N2), or N3. *or*
- 1b. Ik zou N1, en (N2 of N3) uitnodigen;
I would invite N1, and (N2 or N3).
- 2a. 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*
- 2b. 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*
 2c. 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 section 5.4).

Fifteen different proper nouns were included. These names varied in complexity (including or excluding plosives) and length (monosyllabic or polysyllabic), for the reason explained below:

In the Netherlands, esophageal speakers are generally taught to use the injection technique, which combines the glosso-pharyngeal press method, with the production of plosives (Moolenaar-Bijl, 1951; Moolenaar-Bijl, 1953; Van den Berg & Moolenaar-Bijl, 1959). With this method, esophageal speakers rely on plosives to reinflate the esophagus with air. The advantage of this method is its unobtrusiveness: insufflation of the esophagus coincides with the articulation of plosives. The assumption is that this results in fluent phrasing, since plosives can be found at regular intervals in speech. However, in utterances with no plosives, esophageal speakers’ fluency is affected since the injections become separate entities (phonetic events) instead of being integrated in a speech sound (Moolenaar-Bijl, 1951; Moolenaar-Bijl, 1953; Van den Berg & Moolenaar-Bijl, 1959).

The training strategy 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*”).

The stimulus sentences in the present study reflected this increasing degree of complexity:

1. Sentences with monosyllabic names, containing plosives (“*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 x 2 bracketings) + (1 “en” version x 3 bracketings)) x 3 levels of complexity. This resulted in seven sentences per level of complexity. The sentences are presented in Appendix 5.

5.2.2 Speakers

Nine speakers participated. Table 5.1 gives relevant information per speaker.

Table 5.1. Relevant information of speakers participating in this study. Speaker group abbreviations: L = laryngeal; TE = tracheoesophageal; Es = esophageal. Average age at recording; time since operation in years; months; type of surgery: total laryngectomy = TL; radiation: primary or post op; average number of syllables per injection; n.a. = not applicable

group	speaker	age	time since operation	type of surgery	radiation:	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	TL	primary	n.a.
	2	55	5;1	TL+ unilateral neck dissection	post-op	n.a.
Es	3	54	4;6	TL	primary	n.a.
	1	67	9	TL + unilateral neck dissection	primary	6
	2	59	6;4	TL + bilateral neck dissection	primary	7
	3	56	5;11	TL + unilateral neck dissection	primary	9

Three laryngeal speakers produced the stimulus sentences in voiced as well as whispered mode. In this way, we included a condition in which speakers could consistently manipulate duration and voicing, and a condition in which speakers could consistently manipulate duration. The laryngeal speakers functioned as controls.

Three tracheoesophageal speakers and three esophageal speakers participated. All speakers were male. Alaryngeal speakers were proficient speakers, as judged by the author (based on the criteria developed by Bors,

Wicherlink, Schutte & Mahieu, 1986). Esophageal speakers all used the air injection technique (glosso-pharyngeal press in combination with the production of plosives).

5.2.3 Recording procedure

The audio recordings were made in a quiet environment. A Sennheiser MKH 50 P48 microphone was used with a Beyerdynamic DT 250 DAT-recorder, and the microphone-to-mouth distance was approximately 20 cm. 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 section 5.2.1). Speakers were made aware of the ambiguity of the sentences, since speakers 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. When speakers substituted the order of names or mispronounced names, they were asked to repeat the sentence. Speakers' utterances were saved on computer disk.

5.3 PERCEPTION EXPERIMENT

The first research question was: "Can listeners perceive prosodic boundaries in tracheoesophageal and esophageal speech?" The speakers' utterances were presented in a perception experiment, and listeners were asked to identify the phrasing, by matching the utterance (audio) with one of two differently bracketed versions of the sentence (visual orthography).

Based on the literature mentioned above, we expect listeners to accurately perceive prosodic boundaries in the laryngeal voiced and whispered utterances, because durational cues are intact (Lehiste, 1983). Since the air supply in TE speakers is similar to that in laryngeal speakers, we also expect listeners to perceive prosodic boundaries accurately in the TE utterances. We expect the perception of prosodic boundaries to be the least accurate in Es utterances, especially in utterances that contained polysyllabic names without plosives.

5.3.1 METHOD

5.3.1.1 Stimulus material

The speakers' utterances were used as stimulus material. Each stimulus item consisted of a recorded utterance (*sound*) and two (*text*) versions of the utterance's corresponding sentence (where different orthographic bracketing brought about different sentence versions; see section 5.2.1).

5.3.1.2 Listeners

Twenty-seven native speakers of Dutch between the ages of 18 to 27 participated. None reported hearing deficiencies. None were familiar with alaryngeal speech. Listeners were paid for their participation.

5.3.1.3 Procedure

Listeners were seated in a sound-treated booth. A speaker's *spoken utterance* (e.g., 1a, spoken) was presented over headphones, and two versions of the corresponding *written sentence* (e.g., 1b and 1a) 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) was heard. The chosen version was selected by clicking the appropriate text button. Listeners were instructed to guess when uncertain.

In total, listeners had to judge 360 utterances ((12 speakers x 12 "of" utterances) + (12 speakers x (2 x 9 "en" utterances))). Listeners needed, on average, one hour 40 minutes to complete the experiment.

5.3.1.4 Design

As explained above, we used a binary forced choice classification task in this perception experiment: for each spoken utterance, a choice was given between two written response possibilities. The *of* utterances could be matched with one of two differently bracketed sentences (see example 1). However, the *en* utterances could be matched with three differently bracketed sentences (see example 2). Therefore, each *en* utterance was

presented twice, so that the listener had the opportunity to choose among all three written possibilities. For example, spoken utterance (2a) had written sentence versions (2a) and (2b) as response possibilities in one trial, and written sentence versions (2a) and (2c) as response possibilities in a second trial, etc. Trials were presented in random order. This resulted in two listener judgments for each *en* utterance, and one listener judgment for each *of* utterance. Fortunately, there was no significant difference between the two *en* trials as revealed by a Student's *t*-test on the arcsine transformed percentages ($t(214) = 0.246, p = 0.806$). We therefore only included the first *en* trial in further analyses, so that the number of listener judgments for the *of* and *en* utterances was equal.

5.3.2 RESULTS

The expectation was that listeners would accurately perceive prosodic boundaries in the laryngeal and tracheoesophageal speaker groups, but less accurately in the esophageal speaker group. Table 5.2 gives, per speaker group, the average percentage of utterances that listeners perceived correctly.

Table 5.2. Average percentage of correctly perceived phrasing, broken down per speaker group. Averaged over 3 speakers per group, x 21 utterances per speaker, x 27 listener judgments.

Speaker group	N	accuracy (standard error)
Laryngeal Voiced	1701	99% (2.8)
Laryngeal Whispered	1701	97% (5.7)
Tracheoesophageal	1701	96% (15.4)
Esophageal	1701	80% (29)

In the laryngeal as well as the TE speaker groups, listeners accurately identified how the utterances were phrased, as expected. The average percentage correctly identified utterances for the esophageal group, although lower, indicates that listeners still perceived the phrasing rather often, in the majority of the utterances. However, the variation in the esophageal group was much larger than in the other speaker groups. This might be due to the sentences with polysyllabic names not containing any plosives, which we expected to cause phrasing problems in these speakers. The results are therefore broken down by complexity and given in Figure 5.1 for each speaker group.

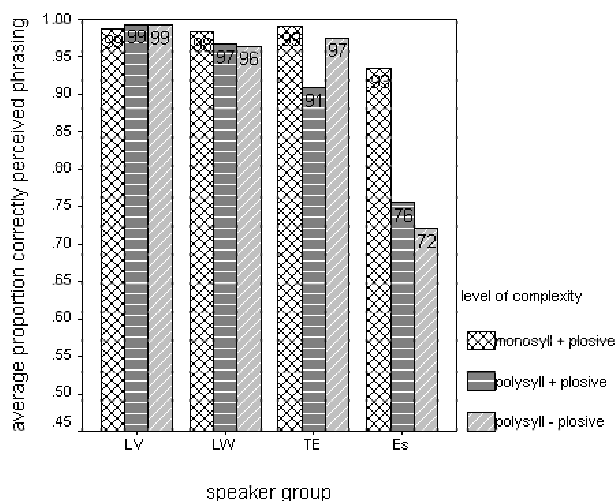


Figure 5.1. Phrasing identification: average proportion of correctly identified utterances given for each speaker group and for each level of complexity, pooled over sentences (7) and listeners (27).

There was little difference between the laryngeal groups (voiced and whispered), indicating that phrasing identification was not adversely affected by lack of voicing. As expected, only the *esophageal* 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. This was contrary to our expectation, which we based on the work by Moolenaar-Bijl, who suggested that the presence of plosives would ensure fluent phrasing (Moolenaar-Bijl, 1951; Moolenaar-Bijl, 1953; Van den Berg & Moolenaar-Bijl, 1959).

Thus, the esophageal group seems to differ from the other speaker groups, and the longer phrases seem to have caused this difference. The (arcsine transformed) percentage of correctly identified phrasing was entered into univariate 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 ($F_1(3,6) = 26.6, p < 0.001$; $F_2(3,8) = 2.99, p = 0.096$), possibly because there were too few sentences or too few speakers per group.

The interaction between ‘level of complexity’ and ‘speaker group’, as illustrated in Figure 5.1, was significant ($F_2(6,16) = 4.4, p = 0.008$).

Furthermore, the main effect of ‘level of complexity’ was significant ($F_1(6,216) = 2.72, p = 0.015$; $F_2(2,16) = 8.41, p = 0.003$). 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, $p < 0.05$), as illustrated in Figure 5.1.

The effect of ‘speaker within speaker group’ was also significant ($F_2(8,216) = 16.4, p < 0.001$). This was somewhat unexpected, because only proficient speakers had been selected. Table 5.2 showed that there was considerable variation within the *esophageal* 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 esophageal speakers.

Because the effect of level of complexity is associated with the esophageal group and we suspect differences among speakers in this group, the levels of complexity are presented for each esophageal speaker in Figure 5.2.

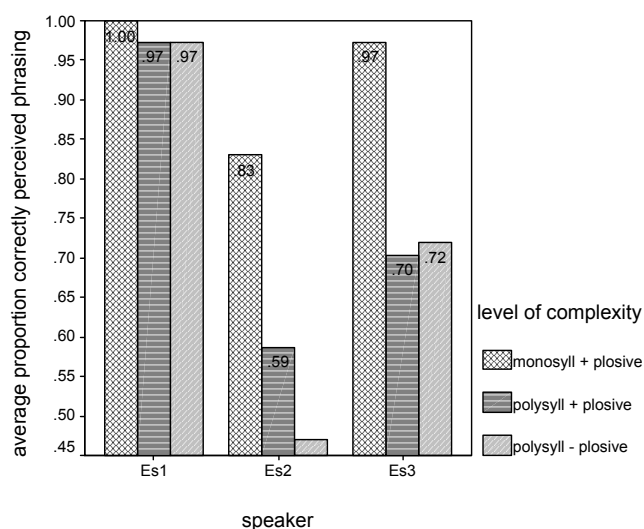


Figure 5.2. Phrasing identification: average proportion of correctly identified utterances, per esophageal speaker, for the different levels of complexity.

Figure 5.2 confirms that there are indeed differences among the esophageal speakers. Es1 accurately conveyed the intended syntactic representation, 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 esophageal speakers more closely mirrored the expectations as explained above: phrases containing monosyllabic names were conveyed quite accurately; but especially Es2 could not disambiguate utterances that contained longer phrases caused by the polysyllabic names.

Table 5.1 showed that Es1 produced the least number of syllables per air injection (6 syllables) when compared to the other speakers (7 and 9 syllables respectively). Thus, it would have been more logical for Es3 to convey phrasing more accurately. Apparently, Es1 followed a different strategy to produce the intended phrasing. This difference in strategy is probably related to the acoustic cues that are used to convey prosodic boundaries: Es1 either used different cues, or used cues more effectively and consistently, when compared to Es2 and Es3.

We tentatively conclude that listeners can disambiguate a syntactically ambiguous utterance accurately – regardless of whether a speaker is laryngeal (either voicing or whispering), tracheoesophageal or esophageal – *provided* that phrasing was unambiguous. In other words, it is not so much the speaker's physiological capability, but also the speaker's strategy that determines if the intended prosodic boundaries in an utterance are conveyed.

In the next section, we investigate which acoustic cues speakers used to convey prosodic boundaries.

5.4 ACOUSTIC ANALYSES

The second research question in the Introduction was: “Which prosodic cues do tracheoesophageal and esophageal speakers manipulate consistently to convey prosodic boundaries?” It is expected that the speakers in the laryngeal group at least used pre-boundary lengthening, and that, when voicing, this group might additionally have used boundary marking pitch tunes. We expect that the TE group also used pre-boundary lengthening, possibly accompanied by boundary marking pitch tunes, and probably by pausing to compensate for poorer speech quality, as explained in the Introduction. We further expect that Es1 followed a different strategy, when compared to the other esophageal speakers.

Prosodic features are used to indicate (syntactic) boundaries. These effects are generally concentrated in the syllable preceding the boundary (Cutler & Butterfield, 1990).

The strongest boundary cues are lengthening of the syllable in the final position of a phrase, and pausing at the boundary (cf., Klatt, 1975; Streeter, 1978; Lehiste, 1983; Wightman, et al., 1992; De Rooij, 1979).

In Dutch the domain of final lengthening was found to be the final (preboundary) syllable, except when the final syllable contained a schwa, in which case the penultimate syllable is also lengthened (Cambier-Langeveld, Nespors & Van Heuven, 1997, Cambier-Langeveld 2000).

Some speakers have been shown to use pitch differentially to disambiguate algebraic expressions similar in structure to the present study's material (Streeter, 1978). Different types of boundary tunes are associated with phrase-final words to signal the end of a phrase (de Rooij, 1979; Swerts, Bouwhuis & Collier, 1994; Blaauw, 1994). 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. F0-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).

Thus, pre-boundary lengthening, pausing and F0-excursions were measured, using *Praat* (Boersma & Weenink, 1996).

5.4.1 METHOD

5.4.1.1 Stimulus material

Per speaker, 15 names occurred in phrase-initial and phrase-final position: as result of the bracketing, names were positioned phrase-initially in one version and phrase-finally in another version of an utterance. Measurements were made on these phrase-initial and phrase-final names, and compared. In total, 360 names were analysed (15 names x 3 speakers x 4 groups x 2 positions).

5.4.1.2 Analyses

5.4.1.2.1 Final Lengthening

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). If the final syllable of a test name contained a schwa, the syllable preceding the final syllable was included in the measurement (Cambier-Langeveld, et al., 1997).

5.4.1.2.2 Pauses

Two types of pauses were differentiated: expected and unexpected pauses. For example, in the utterance: “*Ik zou N1 en (N2 en N3), en (N4 en N5) uitnodigen*”, a pause might be *expected* to follow the phrase-final test names *N3* or *N5*. Similarly, a pause might be *expected* to precede the phrase-initial test names *N2* or *N4*. In contrast, pauses preceding *N3* and *N5* or following *N2* and *N4*, were deemed *unexpected*, 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*’). Esophageal speakers' pauses included any injections that were present. Es speakers might actually use the syntactically motivated pause to inject air, and might differentiate between linguistically motivated pauses and air injection pauses by controlling the duration of the silent interval. In other words, the difference between a syntactic pause and an injection pause is not expected to be the presence or absence of an injection, but the duration of the silent interval.

5.4.1.2.3 Fundamental Frequency

F0 was determined using subharmonic summation (Hermes, 1988). Subsequent F0-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. The distance (in Hertz) between the F0- maximum and F0-minimum in the final syllable was expressed in semitones.

5.4.1.3 Comparison between pre-boundary syllable and its phrase-initial counterpart

Listeners need an actual difference in prosodic realization if they are to distinguish ambiguities via prosody: prosody in the phrase-final position needs to contrast with prosody in the phrase-initial position (Cutler, Dahan & Van Donselaar, 1997).

For pre-boundary lengthening, the duration of the phrase-final syllable should be longer than the duration of the same syllable in phrase-initial position.

Regarding F0-excursions, falling pre-boundary tunes occurred occasionally in the present study, although rising pre-boundary tunes dominated. In contrast, the corresponding phrase-initial syllable tended to be level or carry declination. For F0-excursions, the size of the pre-boundary F0-excursion should therefore be larger than the size of the F0-excursion on the corresponding phrase-initial syllable.

For pausing, there should only be expected pauses, or the expected pauses should be considerably longer in duration than the unexpected (within-phrase) pauses.

As mentioned before, alaryngeal speakers' ability to realize 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 tell you if a speaker uses a cue consistently. We therefore chose to investigate the consistency with which speaker groups produced acoustic cues. For the interested reader, the average values are also given per group in appendix 6.

To determine consistency, we proceeded in the same way as in chapter two (section 2.4.1.6). This will be explained here as well, for the convenience of the reader:

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 perceive¹). The

¹ Although it is conceivable that cues, *individually* below the criteria stated above, might still have a *combined* perceptual effect.

criteria according to which a difference was deemed perceptually meaningful were as follows:

For final lengthening to count as a difference, a JND of 10% was adhered to (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 F0-excursion in the phrase-final syllable had to be at least 1.5 semitones larger than the F0-excursion in the phrase-initial syllable (Rietveld & Gussenhoven, 1985).

No restriction was placed on the duration of pauses, since injection pausing might be considerably shorter than the duration normally associated with linguistically motivated pausing.

Third, based on the criteria above, we totaled the number of times that a cue was perceptually meaningful.

5.4.2 RESULTS

The second research question was: “Which prosodic cues do tracheoesophageal and esophageal speakers consistently manipulate to convey prosodic boundaries?” Based on the results of the Perception experiment, we expect that the laryngeal (whether in voiced or whispered mode) and the tracheoesophageal groups have a similar strategy for conveying prosodic boundaries. We expect Es1 to have a different strategy from Es2 and Es3. For this reason, the results are given separately for Es1, and Es2 and Es3. The Binomial Test (alpha is 0.05) was used to determine if the number of times that a cue was used was indeed significant. If the Binomial Test was significant, the occurrence of a cue was taken to be consistent. Results for the LV, LW and TE groups are presented in Figure 5.3 (top), and results for the Es speakers are presented separately in Figure 5.3 (bottom).

The laryngeal group always lengthened the pre-boundary syllable when voicing (LV), and nearly always when whispering (LW; $p < 0.001$ for both modes). Because the average percentage correctly identified phrases in the Perception experiment was 97% for the LW mode, it seems that pre-boundary lengthening on its own was sufficient to cue the presence of prosodic boundaries. The laryngeal group, when voicing, consistently produced boundary marking pitch tunes ($p = 0.007$), although this cue was used less often than final lengthening.

The tracheoesophageal group also consistently realized pre-boundary lengthening ($p < 0.001$), as well as post-phrase pausing (after the phrase-final name, $p < 0.001$). TE speakers might have used pausing to compensate for poorer speech quality, as mentioned in the Introduction.

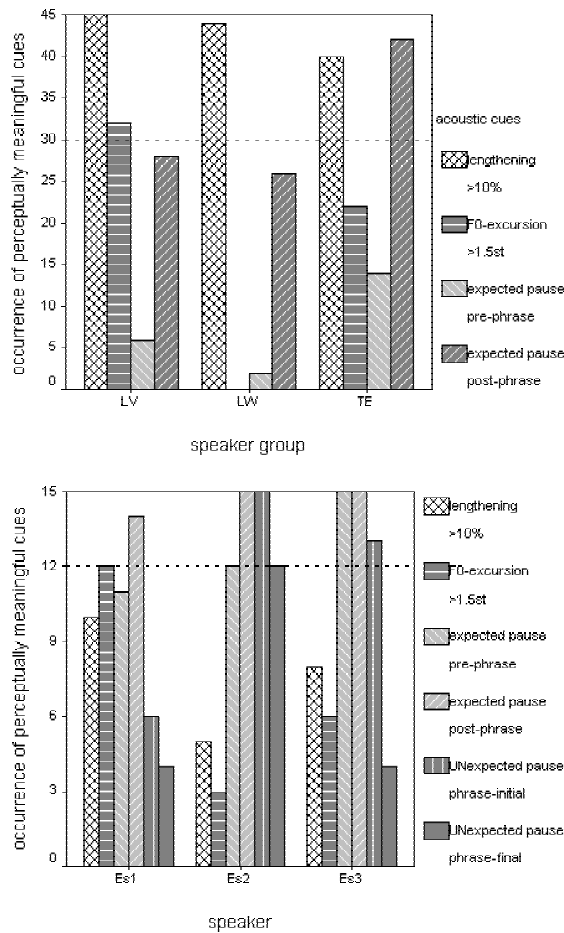


Figure 5.3. Number of times that each speaker group produced an acoustic cue (TOP: LV = laryngeal voiced; LW = laryngeal whispered; TE = tracheoesophageal; 45 = 100%: pooled over level of complexity (3) and speakers (3)), and each Esophageal speaker (BOTTOM: 15 = 100%: pooled over level of complexity); horizontal line and above = significant according to Binomial Test

Es speakers used different cues than the laryngeal speakers. None of the esophageal speakers produced final lengthening consistently, but all used pausing, similar to the TE group. However, as expected, there were also

differences among the esophageal speakers. Es1 consistently used pre-boundary tunes and post-phrase pausing ($p = 0.035$ and $p < 0.001$, respectively). Es2 and Es3 used only pausing consistently, but both speakers also consistently paused *within* phrases (two right-most bars in bottom part, Figure 5.3: number of unexpected phrase-initial and phrase-final pauses). The consistent occurrence of unexpected within-phrase pausing most probably explains the poorer performance of these speakers in the Perception experiment. Es2 even seems to consistently pause *twice* within some phrases. Judging by this speaker's results in the Perception experiment, this seems to have severely impeded the listeners' ability to identify the intended phrasing. The difference in strategy between Es1 and the other Es speakers is at least partly related to pausing. In Table 5.1, we saw that Es1 produced the lowest number of syllables per air injection (six syllables, compared to seven and nine, respectively for Es2 and Es3). Despite this average of six syllables per air injection, speaker Es1 apparently adapted his speaking strategy, thus meeting the requirement of pauseless phrasing more effectively.

Apart from the differences among the esophageal speakers, the results of the Perception experiment also indicated that there were differences among the levels of complexity, especially for speakers Es2 and Es3. These speakers conveyed phrasing less accurately when the phrases contained polysyllabic names. The results of the Es group, for the various complexity levels are therefore given in Figure 5.4.

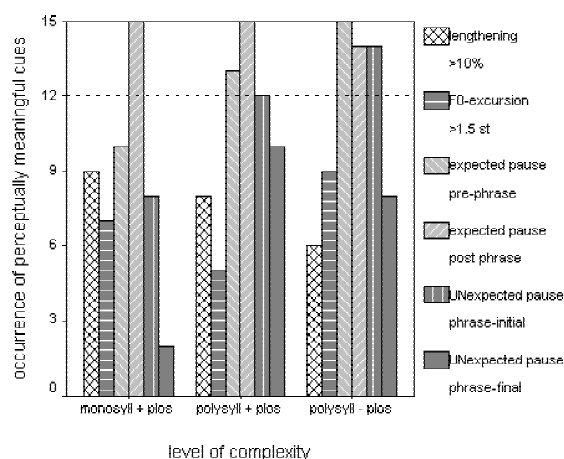


Figure 5.4. Number of times that the Es group produced an acoustic cue, 15 = 100%; given per complexity level; pooled over speakers; horizontal line and above = significant according to Binomial Test.

The results for the different levels of complexity, when pooled over the Es speakers, shows that especially the consistent presence of *unexpected* within-phrase pausing distinguished phrases containing polysyllabic names from phrases containing monosyllabic names. The number of within-phrase pauses is much lower in (short) phrases containing monosyllabic names, and final lengthening occurred more often. However, the fact that even these short phrases contained within-phrase pausing is disturbing, since these Es speakers should be able to produce three syllables with ease, on one injection. The reason why Es speakers (presumably Es2 and Es3) should choose to pause so frequently to inject air is unclear, but we suspect that this might have affected listeners' perception of phrasing negatively. Nevertheless, when we consider the results of the Perception experiment, it has to be noted that listeners still perceived some of the prosodic boundaries correctly in these speakers, apparently despite the presence of within-phrase pauses. This might indicate that speakers differentiate between linguistically motivated pauses and air injection pauses.

We would then expect the linguistically motivated pauses to be significantly longer than the air injection pauses. This is indeed the case (Wilcoxon, $z = -1.989$, $p = 0.047$): syntactically motivated pauses were 443 milliseconds on average (standard error = 22) compared to 286 milliseconds on average (standard error = 17) for air injection pauses. This difference might have aided listeners to some extent in identifying the intended phrasing. Yet, although Es speakers differentiated between syntactically motivated pauses and injection pauses, this was a less effective compensatory strategy than Es1's adaptation in speaking strategy: we conclude that this esophageal speaker conveyed syntactically motivated phrases as effectively as laryngeal speakers, because unexpected within-phrase pauses were kept to a minimum.

5.5 GENERAL DISCUSSION

Speech communication is most effective when the speaker and the listener are attuned to each other. This implies that the speaker provides the information that the listener minimally needs, and that the listener optimally uses the information that the speaker provides (Lindblom, 1990).

The present chapter confirmed that, when bracketing ambiguities had to be conveyed, *all* the speakers provided the listeners with some information

on the intended prosodic boundaries, but not all the speakers provided the information that the listeners minimally needed.

Pre-boundary lengthening on its own, which the whispering laryngeal speakers consistently used as cue, turned out to be sufficient for the accurate perception of prosodic boundaries. This result confirms the findings of De Rooij (1979) and Lehiste (1983). When voicing, laryngeal speakers additionally used pre-boundary pitch movements, and although this strategy did not further improve accuracy of perception, it conforms to the finding that noun phrase boundaries tend to be accompanied by pre-boundary tunes (e.g., Blaauw, 1994). Thus, laryngeal speakers, when voicing, not only conveyed the presence of a prosodic boundary, but also provided information on the depth of the prosodic boundary (e.g., Sanderman & Collier, 1996; Wightman, et al., 1992). This conforms to the notion of a hierarchical prosodic organization, whereby speakers and listeners rely on the same cues in the speech signal to determine the strength of a boundary: the stronger the boundary, the greater the number and strength of the cues (e.g., Shattuck-Hufnagel & Turk, 1996; De Pijper & Sanderman, 1994; Price, et al., 1991; Klatt, 1975).

Tracheoesophageal speakers consistently used final lengthening as well as pausing, which resulted in perception of prosodic boundaries that was equal to the laryngeal groups. These speakers therefore did provide the information that the listeners minimally needed to perceive prosodic boundaries, but they did not consistently manipulate the pitch cue normally associated with this level of syntactic boundary. Instead, these speakers consistently used pausing, which is more often associated with major syntactic boundaries, such as a sentence or clause boundary. In fact, although boundary tunes frequently occur without pauses, pauses are normally accompanied by boundary tunes (Wightman, et al., 1992; De Pijper & Sanderman, 1994). The TE speakers' strategy therefore seems at variance with the accepted prosodic hierarchy as explained above, but TE speakers might have used pausing as a prosodic cue to compensate for the loss of 'normal' speech quality. This would correspond to Nooteboom's (1985) suggestion that the introduction of grammatical pauses helps to maintain intelligibility.

One esophageal speaker consistently combined boundary tunes and pauses, which resulted in the same level of accuracy of perception as pre-boundary lengthening. Similar to the tracheoesophageal speakers, this strategy, although effective in conveying prosodic boundaries, is at variance

with the cues normally associated with phrase boundaries: here final lengthening seems to have been substituted with pausing. This, also, might have been a compensatory strategy, since the limited air-supply associated with esophageal speech probably prohibits consistent lengthening of the pre-boundary syllable, especially in longer phrases.

The two remaining esophageal speakers relied solely on pausing to signal prosodic boundaries. These speakers differentiated between syntactically motivated pauses and air-injection pauses by increasing the silent interval of the syntactically motivated pauses. Results of the Perception experiment indicate that listeners were not always able to use this information optimally.

Listeners might not have been able to fully exploit these speakers' pausing strategy for a number of reasons. Since speakers and listeners adhere to hierarchical internal patterns that govern the durational build-up of speech, according to which the longer the silent interval, the more the pre-boundary lengthening that is expected (i.e., Nooteboom, 1973; Nakatani & Schaffer, 1978), the absence of pre-boundary lengthening in these speakers already violated this expected temporal patterning of speech. Additionally, pause durations of longer than 250 ms are perceived less accurately, and since both injection and syntactic pauses were longer, listeners might not have been able to perceive the difference. Furthermore, there is some evidence that pause durations at prosodic boundaries might cluster around certain values, longer ones being multiples of shorter ones (Fant & Kruckenberg, 1989), but the syntactic pause duration in the present study was too short to be a multiple of the injection pause duration. These esophageal speakers' manipulation of pause duration turned out to be the least effective strategy to convey prosodic boundaries.

In summary, we found that although all the alaryngeal speakers consistently manipulated one or more prosodic cues to signal the presence of a prosodic boundary, none could consistently manipulate *all* the different cues taken for granted in normal laryngeal speech, even though the participating alaryngeal speakers were classified as proficient. Whereas normal laryngeal speakers have the choice to apply different prosodic cues in an hierarchically ordered fashion, we suspect that the majority of alaryngeal speakers will not be able to use *prosodic cues* as a means to indicate the precise depth of a syntactic boundary (whether a syntactic break is meant to be major or minor).

As explained above, we also found that different esophageal speakers used different strategies. Although all the esophageal speakers could

potentially have produced seven or more syllable-phrases per air injection, two of the esophageal speaker did not manage this consistently. As mentioned in the Introduction, Gandour et al. (1986) found that an esophageal speaker who produced on average three syllables per phrase, was capable of seven or more. This speaker's three syllable-phrases did not necessarily coincide with syntactic phrases, a phenomenon that is sometimes found in laryngeal speech: parallel to using prosody to signal syntactic structure, speakers tend to place boundaries so that the number of words and syllables in constituents are equally balanced (Klatt, 1976; Gee & Grosjean, 1983). Two of the esophageal speakers in the present study might have adopted a similar strategy, thus attempting to retain the rhythmic structure of speech, to the detriment of the syntactic structure. However, final lengthening should be incorporated in rhythm, as it allows a gradual slowing down without which speech would sound jerky and unnatural (De Rooij, 1979). The absence of lengthening in these two esophageal speakers probably prohibits the rhythmic effect as well (this was also surmised in chapter two). We conclude that these two speakers did not adapt their speaking strategy effectively to compensate for the loss of an adequate air supply, and were therefore unable to vary rhythmic and durational features normally associated with, and essential to speech communication.

Unlike these two speakers, hardly any within-phrase pausing occurred in the first esophageal speaker. This speaker made an adjustment in speech motor programming similar to the esophageal speaker of Gandour, et al. (1986): in response to the limited air supply associated with esophageal speech, speaking rate was apparently increased, probably by producing more syllables that were shorter and more constant in duration. As this particular esophageal speaker in the present chapter was able to adapt his speaking strategy, it seems entirely plausible that other esophageal speakers can be trained to do so, thereby improving the ability to convey the presence of prosodic boundaries.

In conclusion, we can state that although there were differences between tracheoesophageal and esophageal speakers, this study provides evidence that esophageal speakers, despite their lack of air supply, are potentially capable of conveying prosodic boundaries.

6

Final Discussion

ABSTRACT

In this final chapter, the main findings of this research project on prosody in alaryngeal speech are given and conclusions drawn. Some linguistic and clinical implications are also discussed. Furthermore, limitations of the present project and suggestions for further research are given.

6.1 INTRODUCTION

The general question of this dissertation was whether alaryngeal speakers, given the possible limitations of the alaryngeal voice source, convey prosodic structure in the same manner as normal, laryngeal speakers. The focus was especially on the role of F0 and timing. Less proficient speakers with little control over their voicing source, were included to investigate the effect of an unpredictable F0, and esophageal speakers were included to investigate the effect of a limited air supply on timing. In terms of speech communication, this allowed us to determine if speakers, when tested to the limits of their physical abilities, still attempt to address the needs of the listener. Simultaneously, it allowed us to determine if listeners still attempt to find the necessary prosodic information when they are confronted with speech that might be degraded on a segmental, as well as a suprasegmental level.

In the next section, 6.2, conclusions will be drawn, based on the main findings of this project. In section 6.3 the linguistic implications will be discussed, and in section 6.4 the clinical implications. In 6.5 suggestions for further research will be presented.

6.2 MAIN FINDINGS AND CONCLUSIONS

The first general hypothesis in the General Introduction stated that *all the alaryngeal speakers, regardless of proficiency will conform to the same rules in communication as normal speakers, and will therefore strive to convey necessary prosodic contrasts accurately, although the "hierarchy" of acoustic cues that is used might be dissimilar to the hierarchy found in normal speakers.*

The results of chapter two reveal that alaryngeal speakers who manipulated F0, conveyed accent as accurately as normal speakers. Alaryngeal speakers who did *not* have control over F0, did not convey accent as accurately as speakers with F0, but still conveyed accent rather often. The hierarchy of acoustic cues used by non-F0 speakers was expected to be different, but the acoustic analyses showed that there was no compensation in the sense that non-F0 speakers used duration or intensity as the main cue, instead of F0. Rather, non-F0 speakers seemed to rely primarily on a non-F0 pitch-like phenomenon, which they manipulated to signal accent. This non-F0 pitch-like phenomenon was also perceived to

some degree as the intended speech melody: in chapter three, listeners judged parts of the non-F0 speech melodies to be similar to what had originally been intended by the speakers. The results of the transcription experiment in chapter three showed that abrupt pitch movements, such as a full rise or full fall were indeed transcribed quite accurately. Even in the absence of F0, speakers therefore attempted to convey speech melody. In chapter five, both tracheoesophageal and esophageal speakers manipulated acoustic cues to convey prosodic boundaries, although the hierarchy of acoustic cues used by the alaryngeal speakers was different, when compared to the hierarchy one would have expected, and which was also found in the normal voiced utterances (final lengthening and F0 movement). The tracheoesophageal speakers mainly relied on final lengthening and pauses, whereas one esophageal speaker relied on pauses and F0-movements to signal the presence of a boundary. Two esophageal speakers only distinguished between syntactically motivated pauses and pauses necessary for air injections by increasing the duration of the syntactically motivated pauses.

All in all, these findings support the hypothesis stated above. It may be concluded that alaryngeal speakers, regardless of their proficiency, or whether they were tracheoesophageal, or esophageal, *attempted* to provide the listener with the necessary, and intended prosodic structures.

The second hypothesis in the General Introduction was: *Because the phonetic realization of prosodic structure in alaryngeal speakers is dissimilar when compared to the phonetic realization of prosodic structure in normal speakers, listeners will not perceive the intended prosodic structure in alaryngeal speakers as accurately as the intended prosodic structure in normal speakers.*

In chapter two, the results showed that the presence of accent was not perceived as accurately in the non-F0 alaryngeal speakers, as in the F0 speakers. In chapter three, we saw that the expert transcribers perceived the pitch movements in the non-F0 speaker's utterances far less accurately than the pitch movements in the reference utterances, which contained F0. Although prominence-cueing pitch movements were transcribed more accurately, gradual movements were not. Chapter four also revealed that listeners could not accurately perceive the intended direction of pitch in the absence of F0. The TE speakers and whisperers, who did not have any periodicity in their speech signal, could not convey the intended pitch direction accurately, whereas the only TE speaker who did have a semblance

of periodicity, also managed to convey the intended pitch direction to some degree. Even when the available F0 was inconsistent and unpredictable, listeners were able to perceive the intended direction of pitch. Chapter five showed that, as long as the durational cues were consistently and effectively manipulated, listeners were able to identify the intended prosodic boundaries accurately. Laryngeal and tracheoesophageal speakers used final lengthening consistently, whereas the esophageal speakers primarily relied on pauses to convey the presence of prosodic boundaries. One esophageal speaker's pauses were appropriately located at the boundaries of a phrase, and listeners accurately perceived the prosodic boundaries that this speaker intended. However, two of the esophageal speakers paused within the phrases to inject air. Although these injection pauses were shorter in duration than syntactically motivated pauses, it was not an effective strategy. Listeners could not accurately differentiate between these pause types.

These results confirm the hypothesis above. It is concluded that alaryngeal speakers whose phonetic realization of the intended prosodic contrasts differed from normal speakers' realization, were unable to convey the intended prosodic structures *as accurately* as normal speakers. In other words, because the phonetic realization was different, listeners did not perceive the intended prosodic structures as accurately as in normal speech.

In chapter four, it was shown that whisperers manipulate spectral tilt to convey prominence-lending pitch rises. Both in whisperers and in alaryngeal speakers, spectral tilt is closely related to vocal effort (Glave & Rietveld, 1975; Klatt, 1980; Sluijter, 1995; Nord, Hammarberg & Lundstrom, 1995; Moon & Weinberg, 1987). We propose that non-F0 speakers might increase vocal effort – which manifests itself as a flatter spectral tilt – to convey the presence of prominence.

Although spectral tilt may be a reasonable substitute to convey the presence or absence of accent, some semantic or syntactic information might be lost. As we saw in chapter three and four (also explained above) certain pitch movements were not perceived accurately, and pitch direction was generally perceived even less accurately. Thus, we conclude that spectral tilt cannot adequately replace F0 when different types of pitch movements found in the speech melody, such as type of accent (rise versus rise-and-fall), or type of sentence (question versus statement), need to be conveyed. In that sense, the results of this project have confirmed the unique function of F0 in speech communication.

In chapter five, one esophageal speaker was able to convey the correct phrasing, because he minimised the number of within-phrase pauses and only used pauses at the boundaries of phrases to signal syntactic boundaries. There was no difference between this speaker and the other esophageal speakers in terms of the number of syllables that could potentially be conveyed per air injection. It was reasoned that this speaker adapted his speaking strategy, probably increasing his speaking rate by producing more syllables that were shorter and more constant in duration. It is therefore concluded that esophageal speakers have the potential to adjust their speaking strategy so that timing features are still conveyed during speech communication.

6.3 LINGUISTIC IMPLICATIONS

It was mentioned in the General Introduction that a speaker is expected to provide the information that a listener needs, during speech communication (Lindblom 1990). According to Levelt (1989), a speaker achieves this by producing utterances that a listener will be able to understand, not only in terms of content (what), but also in terms of intent (why). To achieve this, the speaker should realize the prosodic structure such that the listener can deduce the linguistic content of the spoken message. Gandour and Weinberg (1985) concluded after their research project that despite the major differences between normal and proficient alaryngeal speakers, “such differences are irrelevant from a linguistic perspective” (1985: 93). With regard to the present project, one might ask, whether the differences between normal and *less* proficient alaryngeal speakers were also “irrelevant from a linguistic perspective”. We saw that the intention of the alaryngeal speakers was the same as the laryngeal speakers’ intention (first hypothesis): they attempted to convey the prosodic structure. Generally, listeners also perceived the *presence* of an accent and the *presence* of a boundary. However, the second hypothesis was also confirmed. In the absence of F0, listeners were unable to distinguish accurately between different types of pitch movement, as illustrated in chapters three and four. This implies that the absence of F0 undermined accurate perception of, for example, accent type and sentence type; that some semantic information might be lost in everyday speech communication. We further saw in chapter five that alaryngeal speakers did not adhere to the prosodic hierarchical organization that is normally found in speech communication, whereby a specific

combination of acoustic cues is related to the strength of the syntactic boundary (e.g., Shattuck-Hufnagel & Turk, 1996; De Pijper & Sanderman, 1994; Price, et al., 1991). This means that alaryngeal speakers do not have the choice to apply different prosodic cues to signal the precise depth of a syntactic boundary. Some syntactic information might also be lost during everyday speech communication.

If we now reconsider the question posed above, it is clear that the difference between normal speakers and less proficient alaryngeal speakers is not entirely irrelevant from a linguistic perspective. In other words, the difference between laryngeal and alaryngeal speakers does indeed affect the phonetic realization of the prosodic structure, and this may well affect whether the intended semantic or syntactic structure is adequately communicated.

A number of times the word “prominence” has been mentioned in association with accent. Accent, which was investigated in chapter two, is however only one type of prominence. Prominence in normal speech consists of two different linguistic constructs: accent and stress (Sluijter, 1995). Whereas accent makes the important information more prominent on a sentential level, stress is a linguistic property of a word that specifies which syllable in the word is strong or prominent (e.g., “*KINGdom*”, or “*baNAna*”). Although both stress and accent are related to prominence, they have separate acoustic and perceptual correlates. The most important correlate of accent is a change in F₀ or pitch, as we have also seen in this study. Stress is however strongly related to spectral tilt or loudness (Sluijter, 1995). Thus, in normal speech, stress is not a weaker degree of accent and accent is not a stronger degree of stress. In the present research project, non-F₀ speakers seemed to manipulate spectral tilt to make the important information in the sentence more prominent. Thus, in these speakers, spectral tilt was the most important cue to accent, instead of F₀ changes. Therefore, in the absence of F₀, accent can be said to become a more intense degree of stress, and prominence can be said to consist of two different constructs that have similar phonetic realizations.

In the General Introduction, some functions of prosody were discussed. It was explained that prosody might fulfill a function related to syntax, or semantics. The prosodic functions that were explored in this project primarily concerned focus and phrasing. Apart from the fact that alaryngeal speakers attempted to convey these functions, the results of this research project also showed that listeners attempted to identify contrasts in degraded

speech, and even pitch movements in the absence of F0. This supports the notion that listeners actively search for relevant prosodic information, and confirms the interdependence of semantic, syntactic and prosodic structure during speech communication (cf., Cutler, et al., 1997).

6.4 CLINICAL IMPLICATIONS

Based on results that proficient alaryngeal speakers achieved (Gandour and Weinberg 1982, 1983, 1985), it was concluded that alaryngeal speakers have the capacity to produce prosody at proficiency levels exceeding those typically sought by health professionals. One could reason that this does not apply to all alaryngeal speakers, but only to a selection of fortunate individuals who have optimal control over an optimally shaped neoglottis. That only selected speakers have excellent speaking capabilities, seems reasonable, because the alaryngeal population is known for its speaker variability: variation in the shape and size of the neoglottis (Van As 2001), consequent variation in speech quality (e.g., Robbins et al. 1984) and intelligibility (e.g., Doyle et al., 1988) and, as we have seen in this project, variation in prosodic proficiency. This last point was further confirmed by a re-analysis, using MLM, of part of the acoustic data of chapter two (Quené & Van den Bergh 2004). Quené and Van den Bergh found that the variability between alaryngeal speakers was greater than between normal speakers, but variability within alaryngeal speakers was smaller than the variability within normal speakers. There was also no clear relation between duration and intensity, as in normal speakers. Thus, the quality of, and control over the alaryngeal voice source seems to dictate prosodic abilities, at least with regard to the two acoustic cues that were investigated by Quené and Van den Bergh. Hence, some speakers might even have difficulty in communicating semantic and syntactic structures accurately, as was explained in 6.3. This might suggest that professionals should in fact not *overestimate* the capabilities of these speakers.

However, McHenry et al. (1982) commented that it might be unjustifiable to dismiss patients from therapy once the patients' speech is deemed reasonably fluent and intelligible, without attempting to refine, for example, prosodic abilities. Gandour and Weinberg (1985), for example, revealed that none of their speakers had received therapy designed to increase their prosodic ability. Twenty years later, not much seems to have changed: prosody generally seems to receive little, if any structured attention in the

speech rehabilitation process. Similar to Gandour and Weinberg, McHenry et al. (1982) stated that many alaryngeal speakers might have the capability to go well beyond the point of merely functional speech. The question arises if the less proficient speakers in the present study could have achieved better results with training, or if the results were determined only by the speakers' physiological limitations.

Presumably, if F0 is absent in a speaker because the morphology of the neoglottis does not allow vibration to be initiated or sustained, no amount of training will change that. However, in non-F0 alaryngeal speakers, vocal effort, and thus spectral tilt, could play an important communicative role. If speakers are taught to effectively increase and decrease vocal effort to convey the presence and absence of prominence, overall speech communication might well improve. This might be taught, for example, by using the stimulus material in chapter two as training material (appendix 1). In those sentences, prominence is cued by the preceding context. Alternatively, shorter phrases might be designed, in which the words that are meant to be prominent, are visually marked. If the speaker's efforts are also recorded, these can be used as feedback, to indicate which attempts were successfully executed.

The clinician must however first determine that F0 is indeed absent. If some periodicity is still present, it might be more beneficial to concentrate on control over this residual F0, because that is one of the most important prosodic cues. Unfortunately, F0 is hard to detect in the presence of perturbation noise, and alaryngeal speech can be very noisy. However, the present research project showed that simple low-pass filtering increases the accuracy with which the presence of F0 can be determined in these speakers (chapter four).

Non-F0 speakers are not the only speakers who might benefit from training. In chapter five, even though all three Es speakers could potentially produce a similar number of syllables per air injection, only one out of the three Es speakers was able to convey the correct phrasing. This means that physiological properties could not have been the only limitation. Personal communication with the Es speakers revealed that one Es speaker, who conveyed phrasing as accurately as normal speakers, had indeed been taught to synchronize his injection pauses with phrase-boundary pauses. The other two Es speakers' results indicate that appropriate phrasing does not automatically restore itself once Es speakers have been taught the rudimentaries of the injection technique. 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 that were used in chapter five might be suitable as training material, because the speaker is forced to realize phrases properly within the context of a larger utterance. Alternatively, algebraic expressions, such as $(A + B) \times C$ are short and simple, and might be a good starting point in therapy, before attempting longer phrases with difficult names. The recommendations presented in this section are tabulated in Appendix 7, for the convenience of the reader.

Generally speaking, the present research project also confirmed that the Es group did not convey prosodic contrasts as well as the TE group, which again illustrates the superiority of the tracheoesophageal speaking method.

Knowledge of prosody, its role in speech communication and the variation one can expect in alaryngeal speakers should influence a clinician's view on rehabilitation. It is not advisable to employ a standard therapy approach without taking a speaker's prosodic and other speech abilities into account. Once the speaker's limitations have been assessed, it should be possible to systematically improve certain capabilities, thus optimising speech communication.

6.5 LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

The statement that speech communication involves both a speaker and a listener, has been repeated a number of times throughout this dissertation. Indeed, the present research project included both speakers and listeners, but not at the same time. Nooteboom (1983: 183) stated: "Phonetic experiments examining situations in which a speaker speaks to a listener and a listener listens to a speaker, with some real communicative purpose, are extremely rare". The present project unfortunately also suffered from the usual phonetic set-up: speakers speaking in the absence of the listeners, and listeners listening in the absence of the speaker. Perhaps speakers would have made more of an effort to convey the necessary contrasts if they had been confronted with a listener that needed the information.

This research project also suffered from another limitation. It was mentioned a number of times that prosody plays a more important role when the quality and intelligibility of speech is affected, such as in alaryngeal speech. The listeners in the present research project were however instructed to concentrate on the prosody and not on the poor speech quality and intelligibility. Furthermore, the segmental information was visually available

to the listeners: the orthographic representations of the speakers' utterances were always presented on the computer screen. We saw that, generally, listeners perceived the intended prosodic contrasts, although the accuracy with which prosodic intentions were perceived, varied for the alaryngeal speakers. Thus, we do not really know to what extent listeners were affected by the speech quality. Would the prosodic intention have become less conspicuous in the listener's perception because the listener's attention was primarily used to decipher the message's segmental information? Would it have been more insightful to investigate, for example, the effect of accent by determining if accented words are indeed more intelligible, as they are meant to be (e.g., Cutler, et al., 1997)?

A production experiment in which both the speaker and the listener participate at the same time would therefore add valuable information with regard to the alaryngeal speaker's speech communicative abilities. For example, an experiment in which interaction between the speaker and listener is elicited, and conveying the message accurately is of the utmost importance. This type of dialogue might be elicited through a Map Task (Anderson, et al., 1991) in which two participants exchange utterances that are induced by the task. The participants look at similar but significantly different maps of the same region, each unseen by the other. One participant or speaker has a map with a route from a starting point to the goal. This participant instructs the other, so that the other can draw the route through landmarks on the map. Because there are small differences between the maps, and there is no eye contact, a number of issues can be investigated (e.g., questions, interruptions, the need for explanations and repetitions, as well as effective turn-taking skills). This type of task would allow one to investigate many different aspects of communication in a more natural setting than the research conducted in the present project. For example, to what extent the absence of final lengthening and inappropriate pausing in esophageal speakers affects the listener's understanding of the message, or even the listener's processing time (Sanderman & Collier, 1997). It would also allow extralinguistic functions to be investigated in alaryngeal speakers, such as the expression of frustration or impatience when the listener is unable to follow the speaker's instruction.

It is further unclear to what extent alaryngeal speakers adapt other aspects of speech, during spontaneous conversation. It is feasible that a speaker's speaking strategy changes after a laryngectomy, especially if the voice source is not optimal. If speech quality and intelligibility has deteriorated

significantly after a laryngectomy, a speaker might for example compensate by decreasing his speaking rate, or increasing the number of pauses, or simplifying syntactic structures. This would be in line with the H&H theory of Lindblom (1990), in which the speaker is expected to adjust his speech in accordance with the listener's need. One would then expect very proficient speakers to maintain the same speaking strategy as before the laryngectomy, whereas speakers with less proficiency would be expected to adapt their speaking strategy (the poorer the speech quality and intelligibility, the slower the speaking rate, the greater the number of pauses, etc.). Recordings of the speaker's conversational abilities before and approximately six months to a year after the operation would therefore be insightful.

Everyday conversation may take place in favourable, as well as in less favourable communicative environments. Another aspect that therefore needs to be investigated is alaryngeal speakers' ability to adapt to background noise. When circumstances are not favourable, for example at a party, in a bus, over the telephone, are (less) proficient speakers still able to convey a message successfully? Which strategies do they use to communicate effectively under those circumstances?

Furthermore, the present research project revealed that some alaryngeal speakers could control the voice source sufficiently, so that clear F0 excursions were measured in their utterances. The question is exactly how the neo-glottis is controlled and adjusted to produce these F0-movements. In a pilot study it was noted that patients raised their heads and moved them backwards or lowered their chins to their chest when producing glides, or short questions and statements in isolation (personal communication, Jongmans, 2004). This shows that speakers relied on bodily adjustments external to the voice source to increase or decrease the tension or resistance in the neoglottis, thus influencing the height of F0. Such drastic and distracting movements have however not been signalled during normal speech communication, indicating that there might be different means through which F0 can be manipulated. If we know how F0 is controlled, it might be possible to improve this control through specific, goal-oriented exercises.

In conclusion, despite the adverse effect that the alaryngeal voice source may have on speech quality and prosodic expression, speakers still attempted to address the needs of the listener, and listeners still attempted to find the relevant information in the speaker's message, which again shows how

finely speaker and listener are attuned to each other during speech communication.

REFERENCES

- American Standards Association (1960). *Acoustical Terminology* SI, 1-1960, American Standards Association, New York.
- Anderson, A.H., Bader, M., Bard, E.G., Doherty, G., Garrod, S., et al. (1991). The HCRC Map Task corpus. *Language and Speech*, 34, 351-366.
- 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.
- Bertino, G., Bellomo, A., Miani, C., Ferrero, F. & Staffieri, A. (1996). Spectrographic differences between tracheoesophageal and esophageal voice. *Folia Phoniatrica et Logopedica*, 48, 255-261.
- Birch, S.L. & Garnsey, S.M. (1995). The effect of focus on memory for words in sentences. *Journal of Memory and Language*, 34, 232-267.
- 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. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. Doctoral dissertation, University of Amsterdam, the Netherlands.
- Boersma, P. & Weenink (1996). PRAAT manual: a system for doing phonetics by computer. Institute of Phonetic Sciences, University of Amsterdam, report 132 (<http://www.praat.org>)
- Bolinger, D.L. (1958). A theory of pitch accent in English. *Word*, 14, 109-149.
- 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.
- Cambier-Langeveld, G.M. (2000). *Temporal marking of accents and boundaries*. Doctoral dissertation, University of Leiden.
- Casper, J.K. & Colton, R.H. (1993). *Clinical manual for laryngectomy and head and neck cancer rehabilitation*. San Diego, CA: Singular.
- Ching, T.Y.C. & Williams, R. (1994). Communication of lexical tone in Cantonese alaryngeal speech. *Journal of Speech and Hearing Research*, 37, 557-564.

- Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.
- Christensen, M.J. & Dwyer, P.E. (1990). Improving alaryngeal speech intelligibility. *Journal of Communication Disorders*, 23, 445-451.
- Cohen, A. (1968). Errors of speech and their implications in understanding the strategy of language users. *Phonetica*, 21, 177-181.
- Cohen, A. & 't Hart, J. (1967). On the anatomy of intonation. *Lingua*, 19, 177-192.
- Coleman, R.O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech and Hearing Research*, 14, 565-577.
- Collier, R. (1970). The optimum position of prominence lending pitch rises. *IPO Annual Progress Report*, 5, 82-85.
- Cutler, A. & Butterfield, S. (1990). Durational cues to word boundaries in clear speech. *Speech Communication*, 9, 485-495.
- Cutler, A., Dahan, D. & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40, 141-201.
- Damsté, P.H. (1958). *Oesophageal speech after laryngectomy*. Doctoral dissertation, Groningen University, the Netherlands.
- Debruyne, F., Delaere, P., Wouters, J. & Uwents, P. (1994). Acoustic analysis of tracheoesophageal speech. *The Journal of Laryngology and Otology*, 108, 325-328.
- De Pijper, J.R. (1983). *Modelling British-English Intonation*. Dordrecht: Foris Publications.
- 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, 2037-2047.
- De Rooij, J.J. (1979). *Speech punctuation. An acoustic and perceptual study of some aspects of speech prosody in Dutch*. Doctoral dissertation. Utrecht: University of Utrecht.
- Diedrich, W.M. (1968). The mechanism of esophageal speech. *Annals of the New York Academy of Sciences*, 155, 303-317.
- Doyle, P.C., Danhauer, J.L. & Reed, C.G. (1988). Listeners' perception of consonants produced by esophageal and tracheoesophageal talkers. *Journal of Speech and Hearing Research*, 53, 400-407.
- Dworkin, J.P., Meleca, R.J., Zormeier, M.M., Simpson, M.L., Garfield, I., Jacobs, J.R. & Mathog, R.H. (1998). Videostroboscopy of the pharyngoesophageal segment in total laryngectomees. *Laryngoscope*, 108, 1773-1781.
- Eady, S. (1982). Differences in the F0 patterns of speech: tone language versus stress language. *Language and Speech*, 25, 29-45.

- Eefting, W.Z.F. (1991). *Timing in talking*. Doctoral dissertation, Utrecht University, the Netherlands.
- Eklund, I. & Traunmüller, H. (1997). Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *Phonetica*, 54, 1-21.
- Eriksson, A. & Traunmüller, H. (2002). Perception of vocal effort and distance from the speaker on the basis of vowel utterances. *Perception & Psychophysics*, 64, 131-139.
- Fant, G. & Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *Speech Transmission Laboratory-Quarterly Progress and Status Report*, 2, 1-83. RIT:Stockholm.
- Gandour, J. & Weinberg, B. (1982). Perception of contrastive stress in alaryngeal speech. *Journal of Phonetics*, 10, 347-359.
- Gandour, J. & Weinberg, B. (1983). Perception of intonational contrasts in alaryngeal speech. *Journal of Speech and Hearing Research*, 26, 142-148.
- Gandour, J. & Weinberg, B. (1985). Production of speech melody and contrastive stress in esophageal and tracheoesophageal speech. *Journal of Phonetics*, 13, 83-85.
- Gandour, J., Weinberg, B. & Garziona, B. (1983). Perception of lexical stress in alaryngeal speech. *Journal of Speech and Hearing Research*, 26, 418-424.
- Gandour, J., Weinberg, B. & Kosowsky, A. (1982). Perception of syntactic stress in alaryngeal speech. *Language and Speech*, 25, 299-304.
- Gandour, J., Weinberg, B., Petty, S.H. & Dardarananda, R. (1986). Rhythm in Thai esophageal speech. *Journal of Speech and Hearing Research*, 29, 563-568.
- Gandour, J., Weinberg, B., Petty, S.H. & Dardarananda, R. (1988). Tone in Thai alaryngeal speech. *Journal of Speech and Hearing Disorders*, 53, 23-29.
- Gauffin, J. & Sundberg, J. (1989). Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech and Hearing Research*, 32, 556-565.
- Gee, J.P. & Grosjean, F. (1983). Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15, 411-458.
- Giet, F. (1956). Kann man in einer Tonsprache flüstern? *Lingua*, 5, 372-381.
- Glave, R.D. & Rietveld, A.C.M. (1975). Is the effort dependence of speech loudness explicable on the basis of acoustical cues? *Journal of the Acoustical Society of America*, 58, 875-879.
- Grant, K.W. & Walden, B.E. (1996). Spectral distribution of prosodic information. *Journal of Speech, Language and Hearing Research*, 39, 228-239.

- Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.
- Gussenhoven, C. & Rietveld, A.C.M. (1992). Intonation contours, prosodic structure and preboundary lengthening. *Journal of Phonetics*, 20, 283-303.
- Gussenhoven, C., Terken, J. & Rietveld, A.C.M. (1999). *Transcription of Dutch intonation-courseware*. <<http://lands.let.kun.nl/todi>>
- Harbold, G. (1958). Pitch ratings of voiced and whispered vowels. *Journal of the Acoustical Society of America*, 30, 600-601.
- Hasegawa, Y. & Hata, K. (1992). Fundamental frequency as an acoustic cue to accent perception. *Language and Speech*, 35, 87-98.
- Hermes, D.J. (1988). Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America*, 83, 257-264.
- Heeren, W.F.L. (2001). *Intonation in whispered Dutch: correlates of production and perception*. Unpublished Master's Thesis, Leiden University, the Netherlands.
- Hermes, D.J. (1997). Timing of pitch movement and accentuation of syllables in Dutch. *Journal of the Acoustical Society of America*, 102, 2390-2402.
- Hermes, D.J. & Rump, H.H. (1994). Perception of prominence in speech intonation induced by rising and falling pitch movements. *Journal of the Acoustical Society of America*, 69, 83-92.
- Higashikawa, M., Nakai, K., Sakakura, A. & Takahashi, H. (1996). Perceived pitch of whispered vowels – relationship with formant frequencies: a preliminary study. *Journal of Voice*, 10(2), 155-158.
- Higashikawa, M. & Minifie, F.D. (1999). Acoustical-perceptual correlates of “whisper pitch” in synthetically generated vowels. *Journal of Speech, Language and Hearing Research*, 42, 583-591.
- Hilgers, F.J.M. & Schouwenburg, P.F. (1990). A new, low-resistance, self-retaining prosthesis for voice rehabilitation after total laryngectomy. *Laryngoscope*, 100, 1202-1207.
- Hilgers, F.J.M., Ackerstaff, A.H., Balm, A.J.M. & Gregor, R.T. (1996). A new heat and moisture exchanger with speech valve (Provox[®] Stomafilter). *Clinical Otolaryngology*, 21, 414-418.
- Hirano, M. & Bless, D.M. (1993). *Videostroboscopic examination of the larynx*. San Diego CA: Singular.
- Horii, Y. & Weinberg, B. (1975). Intelligibility characteristics of superior esophageal speech presented under various levels of masking noise. *Journal of Speech and Hearing Research*, 18, 413-419.
- Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: Wiley.

- Jensen, M.K., (1958). Recognition of word tones in whispered speech. *Word*, 14, 187-196.
- 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.
- Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67, 971-996.
- Kleinbaum, D.G. (1992). *Logistic Regression: A self-learning text: Statistics in the health sciences*. New York: Springer.
- Krull, D. (2001). Perception of Estonian word prosody in whispered speech. In W.A. van Dommelen & T. Fretheim (eds.) *Nordic Prosody: Proceedings of the VIIIth Conference*, Frankfurt, 153-164.
- Ladd, D.R. (1996). *Intonational Phonology*. Cambridge, U.K.: Cambridge University Press.
- Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- 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.
- Levelt, W.J.M. (1989). *Speaking: from intention to articulation*. Cambridge MA: MIT Press.
- Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6, 172-187.
- Lieberman, P. (1977). *Speech physiology and acoustic phonetics: an introduction*. New York: Macmillan.
- Lienard, J. & Di Benedetto, M. (1999). Effect of vocal effort on spectral properties of vowels. *Journal of the Acoustical Society of America*, 106 411-423.
- Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W.J. & Marchal, A. (Eds), *Speech production and speech modelling*. (pp.403-439) Dordrecht: Kluwer Academic publishers.
- Max, L., Steurs, W. & De Bruyn, W. (1996). Vocal capacities in esophageal and tracheoesophageal speakers. *Laryngoscope*, 106, 93-96.
- McGlone, R.E., & Manning, W.H. (1979). Role of second formant in pitch perception of whispered vowels. *Folia Phoniatica*, 31, 9-14.

- McHenry, M., Reich, A. & Minifie, F. (1982). Acoustical characteristics of intended syllabic stress in excellent esophageal speakers. *Journal of Speech and Hearing Research*, 25, 564-573.
- Meyer-Eppler, W. (1957). Realization of prosodic features in whispered speech. *Journal of the Acoustical Society of America*, 29, 104-106.
- Miller, J.D. (1961). Word tone recognition in Vietnamese whispered speech. *Word*, 17, 11-15.
- 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 Phoniatica et Logopaedica*, 3, 20-24.
- Moolenaar-Bijl, A. (1953). The importance of certain consonants in esophageal 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.
- Moore, B.J.C. (1989) *An Introduction to the Psychology of Hearing* (third edition). London, U.K.: Academic Press.
- Moulines, E. & Laroche, J. (1995). Non-parametric techniques for pitch scale and time-scale modification of speech. *Speech Communication*, 16, 175-205.
- Nakatani, L.H. & Schaffer, J.A. (1978). Hearing "words" without words. *Journal of the Acoustical Society of America*, 63, 234-245.
- Nieboer, G.L.J., De Graaf, T. & Schutte, H.K. (1988). Esophageal voice quality judgements by means of the semantic differential. *Journal of Phonetics*, 16, 417-436.
- Nooteboom, S.G. (1973). The perceptual reality of some prosodic durations. *Journal of Phonetics*, 1, 25-45.
- Nooteboom, S.G. (1983). Is speech production controlled by speech perception? In: M.P.R. Van den Broecke, V.J. Van Heuven & W. Zonneveld (eds.) *Sound structures*. Dordrecht: Foris, 183-194.
- Nooteboom, S.G. (1985). A functional view of prosodic timing. In Michon, J.A. & Jackson, J.L. (eds.), *Time, mind and behavior*. New York: Springer-Verlag.
- Nooteboom, S.G., Scharff, P. and 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.

- Nooteboom, S.G., & Terken, J.M.B. (1982). What makes speakers omit pitch accents? An experiment. *Phonetica*, 39, 317-336.
- Nord, L., Hammarberg, B. & Lundstrom, E. (1995). Laryngectomy speech in noise – voice effort, speech rate and intelligibility. *Scandinavian Journal of Logopedics and Phoniatrics*, 20, 107-112.
- Panconcelli-Calzia, G. (1955). Das Flüstern in seiner physio-pathologischen und linguistischen Bedeutung. *Lingua*, 4, 369-378.
- Pindzola, R.H. Cain, B.H. (1989). Duration and frequency characteristics of tracheoesophageal speech. *Annals of Otolaryngology, Rhinology and Laryngology*, 98, 960-964.
- Plomp, R. & Mimpen, A.M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 18, 43-52.
- Prater, R.J. & Swift, R.W. (1984). *Manual of Voice Therapy*. : Boston: Little, Brown & Hamer Co.
- 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, 2956-2970.
- Qi, Y. & Weinberg, B. (1991). Spectral slope of vowels produced by tracheoesophageal speakers. *Journal of Speech and Hearing Research*, 34, 243-247.
- Qi, Y. & Weinberg, B. (1995). Characteristics of voicing source waveforms produced by esophageal and tracheoesophageal speakers. *Journal of Speech and Hearing Research*, 38, 536-548.
- Quené, H. & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication*, 43, 103-121.
- Remez, R.E. & Rubin, P.E. (1993). On the intonation of sinusoidal sentences: contour and pitch height. *Journal of the Acoustical Society of America*, 94, 1983-1988.
- Rietveld, A.C.M. & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics*, 13, 299-308.
- Ritsma, R.J. (1967). Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America*, 42, 191-198.
- Robbins, J. (1984). Acoustic differentiation of laryngeal, esophageal and tracheoesophageal speech. *Journal of Speech and Hearing Research*, 27, 577-585.
- Robbins, J., Fisher, H.B., Blom, E.C. & Singer, M.I. (1984). A comparative acoustic study of normal, esophageal and tracheoesophageal speech production. *Journal of Speech and Hearing Disorders*, 49, 202-210.

- Rump, H.H. & Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech*, 39, 1-17.
- Sanderman, A.A. & Collier, R. (1996). Prosodic rules for the implementation of phrase boundaries in synthetic speech. *Journal of the Acoustical Society of America*, 100, 3390-3397.
- Sanderman, A.A. & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech*, 40, 391-409.
- Sataloff, R.T. (1997). *Professional voice: the science and art of clinical care*. San Diego, CA: Singular.
- 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, 996-1007.
- Shattuck-Hufnagel, S. & Turk, A.E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 25, 193-247.
- Sisty, N.L. & Weinberg, B. (1972). Formant frequency characteristics of esophageal speech. *Journal of Speech and Hearing Research*, 15, 423-438.
- Slavin, C.S. & Ferrand, C.T. (1995). Factor analysis of proficient esophageal speech: toward a multidimensional model. *Journal of Speech and Hearing Research*, 38, 1224-1231.
- Slis, I.H. & Cohen, A. (1969). On the complex regulating the voiced-voiceless distinction. *Language and Speech*, 12, 80-102; 137-155.
- Sloate, P.L. & Voyat, G. (1983). Language imitation in development. *Journal of Psycholinguistic Research*, 12, 199-222.
- Sluijter, A.M.C. (1995). *Phonetic correlates of stress and accent*. Holland Academic Graphics, The Hague.
- Sluijter, A.M.c. & Van Heuven, V.J. (1995). Effects of focus distribution, pitch accent and lexical stress on temporal organization of syllables in Dutch. *Phonetica*, 52, 71-89.
- Sluijter, A.M.C. & Van Heuven, V.J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471-2485.
- Snidecor, J.C. & Curry, E.T. (1959). Temporal and pitch aspects of superior esophageal speech. *Annals of Otology, Rhinology and Laryngology*, 68, 1-14.

- Speer, S.R., Kjelgaard, M.M. & Dobroth, K.M. (1996). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of Psycholinguistic Research*, 25, 249-271.
- Streeter, L.A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America*, 64, 1582-1592.
- Swerts, M., Bouwhuis, D.G. & Collier, R. (1994). Melodic cues to the perceived “finality” of utterances. *Journal of the Acoustical Society of America*, 96, 2064-2075.
- Tardy-Mitzell, S., Andrews, M.L. & Bowman, S. (1985). Acceptability and intelligibility of tracheoesophageal speech. *Archives of Otolaryngology*, 111, 213-215.
- Terken, J.M.B. & Collier, R. (1992). Syntactic influences on prosody. In Tokhura, E. Vatikiotis-Bateson, & Sagisaka, Y. (eds.), *Speech Perception, Production and Linguistic Structure*, Amsterdam: IOS.
- Titze, I.R. (1994). *Principles of Voice Production*, Englewood Cliffs, NJ: Prentice Hall.
- ‘t Hart, J. & Cohen, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics*, 1, 309-327.
- ‘t Hart, J. & Collier, R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics*, 3, 235-255.
- ‘t Hart, J., Collier, R. & Cohen, A. (1990) A perceptual study of speech melody. Cambridge, U.K.: Cambridge University Press.
- Thomas, I.B. (1969). Perceived pitch of whispered vowels. *Journal of the Acoustical Society of America*, 46, 468-470.
- Trautmüller, H. & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, 107, 3438-3451.
- Van As, J.C. (2001) *Tracheoesophageal speech. A multidimensional assessment of voice quality*. Budde-Elinkwijk Grafische producties, Nieuwegein.
- Van As, J.C., Hilgers, F.J.M., Koopmans-van Beinum, F.J. & Ackerstaff, A.H. (1998). The influence of stoma occlusion on aspects of tracheoesophageal voice. *Acta Otolaryngologica* (Stockholm), 118, 732-738.
- Van den Berg, J. (1958). Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research*, 1, 227-244.
- Van den Berg, J. & Moolenaar-Bijl, A.J. (1959). Crico-pharyngeal sphincter, pitch, intensity and fluency in esophageal speech. *Practica Oto-rhino-laryngologica*, 21, 298-315.

- Van Donselaar, W. & Lentz, J. (1994). The function of sentence accents and given/new information in speech processing. *Language and Speech*, 37, 375-391.
- Van Donzel, M. (1999). *Prosodic aspects of information in discourse*. Doctoral dissertation, University of Amsterdam, the Netherlands.
- Van Katwijk, A. (1974). *Accentuation in Dutch; an experimental linguistic study.*, Amsterdam: Van Gorcum.
- 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 Weissenbruch, R. (1996). *Voice restoration after laryngectomy*. Doctoral dissertation, University of Groningen, the Netherlands.
- Van Zanten, E., Damen, L. & van Houten, E. (1991). The ASSP speech database. *SPIN/ASSP-report 41*, Speech Technology Foundation, Utrecht, the Netherlands.
- Vilkman, E., Aaltonen, O., Raimo, I, Ignatius, J. & Komi, P.V. (1987). On stress production in whispered Finnish. *Journal of Phonetics*, 15, 157-168.
- Weinberg, B., Horii, Y. & Smith, B.E. (1980). Long-time spectral and intensity characteristics of esophageal speech. *Journal of the Acoustical Society of America*, 67, 1781-1784.
- Weinberg, B., Horii, Y. Blom, E. & Singer, M. (1982). Airway resistance during esophageal phonation. *Journal of Speech and Hearing Disorders*, 47, 194-199.
- 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, 1707-1717.
- Williams, S.E. & Watson, J.B. (1987). Speaking proficiency variations according to method of alaryngeal voicing. *Laryngoscope*, 97, 737-740.
- Wong, S.H.W., Cheung, C.C.H, Yuen, A.P.W., Ho, W.K. & Wei, W.I. (1997). Assessment of tracheoesophageal speech in a tonal language. *Archives of Otolaryngology, Head and Neck Surgery*, 123, 88-92.
- Yiu, E., Van Hasselt, C.A., Williams, S.T. & Woo, J.K.S. (1994). Speech intelligibility in tone language laryngectomy speakers. *European Journal of Disorders of Communication*, 29, 339-347.

APPENDIX 1

Stimulus material used in chapter two

1. De vlieger vloog niet over de schutting, de **BAL** vloog over de schutting.
2. De bal vloog niet over de muur, de bal vloog over de **SCHUTTING**.
3. De nieuwe fiets is niet geleend, de nieuwe fiets is **GESTOLEN**.
4. De oude fiets is niet gestolen, de **NIEUWE** fiets is gestolen.
5. De bomen waren niet weg, de bomen waren **KAAL**.
- *****
6. Het slot van de voordeur is niet kapot, de **BEL** van de voordeur is kapot.
7. De bel van de fiets is niet kapot, de bel van de **VOORDEUR** is kapot.
8. De rode appels waren niet zuur, de **GROENE** appels waren erg zuur.
9. De groene appels waren niet bitter, de groene appels waren erg **ZUUR**.
10. De bloemen waren niet verdroogd, het **GRAS** was helemaal verdroogd.
11. Het gras was niet verrot, het gras was helemaal **VERDROOGD**.
12. Rennen is niet gezonder dan fietsen, **LOPEN** is gezonder dan fietsen.
13. Lopen is niet gezonder dan zwemmen, lopen is gezonder dan **FIETSEN**.
14. De peren aan de boom zijn niet rijp, de **APPELS** aan de boom zijn rijp.
15. De appels in de mand zijn niet rijp, de appels aan de **BOOM** zijn rijp.
16. De boer heeft het gras niet gemaaid, de **TUINMAN** heeft het gras gemaaid.
17. De tuinman heeft het gras niet gezaaid, de tuinman heeft het gras **GEMAAID**.
18. De pijp ligt niet in de asbak, de **SIGAAR** ligt in de asbak.
19. De sigaar ligt niet op het bord, de sigaar ligt op de **ASBAK**.
20. Het licht is wel aan, de **KACHEL** is nog steeds niet aan.
21. De kachel is niet kapot, de kachel is nog steeds niet **AAN**.
22. De kok ging niet met vakantie, de **PORTIER** ging met vakantie.
23. De portier ging niet met pensioen, de portier ging met **VAKANTIE**.
24. De tomaten liggen niet in de schuur, de **AARDAPPELS** liggen in de schuur.
25. De aardappels liggen niet in de kelder, de aardappels liggen in de **SCHUUR**.
- *****
26. Dat dorp heeft geen slechte naam, dat **HOTEL** heeft een slechte naam.

APPENDIX 1 (CONTD.)

27. De lakens waren wel gewassen, de **KLEREN** waren niet gewassen.
28. Het natte hout knettert niet in het vuur, het natte hout **SIST** in het vuur.
29. Dat hotel heeft geen goede naam, dat hotel heeft een **SLECHTE** naam.
30. De bomen waren niet weg, de bomen waren **KAAL**.

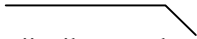
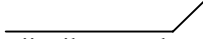
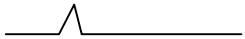
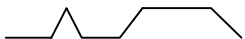

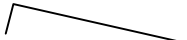
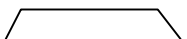
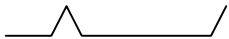


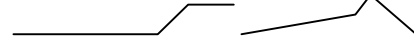
APPENDIX 2

For each group, the mean values and standard error of acoustic cues, given separately for the accented and unaccented versions F0 given in semitones, duration given in ms, intensity and spectral tilt given in dB

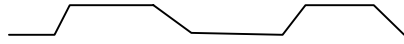
group	cue	Accented		UNaccented	
		mean	s.e.	mean	s.e.
laryngeal	F0	11.15	.14	1.16	.25
	duration	454.83	23.46	411.10	9.75
	intensity	72.32	.31	68.11	3.06
	Spectral tilt	-6.89	.54	-8.42	.52
TE	F0	5.69	.44	.98	.15
	duration	590.01	12.06	461.18	9.44
	intensity	69.79	.30	63.09	.39
	Spectral tilt	-.70	.57	-.46	.47
Es	F0	2.84	.35	.61	.12
	duration	655.63	16.89	553.47	16.16
	intensity	68.91	.38	64.23	.41
	Spectral tilt	6.66	.50	-.90	.53

APPENDIX 3

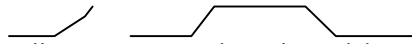
Stimulus material used in chapter three: sentences with melodic recipes

1.  1. Hij wil een meloen
2.  2. Hij wil een meloen
3.  3. Leen jij nu haar roman
4.  4. Leen jij nu haar roman
5.  5. Hij nam haar mee
6.  6. Hij nam haar mee
7.  7. Hij nam haar mee
8.  8. Wil hij wel weer mee?
9.  9. Ja, wij willen vooral winnen.
10.  10. Ja, wij willen vooral winnen.
11.  11. Marianne en Willem doen allebei raar

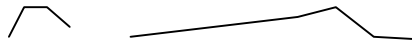
APPENDIX 3 (CONTD.)



12. Marianne en Willem doen allebei raar



13. Vrij warm, maar wel mooi wandelweer



14. Vrij warm, maar wel mooi wandelweer.

APPENDIX 4

CONTROL EXPERIMENT: INTONATION BIAS

The imitators' responses in the Imitation experiment (chapter four, section 4.2) were possibly based on what native speakers intuitively produce in terms of pitch events, when the sentences are spoken aloud (Lieberman, 1963; Chomsky & Halle, 1968). In this experiment, therefore, subjects were asked to read the sentences aloud, and the pitch events of the stretches of speech embedded in the read-aloud sentences were subsequently analysed to reveal the existence of an intonational bias.

A.4.1 METHOD

A.4.1.1 Stimulus material

The six sentences containing the stretches of speech described in 4.2.1.2 were used, but no punctuation was added to the sentences.

A.4.1.2 Subjects

Eighteen Native Dutch speakers participated. None of the subjects had participated in the previous experiment.

A.4.1.3 Procedure

Subjects were seated in a quiet environment, with a microphone placed in front of them. The subjects read the sentences (see above) aloud. If there was a dysfluency (omission, repetition or substitution), they were asked to repeat the sentence. The read-aloud sentences were recorded and stored on disk.

A.4.1.4 Transcription of read-aloud sentences

Per subject, the stretches of speech were transcribed in terms of the direction of the pitch change (rise, fall, rise-and-fall or no movement). Exactly the same procedure as in the Imitation experiment (4.2) was used to transcribe the utterances.

A.4.2 RESULTS

The control experiment was done to determine if certain pitch events were generally associated with the test items when embedded in sentences and read aloud. Figure A.4.1 gives the results, per stretch of speech.

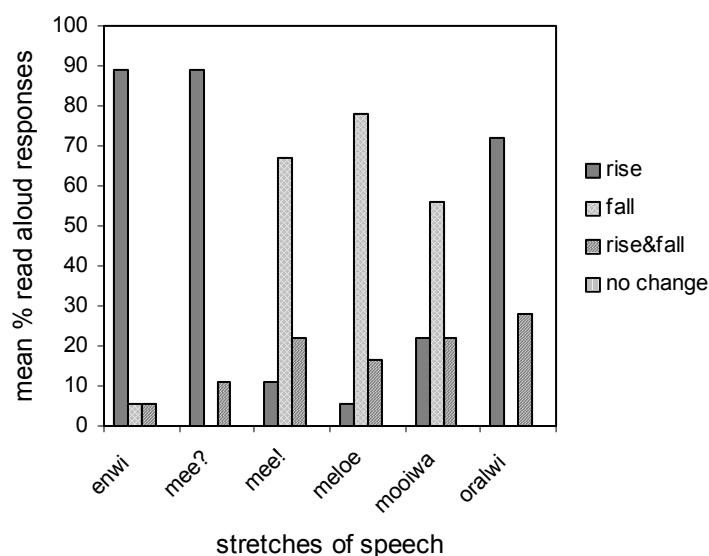


Figure A.4.1. For stretches of speech (x-axis): Percentage of responses produced by listeners per transcription category (y-axis).

For most stretches of speech strong preferences existed towards a specific pitch event. The rise bias for ‘enwi’ and the fall bias for ‘mooiwa’ confirm the biases found in the Imitation experiment (Figure 4.3). The results for ‘mee’ interrogative and ‘mee’ declarative also correspond to the results revealed in Figure 4.3. There was a strong bias towards fall for ‘meloe’, whereas in the Imitation experiment imitators produced falls or rise-and-falls. In the Control experiment, the subjects generally produced a flat hat, which would have prohibited the ‘rise-and-fall’ bias seen in the Imitation experiment. The rise bias seen in Figure A.4.1 for ‘oralwi’ was not found in the Imitation experiment. Overall, this experiment indeed confirms that an intonation bias exists for a number of the stretches of speech.

APPENDIX 5

Stimulus material used in chapter five

IK ZOU (PIET EN TOOS), OF KEES UITNODIGEN.
IK ZOU PIET EN (TOOS OF KEES) UITNODIGEN.
IK ZOU PIET, OF (TOOS EN KEES) UITNODIGEN.
IK ZOU (PIET OF TOOS) EN KEES UITNODIGEN.

IK ZOU (PATRICIA EN CORNELIUS), OF CATHARINA UITNODIGEN.
IK ZOU PATRICIA, EN (CORNELIUS OF CATHARINA) UITNODIGEN.
IK ZOU PATRICIA, OF (CORNELIUS EN CATHARINA) UITNODIGEN.
IK ZOU (PATRICIA OF CORNELIUS) EN CATHARINA UITNODIGEN.

IK ZOU (JOZEFIEN EN JOHANNES), OF WILLEMIJN UITNODIGEN.
IK ZOU JOZEFIEN, EN (JOHANNES OF WILLEMIJN) UITNODIGEN.
IK ZOU JOZEFIEN, OF (JOHANNES EN WILLEMIJN) UITNODIGEN.
IK ZOU (JOZEFIEN OF JOHANNES) EN WILLEMIJN UITNODIGEN.

IK ZOU PIET, EN (TOOS EN KEES), EN (PLEUN EN THIJS) UITNODIGEN.
IK ZOU (PIET EN TOOS), EN (KEES EN PLEUN), EN THIJS UITNODIGEN.
IK ZOU (PIET EN TOOS), EN KEES, EN (PLEUN EN THIJS) UITNODIGEN.

IK ZOU PATRICIA, EN (CORNELIUS EN CATHARINA) EN (CHRISTOFFEL
EN PETRONELLA), UITNODIGEN.
IK ZOU (PATRICIA EN CORNELIUS), EN (CATHARINA EN CHRISTOFFEL),
EN PETRONELLA UITNODIGEN.
IK ZOU (PATRICIA EN CORNELIUS), EN CATHARINA, EN (CHRISTOFFEL
EN PETRONELLA) UITNODIGEN.

IK ZOU JOZEFIEN, EN (MARIUS EN WILLEMIJN), EN (JOHANNES EN
ANNEMARIE) UITNODIGEN.
IK ZOU (JOZEFIEN EN MARIUS), EN (WILLEMIJN EN JOHANNES), EN
ANNEMARIE UITNODIGEN.
IK ZOU (JOZEFIEN EN MARIUS), EN WILLEMIJN, EN (JOHANNES EN
ANNEMARIE) UITNODIGEN.

APPENDIX 6

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 = esophageal); 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	Expected	UNexpected
LV	390 (13)	244 (7)	6 (0.5)	1 (0.3)	288 (24)	n.a.
LW	406 (11)	267 (8)	n.a.	n.a.	345 (34)	n.a.
TE	417 (11)	309 (12)	3 (0.5)	1 (0.2)	470 (26)	n.a.
Es	394 (13)	364 (13)	4 (0.6)	2 (0.4)	443 (22)	286 (17)

APPENDIX 7

Summary of findings which might be relevant in clinical practice

Features investigated (chapter)	Suggestions that might improve alaryngeal speaker's prosodic ability
Presence of Accent (2)	<p><i>When F0 apparently or mostly absent:</i> low-pass filter the speech then re-determine presence / consistency of F0</p> <p><i>When F0 inconsistent:</i> use sentences with highlighted words as training material to raise awareness of accent. Use auditory and visual feedback to train more consistent control over F0 (e.g., pitch contour).</p> <p><i>When F0 absent:</i> use contrastive words/sentences (e.g., KAnon vs kaNON; mee?? vs mee!!) to train consistent increase / decrease of vocal effort, using auditory and visual feedback to raise awareness (e.g., intensity contour).</p>
Different types of accents and boundary tones (3); Pitch direction (4)	<p><i>When F0 inconsistent:</i> practice glides (gradual but consistent increase / decrease of F0), first on vowels, then all-voiced words and short phrases, then longer questions versus statements. Auditory and visual feedback might again be beneficial</p> <p><i>When F0 absent:</i> training not recommended (see results & discussion chapter 4)</p>
Boundaries / phrasing (5)	<p><i>In TE speakers :</i> raise awareness of final lengthening (e.g., contrast words produced in initial position of a phrase with same word produced in phrase-final position) encourage the use of pauses only at major boundaries, except when the speaker's intelligibility is poor.</p> <p><i>In Es speakers:</i> explain adverse effect of wrongly positioned, within-phrase pauses. Train proper phrasing, starting with short, algebraic expressions (e.g., 1+{2 *3} vs {1+2} * 3). Increase the length and complexity of the phrases, using the stimulus sentences given in chapter 5 (appendix 5)</p>

SAMENVATTING

Deze dissertatie gaat over prosodie in sprekers bij wie het strottenhoofd, de larynx, is verwijderd. Het verwijderen van de larynx houdt in dat ook de stemplooien zijn verwijderd en dat een normale manier van stemgeven niet meer mogelijk is.

Normale sprekers hebben een goede controle over de stem, omdat ze de spanning van de larynxspieren en de luchtstroom vanuit de longen heel precies op elkaar af weten te stemmen. Zo kunnen sprekers bijvoorbeeld de trillingsfrequentie (fundamentele frequentie, oftewel F0) van de stemplooien nauwkeurig aansturen om de gewenste toonhoogte te bereiken. Behalve de toonhoogte kan de normale spreker ook de luidheid en de duur van een woord aanpassen zodat dit woord harder of langer wordt dan de aangrenzende woorden. Bij alaryngeale sprekers fungeert het bovenste deel van de slokdarm, ook wel de neoglottis genoemd, als nieuwe “stem”. Net als bij normale sprekers is deze alaryngeale stem afhankelijk van een luchtstroom. Door het verwijderen van de larynx is de normale luchtweg, van de longen via de mond en/of de neus, onderbroken: de luchtpijp is naar voren gebogen en in de huid van de hals vastgehecht. Deze opening in de hals wordt tracheostoma genoemd. Een stemprothese (eenrichtingsklep) wordt in een fistel tussen de luchtpijp (trachea) en de slokdarm (oesophagus) geplaatst zodat de lucht uit de longen in de slokdarm kan worden geblazen. Tegenwoordig maakt de meerderheid van de alaryngeale populatie gebruik van zo'n stemprothese (in deze dissertatie: TE sprekers). Een alternatief is om met de mond lucht te “happen” en deze lucht met behulp van de tong, wangen en mondbodem in de slokdarm te persen (ook wel bekend als de injectiemethode; in deze dissertatie: Es sprekers). De lucht die in de slokdarm geblazen wordt, hetzij uit de longen of uit de mond, zorgt ervoor dat de neoglottis gaat trillen. Dit geluid is de alaryngeale stem. Er is veel variatie in de vorm, grootte en plaats van de neoglottis, maar ook de controle die alaryngeale sprekers kunnen uitoefenen over de neoglottis wisselt. Sommige alaryngeale sprekers zijn er redelijk goed in bijvoorbeeld toonhoogtebewegingen aan te geven, terwijl dat bij andere sprekers niet, of niet consequent lukt. Verder hebben sprekers die gebruikmaken van de injectiemethode minder lucht tot hun beschikking, waardoor het produceren van meerdere woorden op één “spreekadem” niet vanzelfsprekend is. In dit project werd onderzocht of de manier waarop alaryngeale sprekers prosodie produceren vergelijkbaar is met de manier waarop normale sprekers dat

doen. Een tweede vraag was of luisteraars de prosodie in alaryngeale spraak net zo goed kunnen waarnemen als de prosodie in normale spraak.

In normale spraak zijn toonhoogte, luidheid en duur de “dragers” van prosodie. De prosodische structuur helpt de luisteraar om de spraak goed en snel te analyseren. Dit onderzoek richtte zich op twee prosodische functies, namelijk focus en frasering. Focus is, eenvoudig gezegd, het accentueren van belangrijke of nieuwe woorden in een zin. Toonhoogtebewegingen vertellen de luisteraar welk woord geaccentueerd en dus belangrijk is. Door minder goede sprekers te laten participeren, konden we vaststellen in welke mate een afwezige of inconsistente F0 een belemmering zou zijn voor de communicatie. Frasering is het groeperen van woorden in betekenisvolle eenheden. Een verlenging van de laatste lettergreep geeft aan waar de grens van een korte frase ligt, bij grotere (syntactische) eenheden worden toonhoogtebewegingen en pauzes toegevoegd. Door injectiesprekers in te sluiten, konden we nagaan tot welke mate frasering negatief wordt beïnvloed door een beperkte luchtvoorraad. De achtergrondinformatie, die hierboven kort is samengevat, werd in **hoofdstuk een** uitvoerig gepresenteerd.

In **hoofdstuk twee** werd onderzocht of alaryngeale sprekers zinsaccent communiceren in dezelfde mate en op dezelfde wijze als normale sprekers. Normale, TE en Es sprekers produceerden zinnen waarin de zinsaccenten door de voorafgaande context werden aangegeven. Vervolgens identificeerden luisteraars voor iedere gesproken zin welk woord volgens hen het accent droeg. Uit de resultaten van dit luisterexperiment bleek dat het vermogen van alaryngeale sprekers om een zinsaccent te communiceren erg wisselt: er was veel variatie tussen sprekers. Als groep verschilden de Es sprekers significant van de normale sprekers, maar niet van de TE sprekers. Een akoestische analyse van het spraaksignaal wees uit dat F0 in ongeveer de helft van de alaryngeale sprekers afwezig, of niet consistent aanwezig was. Luisteraars konden de aanwezigheid van accent minder goed waarnemen in deze ‘niet-F0’ sprekergroep. Dit bevestigt het belang van F0 voor het waarnemen van zinsaccent. Toch konden luisteraars vaak wel de zinsaccenten in ‘niet-F0’ spraak waarnemen. De verwachting was dat de ‘niet-F0’ sprekers voor hun gebrek aan F0 zouden compenseren door de geaccentueerde woorden met bijvoorbeeld een grotere duur of intensiteit te produceren, maar dat was niet het geval. Een tweede luisterexperiment liet zien dat luisteraars een toonhoogte-achtig fenomeen konden waarnemen in de zinnen van de ‘niet-F0’ sprekers. In plaats van compensatie door duur of

luidheid, leken ‘niet-F0’ sprekers eerder gebruik te maken van een alternatief toonhoogtesysteem, waarmee ze de aanwezigheid van accenten grotendeels konden aangeven.

In **hoofdstuk drie** werd daarom onderzocht of luisteraars inderdaad spraakmelodie konden waarnemen in ‘niet-F0’ spraak. De spraakmelodie is namelijk opgebouwd uit een opeenvolging van toonhoogtebewegingen die verschillende soorten zinsaccenten, grensmarkeringen en de zinstype (bijvoorbeeld een vraag) kunnen aangeven. Drie TE sprekers en twee normale sprekers die fluisterden werden gevraagd modelzinnen te imiteren waarvan bekend was welke spraakmelodie (opeenvolgende toonhoogtebewegingen) ze bevatten. De fluisteraars waren controlesprekers: F0 komt in hun spraak helemaal niet voor. De eerste twee luisterexperimenten gaven aan dat luisteraars niet alleen een spraakmelodie konden horen in de gesproken zinnen van de ‘niet-F0’ sprekers, maar ook oordeelden luisteraars dat delen van deze ‘niet-F0’ spraakmelodieën inderdaad overeenstemden met de originele spraakmelodieën van de modelzinnen. In een derde experiment werden expertfonetici gevraagd de spraakmelodieën te transcriberen van zowel de gesproken ‘niet-F0’ zinnen als de modeluitingen. Daaruit bleek dat de toonhoogtebewegingen in de ‘niet-F0’ zinnen minder accuraat getranscribeerd werden dan de toonhoogtebewegingen in de modelzinnen. Ook bleek voor de ‘niet-F0’ zinnen dat sommige toonhoogtebewegingen beter werden waargenomen dan anderen. Meer geleidelijke toonhoogtebewegingen werden bijvoorbeeld eerder verward met abrupte toonhoogtebewegingen dan andersom.

In **hoofdstuk vier** werd onderzocht of naïeve luisteraars onderscheid konden maken tussen stijgende en dalende toonhoogtebewegingen in ‘niet-F0’ spraak, en werd uitgezocht hoe de sprekers stijgingen en dalingen realiseren. Uit het eerste luisterexperiment werd duidelijk dat luisteraars tot op zekere hoogte de richting van een toonhoogtebeweging konden waarnemen, maar dat hun respons grotendeels gestuurd werd door de omliggende context (de rest van de zin). De stukjes spraak die een stijging of daling bevatten werden voor het tweede luisterexperiment uit de omliggende zincontext gehaald. De “uitgeknipte” spraakfragmenten werden ook gefilterd in 5 banden, die ieder een bepaald deel van het frequentiedomein omvatten. Aan luisteraars werd gevraagd de richting van de spraakfragmenten en van de gefilterde banden te identificeren. Uit dit experiment bleek dat de

luisteraars bij slechts één ‘niet-F0’ spreker een stijging van een daling konden onderscheiden en dat ze zich baseerden op informatie uit de laagste frequentieband. Een akoestische analyse bevestigde dat bij deze spraakfragmenten een soort “pseudo-periodiciteit” kon worden gemeten, in de laagste band. De spraakfragmenten van de andere twee ‘niet-F0’ sprekers bevatten weinig toonhoogte-informatie. De fluisteraars varieerden de spectrale helling, en dit werd waargenomen als toonhoogtebeweging. De spectrale helling is sterk gerelateerd aan spreekinspanning en het vermoeden is dat sprekers, wanneer ze geen F0 tot hun beschikking hebben, de spreekinspanning zouden kunnen variëren om toonhoogtebewegingen aan te geven.

In **hoofdstuk vijf** werd onderzocht hoe TE en Es sprekers frasegrenzen realiseren en overdragen. Uit een luisterexperiment werd duidelijk dat luisteraars de bedoelde frasing minder goed konden waarnemen in de Es sprekergroep dan in de normale en TE sprekergroepen. Eén van de Es sprekers behaalde echter wel goede resultaten. Akoestische analyses lieten zien dat de verschillende sprekergroepen andere combinaties van akoestische parameters gebruikten om een frasegrens aan te geven. Normale sprekers gebruikten finale verlenging en toonhoogtebewegingen; fluisteraars gebruikten alleen verlenging; TE sprekers gebruikten finale verlenging en pauzes en de Es sprekers gebruikten alleen pauzes. Eén Es spreker paste zijn spreekstijl aan om pauzes binnen een frase te voorkomen. Bij de andere twee Es sprekers kwamen pauzes binnen een frase vaak voor. Deze twee Es sprekers maakten wel een duuronderscheid tussen grammaticale pauzes en injectiepauzes, maar dit was perceptief niet goed te onderscheiden, zoals bleek uit het luisterexperiment.

Tot slot: om melodische en ritmische informatie te communiceren, gebruiken alaryngeale sprekers niet persé dezelfde akoestische kenmerken als normale sprekers. De intentie van de alaryngeale spreker wordt hierdoor niet altijd correct waargenomen door luisteraars. Klinici (zoals logopedisten) zouden kunnen vaststellen welke akoestische kenmerken nog aanwezig zijn en door specifieke training van de aanwezige kenmerken de effectiviteit van het communicatieve vermogen kunnen verhogen. De conclusies, die hierboven heel kort genoemd zijn, werden in **hoofdstuk zes** beschreven en ook werden de beperkingen van dit onderzoek besproken en suggesties voor verder onderzoek gedaan.

CURRICULUM VITAE

Maya van Rossum was born in Maasluis, the Netherlands, on October 29, 1965. She received her (pre-) primary and secondary education in Durban, South Africa, where she finished high school in December 1983. In September 1988, she obtained her degree in Logopaedics and Audiology at Pretoria University. Between October 1988 and June 1993, she was employed as a speech pathologist and audiologist at the ENT-Department, Tygerberg Academic Hospital, in Cape Town. From August 1993 until October 1995, she held various short-term posts as audiometrist and speech pathologist in England. She was a student at Utrecht University, the Netherlands, from January 1996 to July 1998, where she obtained her Masters degree in Phonetics. She worked part-time as a speech pathologist at the ENT-Department, Leiden University Medical Centre between February 1998 and May 2003. From September 1999 until January 2005 she was a (part-time) PhD student at the Utrecht institute of Linguistics, where she conducted the research that is described in this dissertation. She is currently working as a post-doc researcher at the Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, in Amsterdam.