Speech quality in oral or oropharyngeal cancer patients

The development and evaluation of objective speech assessment methods

The studies described in this thesis were carried out at the Department of Otolaryngology / Head and Neck Surgery, *Vrije Universiteit* Medical Center, Amsterdam, The Netherlands.
The research was supported by the Dutch Cancer Society, grant number **RUG 2008-3983**: Prediction and prevention of swallowing dysfunction after curative (chemo)-Radiation in head and neck cancer.

Publication of this thesis was financially supported by: Rieta Mulder fotografie www.rietamulder.nl, Dutch Cancer Society
Cover illustration: J. Nederhof
Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress

Speech quality in oral or oropharyngeal cancer patients
Thesis, Vrije Universiteit Medical Center, Amsterdam, The Netherlands
ISBN: 978-90-8891-645-8

VRIJE UNIVERSITEIT

Speech quality in oral or oropharyngeal cancer patients

The development and evaluation of objective speech assessment methods

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op maandag 10 juni 2013 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Marieke Jitske de Bruijn

geboren te Amsterdam

promotoren:	prof.dr. I.M. Verdonck-de Leeuw
	prof.dr. C.R. Leemans
	prof.dr. J.A. Langendijk
copromotor:	dr. L. ten Bosch

Speech quality in oral or oropharyngeal cancer patients

The development and evaluation of objective speech assessment methods

The studies described in this thesis were carried out at the Department of Otolaryngology / Head and Neck Surgery, *Vrije Universiteit* Medical Center, Amsterdam, The Netherlands.
The research was supported by the Dutch Cancer Society, grant number **RUG 2008-3983**: Prediction and prevention of swallowing dysfunction after curative (chemo)-Radiation in head and neck cancer.

Publication of this thesis was financially supported by: Rieta Mulder fotografie www.rietamulder.nl, Dutch Cancer Society
Cover illustration: J. Nederhof
Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress

Speech quality in oral or oropharyngeal cancer patients
Thesis, Vrije Universiteit Medical Center, Amsterdam, The Netherlands
ISBN: 978-90-8891-645-8

VRIJE UNIVERSITEIT

Speech quality in oral or oropharyngeal cancer patients

The development and evaluation of objective speech assessment methods

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op maandag 10 juni 2013 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Marieke Jitske de Bruijn

geboren te Amsterdam

promotoren:	prof.dr. I.M. Verdonck-de Leeuw
	prof.dr. C.R. Leemans
	prof.dr. J.A. Langendijk
copromotor:	dr. L. ten Bosch

Contents

Chapter 1

1

General Introduction

1

Head and neck cancer

Head and neck cancer (HNC) comprises tumours in the upper air- and digestive ways in the oral cavity (tongue, gingiva, floor of the mouth, palatum durum and buccal mucous membrane), oropharynx, nasopharynx, larynx, parotids and glands but excludes tumours of the brains and spine. Annually, approximately 2,900 individuals in the Netherlands are diagnosed with HNC of which the male-female ratio is 1.5. The incidence in males tends to stabilize while the incidence in females ascends due to expenditure of alcohol and smoking. Most newly diagnosed HNC patients are over 45 years of age. The majority of tumours is diagnosed in an advanced stage because of the absence of complaints and late awareness of suspicious symptoms.[1]

In general, the TNM classification system is used to describe tumour extension, including three anatomical components of the tumour: the extent of the tumour (T), regional lymph node metastasis (N) and distant metastasis (M). Tumour classification may serve as the basis for choice of treatment, provides information about the potential path of tumour growth, simplifies exchange of information between physicians, and enables comparison between groups of patients such as in clinical research. These three components form the basis for defining the stage categories (ranging from I to IV) with higher stages reflecting worse prognosis.[2] See figure 1 for an overview of the head and neck region.

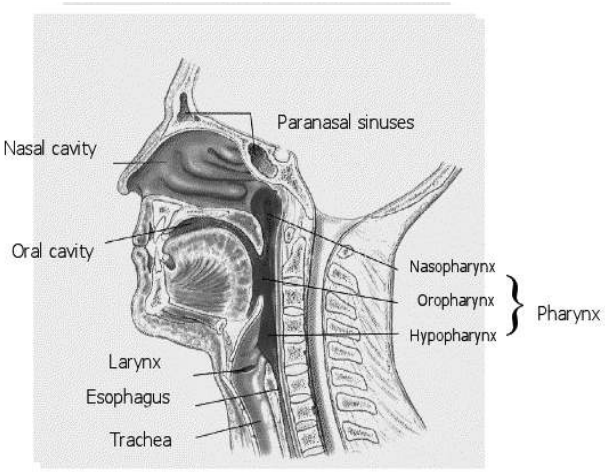

Figure 1. Overview of the head and neck region.

Diagnostics and treatment

The first suspicion of a tumour in the head and neck area often arises from the presence of ulcers, hemorrhage, growths in the mouth or throat, pain, swelling of the neck, or voice or swallowing problems.[1] Diagnostic imaging with MRI or CT-scan is performed to identify the location and size of the tumour next to physical examination and biopsy. In general, treatment of HNC consists of surgery followed by adjuvant (chemo)radiation and if necessary followed by reconstructive surgery[3], or primary (chemo)radiation. The choice of treatment depends on the risk of complications, the presence of prognostic factors, the risk of secondary tumours, the probability to control tumour growth and expected quality of life during and after treatment.

Surgery is chosen when a total excision of the primary tumour can be carried out, when metastases of the lymph nodes in the neck area can be controlled and when reconstruction is possible [4]. Reconstructive surgery is used to close the defects of the tissue after removal of the tumour and surgical margins. There are multiple possibilities to close the wounds and to obtain an esthetic result, ranging from primary closure to more complicated surgery using donor tissue such as skin flaps or bone when the wound is too large for primary closure. Skin flaps or bone originate from the radial fore arm or the leg of the patient[3].

Radiotherapy can be given as primary therapy, or in addition to surgery or concurrently with chemotherapy. During radiotherapy, ionizing radiation is delivered to the tumour. The effect of radiotherapy is based on the principle that radiation causes DNA damage in both normal cells and malignant cells. Cancer cells, however, are less capable to repair radiation-induced DNA damage than normal cells and are, therefore, more likely to undergo several ways of cell death. This radiation effect depends on the total radiation dose, the dose per fraction, and the overall treatment time of radiation. Radiation-induced side-effects commonly seen in head and neck cancer patients include pain, fatigue, dry mouth (xerostomia), fibrosis, mucositis, altered sense of taste and swallowing problems.[5, 6, 7]
Concurrent chemoradiation is indicated in younger patients with locally advanced disease (stage III-IV). The results of a meta-analysis indicated that chemotherapy when given concurrently with radiation significantly improves overall survival with 6.5% after 5 years.[8]

1

Health related Quality of Life

Due to more intensified treatment regimens, the prevalence of head and neck cancer survivors gradually increases.[9] Therefore, quality of life after treatment for HNC becomes increasingly important. Quality of life can be measured using questionnaires such as those developed by the European Organisation for Research and Treatment of Cancer (EORTC).[10, 11] The core questionnaire is the EORTC QLQ-C30, that measures general characteristics of quality of life of patients with cancer, such as emotional functioning, social functioning, and pain and fatigue. The core EORTC QLQ-C30 questionnaire can be supplemented by additional tumour-specific modules, such as the module for patients with head and neck cancer, the EORTC QLQ-H&N35, including items as oral pain, sticky saliva, trismus and speech and swallowing problems

Tumour growth and treatment of HNC often results in changes in the upper aerodigestive tract which may lead to a wide variety of complaints, such as pain, dry mouth, less flexibility of structures and, in some cases, the (partial) loss of organs in the head and neck area such as the larynx. These changes may result in loss of function. Since the head and neck area accommodates various functions, tumour-related or treatment-related anatomical and functional changes may have devastating consequences in daily life of patients. Social events such as eating (swallowing) in public and speaking as well as changes in facial appearance may be experienced by the patient as unpleasant and awkward to seriously problematic. The consequences of HNC often result in a lower quality of life of patients.[12-14] In head and neck cancer patients, oral dysfunction, swallowing and speech problems are often the main outcomes negatively affecting health related quality of life (HRQOL). Although treatment of HNC improved over the last decades, as evidenced by a higher survival rates, it still comes at the expense of the aforementioned treatment-related side effects. Research is needed to improve treatment and to monitor functionality after treatment. Examples of on-going research include patients treated with reconstructive surgery, intensity-modulated radiation therapy (IMRT) and chemo-radiation. IMRT is a radiation technique that is aimed at sparing salivary glands and thus prevention of xerostomia.
Chemo-radiation is an organ sparing technique. In HNC chemo-radiation may be used to preserve the larynx, which is necessary to maintain voice production. However, it is not yet fully unravelled if and how these

techniques contribute to a better conservation of organs and to a higher quality of life. Standardized assessment protocols are needed to monitor functional outcomes in patients after altered treatment modalities.

Head and neck cancer and speech quality

The present study focuses on speech problems related to HNC. Patients may experience problems that substantially affect quality of speech- because of the direct effect of the cancer on the patient's capabilities and anatomy for speech production. Quality of articulation and intelligibility frequently deteriorate after treatment. Speech quality is the global impression of the observer in the perspective of the framework of every day speech. Quality of articulation is principally an assessment of speech production whereas quality of intelligibility is the accuracy of perception by the listener and can be measured or quantified in terms of units such as syllables, words or entire sentences. There is not always a direct correlation between good speech quality and a high score on intelligibility.[15]

Many patients encounter problems in daily life concerning conversation and social encounters. The former fluency of speech production is frequently degraded by speech problems leading to experiencing less satisfaction with life and less emotional well-being in serious cases.[5, 16]

The production of speech is a complicated process as many speech organs and muscles must act in a strict synchronous and timely manner. In short, during the production of speech, air originating from the lungs passes the vocal folds. Depending on the tension on the vocal folds, they vibrate with a specific frequency, or let the air flow pass further into the vocal tract.. Vibration of the vocal folds adds a frequency pulse to the air stream which results in the presence of a fundamental frequency. The air stream moves further through the oropharynx and leaves the body through the mouth and depending on the speech sound, also the nose. While for speech production all cavities in upper body and head play their role in the acoustic resonance, the oropharynx, oral and nasal cavities are the most relevant. The oropharynx and oral cavity contain speech organs such as muscles, cavities, tongue, teeth and lips which give shape to the vocal tract. The shape of the vocal tract act as an acoustic filter that determines the resonance frequencies

1

of the tract which largely determine the timbre (spectral quality) of the produced speech sounds. While for voiced speech sounds the vibration of the glottis is important, in general the timbre of the speech sounds is determined by the shape of the oral and nasal cavity and the way they resonate together. Due to the tongue movement and jaw motility this filter is pliable and flexible. The alterations in the shape of the filter cause deviations in the character of the air stream and therefore results in variation in the resulting speech sounds (see figure 2). While vowels are characterized by an open oral cavity, consonants are determined by place and manner of constrictions that hamper or even may block the airflow, which is the case in e.g. fricatives such as /f,s,v,z/ and plosives /p,t,k,b,d/, respectively. For instance, elevating the back of the tongue to the velum leads to the production of velar sounds as /x/ (as in the Dutch word 'rug'), /k/ and the velar nasal /ng/ (as in Dutch 'ring'). Pushing the tip of tongue to the alveolar ridge leads to the production of alveolar sounds as /d/ and /z/ (voiced alveolar sounds) and /t/ and /s/ (voiceless alveolar sounds). The speech sound /r/ is represented twice, which is due to the different places of articulation of /r/ in Dutch.

It will be clear that changes in the head and neck area due to the presence of a tumour or induced by treatment may cause alterations in the speech production system. In healthy persons the speech organs are intact and flexible, whereas in patients treated for HNC the speech organs and tissue are often damaged and stiffened. Speech sounds produced by patients treated for HNC may become distorted and results in articulation and nasality problems and thereby worsened intelligibility.[17] At the level

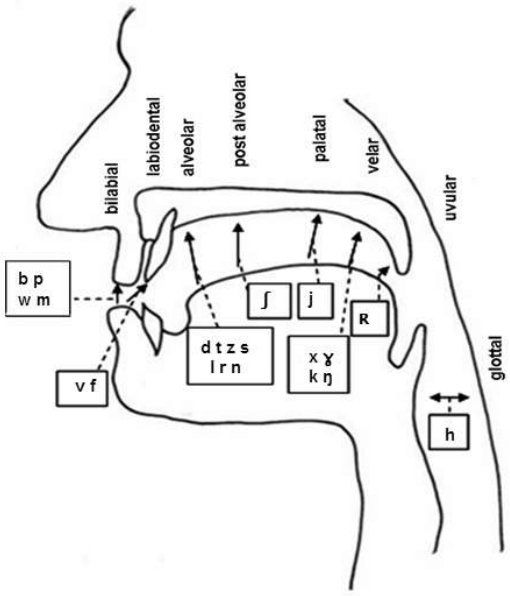

Figure 2. Schematic representation of the vocal tract. Various places of constrictions leading to different consonants are depicted.

of individual speech sounds, it often appears that not all speech sounds produced by one individual patient are distorted and that the level of problematic production of speech sounds depends on the tumour stage and location.[17-20] To a large extent, distortions are patient-specific. For instance, velar speech sounds are likely to become distorted when a patient is treated for a large tumour in the oropharynx, whereas a patient treated for a lip tumour may experience trouble producing labial speech sounds.

Apart from their impact on *speech production*, the changes in the head and neck area may also be reflected in difficulties in all stages of the *swallowing process*. Swallowing dysfunction (dysphagia) also largely depends on tumour stage, location, and treatment modality.

Speech assessment

The effect of HNC and its treatment on speech quality can be measured by different methods.

Subjective judgment is often considered to be the gold standard and is the most frequently used method to assess speech quality.[21-24] Subjective judgment is performed by either naive listeners without formal training or by expert listeners, such as speech pathologists. The listeners usually rate various aspects of speech quality such as general intelligibility, articulation, nasality, or speech sounds in isolation. However, in clinical practice, these judgments are not sufficiently accurate and consistent to allow wide sharing between hospitals when multiple therapists are involved during or after treatment.

Another example of subjective judgment -besides judgment by listeners- is the judgment of speech quality by patients themselves, rating their own quality of speech and related quality of life. Examples of patient reported speech outcome questionnaires are the speech subscale of the EORTC-QLQ-H&N35[13] and the Speech Handicap Index (SHI).[25]

Next to subjective assessment measures, one can make use of objective assessment measures. For objective measures, instruments to automatically determine the quality of speech are used instead of the combination of the

human ear and human decision making. Objective assessment of speech is not as often used as subjective assessment and unambiguous standardization to perform objective speech analyses is lacking.[26-28]

In this dissertation, validity and feasibility of two objective assessment methods are investigated: (1) 'classical' acoustic-phonetic analysis and (2) analysis by means of automated classification methods on the basis of short-term estimates of *articulatory features* (AF). Compared to spectral features, articulatory features form another representation of speech that specifically focus on the properties of speech from the production point of view. Such features, and especially the automatic estimation of these features from the audio signal, have been investigated for the purpose of improving automatic speech recognition as an alternative for the conventional mel-frequency cepstrum coefficients (MFCC) based features[29-36] and more recently for the assessment of pathological speech.[37] Middag et al. (2010)[37] used these features to estimate the running speech intelligibility as relevant indicator of the communication efficiency of a speaker, requiring no knowledge of what the speaker was supposed to say. Regularities and irregularities in the articulatory feature patterns over time can be detected by classification approaches (algorithms) that are able to find structures in complex data sets. By studying the properties and performance of such algorithms, structure in data sets can be found that would otherwise be difficult to detect manually.

Articulatory Features

In this thesis, to estimate articulatory features from the speech signal we explored one particular classification method that is well-known in the area of classifiers, known as Artificial Neural Networks (ANN). ANNs have been developed since the nineteen eighties and are now one of the conventional machine learning algorithms. The layered architecture of an ANN is inspired by biological neural networks as found in the brain. Biological neural networks are made up of real biological neurons that are interconnected.[38-41] In the field of neuroscience, they are often identified as groups of neurons that perform a specific physiological or neurological function. Inspired by this neural connectivity, artificial neural networks are composed of interconnecting artificial neurons, organised in interconnected layers.

Nowadays, broadly speaking, artificial neural networks are applied for two different purposes. The first purpose is to gain understanding in the

functioning of real biological neural networks. The second purpose is to solve problems in the field of artificial intelligence. ANNs are potentially interesting because, unlike many other algorithms, ANNs are able to model non-linear mappings. With the advent of increasing computing power, they became popular in the early nineteen nineties. Since then, they are among the most-used algorithms. ANN can be used as a multi-class classifier and is able to assign a graded membership to pre-specified classes. It is particularly this property that will be used in this thesis. ANNs are, just like Support Vector Machines -another method-, one way of automatic estimation of characteristics of the speech signal. Both methods fall under the realm of Articulatory Feature analysis (AF). It is more correct to use the term Articulatory Feature analysis in comparison with acoustic-phonetic analysis. However, the method of 'Artificial Neural Network' (ANN) is used in this thesis. Therefore we refer to ANN when speaking about the method of automatic feature estimation.

In acoustic-phonetic analysis acoustic aspects of speech sounds are measured in the speech signal of which we know they matter, such as duration of speech sounds, fundamental frequency and other characteristics of the frequency spectrum. These parameters have been used in acoustic phonetic research since the fifties. Articulatory feature estimation is a technique that became available much more recently and their use to assess pathological speech has matured only over the last couple of years.
In this thesis, ANNs have been applied to label an unknown utterance in terms of time-varying articulatory relevant properties.[42, 43] These methods focus on basic properties of the speech signal that characterize the speech signal in a way more similar to acoustic features or phonological features such as 'voicing', 'manner' and 'place' of articulation.[44, 45] See figure 3.

In phonetics and phonology phonological/phonetic features represent distinctive properties of speech sounds. One of the well-known proposals has been put forward by Chomsky & Halle (1986)[46], distinguishing properties such as *manner* of articulation (e.g. vowel, semivowel, fricative, nasal, liquid, stop), *place* of articulation (e.g. labial, bilabial, velar, front, back) and *voicing* in a formal framework. Compared to the Mel-Frequency Cepstral Coefficients (MFCCs) used in conventional ASR, Articulatory Features (AF) can provide more information about how speech sounds are produced, rather than represent spectral details, as MFCCs do. At least in theory AFs are therefore

considered to be more appropriate to indicate deviations in speech production. Another potential advantage of an AF description is that AFs are asynchronous, that is, they allow speech parts to be classified as 'nasal vowels', which is an advantage compared to the conventional 'beads-on-a-string' description of speech in terms of sequences of phone like symbols. Nowadays several machine learning approaches are available to estimate AFs on the basis of real acoustic waveforms as input, including Artificial Neural Nets (ANNs)[29] and Support Vector Machines (SVMs).[34]. AFs have been used for improving Automatic Speech Recognition in noisy conditions.[30, 35] More recently, AFs have been applied to objectively assess pathological speech via automatic phone intelligibility rating.[37, 47]

The AFs used in this dissertation are similar to those described in Middag et al.[37] They are output from ANNs available in the NICO toolkit.[48] The ANNs take a mono waveform, (re)sampled at 16 kHz, as input. In the first step, the waveform is represented using Mel Frequency Cepstral Coefficients (MFCCs). Via a shifting analysis window (with a shift of 10 ms), 100 MFCC frames per second are computed. In the second step, this MFCC sequence is input for the ANNs. By using the outputs of all ANNs, the entire speech signal is eventually represented as a sequence of vectors consisting of articulatory feature values (estimations). The topology of each ANN was modeled by a feed-forward network consisting of one input layer, one hidden layer and one output layer. This topology is fixed and kept the same during training and test.

Table 1 provides an overview of the different ANNs used. Each ANN corresponds to one articulatory feature, specified per row in table I. The name of the feature is given in the first column, while the feature values are mentioned in the second column.

- *Training data*

The ANNs used in this study were trained on the IFA corpus.[49] This corpus is manually segmented on the word and phone level. The ANN training is fully supervised: the input of the ANN during training consists of the sequence of MFCC frames, in combination frame-by-frame with the reference feature values for a particular articulatory feature. These reference AF values were determined by translating the provided phone labelling into an AF sequence, by using a predefined canonical phone-feature translation table.

Table 1. Overview of the ANNs used in this study to estimate the articulatory features. By aggravating all outputs, the result is a 28 dimensional feature vector, generated each 10 ms. Next to the real output values, each feature may have a 'NULL' output in the case the feature does not apply, such as in silence portions; 'nill' meaning that a value is not defined. 'Approx' means approximant; 'fric' means fricative; 'alveol' means alveolar; 'labiodent' means labiodental. The effect of this can be seen in figure 3.

Feature	Set of output values	component
Manner	NULL-approx-fric-nasal-stop-vowel	1-6
Place	NULL-alveol-high-labiodent-low-mid-velar	7-13
Voice	NULL-unvoiced-voiced	14-16
Front-back	NULL-back-central-front-nil	17-21
Round	NULL-nil-round-unround	22-25
Static	NULL-dynamic-static	26-28

- *Input*

The input of the ANN consisted of 7 consecutive MFCC frames, centered around the MFCC frame in question. This allows the ANN to take context into account into one vectorial representation, which is necessary to e.g. interpret *event* phones such as stops as a single classification unit.

- *Output*

After training, the output of the ANN provides an estimation of the presence of the corresponding articulatory property in the input that is presented to the ANN. For example, the *manner* feature is modeled by the manner-ANN which has 6 units in its output layer. These six output units of the manner-ANN estimate the degree of NULL, approximant, fricative, nasality, stop, and vowel, respectively. The NULL value is a unit that takes positive values if the network is not able to positively assign values to any of the other five units, and is used where the other features make no sense, for example in the case of silence and other non-speech portions in the signal. In total, the manner-ANN classifier provides 6 values each 10 msec. All output values vary between 0 (corresponding property absent) and 1 (property present), but are not constrained to have a sum equal to 1.

The full output of the AF analysis is constructed by applying all 6 ANNs synchronously in parallel and combining the outputs of these ANNs into one 28-dimensioanl AF-vector. This vector is updated each 10msec implying that an entire utterance results in an AF matrix. Of this matrix, the number of rows is determined by the number of features (here 28), while the number of columns is determined by the duration of the input utterance.

- *Hidden units*

One of the model parameters in an ANN is the number of hidden layers and the number of units in each of the hidden layers ('hidden' units). The larger the number of hidden units, the more complex an ANN can model a mapping from the input space to the output space. In this study, we adopted a setting that has been suggested in the literature.[29] Each ANN has one hidden layer, consisting of 300 hidden units.

Table 2 provides the accuracy of the ANNs on held out test speakers (also taken from the IFA corpus), both for the features *manner*, *place*, *voicing*, *rounding*, *front-back* and *static*. The accuracy is measured in terms of frame accuracy. In addition, a confusion matrix for the AF *manner* is presented.
Figure 10 provides an example of the behaviour of the AFs over time. The input is a wave file with duration of 1.53 seconds. Along the horizontal axis, time is displayed (in terms of frames, frame shift is 0.01 sec). The vertical axis shows the 28 AF estimations in the order according to table 2. From the figure, it can be seen that there are 23 frames with leading silence and about 10 frames with trailing silence.

Table 2. This table shows the performance (frame accuracy, in percentages) per ANN on independent test data (independent set of healthy test speakers, from the IFA-corpus), after training.

Manner	84.7
Place	76.7
Voice	93.5
Front-back	83.6
Round	87.4
Static	89.7

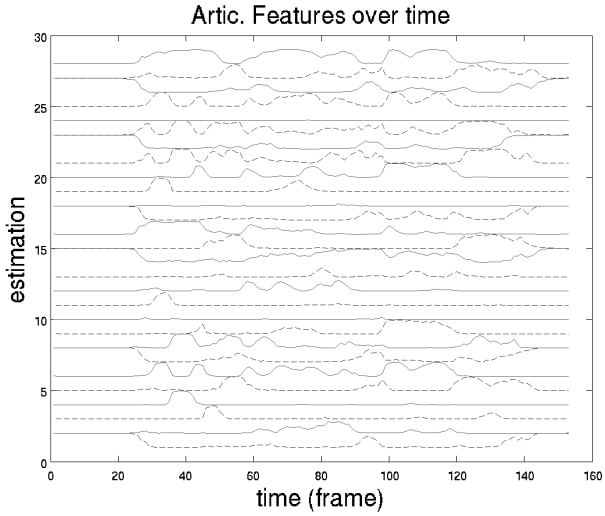

Figure 3. Representation of estimations of articulatory features over time. The horizontal axis represents time; the vertical axis represents the 28 features. The lines of the figure represent the presence and amount of estimated values for the features over time. The duration of one frame is .01 seconds.

Studies using acoustic-phonetic analysis or automatic classification based on ANN to assess speech quality of HNC patients are scarce.[26, 27, 50-54] In this dissertation, these two objective methods are combined into one speech assessment protocol aiming to predict subjective evaluations of speech quality.

1

Speech processing

In order to be able to work with either acoustic-phonetic analysis or ANN, the recorded speech has to be prepared for processing. In this dissertation, we used the computer program Praat.[55] Praat is used because it is a well known and regularly updated program that is frequently used in phonetics and speech research. Also, Praat allows the use of scripts facilitating (semi)automatic extraction of a spectrogram and specific spectral parameters. A spectrogram is a visual representation of the acoustic speech signal and contains information about the amplitude (represented by blackness and frequencies (vertical axis) over time (horizontal axis)). Then the spectrogram is segmented; each segment can be assigned a label (e.g. phone label). See figure 4 for the spectrogram and segmentation of the Dutch word “oppassen” as processed by the computer program Praat.

Careful and accurate segmentation of speech sounds is important in order to obtain spectral information over the intended time frame of a speech sounds and to avoid other sound effects. The boundary between two speech sounds can be identified by audio representation (sound) and supported by visual inspection of the spectrogram (changes in the spectral balance). The use of “visible speech” (spectrogram) provides extra information and is considered an essential requisite when segmenting and rating pathological speech.[56]

After segmentation and labelling of all required speech sounds under investigation, detailed spectral information can be retrieved in Praat by using a script that specifies the type of spectral information and time frame.

Segmentation is also needed when training the Artificial Neural Network. ANN produces estimations of speech features over time. Segmentation is consequently needed to focus onto a specific piece ('frame') of speech to calculate the amount of a specified articulatory feature over a range of frames. A frame depicts a time unit of .01 seconds. Segmentation is not needed during testing.

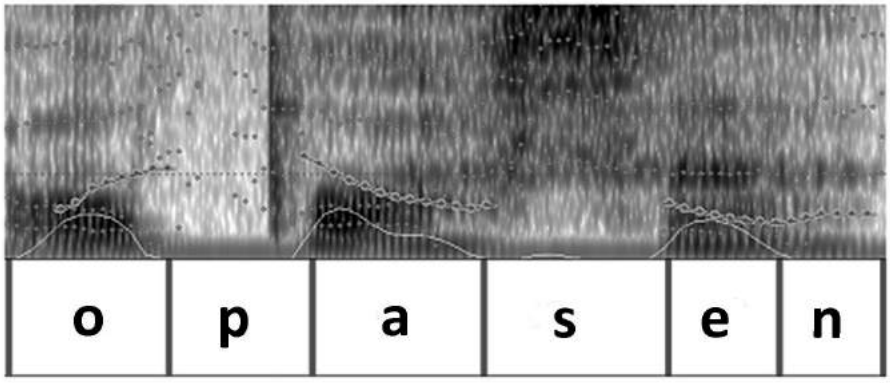

Figure 4. Spectrogram and segmentation of the word "oppassen". The top panel shows the spectrogram; the bottom panel shows the tier in which the segmentations and phone labels are specified. Observe that (in contrast to the Dutch spelling with two 'p'-graphemes) there is only one acoustic realization of the /p/-sound in "oppassen". This sound is characterized by a silence portion followed by short a burst (the vertical dark bar of the spectrogram at ¾ of the duration of /p/).

1

Speech material

In this thesis, the following speech sounds are used:

vowels /a, i, u/ (and the speaker's vowel space)
velar speech sounds /k, x/
stop consonants /b, d, p, t/

Vowels are selected for investigation in this thesis because they provide information about the shape of the vocal tract and are – compared to most consonants– relatively easy to identify in the speech signal. The identity of a vowel like sound (or its spectral color) is characterized by acoustic correlates and is primarily determined by its formants. Broadly speaking, the first formant frequency (F1) is associated with 'height', that is, the degree of opening of the vocal tract, whereas the second formant frequency (F2) is associated with the anterior–posterior tongue position.[57] Together, F1 and F2 provide a reasonable characterization (albeit far from complete) of the timbre of a vowel sound. By plotting the cardinal vowels (i.e. /a, i, u/) onto a graphical F1–F2 representation we obtain the vowel space (more specifically the vowel space). The vertices of the vowel space represent the most extended positions, corresponding to clearly pronounced realisations. The area of the vowel space is a measure for the amount of reduction in the vowel system and can formally be measured in terms of Erb^2, $Bark^2$ or Hz^2.[58] Vowel formant analyses have frequently been used in earlier studies to assess speech quality of HNC patients objectively. Correlates were found between formant values and intelligibility. Patients have a more deviant /i/ than control speakers and a low second formant of this vowel leads to a lower score on subjective intelligibility.[27,50,59] Vowels are also used to determine the presence of nasality in speech. Especially /i/ is useful because of the constriction of tongue to the velum, facilitating air to escape through the nasal cavity in case of inappropriate velar closure.[60–62]

In addition to vowels the velar consonants /k/ and /x/ are selected for investigation because earlier research revealed that patients treated for oral or oropharyngeal cancer often have difficulties with the production of these velar speech sounds.[17, 63] The oropharyngeal area is of influence in building up oral pressure, which is necessary in the production of stop consonants

such as /k/ (and /p, t/). In patients with oropharyngeal tumour this process could be impeded.[64, 65]

Finally, stop consonants /b, d, p, t/ are selected for investigation. In the general population, speakers are usually able to correctly produce speech characteristics such as voicing, silence, building and releasing of air pressure, to build vowels, fricatives, stop consonants, and many more – at different speaking rates. A specific speech characteristic that influences intelligibility and speech quality is voice-onset-time in stop consonants. VOT is defined as the duration that passes between when a stop consonant is released and when voicing, the vibration of the vocal folds, begins. The motivation to focus on voicing is based on medical knowledge in this particular domain of this type of patients: voicing is among the most seriously affected speech characteristics of this type of pathological speech.[66, 67] It is seen that patients treated for an oropharyngeal tumour tend to nasalize alveolar speech sounds. The speech sound /d/ then sounds more like /n/.[17] It is possible that this event affects the amount of voicing during the onset and burst of a stop consonant. It is known that patients treated for HNC experience difficulty with producing the correct amount of nasality and their speech often sounds hypernasal. The articulatory feature 'nasal' is measured on the entire stretch of speech by the Artificial Neural Network.

In this dissertation a number of speech sounds is selected from a 60-second speech recording. The advantage of investigating specific speech sounds is that more detailed information on production and flaws on production caused by HNC and treatment become clear. Selected speech sounds belong to various classes (i.e. vowels, velars and stop consonants) resulting in more in depth understanding of the behaviour of those classes related to subjective assessment of intelligibility, articulation, nasality and patient-reported outcome. The drawback of choosing for individual speech sounds instead of for the entire speech recording is that information on other speech sounds than the selected classes will be lost. However, in certain chapters the speech feature nasalance as measured with ANN is used to investigate the entire stretch of speech. In chapter 6 all 28 speech features are used. The number of selected speech sounds is rather limited but common in medical research.

1

In table 3 words containing the target speech sounds are summarized. In table 4, all speech sounds and characteristics of the two objective methods are depicted.

Table 3. Target speech sounds as taken from of words in the standard text.

	Appearance 1	Appearance 2
/a/	/m a n/	/ w a t ə r/
/i/	/w i/	/d i/
/u/	/m u t/	/t u n/
/x/	/o g ə n/	/g a t/
/k/	/k o m t/	/ij s k ou d ə/
/b/	/b ə h e k s t/	/b o v ə n/
/p/	/o p a s ə n/	/r i m p ə l/
/d/	/d i/	/z o n d ə r/
/t/	/z i t/	/d a t/

Table 4. Overview of objective speech parameters.

	Acoustic-Phonetic Analyses	Artificial Neural Network
/a, i, u/	Formant 1 Formant 2 vowel space	Feature 'nasal'
/x/	Spectral slope	–
/k/	Burst percentage	–
/b, d, p, t/	Duration VOT Duration burst	Feature 'voicing'
Entire text	–	Feature 'nasal All 28 features

Aim of this dissertation

Previous studies have shown that speech quality of HNC patients often deteriorates due to HNC itself and its treatment. A multidimensional speech assessment protocol is often recommended to assess speech outcome in (future) studies on innovative treatment or speech rehabilitation options. Such a multidimensional speech assessment protocol usually includes subjective and objective speech assessment methods and patient-reported speech outcome. In this dissertation we formulated two objectives in the expectation to contribute to the development of a multidimensional speech evaluation protocol.

The first objective was to investigate the feasibility and validity of two objective assessment methods: acoustic-phonetic analysis and articulatory feature analysis by an artificial neural network. These objective methods will be compared to evaluation of articulation, intelligibility and nasality by trained raters and to patient-reported speech outcome.

The second objective of this study was to investigate the contribution of these objective measuring methods to the development of a multidimensional speech evaluation protocol.

Outline of this dissertation

In order to investigate the feasibility of objective speech assessment methods several pilot studies were performed that reflect the process of the first stage of the development of a multidimensional speech protocol. In chapters two, three and four these pilot studies are described, investigating acoustic-phonetic speech analyses of vowels and velar speech sounds (chapter 2), artificial neural network (Articulatory Feature) analysis assessing hypernasality (chapter 3) and acoustic-phonetic and artificial neural network feature analysis assessing stop consonants (chapter 4). In chapter 5 all articulatory speech features of the artificial neural network are investigated. External validation of the findings is carried out by analyzing all previously investigated speech sounds in a larger study cohort (chapter 6). In chapter 7 the associations between specific voice and swallowing parameters are investigated.

1

In the final chapter the results obtained in above described studies are discussed and placed into broader perspective. Recommendations for further research are provided. Finally, English and Dutch summaries are provided. [1]

[1] In most of the following chapters, a patient cohort of 51 patients was used. The segmentation of the speech sounds was done only once and was not repeated for each chapter. This means that parts of those chapters (i.e. Patients and Methods) are identical throughout the dissertation.
Please notice that repetition of information is found throughout the dissertation. This is due to the individually published studies that this dissertation is composed of. Minor changes were made to published texts to enhance clarity.

Reference List

1. CBO NWHHT. Richtlijn Mondholte- en Orofarynxcarcinoom. Nationwide Guideline Oral Cavity and Oropharyngeal Cancer, version 1.4. (in Dutch). 2004.
2. Takes RP, Rinaldo A, Glasg RR, et al. Future of the TNM classification and staging system in head and neck cancer. Head Neck 2010;1693-711.
3. Nederlandse Werkgroep Hoofd-Halstumoren. Richtlijn Mondholte- en Orofarynxcarcinoom. 2004. Aphen aan den Rijn, Van Zuiden Communications.
4. Stell PM, Bowdler DA. Surgery for Head and Neck Cancer. In: Snow GB, Clark JR, editors. Multimodality Therapy for Head and Neck Cancer.New York: Thieme Medical Publishers, Inc.; 1996. p. 23-40.
5. Epstein JB, Emerton S, Kolbinson DA, et al. Quality of life and oral function following radiotherapy for head and neck cancer. Head Neck 1999;(21):1-11.
6. Chambers MS, Garden AS, Kies MS, et al. Radiation-induced xerostomia in patients with head and neck cancer: pathogenesis, impact on quality of life, and management. Head Neck 2004;26:796-807.
7. Clinical Medicine. 6th edition ed. Edinburgh: Elsevier Saunders; 2005.
8. Pignon J-P, A.le Maître, J.Bourhis. Meta-Analyses of Chemotherapy in Head and Neck Cancer (MACH-NC): An Update. International Journal of Radiation Oncology 2007;69(2):S112-S114.
9. www.iknl.nl [computer program]. Integraal Kankercentrum Nederland; 2012.
10. Bjordal K, A.de Graeff, P.Fayers, et al. A 12 country field study of the EORTC QLQ-C30 (version 3.0) and the head and neck cancer specific module (EORTC QLQ-H&N35) in head and neck patients. European Journal of Cancer 2010;36(14):1796-807.
11. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 3;85(5):365-76.
12. Karnell LH, Funk GF, Hoffman HT. Assessing head and neck cancer patient outcome domains. Head Neck 2000 Jan;22(1):6-11.
13. Bjordal K, Hammerlid E, Ahlner-Elmqvist M, et al. Quality of life in head and neck cancer patients: validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35. J Clin Oncol 1999 Mar;17(3):1008-19.
14. Borggreven PA, Aaronson NK, Verdonck-de Leeuw IM, et al. Quality of life after surgical treatment for oral and oropharyngeal cancer: a prospective longitudinal assessment of patients reconstructed by a microvascular flap. Oral Oncol 2007 Nov;43(10):1034-42.

15. Bjordal K, Mastekaasa A, Saasa S. Self-reported satisfaction with life and physical health in long-term cancer survivors and a matched control group. Oral Oncol 1995;(31B):340-5.
16. Borggreven PA, Verdonck-de Leeuw I, Langendijk JA, et al. Speech outcome after surgical treatment for oral and oropharyngeal cancer: a longitudinal assessment of patients reconstructed by a microvascular flap. Head Neck 2005 Sep;27(9):785-93.
17. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. Folia Phoniatr Logop 2009;61(3):180-7.
18. Bressmann T, Sader R, Whitehill TL. Consonant intelligibility and tongue motility in patients with partial glossectomy. Journal of Oral and Maxillofacial Surgery 2004;(62):298-303.
19. Furia C, Kowalski L, Latorre M. Speech intelligibility after glossectomy and speech rehabilitation. Arch Otolaryngol Head Neck Surg 2001;127:877-83.
20. Bodil IK, Lind MG, Aronson AE. Free radial forearme flap reconstruction in surgery of the oral cavity and pharynx: surgical compllication, impairment of speech and swallowing. Clin Otolaryngol Allied Sci 1994;19:28-34.
21. Seikaly H, Rieger J, Wu YN, et al. Functional outcomes after primary oropharyngeal cancer resection and reconstruction with the radial forearm free flap. Laryngoscope 2003;(113):897-904.
22. Knuuttila H, Pukander J, Maatta T, et al. Speech articulation after subtotal glossectomy and reconstruction with a myocutaneous flap. Acta Otolaryngol 1999;(119):621-6.
23. Michi KI, Imai S, Yamashita Y. Improvement of speech intelligibility by a secondary operation to mobilize he tongue after glossectomy. J Craniofac Surg 1989;17:162-6.
24. Rinkel RN, Verdonck-de Leeuw I, van Reij EJ, et al. Speech Handicap Index in patients with oral and pharyngeal cancer: better understanding of patients' complaints. Head Neck 2008 Jul;30(7):868-74.
25. Windrich M, Maier A, Kohler R, et al. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. Folia Phoniatr Logop 2008;60(3):151-6.
26. Lee AS, Ciocca V, Whitehill TL. Acoustic correlates of hypernasality. Clin Linguist Phon 2003 Jun;17(4-5):259-64.
27. Francis AL, Ciocca V, Ching Yu JM. Accuracy and variability of acoustic measures of voicing onset. J Acoust Soc Am 2003;2(113):1025-32.
28. King S, Taylor P. Detection of phonological features in continuous speech using neural networks. Comp Speech Lang 2000;14(4):333-53.
29. Kirchhoff K, Fink GA, Sagerer G. Combining acoustic and articulatory feature information for robust speech recognition. Speech Communication 2002;37:303-19.

30. Frankel J, Cetin O, Morgan N. Transfer Learning for Tandem ASR Feature Extraction. 2007 p. 227–36.
31. Frankel J, Wester M, King S. Articulatory feature recognition using dynamic Bayesian networks. ComputerSpeech & Language 2007;21(4):620–40.
32. Frankel J, Magimai-Doss M, King S, et al. Articulatory Feature Classifiers Trained on 2000 hours of Telephone Speech. 2007.
33. Scharenborg O, Wan V, Moore RK. Towards Capturing Fine Phonetic Variation in Speech using Articulatory Features. Speech Communication 2007;49:811–26.
34. Livescu K. Articulatory Feature-based Methods for Acoustic and Audio-Visual Speech Recognition: 2006 JHU Summer Workshop Final Report. 2006.
35. Parveen S, Green P. Multitask Learning in Connectionist Robust ASR using Recurrent Neural Networks. 2003 p. 1813–6.
36. Middag C, Saeys Y, Martens J-P. Towards an ASR-Free Objective Analysis of Pathological Speech. Proceedings of Interspeech 2010;294–7.
37. MacGregor RJ. Theoretical mechanics of biological neural networks. San Diego, CA, USA: Academic Press Professional, Inc.; 1993.
38. Aleksander I, Morton H. An introduction to neural computing. London, Boston, Melbourne: Intl Thomson Computer Pr (T); 1995.
39. Kohonen T. An introduction to neural computing. Neural Networks 1988;1(1):3–16.
40. Lippmann RP. An introduction to computing with neural nets. IEEE ASSP Magazine 1987;3(4):4–22.
41. Nabil N, Espy-Wilson CY. A signal representation of speech based on phonetic features. 1995 May 22; Inst. of Tech., Utica/Rome 1995 p. 310–5.
42. Nabil N, Espy-Wilson CY. A knowledge-based signal representation for speech recognition. Atlanta, Georgia 1996 p. 29–32.
43. Deng L, Sun DX. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J Acoust Soc Am 1994;(95):2702–19.
44. Erler K, Freeman GH. An HMM-based speech recognizer using overlapping articulatory features. J Acoust Soc Am 1996;100(4):2500–13.
45. Chomsky N, Halle M. The sound pattern of English. MIT Press; 1968.
46. Bocklet T, Haderlein T, Hönig F, et al. Evaluation and Assessment of Speech Intelligibility on Pathologic Voices based upon Acoustic Speaker Models. 2009 p. 89–92.
47. http://nico.nikkostrom.com/ [computer program]. KTH, Stockholm: 1997.
48. van Son RJJH, Binnenpoorte D, van den Heuvel H, et al. The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database. Aalborg 2001 p. 2051–4.
49. Whitehill TL, Ciocca V, Chan JC, et al. Acoustic analysis of vowels following glossectomy. Clin Linguist Phon 2006 Apr;20(2–3):135–40.

50. Kazi R, Prasad VM, Kanagalingam J, et al. Analysis of formant frequencies in patients with oral or oropharyngeal cancers treated by glossectomy. Int J Lang Commun Disord 2007 Sep;42(5):521-32.
51. Zimmerman A, Sader R, Hoole P, et al. The influence of oral cavity tumour treatment of the voice quality and on fundamental frequency. Clin Linguist Phon 2003;5(30):428-37.
52. Yoshida H, Furuya Y, Shimodaira K, et al. Spectral characteristics of hypernasality in maxillectomy patients. J Oral Rehabil 2000 Aug;27(8):723-30.
53. Schuster M, Haderlein T, Noth E, et al. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. Eur Arch Otorhinolaryngol 2006 Feb;263(2):188-93.
54. Praat: doing phonetics by computer [Computer program]. [computer program]. Version Version 5.2.35. University of Amsterdam: 2007.
55. Martens J, Versnel H, Dejonckere Ph. The effect of visible speech in the perceptual rating of pathological voices. Arch Otolaryngol Head Neck Surg 2007;133:178-85.
56. Kent RD. Intelligibility in speech disorders: theory, measurement, and management. Amsterdam, Philadelphia: John Benjamins Publishing; 1992.
57. van Bergem D. On the perception of acoustic and lexical vowel reduction. Berlin 1993 p. 677-80.
58. Sumita YI, Ozawa S, Mukohyama H, et al. Digital acoustic analysis of five vowels in maxillectomy patients. J Oral Rehabil 2002 Jul;29(7):649-56.
59. Fant GMG. Nasal sounds and nasalization. In: Fant GMG, editor. Acoustic theory of speech production.Hague: Mouton; 1970. p. 148-61.
60. Kataoka R, Michi KI. Spectral properties and quantitative evaluation of hypernasality in vowels. Cleft-palate-craniofacial Journal 1996;33(1):43-50.
61. Lee AS, Ciocca V, Whitehill TL. Acoustic correlates of hypernasality. Clin Linguist Phon 2003 Jun;17(4-5):259-64.
62. Markkanen-Leppanen M, Isotalo E, Makitie AA, et al. Speech aerodynamics and nasalance in oral cancer patients treated with microvascular transfers. J Craniofac Surg 2005 Nov;16(6):990-5.
63. Rieger J, Zalmanowitz JG, Li SY, et al. Speech outcomes after soft palate reconstruction with the soft palate insufficiency
repair procedure. Head Neck 2008;30:1439-44.
64. Markkanen-Leppanen M, Isotalo E, Makitie AA, et al. Changes in articulatory proWciency following microvascular reconstruction in oral or oropharyngeal cancer. Oral Oncol 2006;42:646-52.
65. Allen JS, J.L.Miller, D.DeSteno. Individual talker differences in voice-onset-time. J Acoust Soc Am 2003;113(1):544-52.
66. Klatt DH. Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters. Journal of Speech and Hearing Research 1975;18:686-706.

Chapter 2

2

Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer

Marieke J. de Bruijn
Louis ten Bosch
Dirk J. Kuik
Hugo Quené
Johannes A. Langendijk
C. René Leemans
Irma M. Verdonck- de Leeuw

Folia Phoniatrica et Logopaedica, 61 (3), 180-187. (2009)

2

Abstract

Speech impairment often occurs in patients after treatment for head and neck cancer. New treatment modalities such as surgical reconstruction or (chemo)radiation techniques aim at sparing anatomical structures that are correlated with speech and swallowing. In randomized trials investigating efficacy of various treatment modalities or speech rehabilitation, objective speech analysis techniques may add to improve speech outcome assessment. The goal of the present study is to investigate the role of objective acoustic -phonetic analyses in a multidimensional speech assessment protocol. Speech recordings of 51 patients (6 months after reconstructive surgery and postoperative radiotherapy for oral or oropharyngeal cancer) and of 18 control speakers were subjectively evaluated regarding intelligibility, nasal resonance, articulation, and patient-reported speech outcome (speech subscale of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Head and Neck 35 module). Acoustic-phonetic analyses were performed to calculate formant values of the vowels /a, i, u/, vowel space, air pressure release of /k/ and spectral slope of /x/. Intelligibility, articulation, and nasal resonance were best predicted by vowel space and /k/. Within patients, /k/ and /x/ differentiated tumour site and stage. Various objective speech parameters were related to speech problems as reported by patients. Objective acoustic-phonetic analysis of speech of patients is feasible and contributes to further development of a speech assessment protocol.

Introduction

Tumours in the oral cavity and oropharynx may result in damage of various anatomical structures by tumour extension and treatment. Patients often report a decreased use of tongue and perioral muscles and speech organs, such as the lips, tongue and velum, which frequently causes speech difficulty and other problems such as those related to social activities. These problems may ultimately have a negative impact on health-related quality of life.[1] Health-related quality of life significantly deteriorates during the first 6 months after completion of treatment, and may ameliorate by 12 months after treatment. Functionality of the head and neck area frequently remains below pre-treatment level.[2] Speech quality after treatment appears to be highly dependent on tumour size and site.[3-9] Patients who underwent treatment of larger tumours experienced more difficulty with speech than those with smaller tumours. Speech outcome after treatment for an oral tumour often results in articulation difficulties due to tissue loss, and structure alteration of various speech organs, while problems with speech production of patients treated for oropharyngeal cancer often include nasal resonance problems due to velopharyngeal inadequacy.

In the past decades, surgical possibilities of replacing damaged tissues in the oral cavity and oropharynx by different flaps have increased aiming to prevent speech and swallowing impairment. The preferred method of reconstruction of larger defects in the oral cavity or oropharynx is by free flaps. Free fasciocutaneous flaps are thin and pliable and are suitable for reconstruction of dynamic structures, such as the tongue and pharynx.[3-6]
More recently, organ preservation protocols such as chemoradiation are introduced also aiming at prevention of functional impairment. However, a recent literature review reveals that both treatment modalities, reconstructive surgery and organ preservation, still often result in speech and swallowing impairment.[10] New radiation delivery techniques aiming at sparing anatomical structures that are correlated with speech and swallowing may contribute to prevent long-term radiation-induced functional impairment as may speech rehabilitation. Also, new speech rehabilitation approaches such as logopedic exercises in an early stage before or during radiotherapy may improve functional outcome. However, prospective randomized trials are needed to provide evidence-based effectiveness of these approaches. Objective speech analysis techniques may add to improve speech evaluation protocols and enable adequate speech outcome assessment in clinical trials.

Speech quality is most often assessed via subjective evaluation by listeners. Results obtained from subjective assessments reveal correlations between tumour stage, intelligibility and articulation: patients with a smaller tumour (T2) have better intelligibility and articulation than patients with larger tumours (T3–T4). Nasal resonance and articulation of patients are significantly worse than in healthy individuals.[9] Nasal resonance in patients treated for tumours in the oropharynx appears to be worse than in patients with oral tumours. This difference is due to the oropharyngeal area that is involved in the partition between the oral and nasal cavity. In case of failing velar closure, air escapes through the nose, which results in hypernasal characteristics of speech.[11]

Objective measurements of speech quality are less often performed. Acoustic-phonetic analysis of the speech signal appeared to differentiate between healthy speakers and glossectomy patients.[12] Acoustic-phonetic analyses also revealed that patients who underwent partial resection of the tongue have deviant formant values for vowels, especially for /i/.[12, 13] A study using a nasometer revealed that speech of patients after reconstruction with large flaps had worse nasal resonance scores.[5] They also reported that patients with resections of more than half of the soft palate had more nasal resonance than patients with smaller resections of the soft palate.

The aim of this study is to obtain more insight in phonetic-acoustic speech characteristics of patients after microvascular reconstructive surgery for oral or oropharyngeal cancer regarding formant values of the vowels /a, i, u/, and the velar consonants /k/ and /x/. The second aim is to investigate the validity of objective phonetic-acoustic speech parameters. The results contribute to further development of a multidimensional speech assessment protocol that can be used in future prospective trials on efficacy of various treatment modalities and rehabilitation for head and neck cancer.

Patients and Methods

Patients

Patients underwent treatment for advanced oral or oropharyngeal squamous cell carcinoma with microvascular soft tissue transfer for the reconstruction of surgical defects. Surgery consisted of composite resections including excision of the primary tumour with en bloc ipsilateral or bilateral neck dissection. In case of oropharyngeal carcinomas a paramedian mandibular swing approach was used. Defects were reconstructed by a microvascular fasciocutaneous flap; no flap failures were observed. Patients received postoperative radiotherapy in case of advanced (T3-T4) tumours, positive or close surgical margins, multiple lymph node metastases and extra nodal spread. The primary site received a dose of 56-66 Gy in total (2 Gy per fraction, 5 times per week), depending on surgical margins. The nodal areas received a total of 46-66 Gy (2 Gy per fraction, 5 times a week). Exclusion criteria were inability to participate in functional tests, difficulty communicating in Dutch and age above 75 years. Fifty-one patients between 23 and 73 years (mean: 53.8 years, SD: 8.7 years) were included in the study after obtaining written informed consent, as well as 18 gender- and age-matched controls (table 5).

Speech Assessment

Patients (6 months after treatment) and controls read aloud a text with an approximate length of 60 s. The distance between lips and microphone (Sennheiser MKE 212 to 213) was 30 cm. Speech recordings were conducted in a soundproof cabin. For each speaker the recording level was adjusted to optimize signal-to-noise ratio. The recorded speech was digitized with Cool Edit PRO 1.2 (Adobe Systems Incorporated, San Jose, Calif., USA), with a 22-kHz sample frequency and 16-bit resolution.

Subjective Speech Evaluation

Perceptual evaluation of speech quality comprised ratings on intelligibility, articulation and nasal resonance by two speech pathologists. To enable subjective speech evaluation, a computer program was developed to perform blinded randomized listening experiments and score intelligibility, nasality, and articulation. Intelligibility was scored using a 10-point scale, where 1 represents the worst score and 10 represents the best score and 6 is just sufficient. Articulation and nasal resonance were judged using a 4-point

scale, ranging from normal to increasingly deviant speech quality. Interrater agreement for subjective assessment of intelligibility ranged from 40 to 90%. Intrarater agreement for repeated speech fragments of articulation and nasal resonance was high, with 100% equal scores between the ratings. Patient-reported speech outcome was assessed by the speech subscale (including 3 items) of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Head and Neck 35 module. The scores were linearly transformed to a scale of 0-100, with a higher score indicating a higher level of speech problems.[14]

Acoustic-Phonetic Analyses

In the present study, the vowels /a, i, u/ (the cardinal vowels in Dutch) and velar consonants were used as study material. Vowels are -compared to consonants- relatively easy to identify in the speech signal, and easier to analyze acoustically.

Table 5. Overview of gender, tumour site and stage of 51 patients included in the study.

	N	%
Gender		
Male	28	(55)
Female	23	(45)
Tumour site		
Oral cavity	21	(41)
Oropharynx	30	(59)
T-classification		
2	26	(51)
3-4	25	(49)

Vowel formant analyses proved to be valid measures of speech quality in patients with deviant speech originating from oral cancer or other origins in earlier studies.[15, 16] Vowel identity (or its spectral color) is characterized by acoustic correlates and is primarily determined by its formants. Broadly speaking, the first formant frequency (F1) is associated with 'height', that is,

the degree of opening of the vocal tract, whereas the second formant frequency (F2) is associated with the anterior-posterior tongue position.[17] Plotting the vowels /a, i, u/ onto a graphical F1-F2 representation shows the vowel space (more specifically the vowel space). The vertices of the vowel space represent the most extended positions. The area of the vowel space is a measure for the amount of reduction in the vowel system and can (formally) be measured in terms of Hz^2 (see also figure 5).[18]

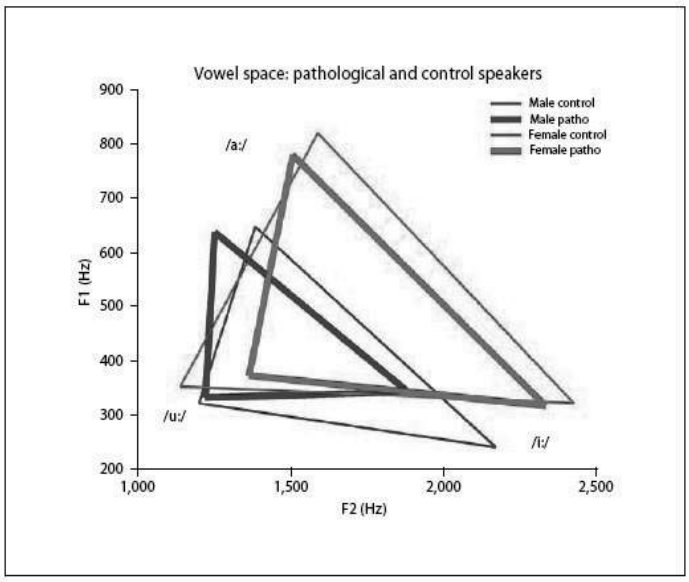

Figure 5. Vowel space of male (blue) and female (pink) patients (fat lines) and of controls (thin lines).

In addition to vowels, the velar consonants /k/ and /x/ were acoustic-phonetically analyzed, because earlier research revealed that patients with an oral or oropharyngeal tumour often have difficulties with the production of velar speech sounds. Speech raters often mistook /k/ for /x/.[9, 11] For /k/ the duration of air pressure release (the so-called plosive) as a percentage of the total duration (short silent period of pressure building + the pressure release) was measured and used as outcome measure. For /x/ the spectral

slope was used as outcome measure because /x/ is poor of harmonics due to a steep slope, so the slope (decreasing amplitude with increasing frequency) was used as characteristic of /x/. The overall spectral shape contains information on the characteristic of the fricative.[19]

For each selected speech sound (/a, i, u, k, x/), two acoustic realizations were segmented from running speech and were acoustic-phonetically analyzed using the speech processing software Praat version 4.0.28.[20] Since the acoustic realization of certain speech sounds may depend on its context, we took different phonological contexts around the target speech sounds into account, in order to improve generalization. A spectrogram functioned as a visual representation of the speech signal, which facilitated recognition of phonemes in the speech signal and facilitated precise extraction of phonemes from running speech. Spectral and acoustic speech analyses were automatically performed using scripts.[20]

Statistical Analysis

Validity of objective speech analyses was tested by means of univariate Pearson correlation coefficients between the subjective speech evaluations of intelligibility, articulation and nasal resonance and objective parameters (formants of the vowels /a, i, u/, size of the vowel space, spectral slope of /x/ and duration of pressure release of /k/). To obtain insight into the role of objective parameters in predicting subjective speech evaluation, multivariate regression analyses were performed. For intelligibility and self-assessments by patients, a linear regression was used, while for articulation and nasal resonance, logistic regression was performed on a binary scale [normal (score 0) vs. deviant (scores 1-3)]. Mann-Whitney tests were performed instead of t tests due to skewed data and were used to determine the validity of the objective speech parameters regarding known group differences: patients versus controls, smaller (T2) versus larger (T3-T4) tumours, and tumour location (oral vs. oropharyngeal).

Results

The two formants of two realisations of each vowel were averaged because inspection of formant values of the two realisations of one vowel revealed that there were no significant differences. For the velar speech sounds /k/ and /x/, however, larger differences were found which made using the average inappropriate. Therefore, /k/1 and /k/2 and /x/1 and /x/2 are analyzed separately and described in the results.

Objective versus Subjective Speech Assessment

Univariate correlations between subjective (self-)evaluations and objective parameters reveal that ratings on intelligibility and articulation are significantly related to objective analyses of /k/, the second formant of /i/, and formant space (table 4).

Table 4. Pearson correlations between objective speech parameters and subjective parameters Intelligibility, Articulation and Nasal Resonance (* $p<.05$).

	Intelligibility	Articulation	Nasal Resonance	EORTC Speech Scale
	r	r	r	r
/x/2	.12	.13	.33*	-.02
/k/1	.50*	.40*	.25*	-.27
/k/2	.36*	.25*	.39*	-.13
/i/F1	-.23	-.19	-.42*	.02
/i/F2	.35*	.36*	.13	-.24
/u/F1	-.11	-.11	-.37*	-.12
vowel space (Hz^2)	.39*	.42*	.15	-.20

To obtain insight into which objective parameters predict subjective (self-)assessments, multiple regression analyses were performed (tables 5-8). The results reveal that /k/, F1 of /i/, and the size of the vowel space predicted best subjective (self-)evaluations. These results reveal adequate validity of objective speech analyses. Especially /k/, /i/, /x/ and the size of the vowel space contribute to a prediction of subjective evaluation by objective speech parameters.

Table 5. Prediction of intelligibility by acoustic-phonetic parameters. (* $p<.05$). $R^2=45\%$.

	Intelligibility	
	B	**t**
/k/1	.038	3.42*
size Δ (Hz²)	7.07	4.11*
/i/F1	-.13	-2.60*

Table 6. Prediction of articulation by acoustic-phonetic parameters. (* $p<.05$). $R^2= 74\%$.)

	Articulation	
	B	**Wald**
/k/1	.11	8.53*
/a/ F1	-.20	6.52*
/a/ F2	.01	4.59*
/i/F1	-.06	8.31*
Vowel space (Hz²)	42.87	9.23*

Table 7. Prediction of nasal Resonance by acoustic-phonetic parameters. (* p<.05). R^2= 52%.

	Nasal Resonance	
	B	**Wald**
/x/2	.20	9.36*
/k/2	.05	7.32*
/i/F1	-.03	7.67*

Table 8. Subjective and objective speech parameters of speech quality that are related to speech problems in daily life as reported by patients. (* p<.05). R^2= 45,4 %.

	EORTC H&N-35 Speech Scale	
	B	**t**
/x/1	1.11	2.29*
/i/F2	-.04	-2.12*

Known Group Differences

To obtain insight into the predictive validity of objective speech analyses, Mann-Whitney tests were performed regarding known group differences: patients versus controls, and within the group of patients regarding tumour classification and tumour site (table 9). Significant differences between patients and controls in acoustic-phonetic parameters revealed that patients have a shorter pressure release for /k/ than controls. Patients have a higher F1 of /i/, but a lower F2 of /i/ than controls. The size of the vowel space is significantly smaller for patients than for controls. Acoustic-phonetic analysis also differentiated regarding tumour stage. Patients with smaller tumours had a longer pressure release compared to patients with a larger tumour. Regarding tumour site, /x/ distinguished between tumour location: patients with an oropharyngeal tumour had a steeper spectral slope than patients with an oral tumour.

Table 9. Significant differences between objective acoustic-phonetic variables measured on vowels (formant values in Hz, vowel space in Hz^2) and consonants (duration of air pressure release (the so-called plosive) as a percentage of the total duration (short silent period of pressure building + the pressure release) of /k/; spectral slope for /x/) between pathological and control speakers, and regarding tumour site and tumour classification, as obtained with a Mann-Whitney test.

	Pathological vs. control speakers		**Patients**		**Controls**	
	Z	p	% / Hz	sd	% / Hz	sd
/k/1	-2.77	.006	28,5%	16	43,5%	18
/k/2	-4.15	<.001	26,4%	19	49,8%	17
F1 /i/	-2.36	.018	334 Hz	54	296 Hz	49
F2 /i/	-2.42	.016	2105 Hz	363	2325 Hz	248
size Δ	-2.42	.015	.143 Hz^2	.12	.213 Hz^2	.11
	Oral tumour vs. oropharyngeal tumour		**Patients**		**Controls**	
/x/	-2.24	.025	-13 Hz	6	-17 Hz	6
	T2 tumour vs. T3-4 tumour		**T2**		**T3-4**	
/k/	-2.09	.037	33%	17	23%	14

Discussion

This study presents an inventory of speech performance 6 months after treatment in a well-defined head and neck cancer patient group after reconstructive surgery and radiotherapy for advanced oral or oropharyngeal cancer. Speech quality was determined with objective acoustic-phonetic analyses and commonly used subjective (self-) evaluations.

The first aim of the present study was to investigate which objective parameters contribute to the prediction of subjective (self-)evaluations of speech. Especially acoustic-phonetic parameters of /k/, /x/, /i/, and the size of the vowel space predicted best subjective assessment of overall intelligibility, articulation, nasal resonance and self-evaluation of speech. The result regarding /k/ is also reported[9], where listeners often judged /k/ as /x/. Production of velar consonants such as /k/ and /x/ require a posterior move of the tongue towards the oropharyngeal region and an adequate motility of the velum. Larger tongue motility corresponds with better intelligibility of consonants, including /k/.[21] No previous studies report on the speech sound /x/, which may be due to the absence of /x/ in other modern Western languages except for Dutch and a few dialects like Scottish.
The size of the vowel space was also found to be a predictor of subjective speech evaluations. The smaller size of the vowel space in patients was caused by the higher F1 and lower F2 of the vowel /i/. These results are in agreement with earlier research, where it was shown that a smaller size of the vowel space – that was also caused by deviant values of F1 and F2 of /i/ – was related to worse intelligibility in glossectomy patients.[12] In the present study, the vowel /i/ itself also proved to predict subjective evaluations: patients had a higher F1 and a lower F2. These results are in agreement with the results of earlier research on pathological speech[13, 18] (both concerned maxillectomy patients), but are not in agreement with results on research concerning partial glossectomy, where it was found that only gender and complication after surgery were of influence on altered F1values.[22]

The second aim of this study was to investigate differences regarding acoustic-phonetic speech characteristics between patients and controls and within the group of patients regarding tumour site and tumour classification. Between patients and controls, pressure release of /k/, F1 and F2 of /i/, and the size of the vowel space differentiated best. Difficulty with production of

/k/ originates from velar function difficulties. The decreased size of the vowel space of patients was mainly caused by deviant formant values of /i/ and is in accordance with earlier studies.[12, 13, 18] Inadequate movement of the tongue regarding height (F1) and anterior–posterior movement (F2) may result in distorted vowels. Acoustic–phonetic analysis also revealed differences between patients regarding tumour stage (/k/) and tumour site (/x/): patients with smaller tumours had a longer pressure release of /k/ compared to patients with a larger tumour. Regarding tumour site, patients with an oropharyngeal tumour had a steeper spectral slope in /x/ than patients with an oral tumour. Due to tumour growth and treatment in the oropharyngeal area, patients with oropharyngeal cancer are likely to experience more difficulty with the production of velar speech sounds. Like /k/, /x/ is also a velar consonant, which appears to be problematic for this patient population. These results are in agreement with earlier research.[9, 21, 23]

The results concerning differentiation between groups can be explained by structure alterations of the vocal tract after tumour involvement and treatment. Patients have more difficulty with proper velar closure, resulting in distorted velar speech sounds. Difficulty regarding production of vowels is also attributable to alterations caused by tumour growth and treatment. Especially patients who underwent treatment involving the tongue may experience more difficulty with the production of vowels. In previous studies, vowels of patients treated for head and neck cancer were considered deviant from vowels produced by healthy individuals: F2 of all vowels was lowered compared to controls, and F1 of /i/ was elevated compared to controls.[12, 18]

In the present study, the velar consonants /k, x/ and the vowels /a, i, u/ were selected from words that were obtained from running speech. The phonological context of the selected speech sounds may be of influence on the perception hereof and could also be of influence on the results obtained in the present study. Further research into speech quality of patients with head and neck cancer could be performed on different speech sounds in order to detect more characteristics of speech quality and more details of specific speech sounds. Also, a different approach to objectively measure the speech quality could be an analysis of speech features present in speech such as nasality or voicing. Such a complex task of calculating speech features could be performed via automatic speech recognition using a neural

network trained in identification of speech features.[24-26] This approach might give additional insight into the speech of patients treated for head and neck cancer. In the present study, the results are based on postoperative data only and no attempts were made to compare these data with preoperative speech. Future research may focus on post- versus preoperative speech quality in order to obtain more insight into sensorimotor adaptation capabilities of patients to compensate for alterations in the vocal tract after treatment.[27, 28]

Conclusion

Speech quality of patients after treatment of an oral or oropharyngeal tumour was investigated. Acoustic-phonetic analyses proved to be valid and are suitable for measuring speech quality of patients. The presented results contribute to further development of a speech analysis protocol to be used in clinical practice and in clinical trials aiming at improving speech outcome in patients with head and neck cancer.

Acknowledgements

The authors wish to thank Li Ying Chao, Pepijn Borggreven and Milou Heiligers for their contributions regarding speech recordings and/or analyses.

2

Reference List

1. Karnell LH, Funk GF, Hoffman HT. Assessing head and neck cancer patient outcome domains. Head Neck 2000 Jan;22(1):6-11.
2. Borggreven PA, Aaronson NK, Verdonck-de Leeuw IM, et al. Quality of life after surgical treatment for oral and oropharyngeal cancer: a prospective longitudinal assessment of patients reconstructed by a microvascular flap. Oral Oncol 2007 Nov;43(10):1034-42.
3. Hara I, Gellrich N, Duker J, et al. Swallowing and speech function after intraoral soft tissue reconstruction with lateral upper arm free flap and radial forearm free flap. Br J Oral Maxillofac Surg 2003;41:161-9.
4. Michi KI, Imai S, Yamashita Y. Improvement of speech intelligibility by a secondary operation to mobilize the tongue after glossectomy. J Craniofac Surg 1989;17:162-6.
5. Seikaly H, Rieger J, Wu YN, et al. Functional outcomes after primary oropharyngeal cancer resection and reconstruction with the radial forearm free flap. Laryngoscope 2003;(113):897-904.
6. Su WF, Hsia YJ, Chang YC, et al. Functional comparison after reconstruction with a radial forearm free flap or a pectoralis major flap for cancer of the tongue. Otolaryngol Head Neck Surg 2003 Mar;128(3):412-8.
7. Pauloski BR, Rademaker AW, Logemann JA, et al. Relationship between swallow motility disorders on videofluorography and oral intake in patients treated for head and neck cancer with radiotherapy with or without chemotherapy. Head Neck 2006 Dec;28(12):1069-76.
8. Furia C, Kowalski L, Latorre M. Speech intelligibility after glossectomy and speech rehabilitation. Arch Otolaryngol Head Neck Surg 2001;127:877-83.
9. Borggreven PA, Verdonck-de Leeuw I, Langendijk JA, et al. Speech outcome after surgical treatment for oral and oropharyngeal cancer: a longitudinal assessment of patients reconstructed by a microvascular flap. Head Neck 2005 Sep;27(9):785-93.
10. van der Molen L, van Rossum MA, Burkhead LM, et al. Functional outcomes and rehabilitation strategies in patients treated with chemoradiotherapy for advanced head and neck cancer: a systematic review. Eur Arch Otorhinolaryngol 2009 Jun;266(6):901-2.
11. Markkanen-Leppanen M, Isotalo E, Makitie AA, et al. Speech aerodynamics and nasalance in oral cancer patients treated with microvascular transfers. J Craniofac Surg 2005 Nov;16(6):990-5.
12. Whitehill TL, Ciocca V, Chan JC, et al. Acoustic analysis of vowels following glossectomy. Clin Linguist Phon 2006 Apr;20(2-3):135-40.
13. Yoshida H, Furuya Y, Shimodaira K, et al. Spectral characteristics of hypernasality in maxillectomy patients. J Oral Rehabil 2000 Aug;27(8):723-30.

14. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 3;85(5):365-76.
15. Lee AS, Ciocca V, Whitehill TL. Acoustic correlates of hypernasality. Clin Linguist Phon 2003 Jun;17(4-5):259-64.
16. Sumita YI, Ozawa S, Mukohyama H, et al. Digital acoustic analysis of five vowels in maxillectomy patients. J Oral Rehabil 2002 Jul;29(7):649-56.
17. Kent RD. Intelligibility in speech disorders: theory, measurement, and management. Amsterdam, Philadelphia: John Benjamins Publishing; 1992.
18. van Bergem D. On the perception of acoustic and lexical vowel reduction. Berlin 1993 p. 677-80.
19. Jesus LMT, Shadle CH. Acoustic analysis of European Portuguese uvular [x,] and voiceless tapped alveolar [] fricatives. Journal of the International Phonetic Association 2005;35(1):27-44.
20. Praat: doing phonetics by computer. [computer program]. Version 5.2.35. University of Amsterdam: 2007.
21. Bressmann T, Sader R, Whitehill TL. Consonant intelligibility and tongue motility in patients with partial glossectomy. Journal of Oral and Maxillofacial Surgery 2004;(62):298-303.
22. Kazi R, Prasad VM, Kanagalingam J, et al. Analysis of formant frequencies in patients with oral or oropharyngeal cancers treated by glossectomy. Int J Lang Commun Disord 2007 Sep;42(5):521-32.
23. Terai H, Shimahara M. Evaluation of speech intelligibility after a secondary dehiscence operation using an artificial graft in patients with speech disorders after partial glossectomy. Br J Oral Maxillofac Surg 2004 Jun;42(3):190-4.
24. Haderlein T, Riedhammer K, Noth E, et al. Application of automatic speech recognition to quantitative assessment of tracheoesophageal speech with different signal quality. Folia Phoniatr Logop 2009;61(1):12-7.
25. Wielgat R, Zielinski TP, Wozniak T, et al. Automatic recognition of pathological phoneme production. Folia Phoniatr Logop 2008;60:323-31.
26. Windrich M, Maier A, Kohler R, et al. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. Folia Phoniatr Logop 2008;60(3):151-6.
27. Savariaux C, Perrier P, Pape D, et al. Speech production after glossectomy and reconstructive lingual surgery:
a longitudinal study. 2001.
28. Houde J, Jordan M. Sensomotor adaptation in speech production. Science 1998;(279):1213-6.

Chapter 3

3

Artificial neural network analysis to assess hypernasality in patients treated for oral or oropharyngeal cancer

Marieke J. de Bruijn
Louis ten Bosch
Dirk J. Kuik
Johannes A. Langendijk
C. René Leemans
Irma M. Verdonck- de Leeuw

Logopedics Phoniatrics Vocology, 36, 4, 168-174 (2011)

3

Abstract

Investigation of applicability of neural network feature analysis of nasalance in speech to assess hypernasality in speech of patients treated for oral or oropharyngeal cancer was the goal of the study. Speech recordings of 51 patients and of 18 control speakers were evaluated regarding hypernasality, articulation, intelligibility and patient reported speech outcome. Feature analysis of nasalance was performed on /a/, /i/ and /u/ and on the entire stretch of speech. Nasalance distinguished significantly between patients versus controls. Nasalance in /a/ and /i/ predicted best intelligibility, nasalance in /a/ predicted best articulation, nasalance in /i/ and /u/ predicted best hypernasality. Feature analysis of nasalance in oral or oropharyngeal cancer patients is feasible; prediction of subjective parameters varies between moderate to poor.

Introduction

Worldwide, each year approximately 500.000 patients are diagnosed with head and neck cancer (HNC). Medical treatment of these patients includes (a combination of) surgery, radiotherapy, and chemotherapy with cure rates varying from 50% to 95%. Due to the removal or damage of tissues that are critical for voice and speech, a large proportion of 'cured' patients (up to 40-70% dependent of tumour stage, location, and treatment modality) experience voice or speech problems that can seriously degrade the 'quality of life' for the patients and their relatives.[1-5] Problems related to voice and speech function can be minimized in two ways. Firstly, in planning surgery and (chemo-) radiation, HNC oncologists may take the (long-term) effects of their interventions on speech production into account. Secondly, speech therapy may be used to reduce or repair part of the detrimental effects that could not be avoided. However, both approaches are hampered by the fact that there are no tools for the objective description of speech quality that allow prediction of the effects of specific interventions, nor to guide therapy programs to focus on the most promising methods.

Tumour growth and -treatment may cause alterations of the vocal tract. For instance, after radiotherapy, a patient may experience stiffening of tissue in the oral cavity due to fibrosis and swelling. This may mean less flexibility of the muscles of the oral cavity and finally, a deteriorated speech quality. From earlier research, one can conclude that speech of HNC patients is characterized (among other features) by hypernasality leading to deteriorated speech intelligibility. In clinical practice, speech quality is mainly assessed by perceptual evaluations made by professionals such as speech therapists or evaluated by patients themselves using standardized questionnaires. Elaborate attempts have been made to devise perceptual rating instruments for voice and speech quality that also can be used to evaluate voice and speech quality of patients treated for head and neck cancer.[1, 2, 6-11] Invariably, the conclusion has been that without elaborate training it is only possible to obtain acceptably high agreement rates on very general characteristics of the speech (such as tempo, loudness, overall pitch) and that raters trained in different groups attach different interpretations to more specific scales, such as nasality, hoarseness, breathiness, etc.[12, 13] Existing instruments and techniques for describing the perceptual quality of pathological speech are not even powerful and accurate enough to allow wide sharing of data and

information between groups of therapists and surgeons working in different hospitals. Therefore, when it comes to understanding the impact of treatment of head and neck cancer on voice and speech quality, the importance of objective quality assessment tools becomes even more relevant. For this purpose, it is not enough to understand the causal relations between spectral and temporal features of the signal and perceptual quality, it is also necessary to understand the relations between pathologies of the speech production apparatus and the resulting speech signals.[1, 5, 7, 14-18]

Trained raters are able to differentiate between patients versus controls, and within patients regarding tumour location.[8] In the study by Borggreven et al. [8] subjective judgments were able to divide patients from controls. Within patients, subjective judgments divided between tumour size (T2 vs. T3-4) but not between tumour sub sites.

Objective classifier-based feature analysis may be able to contribute to the assessment of speech quality of patients after treatment for oral or oropharyngeal cancer. Hereto, we refer to algorithms that are able to estimate values for articulatory-inspired features, such as 'nasal', 'plosive' and 'vowel' directly from the acoustic speech signal. Objective speech analysis of healthy speech of subjects from the general population reveal that scores of 86% correctly identified phonemes and correct classification for healthy and pathological speech by the use of an artificial neural network can be established.[19, 20] Previous research evaluating speech quality after treatment for head and neck cancer via a neural network is scarce and no studies focusing on the feature nasalance in this population were found in the literature. Results of evaluation of other features in speech of patients with head and neck cancer were found in a population combined with other etiologies. Schuster et al. (2006)[21] reported that automatic speech recognition techniques seem to be a good means to objectify and quantify global speech outcome of laryngectomees. Haderlein et al. (2009)[22] also investigated speech of laryngectomees and concluded that automatic speech recognition can be used for objective intelligibility ratings with results comparable to those of human experts. Windrich et al.[23] demonstrated that automatic speech recognition yielded mean word recognition rate of 49% in oral cancer patients and of 76% in controls. Automatic evaluation scores were highly correlated with the experts' perceptual evaluation of intelligibility.

The aim of the present study is to investigate if artificial neural network analyses of the feature nasalance as obtained objectively from the speech signal can distinguish between speech of patients versus controls, and within patients regarding tumour stage (stage 2 versus T3-4) and tumour location (oral versus oropharyngeal cancer).

Patients and methods

Patients

Fifty-one patients between 23 and 73 years (mean: 53.8 years, sd: 8.7 years, table 10) were included in the study after written informed consent. Eighteen gender- and age-matched controls were included. Patients underwent surgery and radiotherapy for advanced oral or oropharyngeal squamous cell carcinoma with microvascular soft tissue transfer for the reconstruction of surgical defects. Defects were all successfully reconstructed by a microvascular fasciocutaneous forearm flap. Almost all (91 %) patients received postoperative radiotherapy. Exclusion criteria were difficulty communicating in Dutch and age above 75 years.

Table 10. Overview of gender, tumour site and stage of 51 patients included in the study.

	N	%
Gender		
Male	28	(55)
Female	23	(45)
Tumour site		
Oral cavity	21	(41)
Oropharynx	30	(59)
T-classification		
2	26	(51)
3-4	25	(49)

Speech recordings and dataset

Patients (6 months after treatment) and controls read aloud a standard Dutch text with an approximate length of 60 seconds. The distance between lips and microphone (Sennheiser MKE 212-3) was 30 centimeters. Speech recordings were conducted in a sound-attenuated booth. For each speaker the recording level was adjusted to optimize signal-to-noise ratio. The recorded speech was digitized with Cool Edit PRO 1.2 (Adobe Systems Incorporated, San Jose, CA, USA), with 22-kHz sample frequency and 16-bit resolution mono (one channel).

For each speaker, two realizations of the vowels /a/, /i/ and /u/ were selected from the 1-minute recording and subsequently analyzed. Vowels are -compared to consonants - relatively easy to identify in the speech signal. Phonetically, vowels are produced without frication or occlusion, giving a strong signal with little noise. This is necessary for analyzing hypernasality in the signal. Two realizations of each vowel yield a larger generalization taking into account different phonological contexts, consisting of stop consonants (/d, t/), liquid consonant (/w/) and nasal consonants (/m, n/).

Subjective speech evaluations

Perceptual evaluation of speech quality comprised ratings of intelligibility, articulation and hypernasality by two speech pathologists on the entire stretch of running speech. To enable subjective speech evaluation, a computer program was developed to perform blinded randomized listening experiments and to store intelligibility, articulation and hypernasality scores in a database. The panel of the two trained listeners rated articulation and hypernasality on a 4-point scale, ranging from normal to increasingly deviant speech quality. Intelligibility was scored using a 10-point scale, following the Dutch educational grading system where 1 represents the worst score and 10 represents the best score and 6 is just sufficient. Interrater agreement for subjective assessment of intelligibility ranged from 40% to 90%. Intrarater agreement for two repeated speech fragments of articulation and hypernasality was high with 100% equal scores between the ratings. Speech problems in daily life as reported by patients were assessed by the Speech Subscale (including 3 items) of the EORTC Quality of Life Questionnaire H&N35 module.[24, 25] The scores are linearly transformed to a scale of 0 to 100, with a higher score indicating a higher level of speech problems. A detailed description of these subjective speech ratings and results can be found in Borggreven et al. (2005).[8] Patient reported speech outcome was

assessed by the Speech Subscale (including 3 items) of the EORTC Quality of Life Questionnaire Head & Neck35 module.[24] The scores were linearly transformed to a scale of 0 to 100, with a higher score indicating a higher level of speech problems.

Artificial neural network and feature analysis

For the objective artificial neural network feature analyses of nasalance, the vowels to be analysed (/a, i, u/) were segmented by hand from running speech using the speech processing software Praat version 5.2.35.[26] A spectrogram functioned as a visual representation of the speech signal, which facilitated recognition of phonemes in the speech signal and facilitated precise extraction of phonemes from running speech. Next, for the nasalance estimation these vowel-segments were input for a specifically trained artificial neural network (see below). In the context of research on Automatic Speech Recognition (ASR), several speech decoding techniques have been developed that are able to automatically segment and label an unknown utterance in terms of phone-like segments.[27, 28] In the last decade however, new approaches were designed that circumvent the use of phone symbols in the segmentation of speech. Instead of using phonetic symbols (mostly predefined by the ASR developer), these methods focus on more basic properties of the speech signal that characterize the speech signal in a way more similar to acoustic features or phonological features such as 'manner' and 'place' of articulation.[29, 30] In the last decade, several techniques have become available, such as Artificial Neural Nets (ANNs) and Support Vector Machines.[20, 31, 32]

In this paper we use ANNs to obtain representation in terms of articulatory features of an input speech signal. ANNs contain a number of model parameters (weights of the connections between network nodes) that determine the relation between input and output of these ANNs. The parameters can be trained on an independent training set in which input and desired output of the ANN are specified for each training data point. The ANNs used in this paper have an input context of 7 input MFCC frames (three preceding target and three following frames) which -during training- are used in combination with a canonical value (0 or 1) of the phonological feature that is being modeled by the ANN. The ANNs have been trained on speech from speakers without reported articulatory problems and -during training- provide estimations of phonological features from the acoustic

input signal. The ANNs that were used in the present study were trained on the basis of speech originating from the corpus of the Institute of Phonetic Sciences, University of Amsterdam, the Netherlands (IFA Spoken Language Corpus).[33] This corpus contains speech in a variety of styles of four normal speaking males and four normal speaking females in combination with accompanying hand-labelled articulatory-inspired features. For the training of ANN used in the present study, however, speech spectra of one healthy male and one healthy female in combination with accompanying segmentations were derived from the IFA-corpus. During the training, several normalisation steps (MFCC mean and variance, energy) were applied, to optimize the generalization of the trained ANNs. ANNs were specifically trained on phone-labelled speech spectra for all phonological features. However, in the present study we only use one phonological feature: nasalance (in the sequel, the resulting ANN is referred to as ANN-nasalance).[34]

Training takes place via error-backpropagation, one of the well-known and commonly used training procedures for ANNs. After training, ANN-nasalance was tested on speech of two other speakers from this IFA corpus. The output of ANN-nasalance varies between 0 (absent) and 1 (present). A value of 0.8 for instance means that nasalance is rather strongly present in that specific speech frame of 0.01 seconds. Depending on the amount of input and of consistency in labels during training, ANN-nasalance achieves high levels of correct classification. In the quality assessment of ANNs, the performance is given in terms of frame accuracy. Performance during testing was 80 % correct at frame level. An accuracy of 80 % means that the classifier correctly classifies the feature value assigned to this frame for 80 % of all frames in the evaluation test set. The degree of the phonological feature nasalance as identified by the artificial neural network was determined for the vowels /a/, /i/ and /u/.

In the experiments described here, ANNs were used to estimate the degree of various phonological features such as manner, place, voicing, front-back and rounding. Each of these properties are modeled by an ANN. Each ANN was modeled by a three layer feed-forward network: one input layer, one hidden layer, and one output layer. The input layer is fed with the MFCCs (mel-frequency cepstral coefficients) obtained from the MFCC extraction step. The units in the output layer represent the estimated values of the various options for that particular feature. For example, the manner feature is

modeled by the manner-ANN which has 6 units on its output layer (see also chapter 1, table 1):

Manner: 0-approx-fric-nasal-stop-vowel

These six units in the manner-ANN estimate the degree of NULL, approximant, fricative, nasality, stop, and vowel, for each 10ms frame in the input speech signal, respectively. The NULL value is a unit that takes positive values if the network is not able to positively assign values to any of the other five units. In total, the manner-ANN provides 6 values for each frame of 0.01 seconds. By taking into account the output of the other five ANNs (place, voicing, front-back, rounding, and static) and stacking all results, a 28-dimensional feature vector for each frame of 0.01 seconds is obtained. In the present study, we particularly focus on the phonological feature nasalance, that is, on specifically one out of this collection of 28 values per 0.01 seconds. For an overview of all 28 features, see chapter 1, table 1. One of the most important model parameters in an ANN is the dimension (number of hidden units) of the hidden layers. In this study, we adopted a setting that has been suggested in the literature and that showed good results in training and test sessions with speech from healthy controls.[20]

The ANNs that were used in this study are available as public domain software.[35] The motivation to use ANNs instead of Support Vector Machines (SVM) is determined by the facts that compared to SVM, ANNs are more elegant and deliver a relatively small model. Moreover, ANNs use continuous mapping and the model does not encounter discrete selections during the classification task, as is the case with SVM.
The goal of this approach is to objectively measure the range of deviation in nasalance between healthy speakers and patients. However, since ANN is trained on speech of (only two) healthy speakers, it should be noted that ANN identifies the amount of nasalance of speech of patients compared to the standardization as obtained by the training. It is not clear how ANN treats speech characteristics of patients that are more or less deviant from this standard.

Statistical analysis
To obtain insight into the role of objective parameters in predicting subjective speech evaluation, multivariate regression analyses were

performed. For intelligibility and self assessments by patients a linear regression was used (these outcome variables have a continuous range of values), while for articulation quality and hypernasality logistic regression was performed (these variables have a very limited range of possible values and the distribution over these values is skewed), on a recoded scale into binary partition.

In order to determine the construct validity of the objective speech parameters regarding known group differences, Mann-Whitney tests were used: patients versus controls, smaller (T2) versus larger tumours (T3-4) -ANN performing a binary classification task-, and tumour location (oral versus oropharyngeal). For all statistical tests a significance level of α=.05 was maintained. ANN-nasalance parameters include two realisations of each vowel /a/, /i/ and /u/ and nasalance on the entire stretch of speech.

Results

Intelligibility was predicted significantly by the amount of nasalance in the second realisation of /a/ and /i/ (R^2= 21.3%). Articulation quality was predicted by the amount of nasalance in the second realisation of /a/ (R^2= 48.7%) and hypernasality was predicted by the amount of nasalance in the first realisations of /i/ and /u/ (R^2= 24.9%). Although significant, reported R^2-values are rather weak and vary between moderate to poor. Patient reported speech outcome was not significantly predicted by any of the objective nasalance parameters (table 11).

Table 11. Results of multivariate regression analyses with hypernasality, articulation, intelligibility as dependent variables and objective feature analyses of the amount of nasalance in vowels and on the entire stretch of speech as predictors (*p<.05).

	Hypernasality R^2= 24.9%*		**Articulation** R^2= 48.7%*		**Intelligibility** R^2= 21.3%*	
	b	P	b	P	b	p
/a/1 ANN-nasal	-1.77	.761	-3.21	.827	.74	.884
/a/2 ANN-nasal	-8.60	.211	-101.08	.050*	-12.01	.030*
/i/1 ANN-nasal	15.12	048*	-3.76	.617	4.11	.405
/i/2 ANN-nasal	-2.51	.181	-12.92	.051	-3.73	.023*
/u/1 ANN-nasal	4.20	.008*	.50	.747	1.92	.081
/u/2 ANN-nasal	-2.57	.528	-3.72	.590	-.54	.863
Speech ANN-nasal	-24.07	.147	7.27	.752	-17.93	.202

Significant differences were found between patients and controls: nasalance in both realisations of /i/ and the second realisation of /a/ distinguished between patients and controls. Within the patient group, regarding tumour classification or tumour site no differences in nasalance were found (table 12). The amount of nasalance on the entire stretch of speech was not able to distinguish between study groups, nor was it involved in the prediction of subjective parameters.

Table 12. Significant differences between objective feature analysis measured on vowels between patients versus and control speakers, and within patients regarding tumour classification and tumour site, as tested with Mann–Whitney analysis. (*p<.05).

	Patients vs. Controls		Tumour classification		Tumour location	
	U	z	U	z	U	z
/a/1 ANN	329.5	−1.89	294.5	−.60	252.5	−1.25
/a/2 ANN	235	−3.30*	294.5	−.60	303	−.24
/i/1 ANN	300	−2.2*	315	−.19	301	−.27
/i/2 ANN	292	−2.31*	272	−1.01	251	−1.23
/u/1 ANN	421	−.52	289	−.69	263.5	−.99
/u/2 ANN	328.5	−1.11	255.5	−.44	274.5	−.56
speech ANN	372.5	−1.20	305.5	−.37	308	−.14

Discussion

Hypernasality is a common speech characteristic in oral and oropharyngeal cancer patients.[8] In the present study, ANN-nasalance trained on healthy speech was applied onto speech of patients treated for head and neck cancer, which is a new approach to analyze pathological speech. It appeared that this technique is feasible and valid in patients treated for oral or oropharyngeal cancer, but R^2-values and construct validity are low. Objective analysis of nasalance in the vowels /a/ and /i/ distinguished patients from control speakers. Furthermore, nasalance in the vowels /a/ and /i/ predicted intelligibility, nasalance in the vowel /a/ predicted articulation quality, and nasalance in the vowels /i/ and /u/ predicted hypernasality as evaluated by speech pathologists. The amount of nasalance on the vowel /i/ was a predictor of subjective evaluated intelligibility and hypernasality. Also, the amount of nasalance on /i/ was significantly different between patients and controls. However, the strengths of correlations were, although statistically significant, poor to moderate.

Previous research concerning objective acoustic-phonetic analysis (spectral analysis) of speech revealed that patients treated for head and neck cancer have more hypernasality than controls. Especially the vowel /i/ seemed to be a predictor for hypernasality. Acoustic-phonetic analysis showed that the first formant of /i/ was higher and the second formant of /i/ was lower in speech that was subjectively judged as nasal. No publications on /a/ and /u/ in combination with hypernasality were found in literature.[9, 15-17, 24] In the study by Borggreven et al.[8] perceptual speech evaluations differentiated patients from controls. Within patients, subjective judgments divided between Tumour size (T2 vs. T3-4) but not for Tumour sub sites. In the present study objective research with ANN was conducted on the same patient cohort. The amount of nasalance as assessed by objective ANN on /a/ and /i/ (but not /u/) differentiates speech produced by patients and controls. Objective measurement with ANN was not able to differentiate between patients for tumour stage and tumour location, while trained listeners are able to differentiate between patients for tumour stage, but not for tumour location.[8] Concluding, ANN is not yet able to substitute trained raters.

There are some limitations to this study that should be acknowledged. Firstly, patient reported speech problems in daily life were not predicted significantly by objective ANN analysis of hypernasality. Apparently, other factors than hypernasality alone account for speech problems as experienced by patients themselves. These factors may include other speech characteristics such as the production of velar consonants[2, 4] and also coping strategies in dealing with speech problems after cancer treatment. Secondly, the amount of variance explained regarding the prediction of perceptual speech characteristics was moderate to poor and ranged between 21% and 49%. Thirdly, no differences were detected regarding known group differences as tumour classification (T2 vs. T3-4) or tumour location (oral vs. oropharyngeal). An explanation may be that the technique of artificial neural network feature analysis is not sensitive enough (yet) to detect more subtle differences within patient groups. Or, perhaps, ANN is able to detect differences between patients, but these results are outweighed when trying to correlate them into fixed scales, such as the scale of tumour classification T1-2 versus T3-4. Further research into other speech features may provide more insight into the aetiology and physiology of hypernasality or otherwise deteriorated speech quality such as spectral characteristics[4, 15-17, 36] or voice quality and intra-oral sensibility.[18]

Furthermore, the patient numbers in the present study are relatively large for research of pathological quality of speech when taking into account the workload to collect a homogeneous group of patients 6 months post-treatment (taking mortality into account) and recording, segmenting, and analyzing the speech samples. However, the number of patients may have lead to underpowered study results: instable predictive values may be caused by a too small number of patients compared to the large amount of variables. Future studies will include larger patient cohorts. Instead of using more data from each patient, the use of a second, larger patient cohort is favoured. In the second cohort, the use of the identical standardized text is recommended. The use of a second patient cohort will correct for coincidences in demographical arrangements such as age, sex, tumour stadium and tumour location, as well as patient specifics, such as the ability to find a strategy to improve intelligibility by a better articulation or by a lower speaking rate.

The goal of this approach is to objectively measure the range of deviation in nasalance between healthy speakers and patients. However, since ANN is trained on speech of only two healthy speakers, it should be noted that it is not clear how the ANN treats speech characteristics of patients that might be very different from the healthy speech in the training set. When it comes to the amount of speech features identified by ANN and the rather poor amount of explained variance and low construct validity, in the present study only one feature was measured: nasalance. In contrast, listeners use all of the features present in the speech signal to come to their judgments on suitability of intelligibility, articulation and hypernasality. In future research, multiple features measured with ANNs could be taken into account and be combined in order to obtain an applicable tool for objective speech evaluation.

The vowels in the present study were taken from running speech. This means that speech sounds were not produced in isolation, but that multiple factors had an influence on production. Such influences could be speaking rate, phonological context (coarticulation or assimilation), pattern of emphasis of syllables or simply a less fluent way of speaking because of difficulty with reading out loud. Further investigation could rule out these variables by analyzing speech sounds produced in isolation. Finally, nasalance on the entire stretch of speech appeared not contribute to the prediction of subjective ratings, nor to differentiation between speech of patients and controls, and within patients, between tumour stage and tumour location. Presumably, other (combined) speech features than nasalance alone are of greater importance in judgments of speech.

Conclusion

Objective artificial neural network feature analysis of nasalance in speech of patients treated for oral or oropharyngeal cancer is feasible and valid, but there are some limitations to the present study.
The amount of variance explained regarding the prediction of perceptual speech characteristics varies between moderate (articulation) to poor (hypernasality and intelligibility). There is no prediction for self evaluations by patients. The artificial neural network is able to differentiate between

speech of patients and control speakers, but not between tumour classification and tumour location.

At this moment, the artificial neural network is not able to substitute a trained rater. Further research is ongoing with larger study samples enabling more in depth analysis and external validation. Results contribute to further development of a multidimensional speech evaluation protocol to be used for research purposes and clinical practice.

Reference List

1. Verdonck-de Leeuw I, Hilgers FJ, Keus RB, et al. Multidimensional assessment of voice characteristics after radiotherapy for early glottic cancer. Laryngoscope 1999 Feb;109(2 Pt 1):241-8.
2. Borggreven PA, Aaronson NK, Verdonck-de Leeuw IM, et al. Quality of life after surgical treatment for oral and oropharyngeal cancer: a prospective longitudinal assessment of patients reconstructed by a microvascular flap. Oral Oncol 2007 Nov;43(10):1034-42.
3. Rinkel RN, Verdonck-de Leeuw I, van Reij EJ, et al. Speech Handicap Index in patients with oral and pharyngeal cancer: better understanding of patients' complaints. Head Neck 2008 Jul;30(7):868-74.
4. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. Folia Phoniatr Logop 2009;61(3):180-7.
5. van der Molen L, van Rossum MA, Burkhead LM, et al. Functional outcomes and rehabilitation strategies in patients treated with chemoradiotherapy for advanced head and neck cancer: a systematic review. Eur Arch Otorhinolaryngol 2009 Jun;266(6):901-2.
6. van As CJ, Koopmans-van Beinum FJ, Pols LC, et al. Perceptual evaluation of tracheoesophageal speech by naive and experienced judges through the use of semantic differential scales. J Speech Lang Hear Res 2003 Aug;46(4):947-59.
7. Lundstrom E, Hammarberg B, Munck-Wikland E, et al. The pharyngoesophageal segment in laryngectomees--videoradiographic, acoustic, and voice quality perceptual data. Logoped Phoniatr Vocol 2008;33(3):115-25.
8. Borggreven PA, Verdonck-de Leeuw I, Langendijk JA, et al. Speech outcome after surgical treatment for oral and oropharyngeal cancer: a longitudinal assessment of patients reconstructed by a microvascular flap. Head Neck 2005 Sep;27(9):785-93.
9. Kazi R, Prasad VM, Kanagalingam J, et al. Analysis of formant frequencies in patients with oral or oropharyngeal cancers treated by glossectomy. Int J Lang Commun Disord 2007 Sep;42(5):521-32.
10. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol 2001 Feb;258(2):77-82.
11. Kempster GB, Gerratt BR, Verdolini AK, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. Am J Speech Lang Pathol 2009 May;18(2):124-32.

12. Yamaguchi H, Shrivastav R, Andrews ML, et al. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. Folia Phoniatr Logop 2003 May;55(3):147-57.
13. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. J Acoust Soc Am 2007 Oct;122(4):2354-64.
14. McConnel FM, Pauloski BR, Logemann JA, et al. Functional results of primary closure vs flaps in oropharyngeal reconstruction: a prospective study of speech and swallowing. Arch Otolaryngol Head Neck Surg 1998 Jun;124(6):625-30.
15. Yoshida H, Furuya Y, Shimodaira K, et al. Spectral characteristics of hypernasality in maxillectomy patients. J Oral Rehabil 2000 Aug;27(8):723-30.
16. Sumita YI, Ozawa S, Mukohyama H, et al. Digital acoustic analysis of five vowels in maxillectomy patients. J Oral Rehabil 2002 Jul;29(7):649-56.
17. Lee AS, Ciocca V, Whitehill TL. Acoustic correlates of hypernasality. Clin Linguist Phon 2003 Jun;17(4-5):259-64.
18. Markkanen-Leppanen M, Isotalo E, Makitie AA, et al. Speech aerodynamics and nasalance in oral cancer patients treated with microvascular transfers. J Craniofac Surg 2005 Nov;16(6):990-5.
19. Wallen EJ, Hansen JHL. A screening test for speech pathology assessment using objective quality measures. 1996 Oct 3; Philadelphia, PA. 1996 p. 776-9.
20. King S, Taylor P. Detection of phonological features in continuous speech using neural networks. Comp Speech Lang 2000;14(4):333-53.
21. Schuster M, Haderlein T, Noth E, et al. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. Eur Arch Otorhinolaryngol 2006 Feb;263(2):188-93.
22. Haderlein T, Riedhammer K, Noth E, et al. Application of automatic speech recognition to quantitative assessment of tracheoesophageal speech with different signal quality. Folia Phoniatr Logop 2009;61(1):12-7.
23. Windrich M, Maier A, Kohler R, et al. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. Folia Phoniatr Logop 2008;60(3):151-6.
24. Bjordal K, Hammerlid E, Ahlner-Elmqvist M, et al. Quality of life in head and neck cancer patients: validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35. J Clin Oncol 1999 Mar;17(3):1008-19.
25. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 3;85(5):365-76.
26. Praat: doing phonetics by computer. [computer program]. Version 5.2.35. University of Amsterdam: 2007.

27. Nabil N, Espy-Wilson CY. A signal representation of speech based on phonetic features. 1995 May 22; Inst. of Tech., Utica/Rome 1995 p. 310-5.
28. Nabil N, Espy-Wilson CY. A knowledge-based signal representation for speech recognition. Atlanta, Georgia 1996 p. 29-32.
29. Deng L, Sun DX. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J Acoust Soc Am 1994;(95):2702-19.
30. Erler K, Freeman GH. An HMM-based speech recognizer using overlapping articulatory features. J Acoust Soc Am 1996;100(4):2500-13.
31. Robinson T, Hochberg M, Renals S. The use of recurrent neural networks in contnuous speech recognition. In: Lee C-H, Soong F, ., editors. Automatic Speech and Speaker Recognition- Advanced Topics. Kluwer Academic Publishers; 1996. p. 233-58.
32. Bridle J, Deng L, Picone J, et al. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. John Hopkins University 1998 p. 1-61.
33. van Son RJJH, Binnenpoorte D, van den Heuvel H, et al. The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database. Aalborg 2001 p. 2051-4.
34. Graupe D. Principles of Artificial Neural networks. Advanced series on circuits and systems - volume 6 ed. Singapore: World Scientific Publishing Company Co. Pte. Ltd.; 2007.
35. http://nico.nikkostrom.com/ [computer program]. KTH, Stockholm: 1997.
36. Whitehill TL, Ciocca V, Chan JC, et al. Acoustic analysis of vowels following glossectomy. Clin Linguist Phon 2006 Apr;20(2-3):135-40.

Chapter 4

Acoustic-phonetic and artificial neural network feature analysis to assess speech quality of stop consonants produced by patients treated for oral or oropharyngeal cancer

4

Marieke J. de Bruijn
Louis ten Bosch
Dirk J. Kuik
Birgit I. Witte
Johannes A. Langendijk
C. René Leemans
Irma M. Verdonck- de Leeuw

Speech Communication, 54, 5, 632-640 (2012)

4

Abstract

Speech impairment often occurs in patients after treatment for head and neck cancer. A specific speech characteristic that influences intelligibility and speech quality is voice-onset-time (VOT) in stop consonants. VOT is one of the functionally most relevant parameters that distinguish voiced and voiceless stops. The goal of the present study is to investigate the role and validity of acoustic-phonetic and artificial neural network analysis (ANN) of stop consonants in a multidimensional speech assessment protocol. Speech recordings of 51 patients 6 months after treatment for oral or oropharyngeal cancer and of 18 control speakers were evaluated by trained speech pathologists regarding intelligibility and articulation. Acoustic-phonetic analyses and artificial neural network analysis of the phonological feature voicing were performed in voiced /b/, /d/ and voiceless /p/ and /t/. Results revealed that objective acoustic-phonetic analysis and feature analysis for /b, d, p/ distinguish between patients and controls. Within patients /t, d/ distinguish for tumour location and tumour stage. Measurements of the phonological feature voicing in almost all consonants were significantly correlated with articulation and intelligibility, but not with self-evaluations. Overall, objective acoustic-phonetic and feature analyses of stop consonants are feasible and contribute to further development of a multidimensional speech quality assessment protocol.

Introduction

Every year, approximately 500.000 patients are diagnosed with head and neck cancer (HNC) worldwide. Medical treatment includes (a combination of) surgery, radiotherapy, and chemotherapy, with cure rates varying from 50% to 95%. Patients often experience a range of discomforts due to tumour growth and treatment. Inflexibility and deteriorated functionality of the head and neck structures often result in difficulty with swallowing and speech, and secondary in problems with social aspects of daily life as eating in public and communication, ultimately resulting in a lower quality of life.[1-3]

Earlier research on speech produced by patients treated for oral or oropharyngeal cancer showed that defects or loss of tissue mass, less flexibility of the tongue and facial muscles, stiffened tissue, nerve damage and velar incompetence are causes of deteriorated speech outcome.[4-20] Speech quality of patients after treatment for oral or oropharyngeal cancer appears to be highly dependent on tumour size and site. Patients who underwent treatment of larger tumours experience more difficulty with speech than those with smaller tumours. Speech outcome after treatment for an oral tumour often results in articulation difficulties due to tissue loss, structure alteration of various speech organs, while problems with speech production of patients treated for oropharyngeal cancer often include nasal resonance problems due to velopharyngeal inadequacy. In case of surgery for advanced tumours in the oral cavity or oropharynx, reconstructive surgery is often performed after treatment for better functionality of organs in the head and neck area. Reconstructive surgery uses thin skin flaps originating from the patients forearm to cover large defects located at dynamic structures. More recently, organ preservation protocols as chemo radiation are introduced also aiming at prevention of functional impairment. However, a recent literature review reveals that treatment modalities, reconstructive surgery and organ preservation, still often result in speech impairment.[21]

In the general population, speakers are usually able to correctly produce speech characteristics such as voicing, silence, building and releasing of air pressure, to build vowels, fricatives, stop consonants, and many more – at different speaking rates. Coordination of glottal activity and locomotion of articulatory muscles usually passes synchronously. A specific speech characteristic that influences intelligibility and speech quality is voice-onset-

time (VOT) in stop consonants. In normal speech, VOT is one of the functionally most relevant parameters that distinguishes voiced and voiceless stops and is a result of the temporal coordination of voicing and oral articulation gestures. VOT is defined as the length of time period that passes between when a stop consonant is released and when voicing, the vibration of the vocal folds, begins. For voiced consonant stops /b, d, g/, voicing starts before the burst of airflow. This voice lead in voiced stop consonants is typical for the Dutch language. For voiceless consonant stops /p, t, k/, a short period of silence precedes the burst (see figure 6). Voiced plosives /b, d, g/ usually have a shorter VOT than voiceless plosives /p, t, k/. VOT is significantly related to speaking rate.[22-26] Patients treated for oral and oropharyngeal cancer may have difficulty with adequate coordination of motor function of articulatory speech structures and vocal fold vibration. Building up oral pressure necessary for stop consonants in combination and synchronously with ceasing vocal fold vibration in case of the voiceless stop consonants may be especially problematic. For patients, it seems problematic to quickly stop the activity of the glottis so that it remains mute. This period of inactivity results in necessary short silent periods during speech production such as the silence during the pressure building period preceding the burst in production of voiceless stops. Because this action is difficult to perform for patients, it is therefore hypothesized that the duration of VOT preceding the burst in voiceless stops in patients is longer compared to controls and that the silence period preceding the burst in voiceless stops show more voicing in patients compared to controls. The motivation to focus on phonological feature value is based on medical knowledge in this particular domain of this type of patients: voicing is among the most seriously affected speech characteristics of this type of pathological speech.

The main objective of the present study is to investigate feasibility and validity of acoustic-phonetic analyses of duration of VOT and of the burst in stop consonants as produced by patients treated for oral or oropharyngeal cancer compared to control speakers. Also, an artificial neural network (ANN) objectively analyses the degree of the phonological feature voicing of VOT (the silent part before the burst) and of the burst in stop consonants. Known group differences will be tested: patients versus controls, and within patients: small versus large tumours, and tumours originating from the oral cavity versus the oropharynx. In addition, correlations between objective speech quality measures of stop consonants and subjective assessments of

intelligibility, articulation and patient reported speech outcome will be tested. The results of this study will contribute to further development of multidimensional speech assessment protocol for research purposes and clinical practice.

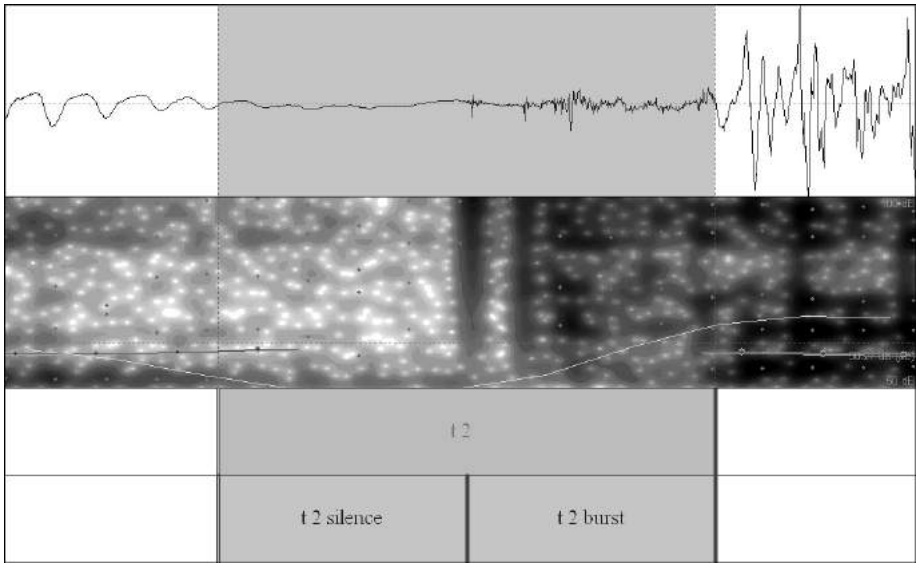

Figure 6. An example of the segmentation of /t/. The first tier shows the entire duration of /t/ ("t 2"); the second tier displays the partition in silence ("t 2 silence") and burst ("t 2 burst"). The spectrogram shows the relative absence of voicing during the pressure-building silence, followed by the outburst of voicing during the burst. This figure is made in Praat.

Methods

Speakers

Patients were treated by surgery and radiotherapy at the Department of Otolaryngology-Head & Neck Surgery of VU University Medical Center in Amsterdam, the Netherlands. In more detail, patients underwent composite resections for advanced oral or oropharyngeal squamous cell carcinoma with microvascular soft tissue transfer for the reconstruction of surgical defects. Surgery consisted of composite resections including excision of the primary tumour with en bloc ipsilateral or bilateral neck dissection. In case of

oropharyngeal carcinomas, a paramedian mandibular swing approach was used. All free flaps were successful. Patients received radiotherapy in case of advanced (T3-4) tumours, positive or close surgical margins, multiple lymph node metastases and/or extra nodal spread. The primary site received a dose of 56 to 66 Gy in total (2 Gy per fraction, 5 times per week), depending on surgical margins. The nodal areas received a total of 46, 56 or 66 Gy (2 Gy per fraction, 5 times a week) in case of N0, N+ without extranodal spread and N+ with extranodal spread, respectively. Exclusion criteria were incapability to participate in functional tests, difficulty communicating in Dutch and age above 75 years. Fifty-one patients between 23 and 73 years (mean: 53.8 years, SD: 8.7 years) were included in the study after written informed consent (table 13). Eighteen gender- and age matched controls were included.

Speech recordings

Patients (6 months after treatment) and controls read-aloud a standardized Dutch text. The distance between lips and microphone was 30 centimetres. Speech recordings were conducted in a sound attenuated booth. For each speaker the recording level was adjusted to optimize signal-to-noise ratio. The recorded speech was digitized with Cool Edit PRO 1.2 (Adobe Systems Incorporated, San Jose, CA, USA) with 22-kHz sample frequency and 16-bit resolution.

Table 13. Overview of gender, tumour site and stage of 51 patients included in the study.

	N	%
Gender		
Male	28	(55)
Female	23	(45)
Tumour site		
Oral cavity	21	(41)
Oropharynx	30	(59)
T-classification		
2	26	(51)
3-4	25	(49)

Subjective speech evaluation

Perceptual evaluation of speech quality comprised ratings on intelligibility and articulation by two speech pathologists on the entire stretch of running speech. To enable subjective speech evaluation, a computer program was developed to perform blinded randomized listening experiments and to store intelligibility and articulation scores in a database. Intelligibility was scored using a 10-point scale, following the Dutch educational grading system where 1 represents the worst score and 10 represents the best score and 6 is just sufficient. Articulation was judged using a 4-point scale, ranging from normal to increasingly deviant speech quality. Interrater agreement for subjective assessment of intelligibility ranged from 40% to 90%. Intrarater agreement for repeated speech fragments of articulation was high with 100% equal scores between the ratings. Speech problems in daily life as reported by patients was assessed by the Speech Subscale (including 3 items) of the EORTC Quality of Life Questionnaire H&N35 module.[3, 27] The scores are linearly transformed to a scale of 0 to 100, with a higher score indicating a higher level of speech problems.

Objective speech evaluation

Acoustic-phonetic analyses

Speaking rate of the entire stretch of speech was calculated in words per minute. The voiced stop consonants /b/ and /d/ and their voiceless counterparts /p/ en /t/ were used as speech material (in the Dutch phonological system, the voiced counterpart /g/ of the voiceless stop consonant /k/ is not present, and therefore these velar stop consonants were not investigated in the present study). For each selected speech sound (voiced /b, d/ and voiceless /p, t/), two realizations were segmented from running speech and were analysed using the speech processing software Praat (version 5.2.35).[28] Duration of VOT and the duration of the following release of air pressure, the so-called burst, were measured (see figure 6). Where this paper refers to the VOT, actually the pre-burst silence portion is meant.

Artificial neural network and feature analysis

In the context of research on Automatic Speech Recognition (ASR), several speech decoding techniques have been developed that are able to

automatically segment and label an unknown utterance in terms of phone-like segments.[29, 30] In the last decade however, new approaches were designed that circumvent the use of phone symbols in the segmentation of speech. Instead of using phonetic symbols (mostly predefined by the ASR developer), these methods focus on more basic properties of the speech signal that characterize the speech signal in a way more similar to acoustic features or phonological features such as 'manner' and 'place' of articulation.[31, 32] In the last decade, several techniques have become available, such as Artificial Neural Nets (ANNs) and Support Vector Machines.[33-35]

In this paper we use ANNs to obtain a feature representation of an input speech signal. ANNs contain a number of model parameters (weights of the connections between network nodes) that determine the relation between input and output of these ANNs. The parameters can be trained on a training set in which input and desired output of the ANN are specified for each training data point. The ANNs used in this paper have an input context of 7 input MFCC frames, which during training are used in combination with a canonical value (0 or 1) of the phonological feature that is being modeled by the ANN. The ANNs have been trained on speech from speakers without reported articulatory problems and in the test provide estimations of phonological features from the acoustic input signal. The ANNs that were used in the present study were trained on the basis of speech originating from the corpus of the Institute of Phonetic Sciences, University of Amsterdam, the Netherlands (IFA Spoken Language Corpus). This corpus contains speech in a variety of styles of four normal speaking males and four normal speaking females in combination with accompanying hand-labelled articulatory-inspired features.[36] For the training of ANN, speech spectra of one healthy male and one healthy female in combination with accompanying segmentations were derived from the IFA-corpus. The various normalisation steps applied during the training of the ANN classifier led to the best possible generalization of the ANNs trained. ANNs were specifically trained on labelled speech spectra for all phonological features. However, in the present study we only use one phonological feature: voicing (in the sequel, the resulting ANN is referred to as ANN-voicing).[37]

Training takes place via error-backpropagation, one of the well-known and commonly used training procedures for ANNs. After training, ANN-voicing

was tested on speech of two other speakers from this IFA corpus. The output of ANN-voicing varies between 0 (absent) and 1 (present). A value of 0.8 for instance means that voicing is rather strongly present in the speech frame. Depending on the amount of input and of consistency in labels during training, ANN-voicing achieves high levels of correct classification. In the quality assessment of ANNs, the performance is given in terms of frame accuracy. For voicing, performance during testing was 80% correct at frame level for independent test speakers. An accuracy of 80% means that the classifier correctly classifies the feature value assigned to this frame for 80% of all frames in the evaluation test set. The degree of the phonological feature voicing as identified by the artificial neural network was determined for VOT as well as for the following burst of the selected stop consonants.

In the experiments described here, ANNs were used to estimate the degree of various phonological features such as manner, place, voicing, front-back and rounding. Each of these properties are modeled by an ANN. Each ANN was modeled by a three layer feed-forward network: one input layer, one hidden layer, and one output layer. The input layer is fed with the MFCCs (mel-frequency cepstral coefficients) obtained from the MFCC extraction step. The units in the output layer represent the estimated values of the various options for that particular feature. For example, the manner feature is modeled by the manner-ANN which has 6 units on its output layer (see also chapter 1, table 1):

Manner: 0-approx-fric-nasal-stop-vowel

These six units in the manner-ANN estimate the degree of NULL, approximant, fricative, nasality, stop, and vowel, for each 10ms frame in the input speech signal, respectively. The NULL value is a unit that takes positive values if the network is not able to positively assign values to any of the other five units. In total, the manner-ANN provides 6 values for each frame of 10 ms. By taking into account the output of the other five ANNs (place, voicing, front-back, rounding, and static) and stacking all results, we obtain a 28-dimensional feature vector for each frame of 10 ms. In the present study, we particularly focus on the phonological feature voicing, that is, on specifically one out of this collection of 28 values per 10ms. One of the most important model parameters in an ANN is the dimension (number of hidden units) of the hidden layers. In this study, we adopted a setting that has been

suggested in the literature[33] and that showed good results in training and test sessions with speech from healthy controls.

The ANNs that were used in this study are available as public domain software.[38] The motivation to use ANNs instead of Support Vector Machines (SVM) is determined by the facts that compared to SVM, ANNs are more elegant and deliver a relatively small model. Moreover, ANNs use continue mapping and the model does not encounter discrete selections during the classification task, as is the case with SVM.

Statistical analysis

Mann-Whitney tests were performed to test differences between patients and controls, and within patients differences regarding tumour stage (smaller (T1-2)) versus larger (T3-4)) and tumour location (oral cavity versus oropharyngeal cavity). Mann-Whitney tests were performed instead of t-tests due to skewed data. Spearman correlation coefficients were used to test correlations between the subjective speech evaluations of articulation and intelligibility, patient reported outcome, and objective parameters (speaking rate, duration of VOT and burst, and ANN-voicing during VOT and burst).

Results

Predictive validity

No significant differences were found between any groups regarding speaking rate. Speaking rate in patients was 171 words per minute versus 175 words per minute in controls. Patients had significant longer VOT in the voiced stop consonants /b/ and /d/, and a shorter burst in the voiceless consonants /t/ and /p/ (table 14). Regarding ANN-voicing, patients had significant more voicing during VOT in the voiced consonant /d/ and voiceless consonant /p/, and more voicing during the burst in the voiced consonants /b/, and voiceless consonants /p/ and /t/ (table 15). However, findings were not always consistent in both realizations of the same consonant.

Table 14. Results of acoustic-phonetic analyses: mean and standard deviation of speaking rate (words per minute), and duration (seconds) of voice onset time (VOT) and burst of two realizations of the voiced stop consonants /b/ (b1 and b2) and /d/ (d1 and d2) and voiceless stop consonants /p/ (p1 and p2) and /t/ (t1 and t2) among patients (n=51) and controls (n=18). Significant differences between patients and controls are indicated with an asterisk. * $p \leq 0.05$, ** $p \leq 0.01$

	Patients	Controls
speaking rate	171 (33)	175 (23)
B1 VOT duration **	.020 (.019)	.008 (.004)
B2 VOT duration	.022 (.027)	.008 (.004)
B1 burst duration	.028 (.014)	.025 (.011)
B2 burst duration	.021 (.011)	.018 (.006)
D1 VOT duration	.006 (.006)	.007 (.005)
D2 VOT duration **	.020 (.020)	.008 (.004)
D1 burst duration	.017 (.014)	.020 (.012)
D2 burst duration *	.014 (.010)	.018 (.008)
P1 silence duration	.083 (.025)	.081 (.012)
P2 silence duration	.043 (.023)	.039 (.021)
P1 burst duration	.028 (.027)	.023 (.006)
P2 burst duration *	.020 (.011)	.028 (.012)
T1 silence duration	.030 (.025)	.027 (.009)
T2 silence duration	.033 (.026)	.033 (.011)
T1 burst duration	.034 (.023)	.035 (.014)
T2 burst duration	.029 (.025)	.031 (.010)

Table 15. Results of artificial neural network analyses of the phonological feature voicing (ANN-voicing): mean and standard deviation of voicing during voice onset time (VOT) and burst of two realizations of the voiced stop consonants /b/ (b1 and b2) and /d/ (d1 and d2) and voiceless stop consonants /p/ (p1 and p2) and /t/ (t1 and t2) among patients (n=51) and controls (n=18).A value of 0 represents absence of voicing; a value of 1 represents maximum amount of voicing. Significant differences between patients and controls are indicated with an asterisk. * $p \leq 0.05$, ** $p \leq 0.01$

	Patients	Controls
B1 ANN-voicing VOT	.68 (.31)	.61 (.28)
B2 ANN-voicing VOT	.82 (.22)	.79 (.18)
B1 ANN-voicing burst *	.84 (.24)	.73 (.23)
B2 ANN-voicing burst	.88 (.17)	.86 (.12)
D1 ANN-voicing VOT *	.67 (.39)	.50 (.34)
D2 ANN-voicing VOT *	.77 (.31)	.68 (.30)
D1 ANN-voicing burst	.68 (.36)	.63 (.30)
D2 ANN-voicing burst	.81 (.30)	.76 (.27)
P1 ANN-voicing silence *	.18 (.21)	.08 (.07)
P2 ANN-voicing silence	.39 (.30)	.26 (.27)
P1 ANN-voicing burst **	.45 (.31)	.19 (.17)
P2 ANN-voicing burst *	.52 (.30)	.33 (.27)
T1 ANN-voicing silence	.21 (.28)	.09 (.08)
T2 ANN-voicing silence	.28 (.28)	.14 (.08)
T1 ANN-voicing burst	.22 (.28)	.06 (.07)
T2 ANN-voicing burst *	.33 (.30)	.12 (.10)

Within patients, patients with larger tumours had significantly less voicing during VOT compared to patients with smaller tumours, (mean 0.77 (s.d. 0.31) versus mean 0.69 (s.d. 0.30)) and during the burst (mean 0.81 (s.d. 0.30) versus mean 0.76 (s.d. 0.27) in the voiced consonant /d/.

Regarding tumour location, patients with a tumour originating in the oral cavity had a shorter burst (mean .029 (s.d. .025)) in the voiceless consonant /t/ compared to patients treated for oropharyngeal cancer (mean .031 (s.d. .010)). See also figures 7 and 8.

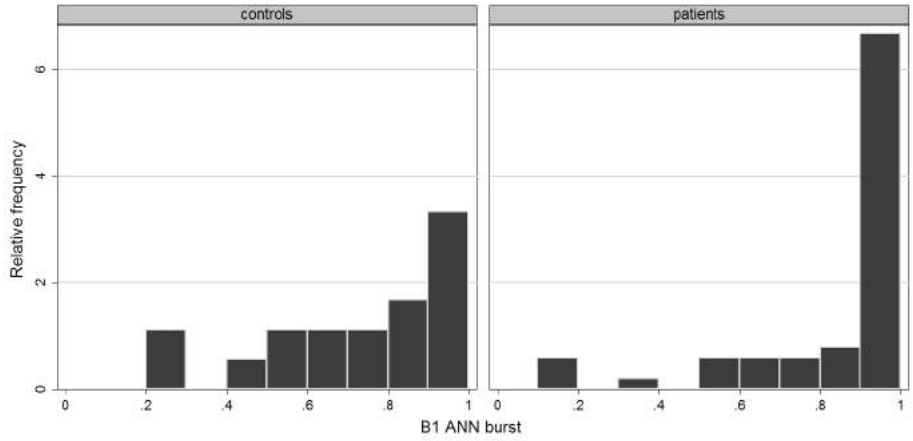

Figure 7. Relative frequency histograms of the dispersion of values of phonological feature voicing on the burst of /b/1 as identified by an Artificial Neural Network: controls and patients. The x-axis represents the amount of the feature voicing from 0 to 1 and the scale of the y-axis is in percentages.

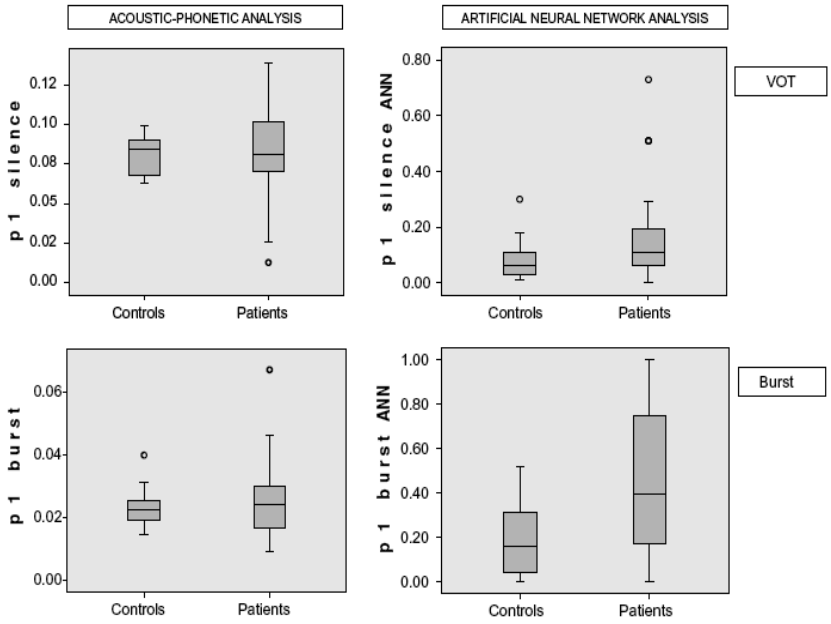

Figure 8. Median duration (in ms), its upper and lower quartiles and the minimum/maximum value within 1.5 times the inter-quartile range distance from the lower/upper quartile of pre-burst silence (VOT) and burst of /p/1 as measured by acoustic-phonetic analysis (left box plots). Median amount of voicing (ranging from 0 - 1), its upper and lower quartiles and the minimum/maximum value within 1.5 times the inter-quartile range distance from the lower/upper quartile of the amount of feature voicing of pre-burst silence (VOT) and burst of /p/1 by artificial neural network analysis (right box plots). Please note the differences in scaling for each individual box plot. The dots in the box plots represent outliers (i.e. more than 1.5 times the inter-quartile range away from the lower/upper quartile).

Concurrent validity

Spearman correlation analyses reveal that objective speech quality parameters are significantly related to subjective ratings of speech quality by speech pathologists, but not to subjective ratings of speech quality by patients themselves. Speaking rate and the phonological feature voicing (ANN-voicing) during VOT and burst of almost all consonants are significantly correlated to subjectively judged articulation and intelligibility (table 16). Correlations coefficients are moderate and vary from .25 to .35.

Table 16. Spearman correlation values between the objectively analyzed stop consonants and subjective judgments by two speech language therapists ("articulation" and "intelligibility") and by patients themselves ("EORTC QLQ H&N35 Speech Subscale"). The correlations between the two realizations of the voiced stop consonants /b/ (b1 and b2) and /d/ (d1 and d2) and voiceless stop consonants /p/ (p1 and p2) and /t/ (t1 and t2) and the subjective judgments of nasal resonance, articulation and intelligibility are displayed. * $p \leq 0.05$, ** $p \leq 0.01$.

	Articulation	Intelligibility	EORTC Speech Scale
	Correlation Coefficient		
speaking rate	.152	.321**	-.253
b1 VOT duration	-.136	-.199	-.041
b1 burst duration	.006	.018	-.189
b2 VOT duration	-.060	-.153	.112
b2 burst duration	-.090	-.234	.166
d1 VOT duration	.123	.143	.272
d1 burst duration	.161	.150	.007
d2 VOT duration	-.136	-.199	-.041
d2 burst duration	.163	.018	.198
p1 silence duration	-.038	-.239*	.112
p1 burst duration	-.108	.110	-.259
p2 silence duration	.005	-.077	-.188
p2 burst duration	.144	.222	.030
t1 silence duration	.079	.019	-.082
t1 burst duration	.027	.106	-.069

t2 silence duration	-.047	.019	-.159
t2 burst duration	.176	.097	-.118
b1 VOT ANN	-.327*	-.149	-.109
b1 burst ANN	-.311*	-.201	-.089
b2 VOT ANN	-.301*	-.262*	-.142
b2 burst ANN	-.314**	-.281*	-.180
d1 VOT ANN	-.310**	-.275*	.016
d1 burst ANN	-.246*	-.261*	.082
d2 VOT ANN	-.271*	-.270*	-.047
d2 burst ANN	-.241*	-.237	-.027
p1 silence ANN	-.288*	-.149	-.118
p1 burst ANN	-.371**	-.348**	.061
p2 silence ANN	-.297*	-.197	-.023
p2 burst ANN	-.351**	-.277*	-.174
t1 silence ANN	-.244*	-.194	.066
t1 burst ANN	-.354**	-.313**	.138
t2 silence ANN	-.165	-.182	-.066
t2 burst ANN	-.254*	-.241	-.112

Discussion

This paper presents an inventory of speech performance 6 months after treatment in a well-defined head and neck cancer patient group after reconstructive surgery and radiotherapy for advanced oral or oropharyngeal cancer. The first aim of the study was to investigate predictive validity of objective analyses of speech quality of stop consonants. Several objective outcome measures differentiated patients from controls: duration of VOT in /b/ and /d/, duration of the burst in /d/ and /p/, and the amount of the phonological feature voicing during VOT in /d/ and /p/ and during the burst of /b/, /p/ and /t/ differentiated patients from controls, although the differences did not reach statistical significance in both realisations of the stop consonants (see table 14). Within patients, predictive validity was less obvious: only the amount of voicing during VOT and the burst of /d/

distinguished tumour stage; duration of the burst in /t/ distinguished tumour site (see section Predictive validity).

In the general population, the voiced stop consonants /b/ and /d/ usually have a shorter VOT than their voiceless counterparts /p/ and /t/. In the present study, results of healthy controls as well as results of patients confirm these findings. It was hypothesized that the duration of VOT preceding the burst in voiceless stops in patients would be longer compared to controls and that the silence period preceding the burst in voiceless stops would show more voicing in patients compared to controls. Argumentation hereof comprises that patients may have more difficulty than controls building up oral pressure in combination and synchronously with ceasing vocal fold vibration to produce voiceless stop consonants. The results in the present study did not show that the silence period preceding the burst in voiceless stops was longer compared to controls. Results in the present study did show indeed that the silence period preceding the burst of all voiceless stops have higher amounts of voicing in patients compared to controls (see table 15 and figure 8 for /p/1). Regarding the voiced stop consonants /b/ and /d/, patients also had a larger amount of voicing during VOT compared to controls (see figure 7) and the duration of VOT was also longer. Compared to previous studies in head and neck cancer, some studies on stop consonants were performed in laryngeal cancer patients, often after laryngectomy. VOT in stop consonants produced by these patients was significantly different from VOT of control speakers.[39-41] One study including oral cancer patients who underwent reconstructive surgery after glossectomy revealed that VOT remained unchanged in most patients, but in some patients the lengths of VOT differed largely after surgery. In these cases, this result was mostly accounted for by the size of the resected tongue and stiffening of the tongue.[42]

No artificial neural network analysis on the phonological feature voicing in stop consonants was performed in previous research in patients treated for head and neck cancer. Results of previous research by automatic speech recognition of speech of patients with head and neck cancer included patients after laryngectomy and oral cancer. Schuster et al. (2006) reported that automatic speech recognition techniques seem good means to objectify and quantify global speech outcome of laryngectomees.[43] Haderlein et al. (2009) also investigated speech of laryngectomized patients and concluded

that automatic speech recognition can be used for objective intelligibility ratings with results comparable to those of human experts.[44] Windrich et al. (2008) demonstrated that automatic speech recognition yielded mean word recognition rate of 49% in oral cancer patients and of 76% in controls. Automatic evaluation highly correlated with the experts' perceptual evaluation of intelligibility.[45] However, a direct comparison of previously found results and results of the present study is not possible. The present study uses one specific phonological feature –voicing– while the previously mentioned studies aimed at the recognition of words.

These findings that neural network feature analyses correlate with subjective speech evaluation were confirmed by the present study. Overall speaking rate and the amount of the phonological feature voicing in almost all consonants was significantly, but moderately, correlated with articulation and intelligibility. No significant correlations were found between objective speech quality measures and patient reported speech outcome (see table 4). Apparently, to be useful for clinical practice purpose, other factors than consonant production alone may be taken in consideration regarding for speech problems as experienced by patients themselves. These factors may include other speech characteristics such as the production of vowels and velar consonants and also coping strategies in dealing with speech problems after cancer treatment.[15, 19, 20]

The results in the present study indicate that patients do maintain a similar overall speaking rate as controls and also similar duration of VOT in voiceless consonants (see figure 8 for /p/1) but not in voiced consonants where patients need more time. In general, patients have a higher amount of voicing during voiceless stop consonants compared to controls (see figure 8 for /p/1). These findings indicate that patients treated for oral or oropharyngeal cancer have more difficulty coordinating articulatory speech movements in conjunction with cessation of vocal fold vibration. However, differences between patients and controls did not always reach statistical significance and were not always consistent for both realizations of the same consonant, indicating that, although there is an overall difference between patients and controls, this difference is not very large or the patient cohort was too small to detect these differences. Other drawbacks of the present study are that the consonants were extracted from running speech meaning that multiple factors may have had influence on consonant production. Such

influences could be speaking rate, phonological context (coarticulation or assimilation), pattern of emphasis of syllables or simply a less fluent way of speaking because of difficulty with reading out loud. Further investigation could rule out these variables by analyzing speech sounds produced in isolation. We would like to emphasize that the current results are obtained by analysis of relatively simple origin. In spite here of, significant results are found which could be considered as promising for future research. In a later developmental stage, a more extensive, dedicated analysis could be used upon a larger data set. In conclusion, the analysis in the present shape is not yet clinically relevant and further research is strongly recommended, but the obtained results are interesting and valuable –for this stage of development.

Conclusion

Objective analysis by acoustic-phonetic measures and artificial neural network analysis of stop consonants in speech of patients treated for oral or oropharyngeal cancer is feasible and valid. Further research is ongoing with larger study samples enabling more in depth analysis and external validation. Results contribute to further development of a multidimensional speech evaluation protocol to be used for research purposes and clinical practice.

Reference List

1. Karnell LH, Funk GF, Hoffman HT. Assessing head and neck cancer patient outcome domains. Head Neck 2000 Jan;22(1):6-11.
2. Borggreven PA, Aaronson NK, Verdonck-de Leeuw IM, et al. Quality of life after surgical treatment for oral and oropharyngeal cancer: a prospective longitudinal assessment of patients reconstructed by a microvascular flap. Oral Oncol 2007 Nov;43(10):1034-42.
3. Bjordal K, Hammerlid E, Ahlner-Elmqvist M, et al. Quality of life in head and neck cancer patients: validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35. J Clin Oncol 1999 Mar;17(3):1008-19.
4. Michi KI, Imai S, Yamashita Y. Improvement of speech intelligibility by a secondary operation to mobilize the tongue after glossectomy. J Craniofac Surg 1989;17:162-6.
5. Pauloski BR, Rademaker AW, Logemann JA, et al. Speech and swallowing in irradiated and nonirradiated postsurgical oral cancer patients. Otolaryngol Head Neck Surg 1998 May;118(5):616-24.
6. McConnel FM, Pauloski BR, Logemann JA, et al. Functional results of primary closure vs flaps in oropharyngeal reconstruction: a prospective study of speech and swallowing. Arch Otolaryngol Head Neck Surg 1998 Jun;124(6):625-30.
7. Yoshida H, Furuya Y, Shimodaira K, et al. Spectral characteristics of hypernasality in maxillectomy patients. J Oral Rehabil 2000 Aug;27(8):723-30.
8. Furia C, Kowalski L, Latorre M. Speech intelligibility after glossectomy and speech rehabilitation. Arch Otolaryngol Head Neck Surg 2001;127:877-83.
9. Sumita YI, Ozawa S, Mukohyama H, et al. Digital acoustic analysis of five vowels in maxillectomy patients. J Oral Rehabil 2002 Jul;29(7):649-56.
10. Hara I, Gellrich N, Duker J, et al. Swallowing and speech function after intraoral soft tissue reconstruction with lateral upper arm free flap and radial forearm free flap. Br J Oral Maxillofac Surg 2003;41:161-9.
11. Seikaly H, Rieger J, Wu YN, et al. Functional outcomes after primary oropharyngeal cancer resection and reconstruction with the radial forearm free flap. Laryngoscope 2003;(113):897-904.
12. Su WF, Hsia YJ, Chang YC, et al. Functional comparison after reconstruction with a radial forearm free flap or a pectoralis major flap for cancer of the tongue. Otolaryngol Head Neck Surg 2003 Mar;128(3):412-8.
13. Bressmann T, Sader R, Whitehill TL. Consonant intelligibility and tongue motility in patients with partial glossectomy. Journal of Oral and Maxillofacial Surgery 2004;(62):298-303.

14. Terai H, Shimahara M. Evaluation of speech intelligibility after a secondary dehiscence operation using an artificial graft in patients with speech disorders after partial glossectomy. Br J Oral Maxillofac Surg 2004 Jun;42(3):190-4.
15. Borggreven PA, Verdonck-de Leeuw I, Langendijk JA, et al. Speech outcome after surgical treatment for oral and oropharyngeal cancer: a longitudinal assessment of patients reconstructed by a microvascular flap. Head Neck 2005 Sep;27(9):785-93.
16. Markkanen-Leppanen M, Isotalo E, Makitie AA, et al. Speech aerodynamics and nasalance in oral cancer patients treated with microvascular transfers. J Craniofac Surg 2005 Nov;16(6):990-5.
17. Whitehill TL, Ciocca V, Chan JC, et al. Acoustic analysis of vowels following glossectomy. Clin Linguist Phon 2006 Apr;20(2-3):135-40.
18. Kazi R, Prasad VM, Kanagalingam J, et al. Analysis of formant frequencies in patients with oral or oropharyngeal cancers treated by glossectomy. Int J Lang Commun Disord 2007 Sep;42(5):521-32.
19. Rinkel RN, Verdonck-de Leeuw I, van Reij EJ, et al. Speech Handicap Index in patients with oral and pharyngeal cancer: better understanding of patients' complaints. Head Neck 2008 Jul;30(7):868-74.
20. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. Folia Phoniatr Logop 2009;61(3):180-7.
21. van der Molen L, van Rossum MA, Burkhead LM, et al. Functional outcomes and rehabilitation strategies in patients treated with chemoradiotherapy for advanced head and neck cancer: a systematic review. Eur Arch Otorhinolaryngol 2009 Jun;266(6):901-2.
22. Klatt DH. Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters. Journal of Speech and Hearing Research 1975;18:686-706.
23. Kent RD. Intelligibility in speech disorders: theory, measurement, and management. Amsterdam, Philadelphia: John Benjamins Publishing; 1992.
24. Ladefoged P, I.Maddieson. The sounds of the world's languages. Blackwell Publishing; 1996.
25. Houde J, Jordan M. Sensomotor adaptation in speech production. Science 1998;(279):1213-6.
26. Allen JS, J.L.Miller, D.DeSteno. Individual talker differences in voice-onset-time. J Acoust Soc Am 2003;113(1):544-52.
27. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 3;85(5):365-76.
28. Praat: doing phonetics by computer. [computer program]. Version 5.2.35. University of Amsterdam: 2007.
29. Nabil N, Espy-Wilson CY. A signal representation of speech based on phonetic features. 1995 May 22; Inst. of Tech., Utica/Rome 1995 p. 310-5.

30. Nabil N, Espy-Wilson CY. A knowledge-based signal representation for speech recognition. Atlanta, Georgia 1996 p. 29-32.
31. Deng L, Sun DX. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J Acoust Soc Am 1994;(95):2702-19.
32. Erler K, Freeman GH. An HMM-based speech recognizer using overlapping articulatory features. J Acoust Soc Am 1996;100(4):2500-13.
33. King S, Taylor P. Detection of phonological features in continuous speech using neural networks. Comp Speech Lang 2000;14(4):333-53.
34. Robinson T, Hochberg M, Renals S. The use of recurrent neural networks in contnuous speech recognition. In: Lee C-H, Soong F, ., editors. Automatic Speech and Speaker Recognition- Advanced Topics. Kluwer Academic Publishers; 1996. p. 233-58.
35. Bridle J, Deng L, Picone J, et al. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. John Hopkins University 1998 p. 1-61.
36. van Son RJJH, Binnenpoorte D, van den Heuvel H, et al. The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database. Aalborg 2001 p. 2051-4.
37. Graupe D. Principles of Artificial Neural networks. Advanced series on circuits and systems - volume 6 ed. Singapore: World Scientific Publishing Company Co. Pte. Ltd.; 2007.
38. http://nico.nikkostrom.com/ [computer program]. KTH, Stockholm: 1997.
39. Robbins J, J.Christensen, G.Kempster. Characteristics of Speech Production after Tracheoesophageal Puncture. Journal of Speech and Hearing Research 1986;29:499-504.
40. Christensen J, B.Weinberg, P.Alfonso. Productive Voice Onset Time Characteristics of Esophageal Speech. Journal of Speech and Hearing Research 1978;21:56-62.
41. Ng M, J.Wong. Voice onset time characteristics of esophageal, tracheoesophageal and laryngeal speech of cantonese. Journal of Speech, Language and Hearing Research 2009;52:780-9.
42. Savariaux C, Perrier P, Pape D, et al. Speech production after glossectomy and reconstructive lingual surgery: a longitudinal study. 2001.
43. Schuster M, Haderlein T, Noth E, et al. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. Eur Arch Otorhinolaryngol 2006 Feb;263(2):188-93.
44. Haderlein T, Riedhammer K, Noth E, et al. Application of automatic speech recognition to quantitative assessment of tracheoesophageal speech with different signal quality. Folia Phoniatr Logop 2009;61(1):12-7.
45. Windrich M, Maier A, Kohler R, et al. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. Folia Phoniatr Logop 2008;60(3):151-6.

Chapter 5

Speech quality in patients treated for oral or oropharyngeal cancer: validation of objective speech analyses

5

Marieke de Bruijn
Louis ten Bosch
Birgit I. Witte
Johannes A. Langendijk
C. René Leemans
Irma Verdonck- de Leeuw

Submitted

Abstract

Speech quality in patients treated for oral or oropharyngeal cancer is most often evaluated subjectively but can also be assessed objectively by analyses of speech recordings. In earlier studies, we developed and tested various acoustic-phonetic (AP) and artificial neural network (ANN) analyses separately. The present study aims to validate these objective speech quality analyses in the same patient cohort and in a new patient cohort (external validation). Speech quality was evaluated subjectively regarding hypernasality, articulation, intelligibility and by patient reported speech outcome. AP analyses were performed on vowels /a, i, u/, stop consonants /k, p, b, d, t/ and fricative /x/. ANN analysis of the feature nasalance was performed on /a, i, u/ and on the entire stretch of speech; ANN analysis of the feature voicing was performed on /p, b, d, t/. In patient cohort 1 Intelligibility was predicted by AP analysis of /p/ and vowel space and by ANN analysis of /d/. Articulation was predicted by AP analysis of vowel space and by measurements of the feature 'voicing' for /b/. Nasal resonance was predicted by AP analysis of /a/, /x/ and /b/. Patient reported speech outcome was predicted by AP analysis of /i/ and /k/ and by ANN analysis of /p/. The amount of variance explained varied from moderate to. In cohort 2 Intelligibility was predicted by AP analysis of /a/, /i/ and /x/. Articulation was predicted by AP analysis of vowel space and by measurements of the feature 'voicing' for /p/. Nasal resonance was predicted by AP analysis of /p/ and /t/. Patient reported speech outcome was predicted by AP analysis of /u/ and /t/ and by ANN analysis of /d/. The amount of variance explained varied from moderate to poor. Objective analyses of speech quality in HNC patients are valid and contribute moderately to a multidimensional speech evaluation protocol.

Introduction

Speech quality is one of the tumour specific quality of life domains that is often compromised in patients treated for oral or oropharyngeal cancer.[1-3] Speech problems occur in 40-70% of these patients, caused by the relatively large amounts of tissue that are damaged or removed in the oral cavity or oropharynx. In addition, radiotherapy may cause fibrosis and stiffening of the tissues of the vocal tract. Surgical tumour resection is associated with a lower intelligibility and worse articulation also after reconstructive surgery. Patients are less able to quickly and correctly produce speech sounds which negatively affect intelligibility and communicational suitability. Speech problems often lead to impaired social functioning and lower quality of life.[4-6]

Perceptual evaluations by professionals such as speech therapists or patient reported outcome through questionnaires are most often used to assess speech quality in clinical practice.[4, 7-9] Perceptual speech quality assessment protocols do not allow to wide sharing between different groups of physicians or speech therapists in various hospitals. Moreover, despite elaborate attempts to design perceptual rating instruments for speech quality, one has to conclude that it is only feasible to obtain high consensus on broad aspects of speech such as tempo and intelligibility but that listeners do not succeed at attaching reliable judgments on more specific aspects of speech, such as nasality or articulation.[10, 11] Therefore, there is an urgent need for an objective speech quality assessment tool to be used in clinical practice and for research purposes.

Three previously performed pilot studies in the first stage of the development of objective speech analyses in patients treated for oral or oropharyngeal cancer, revealed several speech sounds distinguishing patients from controls.[5, 12, 13] Speech sounds were analysed through acoustic phonetic (AP) analyses and through the use of an artificial neural network (ANN).[1] The vowel space regarding the cardinal vowels /a, i, u/, and amount of hypernasality, the stop consonants /p, t, b, d/, and the velar speech

[1] This is an analysis based on a representation as obtained by automatic estimation of articulatory features, performed by means of an artificial neural network.

sounds /k, x/ were identified as potential speech quality markers that categorizes deviant speech as produced by patients from normal speech as produced by healthy controls. Also, within patients these speech sounds differentiated with regard to tumour stage: patients treated for larger tumours having worse speech quality. However, the results of the three pilot studies revealed that the predictive value is moderate. This result may be explained by the fact that the three studies were performed separately, and it is hypothized that better results could be obtained when all variables are combined into one model.
The goal of the present study is to combine the results from the previous studies and to externally validate the findings onto another cohort of patients treated for oral or oropharyngeal cancer.

Patients and methods

Patients
All patients were treated by surgery and postoperative radiotherapy at the Department of Otolaryngology–Head & Neck Surgery and the department of Radiation Oncology of the VU University Medical Center in Amsterdam, the Netherlands, between 1994 and 2003. Patients underwent composite resections for advanced oral or oropharyngeal squamous cell carcinoma with microvascular soft tissue transfer for the reconstruction of surgical defects. Surgery consisted of composite resections including excision of the primary tumour with en bloc ipsilateral or bilateral neck dissection. In case of oropharyngeal carcinomas, a paramedian mandibular swing approach was used. All free flaps were successful. Patients received radiotherapy in case of advanced (T3–4) tumours, positive or close surgical margins, multiple lymph node metastases and/or extra nodal spread. The primary site received a dose of 56 to 66 Gy in total (2 Gy per fraction, 5 times per week), depending on surgical margins. The nodal areas received a total of 46, 56 or 66 Gy (2 Gy per fraction, 5 times a week) in case of N0, N+ without extranodal spread and N+ with extranodal spread, respectively. Exclusion criteria were incapability to participate in functional tests and difficulty communicating in Dutch. Written informed consent was obtained from all patients.

Patient cohort 1 included 51 patients at 6 months after treatment and aged 23 to 73 years (mean: 53.8 years, sd: 8.7 years). Patient cohort 2 included 64

patients at six months to nine years after treatment (median: 1.5 years, range: 8.5 years) and aged 26 to 87 years (mean: 60.4 years, sd: 10.8 years). See table 17 for additional information concerning gender and clinical parameters.

Table 17. Characteristics of 51 patients in cohort 1 and 64 patients in cohort 2.

	Gender		Tumour site		T-classification		(chemo-) Radiation therapy	
	Male	Female	Oral cavity	Oro-Pharynx	2	3-4	Yes	No
	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Cohort 1	28 (55%)	23 (45%)	21 (41%)	30 (59%)	26 (51%)	25 (49%)	47 (92%)	4 (8%)
Cohort 2	35 (55%)	29 (45%)	31 (48%)	33 (52%)	41 (64%)	23 (36%)	43 (67%)	20 (31%)

Speech recordings

Patients read-aloud a standardized Dutch text. The distance between lips and microphone was 30 centimeters. Speech recordings were conducted in a sound attenuated booth. For each speaker the recording level was adjusted to optimize signal-to-noise ratio. The recorded speech was digitized with Cool Edit PRO 1.2 (Adobe Systems Incorporated, San Jose, CA, USA) with 22-kHz sample frequency and 16-bit resolution.

Subjective speech evaluation

Perceptual evaluation of speech quality comprised ratings of intelligibility, articulation and hypernasality by two speech pathologists on the entire stretch of running speech. To enable subjective speech evaluation, a computer program was developed to perform blinded randomized listening experiments and to store intelligibility, articulation and hypernasality scores in a database. The panel of trained listeners rated articulation and hypernasality on a 4-point scale, ranging from normal to increasingly deviant speech quality. Intelligibility was scored using a 10-point scale, following the Dutch educational grading system where 1 represents the worst score and 10 represents the best score and 6 is just sufficient. Interrater agreement for subjective assessment of intelligibility ranged from 40% to 90%. Intrarater agreement for two repeated speech fragments of articulation and hypernasality was high with 100% equal scores between the ratings.

Speech problems in daily life as reported by patients were assessed by the Speech Subscale (including 3 items) of the EORTC Quality of Life Questionnaire H&N35 module.[14, 15] The scores are linearly transformed to a scale of 0 to 100, with a higher score indicating a higher level of speech problems. A detailed description of these subjective speech ratings and results can be found in Borggreven et al.[4]

Objective speech evaluation

In the present study, vowels /a, i, u/, velar consonants /k, x/ and stop consonants /b, d, p, t/ were objectively analysed by acoustic-phonetic analyses and by artificial neural network analysis. See table 18 for an overview of speech analyses specifications.

Table 18. Overview of speech material and Acoustic-Phonetic (AP) and Artificial Neural Network (ANN) analyses.

	Acoustic-Phonetic Analyses	Artificial Neural Network
/a, i, u/	Formant 1 Formant 2 Vowel space	Feature 'nasal'
/x/	Spectral slope	-
/k/	Burst percentage	-
/b, d, p, t/	Duration VOT + burst	Feature 'voicing'
entire text	Speaking rate (words/minute)	Feature 'nasal'

- Acoustic-phonetic analyses

Speaking rate of the entire stretch of speech was calculated in words per minute.

Of the vowels /a, i, u/ the first formant frequency (F1) and the second formant frequency (F2) were studied in the present study. Vowels are – compared to consonants – relatively easy to identify in the speech signal and to analyze acoustically. Vowel formant analyses proved to be valid measures of speech quality in patients with deviant speech originating from oral cancer or other origins in earlier studies.[8, 16] Vowel identity (or its spectral color) is characterized by acoustic correlates and is primarily determined by its formants. Broadly speaking, F1 is associated with 'height': the degree of opening of the vocal tract. F2 is associated with the anterior-posterior tongue position.[17] Plotting the vowels /a, i, u/ onto a graphical F1-F2 representation shows the vowel space (more specifically the vowel space) (see figure 9). The vertices of the vowel space represent the most extended positions. The area of the vowel space is a measure for the amount of reduction in the vowel system and can (formally) be measured in terms of Hz^2.[18]

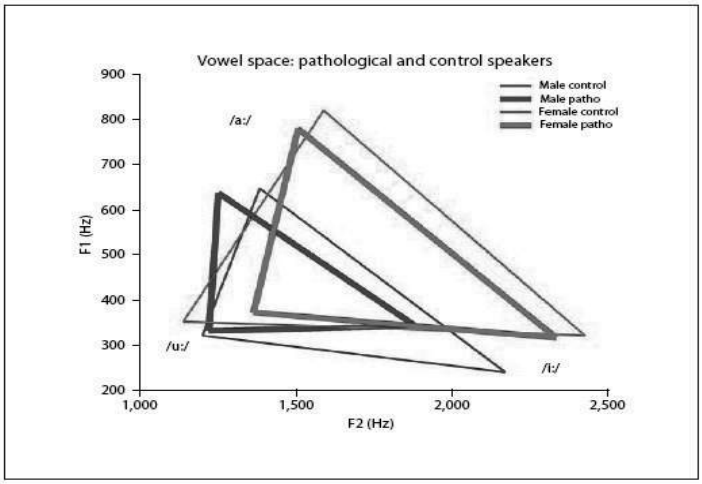

Figure 9. Vowel space of male (blue) and female (pink) patients (fat lines) and of controls (thin lines).

The velar consonants /k/ and /x/ were analyzed because earlier research revealed that patients with an oral or oropharyngeal tumour often have difficulties with the production of velar speech sounds: speech raters often mistook /k/ for /x/.[4, 19] For /k/ the duration of air pressure release (the so-called plosive) as a percentage of the total duration was measured (see figure 10: calculations on /k/ are comparable to those on /t/). For /x/ the spectral slope was used as outcome measure. It describes how quickly the amplitude decreases concurrently with increase of frequency. The spectral slope is of influence on sound quality and timbre and is used for speech discrimination and voice recognition.[20]

Of stop consonants /b, d, p, t/ the duration of voice-onset-time (VOT) and burst were measured. In normal speech, VOT is one of the functionally most relevant parameters that distinguishes voiced and voiceless stops and is a result of the temporal coordination of voicing and oral articulation gestures. VOT is defined as the length of time that passes between when a stop consonant is released and when voicing, the vibration of the vocal folds, begins. For voiced consonant stops /b, d, g/, the voicing starts before the burst of airflow. This voice sound preceding the burst in voiced stop consonants is typical for the Dutch language. For voiceless consonant stops /p, t, k/, a short period of silence precedes the burst. Voiced plosives /b, d, g/ usually have a shorter VOT than voiceless plosives /p, t, k/ (see figure 2). VOT is significantly related to speaking rate.[17, 21-23] The voiced stop consonants /b/ and /d/ and their voiceless counterparts /p/ en /t/ were used as speech material (in the Dutch phonological system the voiced counterpart /g/ of the voiceless stop consonant /k/ is not present and therefore these velar stop consonants were not investigated). Duration of VOT and the duration of the following release of air pressure were measured. Where this paper refers to the VOT, actually the pre-burst silence portion is meant.

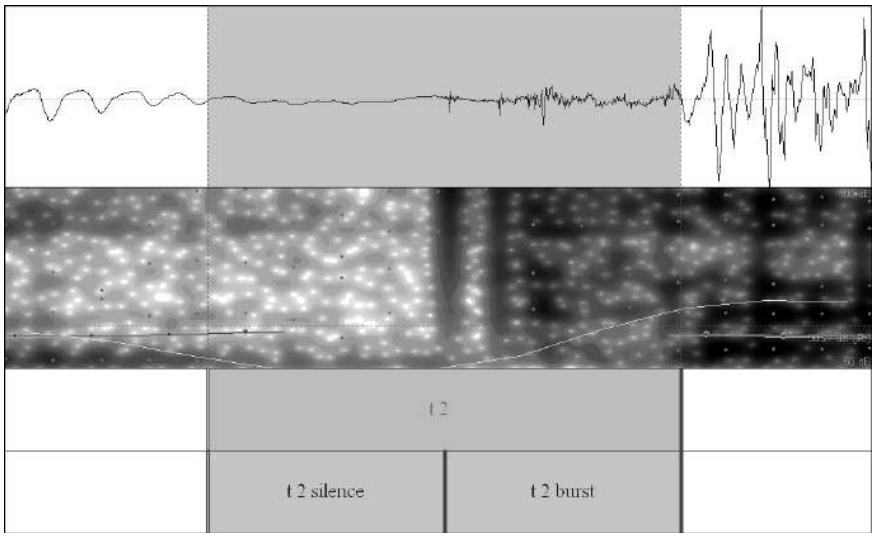

Figure 10. Example of segmentation of /t/. The upper panel tier shows the waveform of the entire duration of /t/ ("t 2"); the second panel displays the spectrogram while the third panel contains the segmentation and labelling partition in silence ("t 2 silence") and burst ("t 2 burst"). The spectrogram shows the relative absence of voicing during the pressure-building silence, followed by the outburst of voicing during the burst. This figure is made in Praat.[24]

Since the acoustic realization of certain speech sounds may depend on its context, we took different phonological contexts around the target speech sounds into account, in order to improve generalization. For each selected speech sound two acoustic realizations were segmented from running speech and were acoustic- phonetically analyzed using the speech processing software Praat.[24] A spectrogram functioned as a visual representation of the speech signal, which facilitated recognition of phonemes in the speech signal and facilitated precise extraction of phonemes from running speech. Spectral and acoustic speech analyses were automatically performed using scripts.[24]

- Artificial neural network and feature analysis

In the context of research on Automatic Speech Recognition (ASR), several speech decoding techniques have been developed that are able to automatically segment and label an unknown utterance in terms of phone-like segments.[25, 26] In the last decade however, new approaches were designed that circumvent the use of phone symbols in the segmentation of speech. Instead of using phonetic symbols (mostly predefined by the ASR development of the ASR system), these methods focus on more basic properties of the speech signal that characterize the speech signal in a way more similar to acoustic features or phonological features such as 'manner' and 'place' of articulation.[27, 28] In the last decade, several classification techniques have become available, such as Artificial Neural Nets (ANNs) and Support Vector Machines.[29-31]

In the present study we used ANNs to obtain an articulatory feature representation of an input speech signal. ANNs contain a number of model parameters (weights of the connections between network nodes) that determine the relation between input and output of these ANNs. The parameters can be trained on a training set in which input and desired output of the ANN are specified for each training data point. The ANNs used in this paper had an input context consisting of 7 consecutive mel-frequency cepstral coefficients (MFCC) frames, which are mathematical coefficients for sound modelling. During training these were used in combination with a canonical value (0 or 1) of the phonological feature that was being modelled by the ANN. The ANNs had been trained on speech from speakers without reported articulatory problems and in the test provided estimations of phonological features from the acoustic input signal. The ANNs used in the present study were trained on the basis of speech originating from the corpus of the Institute of Phonetic Sciences, University of Amsterdam, the Netherlands (IFA Spoken Language Corpus). This corpus contains speech in a variety of styles of four normal speaking males and four normal speaking females in combination with accompanying phoneme labels.[32] For the training of ANN, speech spectra of one healthy male and one healthy female in combination with accompanying segmentations were derived from the IFA-corpus. Several normalisation steps applied during the training of the ANN classifier (including mean and variance normalisation of the MFCCs on utterance level, energy normalisation) led to the best possible generalization

performance of the eventual trained ANNs. ANNs were specifically trained on labelled speech spectra for all phonological features.[33]

Training took place via error-back propagation, one of the well-known and commonly used training procedures for ANNs. After training, ANN-voicing was tested on speech of two other speakers from this IFA corpus. The output of ANN-voicing varies between 0 (absent) and 1 (present). A value of 0.8 for instance means that voicing is rather strongly present in the speech frame. Depending on the amount of input and of consistency in labels during training, ANN-voicing achieves high levels of correct classification. In the quality assessment of ANNs, the performance is given in terms of frame accuracy. Performance during testing was 80% correct at frame level. An accuracy of 80% means that the classifier correctly classifies the feature value assigned to this frame for 80% of all frames in the evaluation test set. The degree of the phonological feature voicing as identified by the artificial neural network was determined for VOT as well as for the following burst of the selected stop consonants.

In the experiments described here, ANNs were used to estimate the degree of various phonological features such as manner, place, voicing, front-back and rounding. Each of these properties was modelled by an ANN. Each ANN was modelled by a three layer feed-forward network: one input layer, one hidden layer, and one output layer. The input layer is fed with the MFCCs obtained from the MFCC extraction step. The units in the output layer represent the estimated values of the various options for that particular feature. For example, the manner feature is modelled by the manner-ANN which has 6 units on its output layer (see also chapter 1, table 1):

Manner: 0/NULL-approx-fric-nasal-stop-vowel

These six units in the manner-ANN estimate the degree of NULL, approximant, fricative, nasality, stop, and vowel, for each 10ms frame in the input speech signal, respectively. The NULL value is a unit that takes positive values if the network is not able to positively assign values to any of the other five units or when that makes no sense such as during silence. In total, the manner-ANN provides 6 values for each frame of 10 ms.

By taking into account the output of the other five ANNs (place, voicing, front-back, rounding, and static) and stacking all results, we obtain a 28-dimensional feature vector for each frame of 10 ms. In the present study, we particularly focus on the phonological features voicing (ANN-voicing) and nasality (ANN-nasal), that is, on specifically two out of this collection of 28 values per 10ms. One of the most important model parameters in an ANN is the dimension (number of hidden units) of the hidden layers. In this study, we adopted a setting that has been suggested in the literature[29] and that showed good results in training and test sessions with speech from healthy controls.

The ANNs that were used in this study are available as public domain software.[34] In the present study we used phonological features voicing and nasal (in the sequel, the resulting ANN is referred to as ANN-voicing and ANN-nasal).The motivation to use ANNs instead of Support Vector Machines (SVM) is determined by the facts that compared to SVM, ANNs deliver a relatively small model and ANNs use continue mapping. The model does not encounter discrete selections during the classification task, as is the case with SVM.

Statistical analysis

Pearson correlation coefficients were used to investigate univariate associations between the objective analyses on the one hand and subjective speech evaluations of intelligibility, articulation, nasal resonance and patient-reported outcome at the other. In case of significant differences of values of the two realisations of each speech sound, both realisations were used separately. In case of no differences of the values of the two realisation of each speech sound, the average of the two realisations was used. Univariate correlations with a p-value $<.20$ were also inserted into the multivariate regression analyses, meaning that only correlations indicating a (significant) coherence were used for further statistical analysis into predictive value of subjective evaluations.

Multivariate regression analyses were performed to obtain insight into the role of objective parameters in predicting subjective speech evaluation. For intelligibility and patient reported speech outcome stepwise linear regression was used, while for articulation and nasal resonance conditional forward

logistic regressions were performed on a binary scale [normal (score 0) vs. deviant (scores 1-3)].
This procedure was performed in cohort 1 and then repeated -as external validation- in cohort 2. As a result hereof, the variables may vary in the final models belonging to subjective (self-)evaluations in cohort 1 and cohort 2.

Results

Univariate Pearson correlations
For cohort 1, both positive and negative Pearson correlations (table 19a) revealed that intelligibility was significantly related to AP analyses of vowel space, /i, u/, /k/ and /b, p/. Subjective evaluation of articulation was significantly related to AP analyses of vowel space, /i/and to ANN-voicing in/d/. Subjective evaluation of nasality was significantly related to AP analyses of /a, i/ and /x/ and to ANN-nasal in /i/. Patient reported speech outcome was not significantly related to AP analyses and not to ANN-voicing nor ANN-nasal.

Table 19 a. Cohort 1: Pearson correlation coefficient r between objective speech parameters and subjective parameters: intelligibility, articulation, nasality and patient-reported speech outcome. "/a/1" means the first realisation of the two /a/'s that are included in the study, "F1" is the first formant, "F2" is the second formant, "VOT" is the voice-onset-time of stop consonants, "ANN" is analysis by an artificial neural network. Phonemes are "averaged" when there is no significant difference between the two realisations of one phoneme. * = p<0.05, ** = p<0.01, Bold = p<.20

N=51	Intelligibility	Articulation	Nasality	EORTC Speech Scale
Acoustic-Phonetic Analysis				
Vowel space	**.34***	**.29***	**-.24**	**-.19**
/a/1 F1	.14	.06	**-.45****	-.12
/a/1 F2	.13	.08	**-.25**	-.16
/a/1 F2	.13	.08	**-.25**	-.16
/a/2 F1	**.19**	.18	-.11	**-.21**
/a/2 F2	-.15	.10	-.07	.09
/i/1 F1	-.13	**-.23**	**-.33***	.07
/i/1 F2	**.31***	.06	**-.19**	**-.22**
/i/2 F1	.04	-.05	**-.26**	-.03
/i/2 F2	**.20**	**.28***	**-.20**	**-.21**
/u/1 F1	.04	.01	**-.19**	-.15
/u/2 F1	-.05	.02	**-.20**	-.05
/u/ F2 averaged	-.13	-.17	.01	.04
/k/1	**.33***	.18	-.06	**-.24**
/k/2	.00	**-.23**	.14	-.09
/x/1	-.01	**-.19**	.10	.13
/x/2	-.03	.13	**.34***	-.02
/b/ VOT averaged	-.03	**.20**	**-.24**	-.13
/b/1 burst	.11	.13	.02	-.17
/b/2 burst	**-.28***	-.09	-.15	.10
/d/1 VOT	-.01	-.02	.08	.17

/d/2 VOT	.14	**.28**	.10	-.18
/d/ burst averaged	.05	-.04	**.22**	-.07
/p/1 silence	**-.29***	-.04	-.15	**.23**
/p/2 silence	-.07	-.02	-.16	-.17
/p/ burst averaged	.07	-.15	**.25**	**-.21**
/t/ silence averaged	-.03	-.10	.15	-.15
/t/1 burst	.14	.09	**.21**	-.12
/t/2 burst	-.10	.03	.16	.08
Artificial Neural Network Analysis				
ann_nasal	-.18	-.18	-.08	-.06
/a/ ANN averaged	-.04	-.02	.02	-.11
/i/1 ANN	.13	.18	**.38****	-.08
/i/2 ANN	-.17	**-.23**	.02	.04
/u/1 ANN	-.05	-.13	**.21**	.02
/u/2 ANN	-.05	-.18	-.04	-.12
/b/1 VOT ANN	-.01	**-.34***	.03	-.06
/b/1burst ANN	.01	**-.23**	.08	-.11
/b/2 VOT ANN	-.16	**-.31***	.12	-.11
/b/2 burst ANN	-.18	**-.29***	.17	-.13
/d/ VOT ANN averaged	**-.21**	**-.20**	-.02	.08
/d/1 burst ANN	**-.23**	**-.22**	.01	.08
/d/2 burst ANN	**-.22**	-.12	.11	.13
/p/1 silence ANN	.02	-.12	-.12	-.07
/p/1 burst ANN	-.10	-.10	.18	.10
/p/2 silence ANN	-.07	-.13	.14	-.03
/p/2 burst ANN	-.10	**-.20**	**.23**	**-.19**
/t/ silence ANN averaged	-.20	-.17	-.07	.07
/t/1 burst ANN	-.15	**-.23**	.06	.15
/t/2 burst ANN	-.18	-.10	.01	-.04

For cohort 2, both positive and negative Pearson correlations (table 19b) revealed that intelligibility was significantly related to AP analyses of /a, i/ and /x/ and to ANN-voicing of /a/ and /d/. Subjective evaluation of articulation was significantly related to AP analyses of /x/ and to ANN-voicing of /p/. Subjective evaluation of nasal resonance was significantly related to AP analyses of the second formant of /p, t/ and to ANN-nasal on the entire stretch of speech. Subjective evaluation of patient-reported outcome was significantly related to AP analyses of the second formant of /u/ and /t/ and not to ANN-analyses.

Table 19 b. Cohort 2: Pearson correlation coefficient r between objective speech parameters and subjective parameters: intelligibility, articulation, nasality and patient-reported speech outcome. "/a/1" means the first realisation of the two /a/'s that are included in the study, "F1" is the first formant, "F2" is the second formant, "VOT" is the voice-onset-time of stop consonants, "ANN" is analysis by an artificial neural network. Phonemes are "averaged" when there is no significant difference between the two realisations of one phoneme. * = p<0.05, ** = p<0.01, Bold = p<.20

N=64	Intelligibility	Articulation	Nasality	EORTC Speech Scale
Acoustic-Phonetic Analysis				
vowel space	**.22**	**.23**	.11	-.02
/a/1 F1	.03	-.04	.16	-.18
/a/1 F2	**-.44****	-.14	-.15	.09
/a/2 F1	-.07	-.07	-.10	.02
/a/2 F2	-.03	.13	-.16	.14
/i/ F1 averaged	-.08	.13	-.05	.02
/i/1 F2	**.27***	**.23**	.08	-.13
/i/2 F2	**.30***	**.20**	**.23**	**-.22**
/u/1 F1	-.07	-.08	-.05	**-.30***
/u/1 F2	-.01	-.13	.08	-.03
/u/2 F1	.10	-.12	-.01	-.09
/u/2 F2	**-.22**	.01	.01	**-.24**
/x/1	**.27***	**.26***	**.21**	-.03
/x/2	**.32***	**.27***	.10	**-.22**

/k/ averaged	**.25**	.04	.01	−.10
/b/1 VOT	−.07	−.10	.02	.05
/b/2 VOT	.00	−.03	.00	−.11
/b/ burst averaged	−.03	.03	.04	.10
/d/ VOT averaged	.05	−.03	−.08	−.14
/d/ burst averaged	.01	**−.19**	.01	.04
/p/ silence averaged	−.06	.05	−.01	.07
/p/ burst averaged	−.12	−.13	**−.28***	**.19**
/t/1 silence	−.03	−.01	.09	**.32***
/t/2 silence	.07	.14	.13	**−.26**
/t/ burst averaged	.11	−.05	**−.27***	.01
Artificial Neural Network Analysis				
ann_nasal	.15	−.03	**−.26***	.06
/a/1 ANN	**.21**	−.14	−.09	−.04
/a/2 ANN	.02	.01	.02	.05
/i/ ANN averaged	**−.21**	−.15	−.08	.17
/u/ ANN averaged	.05	.09	.15	−.12
/b/1 VOT ANN	**.18**	.01	.02	−.07
/b/2 VOT ANN	.07	−.09	−.10	.15
/b/1 burst ANN	**.17**	−.02	.04	−.06
/b/2 burst ANN	.04	−.14	.00	.17
/d/1 VOT ANN	.14	.12	−.06	−.12
/d/1 burst ANN	.15	.02	−.02	.02
/d/2 VOT ANN	**.37****	.15	−.04	**.22**
/d/2burst ANN	**.35****	.08	−.07	**.24**
/p/1 silence ANN	−.05	**−.27***	−.15	−.05
/p/2 silence ANN	**.23**	−.02	−.13	−.13
/t/1burst ANN	**.22**	−.05	.02	−.16
/t/2 burst ANN	.08	**−.23**	−.13	.09
/t/ ANN silence averaged	.07	−.13	**−.18**	.12

Multivariate regression analyses

To obtain insight into which objective parameters predict significantly subjective speech evaluations, multivariate regression analyses were performed in cohort 1 (table 20a). For the first cohort, the results revealed that intelligibility was predicted by AP analysis of /p/ and vowel space, and by ANN-voicing of /d/. Articulation was predicted by AP analysis of vowel space and by ANN-voicing of /b/. Nasal resonance was best predicted by AP analysis of /a/, /b/ and /x/. Patient reported speech outcome was predicted by AP analysis of /i/ and /k/ and by ANN-voicing of /p/. The amount of variance explained varied from moderate (52.0% for Nasal resonance, 37.7% for Intelligibility and 36.2% for Articulation) to poor (21.1% for patient-reported speech outcome).

Table 20 a. Cohort 1: Prediction of intelligibility, articulation, nasal resonance and patient-reported speech outcome by acoustic-phonetic and Artificial Neural Network analyses. * $p < 0.05$, ** $p < 0.01$.

		b	statistic	p	R^2
Intelligibility	/p/1 silence	-32.21	-3.81	.000**	37.7%
	/d/1 burst ANN	-2.61	-3.05	.004**	
	/d/ VOT ANN averaged	2.32	2.20	.033*	
	Vowel space	3.30	2.10	.045*	
Articulation	vowel space	9.59	5.15	.023*	36.2%
	/b/2 VOT ANN	-4.72	5.27	.022*	
Nasal resonance	/a/1 F1	-0.01	4.81	.028*	52.0%
	/x/2	0.24	7.43	.006**	
	/b/ VOT averaged	-78.38	4.26	.039*	
EORTC Speech Scale	/i/2 F2	-0.01	-2.22	.032*	21.1%
	/k/1	-0.39	-1.71	.095	

Multivariate regression analyses were also performed for cohort 2 (table 20 b). The results revealed that intelligibility was predicted by AP analysis of /a/, /i/ and /x/ and not by ANN analysis. Articulation was predicted by AP analysis of vowel space and by ANN-voicing of /p/. Nasal resonance was best predicted by AP analysis of /t/ and /p/ and not by ANN analysis. Patient reported speech outcome was predicted by AP analysis of /t/ and /u/ and by ANN-voicing of /d/. The amount of variance explained differed from moderate (51.9% for patient-reported outcome and 41.3% for Intelligibility) to poor (21.8% for Nasal resonance and 20.9% for Articulation).

Table 20 b. Cohort 2: Prediction of intelligibility, articulation, nasal resonance and EORTC selfevaluations by acoustic-phonetic and Artificial Neural Network analyses.

* $p < 0.05$, ** $p < 0.01$.		B	statistic	p	R^2
Intelligibility	/a/1 F2	-0.004	-4.16	.000**	41.3%
	/i/ F2	0.002	2.75	.009*	
	/x/ 1	0.06	2.06	.047*	
Articulation	Vowel space	6.96	2.58	.108	20.9%
	/p/1 silence ANN	-2.88	3.04	.082	
Nasal resonance	/p/ burst average	-40.07	3.01	.000**	21.8%
	/t/ burst average	-44.87	3.10	.000**	
EORTC Speech Scale	/t/2 silence	-452.62	-3.66	.001**	51.9%
	/t/1 silence	329.00	3.10	.004**	
	/u/1 F1	-0.10	-2.61	.013*	
	/d/2 burst ANN	20.97	2.25	.031*	

Discussion

This paper presents an inventory of speech quality in a well-defined head and neck cancer patient group six months to nine years after reconstructive surgery and postoperative radiotherapy for advanced oral or oropharyngeal cancer. The aim of the present study was to combine the results from previously performed studies on the development of objective speech analyses[5, 12, 13] to judge the validity of objective speech sounds put together into one model. Firstly, this model was tested in the same cohort of 51 patients as we used in our previous studies.[5, 12, 13] Then a second cohort of 64 patients treated for oral or oropharyngeal cancer was used to externally validate the findings.

Results in the present study confirm our earlier pilot studies.[5, 12, 13] For cohorts 1 and 2 Intelligibility was related to AP analyses of vowel space and /a, i, k, b, p/ and to ANN analyses of /a, d/. Articulation was related to AP analyses of vowel space and /i, x/ and to ANN analyses of /p, b/. Nasal resonance was related to AP analyses of /a, i, x, p, t/ and to ANN analyses of /i/ and ANN-nasal on the entire stretch of speech. Patient-reported outcome was related to AP analyses of vowel space and /u, t/. In both cohorts predictive values remained moderate to poor. Several speech sounds with a p-value <.20 were added to this selection and used in the multivariate model.

In the present study, the multivariate model used in the second cohort of 64 patients required another selection of objective speech analyses than the first cohort of 51 patients, indicating that either both cohorts were not comparable or that the multivariate models are not yet very stable (i.e. dependent on the study cohort). These differences originate from the first methodological step in which it was revealed which two realisations of one speech sound were not normally distributed. If there were significant differences, the two realisations were used separately. The differences in the models show that there is variability in the production of speech sounds and that the models are cohort-dependent. It is possible that a too large amount of variables was used in combination with a relatively small population with large differences in patient characteristics. The model then describes the 'noise' of these parameters and not the underlying patterns, in which case over fitting occurs and predictive value remains low.

The differences between the multivariate models of the two study cohorts may also be explained by demographic and clinical characteristics. Patients in the first cohort were younger (mean 54 vs 60 years) and age has been shown to be a significant factor in voice and speech analyses.[35, 36]
The first cohort involved less patients with smaller T1-T2 tumours (51%) than the second cohort (58%). In the first cohort almost all patients (92%) received radiotherapy, versus 78% in the second cohort. Both tumour stage and treatment modality may have been of influence on speech quality due to the amount of surgically removed tissue and/or stiffening of tissue involved in speech production. It is very well possible that the technique of radiation has been improved over the years between inclusion of patients of both cohorts which causes a difference in clinical characteristics. Also the difference in time between treatment and speech recording (follow-up) was different which may have had influence on the development of fibrosis in the cohort with a longer time until speech recording. In cohort 1, speech recordings were made 6 months after treatment in the first cohort, versus 27 months in the second cohort. Although not investigated, inspection of the recorded speech samples during segmentation gave the impression that speech quality of patients in the second cohort was better compared to speech quality in the first cohort. Also, the scores on the four subjective scales differed significantly and were better for cohort 2 compared to cohort 1. It may be that patients who survived for multiple years (cohort 2) have had more time to adjust to an altered vocal tract and develop strategies to speak more clearly.

Another striking finding in the present study was that, although a higher predictive value was expected now that all speech data as assessed by the two objective methods (AP analyses and ANN analyses) were combined into one model, in both study cohorts the value of objective speech analyses to predict subjective speech evaluation by raters or patients themselves remained moderate at best. In the earlier pilot studies[5, 12, 13] speech sounds were investigated based on classes of speech sounds: phonemes belonging to the class of velar speech sounds, vowels and stop consonants. In the present study these separate phonemes were used concurrently in one model. There may be an interaction between the studied phonemes and other -yet unknown- factors influencing subjective apprehension of speech that could clarify the lower than expected predictive value. Factors that were not yet controlled for but could be taken into account in further research

include a variety of mainly phonetic items. The interaction between phonemes -known as coarticulation and assimilation- is of importance since phonemes are pronounced slightly different due to neighbouring speech sounds. Since phonemes are not pronounced in identical ways it is necessary to use and investigate many different phonological contexts in further research.[37] Second, inspection of word- and sentence composition structure patterns (prosodic categories) is needed because stressed syllables "differ from those that are unstressed along at least four parameters: duration, fundamental frequency, overall intensity, and spectral composition"[38] and thus may influence the dataset. In further research a larger variety of phonemes is therefore needed. Speech sounds belonging to the classes of velar speech sounds, vowels and stop consonants were already investigated in the previous and present studies but could be completed by other phonemes belonging to these classes. Following this expansion other classes of speech sounds could be investigated, such fricatives (/s,z,v,f/), nasals (/m,n,ng/) and trills (/r/). Finally, speaker dependent characteristics should be carefully controlled for, such as influence by personal characteristics (gender, age, health) as well as cultural aspects (sociolect, geographical background, language, accent and dialect of the speaker).[39]

Observing the results and interpretation hereof, it is argued that a speech quality assessment protocol to be used for clinical and research purposes in patients treated for HNC ideally should be multidimensional including subjective, objective and patient-reported speech assessment methods. These seem to be complementary to each other and provide different information. How these sources of information are related to each other is not yet fully understood. Future prospective studies are needed including larger samples of HNC patients.

Conclusion

Combined objective analyses of speech quality in HNC patients by acoustic-phonetic analyses and by Artificial Neural Network analyses are valid and contribute moderately to a multidimensional speech evaluation protocol. More prospective research using more phonemes and larger study samples is needed to improve performance of objective speech quality analysis.

Reference List

1. van der Molen L, van Rossum MA, Burkhead LM, et al. Functional outcomes and rehabilitation strategies in patients treated with chemoradiotherapy for advanced head and neck cancer: a systematic review. Eur Arch Otorhinolaryngol 2009 Jun;266(6):901-2.
2. Borggreven PA, Aaronson NK, Verdonck-de Leeuw IM, et al. Quality of life after surgical treatment for oral and oropharyngeal cancer: a prospective longitudinal assessment of patients reconstructed by a microvascular flap. Oral Oncol 2007 Nov;43(10):1034-42.
3. Verdonck-de Leeuw I, Borggreven PA, Eerenstein S, et al. Psychosocial and functional consequences of head and neck cancer. Ned Tijdschr Oncol 2006;3(5):185-91.
4. Borggreven PA, Verdonck-de Leeuw I, Langendijk JA, et al. Speech outcome after surgical treatment for oral and oropharyngeal cancer: a longitudinal assessment of patients reconstructed by a microvascular flap. Head Neck 2005 Sep;27(9):785-93.
5. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. Folia Phoniatr Logop 2009;61(3):180-7.
6. LaBlance GR, Kraus K, Steckol KF. Rehabilitation of swallowing and communication following glossectomy. Rehabil Nurs 1991 Sep;16(5):266-70.
7. Bressmann T, Sader R, Whitehill TL. Consonant intelligibility and tongue motility in patients with partial glossectomy. Journal of Oral and Maxillofacial Surgery 2004;(62):298-303.
8. Sumita YI, Ozawa S, Mukohyama H, et al. Digital acoustic analysis of five vowels in maxillectomy patients. J Oral Rehabil 2002 Jul;29(7):649-56.
9. Michi KI, Imai S, Yamashita Y. Improvement of speech intelligibility by a secondary operation to mobilize the tongue after glossectomy. J Craniofac Surg 1989;17:162-6.
10. Yamaguchi H, Shrivastav R, Andrews ML, et al. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. Folia Phoniatr Logop 2003 May;55(3):147-57.
11. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. J Acoust Soc Am 2007 Oct;122(4):2354-64.
12. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Neural network analysis to assess hypernasality in patients treated for oral or oropharyngeal cancer. Logopedics Phoniatrics Vocology 2011;36(4):168-74.
13. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Acoustic-phonetic and artificial neural network feature analysis to assess speech quality of stop consonants produced by patients treated for oral or oropharyngeal cancer. Speech Communication 2012;54(5):632-40.

14. Bjordal K, Hammerlid E, Ahlner-Elmqvist M, et al. Quality of life in head and neck cancer patients: validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35. J Clin Oncol 1999 Mar;17(3):1008-19.
15. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 3;85(5):365-76.
16. Lee AS, Ciocca V, Whitehill TL. Acoustic correlates of hypernasality. Clin Linguist Phon 2003 Jun;17(4-5):259-64.
17. Kent RD. Intelligibility in speech disorders: theory, measurement, and management. Amsterdam, Philadelphia: John Benjamins Publishing; 1992.
18. Bergem Dv. On the perception of acoustic and lexical vowel reduction. Berlin 1993 p. 677-80.
19. Markkanen-Leppanen M, Isotalo E, Makitie AA, et al. Speech aerodynamics and nasalance in oral cancer patients treated with microvascular transfers. J Craniofac Surg 2005 Nov;16(6):990-5.
20. Tsang CD, Trainor LJ. Spectralslope discrimination in infancy: sensitivity to socially important timbres. Infant behaviour and development 2002;25(2):183-94.
21. Klatt DH. Voice Onset Time, Frication, and Aspiration in Word-Initial Consonant Clusters. Journal of Speech and Hearing Research 1975;18:686-706.
22. Ladefoged P, I.Maddieson. The sounds of the world's languages. Blackwell Publishing; 1996.
23. Houde J, Jordan M. Sensomotor adaptation in speech production. Science 1998;(279):1213-6.
24. Praat: doing phonetics by computer. [computer program]. Version 5.2.35. University of Amsterdam: 2007.
25. Nabil N, Espy-Wilson CY. A signal representation of speech based on phonetic features. 1995 May 22; Inst. of Tech., Utica/Rome 1995 p. 310-5.
26. Nabil N, Espy-Wilson CY. A knowledge-based signal representation for speech recognition. Atlanta, Georgia 1996 p. 29-32.
27. Deng L, Sun DX. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J Acoust Soc Am 1994;(95):2702-19.
28. Erler K, Freeman GH. An HMM-based speech recognizer using overlapping articulatory features. J Acoust Soc Am 1996;100(4):2500-13.
29. King S, Taylor P. Detection of phonological features in continuous speech using neural networks. Comp Speech Lang 2000;14(4):333-53.
30. Robinson T, Hochberg M, Renals S. The use of recurrent neural networks in contnuous speech recognition. In: Lee C-H, Soong F, ., editors. Automatic

Speech and Speaker Recognition- Advanced Topics. Kluwer Academic Publishers; 1996. p. 233-58.

31. Bridle J, Deng L, Picone J, et al. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. John Hopkins University 1998 p. 1-61.
32. The IFA Spoken Language Corpus. de Nederlandse Taalunie 2001;v.1.
33. Graupe D. Principles of Artificial Neural networks. Advanced series on circuits and systems - volume 6 ed. Singapore: World Scientific Publishing Company Co. Pte. Ltd.; 2007.
34. http://nico.nikkostrom.com/ [computer program]. KTH, Stockholm: 1997.
35. Verdonck-de Leeuw I, Mahieu HF. Vocal aging and the impact on social life: a longitudinal study. Journal of Voice 2004;18(2):193-202.
36. Gorham-Rowan MM, Laures-Gore J. Acoustic-perceptual correlates of voice quality in elderly men and women. J Commun Disord 2006;39(3):171-84.
37. W.J.Hardcastle, N.Hewlett. Coarticulation: Theory, Data And Techniques. New York: Cambridge University Press; 1999.
38. Gay T. Physiological and Acoustic Correlates of Perceived Stress. Language and Speech 1978;21(4):347-53.
39. Schultz T. Speaker Characteristics. Lecture Notes in Computer Science. 4343 ed. 2007. p. 47-74.

Chapter 6

Characterization of speech pathologies in patients after vocal tract surgery and radiotherapy for oral or oropharyngeal cancer using articulatory features

6

L. ten Bosch
M.J. de Bruijn
C.R. Leemans
I.M. Verdonck- de Leeuw

Submitted

Abstract

Speech produced by patients treated for oral or oropharyngeal cancer shows deteriorated speech outcome due to loss of tissue, less flexibility of muscles, stiffened tissue, nerve damage, and velar dysfunction. The speech quality of patients after treatment for oral or oropharyngeal cancer appears to be highly dependent on tumour size and subsite: surgical treatments for oral tumour often results in articulation difficulties due to tissue loss or structure alteration of articulators, while treatment for oropharyngeal cancer often yields nasal resonance problems due to velopharyngeal inadequacy. Speech impairment often persists even after reconstructive surgery. Articulatory problems due to surgical modifications may be manifested as changes in temporal patterns of in trajectories in the acoustic space. Deviations in these patterns can be assessed semi-automatically in an objective way by comparing speech samples uttered by the patient to those realized by controls. In this paper we examined deviations of trajectories as defined by sequences of vectors with articulatory features. The comparison between the temporal structure and articulatory feature streams can be quantitatively represented as deviations from the control group on the basis of differences between the acoustic realisations of specific speech sounds. The analysis quantifies the amount of heterogeneity of a patient group compared to a group of controls.

Introduction

Every year, approximately 500.000 patients are diagnosed with head and neck cancer (HNC) worldwide. Medical treatment includes (a combination of) surgery, radiotherapy and chemotherapy. After therapy, patients often experience a range of long-term physical and social discomforts. Inflexibility and deteriorated functionality of the head and neck structures often result in speech problems leading to social dysfunction and deteriorated quality of life.[1-4] The question to what extent this deterioration can be foreseen during treatment and how the degree of speech impairment can adequately be monitored after surgery raises the importance of adequate measures to assess the quality of patients' speech. To that end, intelligibility ratings based on human judgments (annotations) are widely used.[5, 6]
The drawback of these human annotation methods, however, is their subjectivity and the dependence of a certain protocol that must be adhered in order to be able to maintain a minimum inter-rater agreement.[7] An alternative is therefore to use more objective, signal-based computational approaches to analyse speech material produced by patients after surgery. Often these computational approaches are closely related to speech recognition techniques developed in speech technology, in particular the domain of Automatic Speech Recognition.[8]

With the advent of speech technology, computational assessment methods have been developed aiming at estimating 'speech intelligibility scores'. Examples of such methods are the Dutch Intelligibility Assessment (DIA) tool[9] and the assessment tool provided by the PEAKS platform.[10] The functionality of these algorithms is based on model parameters that often need to be adjusted by using a labeled reference corpus. In general, a computational approach provides a more objective way than human transcribers to compare intelligibility of speech by patients to speech by control speakers. In addition, they provide the same functionality across different clinical working environments. An additional advantage of many computational methods is their scalability: often their performance can be improved or generalized as more annotated data become available for training, similar to training methods in ASR.[8] Windrich et al. (2008) showed that application of ASR word recognition rates highly correlated with expert listener evaluations of intelligibility, the exact correlation depending on the type of tumour.[11]

Next to methods for assessing speech intelligibility, the automatic assessment of the pronunciation quality of pathological speech received considerable interest for different patient groups (apraxia[12]; patients with oral or oropharyngeal cancer[13-15]; hearing impaired, dysarthria[16-18]; dysarthria[19-23]; pathology in general[24]). In this domain, many results have been formulated based in terms of conventional articulatory-phonetic measures such as number of words/second, phones/second, phone duration and VOT[11, 13, 14, 25], and articulatory measures.[23] Methods based on the use of conventional features provide already results that are useful to assess certain pathologies, especially of problematic phonemes such as plosives. Compared to controls, patients had significant longer Voice Onset Time (VOT) in the voiced stop consonants /b/ and /d/, and a shorter burst in the voiceless consonants /t/ and /p/. Furthermore, patients showed significant more voicing during VOT in the voiced consonant /d/ and voiceless consonant /p/, and more voicing during the burst in the voiced consonants /b/, and voiceless consonants /p/ and /t/. Within the patients group, patients with larger tumours had significant less amount of voicing during the pre-burst silence portions compared to patients with smaller tumours.[13, 14, 23]

De Bruijn et al's (2011, 2012)[13, 14] analysis is an example in which measurements were based on segments that were manually segmented from the speech material from patients and controls. The PEAKS approach[10] and the DIA tool[9] take a very different stance by relying on methods similar to those developed for Automatic Speech Recognition (ASR). ASR-based algorithms are able to transcribe an input speech signal as a sequence of symbols (in terms of e.g. words or phonemes). To that end, ASR employs statistical acoustic models for monophones or triphones that are trained on a training corpus, usually consisting of speech recordings from a large group of (healthy) speakers. In conventional ASR, the speakers in this training corpus are normal healthy speakers. In standard ASR applications, these acoustical models are used to find the best matching word sequence given a speech signal, and so to make word recognition possible. At the same time, these acoustic models could be applied to estimate statistical dissimilarity of pathological patterns from the 'normal' patterns by comparing the corresponding HMM sequences[16, 20] shows that the degree of pathology in a patient's speech can be assessed by making use of a tailored dictionary to recognize the prompted patients' utterance. Since in ASR the best performance is achieved in acoustic test conditions that match the acoustic

conditions observed during the training, deviation from normal speech due to pathology will have negative impact on the recognition performance of the ASR system. PEAKS exploits this effect by considering the word error rate of the ASR system as a measure of intelligibility of the patient's speech.
This paper presents another way of studying the problem of assessing deviant speech. It presents a method to automatically assess speech quality in terms of speech production deficiency, by using articulatory trajectories. Instead of using statistical models of speech, it uses episodic traces: explicit short stretches segmented from speech signals. Our approach models speech signal evolving over time as a trajectory in an articulatory or acoustic space, and then compares trajectories realized by patients (e.g. by uttering one specific word) with the associated trajectories by control speakers uttering the same word. By comparing different trajectories on the basis of a distance function, controls and patients can be compared. Usually healthy speakers have adequate control over a large number of subglottal, glottal and supraglottal movements that must all temporally be organised, resulting into adequate voice onset and offset, building and releasing of air pressure, and vocal tract configuration. Any deviation from this temporal organisation will therefore be manifest in deviations in terms of articulatory and acoustic trajectories. By zooming in on particular stretches along the trajectories, comparisons are possible on subword-, syllable- or phone level, as long as the corresponding segmentations are known.

In the trajectory-based approach, individual trajectories in articulatory space are the basis for analysis. In this analysis it is essential to have a clear view of the variation intrinsic to speech production. From phonetics it is well known that the acoustic realisations of a single word spoken by healthy speakers with the same mother language, recorded under noise-free conditions, show a large acoustical variation due to both speaker-intrinsic and between-speaker differences. Actually, this variation within and across speakers constitutes one of the major open problems in research on Automatic Speech Recognition. It can be expected that this type of between-speaker variation increases in the case of pathological speech production among patients treated for head and neck cancer (HNC), since HNC pathologies can be diverse, depending on the location of the surgery and reconstruction.[25] The type of problems patients experience in speech production largely depends on the type of disorder and size of the surgery area. The difficulties HNC speakers have in producing specific speech sounds can to a large extent be

explained on the basis of articulatory phonetic knowledge. Problematic cases are often plosives, the distinction between voiceless and voiced plosives, excessive nasalisation, the quality of the (stable cardinal) vowels such as /i/, /a/ and /u/, and the quality of the velars /k/, /x/ and the velar nasal as in 'ring'. While plosives need a timely and time-critical organisation of many articulators, cardinal vowels show how and to what extent articulatory target positions can be reached. HNC patients after surgery often have problems building up the inter-oral pressure that is required for stop consonants; mostly in combination with ceasing vocal fold vibration in case of the voiceless stop consonants. Tissue loss around the velum may result in poor functioning of the separation between oral and nasal cavities, leading to a nasalised sound.

The aim of this paper is to present a computational method for objectively assessing pathological speech, spoken by Dutch patients after treatment for HNC. The goal of the method is to automatically provide a phonetic assessment of a patient's speech, by quantifying the difference between acoustic realisations by patients with comparable realisations by controls.

Methodology

- **Articulatory Features**

In phonetics and phonology phonological/phonetic features represent distinctive properties of speech sounds. One of the well-known proposals has been put forward by Chomsky & Halle (1986)[26], distinguishing properties such as manner of articulation (e.g. vowel, semivowel, fricative, nasal, liquid, stop), place of articulation (e.g. labial, bilabial, velar, front, back) and voicing in a formal framework. The Compared to the Mel-Frequency Cepstral Coefficients (MFCCs) used in conventional ASR, Articulatory Features (AF) can provide more information about how speech sounds are produced, rather than represent spectral details, as MFCCs do. At least in theory AFs are therefore considered to be more appropriate to indicate deviations in speech production. Another potential advantage of an AF description is that AFs are asynchronous, that is, they allow speech parts to be classified as 'nasal vowels', which is an advantage compared to the conventional description of speech in terms of sequences of phone like symbols.
Nowadays several machine learning approaches are available to estimate AFs on the basis of real acoustic waveforms as input, including Artificial Neural

Nets (ANNs)[27] and Support Vector Machines (SVMs).[28] AFs have been used for improving Automatic Speech Recognition in noisy conditions.[29] More recently, AFs have been applied to objectively assess pathological speech via automatic phone intelligibility rating.[17, 30]

The AFs that we will use in this paper are similar to those described in Middag et al. (2010).[17] They are output from ANNs available in the NICO toolkit.[31] The ANNs take a mono waveform, sampled at 16 kHz, as input. In the first step, the waveform is represented using Mel Frequency Cepstral Coefficients (MFCCs). Via a shifting analysis window (with a shift of 10 ms), 100 MFCC frames per second are computed. In the second step, this MFCC sequence is input for the ANNs. By using the outputs of all ANNs, the entire speech signal is eventually represented as a sequence of vectors consisting of articulatory feature values (estimations). The topology of each ANN was modeled by a feed-forward network consisting of one input layer, one hidden layer and one output layer. This topology is fixed and kept the same during training and test.

Table 21 provides an overview of the different ANNs used. Each ANN corresponds to one articulatory feature, specified per row in table 2I. The name of the feature is given in the first column, while the feature values are mentioned in the second column.

Training data

The ANNs used in this study were trained on the IFA corpus.[32] This corpus is manually segmented on the word and phone level. The ANN training is fully supervised: the input of the ANN during training consists of the sequence of MFCC frames, in combination frame-by-frame with the reference feature values for a particular articulatory feature. These reference AF values were determined by translating the provided phone labelling into an AF sequence, by using a predefined canonical phone-feature translation table.

Input

The input of the ANN consisted of 7 consecutive MFCC frames, centered around the MFCC frame in question. This allows the ANN to take context into account into one vector representation, which is necessary to e.g. interpret *event* phones such as stops as a single classification unit.

Output
After training, the output of the ANN provides an estimation of the presence of the corresponding articulatory property in the input that is presented to the ANN. For example, the manner feature is modeled by the manner-ANN which has 6 units in its output layer. These six output units of the manner-ANN estimate the degree of NULL, approximant, fricative, nasality, stop, and vowel, respectively. The NULL value is a unit that takes positive values if the network is not able to positively assign values to any of the other five units, and is used where the other features make no sense, for example in the case of silence and other non-speech portions in the signal. In total, the manner-ANN classifier provides 6 values each 10 msec. All output values vary between 0 (corresponding property absent) and 1 (property present), but are not constrained to have a sum equal to 1.

The full output of the AF analysis is constructed by applying all 6 ANNs synchronously in parallel and combining the outputs of these ANNs into one 28-dimensioanl AF-vector. This vector is updated each 10msec implying that an entire utterance results in an AF matrix. Of this matrix, the number of rows is determined by the number of features (here 28), while the number of columns is determined by the duration of the input utterance.

Hidden units
One of the model parameters in an ANN is the number of hidden layers and the number of units in each of the hidden layers ('hidden' units). The larger the number of hidden units, the more complex an ANN can model a mapping from the input space to the output space. In this study, we adopted a setting that has been suggested in the literature.[27] Each ANN has one hidden layer, consisting of 300 hidden units.

Table 22 provides the accuracy of the ANNs on held out test speakers (also taken from the IFA corpus), both for the features *manner, place, voicing, rounding, front-back* and *static*. The accuracy is measured in terms of frame accuracy. In addition, a confusion matrix for the AF *manner* is presented.

Figure 11 provides an example of the behaviour of the AFs over time. The input is a wave file with a duration of 1.53 seconds. Along the horizontal axis, time is displayed (in terms of frames, frame shift is 0.01 sec). The vertical axis shows the 28 AF estimations in the order according to table 22.

From the figure, it can be seen that there are 23 frames with leading silence and about 10 frames with trailing silence.

Table 21. Overview of the ANNs used in this study to estimate the articulatory features. By aggravating all outputs, the result is a 28 dimensional feature vector, generated each 10 ms. Next to the real output values, each feature may have a 'NULL' output in the case the feature does not apply, such as in silence portions; 'nill' meaning that a value is not defined. 'Approx' means approximant; 'fric' means fricative; 'alveol' means alveolar; 'labiodent' means labiodental.

Feature	Set of output values	component
Manner	NULL-approx-fric-nasal-stop-vowel	1-6
Place	NULL-alveol-[dent-]high-labiodent-low-mid-velar	7-13
Voice	NULL-unvoiced-voiced	14-16
Front-back	NULL-back-central-front-nil	17-21
Round	NULL-nil-round-unround	22-25
Static	NULL-dynamic-static	26-28

Table 22. This table shows the performance (frame accuracy, in percentages) per ANN on independent test data (independent set of healthy test speakers, from the IFA-corpus), after training.

Manner	84.7
Place	76.7
Voice	93.5
Front-back	83.6
Round	87.4
Static	89.7

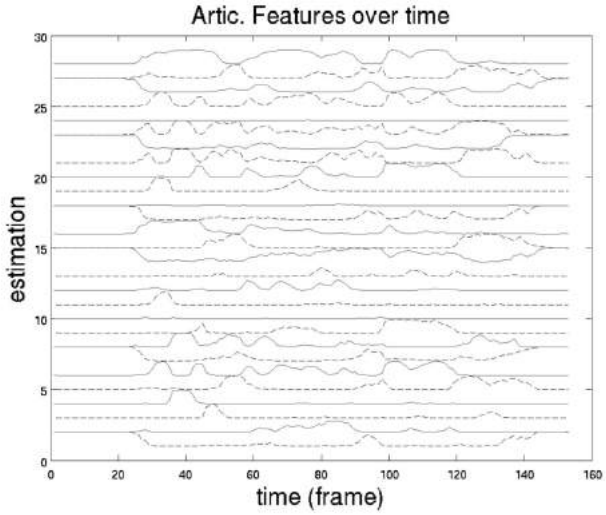

Figure 11. Articulatory feature estimation unfolding over time. Along the horizontal axis, time is displayed (in terms of frames, each frame is 0.01 sec). The vertical axis shows the 28 AF estimations in the order according to table 22. The recording starts with 23 frames with silence, and finishes with about 10 frames of trailing silence. It can be seen that during silence, the features corresponding to NULL in table 22 have a value of 1, in accordance with the definition of these NULL features as 'not applicable'. Each feature has a range between 0 (absent) and 1 (present). For the sake of clarity, feature plots are given alternating colours and line type, and feature plots are displayed with an increasing offset.

- **HNC speakers**

The analysis has been applied onto patients recorded in a database that is maintained at the Department of Otolaryngology-Head & Neck Surgery of VU University Medical Center in Amsterdam, the Netherlands. Fifty-one patients between 23 and 73 years (mean: 53.8 years, sd: 8.7 years) were included in the study after written informed consent. Furthermore, eighteen gender- and age matched controls were included (table 23). All patients suffered from oral or oropharyngeal cancer. They are member of a cohort of patients treated by surgery and radiotherapy at the Department between 1999 and 2001. Patients underwent medically successful composite resections for advanced oral or oropharyngeal squamous cell carcinoma with microvascular soft tissue transfer for the reconstruction of surgical defects. Patients received radiotherapy in the more severe cases such as an advanced (T3-4) tumour.

Exclusion criteria were incapability to participate in functional tests, difficulty communicating in Dutch and age beyond 75 years.

Table 23. Overview of gender, tumour site and stage of 51 patients included in the study

	N	%
Gender		
Male	28	(55)
Female	23	(45)
Tumour site		
Oral cavity	21	(41)
Oropharynx	30	(59)
T-classification		
2	26	(51)
3-4	25	(49)

- **Speech recordings**

Patients (6 months after treatment) and controls read-aloud a standardized Dutch text starting from the beginning of that text, until a recording time of 60 seconds was reached. The text was presented on paper and was the same for each speaker. During recording, the distance between lips and microphone was kept as constant as possible (30 cm). Speech recordings were conducted in a sound attenuated booth at the hospital. For each speaker the recording level was adjusted to optimize signal-to-noise ratio. The recorded speech was digitized with Cool Edit PRO 1.2 (Adobe Systems Incorporated, San Jose, CA, USA) to a mono signal with a sample frequency of 22 kHz (22050 Hz) and a resolution of 16-bit/sample. For each patient, the conversion of the resulting 60-sec wave file into an Articulatory Feature representation resulted in a matrix of size 28 (number of rows) times 6000 (number of columns). The 6000 column vectors correspond to the duration (60 sec) of the entire signal. These matrices are the starting point of the trajectory analysis.

- **Between-speaker alignment**

Because of speaker dependent aspiration difficulties, hesitations and pause durations, the amount of actual speech available in each 60-sec recording differed per speaker. The next step in the processing was therefore the alignment of corresponding stretches across speakers. This alignment was done by applying Dynamic Time Warping (DTW). DTW is a dynamic programming technique that allows the warping of two signals by locally stretching and shrinking; the resulting output alignment path provides the information how this stretching/shrinking must be done to minimize the alignment error. The DTW output is more reliable if the input sequences are not too long and if the two sequences can in principle be warped by a series of progressive steps (i.e. no loops).
To optimize the quality of the alignment, the DTW between two speakers was carried out on multiple substretches, instead of directly on the entire signal. These substretches were manually chosen to correspond to short word sequences, in such a way such that no restart, hesitation of word truncation appeared in any of the stretches.

In general, the resulting stretches were sufficiently pronounced for reliable alignment. For 13 out of 51 patients, however, the resulting DTW paths showed glitches due to sloppy articulation. These glitches were detected by locating significantly longer durations of horizontal and vertical alignment steps. In these cases, the automatic alignment was therefore further improved by forcing the DTW to make use of a number of marked anchor points in the signal on the basis of a manual segmentation of the plosives /p/, /t/, /k/, /b/, /d/ and the vowels /i/, /a/ and /u/ and their direct segmental pre- and post-context.

Figure 12 shows an example of the best alignment path based on the first 20 seconds of two speakers. The optimal alignment path is indicated as red line. The deviation of this path from the diagonal shows the different reading pace of the speakers.

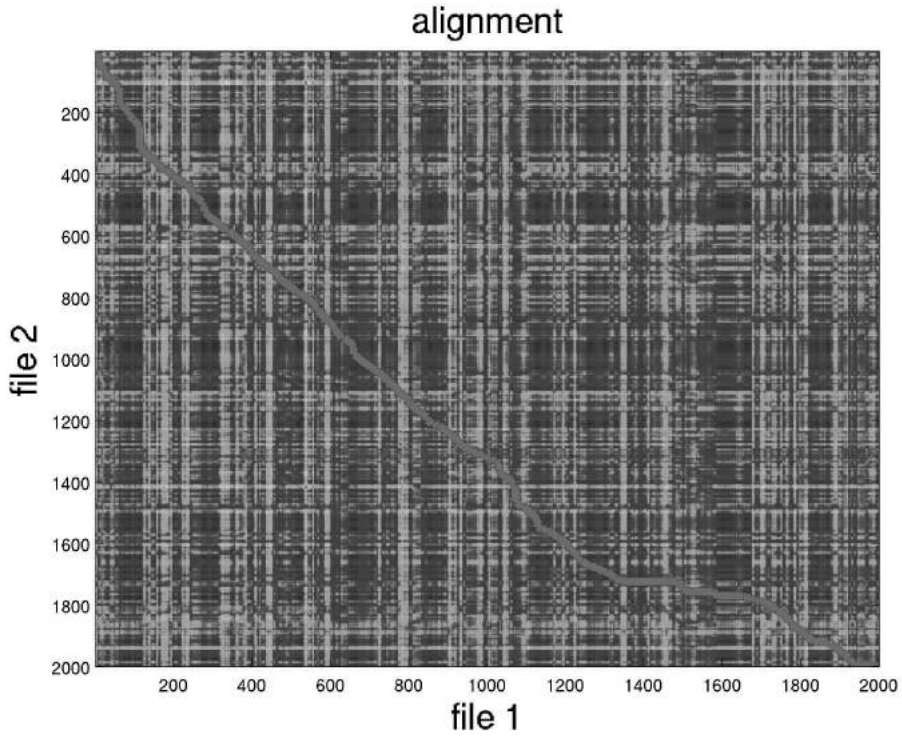

Figure 12. Example of an alignment via DTW between two recordings (file 1 displayed horizontally and file 2 displayed vertically) from two different speakers in the control group. The recording starts at the point (1, 1) in the left upper corner. Deviations from the diagonal reflect the difference in reading pace between the two aligned speakers.

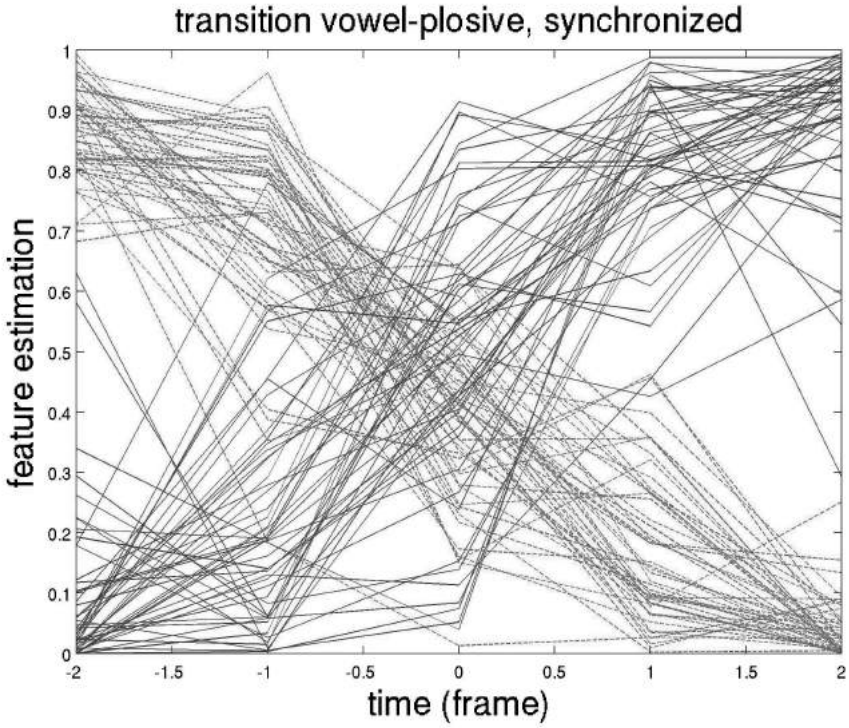

Figure 13. This figure presents transitions between vowels (dashed lines) and plosives (solid lines), represented by overlaying two of the 6 manner features (vowel and plosive), across control speakers. Time is displayed horizontally (one unit is a 10 msec step, corresponding to one frame); the feature values are displayed vertically.

After the DTW alignment, it is possible to closely compare similar stretches of speech across speakers. This comparison was done by first aligning two AF vector sequences according to the DTW alignment path, followed by selecting specific segments (stretches). Figure 13 presents an example of such a comparison, in which speech segments are shown that adhere to the following combination

{ +vowel, -plosive} - {-vowel, +plosive}

The figure presents an overlay of different 40-ms long sub-trajectories in the AF space, which correspond to transitions between vowels and plosive.

Along the horizontal axis, the time is displayed in terms of frames. The vertical axis displays the estimated AF values (where 0 and 1 mean 'absent' and 'present', respectively). The dashed (red) curves indicate the value of the *vowel*-feature varying over time, while the solid (blue) curves) indicate the estimated value of *plosive*. The overlay is based on realisations by control speakers. This figure only shows a 2-dimensional snapshot of vowel-plosive transitions; the actual trajectories are 28-dimensional.
The figure shows a clear temporal relation between the two features. There are, however, a number of transitions that deviate from the overall pattern. These deviations are due to locally imprecise AF estimations in combination with local glitches in the DTW alignment. Despite these deviations from the overall pattern, we decided to use this type of analyses as a basis for two experiments aiming at the assessment of the dissimilarity between a patient's speech and speech by a group of controls. These experiments will be described in the following section.

Experiments

To interpret anomalies of patients' speech, we performed two experiments. In experiment A we focused on two types of speech sounds that are known to be problematic for a large group of HNC patients: the plosives and velars. In this experiment, we investigate this by zooming in on transitions from vowels to plosives and from vowels to velars. These CV-combinations were chosen because articulatory deficiencies are most prominent in such transients. The analysis was done by investigating the temporal behavior of the plosive and velar component in the AF vector, after temporal alignment with the vowel feature. In experiment B, we investigated to what extent the patients group can be distinguished from the controls by quantifying the distances between trajectories in the AF space, by using the entire AF vector, instead of two specific components, as was done in experiment A.

- **Experiment A**

Method

The method is on based on contrasting DTW-associated segments from two trajectories in AF-space. First, we used the time aligned stretches based on the processing presented in the previous section. Next we collected all

stretches in the available speech material that matched a temporal articulatory feature (AF) pattern described by a feature specification

{ +vowel, -plosive} - {-vowel, +plosive}

The stretches were found by using this template on the feature values for *vowel* and *plosive* for the frames centered around time T. The selection procedure and the signal analysis were the same for controls and patients. The resulting stretches are then statistically compared and displayed visually.

Plosives

All stretches in the available speech material were collected that matched the following feature specification

{ +vowel, -plosive} - {-vowel, +plosive}

The stretches were found by using this template on the feature values for *vowel* and *plosive* for the 5 frames centered around time T, and sliding this 5-frame window along the entire AF representations. The selection procedure was the same for controls and patients. Across all speakers, this resulted in a collection of 141 short stretches of trajectories (all with the requested pattern). Of these, 122 correspond to 'real' phonetically valid transients from vowels to plosives in the speech material, the remaining 19 could be considered as 'false alarms' (false positives) and were due to incorrectly estimated feature values.
The trajectory analysis described here is based on all 141 found instances. It was decided to keep the false positives in the analysis set, with the motivation that this makes the modeling entirely data driven (that is, applicable without additional top-down knowledge) and therefore applicable in a clinical setting.

Figure 14 shows an overlay of the curves of the component *plosive* over time. Here, only the plosive component is plotted: the vowel component (which played a role in the selection of the stretches) is irrelevant here and therefore omitted. The thick lines indicate the average trajectory (upper curve for control speakers, lower curve for patients). The bars indicate the 95% confidence interval, assuming a binomial distribution per instant. The figure shows that patients (considered as a group) produce the plosives less

complete (i.e. the patient curves do not reach the plosive values reached by the controls), and that their production is slower (on average, the patient curves increase with a smaller rate than the curves by control speakers).

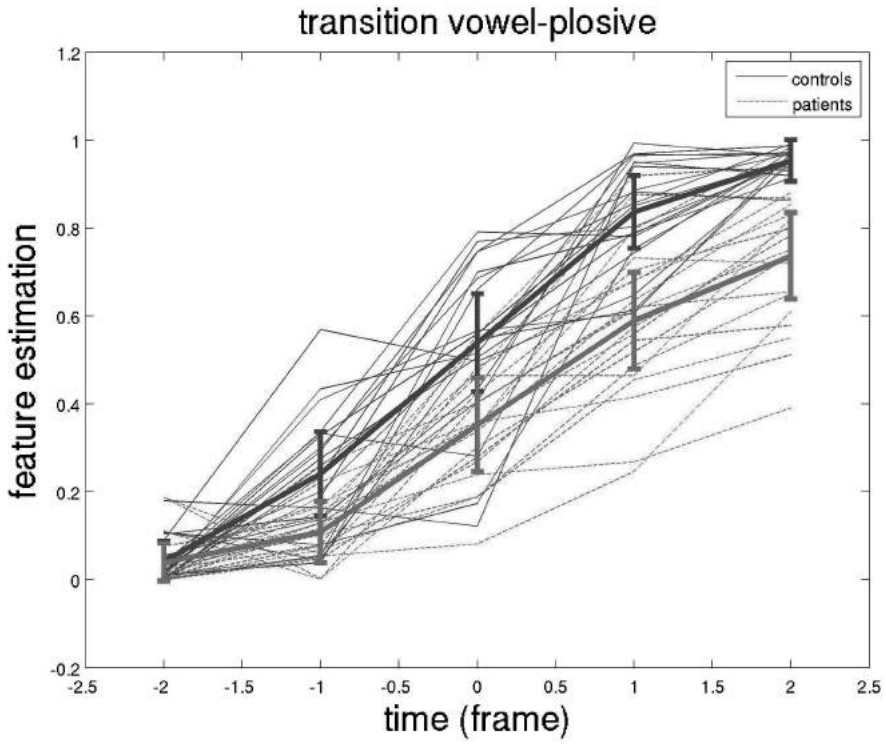

Figure 14. This figure shows the development over time of the plosive feature for vowel–plosive transitions for patients (the lower dashed lines) and control speakers (the upper solid lines). The thick lines represent group averages. The bars indicate the 95% confidence interval. For the sake of clarity, only one third of all individual trajectories are shown.

Velars

For the velars, we collected stretches that matched the pattern {+vowel, -velar} - {-vowel, +velar}. The overlay of the trajectories for the velar component in the AF vector after DTW-alignment is shown in figure 15. The figure shows that realizations can be compared against realizations by control speakers after proper DTW alignments. Similarly for the vowel-plosive transitions, figure 15 suggests that in general patients produce the

velars less complete (the patient curves do not reach the values as reached by the controls) and delayed (on average, the patient curves increase with a smaller rate than the curves by control speakers). The variation in acoustic realizations by patients shows, however, that not all patients underperform compared to the control speakers.

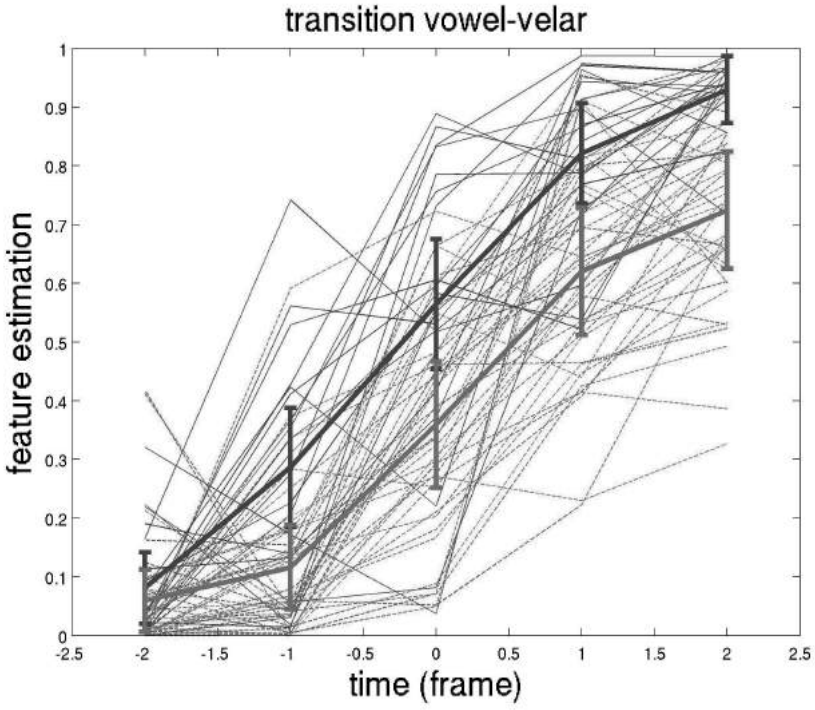

Figure 15. Overlay of the velar component in the AF representation over time, for a set of vowel-velar transitions. The patient group is represented by the dashed lines; the controls are represented by solid lines. The thick lines present the group averages; the vertical bars indicate the 95 percent intervals.

The incompleteness of the plosives and velars of patients compared to the controls, as well as the slowness of the production is in line with other findings in the literature.[13, 23] In addition, however, the analysis is based on the direct comparison of trajectories in AF-space, and provides an analysis framework to quantify this effect. The method is based on an analysis of

short specific stretches after DTW alignment, while the stretches are selected on the basis of target values of specific components in the AF vectors.
The observable variation in acoustic realizations by patients suggests that it is not necessarily the case that all patients underperform, compared to control speakers. Actually a number of trajectories from patients are indistinguishable from realizations by controls. This again supports that deviations of a patient in contrast with controls are specific for particular speech units and that a speaker might only properly be rated in comparison to a group of control speakers by contrasting specifically selected synchronized stretches of speech. This will be further investigated in experiment B.

- **Experiment B**

The previous experiment A aimed at assessing speech from patients by contrasting it to speech from controls, and by focusing on particular stretches in the speech signal that were known to be problematic(plosives and labials). The results suggested that patients are likely to form a rather heterogeneous group (more heterogeneous than control speakers), in which each individual is characterized by his/her own idiosyncrasy in terms of speech production details. This was done by comparing specific CV transients from patients and controls after DTW-alignment, and taking two features (vowel versus plosive, and vowel versus velar, respectively). Since speech pathologies may have a very individual character, on an individual basis a certain patient might be very similar to a control speaker (or a group of control speakers), while being deviant from another control speaker. This suggests that taking all features into account might lead to a more global assessment of patients versus controls.
Experiment B aims to address this by quantifying the dissimilarity between patients and controls by taking the entire AF-trajectory into account. The methodology is based on the use of a classifier to contrast one specific patient against one specific control speaker, after first training the classifier on held-out patient and control data. Since the data set is not very large, overtraining was avoided by using a leave-one-out training and test scheme.

Data

Speech samples of 51 patients and 18 controls were used in this experiment, the same speech samples as in Experiment A.

Method

In order to investigate to what extent a specific patient differs from controls, we performed a classification experiment. The task of the classifier is to distinguish patients from control speakers. The procedure is based on cross-validation, in which a classifier is first trained to discriminate the two groups, and next to specifically contrast a held-out patient and held-out control speaker by classifying them as member of their own respective group.

Because there is an unequal number of patients and controls, which may result in a classification bias, the groups were balanced during training in the following seven-step way.

[1] A specific patient *P* and control *C* were chosen. These speakers were contrasted.
[2] A subgroup of 17 patient speakers was defined by randomly choosing speakers from the patient group – this selection excluded patient *P*. This defined patient group **P**.
[3] The 17 other controls defined control group **C**.
[4] A machine learning algorithm was applied to discriminate groups **P** and **C**. The algorithm was set up in such a way that its outcome can be interpreted as a binary linear classifier discriminating two classes. It was implemented as an MLP without hidden layer. When used in the test, the resulting classifier is able to classify an unknown speaker in terms of belonging to the 'patient' or 'control' group by providing a graded membership for the group 'control' and idem for 'patient'. The membership is between 0 (absolutely no member) to 1 (absolutely a member).
[5] In the actual test, this trained classifier was applied to both unseen speakers *P* and *C*. If these speakers are highly similar, the classifier would not be able to discriminate them, even with prior knowledge about 17 patients and 17 controls. If P and C are sufficiently different, the classifier should be able to classify them both in their own correct category, with a clear margin from the class boundary.
[6] For each pair of *P* and *C*, this procedure was repeated 10 times, every time with another randomly drawn set for **P**. This reduces the risk of classification results being influenced by local minima during the optimization of the classifier to a minimum.
[7] For each patient *P* and control speaker *C*, it was determined how often the classifier was able to separate P and C by assigning them to their own group

with correct membership grades of at least 0.6 (i.e. avoiding a classification result within a margin of 0.1 from the class boundary). (This margin threshold was determined on a number of preliminary test runs; 0.6 led to very interpretable and stable results.)

The result of this analysis is shown in figure 16.

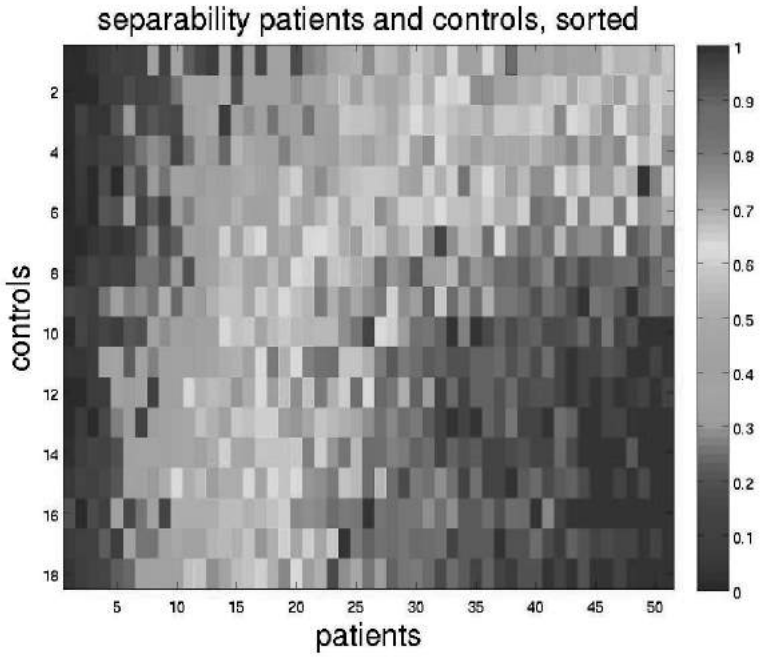

Figure 16. This figure presents a patient-to-control speaker separability matrix. Patients are arranged horizontally, while control speakers are arranged vertically. Each cell displays the probability of a certain patient and control speaker being distinguishable on the basis of the corresponding patterns in their AF trajectories. For this figure, both the 51 patients and the 18 control speakers have been sorted in such a way that the best separation (the highest score in each cell) is for the ones with the highest index (corresponding to the brightest cells, located in the area bottom-right in the figure), while the combinations of patients and controls with the worst separation is located in the top-left part of the matrix. For example, a value of 0.8 in a cell means that a combination of patient (x-axis) and control (y-axis) can be clearly distinguished in 80 percent of all classification trials. Patients situated at the left half of the matrix are less distinguishable from controls than patients located at the right-hand side.

From figure 16 a number of observations can be drawn. From the distribution of high-valued cells (cells with value > 0.8, bottom-right), it can be observed that many of the patients can be clearly distinguished from the majority of controls. Compared to the bottom-right part, the top-right part has lower scores. These lower scores indicate that there are control speakers that less distinguishable from a subgroup of patients. This suggests that these control speakers may be located closer to the group of patients than the other control speakers. The low scores at the very left part of figure 16 show that a few patient speakers actually use pronunciations that are hardly distinguishable from the control speakers. In combination, the figure shows the complex varying degree of distinguishability between patients and controls, by presenting the discriminability distribution of patients and control speakers, in particular the variation within the patient group compared to the variation in the control speakers.

To a certain extent, the position of each patient in figure 16 can be related to the pathology of each patient. In the patient database, each patient is characterized by a number of typological factors such as the tumour size (T1/2 or T3/4), the location of the tumour along the vocal tract (oro-pharyngeal or oral), and the tumour's medical basis (floor of mouth, tongue basis, tonsil, palatum molle, the mobile part of the tongue). Table 24 shows these typological factors for all patients in a top-down order corresponding to the order from left to right in figure 16. It can be seen that tumour size is likely to play a role in the ranking (10 patients with advanced tumour among the first half, 15 among the second half). Also the primary location of the tumour (third column in table 24) seems to be a factor: all patients with advanced tumour at the mobile part of the tongue figure in the lower half of the list. Overall, however, the relation is quite fuzzy – apparently the individual characteristics of patients vary largely.

Table 24. Ordering of patients in accordance with the ordering displayed along the horizontal axis in Fig 6. The typological characteristics for all patients are shown in a top-down order corresponding to the order from left to right in Figure 6. At the top, patients are ranked that are most similar to controls; patients who are clearly distinguishable from controls are ranked at the bottom of the list.

Tumour size	Tumour location along vocal tract	Primary tumour
1-2	oral	floor
1-2	oropharyngeal	tonsil
1-2	oropharyngeal	tonsil
1-2	oropharyngeal	palatum molle
1-2	oral	floor
1-2	oropharyngeal	tonsil
1-2	oral	floor
1-2	oropharyngeal	tonsil
3-4	oral	floor
1-2	oropharyngeal	tonsil
1-2	oropharyngeal	tonsil
1-2	oropharyngeal	tongue basis
3-4	oropharyngeal	palatum molle
3-4	oropharyngeal	tonsil
1-2	oropharyngeal	palatum molle
3-4	oropharyngeal	tonsil
3-4	oropharyngeal	palatum molle
3-4	oral	floor
1-2	oropharyngeal	tongue basis
1-2	oral	tongue (mobile part)
1-2	oral	tongue (mobile part)
3-4	oropharyngeal	tongue basis
3-4	oropharyngeal	tonsil
3-4	oropharyngeal	tonsil

3–4	oral	floor
1–2	oropharyngeal	tongue basis
1–2	oral	tongue (mobile part)
1–2	oropharyngeal	tongue basis
1–2	oral	floor
3–4	oropharyngeal	tonsil
1–2	oral	floor
3–4	oral	tongue (mobile part)
3–4	oral	tongue (mobile part)
3–4	oral	floor
3–4	oropharyngeal	tonsil
1–2	oropharyngeal	tonsil
3–4	oropharyngeal	tonsil
3–4	oropharyngeal	palatum molle
1–2	oropharyngeal	palatum molle
1–2	oropharyngeal	tonsil
3–4	oral	floor
3–4	oropharyngeal	tonsil
3–4	oral	floor
1–2	oral	tongue (mobile part)
1–2	oropharyngeal	tongue basis
1–2	oral	tongue (mobile part)
3–4	oropharyngeal	tonsil
3–4	oral	tongue (mobile part)
3–4	oropharyngeal	tongue basis
3–4	oral	tongue (mobile part)
3–4	oral	tongue (mobile part)

Discussion

The method presented in this paper aims at the quantification of the dissimilarity of an HNC patient's speech, on the basis of high dimensional trajectories in a space spanned by articulatory features. While articulatory features have been used in previous research[29, 33, 34] the use of *trajectories* in AF space is new. The two experiments presented in this paper present different ways to assess the difference between a patient and a group of controls. Experiment A is based on an analysis in which a predefined 'phonological pattern' is used to automatically select small stretches along the AF trajectory, after which a cross-speaker comparison can take place by zooming in on the features that are relevant for the phonetic contrast. This analysis showed that in general patients produce velars and plosives less clearly and more slowly than controls. At the same time, however, there is a quite a large variation within the patient group, which supports the findings that an HNC patient may have individual idiosyncratic speech production anomalies.
The aim of the second experiment (B) was to address the degree of anomaly by specifically contrasting a patient with a control and to investigate how well this patient can be distinguished from this control speaker using a 2-way classifier that generates estimates for graded membership of speakers as member of the 'patient class' and 'control class'. This procedure was repeated for all patients and all control speakers. The results show that both the patient group and control group can be better considered as partially overlapping clouds. Some patients showed a relatively small dissimilarity with some of the controls.

In both experiment A and B, the approach is based on analysis of (a part of) the AF trajectories. The alignment method applied here can be compared to other AF-based methods, such as Middag et al. (2010).[17] Also Middag et al. used features that are closely related to articulatory dimensions. A direct comparison with our results, however, is difficult: their target group consisted of hearing impaired speakers, and their aim was different, namely to show that a limited number of features are sufficient to get a detailed characterization of the type and severity of the articulatory problems of a certain speaker.
The use of AFs is in contrast with other methods such as PEAKS in which more conventional statistical acoustic models are applied. PEAKS makes use

of a tailored dictionary to recognize the prompted patients' utterance. Since in ASR the best performance is achieved in acoustic test conditions that match the acoustic conditions observed during the training, deviation from normal speech due to a pathology will have negative impact on the recognition performance of the ASR system. PEAKS exploits this effect and uses the word error rate of the ASR system as a measure of intelligibility of the patient's speech.
Furthermore, in contrast to ASR-related assessment methods, no predefined dictionary with phonetic pronunciations is required. In ASR, these acoustical models are used to discriminate words and so to make word recognition possible. At the same time, these acoustic models can be applied to estimate statistical dissimilarity of pathological patterns from the 'normal' patterns by evaluating the corresponding HMM sequence. Middag et al. (2009) showed that such models can be used to assess the degree of pathology in a patient's speech on a patient population with a wide range of speech production disorders.[16]

The application of AFs has a potential advantage compared to the use of a phonetic/symbolic description of speech. AFs allow asynchronous changes across the feature streams which make these features useful for assessing coarticulation phenomena and pathologies in speech due to temporal mismanagement. Anomalous production can therefore be interpreted in terms of deviations between trajectories in AF space. In the experiments we made use of this property by investigating the statistical properties of the trajectories of specific segments, aggravated over all speakers groups (patients and controls).
The results of this study are in line with general findings in other studies focussing on pronunciation quality[5, 12-14] but provide a platform for extending the scope of the analysis based on trajectories in a space defined by the articulatory features. Other, more classical phonetic features were often shown to be useful for distinguishing specific pathological realizations of phonemes to distinguish patients and controls. This type of phonetic analyses was mainly performed on specific phones.[14, 23] The AF approach in this paper extends this type of phone-based analysis by taking the entire AF trajectory and the use of all features as a starting point.

Conclusion

In this study, we explored two different methods to study articulatory patterns in speech produced by 51 HNC patients by contrasting these patterns with those from a group of 18 control speakers. The first experiment (A) in this paper is based on a comparison of specific stretches of speech that match a patterns defined in terms of these features (vowel-plosive combinations, and vowel-velar combinations). The results of experiment A show that, on average, patients do not fully reach the articulatory targets as realized by controls. They approximate these targets – on average patients reach only about 80 percent of the target values realized by controls. In addition, this approximation is slower, both for the vowel-plosive transients and for the vowel-velar transients. This objective analysis of articulatory patterns estimated on the basis of articulatory features (AF) in speech of HNC patients is feasible and valid, but leaves room for improvement to make specific conclusions about individual speakers.

In experiment B, the focus is on contrasting an individual patient to an individual control speaker, by using a classifier that attempts to distinguish the two based on two balanced independent groups of patients and control speakers. Figure 16 shows the results of a quantitative separability analysis based on AF trajectories, showing that there is a large variation in this separability measure across patients and across controls. Most patients can be clearly distinguished from most control speakers, but a few patients are hardly distinguishable from any control speaker. This result provides quantitative evidence for the heterogeneous character of pathological speaker groups.[24]

In conclusion, automatic methods for feature extraction in combination with a pattern analysis based on the alignment of episodic corresponding stretches can serve as a basis for research aiming at an in-depth, articulatory-inspired analysis of pathologies in speech.

Reference List

1. Bjordal K, Hammerlid E, Ahlner-Elmqvist M, et al. Quality of life in head and neck cancer patients: validation of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-H&N35. J Clin Oncol 1999 Mar;17(3):1008-19.
2. Karnell LH, Funk GF, Hoffman HT. Assessing head and neck cancer patient outcome domains. Head Neck 2000 Jan;22(1):6-11.
3. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993 Mar 3;85(5):365-76.
4. Verdonck-de Leeuw I, ten Bosch L, Chao L-Y, et al. Speech quality after major surgery of the oral cavity and oropharynx with microvascular soft tissue reconstruction. Proceedings of Interspeech 2007;1186-9.
5. Borggreven PA, Verdonck-de Leeuw I, Langendijk JA, et al. Speech outcome after surgical treatment for oral and oropharyngeal cancer: a longitudinal assessment of patients reconstructed by a microvascular flap. Head Neck 2005 Sep;27(9):785-93.
6. van der Molen L, van Rossum MA, Burkhead LM, et al. Functional outcomes and rehabilitation strategies in patients treated with chemoradiotherapy for advanced head and neck cancer: a systematic review. Eur Arch Otorhinolaryngol 2009 Jun;266(6):901-2.
7. Yamaguchi H, Shrivastav R, Andrews ML, et al. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. Folia Phoniatr Logop 2003 May;55(3):147-57.
8. Holmes JN, Holmes W. Speech Synthesis and Recognition. second ed. London: Taylor and Francis; 2001.
9. Middag C, Martens J-P, van Nuffelen G, et al. Dia: a tool for objective intelligibility assessment of pathological speech. 2009 p. 165-7.
10. Maier A, Haderlein T, Eysholdt U, et al. PEAKS -A system for the automatic evaluation of voice and speech disorders. Speech Communication 2009;51:425-37.
11. Windrich M, Maier A, Kohler R, et al. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. Folia Phoniatr Logop 2008;60(3):151-6.
12. Hosom J-P, Shriberg L, Green JR. Diagnostic assessment of childhood apraxia of speech using automatic speech recognition (ASR) methods. Journal of Medical Speech- Language Pathology 2004;12(4):167-71.
13. de Bruijn M, ten Bosch L, Kuik DJ, et al. Artificial neural network analysis to assess hypernasality in patients treated for oral or oropharyngeal cancer. Logoped Phoniatr Vocol 2011 Aug 24.

14. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Acoustic-phonetic and artificial neural network feature analysis to assess speech quality of stop consonants produced by patients treated for oral or oropharyngeal cancer. Speech Communication 2012;54(5):632-40.
15. Borggreven PA, Verdonck-de Leeuw I, Rinkel RN, et al. Swallowing after major surgery of the oral cavity or oropharynx: a prospective and longitudinal assessment of patients treated by microvascular soft tissue reconstruction. Head Neck 2007 Jul;29(7):638-47.
16. Middag C, Martens J-P, van Nuffelen G, et al. Automated intelligibility assessment of pathological speech using phonological features. EURASIP Journal on advances in Signal Processing - special issue on analysis and signal processing of oesophageal and pathological voices 2009;1-9.
17. Middag C, Saeys Y, Martens J-P. Towards an ASR-Free Objective Analysis of Pathological Speech. Proceedings of Interspeech 2010;294-7.
18. Middag C, Bocklet T, Martens J-P, et al. Combining Phonological and Acoustic ASR-Free Features for Pathological Speech Intelligibility Assessment. Proceedings of Interspeech 2011;2005-8.
19. Carmichael J, Green P. Revisiting dysarthria assessment intelligibility metrics. 2004 p. 742-5.
20. Mengistu KT, Rudzicz F, Falk TH. Using acoustic measures to predict automatic speech recognition performance for dysarthric speakers. Proceedings of Interspeech (MAVEBA) 2011;75-8.
21. van Nuffelen G, de Bodt M, Guns C, et al. Reliability and clinical relevance of segmental analysis based on intelligibility assessment. Folia Phoniatr Logop 2008;60(5):264-8.
22. van Nuffelen G, Middag C, de Bodt M, et al. Speech technology based assessment of phoneme intelligibility in dysarthria. International Journal of Language and Communication Disorders 2009;44(5):716-30.
23. Rudzicz F. Using articulatory likelihoods in the recognition of dysarthric speech. Speech Communication 2012;54(3):430-44.
24. Arias-Lodoño JD, Godino-Llorente JI, Sáenz-Lechón N, et al. Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. IEEE transactions on bio-medical engineering 2011;58(2):370-9.
25. de Bruijn MJ, ten Bosch L, Kuik DJ, et al. Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. Folia Phoniatr Logop 2009;61(3):180-7.
26. Chomsky N, Halle M. The sound pattern of English. MIT Press; 1968.
27. King S, Taylor P. Detection of phonological features in continuous speech using neural networks. Comp Speech Lang 2000;14(4):333-53.
28. Scharenborg O, Wan V, Moore RK. Towards Capturing Fine Phonetic Variation in Speech using Articulatory Features. Speech Communication 2007;49:811-26.

29. Kirchhoff K, Fink GA, Sagerer G. Combining acoustic and articulatory feature information for robust speech recognition. Speech Communication 2002;37:303-19.
30. Bocklet T, Haderlein T, Hönig F, et al. Evaluation and Assessment of Speech Intelligibility on Pathologic Voices based upon Acoustic Speaker Models. 2009 p. 89-92.
31. http://nico.nikkostrom.com/ [computer program]. KTH, Stockholm: 1997.
32. van Son RJJH, Binnenpoorte D, van den Heuvel H, et al. The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database. Aalborg 2001 p. 2051-4.
33. Schuster M, Haderlein T, Noth E, et al. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. Eur Arch Otorhinolaryngol 2006 Feb;263(2):188-93.
34. Haderlein T, Riedhammer K, Noth E, et al. Application of automatic speech recognition to quantitative assessment of tracheoesophageal speech with different signal quality. Folia Phoniatr Logop 2009;61(1):12-7.

Chapter 7

Associations between voice quality and swallowing function in patients treated for oral or oropharyngeal cancer

Marieke J. de Bruijn
Rico N.P.M. Rinkel
Ingrid C. Cnossen
Birgit I. Witte
Johannes A. Langendijk
C. René Leemans
Irma M. Verdonck- de Leeuw

Accepted for publication in Supportive Care in Cancer (2013)

7

Abstract

The purpose was to investigate associations between voice quality and swallowing function in patients treated for oral or oropharyngeal cancer. Recordings of speech and videofluoroscopy of 51 patients after treatment for oral or oropharyngeal cancer were analysed. Acoustic voice parameters (fundamental frequency (F0), perturbation (jitter %, shimmer %), harmonics-to-noise ratio (HNR) and intensity (extracted from the vowels /a/, /i/ and /u/)) were compared to swallowing function parameters as assessed via videofluoroscopy (oral, oropharyngeal, and total transit time, estimated percent of oral, oropharyngeal, and total residue, oropharyngeal swallowing efficiency (OPSE) and the Penetration-Aspiration (PA-)scale).

Stepwise multivariate regression analyses revealed that jitter, shimmer and HNR were not associated with swallowing function. Higher voice intensity in all three vowels /a/, /i/ and /u/ was significantly associated with a higher score on the PA-scale (more penetration and aspiration). Higher voice intensity and F0 was significantly associated with lower OPSE. Higher voice intensity was significantly associated with higher amount of penetration/aspiration, and higher voice intensity and fundamental frequency were significantly associated with swallowing inefficiency. Possible explanations may be found in overcompensation by increased laryngeal muscular strength resulting in increased intensity and pitch during phonation. However, the physiology of associated voice and swallowing function is yet unclear and more research is recommended.

Introduction

Swallowing dysfunction or dysphagia often occurs after treatment for head and neck cancer (HNC) and largely depends on tumour stage, location, and treatment modality. The normal swallowing process comprises three consecutive stages, in which the oral phase is followed by the oropharyngeal and esophageal stage. The processes within and between each of these stages are characterized by complex and accurate collaboration of muscles and nerves. Although the complex event of swallowing passes quickly and smoothly, deviations may occur at each stage of the process. Incorrect swallowing movements –at each of the stages– may have large consequences upon following stages of the swallowing process. Previous research on swallowing in HNC patients showed that swallowing impairment is commonly observed in all stages of the swallowing process with a large variation between patients. Generally speaking, patients with larger tumours experience more difficulty with swallowing. Swallowing difficulties of HNC patients that are observed in the oral stage of the swallowing process include decreased lip closure, reduced tongue volume and –motility and reduced tongue base retraction, as well as occlusion of teeth which is also important for oral swallow function. Delayed pharyngeal swallows and impaired pharyngeal constrictor motility are observed at the pharyngeal stage, while at the level of the larynx, reduced epiglottic inversion, decreased laryngeal elevation and delayed laryngeal vestibule closure are reported.[1-3] At the laryngeal level, incomplete epiglottic inversion, inadequate and late elevation of the larynx and inadequate rotation of the arytenoids cause insufficient protection of the trachea which allows the bolus to enter the airway and penetration and aspiration may occur. In case of failing closure of the airways, invasion hereof could occur in which the bolus does not pass below the vocal folds (penetration) or, worse, in which the bolus does pass below the vocal folds into the lower airways (aspiration).[4] Individuals who are subject to penetration or aspiration usually react at the impaired swallow by discharging the bolus through coughing, either as a reflex of the swallowing system or by self-stimulated action. In case of silent aspiration, the patient is not aware of entry of the bolus into the trachea and runs the accompanying risk of potential development of pneumonia.[4, 5]

A subjective method of identifying swallowing problems is clinical judgment ‘by a physician or speech-language pathologist at the patients’ bedside (bedside test). Objective evaluation of dysphagia takes place by the use of

fiberoptic endoscopic examination of swallowing (FEES) or by videofluoroscopic swallowing study (VF).[6-8] Subjective and objective investigations of swallowing do not always reflect the same results. It is assumed that subjective methods are less reliable than objective methods and research has shown that subjective assessment is not very accurate and does not detect, for instance, 40% of aspiration occurrences.[9, 10] However, disadvantages of FEES and VF include high costs of medical examination and selective availability hereof in hospitals. In addition, these examinations are invasive for patients, may include radiation exposure and are generally strenuous.

Since another function of the larynx is phonation and swallowing and voice problems often co-occur, earlier studies have examined the association between voice quality parameters and dysphagia.[11-16] The vocal folds provide the sound source for phonation as they oscillate in a series of compressions and rarefactions, modulating the subglottal pressure and transglottal flow as short pulses of sound energy.[17] Voice parameters as assessed by acoustic analyses of the voice sound signal include fundamental frequency, intensity, cycle-to-cycle pitch variability (jitter) and amplitude variability (shimmer), and the harmonics to noise ratio.[8, 18-20] The presence of (a portion of) the bolus on the true vocal cords might affect the character profile and quality of the voice. The voice may become gurgly and the patient may respond with a "wet cough" to remove the material. While some studies revealed an association between voice and swallowing,[11, 12] others did not.[13-15] In these studies heterogeneous patient groups were included varying from patients treated for neurological diseases and head and neck cancer.

The goal of the present study is to investigate the associations between voice and swallowing in a homogenous group of patients treated for advanced oral or oropharyngeal cancer. The results contribute to a better understanding of laryngeal function involved in both voice and swallowing in HCN cancer, and may help to develop a noninvasive dysphagia screening tool.

Patients and methods

Patients

Patients who underwent surgery and radiotherapy for advanced oral or oropharyngeal squamous cell carcinoma were asked to participate in the study.[21] The cohort of patients was treated over 10 years ago and only two patients received speech therapy. Surgery consisted of composite resections including excision of the primary tumour with en bloc ipsilateral or bilateral neck dissection. In case of oropharyngeal carcinomas a paramedian mandibular swing approach was used. Defects were reconstructed by a microvascular fasciocutaneous flap; no flap failures were observed. 92% of the patients received postoperative radiotherapy in case of advanced (T3-4) tumours, positive or close surgical margins, multiple lymph node metastases and/or extranodal spread. The primary site received a dose of 56-66 Gy in total (2 Gy per fraction, 5 times per week), depending on surgical margins. The nodal areas received a total of 46-66 Gy (2 Gy per fraction, 5 times a week). Exclusion criteria were incapability to participate in functional tests, difficulty communicating in Dutch and age above 75 years.

Fifty-one patients between 23 and 73 years (mean: 54 years, sd: 9 years) were included in the study after written informed consent (table 25).

Table 25. Overview of gender, and tumour site and stage of 51 patients included in the study

	N	%
Gender		
Male	28	(55)
Female	23	(45)
Tumour site		
Oral cavity	21	(41)
Oropharynx	30	(59)
T-classification		
2	26	(51)
3-4	25	(49)
Surgery	51	(100)
Radiotherapy	47	(92)

Voice recording and analysis

Patients read-aloud a text with an approximate length of 60 seconds, six months after treatment. Patients were asked to speak at a comfortable manner. The distance between lips and microphone was 30 centimeters. Speech recordings were conducted in a sound-proof cabin. The recorded speech was digitized with Cool Edit PRO 1.2 (Adobe Systems Incorporated, San Jose, CA, USA), with 22-kHz sample frequency and 16-bit resolution. Of each patient two realisations of /a/, /i/ and /u/ were extracted from running speech. These vowels are the cardinal vowels in Dutch: the positions of the articulators in these three vowels are the most extreme compared to each other. The vowels were taken from the same words for each subject. The vowel /a/ was extracted from the words /m a n/ (English: moon) and / w a t ə r/ (English: water) with mean duration of the vowel /a/ of 124 ms and 125 ms, respectively. The vowel /i/ was extracted from the words /w i/ (English: who) and /d i/ (English: that) with a mean duration of 92 ms and 83 ms respectively. The vowel /u/ was extracted from the words /m u t/ (English: must) and /t u n/ (English: then) with a mean duration of 72 ms and 72 ms respectively. These two realisations of each vowel were averaged and processed using speech processing software Praat version 4.0.28.[22] A spectrogram functioned as a visual representation of the speech signal, which facilitated recognition of phonemes in the speech signal and facilitated precise extraction of phonemes from running speech. Production of vowels is characterized by unobstructed airflow and can be produced continuously by the speaker. In earlier studies, the vowel /a/ was used as speech material.[11, 12, 14] In the present study we used the vowels /a/ and also /i/ and /u/ to compare differences.

Acoustic voice analyses were performed automatically. Acoustic variables of voice that were measured were fundamental frequency (F0), jitter (percentage), shimmer (percentage), harmonics-to-noise ratio (HNR) and intensity. Jitter represents the irregular duration of the voicing cycles (perturbation in F0) and is related to irregular trembling of the vocal cords. A high jitter means that the consecutive cycles are not identical. Shimmer represents amplitude perturbation. A high shimmer value means that the speaker is not able to maintain constant amplitude in the consecutive voicing cycles. HNR represents the harmonicity of the wave-like shape of consecutive voicing cycles and describes the amount of harmonicity relative to the amount of white noise in the signal. Intensity was measured in dB.

Videofluoroscopic recordings and analyses

Swallowing was assessed by videofluoroscopic swallowing study (VF). VF recordings were performed by modified isovist® swallowing. Isovist® is a radiocontrast substance, used to improve the visibility of body structures by X-ray techniques. A video-recorder with frame-by-frame and slow motion analysis capabilities was used for recording. During the VF, subjects stood upright and were viewed in anterior/posterior and in the lateral position. Each subject was asked to swallow two swallows each of 10 ml liquid isovist, two swallows each of 10 ml thick liquid isovist and two half-teaspoon amounts of one fourth of a cookie coated with isovist paste. Video recordings were digitized (Adobe Premiere Pro 1.5) and evaluations were performed on swallowing liquids of all patients presented in random order by two raters (an otolaryngologist and a phoniatrician) who were blinded for the clinical data.

Swallowing evaluation included oral transit time (OTT; time it takes for the bolus to move through the oral cavity), pharyngeal transit time (PTT; time it takes for the bolus to move through the pharynx) and the total transit time (TTT; the sum of OTT and PTT), oral residue (OR; approximate percent OR after the first swallow), pharyngeal residue (PR; approximate percent PR after the first swallow), total residue (TR; the sum of OR and PR) and nasal regurgitation. The oropharyngeal swallow efficiency (OPSE) was calculated by measuring the percentage of bolus swallowed into the esophagus divided by the total transit time. The OPSE scores typically range from 75 to 125 in normal subjects, meaning that 100% of the bolus is swallowed in 0.8 to 1.3 seconds. In patients, the OPSE often drops below 60, as the percentage of bolus swallowed reduces and the time increases. A lower score means a worse efficiency. In addition to the OPSE analyses, evaluations using the Penetration-Aspiration (P/A) Scale were performed.[2, 23-28] PA-scale is an 8-pointscale with a range of 1-8, a higher score means more penetration or aspiration: 1- doesn't enter airway, 2-above true vocal cords/ejected, 3-above true vocal cords/not ejected, 4- contacts true vocal cords/ejected, 5-contacts true vocal cords/not ejected, 6- below true vocal cords/ejected, 7-below true vocal cords/not ejected, 8- below true vocal cords/no effort to eject. PA-scale has an IQR of 1-7 and a median of 2. Intrarater reliability was tested and proved to be good with Pearson test- retest correlations ranging from 0.61 (OR) to 0.89 (OPSE) to over 0.95 (OTT and PTT and residue), and Spearman correlation of 0.87 for the Penetration/Aspiration (P/A) Scale. The voice recordings and the VF recordings were made at the same day but separately from each other.

Statistical analyses

Pearson correlation coefficients were used to assess the univariate associations between voice and swallowing parameters. Stepwise multivariate linear regression analyses were performed between swallowing variables PA-scale and OPSE as dependent variables and acoustic voice parameters as independent variables. The voice parameters were Log-transformed due to skewed data. A p-value <0.05 was considered to be significant. Statistical analyses were performed with the package SPSS 15.0. During statistical analyses, the demographic variables age, gender, tumour location and tumour stage were taken into account.

Results

Univariate correlation analysis

Descriptive statistics of the voice and swallowing parameters are shown in table 26. Univariate Pearson correlations (table 27) showed that various voice parameters were significantly related to swallowing parameters. Intensity as measured in the vowels /a/, /i/ and /u/ correlated significantly to almost all swallowing variables. A higher intensity is associated with a higher score on PA-scale: patients with a louder voice had a worse swallow (more aspiration). A higher intensity was associated with a lower score on OPSE: patients with a louder voice had a worse swallow (less efficiency). F0 of /a/ and /i/ were related to a few swallowing variables. A higher F0 of /a/ was related with less efficient swallowing and a higher F0 of /i/ was related to a longer total transit time.

Table 26. Mean, standard deviation, median and inter quartile range of swallowing parameters. Mean and standard deviation of voice parameters. For the variable PA-scale the value "2" represents the score 'penetration of bolus above true vocal cords and ejected'. Oral Transit Time, Pharyngeal Transit Time and Total Transit Time were measured in seconds. Oral Residue and Pharyngeal Residue were measured as a percentage, subtracted from the percentage Swallowed Bolus. F0 was measured in Hz, jitter and shimmer as a percentage, HNR (ratio) and Intensity in dB.

Variable	Mean (sd)	Median	IQR
PA-scale	-	2	1-7
OPSE	46.05 (21.03)		
Oral Transit Time	(0.39)	.53	.30 - .69
Pharyngeal Transit Time	1.03 (0.41)	.96	.80 - 1.06
Total Transit Time	1.59 (0.70)	1.44	1.25 - 1.65
Oral Residue	14.80 (17.74)	10	0 - 15
Pharyngeal Residue	19.80 (17.47)	15	10 - 27.5
Total Residue	34.59 (24.85)	25	15 - 50
Swallowed Bolus	65.41 (24.85)	75	50 - 85
F0 /a/	134.13 (27.44)		
Jitter /a/	1.52 (.95)		
Shimmer /a/	.07 (.03)		
HNR /a/	14.27 (4.10)		
Intensity /a/	59.34 (5.00)		
F0 /i/	167.25 (51.79)		
Jitter /i/	1.76 (.77)		
Shimmer /i/	.08 (.03)		
HNR /i/	17.96 (3.45)		
Intensity /i/	60.61 (5.99)		
F0 /u/	164.32 (54.13)		
Jitter /u/	1.88 (1.47)		
Shimmer /u/	.08 (.03)		
HNR /u/	17.05 (3.81)		
Intensity /u/	60.39 (5.21)		

Table 27. Overview of significant (** = p<.01; * = p<.05) univariate Pearson correlations between voice and swallowing variables. Oral Transit Time, Pharyngeal Transit Time and Total Transit Time were measured in seconds. Oral Residue and Pharyngeal Residue were measured as a percentage, subtracted from the percentage Swallowed Bolus. F0 was measured in Hz, jitter and shimmer as a percentage, HNR (ratio) and Intensity in dB.

	Intensity /a/	F0 /a/	Intensity /i/	F0 /i/	Intensity /u/	F0 /u/
PA-scale	.53**		.54**		.47**	
OTT	.36*		.30*			
TTT	.34*			.293*		
Oral residue			.33*		.41**	
Pharyngeal residue	.41**		.40**			
Residue total	.49**		.52**		.48**	
Swallowed bolus	-.49**		-.52**		-.48**	
OPSE	-.46**	-.338*	-.48**		-.40**	
Regurgitation	.33*		.33*			

Multivariate regression analysis of PA-scale

All voice parameters (F0, intensity, jitter %, shimmer % and HNR of /a/, /i/ and /u/) were entered in a multivariate regression model. Stepwise multivariate regression showed that intensity was solely significantly associated with penetration-aspiration in all three vowels /a/, /i/ and /u/ (table 28). Higher intensity in the vowels was associated with more penetration/aspiration (figure 17). Age, gender, tumour location and stage were not confounding factors.

Table 28. Prediction of PA-scale by multiple voice parameters of /a/ (R^2 = 27.2%), /i/ (R^2 = 26.6%) and /u/ (R^2 = 22.7%). ** p< 0.01. Type: stepwise. The correlation of PA-scale versus voice parameter Intensity is represented. The unstandardized beta (regression coefficient) and test statistic (t) were used.

	B	t
Intensity /a/	17.76	4.23**
Intensity /i/	14.76	4.38**
Intensity /u/	14.68	3.59**

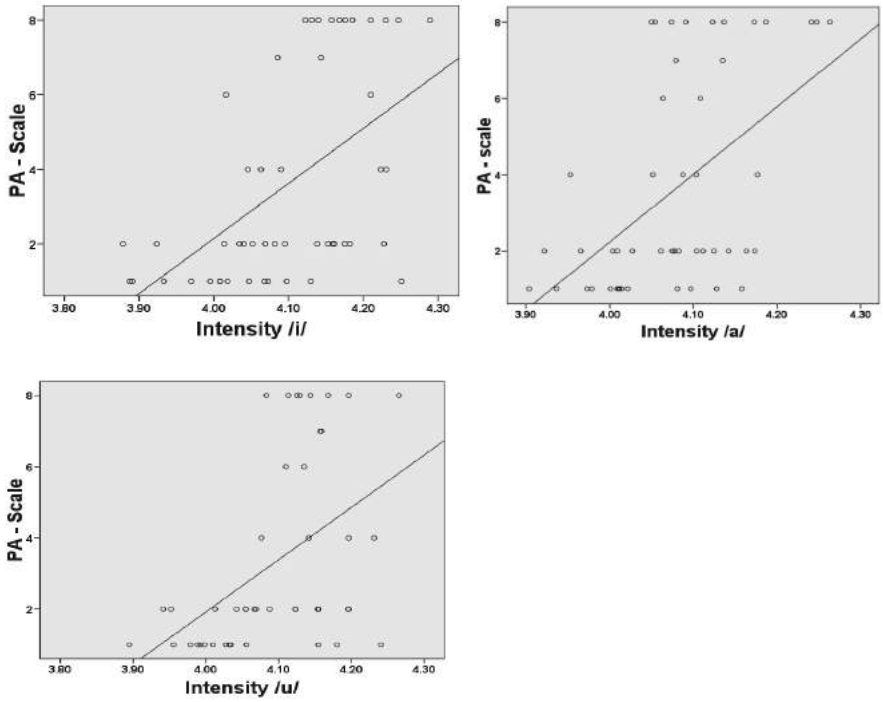

Figure 17. Graphical representation of associations between voice parameter Intensity and swallowing parameter penetration-aspiration (PA-scale). Lines represent regression lines for PA-scale. The x-axis represents the log-transformed value of Intensity; the y-axis represents the score on the PA-scale.

Multivariate regression analysis of OPSE

All voice parameters (F0, intensity, jitter %, shimmer % and HNR of /a/, /i/ and /u/) were entered in a multivariate regression model. Stepwise multivariate regression revealed that intensity was solely significantly associated with oropharyngeal swallowing efficiency (OPSE) in the vowels /i/ and /u/. Regarding the vowel /a/, intensity together with fundamental frequency was significantly associated with OPSE (table 29). Higher intensity and higher fundamental frequency in the vowels were associated with worse swallowing efficiency (figure 18). Age, gender, tumour location and stage were not confounding factors.

Table 29. Prediction of OPSE by multiple voice parameters of /a/ (R^2 = 33.7%), /i/ (R^2 = 21.5%) and /u/ (R^2 = 15.3%) ** p< 0.01. Type: stepwise. The correlation of OPSE versus voice parameter Intensity is represented. The unstandardized beta (regression coefficient) and test statistic (t) were used.

	B	t
Intensity /a/	-118.17	-3.82**
F0 /a/	-36.74	-2.97**
Intensity /i/	-94.56	-3.55**
Intensity /u/	-91.21	-2.79**

A significant difference was found for tumour location on the swallowing outcome as measured with the PA-scale. In oral cancer patients a small majority (n=11) did not have any material in the laryngeal area whereas five patients had material above the true vocal cords and five patients aspirated. In contrast, in oropharyngeal cancer patients only a very small minority (n=3) did not have any material in the laryngeal area, whereas 16 patients had material above and/or contacting the vocal cords and ten patients aspirated (data not shown).

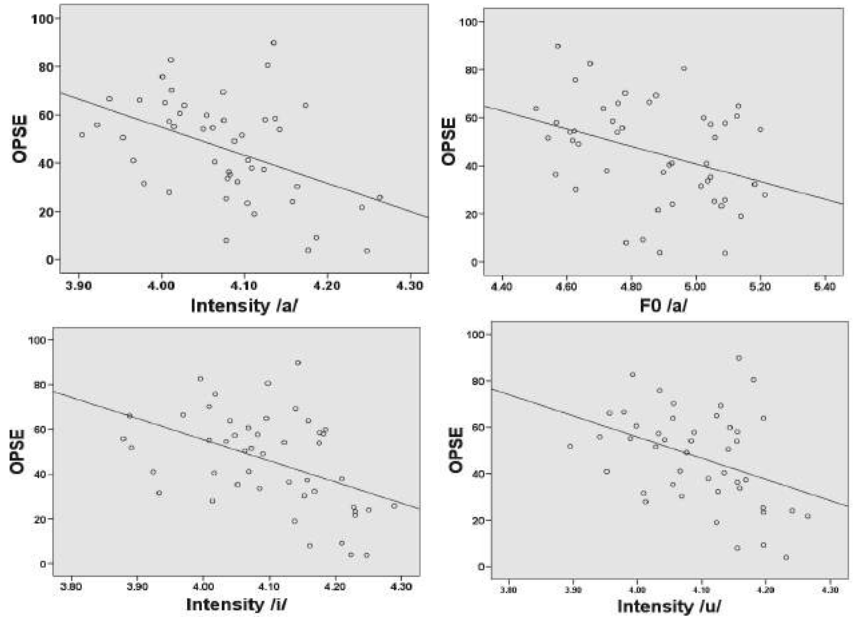

Figure 18. Graphical representation of associations between voice parameter Intensity and swallowing parameter OPSE. Lines represent regression lines for OPSE. For OPSE, for the regression line of /a/ we used the mean value of F0 /a/ for intensity and vice versa. Log-values were used instead of normal values due to scewed data. The x-axis represents the log-transformed value of Intensity; the y-axis represents the score on OPSE.

Discussion

The goal of the present study was to investigate associations between voice quality and swallowing function in patients treated for oral or oropharyngeal cancer. No associations were found between acoustic voice perturbation or noise parameters and swallowing function as assessed via videofluoroscopy. However, the acoustic voice parameter intensity appeared to be associated with oropharyngeal swallowing efficiency and penetration/aspiration: a higher intensity was associated with more penetration and aspiration and less efficient swallowing. Besides voice intensity, also higher fundamental frequency of the vowel /a/ was related to worse swallowing efficiency. A significant difference was found concerning tumour location and swallowing outcome as measured with the PA-scale: patients treated for oropharyngeal cancer have more laryngeal aspiration and/or penetration than patients treated for oral cancer. This finding is confirmed by the study of Borggreven et al. (2007).[24]

Although two earlier studies revealed significant associations between voice and swallowing parameters,[11,12] two other studies did not.[13-15] In these four studies heterogeneous patient groups were included with patients treated for various neurological diseases[12-14] and head and neck cancer.[11,15] Groves-Wright (2007) reported elevated values of jitter and shimmer when barium was visible in the larynx.[11] In contrast, Chang (2012) reported no significant changes of the acoustic signal after swallowing which results are thus in line with our findings.[15] An important difference, however, in the present study, is that the voice recordings were performed several hours after swallowing during videofluoroscopy (VF) while in the above mentioned four studies the voice recordings were made directly following each swallow during VF. From all these studies, no conclusive remarks could be made regarding the associations between swallowing and acoustic voice perturbation and noise parameters.

A new finding in the present study was that swallowing function was significantly associated with (increased) vocal intensity and fundamental frequency, both voice parameters not included in previous studies.[11-15] Although the findings were consistent in various vowels extracted from running speech, an explanation from a physiological point of view is difficult. It may be hypothesized that patients who experience worse swallowing

function may have a tendency to compensate by increased laryngeal muscular strength which results in a louder and higher tone of voice.[29] This hypothesis is supported by research on the efficacy of intensive voice treatment (the Lee Silverman Voice Training (LSVT)) of the muscles involved in voice and swallowing in patients with Parkinson Disease. Fox et al. (2012) stated in their review that LSVT leads to various physiologic changes such as increased movement amplitude of the rib cage (larger excursions) during speech breathing, increased subglottal air pressure, and improved closure and larger/more symmetrical movements of the vocal folds, improvements in orofacial movements, tongue strength and motility, speech rate, ratings of improved facial expression and improvements in some aspects of the oral phase of swallowing even though these functions were not specific targets in therapy.[30] It can be hypothesized that these improvements in both voice and swallowing are due to overall neural and biomechanical coupling of speech (and swallowing) subsystems and increased activation of the entire speech (and swallowing) neuromuscular system. De Swart et al. (2003) reported that training increased phonatory–respiratory effort had adverse effects because it raises vocal pitch and laryngeal muscle tension. They compared the Lee Silverman Voice Treatment ("think loud, think shout") and PLVT ("speak loud and low") and found that both treatments produce the same increase in loudness, but PLVT limits an increase in vocal pitch and prevents a strained or pressed voicing.[21]

In the present study on a cohort of patients treated more than 10 years ago, only two patients received speech therapy. Although it is hypothesized that untrained patients with swallowing difficulties may try to compensate by increased muscle strength leading to an increase in vocal loudness and pitch, this hypothesis needs to be tested in future research. However, in future research it is possible that the findings from the presently used cohort included in this study may differ from findings from patients treated more recently, since treatment protocols as well as supportive care regimens may have changed considerably over the last years.
A final comment on further research concerning the associations between swallowing function and voice parameters include the hypothesis that the voice alterations that were found to be associated with lower OPSE represent a coping mechanism, so this association should be studied in other groups of patients, particularly those with aspiration that were treated more recently.

Conclusion

Higher voice intensity is significantly associated with higher amount of penetration and aspiration, and higher voice intensity and fundamental frequency is significantly associated with swallowing inefficiency. Possible explanations may be found in overcompensation by increased laryngeal muscular strength resulting in increased intensity and pitch during phonation. However, the physiology of associated voice and swallowing function is yet unclear and more research is recommended.

Reference List

1. Pauloski BR, Rademaker AW, Logemann J, et al. Surgical variables affecting swallowing in patients treated for oral/oropharyngeal cancer. Head Neck 2004;(26):625-36.
2. Mittal BB, Pauloski BR, Haraf DJ, et al. Swallowing dysfunction--preventative and rehabilitation strategies in patients with head-and-neck cancers treated with surgery, radiotherapy, and chemotherapy: a critical review. Int J Radiat Oncol Biol Phys 2003 Dec 1;57(5):1219-30.
3. Lazarus CL. Effects of chemoradiotherapy on voice and swallowing. Curr Opin Otolaryngol Head Neck Surg 2009 Jun;17(3):172-8.
4. Robbins J, Coyle J, Rosenbek J, et al. Differentiation of normal and abnormal airway protection during swallowing using the penetration-aspiration scale. Dysphagia 1999;14(4):228-32.
5. J.Logemann. Slikstoornissen. / Evaluation and Treatment of Swallowing Disorders (PRO-ED, Austin, Texas). Amsterdam: Harcourt Assessment; 2000.
6. Logemann J. Manual for the videofluoroscopic study of swallowing (2nd ed.). Austin, Texas: PRO-ED; 1993.
7. Aviv JE, Kim T, Thomson JE, et al. Fiberoptic endoscopic evaluation of swallowing with sensory testing (FEESST) in healthy controls. Dysphagia 1998;13(2):87-92.
8. Verdonck-de Leeuw I, Rinkel RN, Leemans CR. Evaluating the impact of cancer of the head and neck. In: Ward EJ, van As-Brooks CJ, editors. Head and neck cancer: treatment, rehabilitation and outcomes.San Diego, USA: Plural publishing; 2007. p. 27-56.
9. Splaingard ML, Hutchins B, Sulton LD, et al. Aspiration in rehabilitation patients: videofluoroscopy vs bedside clinical assessment. Arch Phys Med Rehabil 1988 Aug;69(8):637-40.
10. Groves-Wright KJ, Boyce S, Kelchner L. Perception of wet vocal quality in identifying penetration/aspiration during swallowing. J Speech Lang Hear Res 2010 Jun;53(3):620-32.
11. Groves-Wright KJ. Acoustics and perception of wet vocal quality in identifying penetration/aspiration during swallowing. Communication Sciences and Disorders, University of Cincinnati, Ohio; 2007.
12. Ryu JS, Park SR, Choi KH. Prediction of laryngeal aspiration using voice analysis. Am J Phys Med Rehabil 2004 Oct;83(10):753-7.
13. Warms T, Richards J. "Wet Voice" as a predictor of penetration and aspiration in oropharyngeal dysphagia. Dysphagia 2000;15(2):84-8.
14. Waito A, Bailey GL, Molfenter SM, et al. Voice-quality abnormalities as a sign of dysphagia: validation against acoustic and videofluoroscopic data. Dysphagia 2011 Jun;26(2):125-34.
15. Chang H-Y, Torng P-C, Wang T-G, et al. Acoustic voice analysis does not identify presence of penetration/aspiration as confirmed by videofluoroscopic swallowing study. Arch Phys Med Rehabil 2012;epub ahead of print.

16. Linden P, Kuhlemeier KV, Patterson C. The probability of correctly predicting subglottic penetration from clinical observations. Dysphagia 1993;8(3):170-9.
17. Stemple JC, Glaze LE, Gerdeman BK. Clinical voice pathology: theory and management. third edition ed. San Diego, CA, USA: Singular Publishing Group; 2000.
18. Robert D, Pouget J, Giovanni A, et al. Quantitative voice analysis in the assessment of bulbar involvement in amyotrophic lateral sclerosis. Acta Otolaryngol 1999;119(6):724-31.
19. Pribuisiene R, Uloza V, Kardisiene V. Voice characteristics of children aged between 6 and 13 years: Impact of age, gender, and vocal training. Logoped Phoniatr Vocol 2011 Mar 30.
20. Fung K, Yoo J, Leeper HA, et al. Vocal function following radiation for non-laryngeal versus laryngeal tumors of the head and neck. Laryngoscope 2001 Nov;111(11 Pt 1):1920-4.
21. Swart de BJ, Willemse SC, Maassen BA, et al. Improvement of voicing in patients with Parkinson's disease by speech therapy. Neurology 2003;60(3):498-500.
22. Praat: doing phonetics by computer [Computer program]. [computer program]. Version Version 5.2.35. University of Amsterdam: 2007.
23. Logemann JA, Pauloski BR, Colangelo L. Light digital occlusion of the tracheostomy tube: a pilot study of effects on aspiration and biomechanics of the swallow. Head Neck 1998 Jan;20(1):52-7.
24. Borggreven PA, Verdonck-de Leeuw I, Rinkel RN, et al. Swallowing after major surgery of the oral cavity or oropharynx: a prospective and longitudinal assessment of patients treated by microvascular soft tissue reconstruction. Head Neck 2007 Jul;29(7):638-47.
25. Campbell BH, Spinelli K, Marbella AM, et al. Aspiration, Weight Loss, and Quality of Life in Head and Neck Cancer Survivors. Arch Otolaryngol Head Neck Surg 2004;(130):1100-3.
26. Pauloski BR, Rademaker AW, Logemann JA, et al. Swallow function and perception of dysphagia in patients with head and neck cancer. Head Neck 2002 Jun;24(6):555-65.
27. Rademaker AW, Pauloski BR, Logemann JA, et al. Oropharyngeal swallow efficiency as a representative measure of swallowing function. J Speech Hear Res 1994 Apr;37(2):314-25.
28. Rosenbek JC, Robbins J.A., Roecker E, et al. A penetration-aspiration scale. Dysphagia 1996;(11):93-8.
29. Van Houtte E, Van Lierde KM, Claeys S. Pathophysiology and Treatment of Muscle Tension Dysphonia: A Review of the Current Knowledge. Journal of Voice 2011;25(2):202-7.
30. Fox C, Ebersbach G, Ramig L, et al. LSVT LOUD and LSVT BIG: Behavioral Treatment Programs for Speech and Body Movement in Parkinson Disease. Parkinson's Disease 2012.

Chapter 8

General Discussion

8

Introduction

In clinical practice, speech quality is mainly assessed by perceptual evaluations by professionals such as speech therapists or evaluated by patients themselves using standardized questionnaires. Elaborate attempts have been made to develop perceptual rating instruments for voice and speech quality that also can be used to evaluate voice and speech quality of patients treated for head and neck cancer (HNC).[1-8] Invariably, the conclusion has been that without elaborate training it is only possible to obtain acceptably high agreement rates on very general characteristics of the speech (such as tempo, loudness, overall pitch, or intelligibility) and that raters attach different interpretations to more specific scales, such as nasality, hoarseness, breathiness, among others.[9, 10] Existing instruments and techniques for describing the perceptual quality of pathological speech are not powerful and accurate enough to allow wide sharing of data and information between groups of therapists and surgeons working in different hospitals. Therefore, when it comes to understanding the impact of treatment of head and neck cancer on voice and speech quality, the importance of objective quality assessment tools becomes even more relevant.

The main objective of this thesis was to investigate the role and validity of objective speech analysis methods in a multidimensional speech assessment protocol, including subjective speech evaluation by experts and by patients themselves, and objective speech analyses. More specifically, the objective methods acoustic-phonetic (AP) and Articulatory Features analysis (AF – under which the presently used Artificial Neural Network (ANN)) were used to research into what extent objective speech analyses are able to predict commonly used subjective speech evaluations.

In this General Discussion, the main results and the fulfillment of the goals of the present research are discussed. Also, methodological considerations, clinical implications of the results and recommendations for further research into the field of objective speech evaluations are discussed.

Fulfillment original goals

Two methods were investigated to investigate speech quality of HNC patients and the role and validity of objective speech analyses: Acoustic-Phonetic (AP) and Artificial Neural Network (ANN)-based speech analyses. The ANNs were applied to obtain a representation of the speech signal in terms of estimates of articulatory features. Results as described in the first three chapters showed that all phonemes distinguish HNC patients from controls: for vowels, AP analysis of /i/ and the vowel space distinguished patients from controls, while for vowels processed by ANN, /i/ and /a/ distinguished patients from controls. For velar phonemes, /k/ showed to discriminate between these populations. Finally, all stop consonants that were analyzed (/b, d, p, t/) discriminated between patients and controls and did so either by AP or by ANN analyses. Concerning differences within the patient group, fewer phonemes differentiated regarding tumour location and tumour stage. For vowels, no differences were found, for velars both /k/ and /x/ differentiate and for stop consonants only /d/ and /t/ differentiated.

These findings confirm earlier studies that speech quality of HNC patients is deviant regarding the speech features voice onset time, nasality and velar. These problems are probably caused by difficult coordination of velar and nasopharyngeal closure and often lead to hypernasality and incorrectly produced velar consonants such as /k/ and /x/. Caused by difficult production of the speech feature 'voicing', voiced consonants such as /b/ and /d/ often are produced as voiceless consonants, possibly perceived as /p/ or /t/ by listeners.[5, 11] In chapter 6 all 28 articulatory features of the Dutch language were used on the entire stretch of speech. This study revealed that besides the features 'nasal' and 'voicing' known to be affected in HNC patients, also the feature 'labio-dental' was found to distinguish patients from controls, which may be caused by (maxillofacial) surgery that may have influenced the production of labio-dental speech sounds.

When the results obtained in the present thesis are placed into a broader perspective, a direct comparison with other studies concerning classifier-based analyses is rather difficult to make. The use of Articulatory Features is relatively new (last decade), and is most often applied onto speech of healthy speakers. In studies targeting pathological speech, diseases of different etiologies were present and HNC was used in only one study.[12] In all these studies, automatic speech recognition was oftentimes used at word recognition level while we investigated speech of HNC patients at feature

level by the specific features 'nasal' and 'voicing'. Moreover, those features were in most cases calculated on specific phonemes extracted from speech. Only in case of the feature 'nasal' values were calculated on the entire stretch of speech.
Concerning acoustic-phonetic (AP) analyses, it seems that our results confirm earlier performed research. Regarding vowels, our observations of a smaller vowel space in patients than in controls and altered formant values were also reported in earlier studies.[13-15] In other studies identical findings regarding difficulty with production of velar phonemes were reported.[5, 16, 17] Our obtained results on stop consonants are also found in previous literature and concern a significant difference in duration of the voice-onset-time between patients and controls.[18, 19] It is important to keep in mind that our research was performed on Dutch speech while other studies were performed in other languages such as English and Cantonese.[18, 19]

Although the objective assessment of speech quality in patients treated for oral or oropharyngeal cancer is a relative young area of research, in groups of patients treated for other types of cancer similar work has been done. In the field of laryngeal cancer and tracheo-esophageal speech, efforts were undertaken to assess speech quality objectively. Laryngectomy means that the entire larynx had to be removed due to tumour growth. The air- and digestive ways then are attached to existing structures, separating the upper airways from the lower airways.[20] A tracheostoma is made through which the patient breathes. Insertion of a prosthesis enables the patient to produce voice and speech.
Various attempts were elaborated to analyze the nature of speech sounds produced by patients after laryngectomy as well as the correlation between perceptual evaluations and objective acoustic parameters. Especially the acoustic variables fundamental frequency, standard deviation of fundamental frequency and amplitude perturbation quotient (comparable to shimmer used in chapter 7 of this thesis) were related to subjective scales. Correlations remained moderate at best, though.[21] Although the nature and characteristics of the acoustic properties of speech quality in laryngectomized patients are not comparable to the speech quality of patients treated for oral or oropharyngeal cancer, to a certain degree the findings from research on patients after laryngectomy are somewhat comparable to our research. Results seem to follow the same direction. Results are encouraging for further research into both fields.

An overview of more literature on assessment of speech and voice produced by patients not necessarily treated for HNC is provided below.

A prospective clinical trial investigating subjective speech assessment in patients treated for advanced HNC showed that voice and speech quality deteriorate after treatment and improve in the following year, but speech and voice quality do not correspond with baseline scores.[22] This result is comparable to the presently investigated patient group. Van der Molen et al. argues that a multidimensional assessment -including objective methods- yields a more reliable assessment of voice- and speech quality.

This was done by Clapham (2011)[23] and by Middag (2012)[24]. Clapham et al. (2011) duplicated the study by Jongmans (2008)[25], in which tracheoesophageal speech was evaluated after speech therapy by human listeners. Clapham et al. used the DIA-tool for this task[26]. The Automatic Speech Recognition tool was not yet able to duplicate results as obtained from human listeners. As the DIA-tool and the human listeners were not trained in the same language, it was argued that such a tool needs to be dialect independent before it can be used in clinical practice.

It was found by Middag et al. (2012) that Automatic Speech Recognition is able to establish correlations between subjective and objective assessment on intelligibility in patients treated for HNC. They also showed that this could be done with alignment-free methods (i.e. the entire stretch of running speech without extraction of targeted speech sounds). In their recent study they concluded that the alignment-free method is independent of language and that the alignment method is language-dependent. It is proposed that a combination of both methods may be able to reach human perception rates and that the combination of both methods is able to detect progress or deterioration of speech quality within one patient.[24]

On a dysarthric patient group, different automated speech recognition methods were used to investigate intelligibility scores, grouped into the underlying systems of MFCC (acoustic features) and of articulatory features. It was concluded that the alignment-based approach is more reliable than the recognition-based approach, but more research is needed in order to provide clinical users useful information.

Methodological considerations

In this section methodological considerations and limitations are discussed. First considerations concerning patients are discussed, followed by reflections concerning speech processing. Finally the usability of acoustic-phonetic analyses and analysis by and Artificial Neural Network is discussed.

- Considerations concerning patients

Patient cohorts

Both cohorts were set up in identical ways. Patients treated for HNC were asked to participate in this study when visiting the hospital for routine control. The first cohort consisting of 51 patients (used in the three pilot studies described in chapters 2 – 4) was representative concerning gender, tumour subsite and stage. The second cohort consisting of 64 patients (used for external validation in chapter 5) was also representative concerning gender, tumour subsite and stage. The main difference between the two cohorts was the timing of the speech quality assessment. While the assessment time point of the cohort consisting of 51 patients was 6 months after treatment (short-term assessment), the assessment time point of the other cohort (64 patients) varied from 6 months post treatment to 9 years post treatment with a mean of 27 months (short to long term assessment).

This large difference in speech assessment timing (short term versus mid-term follow up) may have had an influence on speech production by patients and explain the two different predictive models. Although not investigated, inspection of the recorded speech samples during segmentation gave the impression that speech quality of patients at midterm follow up in the second cohort was better compared to speech quality at short term follow up. A first explanation may be that the second cohort included relatively more cancer survivors compared to the first cohort. Cancer survivors in general have better quality of life compared to patients that are likely not to survive.[27] It may be that speech quality is also more often negatively affected in non-survivors compared to survivors. A second argument may be that patients who survived for multiple years have had more time to adjust to an altered vocal tract and develop strategies to speak more intelligible. Speech therapy may have had a positive influence on maintaining voice and speech functions after treatment for HNC.[28-30]

The two cohorts consisting of 51 and 64 patients are relatively small for the purpose of the present study, as is the group of control speakers consisting of 18 speakers. During further development of the speech evaluation protocol larger study cohorts are needed to get more insight into speech quality, individual phonemes and further validation. Larger groups of patients may be -less than smaller groups- prone to outliers and provide a more general insight into the characteristics of speech after treatment for HNC. Maybe even reliable smaller subgroups concerning tumour stage and -location, treatment techniques and time since treatment can be made and consulted to discover patterns in separate speech sounds.
Larger cohorts are also recommended in the perspective of reliable statistical analysis. A serious drawback of this study is the limited amount of patients related to the large amount of predictors, which leads to instable results.
If the random sample is too small for the number of predictors it is possible that overfitting of the model occurs, resulting in a low predictive value. This is especially problematic when new data is introduced, as was the case in the validation study in chapter 5. Also the amount of 28 speech features is rather large for a population of 51 patients, taken into account that only relatively few speech sounds were segmented with different phonological surroundings. This drawback is not of influence when the entire speech recording is used which consists 100 frames per second, totalling up to a much larger amount of data.

Similar to this item is the related problem concerning collinearity. In regression analyses there can be (inter)dependence between predictors. Due to collinearity it is more difficult to establish the role of one single predictor.
One last drawback on the composition of both patient cohorts is that both cohorts are built from patients with different tumour locations and tumour stages. This means that -although cohorts of 51 and 64 patients respectively are rather large compared with literature- the cohorts are composed of smaller subgroups. It is questionable if the number of patients in each of the sub groups is able to give enough weight to the large amount of variables (speech sounds) in the studies.

Patient-reported speech outcome
Objective speech analyses did not succeed very well at predicting subjective patient-reported outcome as measured by the EORTC QLQ H&N-35 Speech Subscale. This finding may be explained by the content of the Speech

Subscale. This scale is based on only three items: “Have you been hoarse?”, “Have you had trouble talking to other people?”, and “Have you had trouble talking on the telephone?” Hoarseness was a voice quality problem often reported by patients treated for laryngeal cancer and not by the patients included in this study that were treated for oral or oropharyngeal cancer. The other two items were on the impact of impaired speech quality in daily life rather than focusing on speech quality as such. This means that the usefulness of this Speech subscale is questionable. A suggestion could be to ask patients to rate the quality of their speech with a marking score. Another option is to use a more elaborated questionnaire on speech quality. A recently published questionnaire concerning patient-reported outcome of speech quality is the Speech Handicap Index (SHI).[31] The SHI consists of 30 items and is a valid tool for assessing speech problems that patients experience. The SHI was not used in the present dissertation because it was not yet available at the time of data collection. However, no validated questionnaire exists that examines speech production into details such as individual phonemes. If needed for future goals, more research into the development of a detailed patient-reported speech outcome questionnaire is recommended for investigation of subjective articulation, intelligibility, nasal resonance and possibly more aspects of speech production.

- Considerations concerning speech material

Running speech

The analyses of phonemes in this dissertation were performed on recorded running speech. Patients read aloud a standardized Dutch text that was (and still is) used in clinical practice at the department of Otolaryngology / Head & Neck Surgery of VU University Medical Center for at least fifteen years. The advantage of this approach is that an identical text was recorded for all patients. A disadvantage is that it is not a phonetically balanced text which makes it sometimes impossible to analyze a specific speech sound because the targeted speech sound is absent in the text. Another disadvantage of using read aloud text as speech material is that also the effect of reading out loud is measured. Speakers who are not skilled readers are likely to speak less fluently and -for instance- to use incorrect emphasizing or to prolong phonemes undeserved. Speech is the result of complex and asynchronous coordination of speech organs. Individual phonemes in running speech are oftentimes affected by neighboring phonemes (phonological context).

Phonemes are influenced by speaking rate, pattern of emphasis of syllables and phonological context (coarticulation and assimilation) in which the spectral characteristics of the phonemes are influenced by surrounding phonemes. It is important to keep these aspects in mind when interpreting the results.

However, there are ways to avoid these interferences, such as measuring spectral and feature analysis at midpoint of the speech sound –assuming that the influence of neighboring speech sounds is minimal at midpoint– as was done for formant analysis of the vowels. For stop consonants, the average of the phonological speech feature voicing was determined. Another option – not used in the present study– is to avoid coarticulation and assimilation by using phonemes produced in isolation. An important first step in carrying out objective research of phonemes is thus the extraction of phonemes from the original speech recording. It is often not easy to identify the exact boundary between phonemes, because it is common that the boundary between two phonemes is rather a gray area than a distinct line which is caused by fluent movement of diverse articulators.[32] It is already admitted in the early literature that segmentation "involves a great deal of human judgment" (Peterson & Lehiste, 1960:694).[33] A complicating factor in selecting boundaries between phonemes in this thesis is that speech of HNC patients is oftentimes less well-articulated and less intelligible than speech produced by healthy speakers. This makes it even more difficult to establish an unambiguous distinction. In this thesis we used the computer program Praat[34] to perform segmentation of phonemes. This program enables precise extraction by the use of a spectrogram which is a visual representation of the speech signal. During segmentation also audio recordings were used to support the decision-making. According to the recommendations by the founders of the IFA corpus[32], the decision for a boundary is taken by detection of the most explicit change in the speech signal; either in waveform or in a spectral change. In case of multiple potential boundaries the human ear is used to take the final decision in which the target speech sound is as pure as possible. The founders of IFA corpus also reported that it is important to work as systematically as possible and that the segmentations should correspond with 'audible differences': in this dissertation the decision for segmentation was determined by a combination of the spectral balance, the wave form and the human ear.[32]

However, in Peterson & Lehiste (1960:694) it was mentioned that "instrumental accuracy is in general considerably greater than the accuracy with which the segmental boundaries can be determined".[33] Especially transitions from consonants to vowels are characterized by an overlap of speech cues so it is not useful to determine the exact segregation between phoneme boundaries.

The speech sounds that were selected in the present research are from different 'classes' (i.e. vowels, velar speech sounds (of which a fricative /x/) and stop consonants) and from different locations in the word. There is a variety of initial and final plosives and a variety of speech sounds surrounding vowels. These characteristics –as well as the voiced/unvoiced distinction– need to be taken into account when performing the segmentation in the speech processing program Praat.[34] In languages such as English, a word-initial voiceless consonant is followed by aspiration. It is necessary to find consensus on how to perform this segmentation. It was found in literature that the length of aspiration after /p, t, k/ differs between those voiceless stop consonants. In conclusion, it was seen in literature that the 'center of the syllable' (syllable nucleus) is affected by surrounding speech sounds. Especially consonants that follow the syllable nuclei is of influence of the duration hereof. When followed by a voiceless consonant, the syllable nucleus is shorter than when the syllable nucleus is followed by a voiced consonant.[33] Applied onto our data, this could imply that –for the vowel /u/– the duration of /u/ may be affected by the following consonants in /m u t/ and /t u n/. This is also the case for the vowel /a/, being extracted from /m a n/ and /w a t e r/.

Phonemes

From literature is known that patients treated for HNC have difficulties producing velar phonemes, stop consonants and the speech feature hypernasality.[5, 35] Hypernasality is especially audible in vowels because vowels are produced without frication or occlusion, giving a strong signal with little noise. In this thesis these phonemes were investigated by using objective assessment methods. The first step was to select two appearances (realizations) of each speech sound in the standardized text. In case of the vowels /a/ and /i/ there was an abundant amount of choices possible in the text while the vowel /u/ and velar phonemes /k/ and /x/ only occurred twice in the text. Therefore, for some phonemes we were forced to select the only

two occurrences in the text while for other phonemes argumentation and selection was needed.

The argument to use two instead of one token was the generalizability of the phonemes and the results. As the neighboring phonemes are of large influence of the target speech sound, it is imperative that different surroundings are used. Given the explorative nature of this thesis two realizations were used to obtain insight into the characteristics of speech production by patients treated for HNC. However, it is arguable if two realizations of each phoneme provided enough information for solid conclusions. In further investigation toward validation of phonemes, multiple tokens should be included in the design with different phonological contexts. Also, in case of investigation of (other) articulatory features by ANN a specific text containing different phonological contexts could be useful. For instance, to investigate the articulatory feature 'nasal' a specific text containing nasal phonemes may provide more insight into the behavior of ANN.
However, it was frequently observed that only one of the two realizations of a speech sound was significantly different between patients and controls. This means that the presently investigated speech sounds are not yet applicable for a screening method. More investigation is needed onto the nature of selected speech sounds, their neighboring speech sounds and the actual pronunciation. Within one language, multiple dialects exist and are of influence on production. It is well possible that we have to account for these phenomena before the speech sounds are able to contribute to a screening protocol.

Language
Production of phonemes are language specific and depend on rhythm and stress pattern as well as on characteristics of the speaker.[36] [37] In this thesis, the focus was on one specific language, Dutch, which implicates that the results cannot be generalized to other languages because of differences in pronunciation of certain speech sounds. An example is /s/ –not examined in this thesis– which is sharper in English than in Dutch. This means that characteristics of speech sounds are not automatically transferrable across languages Future research is needed including other languages and investigating whether speech difficulties encountered by HNC patients are language specific.

- Considerations concerning speech processing

Artificial Neural Network

In this thesis we aimed to assess a speech signal (derived from patients treated for HNC) in terms of estimates of articulatory-acoustic features (such as voicing, nasal, labial, dental, etc). This representation makes it possible to assess the quality of the speech in terms of measures that in principle can be related to the speech production stage. The signal-to-feature mapping can be achieved by applying machine learning approaches. These approaches are capable of finding such relations after training the model parameters on a speech corpus that has been labeled and segmented in terms of phones, from which articulatory features can be derived.

Within the field of machine learning, several techniques are available. Among these, ANN is a well known approach, and applied since about a decade.[38-40] In general, ANN is particularly useful to model complex connections between a groups of 'input' variables and another group of 'output' variables, e.g. acoustic parameters on the one hand [input] and quantified articulatory parameters on the other hand [output]. Because of their properties, ANNs can serve as an essential step in various classification problems. By applying the classifiers on large corpora, it is possible to find trends in large data sets that otherwise would remain unobserved.

In speech research, much work with ANN done so far focused on an analysis of speech of healthy individuals.[38, 41, 42] In this study ANNs were used to assess pathological forms of speech. These ANNs have been applied to automatically find estimates for features mentioned in table 1, in which six ANNs were used. Each of these six ANNs had identical architecture: every ANN had one hidden layer with the same number (300) of hidden units. The function of the hidden layer was to increase the complexity of the training and enhancing its ability to detect speech characteristics. Some ANNs had more than one hidden layer which may have lead to more detailed computation through a more complicated training strategy. However, for the exploring nature of this thesis, the ANNs with one hidden layer that were used showed that ANNs were able to distinguish speech produced by patients from healthy speech.

Each ANN corresponds to a group of features specified in each line of the

following table (see table 30). For example, the first ANN deals with ‘manner’ and provides estimates for six different outputs or vectors: nil, approximant, fricative, nasal, stop, vowel. NIL means ‘not applicable’ and is a convenient vector to use in case when the input is silent.

Table 30. Overview of the ANNs used in this study to estimate the articulatory features. By aggravating all outputs, the result is a 28 dimensional feature vector, generated each 10 ms. Next to the real output values, each feature may have a ‘NULL’ output in the case the feature does not apply, such as in silence portions; 'nill' meaning that a value is not defined. 'Approx' means approximant; 'fric' means fricative; 'alveol' means alveolar; 'labiodent' means labiodental. The six features (ANNs) and their divisions into vectors, counting up to a 28 vector-model.

Feature	Set of output values
Manner	NULL-approx-fric-nasal-stop-vowel
Place	NULL-alveol-[dent-]high-labiodent-low-mid-velar
Voice	NULL-unvoiced-voiced
Front-back	NULL-back-central-front-nil
Round	NULL-nil-round-unround
Static	NULL-dynamic-static

This table is motivated by King.[38] Since then, other researchers have used a slightly different feature organization. For instance, in table 1 the vectors ‘high’ and ‘low’ are part of the feature Place but it is possible to consider these two vectors separately and to divide the place feature into two sub features that were trained on these specific characteristics of speech production.

Apart from ANNs, several another classifiers exist such as the Weka database[43], in which other types of classifiers were assembled. The disadvantage of using the presently used ANN[44] is that it is not always easy to interpret the results in the perspective of the (biological) cause. Speech was classified in a certain class (the features nasal and voicing in this dissertation) but it does not provide a direct explanation why those results were obtained. There is no relation with the physiological reality -these are

only estimations obtained from the speech signal. The argumentation hereof has to be performed afterwards and falls into the realm of subjective interpretation. Future research in the field of pathological speech may be performed by ANNs with more than one hidden layer or by another type of classifier.

The applicability of the presently used ANNs was not yet clear although results seem promising. Much research focused on obtaining insight in the behaviour of ANN on healthy speech[45-48] but since ANN is a relatively new technique research with pathological speech[26, 49-52] was not as often performed as with normal speech. However, pathological voices were assessed more often by the use of ANN.[53-55] Another issue of the applicability of ANN is the composition of the pathological speaker group. Within the small amount of studies performed on pathological speech, there are hardly any studies concentrating on speech produced by patients treated for HNC.[26, 49-52] In future research more work should be done on the behaviour of ANN applied onto groups of pathological speaker, preferably larger groups of patients treated for HNC. The presently used ANN was trained on speech of only two healthy speakers, male and female. Although this number was enough for assessing the ANNs for speech of the same speech type from healthy speakers, This number is likely not enough to guarantee generalizability across pathological speech types. Additional speakers used for the training stage of ANN are recommended to account for a variety of individual speaker styles and more detailed results provided by ANN. The issue how to optimally train classifiers in order to functionally assess a broad range of pathological speech types is actually still unanswered.

Acoustic-phonetic analyses

Next to the relative new technique of ANN, we used the longer existing objective method of acoustic-phonetic (AP) analyses in this thesis. AP analysis of speech produced by patients treated for HNC is frequently used in literature.[6, 13, 56-61] Despite the longer tradition of AP analysis, there are some drawbacks of this objective method that need to be acknowledged.

Before the actual analyses can be performed, segmentation of the target phonemes from running speech is needed which demands punctuality and is time consuming. Unlike ANN where multiple features quantify the entire stretch of speech automatically, AP parameters are phoneme dependent. For

example, there is no use measuring the spectral slope on running speech. This means that phonemes must be extracted from running speech before AP parameters were measured. One solution is the use of speech in isolation where the speaker produces, for instance, an extended vowel /a/. However, results from measuring speech quality based on extended vowels instead of running speech hampers generalizability in daily life.[60]
A second burden of AP analysis was the large amount of outcome measures. There are many variables to choose from when researching phonemes. For instance, relevant variables for vowels include duration and formant values. There is not an unambiguous set of parameters for each phoneme. This means that in literature sometimes a direct comparison is not possible. It is also not clear which variables are considered 'best' to measure the quality of speech. Finally, as is the case with all estimation techniques, there is no direct relation with the physiological reality. Only estimations obtained from the speech signal were acquired but no direct insight to the anatomical and physiological factors of speech production can be provided.

One last comment on the usefulness of the use of formants has to be made. In this thesis formants were measured in Hz. It would have been better if log(Hz) of bark(Hz) were used for perceptive calculations. The disadvantage of Hz is that it gives a distorted perspective on the actual distance between the data. An example from the present study is the second formant of /i/ which has relative high formant values. This problem could be avoided by using log-values. This is a recommendation for further research.

Clinical implications

The implication of objective speech analyses methods into a multidimensional speech assessment protocol can be found in clinical practice. The development hereof provides clinicians as surgeons, radiation oncologists or speech therapists a tool that gives accurate information on different aspects of speech in patients treated for HNC. As a result, the functional speech status of patients can be registered objectively and may provide additional information next to the patient-reported overall quality of life. Also speech and swallowing rehabilitation may become more personalized when problems are recognized. The impact of the application can be found in easy accessible devices for screening of speech-, voice- and swallowing problems. Objective speech analyses may also be less costly than other diagnostic tests and are not bound to the location of the clinic.

There is a variety of intermediate products of a multidimensional speech assessment protocol. First, a full range protocol may be used in hospitals and is combined of subjective and objective methods to assess speech quality. Subjective function tests are the 'Functional Rehabilitation Outcome Grades (FROG)[62] and patient-reported questionnaires concerning speech, voice and swallowing.[15, 31] Objective function tests include evaluation of nasal resonance by the Nasometer[63] and AP analysis and ANN analysis of speech.

A specific intermediate product includes the development of a speech test by telephone. Patients record a stretch of speech over the telephone which is directly automatically evaluated by the Artificial Neural Network. Speech input then exists of phonetically well-balanced sentenced and phonemes in isolation. If the values as calculated by ANN reach a certain threshold, patients are referred to the hospital for follow-up tests. During follow-up the full range protocol can be used to determine the nature of the voice- and speech problems. This specific use of a home test possibility could also be performed by an app for use on smart phones. Another possibility of an intermediate product is the clinical implication of an internet intervention for automatic assessment of speech samples as proposed by Maier et al. (2006).[64] In literature, the development of such an application has been described before.[65] In 1986 a first attempt was performed on speech of laryngectomees[66] by subjective evaluation of intelligibility over the telephone. After repetition of that study with the use of automatic methods[67],

it appeared that the results indicated that only one subjective rater was not enough and reliable results were absent because of variation between raters. Haderlein et al (2011)[68] investigated the possibility to evaluate speech over the telephone by automatic speech recognition (ASR). They reported a high correlation between objective evaluation of intelligibility and subjective assessment by human listeners. According to Haderlein et al. (2011) 'objective measures help to reduce costs' and the use of an algorithm is 'a method which is independent of a particular therapist's experience'.[68] They also mentioned the fact that the method 'can easily be transferred onto other types of pathology'. Another study on the feasibility of ASR in patients treated for HNC was performed by Maier et al. (2010)[12] including 41 laryngectomees and 49 patients treated for oral cancer. The results indicate that the quantification of intelligibility by ASR performed equally well as was done by human raters.

In this thesis also a study on associations between voice parameters and swallowing function was described. The subjective method of identifying swallowing dysfunction is a clinical judgment' by a physician or speech and swallowing therapist performed at the patients' bedside (bedside test). Objective methods are fiberoptic endoscopic examination of swallowing (FEES) or by videofluoroscopic swallowing study (VF).[69, 70] Disadvantages of FEES and VF include high costs of medical examination and selective availability hereof in hospitals. In the perspective of patients, these examinations are invasive, may include radiation exposure and are generally strenuous. Subjective and objective investigations of swallowing do not always yield the same results. It is assumed that subjective methods are less reliable than objective methods and research has shown that subjective assessment is not very accurate and does not detect, for instance, 40 % of aspiration occurrences.[71, 72] In the perspective of strain for patients and of medical costs, an automatic and objective protocol for the detection of dysphagia is useful. The findings in this thesis investigating the validity of voice analysis as screening tool for swallowing dysfunction (i.e. penetration and aspiration and oropharyngeal swallowing efficiency) suggested that a louder voice was significantly associated with a higher amount of penetration/aspiration and that a louder and higher pitched voice were significantly associated with swallowing inefficiency. It was concluded that acoustic voice analysis possibly is related to swallowing dysfunction in HNC patients and that this finding may contribute to the development of a

screening tool. However, the preliminary findings request more research into the explanation of a worse swallowing function related to louder voice intensity in HNC patients before such a protocol is useful for clinical practice.

Conclusion and recommendations for future research

The overall conclusion of this thesis is that objective AP and ANN speech analysis methods are feasible and valid in a multidimensional speech evaluation protocol, and that voice analyses may be used as a screening tool identifying swallowing dysfunction in HNC patients. Results contribute to further development of a multidimensional speech evaluation protocol to be used for research purposes and clinical practice. However further research is warranted.

Based on the findings in this thesis, an objective speech quality test can be developed including AP and ANN analyses. The "Dutch Intelligibility Assessment" developed by De Bodt et al. (2006)[73] may be adapted for this purpose. Speech material should comprise at least the features nasal, voicing, velar and labio-dental, and accompanying phonemes including the vowel space. It is useful to investigate other phonemes as well, preferably phonemes that have their place of articulation in common with locations that are oftentimes hampered by cancer and its treatment such as the tip of the tongue. Phonemes that could be affected in this case are alveolar speech sounds such as /s, z, l, r, n/. From the present research was already known that other alveolar speech sounds like /d, t/ are affected. Other phonemes that could be nominated for investigation are phonemes in whose production the velum is involved such as /ŋ/. Another option for investigation is the transition between phonemes because the movement of articulators is involved during production. This could be measured by ANN. To rule out the influence of the phonological context speech in isolation could be recorded and processed.
Concerning the training of the ANN, the presently used ANN was trained on speech of only two healthy speakers. For more detailed training enabling detection of more information, ANN could be trained on speech of multiple speakers to account for speaker style variation and pathological speech production. Future results may become more detailed as result of extended training, also when other types of classifiers are used, such as classifiers consisting of more hidden layers and hidden units. These are to be found in the data-mining database of Hall (2009).[43]
Concerning the association between a louder voice and swallowing dysfunction, more research is needed to confirm this finding and to

investigate the physiology of voice, speech and swallowing function in more detail.

Besides objective speech analyses, subjective speech evaluation and patient reported speech outcome, a multidimensional speech assessment tool may also involve an oral function test such as the Functional Rehabilitation Outcome Grades scale (FROG), The FROG has been developed as a structured and easy to use clinical assessment tool to assess oral function[62] and comprises seven subscales: shoulder, mandible, teeth, lips, tongue, oropharynx and saliva. Incorporating the FROG into a multidimensional speech (and swallowing) assessment tool will give insight into the association between anatomical and physiological factors of speech production, objective and subjective speech quality and patient reported speech problems in daily life.

Although more research is needed, a multidimensional speech assessment protocol is useful in clinical studies on the evaluation of treatment and rehabilitation for patients with head and neck cancer.

Reference List

1. Verdonck-de Leeuw I, Hilgers FJ, Keus RB, et al. Multidimensional assessment of voice characteristics after radiotherapy for early glottic cancer. Laryngoscope 1999 Feb;109(2 Pt 1):241-8.
2. Borggreven PA, Aaronson NK, Verdonck-de Leeuw IM, et al. Quality of life after surgical treatment for oral and oropharyngeal cancer: a prospective longitudinal assessment of patients reconstructed by a microvascular flap. Oral Oncol 2007 Nov;43(10):1034-42.
3. van As CJ, Koopmans-van Beinum FJ, Pols LC, et al. Perceptual evaluation of tracheoesophageal speech by naive and experienced judges through the use of semantic differential scales. J Speech Lang Hear Res 2003 Aug;46(4):947-59.
4. Lundstrom E, Hammarberg B, Munck-Wikland E, et al. The pharyngoesophageal segment in laryngectomees--videoradiographic, acoustic, and voice quality perceptual data. Logoped Phoniatr Vocol 2008;33(3):115-25.
5. Borggreven PA, Verdonck-de Leeuw I, Langendijk JA, et al. Speech outcome after surgical treatment for oral and oropharyngeal cancer: a longitudinal assessment of patients reconstructed by a microvascular flap. Head Neck 2005 Sep;27(9):785-93.
6. Kazi R, Prasad VM, Kanagalingam J, et al. Analysis of formant frequencies in patients with oral or oropharyngeal cancers treated by glossectomy. Int J Lang Commun Disord 2007 Sep;42(5):521-32.
7. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol 2001 Feb;258(2):77-82.
8. Kempster GB, Gerratt BR, Verdolini AK, et al. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. Am J Speech Lang Pathol 2009 May;18(2):124-32.
9. Yamaguchi H, Shrivastav R, Andrews ML, et al. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. Folia Phoniatr Logop 2003 May;55(3):147-57.
10. Kreiman J, Gerratt BR, Ito M. When and why listeners disagree in voice quality assessment tasks. J Acoust Soc Am 2007 Oct;122(4):2354-64.
11. Pauloski BR, Rademaker AW, Logemann JA, et al. Speech and swallowing in irradiated and nonirradiated postsurgical oral cancer patients. Otolaryngol Head Neck Surg 1998 May;118(5):616-24.

12. Maier A, Haderlein T, Stelzle F, et al. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. EURASIP Journal on Audio, Speech and Music Processing 2010;7 pages.
13. Whitehill TL, Ciocca V, Chan JC, et al. Acoustic analysis of vowels following glossectomy. Clin Linguist Phon 2006 Apr;20(2-3):135-40.
14. Yoshida H, Furuya Y, Shimodaira K, et al. Spectral characteristics of hypernasality in maxillectomy patients. J Oral Rehabil 2000 Aug;27(8):723-30.
15. Bjordal K, A.de Graeff, P.Fayers, et al. A 12 country field study of the EORTC QLQ-C30 (version 3.0) and the head and neck cancer specific module (EORTC QLQ-H&N35) in head and neck patients. European Journal of Cancer 2010;36(14):1796-807.
16. Bressmann T, Sader R, Whitehill TL. Consonant intelligibility and tongue motility in patients with partial glossectomy. Journal of Oral and Maxillofacial Surgery 2004;(62):298-303.
17. Terai H, Shimahara M. Evaluation of speech intelligibility after a secondary dehiscence operation using an artificial graft in patients with speech disorders after partial glossectomy. Br J Oral Maxillofac Surg 2004 Jun;42(3):190-4.
18. Ng M, J.Wong. Voice onset time characteristics of esophageal, tracheoesophageal and laryngeal speech of cantonese. Journal of Speech, Language and Hearing Research 2009;52:780-9.
19. Robbins J, J.Christensen, G.Kempster. Characteristics of Speech Production after Tracheoesophageal Puncture. Journal of Speech and Hearing Research 1986;29:499-504.
20. van As CJ. Tracheoesophageal Speech. A Multidimensional Assessment of Voice Quality. 2001.
21. van As-Brooks CJ, Hilgers FJ, Verdonck-de Leeuw IM, et al. Acoustical analysis and perceptual evaluations of tracheoesophageal prosthetic voice. Journal of Voice 1998;12(239):248.
22. van der Molen L, van Rossum MA, Jacobi I, et al. Pre- and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: expert listeners' and patients' perception. Journal of Voice 2012;26(5):664.e26-664.e33.
23. Clapham R, Hilgers FJM, van den Brekel M, et al. An exploration into automatic phonological feature evaluation of tracheoesophageal speech. In: W.Zonneveld HQWH, editor. Sound and sounds: studies presented to M.E.H. (Bert) Schouten on the occasion of his 65th birthday.Utrecht: Utrecht Institute of Linguistics OTS; 2011. p. 69-79.
24. Middag C, Clapham R, van Son R, et al. Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer . Computer Speech & Language 2012;epub ahead of print.
25. Jongmans P, . The intelligibility of tracheoesophageal speech: an analytic and rehabilitation study. University of Amsterdam; 2008.

26. Middag C, Martens J-P, van Nuffelen G, et al. Automated intelligibility assessment of pathological speech using phonological features. EURASIP Journal on advances in Signal Processing - special issue on analysis and signal processing of oesophageal and pathological voices 2009;1-9.
27. Verdonck-de L, I, van NA, Leemans CR. The value of quality-of-life questionnaires in head and neck cancer. Curr Opin Otolaryngol Head Neck Surg 2012 Jan 12.
28. van Gogh CD, Verdonck-de L, I, Boon-Kamma BA, et al. The efficacy of voice therapy in patients after treatment for early glottic carcinoma. Cancer 2006 Jan 1;106(1):95-105.
29. Retel VP, van der ML, Hilgers FJ, et al. A cost-effectiveness analysis of a preventive exercise program for patients with advanced head and neck cancer treated with concomitant chemo-radiotherapy. BMC Cancer 2011;11:475.
30. Zuydam AC, Rogers SN, Brown JS, et al. Swallowing rehabilitation after oro-pharyngeal resection for squamous cell carcinoma. Br J Oral Maxillofac Surg 2000 Oct;38(5):513-8.
31. Rinkel RN, Verdonck-de Leeuw I, van Reij EJ, et al. Speech Handicap Index in patients with oral and pharyngeal cancer: better understanding of patients' complaints. Head Neck 2008 Jul;30(7):868-74.
32. de Nederlandse Taalunie. The IFA Spoken Language Corpus v1.0. 2001.
33. Peterson GE, Lehiste I. Duration of syllable nuclei in English. Journal of the Acoustical Society of America 1960;32(6):693-703.
34. Praat: doing phonetics by computer [Computer program]. [computer program]. Version Version 5.2.35. University of Amsterdam: 2007.
35. Markkanen-Leppanen M, Isotalo E, Makitie AA, et al. Speech aerodynamics and nasalance in oral cancer patients treated with microvascular transfers. J Craniofac Surg 2005 Nov;16(6):990-5.
36. Jong de K. Stress, lexical focus, and segmental focus in English: patterns of variation in vowel duration. Journal of Phonetics 2004;32(4):493-516.
37. Scott SK, Clegg FC, Rudge PC, et al. Foreign accent syndrome, speech rhythm and the functional neuronatomy of speech production. Journal of Neurolinguistics 2006;19(5):370-84.
38. King S, Taylor P. Detection of phonological features in continuous speech using neural networks. Comp Speech Lang 2000;14(4):333-53.
39. Erler K, Freeman GH. An HMM-based speech recognizer using overlapping articulatory features. J Acoust Soc Am 1996;100(4):2500-13.
40. Bourlard H, Dupont S. A new ASR approach based on independent processing and recombination of partial frequency bands. Philadelphia 1996 p. 426-9.
41. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall; 2000.
42. Golden RM. Mathematical Methods for Neural Network Analysis and Design. MIT Press; 1996.

43. Hall M, Frank E, Holmes G. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009;11(1).
44. http://nico.nikkostrom.com/ [computer program]. KTH, Stockholm: 1997.
45. Parveen S, Green P. Multitask Learning in Connectionist Robust ASR using Recurrent Neural Networks. 2003 p. 1813-6.
46. Stadermann J, Koska W, Rigoll G. Multi-task Learning Strategies for a Recurrent Neural Net in a Hybrid Tied-Posteriors Acoustic Model. 2005 p. 2993-6.
47. Frankel J, Cetin O, Morgan N. Transfer Learning for Tandem ASR Feature Extraction. 2007 p. 227-36.
48. Frankel J, Wester M, King S. Articulatory feature recognition using dynamic Bayesian networks. ComputerSpeech & Language 2007;21(4):620-40.
49. Schuster M, Haderlein T, Noth E, et al. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. Eur Arch Otorhinolaryngol 2006 Feb;263(2):188-93.
50. Haderlein T, Riedhammer K, Noth E, et al. Application of automatic speech recognition to quantitative assessment of tracheoesophageal speech with different signal quality. Folia Phoniatr Logop 2009;61(1):12-7.
51. Windrich M, Maier A, Kohler R, et al. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. Folia Phoniatr Logop 2008;60(3):151-6.
52. Middag C, Saeys Y, Martens J-P. Towards an ASR-Free Objective Analysis of Pathological Speech. Proceedings of Interspeech 2010;294-7.
53. Ritchings RT, McGillion M, Moore CJ. Pathological voice quality assessment using artificialneuralnetworks. Medical Engineering & Physics 2002;24(7-8):561-4.
54. Godino-Llorente JI, Gomez-Vilda P. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. 2004 p. 380-4.
55. Wang J, Jo C. Vocal Folds Disorder Detection using Pattern Recognition Methods. 2007 p. 3253-6.
56. Robbins J, Fisher HB, Logemann JA. Acoustic characteristics of voice production after Staffieri's surgical reconstructive procedure. J Speech Hear Disord 1982 Feb;47(1):77-84.
57. Seikaly H, Rieger J, Wu YN, et al. Functional outcomes after primary oropharyngeal cancer resection and reconstruction with the radial forearm free flap. Laryngoscope 2003;(113):897-904.
58. Sumita YI, Ozawa S, Mukohyama H, et al. Digital acoustic analysis of five vowels in maxillectomy patients. J Oral Rehabil 2002 Jul;29(7):649-56.
59. Lee AS, Ciocca V, Whitehill TL. Acoustic correlates of hypernasality. Clin Linguist Phon 2003 Jun;17(4-5):259-64.
60. Moon KR, Chung S.M., Park H.S., et al. Materials of acoustic analysis: sustained vowel versus sentence. J Voice 2012;Epub ahead of print.

61. de Carvalho-Teles V, Sennes L.U., Gielow I. Speech evaluation after palatal augmentation in patients undergoing glossectomy. Arch Otolaryngol Head Neck Surg 2008;134(10):1066-70.
62. Thompson WRE, Jackson M, Dawson F, et al. Figuring function with FIGS and FROGs - Functional rehabilitation after oral cancer. 2005 p. 75.
63. Hong KH, Kwon SH, Jung SS. The assessment of nasality with a nasometer and sound spectrography in patients with nasal polyposis. Otolaryngol Head Neck Surg 1997 Oct;117(4):343-8.
64. Maier A, Noth E, Batliner A, et al. Fully automatic assessment of speech of children with cleft lip and palate. Informatica 2006;30(4):477-82.
65. Moran RJ, Reilly RB, de Chazal P, et al. Telephony-based voice pathology assessment using automated speech analysis. 2006 p. 468-77.
66. Zenner HP. The postlaryngectomy telephone intelligibility test (PLTT). In: Herrmann I.F., editor. Speech restoration via voice prosthesis.Berlin: Springer; 1986. p. 148-52.
67. Haderlein T, Riedhammer K, Maier A, et al. An automatic version of the post-laryngectomy telephone test. Lecture Notes in Artificial Intelligence; Berlin, Heidelberg, New York: Springer; 2007 p. 238-45.
68. Haderlein T, Noth E, Batliner A, et al. Automatic intelligibility assessment of pathologic speech over the telephone. Logoped Phoniatr Vocol 2011;36:175-81.
69. Verdonck-de Leeuw I, Rinkel RN, Leemans CR. Evaluating the impact of cancer of the head and neck. In: Ward EJ, van As-Brooks CJ, editors. Head and neck cancer: treatment, rehabilitation and outcomes.San Diego, USA: Plural publishing; 2007. p. 27-56.
70. Aviv JE, Kim T, Thomson JE, et al. Fiberoptic endoscopic evaluation of swallowing with sensory testing (FEESST) in healthy controls. Dysphagia 1998;13(2):87-92.
71. Splaingard ML, Hutchins B, Sulton LD, et al. Aspiration in rehabilitation patients: videofluoroscopy vs bedside clinical assessment. Arch Phys Med Rehabil 1988 Aug;69(8):637-40.
72. Groves-Wright KJ, Boyce S, Kelchner L. Perception of wet vocal quality in identifying penetration/aspiration during swallowing. J Speech Lang Hear Res 2010 Jun;53(3):620-32.
73. de Bodt M, Guns C, van Nuffelen G, et al. NSVO: Nederlandstalig SpraakVerstaanbaarheidsOnderzoek. Vlaamse Vereniging voor Logopedisten (VVL) 2006.

Summary

Summary

The goal of this thesis was to develop and validate objective speech analysis techniques to assess speech quality of head and neck cancer (HNC) patients. Three pilot studies (chapters 2 – 4) and a validation study (chapter 5) of objective speech quality analyses were performed for development of an objective speech assessment protocol. Various aspects of speech sounds such as duration, spectral information, and feature analysis were investigated by acoustic-phonetic analyses and by using artificial neural network (ANN) techniques. In the pilot and validation studies we investigated speech sounds that are known to be difficult to produce for HNC patients. The speech feature voicing is among the most seriously affected speech characteristics of HNC patients and is a characteristic of voice-onset-time in stop consonants such as /p/, /t/ and /b/, /d/. Insufficient velar closure of the nasopharynx often results in hypernasality and in distorted velar consonants as /k/ and /x/. Further exploration of the artificial neural network as speech quality assessment technique is described in chapter 6. Finally, associations between voice quality and swallowing function were investigated (chapter 7). The results of this thesis contribute to the development of a multidimensional speech quality assessment tool.

Chapter 2 describes a pilot study on the role of objective acoustic-phonetic analyses in a multidimensional speech assessment protocol. Speech recordings of 51 patients treated for HNC and of 18 control speakers were subjectively evaluated by trained raters (speech therapists) regarding intelligibility, hypernasality, articulation and by patients themselves using the speech subscale of a questionnaire (European Organisation for Research and Treatment of Cancer (EORTC) QLQ-H&N35). Formant values of the vowels /a/, /i/, and /u/, size of the vowel space, and timing of air pressure release of /k/ and spectral slope of /x/ were measured. Values of the first and second formant are multiples of the frequency of the fundamental frequency (F0) and are of importance in identifying vowels. The value of the first formant is determined by vertical tongue position; the value of the second formant is determined by the horizontal tongue position. The spectral slope is perceived as loudness. A qualitatively good voice is characterized by

strongly produced high frequencies and a gradual decrease of the spectral slope. A weaker voice has a steep decrease of the spectral slope. Treatment of HNC may result in less effective coordination of muscles and deviant anatomy and physiology of the oral-, nasal- and hypo pharynx, resulting in deteriorated speech quality. Results of this study showed that intelligibility, hypernasality and articulation were best predicted by the vowel space and pressure release of /k/. Results of this study revealed that intelligibility, hypernasality and articulation were predicted best by the vowel space and pressure release of /k/. R^2 values varied from 45% to 74%. Objective acoustic-phonetic analyses distinguished between patients and controls: patients having a higher F1 of /i/ and a lower F2 of /i/ than controls. Within patients, pressure release of /k/ and spectral slope of /x/ differentiated patients regarding tumour site and stage: patients with smaller tumours have a longer pressure release of /k/ compared to patients with a larger tumour. Patients with an oropharyngeal tumour have a steeper spectral slope of /x/ than patients with an oral tumour. Objective speech parameters were not significantly related to speech problems as reported by patients. It was concluded that objective acoustic-phonetic analysis of speech of patients is feasible and contributes to further development of a speech assessment protocol. Further investigation is needed to obtain more insight into acoustic-phonetic analysis in combination with other speech sounds that are oftentimes distorted in patients treated for HNC such as hypernasality.

In **chapter 3** the possibilities of a second objective method were described in a pilot study towards nasality. From previously performed research it is known that hypernasality is common among patients treated for HNC. They oftentimes experience incomplete closure of the velum causing airflow into the nasal cavity resulting in hyper nasal speech. Analysis of the articulatory feature nasalance was performed by an Artificial Neural Network (ANN). ANN automatically calculates the amount of nasalance per time frame of .01 seconds. The articulatory feature nasalance was measured on the vowels /a/, /i/ and /u/ and on the entire stretch of speech. Speech recordings of the same cohort of 51 patients and 18 control speakers as in the previous study were subjectively evaluated regarding intelligibility, hypernasality, articulation and patients completed the EORTC QLQ-HN35 including the Speech

subscale. Results indicated that the feature nasalance as measured by ANN on /i/ and /a/ distinguishes between patients and controls. Within the patient group regarding tumour subsite or stage, no differences in nasalance were found. Nasalance in the vowels /a/ and /i/ predicted best intelligibility (R^2= 21.3%), while nasalance in the vowel /a/ predicted best articulation (R^2= 48.7%) and nasalance in the vowels /i/ and /u/ predicted best hypernasality (R^2= 24.9%). It was concluded that nasalance as assessed by ANN contributes moderately to the speech evaluation by trained raters. Further research with larger study samples and other speech features is needed.

In **chapter 4** acoustic-phonetic analyses and the automatic feature detection methods by Artificial Neural Networks (ANNs) were used to analyze the quality of stop consonants (in Dutch /b/, /d/, /p/, /t/). The VOT in stop consonants distinguishes voiced and voiceless stops. Patients treated for HNC may have difficulty with adequate coordination of motor function of articulatory speech structures and vocal fold vibration. Building up oral pressure necessary for stop consonants in combination and synchronously with ceasing vocal fold vibration in case of the voiceless stop consonants may be especially problematic. For patients it seems problematic to quickly stop the activity of the glottis so that no voicing is produced. Because this action is often difficult to perform for HNC patients it is therefore hypothesized that the duration of VOT preceding the burst in voiced stops in patients is longer compared to controls and that the silence period preceding the burst in voiceless stops show more voicing in patients compared to controls. In the present study, stop consonants /p, t, b, d/ were extracted from speech samples of the same cohort of 51 patients and 18 controls as was used in chapter 2-3. Acoustic-phonetic analyses were performed to investigate the duration of VOT and of the burst. The amount of the articulatory feature 'voicing' in VOT and in the burst was measured using ANN. Results revealed that objective acoustic-phonetic analysis and feature 'voicing' analysis for /b, d, p/ distinguish between patients and controls. Within patients, /t, d/ distinguish for tumour location and tumour stage: patients with larger tumours had significantly less voicing during VOT compared to patients with smaller tumours and during the burst in the voiced consonant /d/. Regarding tumour location, patients with a tumour originating in the oral cavity had a shorter burst in the

voiceless consonant /t/ compared to patients treated for oropharyngeal cancer. Measurements of the phonological feature voicing in almost all consonants were significantly correlated with articulation and intelligibility, but not with self-evaluations of speech problems in daily life (EORTC QLQ-HN35 Speech Subscale). It was concluded that objective acoustic-phonetic and feature analyses of stop consonants are feasible and contribute to further development of a multidimensional speech quality assessment protocol.

Chapter 5 describes an external validation study of the speech analyses techniques previously investigated in the pilot studies (chapters 2-4). In these previously performed studies we tested the objective acoustic-phonetic and ANN analyses separately (chapters 2 and 3) and then we tested these two objective methods combined onto a selection of speech sounds (chapter 4). This study is aimed at multivariate validation of these objective speech analyses methods together with all previously used speech sounds onto the previously used patient cohort of 51 patients six months after treatment as well as onto a new patient cohort (external validation). This second patient cohort was composed of 64 patients, six months to nine years after treatment for HNC. Speech quality was subjectively evaluated for intelligibility, articulation, hypernasality and by self-evaluations of patients (the speech-subschaal of the European Organisation for Research and Treatment of Cancer (EORTC) QLQ-H&N35). Acoustic-phonetic analysis were performed on vowels /a, i, u/, stop consonants /k, p, b, d, t/ and fricative /x/. ANN analysis of the feature 'nasalance' was performed on vowels /a, i, u/ and the entire stretch of running speech; ANN analysis of the feature 'voicing' was performed on consonants /p, b, d, t/.
In patient cohort 1subjective intelligibility was predicted by acoustic-phonetic analysis of /p/ and vowel space and by ANN analysis of /d/. Articulation was predicted by acoustic-phonetic analysis of vowel space and by ANN analysis of the feature 'voicing' of /b/. Hypernasality was predicted by acoustic-phonetic analysis of /a/, /x/ and /b/. Self-evaluations by patients were predicted by acoustic-phonetic analysis of /i/ and /k/ and by ANN analysis of /p/. The amount of explained variance varied between moderate (52.0% for hypernasality, 37.7% for intelligibility and 36.2% for articulation) to poor (21.1% for self-evaluations by patients).

In cohort 2 intelligibility was predicted by acoustic-phonetic analysis of /a/, /i/ and /x/. Articulation was predicted by acoustic-phonetic analysis of vowel space and by ANN analysis of the feature 'voicing' of /p/. Hypernasality was predicted by acoustic-phonetic analysis of /p/ and /t/. Self-evaluations by patients were predicted by acoustic-phonetic analysis of /u/ and /t/ and by ANN analysis of /d/. The amount of explained variance varied between moderate (51.9% for self-evaluations by patients and 41.3% for intelligibility) to poor (21.8% for hypernasality and 20.9% for articulation).
The conclusion is that the combination of previously used analysis techniques and speech material in both cohorts lead to two different predictive models, that both are moderately predictive but are not better that previously tested models.

In **chapter 6,** further investigation described the possibilities of the Artificial Neural Network (ANN) analyses to investigate the speech quality of patients treated for HNC. In chapters 3, 4 and 5 two specific articulatory features were investigated ('nasal' and 'voicing') known to be possible features affected in HNC patients. In the present chapter all 28 articulatory features of Dutch were investigated. From the results of the present study was revealed that the features nasalance, voicing and labio-dental appeared to be the most relevant speech features in HNC patients: the speech of patients was significantly different from control speakers for these speech features. The results for nasalance and voicing were in accordance with previously performed studies, as was described in chapters 3 and 4: the difference in voicing between patients and controls speakers was on average 0.16 and the averaged delay between patients and control speakers was more than 0.005 seconds. For the feature nasalance, 8 out of 51 patients appear to have an average nasality that is more than two standard deviations away from the control's mean. For the feature labio-dental -a feature that is place-bounded- it was seen that four patients had trajectories (from vowel to labiodental) different from the controls. These trajectories were all related to the transitions between vowels and the *voiced* labio-dental, strongly suggesting that this effect is actually a side-effect related to the feature *plosive*. That the feature labio-dental is of importance is considered new information and was not known from previously performed research. This result could be possibly explained because

part of the patients underwent (maxillofacial) surgery that was of influence on production of labiodentals speech sounds. Oppositely of the previously performed pilot studies, in the present study no distinction was found between patients and control speakers concerning the speech feature velar (as measured by ANN), while in the velar speech sound /k/ as measured by acoustic-phonetic analysis a difference was found between patients and control speakers (chapter 2). This difference could be explained because ANN calculates the amount of the speech feature velar in running speech, while the acoustic-phonetic method measures the duration of air pressure release as a percentage of the duration of the specific speech sound /k/.

Patients treated for HNC oftentimes experience, next to speech difficulty, difficulty with swallowing. Because in literature a relation between swallowing problems and voice quality was assumed, in **chapter 7** a study is described investigating a possible association between voice- and swallowing parameters in patients treated for HNC. Acoustic variables of voice that were measured were fundamental frequency (F0), jitter (percentage), shimmer (percentage), harmonics-to-noise ratio (HNR) and intensity. Jitter is the temporal deviation of cycles produced by the vocal folds (perturbation or disturbance in F0). A high jitter means large frequency perturbation of consecutive cycles in frequency. Shimmer means a disturbance of the amplitude cycles. HNR represents the harmonicity of the wave-like shape of consecutive voicing cycles and describes the amount of harmonicity relative to the amount of white noise in the signal. Vocal intensity was measured in decibel (dB) and represents the sound pressure level. These acoustic parameters were measured in the vowels /a/, /i/ and /u/ and compared to swallowing function parameters as assessed via videofluoroscopy (oral, oropharyngeal and total transit time, estimated percent of oral, oropharyngeal, and total residue, oropharyngeal swallowing efficiency (OPSE) and the Penetration-Aspiration (PA-) scale). Results revealed that Intensity in all three vowels /a/, /i/ and /u/ was significantly associated with OPSE and the score on the PA-scale: a worse swallowing function is correlated with louder voice. A possible explanation may be found in overcompensation by increased laryngeal muscular strength resulting in increased intensity and pitch during phonation. However, more research is needed to examine this explanation.

In the general discussion (**chapter 8**) of this dissertation the main findings, methodological considerations and clinical implications are described, followed by recommendations for future research.

The main goal of the present research was to develop and validate objective speech analyses techniques to evaluate speech quality among patients treated for HNC. The applied methods (acoustic-phonetic and ANN analyses), as well as a variety of phonemes contributed to this goal. However, correlations with subjective assessments by listeners or by patients themselves were limited. There is a number of remarks to make concerning the studies. Speech of relatively small cohorts were used (51 patients (cohort 1), 64 patients (cohort 2) and 18 control speakers). Considerations concerning speech material include reading aloud text of which reading skills could have influenced speech production. A drawback of running speech is the presence of coarticulation and assimilation of speech sounds in which neighboring speech sounds influence the target phoneme. Concerning the two objective measuring methods, the used ANN is of relatively simple origin and was trained on speech of only two speakers. In the future this technique possibly may be improved by using a larger amount of test speakers.
For clinical application the further development of ANN seems to be a better choice than the further development of acoustic-phonetic analyses, although segmenting of target sounds from running speech can be performed automatically through 'forced alignment'. In forced alignment the signal is lined up onto a sequence of acoustic models that were trained preceding the alignment. In a successful alignment it is expected that 80% of all found phone boundaries are located within a boundary of 20 milliseconds of human-annotated-boundaries. For development of an application that objectively measures voice intensity as an indication of swallowing problems in the oropharyngeal stage, more research is needed to confirm our results and to establish a threshold as a criterion for referral to the clinical setting for further swallowing assessment. Also more research is needed to investigate the (causal) relation between louder voice and worse swallowing.

Easily accessible tools for screening of speech-, voice- and swallowing problems are relevant for clinical practice. Possible applications in the

future are the development of a speech test through the telephone. Patients record speech through the telephone where after the speech recording immediately and automatically is processed by, p.e., an Artificial Neural Network. However, from the present research is known that further pilot studies are needed for validation of objective speech analysis methods. For further research it was advised to use more speech material of larger bodies of speakers –both patients and control speakers– and to take into account the differences in speaker style and demographic and clinical variables such as tumour location, tumour stage and treatment modality. The final conclusion of this dissertation is that objective analysis of speech of patients treated for HNC through acoustic-phonetic analyses and Artificial Neural Network analyses is feasible and valid. Using these findings, medical sciences and speech technology can perform further research that finally may lead to a multidimensional speech evaluation protocol that is usable in clinical practice.

Samenvatting
(summary in Dutch)

Spraakkwaliteit bij patiënten met een orale of oropharyngeale tumor

De ontwikkeling en evaluatie van objectieve spraakbeoordelingsmethoden

Samenvatting (summary in Dutch)

Het doel van dit proefschrift is het ontwikkelen en valideren van spraakanalyse-technieken om de spraakkwaliteit van patiënten behandeld voor hoofd-halskanker (HHK) objectief te kunnen meten. Drie pilotstudies (hoofdstuk 2-4) en één validatiestudie (hoofdstuk 5) zijn uitgevoerd ter ontwikkeling van een objectief spraakanalyseprotocol. Diverse aspecten van spraakklanken zoals duur, spectrale informatie en 'feature-analyse' (analyse van specifieke kenmerken van spraakklanken) werden onderzocht door middel van akoestisch-fonetische metingen en door middel van een artificieel neuraal netwerk (ANN). In de pilot- en validatiestudies zijn spraakklanken geanalyseerd waarvan bekend was dat die moeilijk te produceren zijn voor patiënten behandeld voor HHK. Coördinatieproblemen bij de afsluiting van het velum en de nasofarynx leiden vaak tot hypernasaliteit en incorrect geproduceerde velaire medeklinkers zoals /k/ en /x/. Door problemen met het spraakkenmerk stemhebbendheid worden stemhebbende medeklinkers zoals de /b/ and /d/ vaak stemloos geproduceerd en horen luisteraars een /p/ respectievelijk /t/. Nader onderzoek naar de inzetbaarheid van ANN ter beoordeling van spraakkwaliteit is beschreven in hoofdstuk 6. Tenslotte is de relatie tussen stemkwaliteit en slikfunctie onderzocht in hoofdstuk 7. De bevindingen van dit proefschrift dragen bij aan de verdere ontwikkeling en validatie van een multidimensioneel spraakevaluatie-protocol.

Hoofdstuk 2 beschrijft een pilotstudie naar de rol van objectieve akoestisch-fonetische analyse in een multidimensioneel spraakevaluatie-protocol. Spraakopnamen van 51 patiënten en van 18 controlesprekers zijn subjectief beoordeeld door getrainde beoordelaars (logopedisten) op verstaanbaarheid, hypernasaliteit, articulatie en door patiënten zelf aan de hand van vragen over spraak in een gevalideerde vragenlijst (de Spraak subschaal van de European Organisation for Research and Treatment of Cancer (EORTC) QLQ-H&N35). Formantwaarden van de eerste twee formanten (F1 en F2) van de klinkers /a/, /i/ en /u/, de oppervlakte van de klinkerdriehoek, en duur van de pressure release (drukopheffing) van /k/ en de spectrale helling van /x/ zijn gemeten. De waarden van de eerste en tweede formant zijn veelvouden van de frequentie van de grondtoon (F0) en zijn van belang voor de identificatie

van klinkers. De waarde van de eerste formant wordt bepaald door de verticale tongpositie; de waarde van de tweede formant wordt bepaald door de horizontale tongpositie. De spectrale helling wordt waargenomen als luidheid. Een kwalitatief goede stem kenmerkt zich door sterke boventonen en een geleidelijke afname van de spectrale helling. Een slechtere stem heeft een sterke afname van de spectrale helling.

Behandeling voor HHK kan resulteren in een minder goede aansturing van spieren en afwijkende anatomie en fysiologie van mond-, neus- en en keelholte met verminderde spraakkwaliteit tot gevolg. Resultaten van deze studie lieten zien dat verstaanbaarheid, hypernasaliteit en articulatie het best werden voorspeld door de klinkerdriehoek en de pressure release van /k/. Verklaarde variantie varieerde van 45% tot 74%. Objectieve akoestisch-fonetische analyse onderscheidden patiënten van controlesprekers: patiënten hadden een hogere F1 van /i/ en een lagere F2 van /i/ dan controlesprekers. Binnen de patiëntengroep bleek de pressure release van /k/ en de spectrale helling van /x/ het best onderscheid maakten wat betreft tumorlocatie en -stagiëring: patiënten met kleinere tumor hebben een langere pressure release van /k/ dan patiënten met grotere tumor. Patiënten met een orofaryngeale tumor hebben een steilere spectrale helling van /x/ dan patiënten met een tumor in de mondholte. Objectieve spraakparameters waren niet significant gecorreleerd aan spraakproblemen zoals gerapporteerd door patiënten. Er werd geconcludeerd dat objectieve akoestisch-fonetische analyse van spraak van patiënten behandeld voor HHK uitvoerbaar is en bijdraagt aan een multidimensioneel spraakevaluatieprotocol. Nader onderzoek is nodig om meer inzicht te verkrijgen in akoestisch-fonetische analyse gecombineerd met andere spraakklanken die problematisch zijn voor patiënten behandeld voor HHK, zoals hypernasaliteit.

In **hoofdstuk 3** zijn de mogelijkheden van een tweede objectieve meetmethode beschreven in een pilotstudie naar hypernasaliteit. Uit eerder onderzoek is gebleken dat hypernasaliteit een veelvoorkomend probleem is voor patiënten behandeld voor HHK. Ze hebben vaak onvoldoende coördinatiemogelijkheden van het velum wat resulteert in hypernasale spraak. Analyse van het spraakkenmerk feature nasaliteit

werd uitgevoerd door middel van een artificieel neuraal netwerk (ANN). ANN berekent automatisch de hoeveelheid nasaliteit per tijdframe van .01 seconde. Het spraakkenmerk nasaliteit is gemeten in de klinkers /a/, /i/ en /u/, evenals over lopende spraak (voorgelezen tekst). Spraakopnamen van hetzelfde cohort van 51 patiënten en 18 controlesprekers in hoofdstuk 2 werden subjectief beoordeeld op verstaanbaarheid, hypernasaliteit, articulatie en beoordelingen door patiënten zelf. Uit de resultaten is gebleken dat het spraakkenmerk nasaliteit zoals gemeten met ANN in /i/ en /a/ patiënten en controlesprekers van elkaar kan onderscheiden. Binnen de patiëntengroep zijn geen verschillen in nasaliteit gevonden wat betreft tumorstadiëring en -locatie. Nasaliteit in de klinkers /a/ en /i/ voorspelden het best de verstaanbaarheid (R^2= 21.3%), terwijl nasaliteit in de klinker /a/ het best articulatie (R^2= 48.7%) voorspelde. Nasaliteit in de klinkers /i/ en /u/ voorspelden het best hypernasaliteit (R^2= 24.9%). Geconcludeerd werd dat nasaliteit zoals beoordeeld met ANN ook bijdraagt aan een multidimensioneel spraakevaluatieprotocol.

In **hoofdstuk 4** zijn akoestisch-fonetische en ANN analyses gebruikt om de kwaliteit van de medeklinkers /b/, /d/, /p/, /t/) te analyseren. De 'voice-onset-time' in deze medeklinkers maakt onderscheid tussen stemhebbende (/b/ en /d/) en stemloze (/p/ en /t/) medeklinkers. Patiënten behandeld voor HHK kunnen moeilijkheden hebben met adequate coördinatie van articulatorische spraakstructuren en stembandtrilling. Vooral de orale drukopbouw die benodigd is voor deze medeklinkers gecombineerd met synchrone stopzetting van stembandtrilling in geval van stemloze medeklinkers kan lastig zijn. De hypothese is dat de duur van de VOT, voorafgaand aan de drukopheffing in stemhebbende medeklinkers, langer is bij patiënten vergeleken met controlesprekers, en dat de stilte voor de drukopheffing in stemloze medeklinkers meer stemhebbendheid bevat bij patiënten dan bij controlesprekers. In de huidige studie werden de medeklinkers /p, t, b, d/ geëxtraheerd uit spraakopnamen van 51 patiënten en de 18 controlesprekers (hetzelfde cohort als in de eerdere studies). Akoestisch-fonetische analyses werden uitgevoerd om de duur van de VOT en de drukopheffing te meten. ANN werd gebruikt om het spraakkenmerk stemhebbendheid te meten. Uit de resultaten bleek dat objectieve akoestisch-fonetische en analyse van het spraakkenmerk

stemhebbendheid voor /b, d, p/ onderscheid maken tussen patiënten en controlesprekers. Binnen de patiëntengroep bleek dat /t, d/ onderscheid maken voor wat betreft tumorlocatie en -stagiëring. Wat betreft tumorlocatie bleek dat patiënten met een tumor in de mondholte een kortere drukopheffing van de stemloze medeklinker /t/ hadden vergeleken met patiënten behandeld aan een tumor in de orofarynx. Patiënten met grotere tumoren hadden minder stemhebbendheid gedurende VOT dan patiënten met kleinere tumoren. Deze patiënten hadden ook minder stemhebbendheid tijdens de pressure release van /d/. Metingen van het spraakkenmerk stemhebbendheid op bijna alle onderzochte medeklinkers waren significant gecorreleerd met articulatie en verstaanbaarheid, maar niet met zelfbeoordelingen door patiënten. Geconcludeerd werd dat objectieve akoestisch-fonetische analyse en metingen met ANN op de medeklinkers /p, t, b, d/ haalbaar zijn en ook bijdragen aan een multidimensioneel spraakevaluatieprotocol.

Hoofdstuk 5 beschrijft een externe validatiestudie van de spraakanalysetechnieken in de hierboven beschreven pilotstudies (hoofdstukken 2-4). In deze eerdere studies hebben we de objectieve akoestisch-fonetische en ANN analyses gescheiden getest (hoofdstukken 2 en 3) en gezamenlijk getest op een aantal spraakklanken (hoofdstuk 4). Deze studie is erop gericht deze objectieve spraakanalyses en alle spraakklanken multivariaat te valideren in het eerder gebruikte patiëntencohort van 51 patienten 6 maanden na behandeling en in een nieuw patiëntencohort (externe validatie). Dit tweede patiëntencohort bestond uit 64 patiënten, zes maanden tot negen jaar na behandeling voor HHK. Spraakkwaliteit werd subjectief geevalueerd op verstaanbaarheid, articulatie, en hypernasaliteit en door zelfevaluaties van patiënten (de spraak-subschaal van European Organisation for Research and Treatment of Cancer EORTC QLQ-H&N35). Akoestisch-fonetische analyses werden verricht op de klinkers /a, i, u/, stop consonanten /k, p, b, d, t/ en fricatief /x/. ANN analyse van de feature 'nasalance' werd gedaan over de klinkers /a, i, u/ en over lopende spraak; ANN analyse van de feature 'stemhebbendheid' werd verricht over de consonanten /p, b, d, t/.
In patiënten cohort 1 werd subjectieve verstaanbaarheid voorspeld door akoestisch-fonetische analyses van /p/ en de klinkerdriehoek en door ANN analyse van /d/. Articulatie werd voorspeld door akoestisch-

fonetische analyse van de klinkerdriehoek en ANN metingen van de feature 'stemhebbendheid' van /b/. Hypernasaliteit werd voorspeld door akoestisch-fonetische analyses van /a/, /x/ and /b/. Zelfevaluaties door patiënten werd voorspeld door akoestisch-fonetische analyses van /i/ en /k/ en door ANN analyse van /p/. De verklaarde variantie varieerde van redelijk (52.0% voor hypernasaliteit, 37.7% voor verstaanbaarheid en 36.2% voor articulatie) tot matig (21.1% voor zelfevaluaties door patiënten).
In cohort 2 werd verstaanbaarheid voorspeld door akoestisch-fonetische analyses van /a/, /i/ en /x/. Articulatie werd voorspeld door akoestisch-fonetische analyses van de klinkerdriehoek en door ANN analyses van de feature 'stemhebbendheid' van /p/. Hypernasaliteit werd het best voorspeld door akoestisch-fonetische analyses van /p/ en /t/. Zelfevaluaties door patiënten werd het best voorspeld door akoestisch-fonetische analyses van /u/ en /t/ en door ANN analyse van /d/. De verklaarde variantie varieerde van redelijk (51.9% voor zelfevaluaties door patiënten en 41.3% voor verstaanbaarheid) tot matig (21.8% voor hypernasaliteit en 20.9% voor articulatie).
De conclusie is dat de combinatie van de eerder gebruikte analyse-technieken en spraakmateriaal in beide cohorten tot twee verschillende voorspellende modellen leiden, die beiden redelijk voorspellend zijn maar niet beter dan de eerdere getoetste modellen.

In **hoofdstuk 6** beschrijft verder onderzoek naar de mogelijkheden van het artificieel neuraal netwerk om de spraakkwaliteit van patiënten behandeld voor HHK te analyseren. In hoofdstukken 3, 4 en 5 werden twee specifieke spraakkenmerken 'features' onderzocht (nasaliteit en stemhebbendheid) waarvan bekend is dat deze problematisch kunnen zijn voor patiënten behandeld voor HHK. In dit hoofdstuk werden alle 28 spraakkenmerken van het Nederlands onderzocht.
Uit de resultaten van deze studie bleek dat nasaliteit, stemhebbendheid en labio-dentaal de meest relevante spraakkenmerken zijn bij patiënten behandeld voor HHK: de spraak van patiënten was op deze spraakkenmerken significant ander dan van controlesprekers. Deze resultaten wat betreft nasaliteit en stemhebbendheid komen overeen met de eerder uitgevoerde studies beschreven in hoofdstuk 3 en 4: het verschil in stemhebbendheid tussen patiënten en controlesprekers bedraagt gemiddeld 0.16 en de gemiddelde vertraging tussen patiënten

en controlesprekers is meer dan 0.005 seconden. Voor de feature nasalance lijken 8 van de 51 patiënten een gemiddelde nasalance te hebben die verder dan 2 standaard deviaties is verwijderd van de gemiddelde nasalance van controlesprekers. Voor de feature labio-dental -een plaatsgebonden feature- werd gobserveerd dat bij vier patiënten de overgang van klinker naar labiodentaal verschilde van die van controlesprekers. Deze overgangen waren allen gerelateerd aan de overgang tussen klinker en *stemhebbende* labiodentaal, hetgeen sterk suggereert dat dit effect eigenlijk een bijkomstigheid is van de relatie met de feature 'plosief'. Dat de feature labiodentaal van belang is, is nieuwe informatie en is niet eerder uit onderzoek gebleken. Dit kan mogelijk verklaard worden doordat een deel van de patiënten (kaak-) chirurgische behandeling heeft ondergaan die van invloed is geweest op de productie van labiodentale spraakklanken.
In tegenstelling tot de eerdere pilotstudie werd in de huidige studie geen onderscheid gevonden tussen patiënten en controlesprekers wat betreft het spraakkenmerk velair (zoals gemeten met ANN), terwijl de velaire spraakklank /k/ zoals akoestisch-fonetisch gemeten wel verschilde tussen patiënten en controlesprekers (hoofdstuk 2). Dit kan onder andere verklaard worden doordat ANN de hoeveelheid van het spraakkenmerk velair berekent in lopende spraak, terwijl de akoestisch-fonetische methode de duur van de drukopheffing als percentage van de duur van de specifieke klank /k/ meet.

Patiënten behandeld voor HHK hebben naast spraakproblematiek ook vaak slikproblemen. Omdat in de literatuur gesuggereerd wordt dat er een relatie is tussen slikproblemen en stemkwaliteit, wordt in **hoofdstuk 7** een studie beschreven waarin een mogelijke relatie tussen stem- en slikparameters bij HHK patiënten wordt onderzocht. Akoestische stemvariabelen betroffen grondfrequentie (F0), jitter (percentage), shimmer (percentage), harmonics-to-noise ratio (HNR) en intensiteit. Jitter is de temporele cyclusafwijking veroorzaakt door de stembanden (perturbatie of verstoring in F0). Een hogere jitter betekent een grotere cyclusafwijking in frequentie. Shimmer betekent een verstoring in de amplitudecyclus. HNR is de harmoniciteit van de golfvorm van de opeenvolgende stemcycli en beschrijft de hoeveelheid harmoniciteit ten opzichte van de hoeveelheid ruis in het signaal. Stemintensiteit werd gemeten in decibel (dB) en representeert de

geluiddruk. Deze akoestische parameters zijn gemeten in de klinkers /a/, /i/ en/ /u/ en zijn vergeleken met slikfunctieparameters beoordeeld via videofluoroscopische opnames van het slikproces (orale, orofaryngeale en totale passagetijd, geschat percentage van oraal, orofaryngeaal en totaal residu, orofaryngeale slikefficiëntie (OPSE) en de Penetratie-Aspiratie (PA-)schaal. Uit de resultaten is gebleken dat stemintensiteit in de drie klinkers /a/, /i/ en /u/ significant geassocieerd is met OPSE en de score op de PA-schaal: een slechtere slikfunctie hangt samen met een luidere stem. Een mogelijke verklaring voor deze bevinding wordt gezocht in overcompensatie door verhoogde laryngeale spierspanning leidend tot een verhoogde intensiteit. Maar meer onderzoek is nodig om deze verklaring te toetsen.

In de algemene beschouwing (**hoofdstuk 8**) van deze studie werden de doelstellingen, bevindingen, methodologische kanttekeningen en de klinische implicaties beschreven, gevolgd door aanbevelingen voor toekomstig onderzoek.
De doelstelling van dit onderzoek was het ontwikkelen en valideren van spraakanalysemethoden teneinde de spraakkwaliteit van patiënten behandeld voor HHK objectief te kunnen meten. De toegepaste methoden (akoestisch-fonetische analyses en ANN) en een verscheidenheid aan fonemen droegen bij aan deze doelstelling. Echter, de correlaties met subjectieve beoordelingen door luisteraars of de patiënten zelf waren beperkt. Er is een aantal kanttekeningen te plaatsen bij het onderzoek.
In het onderzoek is spraakmateriaal van relatief kleine cohorten gebruikt (51 (cohort 1), 64 patiënten (cohort 2) en 18 controlesprekers). Het spraakmateriaal betrof voorgelezen tekst, waarin de mate van voorleesvaardigheid een rol kan hebben gespeeld bij de spraakproductie. Het nadeel van lopende spraak is het voorkomen van coarticulatie en assimilatie van spraakklanken waarbij de naburige spraakklanken het doelfoneem kunnen beïnvloeden. Wat betreft de twee objectieve meetmethoden wordt opgemerkt dat voor de ANN techniek slechts twee sprekers zijn gebruikt in de trainingsfase. Mogelijk dat in de toekomst deze techniek verbeterd kan worden door meer sprekers te gebruiken. Voor de klinische toepassing lijkt het verder doorontwikkelen van ANN analyses een beter plan dan het doorontwikkelen van akoestisch-fonetische analyses, hoewel het segmenteren van doelklanken uit

lopende spraak voor akoestisch-fonetische analyses automatisch gedaan kan worden via 'forced alignment'. Bij forced aligment wordt het signaal opgelijnd aan een sequentie van akoestische modellen die vooraf getraind zijn. Bij een succesvolle alignment wordt verwacht dat 80 procent van alle gevonden phone boundaries binnen de marge van 20 milliseconden van de human-annotated-boundaries vallen. Voor de ontwikkeling van een applicatie die objectief stemintensiteit meet indicatief voor slikproblemen in de orofaryngeale fase is meer onderzoek nodig die de gevonden resultaten bevestigen en om een drempelwaarde te bepalen als criterium voor doorverwijzing naar de kliniek voor verder slikfunctieonderzoek. Ook is meer onderzoek nodig naar de (causale) relatie tussen een luidere stem en een slechtere slikfunctie.

Gemakkelijk toegankelijke hulpmiddelen voor het screenen van spraak-, stem- en slikproblematiek zijn relevant voor de klinische praktijk. Mogelijke toepassingen in de toekomst zijn de ontwikkeling van een spraaktest via de telefoon. Patiënten spreken tekst in via de telefoon waarna dit spraakmateriaal direct en automatisch wordt verwerkt door bijvoorbeeld een artificieel neuraal netwerk. Echter, uit het huidige onderzoek blijkt dat er nog verder vooronderzoek moet worden gedaan naar de validiteit van objectieve spraakanalysemethoden. Voor verder onderzoek wordt geadviseerd om meer spraakmateriaal te gebruiken van grotere groepen sprekers -zowel patiënten als controlesprekers- waarbij rekening wordt gehouden met verschillen in spreekstijl en demografische variatie en klinische variabelen als tumorlocatie en grootte en behandelingsmodaliteit.

De uiteindelijke conclusie van deze thesis luidt dat objectieve analyse van spraak van patiënten behandeld voor hoofd-halskanker door middel van akoestisch-fonetische analyse en een artificieel neuraal netwerk haalbaar en valide is. Aan de hand van deze bevindingen kan vanuit de medische wetenschappen en spraaktechnologie verder onderzoek gedaan worden dat uiteindelijk kan leiden tot een multidimensionaal spraakevaluatieprotocol dat bruikbaar is in de klinische praktijk.

List of publications

De Bruijn, MJ; Chao, L-Y; Rinkel, RNPM; Borggreven, PA; Boves, L; Leemans, CR; Verdonck-de Leeuw, IM. Speech quality after major surgery of the oral cavity and oropharynx with microvascular soft tissue reconstruction. Proceedings Interspeech. 2007; 1186-9

De Bruijn, MJ; Verdonck-de Leeuw, IM; ten Bosch, L; Kuik, DJ; Quené, H; Boves, L; Langendijk, JA; Leemans, CR. Phonetic-acoustic and feature analyses by a neural network to assess speech quality in patients treated for head and neck cancer. Proceedings Interspeech. 2008; 1753-6

De Bruijn, MJ; ten Bosch, L; Kuik, DJ; Quené, H; Langendijk, JA; Leemans, CR; Verdonck-de Leeuw, IM. Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. Folia Phoniatr Logop. 2009; 61(3):180-7

De Bruijn, MJ; ten Bosch, L; Kuik, DJ; Langendijk, JA; Leemans, CR; Verdonck-de Leeuw, IM. Artificial neural network analysis to assess hypernasality in patients treated for oral or oropharyngeal cancer. Logoped Phoniatr Vocol. 2011; 36(4):168-74.

De Bruijn, MJ; ten Bosch, L; Kuik, DJ; Witte, BI; Langendijk, JA; Leemans, CR; Verdonck-de Leeuw, IM. Acoustic-phonetic and artificial neural network feature analysis to assess speech quality of stop consonants produced by patients treated for oral or oropharyngeal cancer. Speech Communication. 2012; 54(5), 632-640

Christianen, ME; Schilstra, C; Beetz, I; Muijs, CT; Chouvalova, O; Burlage, FR; Doornaert, P; Koken, PW; Leemans, CR; Rinkel, RN; de Bruijn, MJ; de Bock, GH; Roodenburg, JL; van der Laan, BF; Slotman, BJ; Verdonck-de Leeuw, IM; Bijl, HP; Langendijk, JA. Predictive modelling for swallowing dysfunction after primary (chemo)radiation: Results of a prospective observational study. Radiother Oncol. 2012; 105(1):107-14

De Bruijn, MJ; Rinkel, RNPM; Cnossen, IC; Witte, BI; Langendijk, JA; Leemans, CR; Verdonck-de Leeuw, IM. Associations between voice quality and swallowing function in patients treated for oral or oropharyngeal cancer. Accepted for publication in Supportive Care in Cancer. 2013.