Automatic Evaluation of Voice and Speech Intelligibility After Treatment of Head and Neck Cancer

Renee P. Clapham

UNIVERSITY OF AMSTERDAM
Amsterdam Center for Language and Communication

Universiteit van Amsterdam
Faculteit Geesteswetenschappen
ACLC

Automatic evaluation of voice and speech intelligibility after treatment of head and neck cancer

Renee P. Clapham

Automatic evaluation of voice and speech intelligibility after treatment of head and neck cancer

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in Brisbane, Australië
op woensdag 1 november 2017, te 18:00 uur

door Renee Peje Clapham

geboren te Brisbane, Australië

Promotiecommissie

Promotor:	prof. dr. M.W.M. van den Brekel	Universiteit van Amsterdam
Co-promotoren:	prof. dr. F.J.M. Hilgers	Universiteit van Amsterdam
	dr. R.J.J.H. van Son	Universiteit van Amsterdam
	dr. C. Middag	Universiteit Gent
Overige leden:	prof. dr. E.C. Ward	The University of Queensland
	prof. dr. M. De Bodt	Universiteit Antwerpen
	prof. dr. A.C.M. Rietveld	Radboud Universiteit Nijmegen
	prof. dr. P.P.G. Boersma	Universiteit van Amsterdam
	dr. L.J. Beijer	St. Maartenskliniek, Hogeschool Arnhem Nijmegen
Faculteit:	Geesteswetenschappen	

Acknowledgments

Renee Clapham was partially funded by the Verwelius Foundation, Huizen, The Netherlands.

Author contributions

1 General introduction

Renee P. Clapham wrote the text. Rob van Son, Catherine Middag, Frans Hilgers, and Michiel van den Brekel contributed to the final version.

2 NKI-CCRT Corpus - Speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy

Renee P. Clapham, Lisette van der Molen, Rob J.J.H. van Son, Michiel W.M. van den Brekel, F.J.M. Hilgers. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12): 23-25 May, 2012, Istanbul, Turkey, 3350-3355.*

RC and RvS posed the research question and designed the experiments. FH, LvdM, and MvdB contributed to the design of the study. RvS prepared the speech stimuli and did the technical set-up of the perceptual experiment. RC recruited the listeners for the perceptual experiment. RvS prepared the response data. RC did the statistical analysis and interpretation. RC wrote the initial versions of the text and all authors contributed to the final version of the text.

3 Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer

Catherine Middag, Renee P. Clapham, Rob van Son, Jean-Pierre Martens. *Computer Speech and Language 28 (2), March 2014, 467-482.*

CM and RC contributed equally to this study. CM and RC developed the research questions and designed the experiments. JPM and RvS contributed to the design of the study. CM developed the computer models. CM and RC then evaluated the results. CM and RC wrote the initial versions of the text and all authors contributed the the final text.

4 Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer

Renee P. Clapham, Catherine Middag, Frans J.M. Hilgers, Jean-Pierre Martens, Michiel W.M. van den Brekel, Rob J.J.H. van Son. *Speech Communication, 59, April 2014, 44-54.*

RC and CM contributed equally to this study. RC and CM developed the research questions and designed the experiments. JPM, FH, MvdB, and RvS contributed to the design of the study. CM developed the computer models. RC and CM evaluated the data. RC and CM wrote the initial versions of the text and all authors contributed to the final text.

5 Automatic tracheoesophageal voice typing using acoustic features

Renee P. Clapham, Corina J. Van As-Brooks, Michiel W. M. van den Brekel, Frans J. M. Hilgers, Rob J. J. H. Van Son. *Proceedings of INTERSPEECH 2013, Lyon, France, 2162-2166.*

RC and RvS posed the research questions. RC, RvS, FH and MvdB developed the study design. RvS prepared the speech data and wrote the scripts. RC and CvAB evaluated the recordings. RC and RvS performed the analysis of the data. RC and RvS wrote the first version of the text and all authors contributed to the final text.

6 The relationship between acoustic signal typing and perceptual evaluation of tracheoesophageal voice quality for sustained vowels

Renee P. Clapham, Corina J. Van As-Brooks, Rob J.J.H. van Son, Frans J.M. Hilgers, Michiel W.M. van den Brekel, *Journal of Voice, 29 (4), July 2015, 517.e23âĂŞ517.e29.*

RC and CvAB posed the research question and designed the experiments. FH and MvdB contributed to the study design. RvS prepared the speech stimuli and developed the technical set-up for the experiments. RC performed the data analysis. RC wrote the first version of the text and all authors contributed to the final version of the text.

7 Computing scores of voice quality and speech intelligibility in tracheoesophageal speech for speech stimuli of varying lengths

Renee P. Clapham, Jean-Pierre Martens, Rob J.J.H. van Son, Frans J.M. Hilgers, Michiel W.M. van den Brekel, Catherine Middag. *Computer Speech and Language, Accepted November 2015.*

RC and CM developed the research questions and developed the study design. JPM, FH, RvS, and MvdB contributed to the design of the study. CM developed the computer models. RC wrote the first version of the text and all authors contributed to the final version of the text.

8 General discussion

Renee P. Clapham wrote the text. Rob van Son, Catherine Middag, Frans Hilgers, and Michiel van den Brekel contributed to the final version of the text.

Funding

1 General introduction

Renee P. Clapham
RC was partially funded the Verwelius Foundation, Huizen, The Netherlands.

2 NKI-CCRT Corpus - Speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy

Renee P. Clapham, Lisette van der Molen, Rob J.J.H. van Son, Michiel W.M. van den Brekel, F.J.M. Hilgers. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12): 23-25 May, 2012, Istanbul, Turkey, 3350-3355.*

RC was partially funded by the Verwelius Foundation, Huizen, The Netherlands. RvS, LvdM, FH and MvdB were partly funded through a research grant that the Netherlands Cancer Institute receives from Atos Medical, Sweden. This grant from Atos contributes to the Head and Neck Oncology and Surgery department's existing infrastructure for health-related quality of life research.

3 Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer

Catherine Middag, Renee P. Clapham, Rob van Son, Jean-Pierre Martens. *Computer Speech and Language 28 (2), March 2014, 467-482.*

RC was partially funded by the Verwelius Foundation, Huizen, The Netherlands. RvS was partly funded through a research grant that the Netherlands Cancer Institute receives from Atos Medical, Sweden. This grant from Atos contributes to the Head and Neck Oncology and Surgery department's existing infrastructure for health-related quality of life research.

4 Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer

Renee P. Clapham, Catherine Middag, Frans J.M. Hilgers, Jean-Pierre Martens, Michiel W.M. van den Brekel, Rob J.J.H. van Son. *Speech Communication, 59, April 2014, 44-54.*

RC was partially funded by the Verwelius Foundation, Huizen, The Netherlands. FH, MvdB and RvS were partly funded through a research grant that the Netherlands Cancer Institute receives from Atos Medical, Sweden. This grant from Atos contributes to the Head and Neck Oncology and Surgery department's existing infrastructure for health-related quality of life research.

5 Automatic tracheoesophageal voice typing using acoustic features

Renee P. Clapham, Corina J. Van As-Brooks, Michiel W. M. van den Brekel, Frans J. M. Hilgers, Rob J. J. H. Van Son. *Proceedings of INTERSPEECH 2013, Lyon, France, 2162-2166.*

RC was partially funded by the Verwelius Foundation, Huizen, The Netherlands. CvAB, MvdB, FH and RvS were partly funded through a research grant that the Netherlands Cancer Institute receives from Atos Medical, Sweden. This grant from Atos contributes to the Head and Neck Oncology and Surgery department's existing infrastructure for health-related quality of life research.

6 The relationship between acoustic signal typing and perceptual evaluation of tracheoesophageal voice quality for sustained vowels

Renee P. Clapham, Corina J. Van As-Brooks, Rob J.J.H. van Son, Frans J.M. Hilgers, Michiel W.M. van den Brekel, *Journal of Voice, 29(4), July 2015, 517.e23âĂŞ517.e29.*

RC was partially funded by the Verwelius Foundation, Huizen, The Netherlands. CvA, RvS, FH and MvdB were partly funded through a research grant that the Netherlands Cancer Institute receives from Atos Medical, Sweden. This grant from Atos contributes to the Head and Neck Oncology and Surgery department's existing infrastructure for health-related quality of life research.

7 Computing scores of voice quality and speech intelligibility in tracheoesophageal speech for speech stimuli of varying lengths

Renee P. Clapham, Jean-Pierre Martens, Rob J.J.H. van Son, Frans J.M. Hilgers, Michiel W.M. van den Brekel, Catherine Middag. *Computer Speech and Language, Accepted November 2015.*

RC was partially funded by the Verwelius Foundation, Huizen, The Netherlands. RvS, FH and MvdB were partly funded through a research grant that the Netherlands Cancer Institute receives from Atos Medical, Sweden. This grant from Atos contributes to the Head and Neck Oncology and Surgery department's existing infrastructure for health-related quality of life research. CM was partially funded by 'Kom op tegen Kanker' the campaign of the Flemish League Against Cancer foundation.

8 General discussion

Renee P. Clapham

RC was partially funded by the Verwelius Foundation, Huizen, The Netherlands.

Table of Contents

List of Figures

List of Tables

1 General introduction

Cancer of the head and neck and its medical treatment and management, often has long-term negative consequences on the structures and tissues involved in a person's swallowing, speech and voice production. As a result, the way a person looks, sounds, talks and eats may change. For the speech pathologist, evaluation of swallowing, speech and voice is an important part of patient management and is necessary for documenting an individual's long-term outcome (Verdonck-de Leeuw et al., 2007a). An important aspect of documenting outcomes is measurement of functional speech and voice throughout the treatment process (Verdonck-de Leeuw et al., 2007b).

When assessing voice, a multidimensional approach combining acoustic, imaging, aerodynamic and patient-reported data in addition to perceptual evaluation is recommended (Dejonckere, 2010a,b; Dejonckere et al., 2001). Perceptual assessment of speech can include the components such as respiration, phonation/vocal quality, resonance, articulation, loudness, and prosody (Freed, 2000; Hodge and Whitehall, 2010). Rating scales of voice quality can consider the parameters *G* overall grade or severity, *R* roughness, *B* breathiness, *A* asthenia and *S* strain per the GRBAS scale (Hirano, 1981), or the parameters intelligibility, noise, fluency and voicing per the INFVo scale (Moerman et al., 2006).

Although the exact speech and voice assessment protocol can vary according to the patient's presenting symptoms, medical history and the preferred protocol used within an institution, perceptual evaluation of speech intelligibility and voice quality are common components. An advantage of perceptual evaluation is that the equipment and time demands are low and this, as Moerman et al. (2014) noted, is important if a protocol is to be clinically feasible. In this study, we primarily focus on the global perceptual components phonation/voice quality and speech intelligibility.

Articulation is the process of shaping the airstream into speech units by using articulators (i.e., tongue, lips) to block or restrict the airstream (Freed, 2000). Successful speech production requires the voluntary movement and control of the articulators with correct timing, force, placement and speed. Speech intelligibility, on the other hand, is a dynamic process between speaker and listener and a measure of speech intelligibility should reflect the degree to which the acoustic signal is decoded/identified by a listener (Kent et al., 1989; Yorkston et al., 1996).

Although articulation is a major component of intelligibility (see De Bodt et al., 2002; de Bruijn et al., 2009), intelligibility is more than an articulation assessment: mispronounced sounds produced in a way that make them (a) indistinctive from other sounds, or (b) inconsistent in their production confuse a listener. We include the aspect articulation as a variable in part of this study given this aspect of speech production is strongly correlated with speech intelligibility.

Phonation/voice quality reflects the integrity of the voicing source (vocal folds or neoglottis) for speech sound production based on the perceptual characteristics of the acoustic signal when compared to accepted and cultural norms for that group. An important aspect to note is that when considering speaker group, alaryngeal voice should not be perceptually compared to laryngeal voice (Moerman et al., 2006). As such, alaryngeal voice is unsuitable for evaluation using generic scales (Dejonckere, 2010a).

The need for an objective listener

In the clinical setting, speech intelligibility can be evaluated using a recognition task (e.g., open or closed identification) or a rating scale (e.g., 7-point scale). Voice quality is often evaluated on a rating scale. Although perceptual evaluation is cited as *the gold standard* (De Bodt et al., 2010; Moerman et al., 2006) and is a relatively fast procedure, perceptual evaluation requires involvement from a listener. The question is, however, who is the listener and how does the listener influence a measurement?

Measures can be influenced by a listener's background (Mády et al., 2003), familiarity with the speaker (Hustad and Cahill, 2003), familiarity with the test material (Yorkston and Beukelman, 1980), and knowledge of whether a speech sample is from before or after an intervention (Ghio et al., 2013). Unlike a human listener, tools allowing automatic evaluation for clinical purposes are not vulnerable to these influences. In the clinical setting, there is a need for a clinician/listener to evaluate quality measures of speech and voice with minimal influence from the listener. Evaluation of speech and voice quality by means of computerized assessment models may provide an objective and reliable adjunct to a clinician's subjective evaluation(s).

Study aims

The objective of this thesis is to investigate whether and how existing automatic evaluation tools can be adapted for clinical use in measuring voice quality and speech intelligibility of patients after treatment for head and neck cancer. Our primary aim is to investigate which tool, or parameters within a tool, can be used for Dutch speakers with head and neck cancer with regard to evaluating speech intelligibility and voice quality.

We envisaged that if such a tool could be adapted, it might assist in long-term patient monitoring (i.e., detection of therapy-induced changes), allow comparison of outcome measures and, in the clinical situation, act as an adjunct to a clinician's perceptual evaluation.

If these objectives are met, follow-up research could include investigate whether (1) automatically generated output is sensitive to changes as a result of speech therapy and/or surgical reconstruction technique, (2) automatic evaluation tools can be applied to the clinical environment to assist clinicians in designing therapy plans and (3) automatic tools can be adapted easily to cross-linguistic situations.

To appreciate the difficulty applying automatic evaluation tools in the clinical setting, we first discuss the general changes to speech and voice that occur with head and neck cancer before discussing existing automatic tools to evaluate speech intelligibility and voice quality.

1.1 Head and neck cancer

Head and neck cancer types can be broadly classified according the anatomic region of the lesion. As illustrated in Figure 1.1, the three main types are (1) oral cancer (comprising the tongue and floor of the mouth), (2) pharyngeal cancer (comprising the nasopharynx, hypoharynx and oropharynx, which includes the base of tongue) and (3) laryngeal cancer (comprising the larynx). Given that normal speech and voicing requires air movement from the lungs to the oral cavity and nasal cavity via the larynx and pharynx, it is not surprising that cancer in these regions and its subsequent medical treatment result in complex changes to the anatomy and physiology of this pathway.

Before medical management of the cancer, structural changes can impact on speech and voice and people can present with pain, altered sensation and dif-

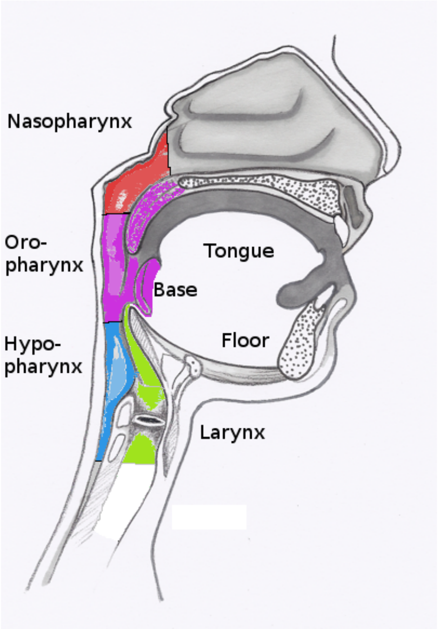


Figure 1.1: The three main anatomic locations of cancer of the head and neck: oral cancer (white, includes tongue and floor of mouth), laryngeal cancer (green) and pharyngeal cancer. Pharyngeal cancer comprises the nasopharynx (red), oropharynx (purple, also includes base of tongue) and hypopharynx (blue).

ficulty swallowing (Cnossen et al., 2012; Hoofd-Halstumoren, 2004; Kazi et al., 2008b; Pederson et al., 2010). Although changes in voice quality are not frequently reported when tumors are above the level of the larynx, it may be that changes in voice quality due to lifestyle factors associated with head and neck cancer, such as smoking and alcohol use (Jacobi et al., 2010c) mask pathology-related changes. When tumors are at the level of the larynx, people can present with increased vocal effort, breathiness and hoarseness as tumor(s) can limit the movement of the vocal folds and/or cause changes to airflow (Jacobi et al., 2010b).

Medical management for head and neck cancer can be surgical (i.e., removal of tissue) or non-surgical (e.g., radiation therapy) or a combination of both surgical and non-surgical, termed multi-modality treatment. Surgical treatment involves tissue removal and often also includes reconstruction, a process in which nerves involved in speech and swallowing can be compromised (Korpijaakko-Huuhka et al., 1998). Surgical treatment can result in decreased control and movement of structures requires for speech, such as the tongue (Hoofd-Halstumoren, 2004). Likewise, radiation therapy can result in decreased movement and strength of the articulators as well as other changes to the mucous membranes in the oral cavity (Hoofd-Halstumoren, 2004; Weber et al., 2010).

After medical treatment, many patients report speech and swallowing difficulties (Cnossen et al., 2012; Oozeer et al., 2010) and there is an association between decreased speech intelligibility and decreased quality of life for people treated for head and neck cancer (Meyer et al., 2004). Long-term tissue changes, such as radiation-induced scarring, can continue to negatively impact speech and voice (Kazi et al., 2008b; Kraaijenga et al., 2016). Larger tumors, advanced tumors, radiotherapy and extensive resections are associated with poorer speech outcomes (de Bruijn et al., 2009; Furia et al., 2001; Korpijaakko-Huuhka et al., 1998; Mády et al., 2003; Nicoletti et al., 2004; Zuydam et al., 2005) and treatment effects are associated with decreased quality of life (Weber et al., 2010).

We restrict our discussion on speech and voice changes to two groups: people who undergo the non-surgical combination of radiation therapy and chemotherapy (Section 1.1.1) and people who undergo the surgical procedure total laryngectomy (TL) (Section 1.1.2) . This division reflects the data and speech material available to our research group and discussed in this thesis.

1.1.1 Non-surgical treatment: chemoradiation therapy

Depending on tumor location and size, non-surgical cancer management involves radiotherapy with or without chemotherapy. One of the most widely applied protocols is **concomitant chemoradiation therapy (CCRT)**; when radiotherapy is administered simultaneously with chemotherapy. Although non-surgical management is viewed as an organ preservation treatment, it is not synonymous with the preservation of organ function as speech, swallowing and voice are often negatively impacted by treatment (see review by Jacobi et al., 2010a).

Before treatment, however, changes may already be present in speech and voice. Laryngeal tumors can cause perceptual changes to phonation/voice quality (see review paper by Jacobi et al., 2010a and subsequent studies Jacobi et al., 2010b) and tumors in the speech tract can impact on aspects of speech production such as tongue movement and precision (articulation, especially when tumor is in the base of tongue) and velum control and movement (nasality, especially when tumor is in the nasopharynx or oropharynx) (Jacobi et al., 2010b, 2013; Kraaijenga et al., 2015).

Articulation difficulties are associated with decreased tongue motion, particularly for sounds that require elevation of the tongue tip and control to create speech sounds with complete or partial constriction (e.g., /t, s, l/) (Bressmann et al., 2004; Jacobi et al., 2013; Korpijaakko-Huuhka et al., 1998), decreased range of motion for the tongue (de Bruijn et al., 2009; Jacobi et al., 2013; Korpijaakko-Huuhka et al., 1998; Whitehill et al., 2006) and increased nasality (Jacobi et al., 2013). Recent studies from our department support the notion that changes in tongue mobility play a central role of in speech intelligibility outcomes after treatment for head and neck cancer (Jacobi et al., 2010a, 2013, 2015a; van der Molen et al., 2012).

The general trend is that function (speech, voice, swallowing) can be impaired before treatment, can decrease during treatment and, although function improves in the first year, impairments can be long-term (i.e., $\geqslant$ 5 years) (Jacobi et al., 2010a, 2013; Kraaijenga et al., 2016, 2015; Pederson et al., 2010; van der Molen et al., 2012). The effect and impact of treatment varies depending on tumor location, associated radiation fields and radiation dosage (Jacobi et al., 2010a,b; Kraaijenga et al., 2016, 2015; van der Molen et al., 2012).

When the radiation field includes the jaw and tongue, changes in movement and strength of the tongue and jaw, changes in saliva production and consistency, and inflammation of the mucous membranes can be expected (Hoofd-Halstumoren, 2004; Mowry et al., 2006; Pederson et al., 2010; Weber et al., 2010). These changes can lead to impairments in speech and swallowing. Short-term, speakers treated for oropharyngeal cancer have more impaired articulatory precision compared to other tumor locations (Jacobi et al., 2013). Long-term many acoustic measures are similar to baseline and this positive outcome is reflected in speaker self-reported voice handicap scores (Kraaijenga et al., 2015).

When the radiation field includes the larynx, treatment effects result in changes to the vocal folds (such as vibration patterns and movement), which are perceived as impaired vocal quality. Radiation fields encompassing the larynx are not restricted to laryngeal cancer. For example, 90% of the study group discussed in Kraaijenga et al. (2015) received $\geqslant$ 43.5 Gy to the larynx despite only 36% of the study group being treated for hypopharyngeal or laryngeal cancer.

When chemotherapy is included in treatment protocols, radiation side-effects can become more pronounced (Hoofd-Halstumoren, 2004; Kelly, 2007).

1.1.2 Surgical treatment: total laryngectomy

Background

Although medical treatment options vary for early stage laryngeal cancer, total laryngectomy (TL) with or without additional non-surgical treatment (i.e., radiation therapy) remains the standard treatment for infiltrative advanced laryngeal cancer (Elmiyeh et al., 2010; Timmermans et al., 2015). TL involves the surgical removal of the larynx, epiglottis, hyoid bone, thyroid and the two top rings of the trachea (Labaere and Laeremans, 2009). This means that the connection between the oral cavity and the lungs is severed: the person no longer breathes through his/her nose or mouth as the trachea (airway) is pulled forwards to the base of the neck and breathing now occurs through a created, permanent stoma.

With the removal of the larynx, the vibratory sound source for speech is also removed. Voice restoration is either via an external vibratory source (e.g., an electrolarynx placed against the neck) or an internal vibratory source. This internal voicing option involves the vibration of a new voicing source, the neoglottis, which is located within the pharyngeal cavity. To redirect pulmonary airflow towards the neoglottis (also termed the pharyngoesophageal segment), a puncture between the posterior wall of the trachea and the anterior wall of the esophagus is surgically created and a device is placed in this opening. This device, referred to as a voice prosthesis, connects the trachea and the esophagus and ensures that the puncture remains open and that movement between the two cavities is in one direction (i.e., air can flow from the trachea to the esophagus via the prosthesis but food and fluids can not pass from the esophagus to the trachea). See Figure 1.2 for a schematic illustration.

When the tracheostoma is occluded, pulmonary air is redirected through the voice prosthesis to the esophagus where it passes the neoglottis and sets the mucosa into vibration. This type of voice and speech restoration is termed **tracheoesophageal (TE)** speech or TE voice. Adequate phonation requires the neoglottis vibrate and that the speaker has some control over the neoglottis. This ability depends on the structure and tension/movement of the neoglottis, however, these aspects among speakers (Schuster et al., 2005; Van As et al., 2004). The speaker can manipulate this vibrating signal further in the speech tract via normal articulation.

TE speech and voice

Although TL does not involve the speech articulators, speakers can have decreased speech intelligibility after surgery (Bussian et al., 2010; D'Alatri et al., 2012; Jongmans et al., 2006; Searl et al., 2001). This decrease is often attributed to difficulty

- producing contrasts voiced/voiceless sounds (e.g., /t/ is the voiceless partner of /d/);
- manipulating airflow to create plosive sounds (e.g., /p,b,t,d/ require a momentary blockage of airflow) and fricative sounds (e.g., /f, v, s/ require constriction without blockage of airflow); and
- producing sounds in certain locations (e.g., Dutch sound /h/ is produced in the glottis and this sound is often incorrectly perceived by listeners)

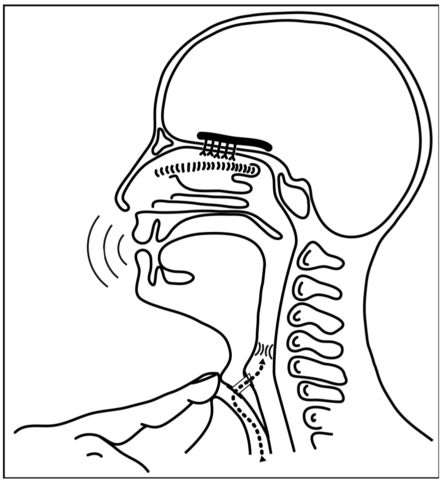


Figure 1.2: Schematic depiction of tracheoesophageal communication. The dashed line indicates the direction of pulmonary airflow through the one-way voice prosthesis towards the neoglottis. This redirected air sets the mucosa of the neoglottis into vibration and the resulting sound is further manipulated in the speech tract.

(Jongmans et al., 2006, 2010; Moerman et al., 2004; Searl et al., 2001; van Rossum et al., 2009).

TE voice is described as low, breathy, weak, gurgly, bubbly and unsteady (Kazi et al., 2008a; Lundström et al., 2008; van As-Brooks, 2008; van Rossum et al., 2009) and only a small proportion of TE speakers self-report good voice quality (D'Alatri et al., 2012). There is a relationship between voice quality and the tonicity of the neoglottis with moderate to good voice quality associated with a hypotonic-normotonic neoglottis (Lundström et al., 2008) and poor voice quality associated with a lack of tonicity (Op de Coul et al., 2003; van As-Brooks et al., 2005). The relationship between this variability and the reconstructive surgical technique applied is still unclear (Jacobi et al., 2015b; van As et al., 1999).

When physiological or imaging data is combined with acoustic information, the results show that fundamental frequency for TE speakers can range from approximately 80 to 200 Hz, regardless of speaker gender (Kazi et al., 2009, 2008a; Lundström and Hammarberg, 2011; Lundström et al., 2008; Schuster et al., 2005). This means that fundamental frequency of female TE speakers is generally within the range of male TE speakers. Voice quality can be perceived as less favorable for female speakers if the listener is aware of the speaker's gender (Eadie and Doyle, 2004).

In terms of clinical voice evaluation, including acoustic data in a multidimensional approach can be challenging as the irregularities in vibrating characteristics of the neoglottis mean that the common pitch-detection algorithms of general acoustic programs fail when the signal has low/no fundamental frequency or has high levels of noise. In this situation, a clinician could consider incorporating a visual evaluation of a voice sample in a protocol, termed **acoustic signal typing (AST)**.

AST was developed as a method of categorizing speech signals for laryngeal voices and was adapted by van As-Brooks et al. for alaryngeal speakers. Using AST, TE voices are categorized into four types reflecting the stability of the acoustic signal. The suggestion is that AST may be a useful indicator of perceptual voice quality as there is a relationship between AST and perceptual ratings of TE voice quality (D'Alatri et al., 2012; van As-Brooks et al., 2006).

1.2 Tools for automatic evaluation

In the last decade, the application of speech technology to perform *perceptual like* evaluation has become a research area of interest and results have been published on several populations. Unlike commercial tools using automatic speech recognition where the goal is to achieve maximum recognition, in the clinical setting the primary goal is to develop tools that reflect perceptual ratings. In other words, to predict how a listener would evaluate a voice or speech sample.

This thesis is not about developing new speech recognition tools or comparing acoustic or language models within the tools; this thesis is about the application and/or expansion of existing tools with a view towards implementation in the clinical setting, specifically for speech pathologists working with speakers treated for head and neck cancer. To this end, an informal literature search was undertaken to identify key stakeholders in this area and assess the current standing of the clinical application of automatic evaluation of speech intelligibility and voice quality.

1.2.1 Literature review

A search in the PubMed database using the search strategy displayed in Figure 1.3, yielded 96 papers of which 7 involved automatic prediction of listener-derived perceptual scores for voice quality and/or speech intelligibility. In addition, several journal papers were included by authors identified in the search but whose papers were not captured in the search strategy. Note that conference proceedings were not included as additional papers as these papers often presented preliminary data of a later-published work. Table 1.4 lists a detailed description of the papers.

The majority of papers investigated automatic evaluation of speech intelligibility (79%), two papers investigated automatic evaluation of voice quality

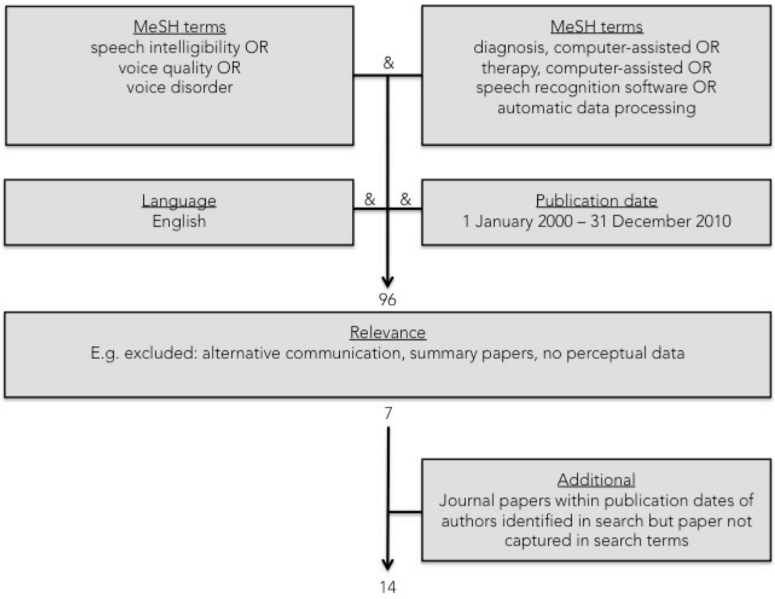

MeSH terms
speech intelligibility OR
voice quality OR
voice disorder
&
MeSH terms
diagnosis, computer-assisted OR
therapy, computer-assisted OR
speech recognition software OR
automatic data processing
Language
English
&
&
Publication date
1 January 2000 – 31 December 2010
96
Relevance
E.g. excluded: alternative communication, summary papers, no perceptual data
7
Additional
Journal papers within publication dates of authors identified in search but paper not captured in search terms
14

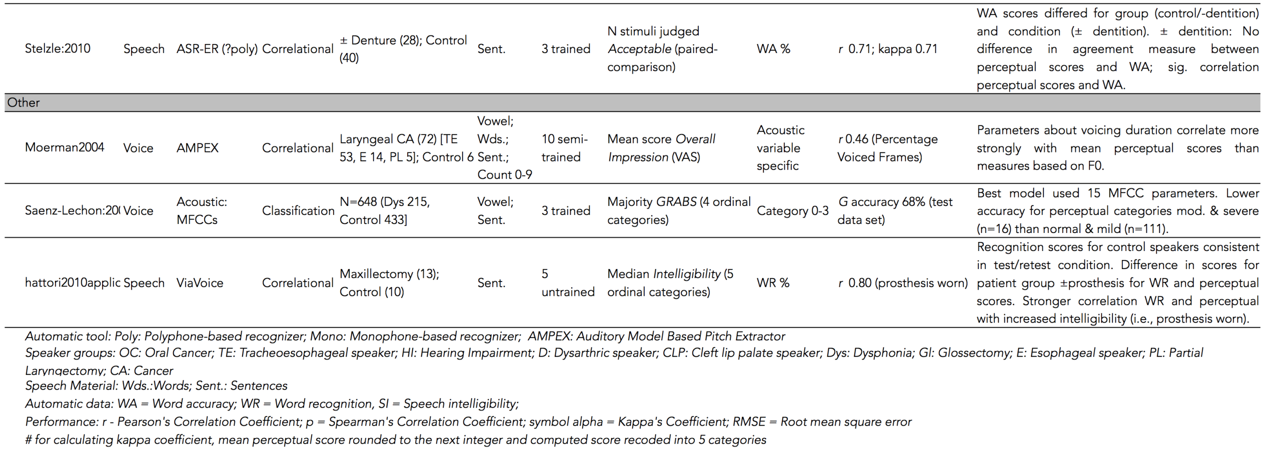
Figure 1.3: Illustration of the PubMed literature review search strategy

(14%) and one paper investigated both speech and voice quality (7%). Of the papers measuring speech intelligibility, one study utilized a commercially available automatic speech recognizer (ASR) (Hattori et al., 2010) and the remaining used ASRs developed by research groups at universities. The three studies with voice quality data utilized extracted acoustic data with dedicated software such as AMPEX (Moerman et al., 2004) or a prosody module that could be combined with ASR data (Haderlein et al., 2007b).

Speech corpora are often re-used within research groups. The largest speaker groups are speakers with dysarthria (n=60) or a hearing impairment (n=42), speakers treated for oral cancer (n=46) or cleft lip and palate (n=31), or speakers who use TE speech (n=18-41). With the exception of the cleft lip and palate speakers, all speakers were adult. One study included Japanese speakers (Hattori et al., 2010), another Spanish speakers (Sáenz-Lechón et al., 2006). Given that the majority of the research comes from research groups in Germany and in Belgium, the remaining papers included German or Flemish/Dutch speakers.

Speech intelligibility data includes intelligibility ratings made on a 5-point scale for sentence level and word level speech material (Haderlein et al., 2007b, 2009; Hattori et al., 2010; Maier et al., 2007, 2009, 2010; Schuster et al.,

Author	Variable of	Automatic tool	Study design	Speaker group(s) (N)	Speech material	Listeners	Perceptual data	Automatic data	Performance	Conclusion
ELIS & ESAT										
van2009speech	Speech	ASR-ELIS, ASR-ESAT	Prediction	N=211 [Dysarthria 60, Control 51, Other 100]	Wds.	1 trained	% correct phonemes	SI %	*r* 0.94 (PMF+PLF)	Inclusion of features from two alignment systems (PMF-ESAT + PLF-ELIS) results in stronger performance than single systems.
middag2009auto	Speech	ASR-ELIS, ASR-ESAT	Prediction	N=211 [Dysarthria 60, HI 42, TE 37, Control 51, Others 21]	Wds.	1 trained	% correct phonemes	SI %	General: RMSE 7.9 (PLF+CD-PLF); Disorder specific: RMSE < 6.0	CD-PLFs (ELIS) model achieved comparable performance to PMF (ESAT) +PLF (ELIS) but is computationally simpler.
PEAKS, ASR-ER, & Prosody module										
schuster2006intel	Speech	ASR-ER (mono)	Correlational	TE (18); Control (18)	Sent.	5 trained	Mean *Intelligibility* (5-pt scale)	WA %	*r* 0.84, kappa$^{\#}$ 0.43	Sig difference in WA between speaker groups. TE: Strong correlation WA & mean rating; agreement WA & mean rater similar to level among the raters (67% rounded data correct)
schuster2006eval	Speech	ASR-ER (poly)	Correlational	CLP (31)	Wds.	3 semi-trained	Mean *Intelligibility* (5-pt scale)	WA %	*r* 0.90, kappa$^{\#}$ 0.52	Strong correlation WA & mean perceptual score. Agreement WA and mean rater similar to level among the raters.
Haderlein:2007dc	Speech & Voice	ASR-ER + Prosody	Correlational	TE (18); Control (18)	Sent.	5 trained	Mean *Intelligibility* (5-pt scale); *Overall Quality* (VAS)	Acoustic variable specific	*Intelligibility r* < 0.70; *Overall Quality r* 0.72-0.78	Largest difference in prosody feature TE vs Control for average pause duration before current word. TE: r < .7 for *Intelligibility* and prosody features; r > .70 for *Overall quality* and some duration and pause features.
Maier:2007	Speech	ASR-ER (poly) + Prosody	Prediction	TE (41); CLP (31)	TE: Sent.; CLP: Wds.	5 trained	Mean Intelligibility (5-pt scale)	WA %; WR %; Acoustic variable specific.	TE *r* 0.90; CLP *r* 0.86	Prosody features + WA improved peformance for TE but not CLP. WA stronger correlate with *intelligibility* than WR and *intellgibility.*
windrich2008autc	Speech	ASR-ER (poly)	Correlational	OC (46); Control (40)	Sent.	4 trained	Mean Intelligibility (5-pt scale)	WR %	*r* 0.93; kappa$^{\#}$ 0.58	Significant difference in WR between speaker groups. OC: Strong correlation and agreement WA and average perceptual score. Agreement coefficient WR and perceptual similar to that of among the raters.
haderlein2009apr	Speech	ASR-ER (poly; mono)	Correlational	TE (41)	Sent.	5 trained	Mean *Intelligibility* (5-pt scale)	WA %	*r* 0.87 (poly)	Stronger correlations between automatic and mean perceptual scores with better signal quality. Polyphone-based recogniser stronger correlations with perceptual scores than monophone-based recogniser.
maier2009peaks	Speech	PEAKS [ASR-ER (poly)] + Prosody	Prediction	TE (41); CLP (31)	TE: Sent.; CLP: Wds.	5 trained	Mean *Intelligibility* (5-pt scale)	WA %; WR %; Acoustic variable specific.	TE *r* 0.90 (WA+Prosody); CLP *r* 0.87 (WA+Prosody)	Inclusion of features from two systems (WA+Prosody) results in slight increases (.03 TE, .01 CLP) in performance than single-system.
maier2010automa	Speech	PEAKS [ASR-ER (poly)]	Correlational	TE (41); OC (49); Control (40)	Sent.	TE 5 trained; OC 4 trained	Mean *Intelligibility* (5-pt scale)	WR %	OC *p* 0.90; TE *p* 0.83	Increase in WR% with growing *n*-gram language model does not equate with increased correlation with perceptual scores. Strongest correlation WR and perceptual scores for unigram model. Unigram: Sig. difference WR% between TE/Control and OC/Control.



Stelzle:2010	Speech	ASR-ER (?poly)	Correlational	± Denture (28); Control (40)	Sent.	3 trained	N stimuli judged *Acceptable* (paired-comparison)	WA %	*r* 0.71; kappa 0.71	WA scores differed for group (control/-dentition) and condition (± dentition). ± dentition: No difference in agreement measure between perceptual scores and WA; sig. correlation perceptual scores and WA.
Other										
Moerman2004	Voice	AMPEX	Correlational	Laryngeal CA (72) [TE 53, E 14, PL 5]; Control 6	Vowel; Wds.; Sent.; Count 0-9	10 semi-trained	Mean score *Overall Impression* (VAS)	Acoustic variable specific	*r* 0.46 (Percentage Voiced Frames)	Parameters about voicing duration correlate more strongly with mean perceptual scores than measures based on F0.
Saenz-Lechon:20(	Voice	Acoustic: MFCCs	Classification	N=648 (Dys 215, Control 433]	Vowel; Sent.	3 trained	Majority *GRABS* (4 ordinal categories)	Category 0-3	*G* accuracy 68% (test data set)	Best model used 15 MFCC parameters. Lower accuracy for perceptual categories mod. & severe (n=16) than normal & mild (n=111).
hattori2010applic	Speech	ViaVoice	Correlational	Maxillectomy (13); Control (10)	Sent.	5 untrained	Median *Intelligibility* (5 ordinal categories)	WR %	*r* 0.80 (prosthesis worn)	Recognition scores for control speakers consistent in test/retest condition. Difference in scores for patient group ±prosthesis for WR and perceptual scores. Stronger correlation WR and perceptual with increased intelligibility (i.e., prosthesis worn).

Automatic tool: Poly: Polyphone-based recognizer; Mono: Monophone-based recognizer; AMPEX: Auditory Model Based Pitch Extractor

Speaker groups: OC: Oral Cancer; TE: Tracheoesophageal speaker; HI: Hearing Impairment; D: Dysarthric speaker; CLP: Cleft lip palate speaker; Dys: Dysphonia; Gl: Glossectomy; E: Esophageal speaker; PL: Partial Laryngectomy; CA: Cancer

Speech Material: Wds.:Words; Sent.: Sentences

Automatic data: WA = Word accuracy; WR = Word recognition, SI = Speech intelligibility;

Performance: r - Pearson's Correlation Coefficient; p = Spearman's Correlation Coefficient; symbol alpha = Kappa's Coefficient; RMSE = Root mean square error

# for calculating kappa coefficient, mean perceptual score rounded to the next integer and computed score recoded into 5 categories

Figure 1.4: Overview of the study design, automatic tool used, speaker groups and numbers, performance outcome and main conclusion of the papers included in the literature review

2006a,b; Windrich et al., 2008) and the percentage of correctly identified phonemes for word level material (Middag et al., 2009a; Van Nuffelen et al., 2009). The three papers with voice quality information include data from VAS (Haderlein et al., 2007b; Moerman et al., 2004) and voice evaluation according to the GRBAS rating scale (Sáenz-Lechón et al., 2006).

1.2.2 The three main automatic systems

Three recognition systems are predominately used by the research groups: ASR-ELIS (developed at the Department of Electronics and Information Systems, Ghent University, Belgium), ASR-ESAT (developed at the Department of Electrical Engineering, University of Leuven, Belgium) and ASR-ER (developed at the Chair of Pattern Recognition at the University of Erlangen-Nurnberg, Germany). These ASRs can be used as stand-alone systems or included as parts of a software package. The German Program for Evaluation and Analysis of all Kinds of Speech Disorders, referred to as PEAKS, uses ASR-ER as the basis for its analysis. Similarly, ASR-ESAT and ASR-ELIS are used to automate the Dutch Intelligibility Assessment, referred to as the DIA.

As research has progressed, various acoustic and language models have been investigated. Clinical application requires a system's output reflect perceptual scores; complicated acoustic or language models do not necessarily entail stronger system performance when compared to listener-derived scores. This is because performance for the clinical application of an automatic system is not based on the absolute recognition accuracy, but rather performance is based on the strength of the relationship to perceptual scores, the ability to classify into perceptual categories or to predict perceptual scores.

1.2.2.1 PEAKS

Over half the papers identified in the literature search used ASR-ER either as a stand-alone system or as part of the Program for Evaluation and Analysis of all Kinds of Speech Disorders, referred to as the PEAKS software package. PEAKS uses a word recognizer, ASR-ER, to evaluate the recognition rate of words from *Nordwind und Sonne*. This German text contains 108 words (71 unique) and all German phonemes.

ASR-ER Phoneme recognition is supported by Hidden Markov Models (HMM) that describe the likelihood an analyzed signal matches a phoneme. The majority of studies using this ASR include polyphone-based models because work supports these acoustic models lead to stronger correlations with perceptual data compared with monophone-based acoustic models (Haderlein et al., 2009).

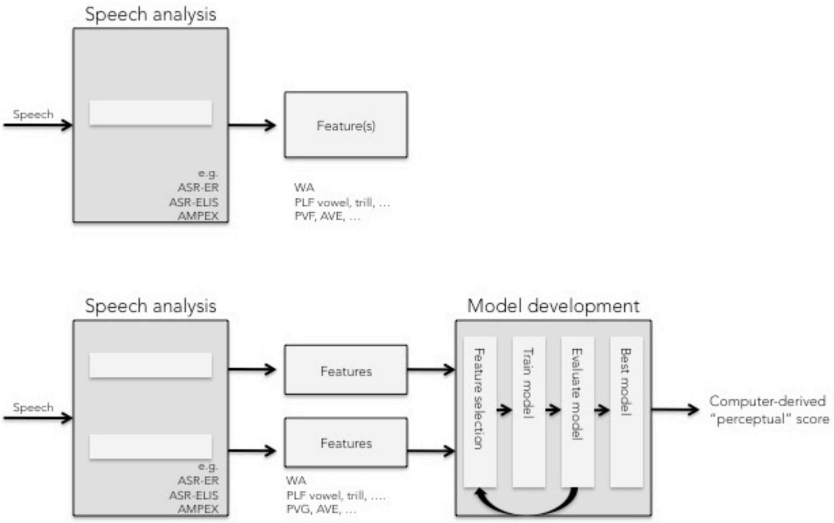


Figure 1.5: Schematic depiction of one-stage (top) and two-stage (bottom) automatic systems

The ASR is supplied with a lexicon of the words in the spoken passage as well as words incorrectly read by speakers (Haderlein et al., 2007b; Schuster et al., 2006a). The system is also be supported with language models, with the majority of studies providing a unigram language model as research from the group found that this model provided the strongest correlations with perceptual data (Maier et al., 2010). This means that the system is only provided the word frequencies within the text and not the patterns of words.

For each sentence in the passage, the ASR output is compared with the target allowing the calculation of two measures expressed as a percentage of the number of reference words: word accuracy (WA) and word recognition (WR). The difference between the two measures is that WA counts the correctly recognized words and penalizes for deleted, inserted or substituted words. WR only reflects the total number of correctly recognized words. WA is reported to have stronger correlations with perceptual scores than WR (Maier et al. 2007, 2009; also see Haderlein et al. 2007a; Riedhammer et al. 2007).

Prosody module To extend the performance capability of the system and expand its use to voice quality, a prosody module is included to extract acoustic and prosodic information from the speech signal. This module extracts two types of information: global prosody features and local prosody features. Global features are measured over the entire utterance and are based on fundamental frequency fluctuations (jitter), intensity fluctuations (shimmer) and the number of voice/unvoiced segments. Local features are measures of duration, energy and fundamental frequency over different reference points (e.g., current word; end of the word). Reference points are determined based on word boundaries identified by the ASR. The prosody features are reported as averages, maximums, minimums and standard deviations.

Prediction model To convert speaker features (e.g., WR, WA, any of the prosody features) to intelligibility scores, the features need to be mapped to perceptual scores. The PEAKS system uses support vector regression (SVR) to create a prediction model. Study designs from this research group have predominately used a leave-one-out strategy (Maier et al. 2009; see also Riedhammer et al. 2007) to develop and test the prediction models.

With this strategy, data from all but one speaker is regarded as training data. As illustrated in Figure 1.5, the first step is to identify a subset of the speaker features that correlate with the reference scores. In this case, the reference scores are the mean perceptual ratings from a group of raters. Features are selected and added to the model until performance no longer improves.

The subset of speaker features providing the strongest performance is used to train a prediction model and is validated on the left-out speaker (i.e., the speaker has a predicted score that was developed on all other speakers). This process is repeated until every speaker has been used as validation speaker. The measure of prediction accuracy used by this research group is the correlation (Pearson correlation coefficient and Spearman rank correlation) between the predicted scores and the mean perceptual scores. See Section 1.2.3 for information on performance.

1.2.2.2 Automated DIA

The original, manual version of the DIA requires a speaker reads 50 consonant-vowel-consonant (CVC) combinations while the clinician identifies the missing sound on a test sheet (e.g., ..op; n .. s). The DIA stimuli were developed so all Dutch consonants, vowels and diphthongs were included at least once in the items. The speech intelligibility score is the percentage of correctly identified sounds.

The computerized version of the DIA allows simultaneous evaluation by clinician and computer. Unlike the manual version in which only the target sound is included in the score, the ASRs used in the automatic version have access to all phonemes. The speech data undergoes acoustic signal analysis in which for consecutive, overlapping frames are analyzed. The acoustic models used in the ASR-ESAT and ASR-ELIS systems were trained on speech samples from control Flemish/Dutch speakers. For each frame, information about the energy and the shape of a segment is calculated. The second stage of processing is when the ASR analyzes this information to generate information on speaker features.

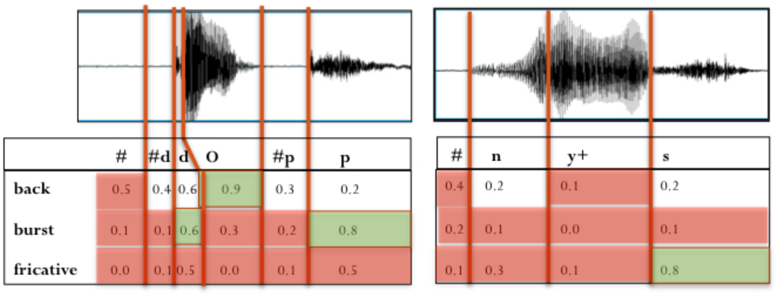
The two types of features used by this research group to develop prediction models are phonological-based features and phonemic/monophone features. These features can be derived either via a process of forced alignment between the speech material and the text or via a process that does not require forced alignment. For the literature review period in question, the forced-alignment approach was more established and as such, we discuss these features in more details. Although both ASR-ELIS and ASR-ESAT can provide phonological features (PLFs) and phonemic/monophone features (PMFs/MPFs) (in addition to WA), the combination of PLFs from ASR-ELIS and PMFs/MPFs features from ASR-EAST provide the strongest results (Middag et al., 2008).

ASR-ESAT: PMFs/MPFs These features are considered to reflect how well monophones such as /s/ or /A/ are realized by a speaker. Note that these features were originally termed phonemic features (PMFs), however this was later changed to monophone features (MPFs) (Middag et al., 2014). To generate the speaker features, the speech is aligned with the canonical transcription of what a speaker should have said. This alignment is supported by a semi-continuous HMM system and acoustic triphone models, meaning that co-articulation effects from sounds to the left and right are taken into account. The theory behind analyzing the phonetic segmentation made by the ASR is that it provides a richer way to characterize a speaker compared to word recognition.

PMFs/MPFs are calculated once all frames have been assigned a triphone state. For each frame identified as belonging to a certain phone, the average posterior probability over these frames is calculated. Meaning that for the 40 Dutch/Flemish phones, 40 PMFs/MPFs can be calculated with each feature value representing how well the phonemic feature was recognized over the entire utterance.

Each PMF/MPF for a given monophone has an associated value representing the average posterior probability for that monophone (calculated over all frames identified as belonging to that sound). High values (max score 1) indicate realizations similar to the acoustic model (i.e., accurately produced, easily identified) whereas low values (min 0) indicate a realization different to the acoustic model expected for that monophone.

The first row in Figure 1.6 displays the alignment between speech signal and target monophones. In the example, the target /d/ from the word *dop* is produced as a /t/. As such, its associated value (e.g., 0.3) would be low compared to the target /p/ realized as /p/ (e.g., 0.8). Note the figure does not display the corresponding PMF/MPF data.



	#	#d	d	O	#p	p
back	0.5	0.4	0.6	0.9	0.3	0.2
burst	0.1	0.1	0.6	0.3	0.2	0.8
fricative	0.0	0.1	0.5	0.0	0.1	0.5

#	n	y+	s
0.4	0.2	0.1	0.2
0.2	0.1	0.0	0.1
0.1	0.3	0.1	0.8

Figure 1.6: Example of the two types of PLFs: positive features (green) and negative features (red). See text for further details. Image used with permission from C. Middag and taken from Middag (2011)

ASR-ELIS: PLFs Phonological features reflect how well binary phonological categories related to manner (e.g., BURST), place (e.g., BILABIAL) and voicing (e.g., VOICED) are present or absent at the expected moments. After all speech frames are assigned a phone during the forced-alignment process, phonological features can be calculated (a) over the frames where the phonological feature should be present and (b) over the frames where the phonological feature should not be present. This results in two types of PLFs: positive PLFs indicate how much a feature is present when it should be present and negative PLFs indicate how much a feature is present when it should be absent. There are 24 phonological features each with a binary option (e.g., should be present /should be absent), which results in 48 PLFs to characterize Flemish/Dutch articulation patterns.

Figure 1.6 illustrates both the identification of PLFs and their calculation for two CVC targets *dop* (spoken as *top*) and *nuis*. Aligned under the speech signal are the respective target phones (e.g., vowel /O/ or closure for /#p/) and the figure displays three possible PLFs: BACK, BURST, FRICATIVE. High values (max 1) for BACK on the /O/ and FRICATIVE on the /s/ (note, these values are the average value for the overlapping frames within this segment) indicate a high probability that the phonological feature was present when it was expected to be present. The values for BURST around the target /d/ are lower (min 0) as the /d/ was produced as a /t/.

For each phone the posterior probability over the entire test set for each positive PLF is calculated and averaged for where the feature should be present (green cells) and for each negative PLF, where the feature should not be present (red cells). Note no calculation for vowel feature BACK for consonants. If the speaker had only produced these two test items, the value for +BURST and −BURST would be 0.7 and 0.1, respectively.

Feature expansion In Middag et al. (2009a), PLFs were extended to context-dependent PLFs (CD-PLFs). The authors hypothesized that speakers with impaired speech may have difficulty producing particular phonological classes in some phonemic contexts rather than across all contexts. In other words, a speaker may have difficulty producing a class of sounds in one sound environment more than in another. To achieve this, CD-PLFs are computed taking the properties of the surrounding phones into account.

Recognising that speech intelligibility measures derived from single words may not capture the speech of a person in a communicative setting and to accommodate reading errors, the PLFs were further extended in Middag et al. (2010) to alignment-free PLFs (ALF-PLFs). This approach does not require an alignment between the speech and text, and, indeed, no reference text is required.

During speech analysis stage for the aligned PLF system, the speech first undergoes short-term acoustic analysis (generating 12 MFCC coefficients and a log-energy) and this data is then provided to the ASR-ELIS together with the speech transcription. The ASR then aligns speech and transcript and calculation and extraction of the features can occur. In an alignment-free approach, however, the data from the initial acoustic analysis is directly converted into phonological feature information describing the feature over the entire utterance. This is achieved with a neural network to compute the posterior probabilities of the phonological properties. Unlike in the standard PLF output that has a single value per positive/negative property, the alignment-free method calculates 12 statistical measures per component, such as the mean value and standard deviation for a feature.

Prediction The second stage requires the speaker features (e.g., PLFs, MPFs/PMFs) be transformed into intelligibility scores that reflect perceptual intelligibility scores. As illustrated in Figure 1.2.3, this second stage requires selecting a subset of speaker features to train and develop a prediction model. The method used in the papers identified in the literature review use linear regression models to predict speech intelligibility. Note that Middag et al. (2010) reported no performance differences according to model type. When selecting speaker features for the feature subset, models can be limited to features from one ASR (e.g., model only able to select from the 40 MPFs) or features from multiple ASRs (e.g., may select from 40 MPFs and the 48 PLFs to create a MPF+PLF model).

Study designs from the research group used 5-fold cross-validation to identify feature subsets yielding optimal model performance. Feature selection was predominately performed using forward feature selection (Middag et al., 2008; Van Nuffelen et al., 2007, 2009) (c.f., forward and backward feature selection

Middag et al., 2009a). In this strategy, the data set is divided into five parts: four parts are used for feature selection and model training and the fifth part is used to test the identified strongest model. This process is repeated until all five parts of the data have been used four times in the training set and once in the test/validation set. Features are added to the model until performance no longer improves.

Earlier work reported performance as the Pearson correlation coefficient between predicted scores and perceptual scores (Middag et al., 2008; Van Nuffelen et al., 2007, 2009). Later work reported performance as the root mean square error (RMSE) between predicted and perceptual scores (Middag et al., 2009a, 2010). The authors argue the RMSE is a stronger measure of performance as it can be directly interpreted because it reflects the distance of predicted scores from the observed scores and that the RMSE is a stable measure when a prediction model is developed to cover a range in intelligibility scores but is tested on a smaller range (Middag et al., 2009a). Note that perceptual scores are percentage of correctly identified phonemes as perceived by a single rater.

1.2.3 Research trends

1.2.3.1 One-stage systems

Early research investigated system performance as (a) the relationship strength between a single speech analysis tool and perceptual scores and (b) the sensitivity of the automatic data to differentiate speech samples from control speakers and samples from a clinical population. The focus of these studies was to identify optimal acoustic models, language models and output data. Such studies utilized a one-stage system in which the output of the automatic tool (e.g., WA) is directly used and the perceptual scores are independent of the automatic output (see Figure 1.5 for schematic representation).

The majority of studies using the ASR-ER system involve one-stage processing. The drawback of this approach is that the speech recognition systems are trained on control speakers and performance is predominately measured as the strength of the relationship between automatic scores and perceptual scores (e.g., Moerman et al., 2004; Schuster et al., 2006a). This means that a word recognition rate of 80% does not infer that the speaker was evaluated as being 80% intelligible to a listener.

Data from a control speaker group is used to investigate whether automatically derived scores are sensitive to differences in control/normal versus altered speech or voice (e.g., Windrich et al., 2008) or the reliability of a system in a test/retest condition (Hattori et al., 2010). Two studies used a repeated-measures design where automatic scores for speech intelligibility for a speaker with and without dentition or a prosthesis were compared (Hattori et al., 2010;

Stelzle et al., 2010). None of the studies investigated whether automatically derived scores could track changes in speech or voice over time as a result of speech pathology intervention.

Performance In general, the correlation coefficient reported between one-stage systems and observed perceptual scores range r <0.70 to 0.93 for speech intelligibility (Haderlein et al., 2007b; Windrich et al., 2008) and r 0.46 - 0.78 for voice quality (Haderlein et al., 2007b; Moerman et al., 2004). Agreement correlation coefficients between automatic scores and mean perceptual data (i.e., comparing automatic results with an average rater) report κ values around 0.50 (see Table 1.4).

The results indicate that although the recognition rate of a system increases with an increase in language model complexity (*1*-gram, *2*-gram, *3*-gram language models), this does not equate to improved correlations with perceptual ratings (Maier et al., 2010). In general, acoustic models using polyphone-based recognizers achieve stronger correlation coefficients than monophone-based recognizers (Haderlein et al., 2009).

1.2.3.2 Two-stage systems

In order to have an automatically derived score that reflects that provided by a listener, a prediction model is required to map automatically derived data to perceptual scores. This becomes a supervised learning problem. This approach is used by the research group in Belgium (Middag et al., 2009a, 2010; Van Nuffelen et al., 2007, 2009) and in later work by the research group in Germany (Maier et al., 2007, 2009; Riedhammer et al., 2007).

The advantage of two-stage systems is that a subset of features from single or multiple systems can be combined in a prediction model. Although the underlying acoustic models used in the automatic tools discussed in this section are developed on control speakers, the prediction model interprets this data and applies it to a clinical population of speakers. By doing so, the automatic output can reflect speech intelligibility or voice quality scores or ratings as evaluated by listeners. In effect, the computer becomes an additional evaluator. Performance can then be evaluated as the accuracy of the prediction model against perceptual scores.

Performance One of the main trends is that inclusion of features from different systems results in performance scores that are stronger than the performance of the individual systems. Where previously the strongest correlations for TE speakers was using WA scores from ASR-ER, the combination of WA from ASR-ER and prosody information resulted in an increased correlation coefficient

(Maier et al. 2007, 2009, also see Riedhammer et al. 2007). This pattern was also reported in Van Nuffelen et al. (2009) in which features from ASR-ELIS combined with features from ASR-ESAT resulted in stronger performance than the individual systems (also see Middag et al., 2008, 2010).

The development of CD-PLFs is a possible refinement of the phonological features as take the surrounding sound environment into account. In a generalized prediction model (i.e., trained on a variety of speaker groups) in Middag et al. (2009a), a model only making use of the newer CD-PLFs achieved a performance accuracy that was better than models using only PLFs or MPFs/PMFs. Combing CD-PLF information, however, with the other two features resulted in a small, but not significant improvement in model performance when trained and evaluated on mixed-pathologies.

Performance also improved when prediction models were speaker-group specific (e.g., TE speakers) as opposed to general prediction models and the best combination of input speaker features varies by pathology (Maier et al., 2007, 2009; Middag et al., 2008, 2009a). Also see Riedhammer et al. (2007). This supports the hypothesis that speech/voice characteristics are pathology specific and can best be modeled using input features that capture the group in question. For pathology specific models, however, combining features from different systems does not always lead to improved model performance.

With the development of the alignment-free PLFs for Dutch/Flemish speakers and inclusion of prosody information for German speakers, greater opportunities become available for modeling speech and voice quality. Preliminary work by Middag et al. (2010) indicates that alignment-free features can be used to develop a reliable model, however, more data is required to assess the accuracy of prediction models using these features. The work completed by Maier and associates in 2009 indicates consistent correlations with perceptual ratings for running speech intelligibility and agreement values between automatic scores and the average rater are comparable with the level of agreement among a group of raters.

As far as we are aware, no work has been published on predicting voice quality scores although work has been published on the correlation between automatically-derived scores and perceptual ratings (Haderlein et al., 2007b; Moerman et al., 2004) and automatic classification (Sáenz-Lechón et al., 2006).

1.3 Automatic evaluation in the clinical situation

Applying speech technology within the area of speech and language is not a new concept and has been applied in the areas of pronunciation training for language learning (Neri et al., 2006) and speech training for speakers with neurological disorders (Beijer et al., 2010). There is a clinical need for automated tools to provide data that can be used to complement a clinician's subjective evaluation of voice quality and speech intelligibility. One of the advantages of an automated evaluation tool is that derived scores are not influenced by aspects such as familiarity with the speaker or whether a speech sample is from before or after an intervention: recognition scores remain constant in test/retest conditions (Hattori et al., 2010) and the consistent performance of prediction models for the same database of speakers (see performance data in Table 1.4) support the reliability of automatically derived measures.

Tools such as PEAKS and the DIA offer automatic analysis in real-time with clinicians only requiring a laptop/PC, internet connection and quality microphone (Maier et al., 2009; Middag et al., 2009b). The error rates for automatic evaluations can be as low as 8% (Middag et al., 2010) and, as seen in the literature review, computer-derived scores attain levels of reliability considered comparable to that of a group of raters.

The performance of such tools is promising, however, we identified several interconnected trends in the results of the literature that require consideration if automatic evaluation is to be considered in the clinical situation.

Perceptual scores The inter-rater variation in perceptual scores increases as speaker intelligibility/voice quality decreases, which means observed scores evaluated as having lower quality are more difficult to model. This is evidenced by a greater error between the predicted score and the observed score. In addition, most data sets used for model development are skewed and have fewer examples of speakers with low perceptual scores meaning that observed data points are not evenly distributed along the perceptual score continuum. This results in under-training for speech samples with lower perceptual qualities (also see Chapter 4 of this thesis).

Data set size Re-sampling strategies are necessary when it comes to the area of developing models for clinical speech and voice populations because of the relatively small size of speech material with perceptual data available to researchers. Ideally, data would be divided into training, validation and test sets where the proportions of severity are held constant over the sets and where the severities are frequent enough within each category to enable optimal training. Cross-validation strategies with small data sets is a common technique to maximize the size of the training and validation sets while keeping the overlap between sets as small as possible to minimize error (Alpaydin, 2010). Larger data sets, specifically larger data sets of specific clinical groups, would assist developing accurate and reliable models with strong generalization capabilities.

Tracking speaker trends Data sets including multiple recordings from speakers over time could allow the sensitivity of prediction models to be evaluated. To our knowledge, no automatic evaluation tools have included such speech material. If a model could track changes in speech or voice quality, it would offer clinicians a way to collect clinician-independent pre-treatment and post-treatment data. Beyond the use of automatic evaluation tools for therapy outcome measures, automatic tools could be used to follow a patient's progress (e.g., throughout speech pathology intervention(s) or to monitor long-term changes post medical intervention(s)). This is of particular importance in the area of head and neck oncology due to the long-term cancer treatment effects.

Global evaluation The goal of current prediction models has been to predict perceptual information related to speech and voice quality. Group specific models (e.g., TE speakers) provide opportunities for a more fine-grained exploration of voice or speech by considering which speaker-features a model selects. As noted by Middag et al. (2010), models select features that can often be linked to the speech characteristics of the speaker group, such as voicing and fricative production for TE speakers. By developing prediction models utilizing features related to the speech dimension, the suggestion is that clinicians may be able to access the profile of a speaker to characterize the nature of the speech difficulties. Theoretically, clinicians could then use speaker-profile information to support the identification of therapy goals. The risk is, however, that a cause-effect relationship could be linked to feature selection: Features are selected as a model inputs based on correlations and the discriminate strength of a feature and do not imply a causal relationship between feature and speech intelligibility or voice quality.

Summary

In general, the results of the literature review show that automatic evaluation of speech and voice quality by means of computerized assessment models is possible and may provide an objective and reliable adjunct to a clinician's subjective evaluation(s) in the clinical setting. For clinical implementation within the setting of head and neck oncology, the following needs to be considered:

1. Future investigation of automatic evaluation need to focus on detailed measures rather than global measures to ensure that

    (a) results are meaningful for patients and therapist (e.g., to provide feedback on production),

   (b) the tools can be used to follow the voice and speech characteristics of population sub groups such as oral cancer versus laryngeal cancer, and

   (c) the results of an individual patient can be followed.

2. Model performance for speakers with lower perceptual scores needs to be addressed

3. An automatic tool that measures voice quality using F0-based measurements may be unreliable. A non-F0 based measure should be included in automatic processes (see Jacobi et al., 2010c).

4. Voice measures should not be based on comparisons with control speakers because control speakers are not representative of these patients due to (a) a different vocal source and/or (b) lifestyle differences such as alcohol and smoking which cause changes to vocal quality.

1.4 Thesis outline

As stated in this chapter, the study aim of this thesis is to investigate whether and how existing automatic evaluation tools can be used in the clinical situation to measure voice quality and speech intelligibility of speakers after treatment for head and neck cancer. The goal is to develop models of these two variables so that objective, automatically derived quality scores can be used as an adjunct to the perceptual score provided by a clinician.

This thesis focuses on two distinct cohorts of head and neck cancer patients. For both cohorts extensive recording databases have been collected at the Netherlands Cancer Institute. The first cohort comprises patients with advanced head and neck cancer treated with organ-preserving concurrent chemoradiotherapy (CCRT). As discussed in Section 1.1.1, this treatment may negatively impact on speech and voice. The second cohort comprises patients treated for advanced or recurrent laryngeal cancer with total laryngectomy (TL). For these speakers, the insertion of a voice prosthesis can allow speech restoration via tracheoesophageal speech. The speech material and perceptual data for each database are described in **Chapter 2** (CCRT) and **Chapter 5** (TL) and all evaluation results are listed in the **Appendix**.

Chapter 3 presents the results of a novel method to evaluate the running speech intelligibility in the CCRT cohort and this method is extended in **Chapter**

4 to include evaluation of voice quality as well as articulation. In both chapters we also consider whether automatic scores can track changes over times.

In addition to describing the TL database, Chapter 5 investigates the possibilities of categorization of TE vowels according to signal types using acoustic information. The relationship between these categories and perceptual evaluation is further explored using a dedicated internet-based tool in **Chapter 6**. Using the same automatic tools from Chapters 3 and 4, **Chapter 7** presents the automatic assessment models for evaluating TE speech intelligibility and voice quality. In **Chapter 8** the results are discussed and related to recent findings in the literature.

References

E Alpaydin. *Introduction to machine learning*. Massachusetts Institute of Technology, 2nd edition, 2010.

LJ Beijer, T Rietveld, MM van Beers, RM Slangen, H van den Heuvel, BJ de Swart, and AC Geurts. E-learning-based speech therapy: a web application for speech training. *Telemed J E Health*, 16(2):177–180, 2010.

Tim Bressmann, Robert Sader, Tara L Whitehill, and Nabil Samman. Consonant intelligibility and tongue motility in patients with partial glossectomy. *J Oral Maxillofac Surg*, 62(3):298–303, Mar 2004.

Claudia Bussian, Dorit Wollbrück, Helge Danker, Esther Herrmann, Alexander Thiele, Andreas Dietz, and Reinhold Schwarz. Mental health after laryngectomy and partial laryngectomy: a comparative study. *Eur Arch Otorhinolaryngol*, 267:261–266, 2010. doi: 10.1007/s00405-009-1068-7.

Ingrid Cnossen, Remco de Bree, Rico Rinkel, Simone Eerenstein, Derek Rietveld, Patricia Doornaert, Jan Buter, Johannes Langendijk, C. Leemans, and Irma Verdonck-de Leeuw. Computerized monitoring of patient-reported speech and swallowing problems in head and neck cancer patients in clinical practice. *Supportive Care in Cancer*, pages 1–7, 2012. ISSN 0941-4355. URL `http://dx.doi.org/10.1007/s00520-012-1422-y`. 10.1007/s00520-012-1422-y.

Lucia D'Alatri, Francesco Bussu, Emanuele Scarano, Gaetano Paludetti, and Maria Raffaella Marchese. Objective and subjective assessment of tracheoesophageal prosthesis voice outcome. *Journal of Voice*, 26(5):607–613, 2012. doi: 10.1016/j.jvoice.2011.08.013.

M De Bodt, L Heylen, F Mertens, J Vanderwegen, and P Van de Heyning. *Stemstoornissen: Handeiding voor de klinische praktijk*. Garant, Antwerpen (Belgium), 3rd edition, 2010.

Marc S De Bodt, Huici Maria E Hernández-Díaz, and Paul H Van De Heyning. Intelligibility as a linear combination of dimensions in dysarthric speech. *J Commun Disord*, 35(3):283–92, 2002.

Marieke J de Bruijn, Louis ten Bosch, Dirk J Kuik, Hugo Quené, Johannes A Langendijk, C René Leemans, and Irma M Verdonck-de Leeuw. Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. *Folia Phoniatr Logop*, 61(3):180–7, 2009.

P. Dejonckere. Assessment of voice and respiratory function. In Marc Remacle and Hans Edmund Eckel, editors, *Surgery of Larynx and Trachea*, pages

11–26. Springer Berlin Heidelberg, 2010a. ISBN 978-3-540-79136-2. URL http://dx.doi.org/10.1007/978-3-540-79136-2_2.

P. Dejonckere. Voice evaluation and respiratory function assessment. In *Otorhinolaryngology, Head and Neck Surgery*, pages 563–574. Springer, 2010b.

P Dejonckere, Patrick Bradley, Pais Clemente, Guy Cornut, Lise Crevier-Buchman, Gerhard Friedrich, Paul Van De Heyning, Marc Remacle, and Virginie Woisard. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *European Archives of Oto-Rhino-Laryngology*, 258:77–82, 2001. ISSN 0937-4477. URL http://dx.doi.org/10.1007/s004050000299. 10.1007/s004050000299.

Tanya L. Eadie and Philip C. Doyle. Auditory-perceptual scaling and quality of life in tracheoesophageal speakers. *The Laryngoscope*, 114(4):753–759, 2004. ISSN 1531-4995. doi: 10.1097/00005537-200404000-00030. URL http://dx.doi.org/10.1097/00005537-200404000-00030.

B Elmiyeh, R C Dwivedi, N Jallali, E J Chisholm, R Kazi, P M Clarke, and P H Rhys-Evans. Surgical voice restoration after total laryngectomy: an overview. *Indian J Cancer*, 47(3):239–47, 2010.

D Freed. *Motor speech disorders: diagnosis and treatment*. Singular Publishing Group, 2000.

C L Furia, L P Kowalski, M R Latorre, E C Angelis, N M Martins, A P Barros, and K C Ribeiro. Speech intelligibility after glossectomy and speech rehabilitation. *Arch Otolaryngol Head Neck Surg*, 127(7):877–83, Jul 2001.

A Ghio, J Revis, S Merienne, and A Giovanni. Top-down mechanisms in dysphonia perception the need for blind tests. *Journal of Voice*, 27(4):481–485, 2013.

T Haderlein, K Reidhammer, A Maier, E Noth, Hikmet Toy, and F Rosanowski. An automatic version of the post-laryngectomy telephone test. In *Proceedings Text Speech and Dialogue*, pages 238–245, 2007a.

Tino Haderlein, Elmar Nöeth, Hikmet Toy, Anton Batliner, Maria Schuster, Ulrich Eysholdt, Joachim Hornegger, and Frank Rosanowski. Automatic evaluation of prosodic features of tracheoesophageal substitute voice. *European Archives of Oto-Rhino-Laryngology*, 264(11):1315–1321, 2007b. doi: DOI10.1007/s00405-007-0363-4.

Tino Haderlein, Korbinian Riedhammer, Elmar Nöth, Hikmet Toy, Maria Schuster, Ulrich Eysholdt, Joachim Hornegger, and Frank Rosanowski. Application of automatic speech recognition to quantitative assessment of tracheoesophageal speech with different signal quality. *Folia Phoniatrica et Logopaedica*, 61(1):12–17, 2009.

Mariko Hattori, Yuka I Sumita, Shinta Kimura, and Hisashi Taniguchi. Application of an automatic conversation intelligibility test system using computerized speech recognition technique. *Journal of prosthodontic research*, 54 (1):7–13, 2010.

M Hirano. *Clinical examination of voice*. Springer Verlag, 1981.

M Hodge and T Whitehall. *The handbook of language and speech disorders*, chapter Intelligibilty impairments, pages 99–114. 2010.

Nederlandse Werkgroep Hoofd-Halstumoren. *Richtlijn mondholte- en orofarynxcarcinoom*. Van Zuiden, Alphen aan den Rijn (Netherlands), 2004.

Katherine C Hustad and Meghan A Cahill. Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *Am J Speech Lang Pathol*, 12(2):198–208, May 2003.

Irene Jacobi, Lisette van der Molen, Hermelinde Huiskens, Maya A Van Rossum, and Frans JM Hilgers. Voice and speech outcomes of chemoradiation for advanced head and neck cancer: a systematic review. *European Archives of Oto-Rhino-Laryngology*, 267(10):1495–1505, 2010a.

Irene Jacobi, Lisette van der Molen, Maya van Rossum, and Frans Hilgers. Pre- and short-term posttreatment vocal functioning in patients with advanced head and neck cancer treated with concomitant chemoradiotherapy. ISCA, 2010b.

Irene Jacobi, Lisette van der Molen, Maya A van Rossum, and Frans J Hilgers. Pre- and short-term posttreatment vocal functioning in patients with advanced head and neck cancer treated with concomitant chemoradiotherapy. In *Proceedings of Interspeech*, pages 2582–2585. ISCA, 2010c.

Irene Jacobi, Maya A van Rossum, Lisette van der Molen, Frans JM Hilgers, and Michiel WM van den Brekel. Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy. *Annals of Otology, Rhinology & Laryngology*, 122(12): 754–762, 2013.

Irene Jacobi, Arash Navran, Lisette van der Molen, Wilma D Heemsbergen, Frans JM Hilgers, and Michiel WM van den Brekel. Radiation dose to the tongue and velopharynx predicts acoustic-articulatory changes after chemo-imrt treatment for advanced head and neck cancer. *European Archives of Oto-Rhino-Laryngology*, pages 1–8, 2015a.

Irene Jacobi, AJ Timmermans, Frans J M Hilgers, and Michiel W M van den Brekel. Voice quality and surgical detail in post-laryngectomy tracheoesophageal speakers. *European Archives of Oto-Rhino-Laryngology*, Online 22 September 2015, 2015b. doi: 10.1007/s00405-015-3777-4.

Petra Jongmans, Frans J M Hilgers, Louis C W Pols, and Corina J van As-Brooks. The intelligibility of tracheoesophageal speech, with an emphasis on the voiced-voiceless distinction. *Logoped Phoniatr Vocol*, 31(4):172–81, 2006.

Petra Jongmans, Ton G Wempe, Harm van Tinteren, Frans J M Hilgers, Louis C W Pols, and Corina J van As-Brooks. Acoustic analysis of the voiced-voiceless distinction in dutch tracheoesophageal speech. *J Speech Lang Hear Res*, 53(2):284–97, Apr 2010.

R. Kazi, J. Kanagalingam, R. Venkitaraman, V. Prasad, P. Clarke, C.M. Nutting, P. Rhys-Evans, and K.J. Harrington. Electroglottographic and perceptual evaluation of tracheoesophageal speech. *Journal of Voice*, 23(2): 247–254, 2009.

Rehan Kazi, Arvind Singh, Alya Al-Mutairy, Jose de Cordova, Lisa O'Leary, Chris Nutting, Peter Clarke, Peter Rhys Evans, and Kevin Harrington. Electroglottographic analysis of valved speech following total laryngectomy. *Logopedics Phoniatrics Vocology*, 33(1):12–21, 2008a.

Rehan Kazi, Ramachandran Venkitaraman, Catherine Johnson, Vyas Prasad, Peter Clarke, Peter Rhys-Evans, Christopher M Nutting, and Kevin J Harrington. Electroglottographic comparison of voice outcomes in patients with advanced laryngopharyngeal cancer treated by chemoradiotherapy or total laryngectomy. *Int J Radiat Oncol Biol Phys*, 70(2):344–52, Feb 2008b.

LE Kelly. *Head and neck cancer: treatment, rehabilitation, and outcomes*, chapter Radiation and chemotherapy, pages 57–86. Plural Publishing, 2007.

R D Kent, G Weismer, J F Kent, and J C Rosenbek. Toward phonetic intelligibility testing in dysarthria. *J Speech Hear Disord*, 54(4):482–99, Nov 1989.

AM Korpijaakko-Huuhka, AL Soderholm, and M Lehtihalmes. Long-lasting speech and oral-motor deficiencies following ral cancer surgery: a retrospective study. *Logoped Phoniatr Vocol*, 24:97–106, 1998.

S. A. C. Kraaijenga, I. M. Oskam, R. J. J. H. van Son, O. Hamming-Vrieze, F. J. M. Hilgers, M. W. M. van den Brekel, and L. van der Molen. Assessment of voice, speech, and related quality of life in advanced head and neck cancer patients 10-years+ after chemoradiotherapy. *Oral Oncology*, 55: 24–30, 2016/07/02 2016. doi: 10.1016/j.oraloncology.2016.02.001. URL http://dx.doi.org/10.1016/j.oraloncology.2016.02.001.

Sophie A. C. Kraaijenga, Lisette van der Molen, Irene Jacobi, Olga Hamming-Vrieze, Frans J. M. Hilgers, and Michiel W. M. van den Brekel. Prospective clinical study on long-term swallowing function and voice quality in advanced head and neck cancer patients treated with concurrent chemoradiotherapy and preventive swallowing exercises. *European Archives of Oto-Rhino-Laryngology*, 272(11):3521–3531, 2015. ISSN 1434-4726. doi: 10.1007/s00405-014-3379-6. URL http://dx.doi.org/10.1007/s00405-014-3379-6.

A Labaere and M Laeremans. *Spraakrevalidatie na een totale laryngectomie*. Acco, Leuven (Belgium), 2009.

Elisabet Lundström and Britta Hammarberg. Speech and voice after laryngectomy: Perceptual and acoustical analyses of tracheoesophageal speech related to voice handicap index. *Folia Phoniatr Logop*, 63:98–108, 2011. doi: 10.1159/000319740.

Elisabet Lundström, Britta Hammarberg, Eva Munck-Wikland, and Nick Edsborg. The pharyngoesophageal segment in laryngectomees–videoradiographic, acoustic, and voice quality perceptual data. *Logoped Phoniatr Vocol*, 33(3):115–25, 2008.

K Mády, R Sader, P H Hoole, A Zimmermann, and H H Horch. Speech evaluation and swallowing ability after intra-oral cancer. *Clin Linguist Phon*, 17(4-5):411–20, 2003.

Andreas Maier, Tino Haderlein, Maria Schuster, Emeka Nkenke, and Elmar Nöth. Intelligibility is more than a single word: quantification of speech intelligibilty by ASR and prosody. *Lecture Notes in Computer Science*, 4629: 278–285, 2007.

Andreas Maier, Tino Haderlein, Ulrich Eysholdt, Frank Rosanowski, Anton Batliner, Maria Schuster, and Elmar Nöth. Peaks–a system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5): 425–437, 2009.

Andreas Maier, Tino Haderlein, Florian Stelzle, Elmar Nöth, Emeka Nkenke, Frank Rosanowski, Anne Schützenberger, and Maria Schuster. 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*, 2010:1–7, 2010.

Tanya K Meyer, Joan C Kuhn, Bruce H Campbell, Anne M Marbella, Katherine B Myers, and Peter M Layde. Speech intelligibility and quality of life in head and neck cancer survivors. *Laryngoscope*, 114(11):1977–81, Nov 2004.

C Middag. Dia-tool. Powerpoint presentation, Feb 2011.

Catherine Middag, Gwen Van Nuffelen, Jean-Pierre Martens, and Marc De Bodt. Objective intelligibility assessment of pathological speakers. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech'08)*, pages 1745–1748, 2008.

Catherine Middag, Jean-Pierre Martens, Gwen Van Nuffelen, and Marc De Bodt. Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing, Article ID 629030*, 2009a.

Catherine Middag, Jean-Pierre Martens, Gwen Van Nuffelen, and Marc De Bodt. DIA: a tool for objective intelligibility assessment of pathological speech. In *Proceedings of MAVEBA2009*, 2009b.

Catherine Middag, Yvan Saeys, and Jean-Pierre Martens. Towards an ASR-free objective analysis of pathological speech. In *11th Annual conference of the International Speech Communication Association (Interspeech 2010)*, volume 2010, pages 294–297. International Speech Communication Association (ISCA), 2010.

Catherine Middag, Renee Clapham, Rob van Son, and Jean-Pierre Martens. Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. *Computer Speech and Language*, 28(2): 467 – 482, 2014. ISSN 0885-2308. doi: http://dx.doi.org/10.1016/j.csl.2012.10.007. URL http://www.sciencedirect.com/science/article/pii/S0885230812000903.

M. Moerman, Jean Pierre Martens, M. Van der Borgt, M. Peleman, M. Gillis, and P. Dejonckere. Perceptual evaluation of substitution voices: development and evaluation of the (i)infvo rating scale. *European Archives of Oto-Rhino-Laryngology*, 263(2):183–187, 2 2006.

M Moerman, Jean-Pierre Martens, and Philippe Dejonckere. Multidimensional assessment of strongly irregular voices such as in substitution voicing and spasmodic dysphonia: A compilation of own research. *Logopedics Phoniatrics Vocology*, Early online, 2014.

Mieke Moerman, Glenn Pieters, Jean Pierre Martens, Marie-Jeanne Van der Borgt, and Philippe Dejonckere. Objective evaluation of the quality of substitution voices. *European Archives of Oto-Rhino-Laryngology*, 261(10):541–547, 11 2004.

Sarah E Mowry, Allen Ho, Maria M. LoTempio, Ahmad Sadeghi, Keith E. Blackwell, and Marilene B Wang. Quality of life in advanced oropharyngeal carninoma after chemoradiation versus surgery and radiation. *The Laryngoscope*, 116(9):1589–1593, 2006.

A Neri, C Cucchiarini, and H Strik. Asr-based corrective feedback on pronunciation: does it really work? In *Proceedings of Interspeech*, pages 1982–1985, 2006.

Giovanni Nicoletti, David S Soutar, Mary S Jackson, Alan A Wrench, Gerry Robertson, and Chris Robertson. Objective assessment of speech after surgical treatment for oral cancer: experience from 196 selected cases. *Plast Reconstr Surg*, 113(1):114–25, Jan 2004.

N B Oozeer, S Owen, B Z Perez, G Jones, A R Welch, and V Paleri. Functional status after total laryngectomy: cross-sectional survey of 79 laryngectomees using the performance status scale for head and neck cancer. *J Laryngol Otol*, 124(4):412–6, Apr 2010.

Bas Op de Coul, F van den Hoogen, Corina J van As-Brooks, H Marres, F Joosten, J Manni, and Frans J Hilgers. Evaluation of the effects of primary myotomy in total laryngectomy on the neoglottis with the use of quantitative videofluroscopy. *Arch Otolaryngol Head Neck Surg*, 129:1000–1005, 2003.

Aaron W. Pederson, Daniel J. Haraf, Mary-Ellyn Witt, Kerstin M. Stenson, Everett E. Vokes, Elizabeth A Blair, and Joseph K. Salama. Chemoradiotherapy for locoregionally advanced squamous cell carcinoma of the base of tongue. *Head & Neck*, 32(11):1519–1527, 2010. doi: 10.1002/hed.21360.

Korbinian Riedhammer, Georg Stemmer, Tino Haderlein, Maria Schuster, Frank Rosanowski, Elmar Nöth, and A Maier. Towards robust automatic evaluation of pathologic telephone speech. In *Proceedings of the IEEE automatic Speech Recognition and Understandibility Workshop (ASRU '07)*, pages 717–722. IEEE, 2007.

Nicolás Sáenz-Lechón, Juan I Godino-Llorente, Víctor Osma-Ruiz, Manuel Blanco-Velasco, and Fernando Cruz-Roldán. Automatic assessment of voice quality according to the grbas scale. *Conf Proc IEEE Eng Med Biol Soc*, 1: 2478–81, 2006.

Maria Schuster, Frank Rosanowski, Raphael Schwarz, Ulrich Eysholdt, and Jorg Lohscheller. Quantitative detection of substitute voice generator during phonation in patients undergoing laryngectomy. *Arch Otolaryngol Head Neck Surg*, 131(11):945–952, 2005.

Maria Schuster, Tino Haderlein, Elmar Nöth, Jörg Lohscheller, Ulrich Eysholdt, and Frank Rosanowski. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, 263(2):188–193, 2006a.

Maria Schuster, Andreas Maier, Tino Haderlein, Emeka Nkenke, Ulrike Wohlleben, Frank Rosanowski, Ulrich Eysholdt, and Elmar Nöth. Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition. *International journal of pediatric otorhinolaryngology*, 70(10):1741–1747, 2006b.

J P Searl, M A Carpenter, and C L Banta. Intelligibility of stops and fricatives in tracheoesophageal speech. *J Commun Disord*, 34(4):305–21, 2001.

F Stelzle, B Ugrinovic, C Knipfer, T Bocklet, E Nöth, M Schuster, S Eitner, M Seiss, and E Nkenke. Automatic, computer-based speech assessment on edentulous patients with and without complete dentures - preliminary results. *J Oral Maxillofac Surg*, 37(3):209–216, 2010.

AJ Timmermans, BAC van Dijk, LIH Overbeek, MLF van Velthuysen, Harm van Tinteren, F J M Hilgers, and Michiel WM van den Brekel. Trends in treatment and survival of advanced larynx cancer: a 20-year population-based study in the netherlands. *Head & Neck*, 2015. doi: 10.1002/hed.24200.

Corina J van As, M Tigges, T Wittenberg, Bas Op de Coul, U Eysholdt, and Frans J Hilgers. High-speed digital imaging of neoglottic vibration after total laryngectomy. *Arch Otolaryngol Head Neck Surg*, 125:891–897, 1999.

Corina J Van As, Bas M R Op De Coul, Ulrich Eysholdt, and Frans J M Hilgers. Value of digital high-speed endoscopy in addition to videofluoroscopic imaging of the neoglottis in tracheoesophageal speech. *Acta Oto-laryngologica*, 124 (1):82–89, 2004. ISSN 0001-6489 (Print).

C. J. van As-Brooks. Acoustic analyses of postlaryngectomy voice and their perceptual relevance. *Invitational Round Table "Evidence-based Voice and Speech Rehabilitation in Head and Neck Cancer"*, page 8, 2008.

Corina J van As-Brooks, Frans J Hilgers, Koopmans van Beinum, and L C W Pols. Anatomical and functional correlates of voice quality in tracheoesophageal speech. *Journal of Voice*, 19(3):360–372, 2005.

Corina J van As-Brooks, Florien J Koopmans-van Beinum, Louis C W Pols, and Frans J M Hilgers. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *J Voice*, 20(3):355–68, Sep 2006.

Lisette van der Molen, Maya A van Rossum, Irene Jacobi, Rob JJH van Son, Ludi E Smeele, Coen RN Rasch, and Frans JM Hilgers. Pre-and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: expert listeners' and patient's perception. *Journal of Voice*, 26(5):664–e25, 2012.

Gwen Van Nuffelen, Catherine Middag, Jean-Pierre Martens, and Marc De Bodt. Speech technology based assessment of dysarthric speech: preliminary results. In *Proceedings of 27th World Congress of the International Association of Logopedics and Phoniatrics (IALP)*, 2007.

Gwen Van Nuffelen, Catherine Middag, Marc De Bodt, and Jean-Pierre Martens. Speech technology-based assessment of phoneme intelligibility in dysarthria. *International journal of language & communication disorders*, 44 (5):716–730, 2009.

M A van Rossum, Corina J van As-Brooks, Frans J M Hilgers, and M Roozen. Quality of 'glottal' stops in tracheoesophageal speakers. *Clin Linguist Phon*, 23(1):1–14, Jan 2009.

Irma Verdonck-de Leeuw, Louis ten Bosch, Li Ying Chao, Rico NPM Rinkel, Pepijn A Borggreven, Lou Boves, and C René Leemans. Speech quality after major surgery of the oral cavity and oropharynx with microvascular soft tissue reconstruction. In *INTERSPEECH*, pages 1186–1189, 2007a.

Irma M Verdonck-de Leeuw, Rico N Rinkel, and C René Leemans. *Head and neck cancer: treatment, rehabilitation, and outcomes*, chapter Evaluating the impact of cancer of the head and neck, pages 27–56. Plural Publishing, 2007b.

Clemens Weber, Steffen Dommerich, Hans Wilhelm Pau, and Burkhard Kramp. Limited mouth opening after primary therapy of head and neck cancer. *Oral Maxillofac Surg*, 14(3):169–73, Sep 2010.

Tara L Whitehill, Valter Ciocca, Judy C-T Chan, and Nabil Samman. Acoustic analysis of vowels following glossectomy. *Clin Linguist Phon*, 20(2-3):135–40, 2006.

Martin Windrich, Andreas Maier, Regina Kohler, Elmar Nöth, Emeka Nkenke, Ulrich Eysholdt, and Maria Schuster. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. *Folia Phoniatrica et Logopaedica*, 60(3):151–156, 2008.

K Yorkston, E Strand, and Kennedy M. Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *Am J Speech Lang Pathol*, 5(1):55–66, 1996.

Kathryn M. Yorkston and David R. Beukelman. A clinician-judged technique for quantifying dysarthric speech based on single-word intelligibility. *Journal of Communication Disorders*, 13(1):15 – 31, 1980.

A C Zuydam, D Lowe, J S Brown, E D Vaughan, and S N Rogers. Predictors of speech and swallowing function following primary surgery for oral and oropharyngeal cancer. *Clin Otolaryngol*, 30(5):428–37, Oct 2005.

Part I

Aspects of evaluating voice and speech after chemoradiotherapy

2

NKI-CCRT corpus - speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy[0]

Abstract

Evaluations of speech intelligibility based on a read passage are often used in the clinical situation to assess the impact of the disease and/or treatment on spoken communication. Although scale-based measures are often used in the clinical setting, these measures are susceptible to listener response bias. Automatic evaluation tools are being developed in response to some of the drawbacks of perceptual evaluation, however, large corpora judged by listeners are needed to improve and test these tools. To this end, the NKI-CCRT corpus with individual listener judgements on the intelligibility of recordings of 55 speakers treated for cancer of the head and neck will be made available for restricted scientific use. The corpus contains recordings and perceptual evaluations of speech intelligibility over three evaluation moments: before treatment and after treatment (10-weeks and 12-months). Treatment was by means of chemoradiotherapy (CCRT). Thirteen recently graduated speech pathologists rated the speech in-

[0] R.P. Clapham, L. van der Molen, R.J.J.H. van Son, M.W.M. van den Brekel, F.J.M. Hilgers. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12): 23-25 May, 2012, Istanbul, Turkey (pp. 3350-3355).*

telligibility of the recordings on a 7-point scale. Information on recording and perceptual evaluation procedures is presented in addition to preliminary rater reliability and agreement information. Preliminary results show that for many speakers speech intelligibility is rated low before cancer treatment.

2.1 Introduction

A recent randomized controlled clinical trial by van der Molen and colleagues (van der Molen et al., 2012) followed a group of patients prior to and after concomitant chemoradiotherapy (CCRT) for advanced cancer of the head and neck. The Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital (NKI-AVL) has made part of these recordings with speech intelligibility ratings available to researchers to aid the development of automatic methods of evaluating speech intelligibility. This corpus is termed the NKI-CCRT corpus. This paper describes the speech corpus and presents some preliminary results regarding the perceptual evaluation of speech intelligibility.

Developing automatic methods to evaluate speech intelligibility has become a recent research interest and studies have focused on completely automatic assessments (e.g., Maier et al., 2009; Middag et al., 2009; Pitaksirianant et al., 2011; Windrich et al., 2008) or computer-supported evaluation procedures (e.g. Sentence Intelligibility Test, Yorkston et al. (2007); MVP-online, Ziegler and Zierdt (2008)). The move towards complete automatic evaluation is in response to some of the drawbacks of perceptual evaluations of speech intelligibility, such as a listener's familiarity with a speaker or knowledge of test stimuli. Although evaluation of paragraph stimuli provides a more realistic indicator of a speaker's level of speech intelligibility outside the clinical situation, evaluations based on paragraph level stimuli can only be evaluated by means of a scale. Scale-based evaluations, however, are susceptible to listener response bias (e.g. variation in internal anchors). Mean scores are often used to remove some of this 'error'.

In van der Molen et al. (2012) the authors reported a general decrease-increase trend regarding changes in speech and voice quality, however, changes in speech intelligibility for the speaker group between evaluation moments did not reach statistical significance. As van der Molen used a within-speaker paired-comparison evaluation paradigm, these evaluations are not easily transferred for training automatic prediction models. For the corpus to be useful in developing speech intelligibility predication models, we have gathered perceptual speech intelligibility ratings for the recordings presented in and collected by van der Molen et al. (2012).

In addition to presenting the corpus and preliminary information on rater agreement and rater reliability, we investigate whether (a) the decrease-increase trend reported in van der Molen et al. is present for scale measurements of speech intelligibility and (b) speech intelligibility ratings vary according to which fragment of a text the listener rated. This last question has implications for speech technology researchers as it allows researchers to investigate how text dependent a prediction model may be.

Although we use data based on mean scores in this paper to describe the speech intelligibility ratings, the corpus is not limited to mean scores. By making this corpus with listener judgements on speech intelligibility available for restricted scientific use, we hope to progress the work into automatic evaluation of speech intelligibility.

2.2 Method

2.2.1 Speakers

The corpus contains recordings of 55 speakers recorded at three evaluation moments: before CCRT (N = 54[1]), 10-weeks after CCRT (N = 48) and 12-months after CCRT (N = 39). Average speaker age before CCRT was 57 years. Based on perceptual evaluation by a Dutch phonetician (RvS), speakers were categorized as either speakers of Dutch as a first language or Dutch as a second language. This was necessary as language background was not a patient characteristic collected at the time of recordings. Table 2.1 presents speaker characteristics.

	Dutch 1st Language	Dutch 2nd Language	Total (%)
Male	39	6	45 (82)
Female	8	2	10 (18)
Total (%)	47 (85)	8 (15)	

Table 2.1: Language background of speakers based on perceptual evaluation of speech recordings

Speech materials and recordings

Recordings were made in a sound-treated room with a Sennheiser MD421 Dynamic Microphone and portable 24-bit digital wave recorder (Edirol Roland R-1). Sampling frequency was 44.1 kHz and mouth to microphone distance was 30 cm.

[1] Due to an oversight, one speaker's before treatment recording was not included in the perceptual experiment.

All speakers read a 189-word passage from a Dutch fairy tale. We divided the recorded text into three fragments based on natural breaks in the text (fragment A = 70 words, fragment B = 68 words, fragment C = 51 words). Only fragments A and B were used in the perceptual experiment and are included in the corpus. The two fragments are similar regarding number of unique words (A = 49, B = 50), average syllable length (A = 1.3, B = 1.5) and phoneme frequencies (A = 237, B = 247). The phoneme /f/ only appears in fragment A (see appendix for phoneme overview). The text was not balanced for phoneme frequency and the two fragments do not contain all Dutch phonemes.

2.2.2 Annotations and tags

All recordings were annotated with Praat (Boersma and Weenink, 2011). Annotations are stored in Praat TextGrid files. Each annotation contains four tiers:

1. Transliteration: Sentence-aligned transliteration of the spoken utterances using the conventions of the Spoken Dutch Corpus (Oostdijk et al., 2002);
2. Sentences: The original text aligned per sentence (aligned on the previous tier);
3. Text: The complete original text;
4. Interferences: Noise markers.

The corpus contains automatically-generated word alignment and phoneme alignment annotations. Overlapping speech of the clinician has not been transcribed and is marked in the Interferences tier and as silence in the Transliteration tier. Tags used in the Interference tier include Recording Level, Microphone Failure, Other Speaker, and Noise, indicating, respectively, noticeable changes in the recording level, manipulations of the microphone that mask all sound, any speech from other speakers than the patient, and general noise (e.g., phone ringing). All recordings have been evaluated on the presence of noise and extraneous sounds by one of the authors (RvS) using a 3-point scale.

2.2.3 Perceptual evaluation

A group of recently graduated and about to graduate[2] speech pathologists evaluated the speech recordings by means of a 7-point scale. All listeners reported no hearing problems and were Dutch native speakers. Speech intelligibility was

[2] All students were either in their final weeks of the speech pathology course or had graduated several weeks before the perceptual evaluation.

defined as the difficulty/ease with which the listener decodes the speech signal. Listeners were instructed to try to ignore aspects of voice acceptability, reading fluency and any interrupting noises in the files. In addition to speech intelligibility, listeners also rated other aspects of speech production (e.g. articulation and voice quality). This information is not discussed in this paper and is not included in the corpus.

Although 14 listeners took part in this study, one listener's results were removed from analysis as this listener became unwell during the period of completing the evaluations. Average age of the 13 female volunteers was 23.7 (range 21.9-27.6). Listeners received a small financial reward for their participation.

Task familarization

All participants completed an online familiarization module. The module contained examples of good, reasonable and poor speech intelligibility as evaluated by one of the authors (RPC). Audio-stimuli were not restricted to speakers with cancer of the head and neck. Participants used their own anchors and received no feedback on performance.

Experimental design

All stimuli were presented via an online experiment. Audio file intensity was averaged to 70 dB. Participants were requested to complete all evaluations within five days, complete listening sessions at roughly the same time of day and complete evaluations in a quiet environment using the headset provided (Sennheiser HD418). Participants had access to the narrative text and were able to replay a stimulus. Participants were unable to change submitted responses.

Listeners evaluated 4 practice stimuli (fragment C to avoid a learning effect), just under 300 experimental stimuli (fragments A and B), and a repetition of the first 10 experimental stimuli (retest items). Stimuli were presented in a randomized order for each listener. Listeners completed the evaluations over three sessions. Average time to complete a listening session was 70 minutes.

2.2.4 Corpus meta-data

Age before CCRT and gender is available for each speaker ID[3]. For each audio stimulus the meta-data includes speaker ID, recording moment (pre-treatment [T0], 10-weeks post-treatment [T1], 12-months post-treatment [T3]) and intelligibility ratings.

[3] Speaker IDs are not related to patient identification numbers.

Rater	N	Within-rater Reliability PCC (CI)	% Agree. exact (+/-1)	N	Between-rater Reliability PCC (CI)
1	5	0.70 (-0.48-0.98)	20 (80)	39	0.58 (0.32-0.76)
2	9	0.61 (-0.09-0.91)	44 (89)	40	0.68 (0.47-0.82)
3	10	0 .90 (0.63-0.98)	20 (80)	40	0.75 (0.57-0.86)
4	10	0.69 (0.11-0.92)	50 (90)	40	0.76 (0.59-0.87)
5	10	0.73 (0.18-0.93)	40 (80)	40	0.80 (0.66-0.89)
6	10	0.92 (0.68-0.98)	40 (70)	40	0.88 (0.78-0.93)
7	10	0.87 (0.54-0.97)	80 (100)	40	0.71 (0.52-0.84)
8	10	0.92 (0.68-0.98)	50 (100)	40	0.88 (0.78-0.93)
9	10	0.90 (0.62-0.98)	50 (90)	40	0.85 (0.73-0.92)
10	10	0.83 (0.42-0.96)	20 (60)	40	0.80 (0.65-0.89)
11	10	0.79 (0.33-0.95)	60 (100)	39	0.85 (0.73-0.92)
12	10	-	80 (100)	39	0.72 (0.52-0.84)
13	8	0.80 (0.23-0.96)	75 (88)	40	0.85 (0.73-0.92)

Table 2.2: Within-rater reliability and agreement and between-rater reliability. N = number of paired stimuli, CI = 95% confidence interval. Correlations rounded to two decimal places. Percentages rounded to whole numbers.

2.2.5 Data analysis

For all analyses the alpha level was .05. Where multiple comparisons were made, the alpha level was adjusted (see paragraphs below). All statistics were completed with statistics program R (Team, 2012).

Reliability

Reliability was calculated using Pearson's correlation coefficient (PCC). We use this coefficient rather than the non-parametric Kendall's Tau for two reasons: to allow comparison with other studies and to report the strength of the association between the two variables. Reliability of speaker scores averaged over all listeners was calculated with the Interclass Correlation Coefficient (ICC) (two-way random effects model, average consistency).

Within-rater reliability was estimated by comparing each listener's 10 test-retest evaluations. For the between-rater reliability 40 stimuli that were not test-retest items for any listeners were randomly selected. We then compared each listener's evaluations against the average of all other raters.

Agreement

We report the percent exact agreement and the percent close agreement (+/-1 scale score) of each listener's 10 test-retest evaluations.

Independence of text fragment

We investigated if there were differences in speech intelligibility scores (averaged across listeners) for the two text fragments by means of Wilcoxon-Signed Ranks.

Changes in speech intelligibility

Change in speech intelligibility over time was investigated for speakers with three evaluation points by means of Friedman's test with Wilcoxon test for dependent samples as post hoc test.

2.3 Results

2.3.1 Reliability and agreement

Table 2.2 displays all listener reliability and agreement information. Although the correlation coefficient was below 0.7 for two listeners and the lower-bound CI was below 0, we did not remove these listeners given the small number of test-retest cases. For one listener no correlation could be calculated because this listener had no variation in retest scores. Exact agreement ranged from 20 to 80 percent, and percent close agreement ranged from 60 to 100 percent.

Between-rater reliability for the 40 randomly selected audio files ranged from a PCC of 0.58 to 0.88. An ICC of 0.95 (95% CI: 0.92-0.97) for the 13 participants based on ratings of 37 items[4] suggests that the mean score (averaged over all listeners) is reliable.

Although not all subjects completed all the evaluations per protocol (i.e. an entire session in one sitting), these subjects were not excluded from the study as their reliability results indicated that these listeners were no less reliable than those who completed the evaluations following protocol.

2.3.2 Text fragment analysis

To assess if intelligibility scores varied according to fragment, we compared all fragment pairs. As there was no significant statistical difference between ratings for the two fragments (p = 0.18), we report speaker mean scores pooled over fragments.

[4] 3 items removed due to missing values

Evaluation moment	Mean (SD)	Range
Pre-CCRT	5.61 (0.97)	3.03-6.65
10-weeks after CCRT	5.59 (0.95)	2.32-6.73
12-months after CCRT	5.62 (0.92)	2.88-6.69

Table 2.3: Overview of group speech intelligibility evaluations for the 37 speakers with three evaluation moments.

2.3.3 Changes in speech intelligibility ratings

Group Level

Based on the mean scores (averaged over all listeners), mean speech intelligibility is lowest before CCRT (mean 5.41, SD 1.08, N = 54) and highest 12-months after CCRT (5.85, SD 0.91, N = 39).

As displayed in Figure 2.1, listeners rate many speaker's speech intelligibility as low before CCRT. Visual inspection of the figure indicates that for approximately half of the speakers, speech intelligibility ratings peak before CCRT whereas for the other half of the speakers, change in speech intelligibility ratings appears more variable.

Of the 27 speakers with intelligibility scores under the median before CCRT, 59 percent contribute recordings at all evaluation moments; for speakers with scores above the pre-treatment median, this is 78 percent. Analysis by means of Fisher's exact test revealed that the number of complete evaluation moments does not significantly differ for speakers who are above or below the median pre-treatment score (p = .24, CI = 0.10-1.60). We therefore continue our analysis with the 37 speakers with speech intelligibility scores for all evaluation moments.

Based on the group average scores of the 37 speakers with recordings for all evaluation moments, speech intelligibility ratings decreased after treatment but returned to pre-treatment levels 12-months after treatment (see Table 2.3). Friedman's test indicated that there was no significant difference between the three evaluation moments for the group.

Speaker level

Given the variation in score patterns between the listeners, we investigated changes at the level of the speaker. Compared to pre-treatment, the majority of speakers had lower scores at both follow-up moments whereas the pattern between 10-weeks and 12-months was variable (see Figure 2.1). To investigate within-speaker changes in speech intelligibility ratings over time, we compared the scores for each evaluation moment (averaged over the two fragments; 13

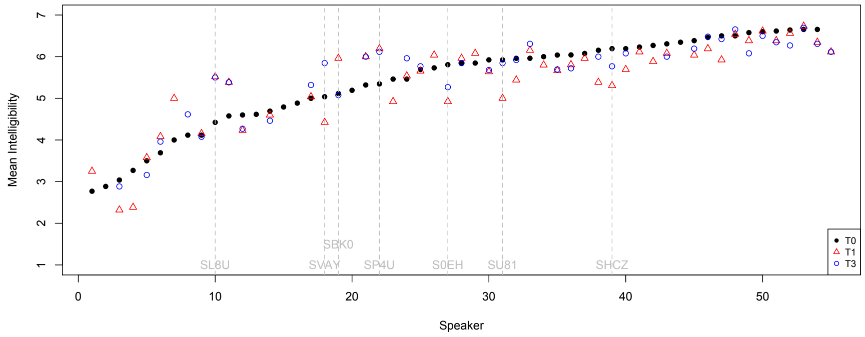


Figure 2.1: Intelligibility scores for individual speakers at each measurement moment. Data is ordered according to pre-treatment intelligibility score. Dashed lines show the speakers with a significant difference between two or more measurement moments ($p < 0.0013$). T0 = pre-treatment, T1 = 10-weeks post treatment, T3 = 12-months post treatment.

observations per evaluation moment).

Table 2.4 displays the mean difference in speech intelligibility rating for the group between all evaluation moments plus the frequency of the direction of change. For seven speakers (see the vertical lines in Figure 2.1) there was a significant difference in scores over time based on Friedman's test (alpha adjusted for multiple comparisons, $p < 0.0013$). There was a significant difference between the pre-treatment and 10-weeks post treatment rank order comparisons for six speakers (3 increase), the pre-treatment and 12-months post treatment rank order comparisons for 2 speakers (2 increase) and 10-weeks and 12-months post-treatment comparisons for 3 speakers (2 increase).

	Mean difference	+ (%)	- (%)
T1-T0	-0.11	15 (41)	22 (59)
T3-T0	.00	12 (32)	25 (68)
T3-T1	0.12	18 (49)	19 (51)

Table 2.4: Mean difference in score between each evaluation moment. For each evaluation pair, number of speakers with positive (+) and negative (-) differences are given. Percentages are presented as whole numbers. T0 = pre-treatment, T1 = 10-weeks post treatment, T3 = 12-months post treatment.

2.4 Discussion

In this paper we have described the recordings and perceptual evaluations of the NKI-CCRT corpus. For full details on the speakers and treatment we refer the reader to van der Molen et al. (2012). Unlike the evaluations in van der Molen et al. who used a paired-comparison paradigm to investigate changes in speech intelligibility, the results in this paper are based on evaluations made by 13 (recently) graduated speech pathologists on a 7-point rating scale. This was necessary as paired-comparison scores allow neither comparison between speakers nor provide an indication of speech intelligibility: for the data to be used as training material for automatic evaluation, this information is desirable.

Comparing results between the evaluations reported in van der Molen et al. and the ratings collected for this corpus is difficult due to the differences in scoring paradigms and analysis. At a group level, both studies agree that there is no significant change in speech intelligibility scores over the evaluation moments. The mean ratings for the 37 speakers who contributed recordings at all evaluation moments, however, support the decrease-increase trend found in van der Molen for speech and voice quality.

The lack of significant results when the speakers are taken as a whole is not surprising given the variability in ratings between the speakers: 59 percent of speakers' speech intelligibility ratings decreased after CCRT, and 49 percent of the speaker's speech intelligibility ratings increased between short-term and long-term evaluation moments. Although for six of the speakers there was a significant effect of time on speech intelligibility ratings, no pattern is apparent regarding the change of direction (i.e., increase or decrease in speech intelligibility rating). This suggests that variety within the group of speakers may mask individual speaker changes.

Although the results indicate that the listeners are, as a whole reliable, the confidence intervals for some listeners' within-rater reliability are low. This raises the question whether speaker scores should be averaged over listeners and, if so, which listeners. We anticipate that future work will investigate the role of the listener in speech intelligibility judgments: a better understanding of this relationship may aid automatic evaluation tools.

2.5 Conclusion

The primary aim of this study was to introduce the NKI-CCRT corpus and present preliminary data on speech intelligibility ratings for the recordings. The findings that perceptual speech intelligibility scores do not differ depending on text fragment and that speech intelligibility scores significantly vary for some speakers over time make this corpus attractive for use in developing speech intelligibility prediction models for Dutch speakers treated for cancer of the head and neck.

Declaration of Interest

Part of this research was funded by an unrestricted research grant from Atos Medical, Horby, Sweden, and the Verwelius Foundation, Naarden.

Acknowledgements

Many thanks to Irene Jacobi for her comments and suggestions. Friedman's test with post-hoc function and graphics was run using R code published by Tal Galili (www.r-statistics.com/2010/02/post-hoc-analysis-for-friedmans-test-r-code).

Availability

Corpus will be available in the latter half of 2012 for restricted scientific use. Parties interested in obtaining a copy of the corpus can contact Michiel van den Brekel (Head & Neck Oncology, The Netherlands Cancer Institute).

Appendix

Consonant	A	B	Consonant	A	B
p	3	3	m	5	9
b	2	4	n	31	28
t	18	22	N	3	1
d	12	9	l	8	6
k	7	4	r	14	17
f	1	0	j	3	3
v	4	4	w	7	9
s	10	10	i	4	3
z	1	5	I	7	5
o	2	2			
x	11	9			
h	9	12			

Table 2.5: Consonant frequency for the two text fragments (A and B) based on on automatic broad transcription of canonical pronunciation.

Vowel	A	B	Vowel	A	B
e	12	3	O	4	4
E	11	14	a	2	8
A	15	17	Eˆ	6	7
@	17	24	Oˆ	4	1
u	3	3	@ˆ	1	1

Table 2.6: Vowel frequency for the two text fragments (A and B) based on on automatic broad transcription of canonical pronunciation.

References

P. Boersma and D Weenink. Praat: doing phonetics by computer [computer program]. http://www.praat.org/, 2011.

A Maier, T Haderlein, U Eysholdt, F Rosanowski, Anton Batliner, M Schuster, and E Noth. Peaks - a system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425–437, May 2009.

C Middag, J-P Martens, G van Nuffelen, and Marc S De Bodt. Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal of Advances in Signal Processing*, 2009.

N. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H Baayen. Experiences from the spoken dutch corpus project. In Araujo, editor, *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 340–347, 2002.

N. Pitaksirianant, K. Saykhum, C. Wutiwiwatchai, A. Chotoimongkol, and A. Pimkhaokham. A study of automatic speech intelligibility testing for Thai oral surgical patients. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2011 8th International Conference on*, volume Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pages 938–941, 2011.

R Development Core Team. R: A language and environment for statistical computing (version 2.12.0) [computer program]. http://www.R-project.org, 2012.

L. van der Molen, M A. van Rossum, I. Jacobi, R. van Son, Ludi E Smeele, Coen R N Rasch, and FJM. Hilgers. Pre- and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: expert listeners' and patient's perception. *Journal of Voice*, online, January 3 2012.

M Windrich, A Maier, R Kohler, E Noth, E Nkenke, U Eysholdt, and M Schuster. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. *Folia Phoniatr Logop*, 60(3):151–6, 2008.

K. M. Yorkston, D. R. Beukelman, M. Hakel, and M. Dorsey. *Speech intelligibility test for windows [computer program]*. Institute for Rehabilitation Science and Engineering at Madonna Rehabilitation Hospital, 2007.

W Ziegler and A Zierdt. Teldiagnostic assessment of intelligibility in dysarthria: A pilot investigation of mvp-online. *J Commun Disord*, 41:553–577, 2008.

3

Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer[0]

Abstract

It is generally acknowledged that an unbiased and objective assessment of the communication deficiency caused by a speech disorder calls for automatic speech processing tools. In this paper, a new automatic intelligibility assessment method is presented. The method can predict running speech intelligibility in a way that is robust against changes in the text and against differences in the accent of the speaker. It is evaluated on a Dutch corpus comprising longitudinal data of several speakers who have been treated for cancer of the head and the neck. The results show that the method is as accurate as a human listener in detecting trends in the intelligibility over time. By evaluating the intelligibility predictions made with different models trained on distinct texts and accented speech data, evidence for the robustness of the method against text and accent factors is offered.

[0] C. Middag, R.P. Clapham, R.J.J.H. van Son, JP. Martens. *Computer Speech and Language 28 (2), March 2014, 467-482.*

3.1 Introduction

Effective verbal communication is an essential aspect of daily life and is often taken for granted. It presents a major bottleneck though for people experiencing speech disorders. Disordered (or pathological) speech can be the consequence of a plurality of causes, but the assessment, treatment and monitoring of pathological speech have been receiving growing attention in the biomedical field.

A widely used measure of the severity of a speech disorder is speech intelligibility, loosely defined as the ease with which a listener is able to lexically decode the utterances of a speaker (Yorkston. et al., 1996). In the clinical setting, measures of speech intelligibility for text level stimuli are often acquired by means of a perceptual test, but the results of such a test are acknowledged to be subjective and influenced by the listener's familiarity with both the patient's voice and the read text.

Previous research indicated that automatic speech recognition (ASR) can be used for intelligibility measurement. Ferrier et al. (1995) experimented with repeated readings of the same passage to a dictation system (Dragon Dictate). In a test on ten dysarthric speakers, they obtained high correlations between mean recognition rate over eight readings and the perceptually measured intelligibility scores. More recently, Vijayalakshmi et al. (2006, 2009) proved that the phone recognition rate is a valuable measure of intelligibility. But again, only nine dysarthric speakers were tested. A handful of other objective intelligibility assessment methods have been reported (Falk et al., 2011; Gu et al., 2005; Hosom et al., 2004), but a major limitation of using ASR to develop a robust intelligibility assessment tool is that it needs many recordings of pathological speech for model training before it reaches a reliable outcome. Unfortunately, large pathological speech corpora are scarce. This explains why it was only recently that effective tools for objective intelligibility assessment could be developed. The tools proposed in Maier et al. (2009) and Middag et al. (2009) for instance were shown to compete well with traditional perceptual evaluations.

In previous work (Middag et al., 2008; Middag et al., 2009), we were successful in automating the Dutch Intelligibility Assessment (DIA) (De Bodt et al., 2006; Van Nuffelen et al., 2008). The DIA requires the patient to utter 50 monosyllabic (partly nonsense) words and, per utterance, a human listener has to identify the tested phoneme. The number of correctly identified phonemes then determines the phoneme intelligibility (PI). With an Interclass Correlation Coefficient of 0.91, the inter-rater reliability for scoring PI is strong (De Bodt et al., 2006). The automated DIA-tool[1] works with the same speech items but uses an automatic system to analyze the utterances and to produce an objective

[1] The automated DIA is currently available (for Flemish only) online at http://diaweb.elis.ugent.be/.

intelligibility score. Experiments have shown (Middag et al., 2008) that these objective scores correlate well with the human scores. Nevertheless, the present DIA suffers from a number of limitations, and in this paper, we address these limitations and we propose new solutions to overcome them.

3.1.1 Isolated word versus running speech analysis

In the current DIA method each speaker reads 50 consonant-vowel-consonant (CVC) words, mostly nonsense words. A fundamental problem with this set-up is that the phoneme intelligibility derived from listening to these isolated utterances is bound to correlate only moderately with the ability of the patient to communicate in a more realistic situation where running speech is the most important speech mode (Kent et al., 1989; Van Nuffelen, 2009). It would therefore be more interesting to extend our current automatic methods of evaluation to the prediction of running speech intelligibility (RSI). This may not be that unattainable given that the acoustic models embedded in the DIA tool are already trained on running speech (normal speakers reading full sentences and text paragraphs). These models are bound to be better suited for the assessment of running speech than they are for the assessment of isolated word utterances.

3.1.2 Text-dependent versus text-independent methods

The present automated tool performs a time-alignment of a spoken utterance with the canonical phonetic transcription of the prompted text (= speech-to-text alignment). By analyzing the alignments for all utterances of a speaker, a set of so-called speaker features is extracted and from this set the objective PI is estimated (Middag et al., 2009). However, since some of the uttered words are nonsense words, the speech material can contain hesitations, reading errors and pronunciation variations (there may be different acceptable pronunciations of the same nonsense word and the speaker may not necessarily use the one described by the canonical transcription of the word). Consequently, a methodology based on speech-to-text alignment is bound to be sensitive to these sources of variation.

The envisioned automatic intelligibility assessment should be able to produce a reliable score, even in the presence of discrepancies between what was spoken and what is encoded in the canonical transcription of the prompted text. In order to achieve that kind of robustness, the method should not rely too much on a speech-to-text alignment but rather work with statistical measures that are presumed to be only weakly dependent on the text that is being spoken. Obviously, good results will only be achieved if the read text is sufficiently rich and phonetically balanced.

Once robustness against text changes is achieved, one can envisage robustness against changes in the accent and the language, even though the latter may be hard to achieve since every language has its own sound system. On the other hand, by employing e.g. phonological descriptors, it may be possible to cross some language boundaries. In this respect we have already demonstrated (Middag et al., 2011) that phonological features learned on one language (Dutch) have predictive power in another Germanic language (German).

The rest of the paper is organized as follows. In Section 3.2 we review the corpus that will be used to experimentally validate the investigated methods. In Section 3.3 we consider some previously developed methods for creating speaker feature sets (see Middag et al., 2009; Middag et al., 2010) and we add a new method leading to a new speaker feature set. In Section 3.4 we describe the experiments we conducted and show that thanks to the new features an automatic intelligibility assessment that meets the envisioned robustness criteria and that is sufficiently accurate to monitor a patient over time is now possible.

3.2 Validation corpus

A very important issue with respect to automated intelligibility analysis is its experimental validation. Although previous research has shown high correlations between automatically generated scores and perceptual ratings, these correlations have always been measured on a large group of pathological speakers where the aim was to compare one speaker against another. However, in a clinical setting there is also a high need for tools that are able to monitor the progress of an individual patient. In order to evaluate whether our methods can accomplish this, we conducted experiments on the recently developed NKI-CCRT corpus.

All speech material in the new corpus was collected as part of a longitudinal study on voice and speech outcomes of patients with advanced head and neck cancer who were treated with concomitant chemoradiotherapy (CCRT, van der Molen et al., 2012). The perceptual evaluations are part of a larger study investigating the automatic evaluation of speech intelligibility and voice quality for speakers treated for advanced head and neck cancer. Here we just provide a synopsis of the information regarding participating speakers and the perceptual evaluations that have been performed on the data. We refer the reader to van der Molen et al. (2012) and Clapham et al. (2012) for more detailed information.

3.2.1 Speakers

The corpus contains recordings and perceptual evaluations of 55 speakers: 54 of them were recorded before CCRT (T0), 48 were recorded ten weeks after

	Text diversity		Syllable length	Sentence length
Fragment	Tokens	Types	Mean (SD, range)	Mean (SD, range)
A	70	49	1.3 (0.6,1-3)	11.7 (6.3,4-21)
B	68	50	1.5 (0.7,1-3)	17.0 (5.8,12-23)
A&B	138	77	1.5 (0.7,1-3)	13.8 (6.4,4-23)

Table 3.1: Characteristics of the two text fragments: number of tokens and number of types. Average syllable and sentence length are denoted in number of syllables and number of words respectively. Data are rounded to one decimal place.

CCRT (T1) and 39 were recorded a third time, twelve months after CCRT (T3)[2]. Average age at pre-treatment was 57 years (range 32-79) and the tumor locations are detailed in van der Molen et al. (2012). Based on perceptual categorization by a Dutch phonetician, 8 speakers were categorized as non-native whereas the other 47 were categorized as native.

As most speakers were recorded before CCRT and at two moments after CCRT, this dataset makes it possible to monitor short-term and long-term changes in a patient's intelligibility. Preliminary results presented by Clapham et al. (2012) show however that not all speakers exhibit statistically significant changes in perceptual speech intelligibility ratings over time.

3.2.2 Stimuli

Two fragments of a 189-word passage from a Dutch fairy tale were selected as fragments A and B. Fragment A contains 70 words (tokens) while fragment B contains 68 words. Fragment A contains 49 unique words (types) and fragment B contains 50 unique words (see Table 3.1). The two fragments have only 22 types in common, which makes them clearly lexically different. Each speaker read at least one of the fragments, but most of them read both: the corpus contains 141 recordings of fragment A and 140 of fragment B. Average durations of the recordings were 26.9 seconds for fragment A and 26.4 seconds for fragment B.

From the phoneme frequencies in fragments A and B (see Clapham et al., 2012), it follows that the two fragments have an almost identical phonetic balance.

[2] There were also recordings for some of the patients at time T2, situated between T1 and T3, but due to time constraints, these recordings were not perceptually rated, and therefore not used in this study. However, to maintain numerical consistency with the publication of van der Molen et al. (2012), we use the term T3 for the last recording moment.

3.2.3 Perceptual analysis

Thirteen recently graduated or about to graduate speech pathologists (all female, native Dutch speakers, average age of 23.7 years) evaluated the speech recordings in an on-line, self-paced experiment. The recordings were presented in a randomized order and listeners could replay a recording as many times as they wished. Each recording contained the reading of a complete fragment by one speaker. The listeners used their own anchors and received no feedback on performance. All listeners completed an on-line familiarization module before evaluating the stimuli for the dataset. The retest recordings (repetitions of formerly rated recording) and items for practicing are not included in the dataset.

Intelligibility was evaluated on a 7-point scale with labels provided for the scale ends ("poor" for 1 and "good" for 7). Preliminary results presented in Clapham et al. (2012) indicate that although some listener's test-retest reliability was low, the Interclass Correlation Coefficient (Shrout and Fleiss, 1979) assessing the between-rater reliability was 0.95 (based on a sample of 37 items). This high value indicates that the mean intelligibility scores are reliable. The percentage exact agreement for the rater's test-retest recordings ranged from 20 to 80 percent. The percent close agreement ($\pm$ 1 difference on the scale) ranged from 60 to 100 percent. In terms of Pearson Correlation Coefficient (PCC), the correlation between the scores of one individual rater and the mean perceptual ratings (= means of the scores of all 13 raters) varies between 0.72 and 0.92, with a mean of 0.84.

Figure 3.1 depicts the histogram of the mean perceptual ratings for all recordings.

3.3 Objective intelligibility assessment

The derivation of an objective intelligibility score is a multi-stage process involving an acoustic analysis, a phonetic or phonological analysis, a speaker feature extraction and an intelligibility prediction.

The **acoustic analysis** extracts a stream of acoustic parameter vectors $X_t, t = 1, ..., T$, with t a multiple of 10 ms, from the waveform.

The **phonetic or phonological analysis** aims at converting the acoustic parameter vectors into phonetic or phonological scores. A phonetic analysis generates scores for a finite set of distinctive speech sounds, called phones, that can be used to annotate how speech is perceived. A compact phone set for American English is the one that was used for the annotation of the TIMIT corpus (Fisher et al., 1986). A much more extended phone set is the set of triphones (context-dependent phones) used in modern ASR systems. A

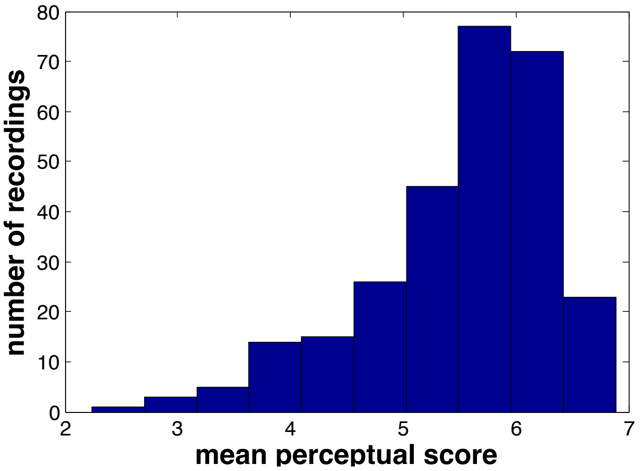


Figure 3.1: Histogram of the mean perceptual scores (= means over 13 raters).

phonological analysis generates scores for a finite set of binary phonological categories that can be used to annotate speech in terms of its production. Two examples of phonological categories are "voiced" and "nasal". The former indicates whether the vocal chords are vibrating, the latter whether the air streams through the nasal cavity. In both analyses, the scores are computed by means of stochastic acoustic models whose free parameters were optimized (trained) on a corpus of normal speech.

The **speaker feature extraction** derives holistic features that characterize the speech of a certain speaker as a whole. One approach is to make a speech-to-text alignment which produces the most likely segmentation into phones, given the acoustic model outputs and knowledge of the text that was spoken. From this segmentation one can then extract, per phone or phonological category, a holistic feature indicating the mean confidence of the acoustic models in time intervals that are assigned to that phone or phonological category (e.g. all back vowels). Another approach is not to perform any alignment but to characterize the acoustic model outputs as they evolve in time in the course of the speaker's utterances. Examples of such holistic features would be the mean nasality, the mean of the peaks in the voicing evidence, etc.

The **intelligibility prediction** is finally responsible for converting the speaker features into a speaker intelligibility score. It does so by means of a so-called intelligibility prediction model (IPM).

In the subsequent sections, we propose a number of approaches for deriving interesting speaker features and for developing robust IPMs on the basis of a limited amount of speaker data. In particular, we will discuss three previously proposed speaker feature sets as well as a novel set that is specifically designed with the aim of increasing the robustness of the IPM against text and accent changes.

3.3.1 Speaker feature extraction

We have investigated our speaker feature extraction methods according to two axes. One is whether or not they involve a speech-to-text alignment. Another is whether they incorporate a phonetic or a phonological analysis.

3.3.1.1 Alignment-based features

In an alignment-based approach we make use of the prompted text to create a Hidden Markov Model (HMM) $\mathcal{M}$ that can generate all possible utterances of that text. The creation of that HMM relies on a pronunciation dictionary containing canonical phonemic transcription of all words, a procedure for converting phonemic transcriptions to phone sequences (see further) and a predefined model architecture for each phone model (e.g. a three state model). The acoustic models are needed to compute the probability $P(X, S|\mathcal{M})$ of generating acoustic parameter sequence X along a state sequence S[3]. The task of the aligner is thus to find the most likely state sequence S along which X can be generated.

We have experimented with two aligners (see also Middag et al., 2008). The first one, called ASR-ESAT, uses an inventory of context-dependent phones (triphones) and Gaussian Mixture Models (GMMs) to compute the $P(X_t|s_t)$. The second one, called ASR-ELIS, uses a compact phone inventory which is the Flemish equivalent of the set that was employed for annotating TIMIT.

Speaker feature extraction with the ASR-ESAT aligner

The ASR-ESAT aligner (Demuynck, 2001) works with acoustic parameter vectors that are created as follows: computation of 24 log-mel-spectral coefficients per time step (Davis and Mermelstein, 1980), application of noise masking and spectral mean normalization (D. Van Compernolle, 1989), addition of first

[3] We are aware that strictly speaking the probability is a likelihood, but in spite of this it is conventional to make no distinctions between likelihoods and probabilities so as to simplify the discussions.

and second order derivatives, decorrelation of the 72-dimensional vectors (Demuynck et al., 1998) and dimensionality reduction via MIDA (Demuynck et al., 1999). The final vectors are 39-dimensional.

The aligner uses context dependent phones, called triphones as they are characterized by a central phone and its left and right neighbors. Each triphone has three states, but a global decision tree clusters the many thousands of triphone states into 1567 tied states each modeled by a separate GMM. However, all GMMs are built on one large set of state-independent Gaussians (= semi-continuous HMM).

Omitting the $\mathcal{M}$ from the formerly introduced notation, the probability $P(X, S)$ is computed as

$$P(X,S) = \sum_{t=1}^{T} P(s_t|s_{t-1})\, P(X_t|s_t) \tag{3.1}$$

with $P(s_t|s_{t-1})$ representing the constraints imposed by the HMM and $P(X_t|s_t)$ the so-called emission probabilities computed by the GMMs.

The speaker feature extraction then works in two stages. First of all it converts the likelihoods $P(X_t|s_t)$ to posterior probabilities:

$$P(s_t|X_t) = \frac{P(X_t|s_t)P(s_t)}{P(X_t)} \tag{3.2}$$

$$P(X_t) = \sum_{j=1}^{N_S} P(X_t|s_t = S_j)P(S_j) \tag{3.3}$$

with the summation taken over the set $\mathcal{S} = \{S_j : j = 1, .., N_S\}$ of all possible triphone states. Then, it takes the mean posterior over all states belonging to a triphone with a particular central phone F_k. Repeating this for all monophones F_k $(k = 1, .., N_F)$ leads to $N_F = 40$ monophone speaker features.

Speaker feature extraction with the ASR-ELIS aligner

The ASR-ELIS aligner works with acoustic parameter vectors that are created as follows: computation of the log-energy + 12 MFCCs per time step ((Davis and Mermelstein, 1980)), cepstral mean normalization and addition of first and second order derivatives. Each vector consists of 39 components.

The system employs 55 phones: 40 monophones, 6 plosive closures, 6 plosive bursts, a glottis and two silence symbols to accommodate inter and intra-sentence pauses. Each phone is modeled by a single-state model. The probability $P(X, S)$ is computed as

$$P(X,S) = \sum_{t=1}^{T} P(s_t|s_{t-1})\, P(X_{t-5}^{t+5}|s_t) \tag{3.4}$$

with X_{t-5}^{t+5} representing the acoustic parameter vectors $X_{t-5}, .., X_{t+5}$.

The acoustic models work in two stages: first of all, a neural network based phonological category detector (Figure 3.2) extracts the posterior probabilities

$$Y_{tm} \doteq P(C_m|X_{t-5}^{t+5}), \qquad m = 1, .., N_C = 24 \tag{3.5}$$

that phonological categories C_m are on/active/present at time t. In the second stage, $P(X_{t-5}^{t+5}|s_t)$ is replaced by $P(Y_t|s_t)/P(Y_t)$ and the latter is computed as

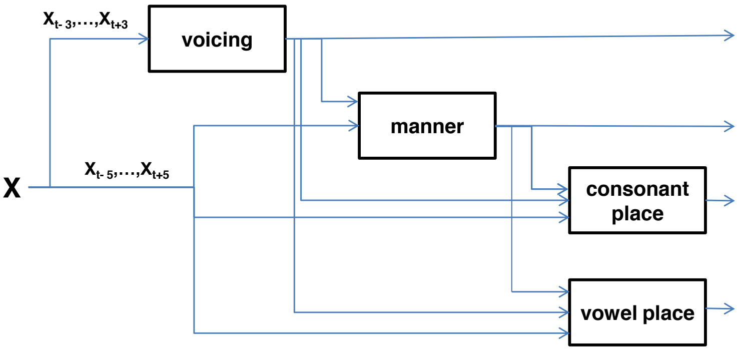


Figure 3.2: Architecture of the phonological feature analyzer: see (Stouten, 2008).

$$\frac{P(Y_t|s_t)}{P(Y_t)} = \frac{P(s_t|Y_t)}{P(s_t)}, \qquad P(s_t = S_j|Y_t) = \left[\prod_{m, V_m(S_j)=1} Y_{tm} \right]^{\frac{1}{N_p(S_j)}} \tag{3.6}$$

where $\mathcal{S} = \{S_j : j = 1, .., N_S = 55\}$ is the set of phone states, $V_m(S_j) = 1$ means that S_j belongs to category C_m and $N_p(S_j)$ is the number of categories S_j belongs to. For more details and motivation and for a complete list of phonological categories, the reader is referred to (Stouten and Martens, 2006). Here we just mention some typical examples such as "voiced" (= vocal source class), "burst" (= manner class), "labial" (= place of a consonant) and "mid-low" (= height of a vowel). The fact that probabilities are estimated over a time interval of 125 ms (= 10 times frame shift + 1 time frame size) means that co-articulations between phones can be handled implicitly, even though the Y_t will be evaluated using monophone state distributions.

The speaker feature extractor now takes the mean of the posterior probabilities $P(s_t|Y_t)$ over all frames that were assigned to a state belonging to category C_m ($V_m = 1$). This leads to N_S positive features $PLF_1(m)$. In a similar vein, it also computes negative feature $PLF_o(m)$ by taking means over frames that were assigned to a state not belonging to category C_m $(V_m = 0)$.

The problem with the straightforward averaging method described above is that the different phones contribute with a different and text-dependent weight to a particular PLF(m). Therefore, we proposed (Middag et al., 2008) to take an average per phone first and to take the mean of these averages over all phones with the right V_m (0 or 1).

Note that certain features can be irrelevant for some of the phones (e.g. a vowel category is irrelevant for a consonant), so that not all frames are necessarily involved in the computation of the positive and negative features of a certain category.

Later we will retrieve phonological features for Dutch speech from either a Flemish or Dutch phonological category detector. However, the two accents of Dutch differ in e.g. the voicing of fricatives and the degree of diphtonguation of long vowels (Nerbonne et al., 1995; Van Compernolle et al., 1991). One can still use a Flemish detector for analyzing Dutch speech provided the the phonological descriptions of the phones are set correctly. E.g. the /g/ that was "voiced" during the training of the Flemish detector must be set "unvoiced" for the assessment of Dutch speech. Otherwise, the speech-to-text alignment might derail for some sentences.

3.3.1.2 Alignment-free features

We conjecture that in order to achieve a robust intelligibility assessment, we should utilize speaker features that can be obtained without exact knowledge of the text that is spoken, i.e., without the requirement of a meticulous speech-to-text alignment. Here we discuss two such feature sets, namely a previously developed one and a newly proposed one.

Phonological features

In Middag et al. (2010) we have developed phonological speaker features that can be computed without any knowledge at all about the read text. As their extraction does not require any speech-to-text alignment, we call the new features alignment-free and we denoted them as ALF-PLF. The only kind of 'alignment' that is needed is one at the level of speech or silence. Silences longer than 1 second are detected by means of an energy-based silence detector and are excluded from further analysis.

As before, the vectors X_t are converted into posterior phonological category probabilities Y_t, but this time we only distinguish categories that can be retrieved from very local information carried by X_{t-1}, X_t and X_{t+1} (see below for a motivation). For instance, a modulation feature like "trill" is not considered.

Three binary categories "voicing", "silence" and "turbulence" are always on or off and are modeled by a single MLP. Nine other categories, like "nasal", can be on, off or irrelevant and are modeled by a tandem of two MLPs: one that distinguishes between relevant (=1) and irrelevant (= 0) and another that distinguishes between on (= 1) and off ($= -1$). The latter MLP also takes the output of the first MLP into account. Two vowel categories "back" and "high" are modeled in a slightly different way. Here the first MLP makes a difference between "non-central" and "central" or "consonant" (and something similarly for "high"). The argumentation is that pathological speakers mainly experience problems with vowels at the extremes of the vowel trapezium, and not with the central vowels, and therefore one can consider "central" as irrelevant for measuring deficiencies along the place or hight dimensions. To sum up, the output of the phonological analyzer consists of 25 components Z_t.

For each speaker, a statistical analysis of the temporal evolution of each individual component of Z_t is performed. This analysis yields 12 measurements per component, e.g. mean value, standard deviation, percentage of positive, negative and close-to-zero values, mean of the peaks and the valleys, mean time needed to reach a peak or valley, etc. In total, the analysis thus yields 12 x 25 = 300 ALF-PLFs. The hypothesis is that temporal fluctuations in the components of Z_t can reveal articulatory deficiencies, regardless of the exact phonetic content of the text that was read, at least as long as this text is sufficiently rich in phonetic content.

Phonological classes are in principle universal (cross-lingual) but they are extracted by models that were trained on data of one language, exposing contextual influences which are typical for that language. We argue that by supplying only very local information (X_{t-1}, X_t and X_{t+1}) to the phonological analyzer, we can achieve that these contextual factors have only a weak impact on the trained models. As such, these models are expected to be predictive in other languages than the one available during training. Our previous work (Middag et al., 2011) confirmed that ALF-PLF derived from the outputs of Flemish phonological detectors can predict the intelligibility of German pathological speakers.

Monophone features

The ALF-PLF are expected to be powerful if the intelligibility reduction due to a certain speech disorder can be attributed to problems with the realization of individual phonological classes. Nevertheless, it may well be that this degradation mainly follows from problems that only arise when a certain combination of phonological classes must be realized, e.g. the realization of "voicing" and "fricative" in phone /z/. In that case, intelligibility prediction could benefit more from features that take these interactions between phonological classes into account.

An obvious way to accommodate this is the following. In the first stage one assigns all frames to the phone F_k (= state S_k) that yields the maximal posterior probability according to Equation 3.6. In the second stage one considers all frames assigned to phone F_k and one measures the mean, the standard deviation, the mean of the valleys and the mean of the peaks as the four speaker features for phone F_k. Note that we return to a context of 5 frames to the left and to the right again because we reckon that contextual modeling is a requisite to get good context-independent phone evidences.

To complete the our feature set, we also count the frames where F_k had the maximum posterior probability and convert it to $P(F_k|U,R)$, the probability that F_k appears in the utterance U (actually the concatenation of all sentences spoken by the speaker) when the text to read was R. Clearly this probability can be decomposed as follows:

$$\begin{aligned} P(F_k|U,R) &= \sum_{u,r=1}^{N_F} P(F_k, F_u, F_r|U,R) & k = 1,..,N_F \quad (3.7) \\ &= \sum_{u,r=1}^{N_F} P(F_r|R)\, P(F_u|F_r)\, P(F_k|F_u) & k = 1,..,N_F \quad (3.8) \end{aligned}$$

The meanings of the probabilities in the right hand side are the following:

- $P(F_k|F_u)$ is the probability that F_k is the winner when the speaker tries to utter F_u. Obviously it depends on the quality of the phonological analyzer, but more importantly, on the difficulties the speaker experiences to pronounce F_u.

- $P(F_u|F_r)$ is the probability that the speaker tries to pronounce F_u when according to the canonical transcription of the text it should have been F_r. It is a measure of how many times the speaker is making a reading error.

- $P(F_r|R)$ is the probability that F_r appears in the canonical transcription of the text. This is strictly a property of the text, but if the text is long enough it will be more like a property of the language.

The above formulation allows us to demonstrate that as long as the number of reading errors is small, the ratio $P(F_k|U,R)/P(F_k|R)$ is bound to be a sensible text-independent feature to add to the other ALF features. Indeed, under the given assumption Equation (3.8) can be simplified and one obtains that

$$\frac{P(F_k|U,R)}{P(F_k|R)} \simeq \sum_{r=1}^{N_F} \frac{P(F_r|R)}{P(F_k|R)}\, P(F_k|F_u = F_r) \qquad (3.9)$$

Obviously, the sum in the right hand side is bound to contain only a few relevant terms. If $P(F_r|R)$ is much lower than $P(F_k|R)$, the term is obviously negligible. On the other hand, if $P(F_r|R)$ is much larger than $P(F_k|R)$ we suppose that this will be true for any text, and in that case one can expect the acoustic model of F_r to be much better trained than that of F_k. Consequently, it is very unlikely then that F_k will be the winner when $F_r \neq F_k$ is uttered. We can thus conclude that the sum will only contain components with a weight $P(F_r|R)/P(F_k|R)$ that is close to 1, and therefore, that the sum will only weakly depend on the text that was read.

By also adding the same probability ratios, but this time with a nominator that the mean posterior probability of F_k over all frames of the utterance, we finally obtain $6N_F = 6 * 55 = 330$ alignment-free monophone speaker features, denoted as ALF-MPF.

3.3.2 Intelligibility prediction model

Once all speaker features have been computed, they need to be converted to an intelligibility score using a regression model, hereafter called the intelligibility prediction model (IPM).

A variety of statistical learners is available for optimizing regression problems. However, in order to avoid over-fitting, only a few of them can be applied to a data set comprising as few as N_{sp} = 55 speakers. We therefore opt for ensemble linear regression (ELR), which combines the low model complexity of linear regression with a bagging strategy (Breiman, 1996). The latter boosts the predictive power by using many simple models (linear regression models in this case) which are each trained on a different random subset of the training data. For the training of our ELR model we create ten random divisions of the training set into two equally large parts: one part for estimating the regression coefficients and the other for assessing the model. As we have a large number of features at our disposal and as every division will only comprise a very restricted number of speakers (not more than 28), some feature selection procedure is indispensable. Every single model is created by adopting a greedy forward feature selection procedure which starts with the feature leading to the best performance and continues to add features as long as that performance rises. The utilized performance criterion is the Root Mean Squared Error (RMSE) between targeted and computed scores. Typically, the number of selected features varies between 2 and 10.

To compute the intelligibility of a test utterance, we employ the ten models emerging from the training set divisions to yield one estimate and we take the average of these ten estimates. In practice, this final score can be achieved more efficiently by constructing a single linear model in the space of the features that were selected by at least one of the ten models.

3.4 Experimental evaluation

The main objectives of the experimental evaluation were to assess the accuracy of the IPMs derived from the different speaker feature sets and their robustness against changes in the read text and the spoken accent (Dutch or Flemish). In order to reach these objectives we have derived IPMs from different text fragments (fragments A and B).

In order to investigate accent dependency, we tested feature sets derived by means of acoustic models trained on Flemish and Dutch normal speech respectively. The Flemish phonological category models were trained on 7 hours of read speech from the CoGeN corpus (Demuynck et al., 1997). The speech came from 174 persons residing in Flanders, the northern part of Belgium. The Flemish phone models were trained on 40 hours of read speech from the Spoken Dutch Corpus (Schuurman et al., 2003), namely speech of 150 speakers residing in Flanders. The Dutch phone and phonological category model sets were both trained on 64 hours of read speech from the Spoken Dutch Corpus, namely speech of 324 speakers residing in the Netherlands.

Before describing our experimental results in more detail, we first take a closer look at the evaluation strategy we have adopted. All IPMs were trained and evaluated using a 5-fold cross validation (CV) strategy. As most speakers were recorded two or three times (at T0, T1 and/or T3) and since two fragments (A and/or B) were recorded in most cases, 281 samples were available in total. These samples were divided into five folds such that all recordings of one speaker always belonged to one fold. Performance is expressed in terms of the RMSE and the Pearson Correlation Coefficient (PCC) between computed and perceptual intelligibilities. The latter were defined as mean scores over all human raters. The Wilcoxon signed-rank test (Sheskin, 2004) is used to investigate whether results are significantly different at a confidence level of 0.05.

3.4.1 Individual speaker feature sets

In a first experiment we tested the four feature sets we proposed in combination with IPMs that were trained and tested on the same fragment. In view of later experiments however, we introduce the notation $A \rightarrow B$ for instance to express that the IPM is trained on fragment A and tested on fragment B. In Table 3.2 one finds the results for the cases $A \rightarrow A$ and $B \rightarrow B$, and for features that were either computed with the help of Dutch or Flemish acoustic models.

The main conclusion is that the monophone features (MPF and ALF-MPF) outperform the corresponding phonological features in most cases. There is only one exception to this rule, namely the Flemish MPF performing worse than ALF-PLF on fragment A. The difference between PLF and MPF can

	$A \rightarrow A$				$B \rightarrow B$			
	FL		DU		FL		DU	
Features	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC
MPF	0.82	0.60	**0.65**	**0.77**	0.68	0.73	**0.60**	**0.77**
PLF	0.83	0.58	0.79	0.60	0.75	0.63	0.68	0.72
ALF-PLF	0.77	0.63	0.77	0.62	0.74	0.66	0.73	0.66
ALF-MPF	0.68	0.73	0.68	0.73	0.70	0.70	0.70	0.70

Table 3.2: Performances of IPMs using from Flemish (FL) or Dutch (DU) feature sets. Per training and test fragment combination, results differing significantly at a level of $p < 0.05$ from the best result (indicated in bold) are underlined.

be partly explained by differences in the systems supplying the text-to-speech alignments that are needed for constructing the speaker features: the state-of-the-art ESAT-ASR usually leads to a better alignment than the much less complex ELIS-ASR. The difference between ALF-MPF and ALF-PLF on the other hand cannot be explained in terms of the alignment (there is none) nor in terms of the phonological analyzers that were used (they were actually very similar). The data seem to support the hypothesis that intelligibility reductions are more correlated with co-occurrences of phonological classes, as they materialize in specific phonetic units, than with individual phonological classes.

The second conclusion we can draw is that the alignment-free features have a more consistent performance across different configurations than the alignment-based features. On the other hand, the alignment based features do usually lead to the highest performance (again with the exception of Flemish MPF on fragment A). The latter is due to the fact that the speakers recorded in the NKI-CCRT corpus were mostly native adults who did not make many reading errors which could have derailed the alignment.

3.4.2 Robustness against speaker accent

A very striking result with respect to the impact of the speaker accent is that the alignment based methods are sensitive to a change of accent whereas the alignment-free methods are not. As expected, the alignment based models clearly perform better when the acoustic models are matched to the accent of the speaker.

That alignment-free features are so robust must mean that the global statistical analysis conducted to retrieve alignment-free parameters is robust against differences in the quality of the phonological analyzer, whereas the state-by-state analysis in an alignment-based method tends to be sensitive to the quality of the best alignment path.

3.4.3 Robustness against changes in the text

In order to investigate this aspect we conducted an additional experiment in which we tested the matched alignment-based feature sets (DU-MPF and DU-PLF) and the matched alignment-free feature sets (DU-ALF-PLF and DU-ALF-MPF) in combination with matched and unmatched IPMs. The IPM is called unmatched if it is not trained and tested on recordings of the same text fragment. The results obtained with different combinations of text fragments can be found in Table 3.3.

The data clearly demonstrate that all feature sets show the same performance on a particular test fragment, irrespective of whether the IPM was trained on the same or on another text. However, the used test set does play a role. The differences between the figures obtained by testing on A and B are much larger for the alignment-based than for the alignment-free feature sets. This proves that the latter feature sets are more robust to changes in the text during evaluation. We argue that this stems from the fact that the quality of the alignment depends to some extent on the phonetic content of the text (it is known that some sound sequences are much more difficult to segment than others). The mismatch between the training and the evaluation text does not seem to be a problem though.

	$A \rightarrow A$		$B \rightarrow A$		$B \rightarrow B$		$A \rightarrow B$	
Feature set	RMSE	PCC	RMSE	PCC	RMSE	PCC	RMSE	PCC
DU-MPF	0.65	0.77	0.68	0.76	0.60	0.77	0.60	0.76
DU-PLF	0.79	0.60	0.80	0.61	0.68	0.72	0.66	0.72
DU-ALF-PLF	0.77	0.62	0.77	0.62	0.73	0.66	0.73	0.65
DU-ALF-MPF	0.68	0.73	0.70	0.72	0.70	0.70	0.71	0.70

Table 3.3: IPMs developed on fragment X (A or B) and tested on fragment Y (A or B) as indicated by the notation $X \rightarrow Y$.

3.4.4 A combination of speaker features

From the former experiments it follows that none of the speaker feature sets leads to a correlation between the objective and the perceptual scores that can compete with the mean correlation of 0.84 observed between individual raters and the mean of these raters. In this respect, we have investigated whether the combination of different feature sets may bridge this gap. We tested combinations of

- the two alignment-based feature sets DU-PLF and DU-MPF,
- the two alignment-free feature sets DU-ALF-PLF and DU-ALF-MPF,

- the two phonological feature sets DU-PLF and DU-ALF-PLF, and
- the two phone(m/t)ic feature sets MPF and ALF-MPF.

The results obtained with these combinations are listed in Table 3.4.

Clearly all feature set combinations perform better than the individual feature sets they are composed of, but in most cases the improvement is not significant. The IPM incorporated in the present DIA tool achieves a good PCC but it does not achieve the envisaged RMSE. The combination of the formerly proposed DU-MPF and the newly developed DU-ALF-MPF on the other hand does. Figure 3.3 shows a convincing scatter plot of the perceptual scores versus the objective scores computed by the IPM designed for this combination.

feature combination	RMSE	PCC
DU-PLF + DU-MPF	0.61	0.80
DU-ALF-PLF + DU-ALF-MPF	0.68	0.73
DU-PLF + DU-ALF-PLF	0.64	0.74
DU-MPF + DU-ALF-MPF	**0.52**	**0.85**

Table 3.4: Predictive power of IPMs built on different combinations of two feature sets. Listed are RMSE and PCC between the computed results and the means of the 13 perceptual raters. Underlined results differ significantly $(p < 0.05)$ *from the best result, denoted in bold.*

For the best combination we have also investigated in more detail how many features and which features were selected. As we adopted a five-fold cross validation strategy, 5 models were created and each of these models was on its turn obtained as a combination of 10 small models each selecting 7 - 8 features. On average, the combined model incorporated 25 features (range 21-29). Per fold, statistics were calculated on how many times a feature was selected in one of the ten small models. Features selected 5 times or more are the MPFs /r/,/A/,/@/,/i/ and the ALF-MPFs /A_min/ and /N_max/, where /A/ is the vowel in the Dutch word "man", /@/ stands for the schwa in "de", /i/ is the long vowel of "tien". Furthermore, /A_min/ is the mean of the valleys for /A/ and /N_max/ is the mean of the peaks for /N/, which is the final nasal sound of the word "koning". Apparently, four out of six features are vowel-related. If we take a closer look at the features, we observe that these vowels define the diagonal of the vowel trapezium in the (place,height) plane: /i/ determines the upper-left corner (as it is *front* and *high*) while /A/ determines the lower-right corner (as it is *back* and *low*) and /@/ represents the center of this diagonal. Consequently, the vowel features can represent the amount of variation from the neutral (central) position the speaker can achieve in two directions. Together they represent the size of the speaker's vowel trapezium as a potential factor affecting his intelligibility.

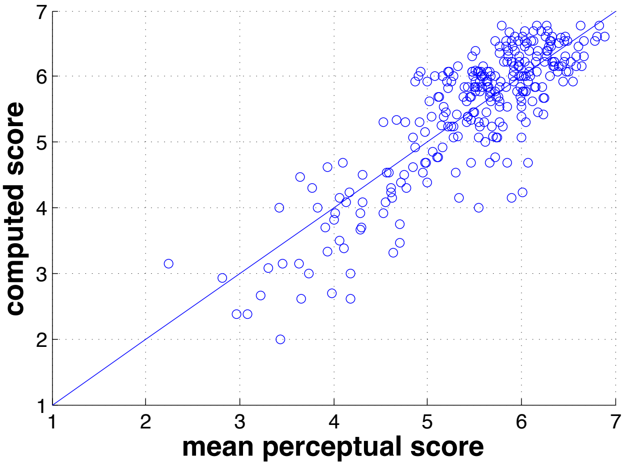


Figure 3.3: Correlation between perceptual and computed scores.

Although few articles describe the speech of people treated with chemoradiation therapy, it is known that even chemo-radiation therapy affects the organic structures and tissues around the tumor location (van der Molen et al., 2012). Swallowing problems are common, and tongue and palate tissues are affected at least for part of the speakers. Persons with reduced tongue motility are known to show a strong correlation between intelligibility and vowel trapezium size (de Bruijn et al., 2009; Neel, 2008).

The fact that tongue motility can be affected in this patient group also explains the selection of features /r/ and /N_max/ as realizations of the uvular /r/ and /N/ need good functioning of the back of the tongue. Secondly, van der Molen et al. (2012) shows that nasality is significantly worsened by CCRT treatment. Nasality is thus an issue in our dataset, and it is not so surprising then to notice that a nasal related feature such as /N_max/ is selected.

3.4.4.1 Patient monitoring

Now that we have established an IPM that can mimic evaluations made by a group of listeners for the comparison of one speaker against another, the next challenge is to prove that this model can also track trends in an individual patient's intelligibility over time.

First of all we have investigated whether such trends are exposed by the perceptual scores. To that end we have determined the differences between the ratings of the same fragment read by the same speaker at times T0 and T1, T1 and T3 and T0 and T3. A former analysis of these data (Clapham et al., 2012) demonstrated that not all speakers show a clear trend (neither progress nor deterioration) over time. For each speaker we computed the rating differences at times T0 and T1, T1 and T3 and T0 and T3 and the PCC between the differences derived from ratings of one rater and those derived from the mean ratings (over 13 raters). As revealed by Table 3.5, the PCC are rather low. Since we did not expect our IPM to outperform human raters, we selected those speakers for which the human raters seemed to agree on the presence and direction of the trend. The correlations between one rater and the mean of the 13 ratings for these speakers are listed in Table 3.5, together with the number of recordings for which this is the case.

	All trends			Only clear trends		
Times	Mean	Range	Number	Mean	Range	Number
T1-T0	0.56	0.45 - 0.70	93	0.70	0.40 - 0.84	26
T3-T1	0.44	0.17 - 0.62	74	0.75	0.43 - 0.96	8
T3-T0	0.62	0.45 - 0.75	78	0.78	0.60 - 0.89	28

Table 3.5: Inter-rater agreements (PCC) (mean and range over tested speakers) about speaker trends measured on all trend data and on the data exhibiting a clear trend. The number of tested speakers is mentioned under the "number" columns.

	All trends			Only clear trends		
Times	**IPM**	Mean	Range	**IPM**	Mean	Range
T1-T0	**0.41**	0.56	0.45 - 0.70	**0.51**	0.70	0.40 - 0.84
T3-T0	**0.62**	0.62	0.45 - 0.75	**0.82**	0.78	0.60 - 0.89

Table 3.6: Correlations on speaker trend level. Results from the IPM are marked in bold.

Based on the data in Table 3.5, we can conclude that one can only measure a clear trend from T1 to T3 for 8 speakers. As this is considered insufficient to measure reliable correlations, we only analyzed the correlations between T0 and T1 and between T0 and T3. The results of this analysis are listed in Table 3.6.

In the case of T3-T0, the mean human-machine-correlation is as good as the mean correlation between one rater and the mean rating, and even better for the clear trends. For T1-T0, it is lower, but nevertheless, the human-machine correlation is in the range of human correlations, at least for the cases with a clear trend. We can therefore conclude that the IPM we developed seems able to follow the progress of an individual speaker as (un)reliably as a human rater can.

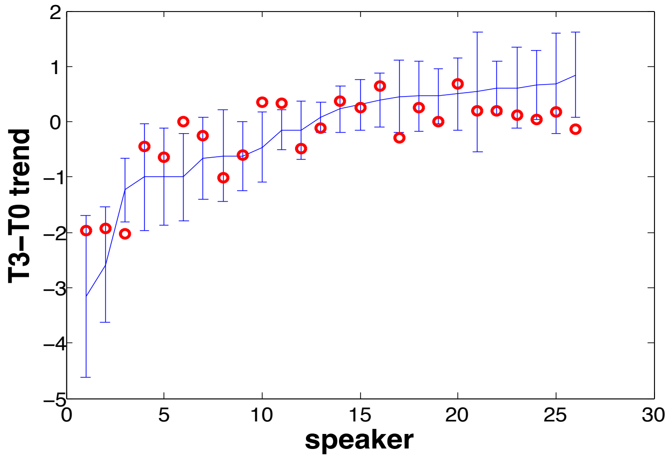


Figure 3.4: Measured and predicted trends between T0 and T3 for the speakers exhibiting a clear trend (see text)

Figure 3.4 shows the means and standard deviations of the T3-T0 differences in the human ratings for the 26 cases that were categorized as exhibiting a trend. Also on the Figure one finds the predicted trends. There is a lot of uncertainty on the human ratings but for 10 out of 13 of the subjects exhibiting a negative trend, the model also predicts a negative trend. The positive trends are less pronounced and, likewise, not so well predicted. Needless to say that the plot for the T1-T0 differences is less convincing given the lower PCC. We conjecture that it takes more reliable human ratings to generate better automatic trend predictions.

3.5 Conclusions and future work

In previous work (Middag et al., 2008) we demonstrated that an alignment based method combining two distinct ASRs can yield good correlations between subjective (human) and objective (computed) intelligibility scores. More recently (Middag et al., 2010) we also succeeded in showing that alignment-free methods have potential as well to predict intelligibility from running speech. In this paper, we extended our work by proposing a new alignment-free feature set which is designed to be text-independent and applicable for different languages. For the time being, we validated this feature set when developed on Flemish data, on a Dutch dataset, called the NKI-CCRT dataset. Although Dutch and Flemish are not really two distinct languages, they do represent two very different regional accents of Dutch.

Comparing results from IPMs (Intelligibility Prediction Models) built on Flemish and Dutch acoustic models respectively, we could establish that the alignment based methods are clearly language sensitive whereas the alignment-free methods are not. Comparing results emerging from IPMs built on different text fragments, we discovered that all feature sets are largely text-independent, at least in the absence of reading errors.

Our experiments show that by using one single speaker feature set, we were unable to create an IPM that is as reliable as a human rater. On the other hand, by combining the Dutch versions of the feature sets currently used in the DIA tool we already get a human-machine correlation of 0.80, which is only slightly worse than human-based evaluations. Combining alignment-free and alignment based monophone features leads to a model that can compete with a human rater for comparing one pathological speaker to another. Moreover, the IPM built on these two feature sets is capable of detecting progress or deterioration of a patient to the same extent humans can.

As the NKI-CCRT dataset not only contains speech intelligibility ratings but also ratings concerning articulation, voicing etc., future work will focus on the further development of a robust diagnosing system that also offers a more detailed speaker profile concerning articulation, voicing etc. From such a profile one could then retrieve objective and detailed information about the progress of a certain patient in the course of a therapy as well as information which could help determining the right personalized therapy for each patient.

Acknowledgements

Audio recordings were collected by Lisette van der Molen and Irene Jacobi. The perceptual scores were collected as part of a research project with Frans Hilgers and Michiel van den Brekel.

References

L. Breiman. Bagging Predictors. In *Machine Learning*, volume 24, pages 123–140, 1996.

R. P. Clapham, L. van der Molen, R. J. J. H. van Son, M. W. M. van den Brekel, and F. J. M. Hilgers. NKI-CCRT Corpus: Speech Intelligibility Before and After Advanced Head and Neck Cancer Treated with Concomitant Chemoradiotherapy. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, page In Press, 2012.

D. Van Compernolle. Noise Adaptation in a Hidden Markov Model Speech Recognition System. *Computer Speech and Language*, 3(2):151–168, 1989.

S. B. Davis and P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-28(4):357–366, 1980.

M. S. De Bodt, C. Guns, and G. Van Nuffelen. *NSVO: Nederlandstalig Spraakverstaanbaarheidsonderzoek*. Herentals: Vlaamse Vereniging voor Logopedisten, 2006.

M. J. de Bruijn, L. ten Bosch, D. J. Kuik, H. Quené, J. A. Langendijk, and C .R. Leemans I. M. Verdonck-de Leeuw. Objective Acoustic-Phonetic Speech Analysis in Patients Treated for Oral or Oropharyngeal Cancer. *Folia Phoniatrica et Logopaedica*, 61(3):180–187, 2009.

K. Demuynck. *Extracting, Modelling and Combining Information in Speech Recognition*. PhD thesis, K.U.Leuven, ESAT, 2001.

K. Demuynck, D. Van Compernolle, C. Van Hove, and J. P. Martens. *Een Corpus gesproken Nederlands voor spraaktechnologisch Onderzoek. Final Report of CoGeN Project*. ELIS UGent, Gent, 1997.

K. Demuynck, J. Duchateau, D. Van Compernolle, and P. Wambacq. Improved Feature Decorrelation for HMM-based Speech Recognition. In *Proceedings of the International Conference on Spoken Language Processing*, volume VII, pages 2907–2910, Sydney, Australia, December 1998.

K. Demuynck, J. Duchateau, and D. Van Compernolle. Optimal Feature Sub-space Selection based on Discriminant Analysis. In *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, volume III, pages 1311–1314, Budapest, Hungary, September 1999.

T. H. Falk, W. Y. Chan, and F. Shein. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication*, 54(5):622–631, 2011.

L. Ferrier, H. Shane, H. Ballard, T. Carpenter, and A. Benoit. Dysarthric Speakers' Intelligibility and Speech Characteristics in Relation to Computer Speech Recognition. *Journal of Augmentative and Alternative Communication*, 11: 165–74, 1995.

W. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The DARPA Speech Recognition Research Database: Specifications and Status. In *DARPA Workshop on Speech Recognition*, pages 93–99, 1986.

L. Gu, J. G. Harris, R. Shrivastav, and C. Sapienza. Disordered Speech Assessment Using Automatic Methods based on Quantitative Measures. *EURASIP Journal on Applied Signal Processing*, 9:1400–1409, 2005.

J. P. Hosom, L. Shriberg, and J. Green. Diagnostic Assessment of Childhood Apraxia of Speech Using Automatic Speech Recognition (ASR) Methods. In *Journal ofMedical Speech Language Pathology*, volume 12, pages 167–171, 2004.

R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek. Toward Phonetic Intelligibility Testing in Dysarthria. *Journal of Speech and Hearing Disorders*, 54:482–499, 1989.

A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth. PEAKS - A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51:425–437, 2009.

C. Middag, G. Van Nuffelen, J. P. Martens, and M. De Bodt. Objective intelligibility assessment of pathological speakers. In *Proceedings of the International Conference on Spoken Language Processing, Brisbane, Australia*, pages 1745–1748, 2008.

C. Middag, J. P. Martens, G. Van Nuffelen, and M. De Bodt. Automated Intelligibility Assessment of Pathological Speech Using Phonological Features. *EURASIP Journal on Advances in Signal Processing*, 2009:9, 2009.

C. Middag, Y. Saeys, and J. P. Martens. Towards an ASR-Free Objective Analysis of Pathological Speech. In *Proceedings of the International Conference on Spoken Language Processing, Tokio, Japan*, pages 294–297, 2010.

C. Middag, T. Bocklet, J. P. Martens, and E. Nöth. Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment. In *Proceedings of the International Conference on Spoken Language Processing, Florence, Italy*, pages 3005–3008, 2011.

A. T. Neel. Vowel Space Characteristics and Vowel Identification Accuracy. *Journal of Speech, Language, and Hearing Research*, 51(3):574, 2008.

J. Nerbonne, W. Heeringa, E. van den Hout, P. van der Kooi, S. Otten, W. van de Vis, and Alfa-informatica Bcn. Phonetic Distance between Dutch Dialects. In *Proceedings of the Computational Linguistics in the Netherlands meeting, Antwerp*, pages 185–202. Available, 1995.

I. Schuurman, M. Schouppe, H. Hoekstra, and T. van der Wouden. CGN, an Annotated Corpus of Spoken Dutch. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, 2003.

D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, 2004.

P. R. Shrout and J. L. Fleiss. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86:420–428, 1979.

F. Stouten. *Feature Extraction and Event Detection for Automatic Speech Recognition* . PhD thesis, Universiteit Gent, 2008.

F. Stouten and J. P. Martens. On the Use of Phonological Features for Pronunciation Scoring. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 329–332, 2006.

D. Van Compernolle, J. Smolders, P. Jaspers, and T. Hellemans. Speaker Clustering for Dialectic Robustness in Speaker Independent Speech Recognition. In *Proceedings of the European Conference on Speech Communication and Technology,Genova, Italy*, pages 723–726, 1991.

L. van der Molen, M. A. van Rossum, I. Jacobi, R. J. J. H. van Son, L. E. Smeele, C. R. N. Rasch, and F. J. M. Hilgers. Pre- and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: Expert listeners' and patient's perception. *Journal of Voice*, online, January 3 2012.

G. Van Nuffelen. *Speech Intelligibility in Dysarthria - Assessment and Treatment*. PhD thesis, Universiteit Antwerpen, 2009.

G. Van Nuffelen, M. De Bodt, F. Wuyts, and P. Van de Heyning. Reliability and Clinical Relevance of a Segmental Analysis based on an Intelligibility Assessment. *Folia Phoniatrica et Logopaedica*, 60:264–268, 2008.

P. Vijayalakshmi, R. Reddy, and D. O'Shaughnessy. Assessment of Articulatory Sub-systems of Dysarthric Speech Using an Isolated-style Phoneme Recognition System. In *Proceedings of the International Conference on Spoken Language Processing*, pages 981–984, 2006.

P. Vijayalakshmi, T. Nagarajan, and M. R. Reddy. Assessment of Articulatory and Velopharyngeal Sub-systems of Dysarthric Speech. *International Journal of Biomedical Soft Computing and Human Sciences, special issue on Biosensors: Data acquisition, Processing and Control*, 14(2):87–94, 2009.

K. M. Yorkston., E. A. Strand, and M. R. T. Kennedy. Comprehensibility of Dysarthric Speech: Implications for Assessment and Treatment Planning. *American Journal of Speech-Language Pathology*, 5(1):55–66, 1996.

4

Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer[0]

Abstract

Purpose: To develop automatic assessment models for assessing the articulation, phonation and accent of speakers with head and neck cancer (Experiment 1) and to investigate whether the models can track changes over time (Experiment 2).

Method: Several speech analysis methods for extracting a compact acoustic feature set that characterizes a speaker's speech are investigated. The effectiveness of a feature set for assessing a variable is assessed by feeding it to a linear regression model and by measuring the mean difference between the outputs of that model for a set of recordings and the corresponding perceptual scores for the assessed variable (Experiment 1). The models are trained and tested on recordings of 55 speakers treated non-surgically for advanced oral cavity, pharynx and larynx cancer. The perceptual scores are average unscaled ratings of a group of 13 raters. The ability of the models to track changes in perceptual scores over time is also investigated (Experiment 2).

[0] R.P. Clapham, C. Middag, F.J.M. Hilgers, JP. Martens, M.W.M. van den Brekel, R.J.J.H. van Son. *Speech Communication, 59, April 2014, 44-54.*

Results: Experiment 1 has demonstrated that combinations of feature sets generally result in better models, that the best articulation model outperforms the average human rater's performance and that the best accent and phonation models are deemed competitive. Scatter plots of computed and observed scores show, however, that especially low perceptual scores are difficult to assess automatically. Experiment 2 has shown that the articulation and phonation models show only variable success in tracking trends over time and for only one of the time pairs are they deemed compete with the average human rater (Experiment 2). Nevertheless, there is a significant level of agreement between computed and observed trends when considering only a coarse classification of the trend into three classes: clearly positive, clearly negative and minor differences.

Conclusions: A baseline tool to support the multi-dimensional evaluation of speakers treated non-surgically for advanced head and neck cancer now exists. More work is required to further improve the models, particularly with respect to their ability to assess low-quality speech.

4.1 Introduction

Cancer of the head and neck and its treatment can have negative consequences on the structures and tissues involved in swallowing and speech and voice production. For the speech-language pathologist, evaluating a patient's speech and voice is an important part of patient management and is necessary for documenting a patient's long-term outcome (Verdonck-de Leeuw et al., 2007). The design and validation of automatic tools to perform "perceptual-like" evaluations has become an area of interest for researchers and recently, interesting results for speech intelligibility (Maier et al., 2009; Middag et al., 2009, 2011, 2014; Van Nuffelen et al., 2009) and phonation (De Bruijn et al., 2009, 2011a; Maryn et al., 2010) have been reported in the literature.

In this study, we investigate whether a machine can reliably evaluate articulation (perception of the precision of speech production), phonation (perception of phonation quality) and accent (perception of degree of accent) (see section 4.2.1.3 for details). If these models were to be combined with an existing model of functional speech intelligibility (Middag et al., 2014), one would have a powerful automatic tool for the multidimensional evaluation of a speaker. For modelling, we include the variables articulation and phonation because they can both be impaired as a result of tumor, cancer treatment such as concomitant chemoradiotherapy (CCRT) or a combination of both tumor and treatment (Jacobi et al., 2010, 2013; Newman et al., 2001; van der Molen et al., 2012). We also include accent because in the Netherlands there is considerable articulatory-acoustic variation as a result of regional variation and social background (Jacobi, 2009) and because of language background in the case of non-native Dutch speakers. Unlike articulation and phonation, accent is not a clinically relevant aspect but there is a risk that an automatic analysis technique will be influenced by the gravity of the accent. By modeling accent, we envisage that clinicians can take the computed accent score into account when interpreting computed scores of speech intelligibility and articulation. In other words, if accent is strongly present caution may be warranted when drawing conclusions on a speaker's computed scores, which may be underestimated.

The aim of this study is to develop assessment models for the perceptual variables articulation, accent and phonation and to compare the assessments of best models with human ratings (Experiment 1). We also investigate whether articulation and phonation assessment models can track trends over time in the human ratings of a single speaker (Experiment 2).

4.2 General method

4.2.1 Validation corpus

All audio recordings are taken from a corpus developed by the Netherlands Cancer Institute (termed the NKI-CCRT corpus). These recordings were collected as part of a preventative rehabilitation study on speech, voice and swallowing outcomes for patients after treatment for advanced head and neck cancer (van der Molen et al., 2012). The perceptual evaluations emerge from a larger study investigating the use of automatic tools to evaluate perceptual aspects of speech production for speakers treated for head and neck cancer. Below we provide an overview of the speakers, stimuli and perceptual data and refer the reader to van der Molen et al. (2012) and Clapham et al. (2012) for more information.

4.2.1.1 Speakers

The corpus contains recordings of 55 speakers who received CCRT over a period of seven weeks for stage III-IV head and neck tumors. Tumors were located in the oral cavity, nasopharynx, oropharynx, hypopharynx or larynx, and recordings were made before treatment (T0) (54 speakers), 10-weeks post-treatment (T1) (48 speakers) and 12-months post-treatment (T3) (39 speakers). The main reason for loss of speakers at follow-up was due to morbidity and mortality (van der Molen et al., 2012). Due to an administrative miss, the T0 recording of one speaker was not included. Average speaker age at T0 was 57 years (range 32-79 years) and approximately 15% of the speakers were non-native Dutch speakers (Middag et al., 2014).

4.2.1.2 Stimuli

All speakers read the same 189-word Dutch text of neutral content. Note that not all speakers contributed recordings at follow-up. The corpus contains only the first 138 words of each recording: the first 70 words are referred to as fragment A and the next 68 words are referred to as fragment B. Fragment A contains 49 unique words and fragment B contains 50 unique words. The corpus contains 141 fragment A recordings and 140 fragment B recordings (one speaker only read fragment A).

4.2.1.3 Perceptual analysis

Thirteen recently graduated or about to graduate speech-language pathologists evaluated all recordings (stimuli) in an online, self-paced experiment. All listeners were female, native Dutch speakers (average age 23.7 years). They could replay a recording as often as they wished and no stimuli anchors were provided. All recordings were presented in a randomized order and the first 10 stimuli reappeared in the final stimuli and were used to check the intra-rater consistency. They were not included in the corpus for the development of assessment models. Although listeners rated several aspects of speech and voice, the variables of interest in this paper are articulation, phonation and accent.

Articulation Listeners were instructed to evaluate the general precision of vowel and consonant production as compared to normal running speech on a 5-point scale with descriptors at 1 (*extremely imprecise articulation*) and 5 (*normal/precise articulation*). Precise articulation was defined as correct manner and place of production and clear coordination between sounds.

Phonation Listeners were instructed to evaluate the degree to which phonation deviated from what they considered normal. Listeners rated phonation on a 5-point scale with descriptors at 1 (*very deviant phonation*) and 5 (*normal phonation*).

Accent Listeners were asked to evaluate the weight of the speaker's dialect or accent as compared to standard Dutch (defined as the speech commonly heard on radio and television). Listeners evaluated their perception of accent on a 5-point scale with descriptors at 1 (*heavy accent*) and 5 (*normal/no accent*).

Agreement and reliability In terms of intra-rater agreement on the 10 test-retest items, the average percentage exact agreement is 87% for articulation (range 60-100%), 55% for phonation (range 40-80%) and 53% for accent (range 30-80%). The average percentage close agreement (differences not larger than 1 point on the 5-point scale) for all variables was above 88% (articulation 95%, accent 94% and phonation 89%).

For this study, we also calculated the inter-rater reliability of the human ratings as the average root mean square error (RMSE) and Pearson Correlation Coefficient (PCC) between the ratings of one individual rater and the group mean scores (see Table 4.1). Although inter-rater reliability is low in some cases, the most deviant rater differs per variable; as such, there is no single rater we can exclude as being unreliable for all perceptual variables. We thus continue to use the average unscaled ratings of the group (means over all 13 perceptual scores) as our reference scores and the average RMSE and PCC as measures of human performance.

Variable	RMSE	PCC
Articulation	0.54 (0.36-0.76) [R04]	0.75 (0.56-0.84) [R11]
Accent	0.57 (0.43-0.91) [R06]	0.78 (0.65-0.89) [R02]
phonation	0.56 (0.36-0.79) [R06]	0.66 (0.47-0.78) [R07]

Note. Values represent the average RMSE and PCC between one individual rater and the group mean.

Table 4.1: Average (and range) of rater reliability data and subject codes of the raters with the highest RMSE and lowest PCC values.

4.2.2 Automatic evaluation tools

Regardless of the variable being modeled, assessment of that variable involves three stages of processing: (1) acoustic front-end analysis of the speech signal, (2) extraction of speaker features and (3) conversion of speaker features to computed scores by means of a so-called assessment model.

During stage one, the acoustic front-end processes Hamming windowed-segments of 25 ms, called frames, and subsequent frames are shifted over 10 ms. Per time step t it extracts an acoustic parameter vector X_t of length 39 describing the total energy and the shape of the spectrum of the segment (for more details see Middag et al. 2014).

During the second stage, all the vectors X_t of a speaker's recording are analyzed and this analysis generates a number of global features that characterize the speaker and that are therefore called speaker features. This paper investigates the power of five speaker feature sets which are described in detail below. Two of the speaker feature sets are derived from speech-to-text alignment by means of an automatic speech recognizer (ASR). Two other speaker feature sets emerge from a plain analysis of the temporal evolutions of the frame-wise outputs of a phonological feature extractor (no speech-to-text alignment involved).

These four feature sets were previously found successful for assessing both single word and running speech intelligibility (Middag et al., 2009, 2011). Given the relationship between speech intelligibility and phonation quality (De Bodt et al., 2002; Moerman et al., 2006), we anticipate that these feature sets will also be effective for assessing articulation and accent. As one of our aims is to assess phonation, we also include a set of pitch and voicing related features that are known to correlate with that variable (Moerman et al., 2004).

During the final stage, speaker features are converted into computed scores by means of an assessment model. In this paper, we employ linear regression models. The training of such models involves an automatic selection of a compact subset of the speaker features (to ensure generalization to unseen data) and a computation of the regression coefficients. The model development approach is described in more detail below.

4.2.2.1 Speaker features emerging from speech-to-text alignment

Two speaker feature sets are derived after a forced alignment. To this end, an ASR matches the acoustic models corresponding to the phonetic transcription of the target text with the speech uttered by the patient. Intuitively, the features emerging from this alignment represent how well, according to the ASR, the target text is realized by the speaker (i.e. how closely the speech fits the models of the ASR given the transcription).

Phonological features (PLFs) PLFs reflect how well binary phonological properties related to manner of articulation (e.g. BURST), place of articulation (e.g. BILABIAL) and voicing (e.g. VOICED) are present or absent at the right times. The extraction of these features involves an ASR encompassing a neural network that computes posterior probabilities of 24 binary phonological properties for each speech frame. To establish (during alignment) how likely a frame is part of a certain phone (defined as a basic speech sound) the ASR computes the geometric mean of the posterior probabilities of the different phonological properties that are assumed to be true for that phone (Stouten and Martens, 2006). The neural network is trained on read speech from the Spoken Dutch Corpus (Schuurman et al., 2003).

Once all frames are assigned to a phone, the PLFs are computed as follows: (1) consider all frames assigned to a phone having a canonical value A (either 1 or 0) for a certain phonological property and (2) compute the mean of the corresponding neural network output over all these frames (see Stouten and Martens, 2006). By repeating this for all 24 phonological properties and for two sets of phones, those with an A = 1 and those with an A = 0, one obtains 48 phonological features to characterize the speech of the investigated speaker. Positive PLFs (derived using A = 1) reflect the presence of a phonological property in the acoustic signal at times the phonological property is supposed to be present (e.g. the presence of voicing during utterance of an /a/). Negative PLFs (derived using A = 0) reflect the presence of a phonological property at times it is not supposed to be present (e.g. the presence of voicing while uttering a /p/). Together, the 48 features are expected to reveal how well the speaker has performed the various articulatory actions involved in the speech production.

Monophone features (MPFs) MPFs reflect how well context-independent phones (also called monophones) such as /s/, /z/ and /A/ are realized. The creation of these features involves an ASR that internally works with approximately 1600 context-dependent phone states (also called triphone states). These states are trained by means of expectation maximization on speech material from the Spoken Dutch Corpus (Schuurman et al., 2003). Once each frame is assigned to a triphone state, the MPFs can be computed as follows: (1) consider all frames assigned to a state of a certain phone and (2) compute the mean posterior probability of this phone over all these frames. Repeating this for all 40 phones of the language results in a set of 40 monophone features that characterize the speech of the investigated speaker.

4.2.2.2 Speaker features not emerging from a speech-to-text alignment

We also discern two speaker feature sets that emerge from a speech analysis that does not require knowledge of the input text.

Alignment-free phonological features (ALF.PLFs) ALF.PLFs follow from a plain analysis of the temporal evolutions of the individual outputs of a neural network that computes posterior probabilities of binary (on/off) and ternary (on/off/intermediate) phonological properties (for details see Middag et al., 2010). Per output, the temporal analysis determines characteristics such as mean, standard deviation, percentage of time high, intermediate and low, mean of the peaks (maxima) and mean time needed to make a transition from low to high. In total, this analysis yields 300 features. The hypothesis is that temporal fluctuations in the network outputs can reveal articulatory deficiencies, regardless of the exact phonetic content of the text that was read, at least as long as this text is sufficiently rich in phonetic content.

Alignment-free phonetic features (ALF.MPFs) ALF.MPFs follow from a plain analysis of posterior phone probabilities that can be retrieved from the outputs of the neural network that gave rise to the ALF.PLFs. The conversion of network outputs to posterior phone probabilities (scores) is achieved in the same way as before (creation of MPFs).

Per phone one considers all frames and one computes (1) the mean and (2) the standard deviation of the corresponding phone score, as well as (3) the mean of the peaks (maxima) and (4) the valleys (minima) found in the temporal evolution of that score. In addition, one also computes (5) the fraction of the time the phone is the winner (gets the maximal score) and (6) the mean score of the considered phone over all frames. The latter two quantities are divided by the expectations one can derive from the expected phone frequencies and the average phone durations. This way, one obtains 6 features per phone. Repeating this for all 55 phones (40 monophones, 6 closures and 6 burst sounds for modeling the plosives, a glottis and two silence symbols to accommodate inter and intra-sentence pauses), we obtain 330 (= 6 * 55) ALF.MPFs in total.

Pitch and voicing related features (AMPEX) Eight pitch and voicing related parameters (the percentage of speech frames being classified as voiced, the jitter of the pitch in voiced frames, etc.) are extracted from the frame level outputs of the AMPEX pitch and voicing detector proposed by Van Immerseel and Martens (1996). These features have already been employed with success for pathological speech assessment (Moerman et al., 2004) and were therefore anticipated to be suitable for the envisaged phonation assessment. The program to extract the features is freely available[1].

4.2.3 Performance criterion

Model performance is characterized by the root mean square error (RMSE) and the Pearson correlation coefficient (PCC) between computed scores and average unscaled ratings of the group (means over 13 individual human ratings are considered as 'ground truth'). The aim is to pursue low RMSE and high PCC values. Human performance is defined as the RMSE and PCC between the ratings of an individual and the same ground truth.

4.3 Experiment 1

4.3.1 Objective

The first objective of this study is to develop models for computing "perceptual-like" scores of articulation, accent and phonation. We discern single-feature assessment models that use only one features set (either PLF, MPF, ALF.PLF, ALF.MPF or AMPEX) and multiple-feature models that use a combination of two or three feature sets (PLF+MPF, PLF+ALF.PLF, MPF+ALF.MPF and

[1] http://dssp.elis.ugent.be/downloads-software

PLF+ALF.PLF+AMPEX). Model performance is considered with respect to (i) other models and (ii) human performance.

4.3.2 Method

The assessment models are trained and evaluated using a 5-fold cross validation strategy. Each fold is used once as a validation set for a model that is developed on the remaining folds as the training set. The recordings of a particular speaker always belong to the same fold. During training, we create ten random divisions of the training set into two equally large parts: one part for selecting the features to use and the other part for estimating the regression coefficients for these features. The ten simple models emerging from this step are then merged into one model which utilizes all features that were selected by the ten simple models with regression coefficients that are equal to the averages of the coefficients of the simple models (0 is assumed for a model not selecting a coefficient). This approach is called Ensemble Linear Regression (Breiman, 1996) and it allows us to inspect how often the simple models select a particular feature (between 0 and 10 times per fold). We refer the reader to Middag et al. (2014) for further details.

When comparing models, we consider the model with the lowest RMSE as the best model. If two models have the same RMSE, we prefer the model with the largest PCC. We use the Wilcoxon signed rank test ($p < .01$, a conservative p-value to account for multiple comparisons) to measure the statistical significance between the performances of the best model and runner-up models.

When comparing model performance to human performance, we consider the model better if both the RMSE is lower and the PCC is higher than that of the average human rater (the latter data are presented in Table 4.1). We consider the model competitive if either its RMSE is lower or its PCC is higher than that of the average human rater.

4.3.3 Results

4.3.3.1 Articulation assessment models

Performances for single-feature and multiple-feature articulation models are listed on the left-hand side of Table 4.2. Four single-feature models are competitive with human raters, but the AMPEX model is unable to reach that level most likely because it does not encompass features related to place and manner of articulation. Although the best multiple-feature model is not significantly better than the best single-feature PLF model, multiple-feature models do consistently yield lower RMSE and higher PCC values than single-feature models. The table also reveals that there is no clear difference between monophone and phonological features.

The best model is MPF+ALF.MPF and it surpasses human performance. Figure 4.1a displays a scatter plot of the computed scores (model) and mean unscaled perceptual scores (human) for that model. Table 4.3 lists the features that were selected more than 10 times (out of 50, coming from 5 folds and maximum 10 times per fold) during ensemble linear regression. The most selected features are six consonant features related to the production of /s/, /n/, /l/, /d/ and five vowel features related to /i/ ('liep'), /A/ ('lat'), /u/ ('voer'), /@/ ('de') and /Au/ ('koud').

4.3.3.2 Accent assessment models

Performances for single-feature and multiple-feature accent models are listed in the middle section of Table 4.2. Only one single-feature model, namely PLF, is competitive with a human rater. Again, AMPEX is the weakest model. All multiple-feature models are competitive with a human rater and there is little difference between models. MPF+ALF.MPF is the best model. Figure 4.1b displays a scatter plot of the computed scores (model) and mean unscaled perceptual scores (human) for this model. The most selected features for this best model are three vowel features related to /9y/ ('huis'), /A/ ('lat') and /y/ ('buut') and two consonant features related to /n/ and /z/.

4.3.3.3 Phonation assessment models

Performances for single-feature and multiple-feature models are listed in the right-hand side of Table 4.2. Only two single-feature models (PLF and AMPEX) can be deemed competitive with a human rater. The AMPEX model now has a slight advantage over the phonological and monophone models. Combining phonological features and monophone features only results in small improvements, but adding AMPEX as a third feature set leads to a significant improvement and yields a model that is competitive with the average human rater. Figure 4.1c displays a scatter plot of the computed scores (model) and mean unscaled perceptual scores (human). The plot shows that this best model fails when the perceptual score is low.

The most frequently selected features are related to the presence of voicing (average voicing evidence in voiced frames [AVE], percentage of voiced speech frames [VSS], percentage of voiced frames [PVF] and mean time required to switch from unvoiced to voiced [unvoiced-voiced]). Other frequently selected features are related to the rate of movement from back to front, the inclusion of the nasal cavity (nasality, vowel nasality) and trill (Dutch /r/ is variable and the trill variant can be produced as an alveolar or uvular trill (Rietveld and Van Heuven, 1997). Difficulty producing uvular trills may indicate insufficient control of the velum).

Feature set(s) in model	Articulation RMSE (PCC)	Accent RMSE (PCC)	Phonation RMSE (PCC)
Single-feature models			
PLF	0.44 (0.75)	0.56 (0.72)	0.55 (0.39) *
MPF	0.45 (0.74) *	0.59 (0.67) *	0.59 (0.24) *
ALF.PLF	0.51 (0.66) *	0.68 (0.55) *	0.58 (0.33) *
ALF.MPF	0.45 (0.75) *	0.65 (0.64) *	0.60 (0.23) *
AMPEX	0.66 (0.24) *	0.82 (0.30) *	0.55 (0.43) *
Multiple-feature models			
PLF+MPF	0.44 (0.75)	0.56 (0.71) *	0.54 (0.42) *
PLF+ALF.PLF	0.44 (0.78)	0.55 (0.74)	0.53 (0.47) *
MPF+ALF.MPF	**0.42 (0.80)**	**0.54 (0.77)**	0.58 (0.27) *
PLF+ALF.PLF+AMPEX	0.44 (0.78)	0.56 (0.71)	**0.46 (0.62)**
Human performance	0.54 (0.75)	0.57 (0.78)	0.56 (0.66)

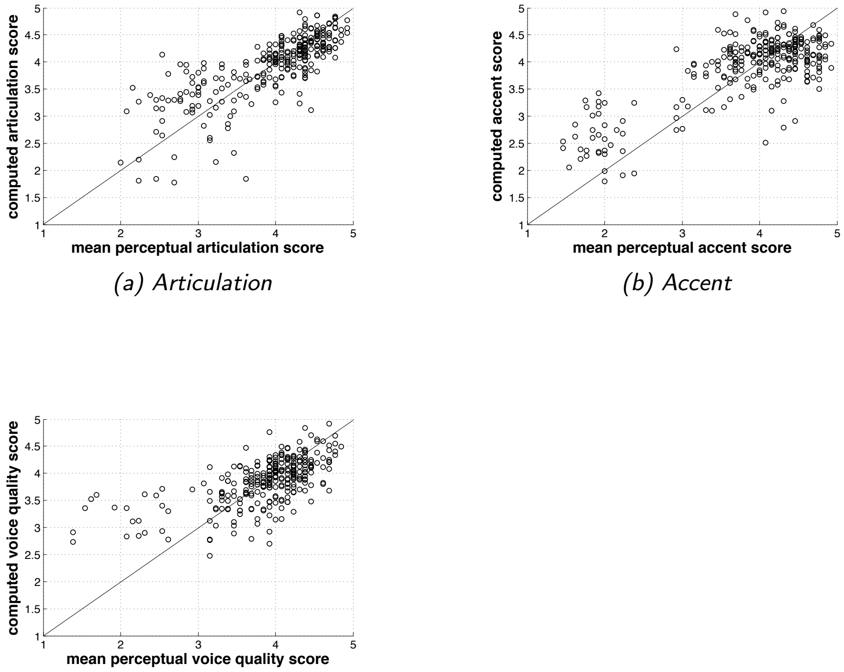
Note. In each column, the best performing model is highlighted in bold and compared to the other models for that variable (* = significantly different at $p < .01$). Double underlining is used to highlight models outperforming the average human rater and single underlining to highlight models with competitive performance. Human performance is provide to aid comparison.

Table 4.2: Performances of assessment models and target performance (human performance) for the three variables

4.3.4 Discussion

The best articulation model (RMSE 0.42; PCC 0.80) outperforms the average human rater's performance (RMSE 0.54; PCC 0.77) and the best accent and phonation models are competitive. The scatter plots of computed versus mean perceptual scores for all best models confirm the finding by Van Nuffelen et al. (2009) that low perceptual scores are difficult to predict. In both Van Nuffelen et al. and our study, this may be due to the low prevalence of speakers with low perceptual scores in the corpus, and in the case of phonation, to the failures of the pitch and voicing detector (Manfredi et al., 2011). Ideally, a corpus for model training would have a balanced distribution of to-be-modelled data but when data comes from clinical populations this is often not the case. Developing separate models for lower and for higher perceptual scores or weighting the perceptual scores may improve prediction accuracy.

For the articulation and accent models, combining feature sets generally improves performance, but this improvement is not significant in a statistical



(a) Articulation

(b) Accent

(c) Voice quality

Figure 4.1: Scatter plots of the computed scores (model) and mean unscaled perceptual scores (human) for the variables (a) articulation (model MPF+ALF.MPF), (b) accent (model MPF+ALF.MPF) and (c) phonation (model PLF+ALF.PLF+AMPEX). The line displays the ideal relationship between the computed and perceptual scores.

Articulation		Accent		Phonation	
MPF+ALF.MPF		MPF+ALF.MPF		PLF+ALF.PLF+AMPEX	
/i/	(44)	/9y/	(41)	AVE$_{AMPEX}$	(35)
/s/	(36)	/A/	(38)	PVS$_{AMPEX}$	(28)
/n/[a]	(28)	/n/[a]	(25)	PVF	(22)
/l/	(25)	/y/[a]	(19)	vowel	(22)
/l/[a]	(23)	/z/[c]	(15)	unvoiced-voiced[d]	(19)
/A/	(22)			nasality[e]	(12)
/u/	(16)			front-back[f]	(12)
/s/[a]	(16)			trill	(11)
/d/	(15)			vowel nasality[g]	(11)
/@/	(15)				
/Au/[b]	(11)				

Note. Monophone features are given in SAMPA notation. AVE = average voicing evidence in voiced frames. PVS = % of voiced speech frames. PVF = % of voiced frames.

[a] percentage of frames x was recognized. [b] mean evidence of feature x over all frames in which x was recognized. [c] SD of probability of /z/. [d] mean time needed to go from unvoiced voiced. [e] mean minimum value for relevance of consonant nasality. [f] mean time needed to go from relevant to not relevant in front-back aspect. [g] SD of vowel nasality probability in frames in which vowel nasality is present.

Table 4.3: Speaker features selected more than 10 times during the cross-validation process of model development for the best articulation, accent and phonation models. The number of times selected is displayed between parentheses. See Section 4.3.2 for details.

sense. Three multiple-feature articulation models outperform the average human rater, whereas all multiple-feature accent models are competitive. The strong performance of the multiple-feature set model which includes the AMPEX features is not surprising as the AMPEX features were designed for the assessment of overall phonation (Moerman et al., 2004). The fact that combining AMPEX with PLF and ALF.PLF leads to higher performance suggests that the feature sets represent partly complementary views on phonation.

It is difficult to compare our articulation and phonation assessment results with other studies as we are unaware of any studies focusing on automatic evaluation of articulation for speakers treated *non-surgically* for cancer of the head and neck. In De Bruijn et al. (2011b) an artificial neural network is used to generate acoustic features related to plosive production by speakers treated with surgery and radiotherapy for oral and oropharyngeal cancer. However, the correlation coefficients between these features and the perceptual scores of articulation were below 0.40. The higher proportion of speakers with low perceptual scores in De Bruijn et al. may be partly responsible for this low correlation coefficient. Note that the surgical procedures within the oral cavity (present in the data of De Bruijn et al.) may have a larger effect on articulation production than non-surgical treatment (present in our data). Haderlein et al. (2007) reported PCCs above 0.70 between computed scores and perceptual scores of phonation made on a 5-point scale for a small group of tracheoesophageal speakers. Maryn et al. (2010) reported a correlation coefficient of 0.796 between the Acoustic phonation Index and a mean perceptual score of the overall grade of dysphonia made on a 4-point scale for a group of 33 dysphonic and 6 control speakers.

Features selected by the articulation assessment models overlap with those selected by a speech intelligibility assessment model in Middag et al. (2014). The overlap includes vowels from the diagonal of the vowel trapezium (/i/, /@/, /A/) and features related to nasality. This is not surprising given the high correlation between articulation and speech intelligibility scores. One of the main differences between articulation and speech intelligibility models is that the former select more consonant-related than vowel-related features whereas this pattern is reversed in the latter. As a whole, the features selected in the articulation models appear to be related to tongue movement in the diagonal of the vowel trapezium and to production of anterior lingual consonants. One explanation as to their inclusion is that the studied speakers have difficulty with producing anterior lingual consonants (Newman et al., 2001) and that realizations deviate from acoustic models based on healthy speakers. An alternative explanation is that the model selects features with the potential to discriminate between scores of speakers with tumors in different locations. A recent study by Jacobi et al. (2013) on our speakers showed significant differences between speakers' acoustic measures related to vowel space, production of stops and fricatives and nasality that can be owed to tumor location.

Unlike the articulation assessment model, the best accent assessment model does not focus on differentiating vowels in the vowel trapezium, but rather suggests that differentiating high and low central vowels is important for predicting perceptual accent scores. Selection of the feature related to the alveolar fricative /z/ may reflect regional voicing variation (Kissine et al., 2003).

Features selected by the phonation model may be understood in terms of the source-filter model of speech production, that is, the effect of tumor/treatment at the level of the vocal-source (i.e. effect on phonation) and at the level of the vocal-tract filter (i.e. effect on resonance). We would expect that at the level of the larynx, a tumor and its treatment would have a greater effect on phonation whereas a tumor in the nasopharynx or oropharynx area is likely to have an effect on the speaker's ability to use the nasal cavity as a filter. The features selected in the assessment models indicate that the human rater of phonation takes both phonation and resonance information into account.

4.4 Experiment 2

4.4.1 Objective

The second objective of this study is to investigate whether the best articulation model (MPF+ALF.MPF) and phonation model (PLF+ALF.PLF+AMPEX) can track changes over time in the perceptual scores of an individual patient. We exclude the variable accent from this experiment, as it is not a clinically relevant variable.

4.4.2 Method

We describe trends by means of differences between perceptual scores on two evaluation moments. Given the preliminary nature of this study, we only consider patients for which all raters agreed on the direction of change (either positive/no change or negative/no change) between a given time-pair. This way, we retain 57 score differences for articulation (19 for T0-T1; 17 for T1-T3; 21 for T0-T3) and 61 for phonation (27 for T0-T1; 18 for T1-T3; 16 for T0-T3). Note that the imbalance in the data pairs is a result of morbidity and mortality and our exclusion criterion.

We investigate the model and rater capacities to perform a three-fold trend classification: clearly negative (score difference $\leq$ -0.5), minor differences, and clearly positive (score difference $\geq$ 0.5). Our choice of class boundaries leads to comparable sizes for the three trend categories. Agreement between computed and mean perceptual score differences is calculated using the kappa statistic (linear weighting). We use a resampling method without replacement to assess whether the observed kappa value is due to chance or not.

As in Experiment 1, we characterize the assessment of changes by means of the RMSE and PCC between the computed score differences (differences between two evaluation moments) and the corresponding mean perceptual score differences. The same criteria as before are used for declaring a model better (i.e. lower RMSE and higher PCC) or competitive (i.e. either the RMSE is lower or the PCC is higher) with respect to the average human rater.

4.4.3 Results

The upper-half of Table 4.4 lists the performance for the MPF+ALF.MPF articulation model and the PLF+ALF.PLF+AMPEX phonation model for tracking changes between two evaluation moments. Both the articulation and phonation model outperform the human rater for T0-T3, but not for the other time-pairs. For T0-T1, the phonation model is just competitive.

Eval.	Articulation			Phonation		
	n	Rater	Model	n	Rater	Model
		RMSE (PCC)	RMSE (PCC)		RMSE (PCC)	RMSE (PCC)
T0-T1	19	0.58 (0.69)	0.58 (0.22)	27	0.73 (0.77)	0.60 (0.76)
T1-T3	17	0.62 (0.74)	0.69 (0.04)	18	0.60 (0.75)	0.61 (0.45)
T0-T3	21	0.57 (0.70)	0.40 (0.72)	16	0.64 (0.79)	0.45 (0.86)

Note. Performance for computing changes between two evaluation moments is for cases where raters agree on the direction of trend. Double underlined values highlight comparisons outperforming the average rater and single underlined values highlight models with competitive performance. T0 = pre-CCRT. T1 = 10-weeks post CCRT. T3 = 12-months post CCRT. n is the number of recordings included in the comparison.

Table 4.4: Human rater and model performance for tracking changes in articulation and phonation over time

Table 4.5 shows that the accuracy of an articulation trend classification (positive, minor, negative change) is 72% (41 out of 57 cases), corresponding to a significant degree of agreement between computed and mean perceptual trends ($\kappa = 0.37$, $p < .001$). The accuracy of a phonation trend classification is 64% (39 out of 61 cases), corresponding to a significant degree of agreement between computed and mean perceptual trends ($\kappa = 0.45$, $p < .001$). The different kappa values for the two variables likely reflect properties of the data distribution, that is, clearer trends for phonation. As can be seen in Table 4.5, all disagreements are between adjacent classes. The models are clearly biased towards deciding that there is no clear trend.

(a) Articulation

Computed	Observed		
	-	±	+
-	3	0	0
±	7	35	8
+	0	1	3

(b) Phonation

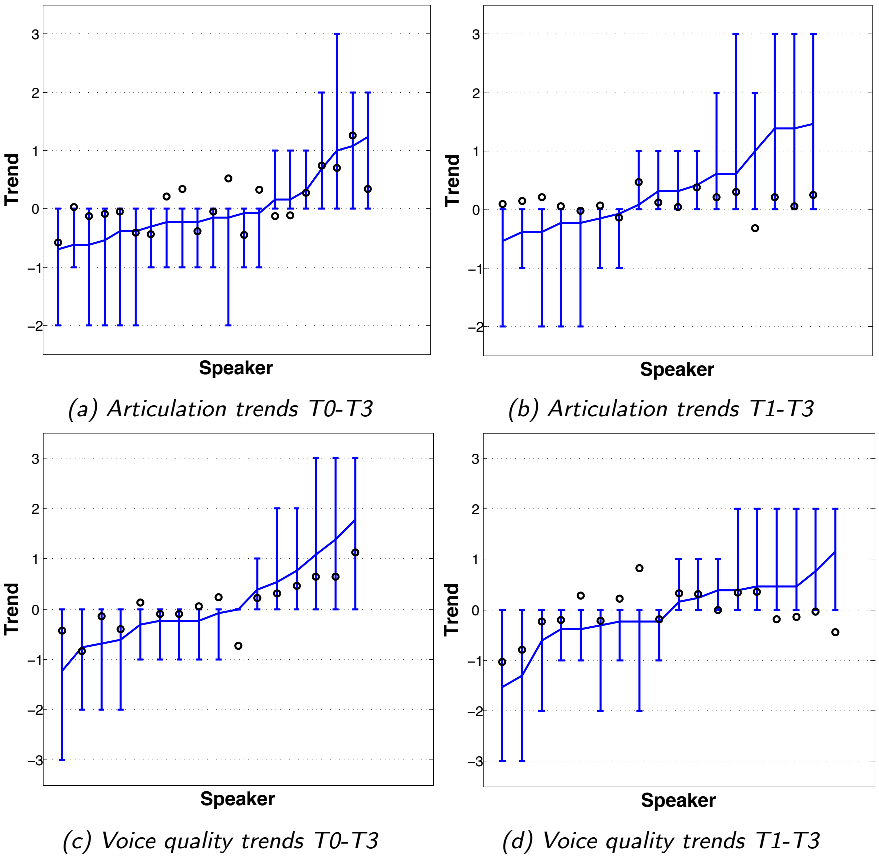
Computed	Observed		
	-	±	+
-	7	3	0
±	14	25	4
+	0	1	7

Table 4.5: Contingency table between computed and observed trends for clearly negative change (-), minor change (±) or clearly positive change (+).

4.4.4 Discussion

Although the RMSE values are relatively high for both articulation and phonation changes, inspection of the trends indicates that most computed trends

Computed	Observed		
	-	±	+
-	7	3	0
±	14	25	4
+	0	1	7



(a) Articulation trends T0-T3 *(b) Articulation trends T1-T3*

(c) Voice quality trends T0-T3 *(d) Voice quality trends T1-T3*

Figure 4.2: Difference in scores between two evaluation moments for articulation and phonation. The continuous line displays the mean difference in perceptual scores (vertical lines indicate the range) and the circles display the difference in computed scores.

fall within the range of perceptual trends and can thus be considered acceptable (see plots in Figure 4.2 which correspond to the best and worst time pair respectively).

The plots also show that the observed trends are often close to zero and that human raters only indicate a clear difference between evaluation moments

($\pm$ 1) in relatively few cases. This means that the PCC between computed and human scores will mainly be determined by how well the assessment model evaluates these few cases.

The trends in pairs involving T1 are badly predicted. Our first hypothesis was that this was because the computed score at T1 differed more from the mean perceptual scores than at T0 and T3. This did not turn out to be the case. Another hypothesis was that there is a larger percentage of low perceptual scores ($\leq$ 3, the point where the model appears to fail) from which to derive the trends in the time pairs including T1. For articulation, these percentages are 7% for T0-T3, 11% for T0-T1 and 12% for T1-T3. For phonation, these percentages are 6% for T0-T3, 15% for T0-T1 and 9% for T1-T3. Although these figures do not confirm this hypothesis, they do not contradict it either.

Acoustic analysis of the speakers from this corpus revealed that there are acoustic differences in vowel and consonant production between evaluation moments (Jacobi et al., 2013), however, these changes may not be perceptually salient and/or the 5-point scale used for collection of perceptual scores may not allow fine-grained differentiation. Transforming the perceptual scores as proposed by Shrivastav et al. (2005) may reduce the inter-rater variation due to differences in listener anchor points on the scale and, thereby, improve the reliability of the ground truth that is used to train the models. An alternative experimental set-up involving paired-comparison evaluations, such as that used in van der Molen et al. (2012), was not used in perceptual data collection as this method allows neither comparison of the speaker to a reference healthy speaker nor to other speakers from the same population.

4.5 Summary and concluding discussion

The assessment models presented in this paper have been developed and tested on a group of Dutch speakers with cancer of the head and neck who contributed speech recordings before and after non-surgical treatment. The aims of this study were (i) to investigate whether perceptual scores of articulation, accent and phonation can be automatically assessed with an accuracy comparable with that of a human rater and (ii) to investigate whether articulation and phonation assessment models can reveal trends in the articulation and phonation over time.

We have shown that speaker features emerging from a forced alignment between the speech and the text as well as speaker features emerging from a plain analysis of the temporal evaluation of acoustic model outputs give rise to good assessment models but that combining feature sets generally leads to improved model performance. The correlations between computed and perceptual scores are often within the range of human performance and, in the case of articulation, the model outperforms the average human rater's performance. Plots of computed and observed perceptual scores do show however that models have a tendency to produce over-optimistic scores for bad speakers.

Despite the overall positive performance of the models in computing perceptual scores, the articulation and phonation assessment models attain varying levels of success in tracking changes over time. Nevertheless, categorization of the trends in three classes (clearly positive, minor, clearly negative) can be achieved at human performance level.

There seems to be some evidence that a part of the problem resides from the fact that human scores assigned to low-quality speech also tend to be unreliable, and consequently, this unreliability is bound to transfer to the models derived thereof. We envisage that future work focuses on removing some of the inter-rater variability by normalizing perceptual scores, by developing rater-specific assessment models or by combining the two approaches (i.e. rater-specific models based on normalized perceptual scores).

Acknowledgements

Part of this research was funded by unrestricted research grants from Atos Medical (Horby, Sweden) and the Verwelius Foundation (Naarden, the Netherlands). All speech recordings were collected by Lisette van der Molen, SLP PhD, and Irene Jacobi, PhD. The authors wish to thank Maya van Rossum, SLP PhD for her input on collecting the perceptual scores. Prof. Louis Pols is greatly acknowledged for his critical and constructive review of the manuscript.

References

L Breiman. Bagging predictors. *Machine Learning*, 24:23–140, 1996.

R.P. Clapham, L. Van Der Molen, R.J.JH. Van Son, M. Van Den Brekel, and F.J.M. Hilgers. NKI-CCRT Corpus - speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources* *and Evaluation*, pages 3350–3355, Istanbul, Turkey, May 2012. European Language Resources Association. ISBN 978-2-9517408-7-7.

Marc S De Bodt, Huici Maria E Hernández-Díaz, and Paul H Van De Heyning. Intelligibility as a linear combination of dimensions in dysarthric speech. *J Commun Disord*, 35(3):283–92, 2002.

Marieke J De Bruijn, Louis Ten Bosch, Dirk J Kuik, Hugo Quené, Johannes A Langendijk, C René Leemans, and Irma M Verdonck-de Leeuw. Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. *Folia Phoniatr Logop*, 61(3):180–7, 2009. doi: 10.1159/000219953.

Marieke J De Bruijn, Louis ten Bosch, Dirk J Kuik, Johannes A Langendijk, C René Leemans, and Irma M Verdonck-de Leeuw. Artificial neural network analysis to assess hypernasality in patients treated for oral or oropharyngeal cancer. *Logopedics Phoniatrics Vocology*, 36:168–174, 2011a.

Marieke J De Bruijn, Louis ten Bosch, Dirk J Kuik, B Witte, Johannes A Langendijk, C René Leemans, and Irma M Verdonck-de Leeuw. 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*, 54(5):632–640, 2011b.

Tino Haderlein, Elmar Nöeth, Hikmet Toy, Anton Batliner, Maria Schuster, Ulrich Eysholdt, Joachim Hornegger, and Frank Rosanowski. Automatic evaluation of prosodic features of tracheoesophageal substitute voice. *European Archives of Oto-Rhino-Laryngology*, 264(11):1315–1321, 2007. doi: DOI10.1007/s00405-007-0363-4.

Irene Jacobi. *On variation and change in dipthongs and long vowels of spoken Dutch*. PhD thesis, University of Amsterdam, 2009.

Irene Jacobi, Lisette Van Der Molen, H Huiskens, M A Van Rossum, and F J M Hilgers. Voice and speech outcomes of chemoradiation for advanced head and neck cancer: a systematic review. *Eur Arch Otorhinolaryngol*, 267:1495–1505, 2010.

Irene Jacobi, Maya A van Rossum, Lisette van der Molen, Frans JM Hilgers, and Michiel WM van den Brekel. Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy. *Annals of Otology, Rhinology & Laryngology*, 122(12): 754–762, 2013.

Mikhail Kissine, Hans Van De Velde, and Roeland Van Hout. An acoustic study of standard dutch /v/, /f/, /z/ and /s/. *Linguistics in the Netherlands*, -(93-104), 2003.

A Maier, T Haderlein, U Eysholdt, F Rosanowski, Anton Batliner, M Schuster, and E Nöth. PEAKS - A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425–437, May 2009.

C Manfredi, A Giordano, S F Schoentgen, L Bocchi, and Philippe Dejonckere. Validity of jitter measures in non-quasi-periodic voices. part ii. the effect of noise. *Logopedics Phoniatrics Vocology*, 36:78–89, 2011.

Youri Maryn, Marc De Bodt, and Nelson Roy. The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders. *Journal of Communication Disorders*, 43(3):161 – 174, 2010. ISSN 0021-9924. doi: DOI:10.1016/j.jcomdis.2009.12.004.

C Middag, J-P Martens, G Van Nuffelen, and Marc S De Bodt. Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal of Advances in Signal Processing*, 2009.

C Middag, Y Saeys, and J-P Martens. Towards an ASR-free objective analysis of pathological speech. In *Proceedings of Interspeech*, pages –. Interspeech, 2010.

C Middag, T Bocklet, J Martens, and E Nöth. Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pages 3005–3008, 2011.

Catherine Middag, Renee Clapham, Rob van Son, and Jean-Pierre Martens. Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. *Computer Speech and Language*, 28(2): 467 – 482, 2014. ISSN 0885-2308. doi: http://dx.doi.org/10.1016/j.csl.2012.10.007. URL http://www.sciencedirect.com/science/article/pii/S0885230812000903.

M Moerman, J-P Martens, M.J van der Borgt, M Peleman, M Gillis, and Philippe Dejonckere. Perceptual evaluation of substitute voices: development and evaluation of the (i)infvo rating scale. *Eur Arch Otorhinolaryngol*, 263: 183–187, 2006.

Mieke Moerman, Glenn Pieters, Jean-Pierre Martens, Marie-Jeanne Van Der Borgt, and Philippe Dejonckere. Objective evaluation of the quality of substitution voices. *European Archives of Oto-Rhino-Laryngology*, 261 (10):541–547, 11 2004.

LA Newman, KT Robbins, J A Logemann, A W Rademaker, CL Lazarus, A Hamner, S Tusant, and CF Huang. Swallowing and speech ability after treatment for head and neck cancer with targeted intraarterial versus intravenous chemoradiation. *Head Neck*, 24:68–77, 2001.

ACM Rietveld and VJ Van Heuven. *Algemene Fonetiek*. Dick Coutinho, Bussem, 1997.

Ineke Schuurman, Machteld Schouppe, Heleen Hoekstra, and Ton Van Der Wouden. Cgn, an annotated corpus of spoken dutch. In *In: Proceedings of 4th International Workshop on Language Resources and Evaluation*, pages 340–347, 2003.

R. Shrivastav, C. Sapienza, and V. Nandur. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res*, 48:323–335, 2005.

F. Stouten and J.-P. Martens. On the use of phonological features for pronunciation scoring. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 329–332, may. 2006. doi: 10.1109/ICASSP.2006.1660024.

Lisette van der Molen, Maya A. van Rossum, Irene Jacobi, Rob J.J.H. van Son, Ludi E. Smeele, Coen R.N. Rasch, and Frans J.M. Hilgers. Pre- and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: Expert listeners' and patient's perception. *Journal of Voice*, 26(5):664.e25–664.e33, 2012. doi: 10.1016/j.jvoice.2011.08.016.

Luc Van Immerseel and Jean-Pierre Martens. Pitch and voiced/unvoiced determination with an auditory model. *J Acoust Soc Am*, 91(6):3511–3526, 1996.

G Van Nuffelen, C Middag, M De Bodt, and J Martens. Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language and Communication Disorders*, 44(5):716 – 730, 2009. ISSN 1368-2822.

Irma M Verdonck-de Leeuw, Rico N Rinkel, and C René Leemans. *Head and neck cancer: treatment, rehabilitation, and outcomes*, chapter Evaluating the impact of cancer of the head and neck, pages 27–56. Plural Publishing, 2007.

Part II

Aspects of evaluating voice and speech after total laryngectomy

5

Automatic tracheoesophageal voice typing using acoustic features[0]

Abstract

The acoustics of isolated vowels, e.g. of /a/, have in many studies been linked to pathological voice types, such as tracheoesophageal (TE) voice. To study the possibilities of objective and automatic classification of pathological TE voice types, the acoustic features of /a/ were quantified and subsequently classified using a suit of machine learning technologies. Best classification was achieved by using a voiced-voiceless measurement and the harmonics-to-noise ratio. Other common acoustic features were correlated to pathological type as well, but were less distinctive in classification. We conclude that for objective and automatic classification of TE voice pathology, voicing distinction and harmonics-to-noise ratio are most relevant.

5.1 Introduction

Cancer of the larynx as well as most treatment modalities have a negative impact on a person's voice and speech quality. In the case of advanced laryngeal cancer, a total laryngectomy is often unavoidable. Although many patients develop functional alaryngeal speech by means of a prosthetic device to direct

[0] R.P. Clapham, C.J. Van As-Brooks, M.W. M. van den Brekel, F.J. M. Hilgers, R.J. J. H. Van Son. *Proceedings of INTERSPEECH 2013, Lyon, France, 2162-2166.*

air towards the neo-glottis, voice quality is variable (Haderlein et al., 2007; Jongmans, 2008; van As, 2001).

Presently, the prospects for the development of an adequate substitute voice due to the use of prosthetic devices are good (Jongmans, 2008; Op de Coul et al., 2005; van As, 2001). Subsequent speech therapy will then aim at further improving voice quality and speech intelligibility. Studies have shown that improvements of speech quality and intelligibility can indeed improve the quality-of-life (QoL) of patients (Jongmans, 2008). To support and evaluate voice quality after total laryngectomy and subsequent speech therapy, efforts have recently been made to introduce objective methods and automatic evaluations of the intelligibility and quality of alaryngeal speech (Haderlein et al., 2007; Moerman et al., 2006).

A three-type classification of voice quality on sustained vowels by Titze (Titze, 1995; Zhang et al., 2008) was adapted by Van As-Brooks (van As, 2001; van As-Brooks et al., 2006) to a four-type classification for tracheoesophageal speech (TES) on sustained /a/, i.e., acoustic signal typing (AST). Both classifications were based on spectrographic information of a sustained vowel. Both classification systems have consistent links to perceptual evaluation of voice quality of these speakers by speech and language therapists (SLTs).

The link between an objective classification system of voice pathology and the auditory perception of voice quality offers an opportunity to link objective and automatic acoustic measurements to the perception of pathology. Many studies have investigated the correlation between individual acoustic measures and TES voice pathology, see Table 5.1 for a short list.

It is clear that many acoustic variables are related to the severity of TES voice pathology. However, it is not clear how these should be weighted and combined to get a better understanding of TES voicing pathology. Ideally, one would like to be able to "predict" the AST class from acoustic parameter measurements alone. Such automatic classifications are the subject of machine learning (Guyon and Elisseeff, 2003; Ladha and Deepa, 2011; Maindonald and Braun, 2006).

The current study is part of an ongoing effort to understand the evaluation of TE speech and the development of diagnostic aids. The question we want to answer here is: To what extent can acoustic features of sustained /a/ contribute to predicting and understanding the severity of voice pathology in TES?

Acoustic measure	Reference	Type	Significant
Percentage voiced	Kazi et al. (2009)	Percept.	+
	Moerman et al. (2004)	Percept	ns
Max. voicing duration	Moerman et al. (2004)	Percept	ns
F_0	van As-Brooks et al. (2006)	AST	ns
	van Gogh et al. (2005)	AST	ns
	Kazi et al. (2009)	Percept	ns
F_0 variability	van As-Brooks et al. (2006)	AST	+
	van Gogh et al. (2005)	AST	+
	Kazi et al. (2009)	Percept	+
Shimmer	Kazi et al. (2009)	Percept	+
Jitter	van As-Brooks et al. (2006)	AST	ns
HNR	Maryn et al. (2009)	Percept.	ns
	Moerman et al. (2004)	Percept	ns
	van As-Brooks et al. (2006)	AST	+
HNR $<$700 Hz	van Gogh et al. (2005)	AST	+
HNR $\geq$700 Hz	van Gogh et al. (2005)	AST	ns
High frequency noise	van Gogh et al. (2005)	AST	ns
GNE	van As-Brooks et al. (2006)	AST	ns
Rahmonic Intensity	Maryn et al. (2009)	Percept	ns
	van Gogh et al. (2005)	AST	ns
BED	van As-Brooks et al. (2006)	AST	+
$D_2^\star$, SampEn*	Yan et al. (2013)	Percept	+

Table 5.1: Overview of acoustic parameters in studies investigating TES vowel voice quality. AST: Acoustic Signal Typing (van As, 2001; van As-Brooks et al., 2006), Percept..: Perceptual evaluation. ns: Not significant, +: Significant (versus normal), ⋆: Not included in this study. See section 5.2.3

5.2 Materials and methods

5.2.1 Speech recordings

We used a corpus containing sustained vowel /a/ of 87 TE speakers. Recordings were made between 1995 and 2009 as part of several unrelated studies. At the time of the recordings all speakers provided informed consent allowing the recordings to be used for research purposes within the institute. In total there were 74 male and 13 female speakers. Age at treatment was 38-85 (median age 57). All speakers produced sustained /a/ vowels as part of a larger assessment battery. As some speakers had provided multiple recordings for various research projects over the 14 years, we selected the /a/ recording with the earliest recording date. At the time of recording, 83 speakers had a Provox1 or Provox2 prosthesis and the remaining four speakers had a Provox Vega prosthesis (three speakers a 22.5 Fr and one speaker a 17 Fr) (Hilgers and Balm, 2007; Hilgers et al., 1997, 2010).

Due to the fact that recordings were made over more than a decade as part of unrelated studies, a range of equipment and media were used for recording and storage, but this is not expected to alter acoustic measures below 5 kHz (van Son, 2005). For this study, all recordings were first digitized and converted to 44.1 kHz sampling rate and 16-bit Signed Integer PCM encoding. No audio compression had been used on the recordings.

5.2.2 TEVA and acoustic signal typing

The NKI developed a computer program (Tracheoesophageal Voice Analysis tool, TEVA; van Son, 2012) to assist researchers and SLPs to identify acoustic signal types. TEVA runs as a Praat extension (Boersma, 2001; Boersma and Weenink, 2009) and both programs are available under an Open Source License (GPL). Acoustic signal type classification for TE speakers requires an observer to classify a segment of a spectrogram into one of four signal types: stable and harmonic (*1*), stable with at least one harmonic (*2*), unstable or partly harmonic (*3*) and barely harmonic (*4*) which corresponds roughly to a severity scale from *good* to *bad* (van As-Brooks et al., 2006). As observers may differ in how they arrive at a classification, a consensus procedure was used for segment selection and classification into signal type.

Using the TEVA program, two experienced SLPs (authors Clapham and Van As-Brooks), classified all 87 recordings into signal type based on visual inspection of the spectrogram. They were blind to speaker characteristics (e.g. prosthesis type or gender) and were unable to listen to the recordings.

The spectrograms were classified according to AST over two steps. During step one, each rater independently classified the segment of the spectrogram that she considered most stable (1.75 seconds) and in step two, a consensus model was used whereby the raters first agreed on the segment of the spectrogram that was the most stable and then agreed on the AST of this stable segment. This interval of 1.75 seconds is shorter than the 2 seconds advised in (van As, 2001; van As-Brooks et al., 2006) because several of the recordings had been segmented (i.e., the original unedited recordings were no longer available) meaning that the margins of the spectrogram would be invisible for stimuli with a length of 2 seconds. Inter-rater agreement was 58% before consensus with a correlation coefficient of $R = 0.75$ between the AST values (p<0.001). See Table 5.3 for the distribution of the speakers over AST classes.

5.2.3 Acoustic measurements

The consensus intervals were used to measure the acoustic features. Table 5.2 lists the acoustic features which were selected for this study, based on the

Feature	Description
VF	Fraction of frames that are voiced
MVD	Maximum voicing duration
F_0	Standard deviation of F_0
Shimmer	
Jitter	
HNR	Harmonics-to-noise ratio (dB)
HNR_{low}	HNR low pass filtered speech (<700Hz)
HNR_{high}	HNR band pass (700Hz - 2300Hz)
GNE	Glottal noise energy
CPP	Cepstral peak prominence
BED	Band energy difference
QF_1-QF_3	F_1-F_3 quality factor (F_i/B_i)

Table 5.2: Overview of acoustic parameters used. With the exception of BED and QF_1-QF_3, all measures depend on the detection of voicing and pitch.

studies presented in Table 5.1. These features were automatically measured with Praat with a pitch floor of 40 Hz and a window size to 25 ms (see *AcousticMeasureScripts.praat*; van Son, 2013). Where possible, we used published settings for measurements (van As, 2001; van Gogh et al., 2005). MVD was determined on the whole /a/ realization. For practical reasons, the HNR_{low}, HNR_{high}, and cepstral rahmonic intensity as used by van Gogh et al. (2005) were substituted with the HNR of low-passed and band-passed speech, and the cepstral peak prominence (CPP), respectively. Formant quality factors (QF_1-QF_3) were added as non-voice measures. D_2 and Sample Entropy as proposed in Yan et al. (2013) could not yet be implemented in Praat.

5.2.4 Acoustic features and machine learning

Automatically evaluating AST based on acoustic information has aspects of both classification (identity) and regression (size): each signal type is distinct and derived from features in the spectrogram (classification), yet the signal types are also ordinal whereby prediction between classes can be seen as an intermediate value (regression). Model performance can be evaluated based on classification error when using a classification algorithm, on the root mean square (RMSE) when using a regression algorithm, or on the explained variance (e.g., correlation coefficient) between observed and predicted signal types.

Although it is not possible to find *the* best classification function in an efficient way, it is still possible to find *an* efficient classification function from examples. Using a variety of machine learning techniques and feature selection, it is also possible to estimate the robustness of the solution under different sets of examples (Guyon and Elisseeff, 2003; Ladha and Deepa, 2011; Maindonald and Braun, 2006). These technologies can also be used to determine the importance or redundancy of individual and combinations of acoustic features for classification. Acoustic features were selected and ordered on explanatory importance using machine learning (ML) techniques as described in Guyon and Elisseeff (2003); Ladha and Deepa (2011). All ML experiments were done using implementations in R (R Core Team, 1998–2012) (see *model_AST.R*; van Son, 2013). Seven ML algorithms were tested: Linear model (*LM*), Linear and Quadratic discriminant analysis (*LDA*, *QDA*), Support Vector Machines (*SVM*), Random Forest (*RF*), CaRT (*RPart*), and Neural nets (*NNet*).

Methods were used with their default settings in R (R Core Team, 1998–2012). The number of possible settings is too large to allow meaningful optimization for our data set. The results presented here should be interpreted as lower bounds on performance. All ML methods were tested in classification and regression mode. Where necessary, regression results were converted to classification, class *1-4*, by rounding (*LM*, *NNet*). Classification probabilities were converted to regression values by calculating the expected value (*LDA*, *QDA*).

A wrapper methodology with forward selection and backward elimination was used for feature selection (Guyon and Elisseeff, 2003; Ladha and Deepa, 2011). This means that each ML method was used as a black box that outputs a figure of merit given a training and feature set. Stratified bootstrap sampling validation, with 40-fold resampling, was used to check robustness of feature selection. Leave-one-out cross-validation (LOOCV), where each sample is predicted using all but this sample as training set, was used to estimate the real predictive power of the models. Three recordings had no measurable voicing, and thus, no pitch related features. These were assigned predicted type *4*.

5.3 Results and discussion

5.3.1 Single feature analysis

A summary of the relationship between acoustic features and observed AST is listed in Table 5.3. Nine of the acoustic features show a main effect for classification type and many of these can differentiate between signal type pairs (see post-hoc results in Table 5.3). A simple linear regression model using VF alone can explain almost 60% of the variance in classification. There are strong correlations between the acoustic features (not shown), the strongest between VF and MVD (R^2=0.71), HNR and HNR$_{low}$ (R^2=0.58), and between VF and HNR (R^2=0.64). Purely random classification, using permutations, results in a correct classification of 0.33 (sd=0.04) and R^2=0.01 (sd=0.016).

ML methods were trained on the link between AST classification and single acoustical features. The best and median performances are presented in Figure 5.1. For both R^2 and correct classification, VF, MVD, HNR and HNR_{low} outperform the other features (in this order). Where median values are higher than chance performance plus two standard deviations (see Figure 5.1), the classification is likely robust. Otherwise, the performance is expected to be erratic. For the leftmost four features (VF, MVD, HNR, HNR_{low}), the classifications seem to be robust. For neither classification nor regression do QF3, QF2 or HNR_{high} reach this level of significance.

5.3.2 Feature combinations

Bootstrap validation versus LOOCV and forward selection versus backward elimination all resulted in comparable feature selection and performance (not shown). Classification outperformed regression slightly, but was otherwise comparable. Only results for classification with LOOCV and forward selection will be reported unless indicated otherwise.

Classification performance is plotted versus the number of acoustic features in Figure 5.2 for all ML algorithms used. The ML methods split into two groups: *LDA*, *QDA*, *SVM*, and *RPart* all reach high correct classification rates with only

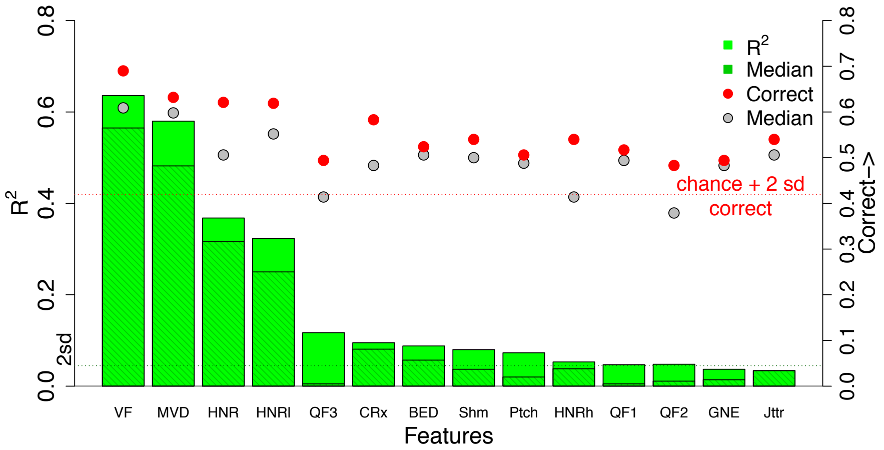


Figure 5.1: Maximum R^2 (green bars) and correct classification (red circles) for single acoustical features over all ML methods (ordered on decreasing R^2, LOOCV, see text). Median R^2 and correct classification are indicated with cross-hatched bars and grey circles, respectively. Added are chance level + 2·sd lines for R^2 (0.045, green) and correct classification (0.42, red).

three acoustic features. The remaining methods, *LM*, *RF*, and *NNet*, perform worse. The inherently stochastic nature of *RF*, and *NNet* might at least partly explain their erratic results on small data sets. "Good" runs were selected for these two methods.

Classification performance as a function of the number of features varies widely between methods. Complex ML methods such as *SVM*, *QDA*, and *RF* are sensitive to the "curse of dimensionality": Including more features leads to marked decreases in performance due to overtraining (Guyon and Elisseeff, 2003; Ladha and Deepa, 2011). However, *RPart* drops features that do not increase performance. This might be an explanation for the stable performance of *RPart*.

The order of selection of features was investigated with bootstrap validation (see section 5.2.4). All methods select either VF or MVD as their first feature. The second feature is then one of the HNR features (HNR, HNR$_{high}$, or HNR$_{low}$). The third feature selected is more varied, either another from MVD, VF, or the HNR group, but also QF2 and BED were selected (*LM*, *LDA*). The LOOCV results were equivalent, but varied somewhat in the third selected feature (Figure 5.2).

From this we conclude that VF and MVD alone supply enough information to get well over 60% correct classification. Including the HNR group of features

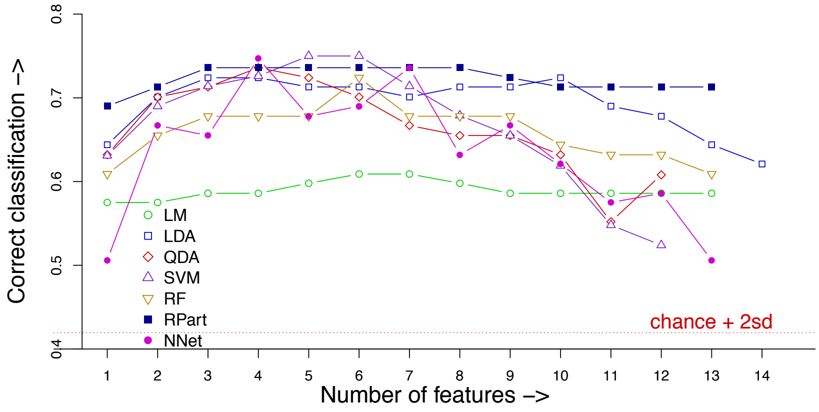


Figure 5.2: Correct classification as a function of the number of acoustic features for all ML algorithms (see section 5.2.4). Features were included using forward selection and leave-one-out cross validation. Added is a chance level + 2·sd line (0.42, red). R^2 values follow comparable curves (not shown, see text).

then allows performance to rise to over 70% correct classification (see Figure 5.2). Members of these groups often appear again as third selected feature, indicating they are not completely interchangeable (redundant). With three features, the high-performance methods get over 70% correct. Increasing the number of features can sometimes improve performance even to 75% correct classification with five features, e.g., for *QDA* and *SVM*. However, differences become rather small and unreliable for our data set. For all ML methods it was found that an analysis which excluded VF, where MVD would substitute for it, resulted in slightly lower performance, still reaching ∼70% correct classification and R^2 up to 0.6 (note, *RPart* performed *better* without VF).

AST classification is also an ordinal scale. Therefore, not all classification errors should be weighted equal. An AST class *1-4* confusion is worse than a class *3-4* confusion. The squared correlation coefficient (R^2) between predicted and consensus classification is a figure of merit that measures such discrepancies. For all ML methods, except *LM*, the R^2 peaked between 0.6 and 0.7 at best classification performance in Figure 5.2. That is, the ML methods were able to explain more than 60% (close to 70% for *SVM*) of the variance in the consensus classification.

Table 5.3 presents several other features beside VF, MVD, and HNR that show statistically significant differences between AST classes, e.g., Shimmer, GNE, CPP. However, it seems the ML algorithms applied here are unable to use

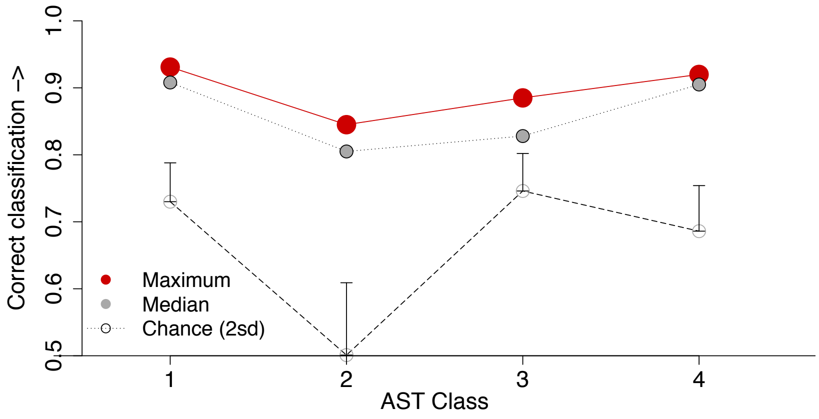


Figure 5.3: Best classification for individual types versus all others (see text). Presented are best and median ML classifier performance. Added is chance level for each class distinction and 2·sd error bars for random classification.

this information to improve classification. The analysis presented in Figure 5.2 was repeated with the exclusion of VF and MVD. With this exclusion, classification was regularly over 0.6 but R^2 came only slightly over 0.4 ($\leq$43% explained variance). The first feature selected was always either HNR or HNR_{low}.

When excluding all of VF, MVD, and the HNR group of features, correct classification peaked at 0.62 (for *QDA* with BED + Shimmer), but was well below 0.6 for all other methods (not shown). This might seem high considering the chance level was 0.33. However, R^2 was rarely above 0.2, and generally lower ($\leq$20% explained variance). This indicates that classification errors became much more random. Information in these acoustic features might mainly identify individual classes (c.f., Table 5.3). The first feature selected under these conditions was four times CPP, and BED, Shimmer, and QF3 each once.

5.3.3 Individual class type identification

Best performance of AST classification might not be attained using a single model for all types. The above analysis was repeated, but now as four two-type classification tasks. All ML methods (see section 5.2.4) were trained and tested on a single type with all other types merged into a single class, e.g., type *2* against types *1*, *3*, and *4* combined. Chance classification performance was recalculated for each combination. The results are presented in Figure 5.3.

As expected, the end-point types *1* and *4* were easier to identify than the inner types *2* and *3*. Behavior of the classifiers was more erratic than with the original four type task. *SVM* could not even classify type *3* versus the others. Number and selection of features varied much more than the patterns seen in Figure 5.2. This is likely caused by the unbalance between positive and negative samples in this task.

5.4 Conclusions

Many acoustic measurements correlate, often strongly, with the AST classification (see Tables 5.1, 5.3). However, our study shows that only two groups of features can perform a classification to any reasonable extent: voice detection (VF and MVD) and the harmonics-to-noise ratio (HNR, HNR_{low}, and HNR_{high}). Other factors improve classification performance only marginally. This indicates that the presence and duration of voicing and the harmonic-to-noise ratio are the most salient acoustic features that can be used to classify a TE signal into its acoustic signal type. In our study, *QDA* and *SVM* performed best, but *RPART* would perform almost as good and can easily be assessed automatically. A practical tool incorporating these methods will be made available online at van Son (2012).

The fact that classical measures of glottal voices, like jitter and shimmer, are less salient in TE speech can possibly be attributed to the inherent instability of neo-glottis vibrations (Yan et al., 2013).

Acknowledgements

This research was made possible by an unrestricted research grant from Atos Medical, Sweden and a grant from the Verwelius Foundation, Naarden. We thank Louis Pols for his helpful comments and suggestions.

References

Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5:341–345, 2001. URL `http://www.Praat.org/`.

Paul Boersma and David Weenink. Praat: doing phonetics by computer. Computer program: http://www.Praat.org/, 2009.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

Tino Haderlein, Elmar Nöth, Hikmet Toy, Anton Batliner, Maria Schuster, Ulrich Eysholdt, Joachim Hornegger, and Frank Rosanowski. Automatic evaluation of prosodic features of tracheoesophageal substitute voice. *European Archives of Oto-Rhino-Laryngology*, 264(11):1315–1321, 2007.

Frans JM Hilgers and Alfons JM Balm. Long-term results of vocal rehabilitation after total laryngectomy with the low-resistance, indwelling provoxtm voice prosthesis system. *Clinical Otolaryngology & Allied Sciences*, 18(6):517–523, 2007.

Frans JM Hilgers, Annemieke H Ackerstaff, Alfons JM Balm, I Bing Tan, Neil K Aaronson, and Jan-Ove Persson. Development and clinical evaluation of a second-generation voice prosthesis (provox® 2), designed for anterograde and retrograde insertion. *Acta oto-laryngologica*, 117(6):889–896, 1997.

Frans JM Hilgers, Annemieke H Ackerstaff, Maya van Rossum, Irene Jacobi, Alfons JM Balm, I Bing Tan, and Michiel WM van den Brekel. Clinical phase i/feasibility study of the next generation indwelling provox voice prosthesis (provox vega). *Acta oto-laryngologica*, 130(4):511–519, 2010.

P. Jongmans. *The intelligibility of tracheoesophageal speech: An analytic and rehabilitation study*. PhD thesis, University of Amsterdam, 2008.

R. Kazi, J. Kanagalingam, R. Venkitaraman, V. Prasad, P. Clarke, C.M. Nutting, P. Rhys-Evans, and K.J. Harrington. Electroglottographic and perceptual evaluation of tracheoesophageal speech. *Journal of Voice*, 23(2): 247–254, 2009.

L. Ladha and T. Deepa. Feature selection methods and algorithms. *International Journal on Computer Science and Engineering*, 3(5):1787–1797, 2011.

J. Maindonald and J. Braun. *Data analysis and graphics using R: an example-based approach*, volume 10. Cambridge University Press, 2006.

Y. Maryn, C. Dick, C. Vandenbruaene, T. Vauterin, and T. Jacobs. Spectral, cepstral, and multivariate exploration of tracheoesophageal voice quality in continuous speech and sustained vowels. *The Laryngoscope*, 119(12):2384–2394, 2009.

M. Moerman, Jean Pierre Martens, M. Van der Borgt, M. Peleman, M. Gillis, and P. Dejonckere. Perceptual evaluation of substitution voices: development and evaluation of the (i)infvo rating scale. *European Archives of Oto-Rhino-Laryngology*, 263(2):183–187, 2 2006.

Mieke Moerman, Glenn Pieters, Jean Pierre Martens, Marie-Jeanne Van der Borgt, and Philippe Dejonckere. Objective evaluation of the quality of substitution voices. *European Archives of Oto-Rhino-Laryngology*, 261(10):541–547, 11 2004.

B. M. R. Op de Coul, AH Ackerstaff, C. J. van As-Brooks, F. J. A. Van Den Hoogen, C. A. Meeuwis, J. J. Manni, and F. J. M. Hilgers. Compliance, quality of life and quantitative voice quality aspects of hands-free speech. *Acta oto-laryngologica*, 125(6):629–637, 2005.

R Core Team. The R project for statistical computing, version 2.15.2 (2012-10-26). Computer program : http://www.r-project.org/, 1998–2012.

I.R. Titze. Workshop on Acoustic Voice Analysis: Summary Statement. pages 1–36. Denver, CO: National Center for Voice and Speech, 1995.

Corina J. van As. *Tracheoesophageal speech. a multidimensional assessment of voice quality*. PhD thesis, University of Amsterdam, September 2001.

C. J. van As-Brooks, F.J. Koopmans-van Beinum, L.C.W. Pols, and F.J.M. Hilgers. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *Journal of Voice*, 20(3):355–368, 2006.

C.D.L. van Gogh, J.M. Festen, I.M. Verdonck-de Leeuw, A.J. Parker, L. Traissac, A.D. Cheesman, and H.F. Mahieu. Acoustical analysis of tracheoesophageal voice. *Speech Communication*, 47(1):160–168, 2005.

R. J. J. H. van Son. A study of pitch, formant, and spectral estimation errors introduced by three lossy speech compression algorithms. *Acta acustica united with acustica*, 91(4):771–778, 2005.

R. J. J. H. van Son. NKI TE-VOICE Analysis tool (TEVA). Computer program: http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/NKIcorpora/NKI_TEVA/, 2012.

R. J. J. H. van Son. Additional Files to Interspeech 13 proceedings. Link: http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/NKIcorpora/NKI_TEVA/, 2013.

Nan Yan, Manwa L. Ng, Dongning Wang, Lan Zhang, Victor Chan, and Rerrario S. Ho. Nonlinear dynamical analysis of laryngeal, esophageal, and tracheoesophageal speech of cantonese. *Journal of Voice*, 27(1):101–110, 2013. ISSN 0892-1997. doi: 10.1016/j.jvoice.2012.06.009. URL `http://www.sciencedirect.com/science/article/pii/S0892199712001014`.

Y. Zhang, J.J. Jiang, et al. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *Journal of voice: official journal of the Voice Foundation*, 22(1):1, 2008.

	Effect	R^2	Median				AST comparisons (p values Mann-Whitney test)					
	p<	(*LM*)	*1*	*2*	*3*	*4*	*1-2*	*1-3*	*1-4*	*2-3*	*2-4*	*3-4*
VF	.000	0.595	1.00	0.95	0.34	0.06	.000 !	.000 !	.000 !	.001 !	.000 !	.002 !
MVD	.000	0.457	5.06	3.57	1.36	0.36	.007*	.000	.000 !	.000	.000 !	.000 !
F$_0$	.623	0.010	4.44	9.41	6.59	4.29						
Shimmer	.002	0.105	0.11	0.21	0.21	0.21	.001 !	.002	.003 !	.835 !	.269 !	.381 !
Jitter	.018	0.018	0.01	0.02	0.01	0.01	.008*!	.033	.359 !	.721 !	.048 !	.134 !
HNR	.000	0.355	10.42	3.74	2.18	1.38	.000	.000	.000	.107	.000	.118
HNR$_{low}$	.000	0.282	28.11	17.43	13.78	10.68	.001	.003	.000	.136	.000	.316
HNR$_{high}$	.000	0.042	15.63	7.46	8.09	9.25	.000	.012	.001	.732	.501	.892
GNE	.003	0.017	0.923	0.850	0.890	0.879	.013 !	.152 !	.152	.613 !	.306 !	.963 !
CPP	.007*	0.084	21.40	17.90	17.41	16.26	.002	.054	.002	.853	.204	.294
BED	.091	0.097	26.94	23.88	17.61	17.73						
QF1	.027	0.190	5.91	3.54	5.23	5.02	.039 !	.560 !	.450 !	.040 !	.027 !	.630 !
QF2	.798	0.010	6.90	6.20	6.70	6.5						
QF3	.307	0.015	5.85	10.00	9.70	7.00						

Table 5.3: Effect of signal type (AST) on each acoustic variable (Kruskal-Wallis test), explained variance (R^2) using a linear model, median variable value, and post-hoc comparisons (Mann-Whitney tests, if Effect significant).

P-values *: p<.0083, shaded: p<.0035 (correction for multiple comparisons). Exclamation mark highlights comparisons where exact significance cannot be computed due to ties within a category. AST class frequencies (c:N) - 1:14, 2:43, 3:13, 4:17. See Table 5.2 for abbreviations.

6

The relationship between acoustic signal typing and perceptual evaluation of tracheoesophageal voice quality for sustained vowels[0]

Abstract

Objectives To investigate the relationship between acoustic signal typing and perceptual evaluation of sustained vowels produced by tracheoesophageal (TE) speakers and the use of signal typing in the clinical setting.

Method Two evaluators independently categorized 1.75-second segments of narrow-band spectrograms according to acoustic signal typing and independently evaluated the recording of the same segments on a visual analogue scale according to overall perceptual-acoustic voice quality. The relationship between acoustic signal typing and overall voice quality (as a continuous scale and as a 4-point ordinal scale) was investigated and the proportion of inter-rater agreement as well as the reliability between the two measures.

Results The agreement between signal type (I-IV) and ordinal voice quality (4-point scale) was low but significant and there was a significant linear relationship between the variables. Signal type correctly predicted less than half of the voice quality data. There was a significant main effect of signal type on continuous

[0] R.P. Clapham, C.J. Van As-Brooks, R.J.J.H. van Son, F.J.M. Hilgers, M.W.M. van den Brekel, *Journal of Voice, 29, 2015, e29.*

voice quality scores with significant differences in median quality scores between signal types I-IV, I-III and I-II.
Conclusions Signal typing can be used as an adjunct to perceptual and acoustic evaluation of the same stimuli for TE speech as part of a multi-dimensional evaluation protocol. Signal typing in its current form provides limited predictive information on voice quality and there is significant overlap between signal type II and III and perceptual categories. Future work should consider whether the current four signal types could be refined.

6.1 Introduction

Functional voice assessment requires a multi-dimensional approach to evaluation and data should allow a clinician to determine whether a voice is classified normal or pathological, the severity and cause of a pathology and allow tracking changes in voice over time (Dejonckere, 2010). It is recommended that an evaluation protocol contain perceptual evaluation combined with acoustic, imaging, aerodynamic and patient self-report measures (Dejonckere, 2010). A specialized protocol for voice assessment is required within the area of tracheoesophageal (TE) speech because the overall voice quality of substitute voicing should be compared to "near normal laryngeal voicing" rather than normal laryngeal voicing and performing acoustic evaluation can lead to unreliable and inaccurate measurements because standard pitch-detection algorithms in general acoustic software fail when the speech signal has low or no fundamental frequency or high levels of noise.

Titze (1995) introduced acoustic signal typing for laryngeal speakers as a decision making tool on whether the researcher/clinician could collect reliable acoustic data. Signal typing involves categorising recorded speech samples based on visual characteristics observed on narrow-band spectrograms. van As-Brooks et al. (2006) adapted Titze's signal-typing technique for TE voice and identified four signal types based on the spectral characteristics of this speaker group. Although the use of signal typing is recommended as a decision making tool (Titze, 1995; van As-Brooks et al., 2006), there is a relationship between signal type of sustained vowels and auditory-perceptual judgements of voice quality for running speech (D'Alatri et al., 2011; van As-Brooks et al., 2006) and as such, signal typing has been proposed as an indicator of the overall perception of voice quality or of functional voice outcome (D'Alatri et al., 2011; Sprecher et al., 2010; van As-Brooks et al., 2006). The use of signal typing as part of a multi-dimensional evaluation of TE voice can be useful as it is estimated that 77 percent of TE speakers have a measurable fundamental frequency (van As-Brooks et al., 2006) and many acoustic measures will fail this population because of the lack of periodicity in the speech signal.

As noted by Van Gogh et al. (2005), there is a subjective component when performing signal typing and reliability and agreement measures warrant reporting just as auditory-perceptual reliability and agreement measures are generally reported. Many studies investigating signal type for TE speech, however, have used classifications from a single evaluator or do not include procedural information on who performed classifications and do not include reliability information (D'Alatri et al., 2011; Lawson et al., 2001; Sprecher et al., 2010; van As-Brooks et al., 2006; Van Gogh et al., 2005). The present study is unique in that we (a) consider the relationship of signal type and perceptual evaluation of the same stimuli and (b) use a scoring procedure that reflects the clinical setting. That is, rather than use mean scores of a large group of raters, we use consensus scores made by two speech pathologists.

This paper explores the use of signal typing in its current form for TE voice and the relationship of signal type to perceptual scores of voice quality of the same stimuli. Our principal research line investigates the association between signal type and voice quality for the same stimuli and whether there is a predictive relationship between the two variables. Our secondary research line was to compare the inter-rater agreement and reliability of signal type evaluations with voice quality evaluations. The key variables are consensus acoustic signal type (ordinal data containing four categories) and consensus voice quality scores (continuous data 0-1000). We also utilize each rater's individual evaluations (i.e. pre-consensus evaluations) to report inter-rater agreement and reliability.

6.2 Method

6.2.1 Audio stimuli

Audio recordings were collected at the Netherlands Cancer Institute (Amsterdam, the Netherlands) as part of various research studies between 1996 and 2009. All speakers produced a sustained /a/ as part of the recording procedure. All speakers provided informed consent at the time of data collection and granted use of the recordings for research purposes. As the recording conditions, settings and equipment varied across the past studies, for the current study we digitalized analogue recordings and all recordings were converted to 44.1 kHz sampling rate with 16-bit Signed Integer PCM encoding. No compression had been used on the recordings. Where possible, we used original recordings, but in several cases, only 2-second segments of the vowels were available.

The collection contains recordings from 87 TE speakers. The majority of speakers were male (74 [85%]) and median age at time of laryngectomy was 57 years (range 38-85 years; age at time of laryngectomy was not recorded for one speaker). Age at the time of the recordings could be retraced for 37 (43%)

of the speakers (median age 66 years, range 46-81 years). As many speakers provided recordings for multiple studies, we selected the stimuli with the earliest recording date. For the recordings used in the present study, 83 (95%) speakers used a Provox1 or Provox2 prosthesis and the remaining 4 (5%) speakers used a Provox Vega prosthesis.

6.2.2 Acoustic signal typing

Procedure The four signal types are Type 1 (Stable and harmonic), Type II (Stable and at least one harmonic, Type III (Unstable or partly harmonic) and Type IV (Barely harmonic). During the evaluation of 12 practice items, two speech pathologists (authors RPC and CVAB) discussed and adapted scale definitions. The signal typing criteria presented in van As-Brooks et al. (2006) was adjusted to account for the minimum length of the pre-segmented stimuli and perceived ambiguity in the definition of 'stable' (see Table 6.1). For this present study, 'stable' was defined as a continuous signal at the fundamental frequency harmonic. Note that the original signal typing criteria of 2 seconds was adjusted to 1.75 seconds, as pre-edited 2-second recordings would have had missing margins in the spectrograms. Note also that the 2-second rule used in van As-Brooks et al. was based on the minimum length of the stimuli.

Acoustic signal type		Criteria
I	Stable and harmonic	• Stable signal for $\geq$1.75 seconds, & • Clear harmonics from 0 to 1000 Hz
II	Stable and at least one harmonic	• Stable signal for $\geq$ 1.75 seconds, & • At least one stable harmonic at the fundamental frequency for $\geq$ 1.75 seconds
III	Unstable or partly harmonic	• No stable signal for $>$ 1.75 seconds, or • Harmonics in only part of the sample (for longer than 1 second)
IV	Barely harmonic	• No detectable harmonics or only short-term detectable harmonics for $<$ 1 second

Table 6.1: Criteria for each of the four acoustic signal types

Spectrograms were presented via a custom-made program termed the NKI TE-Voice Analysis tool (TEVA; van Son, 2012), which runs as a Praat extension (Boersma and Weenink, 2009). The entire recording was visualized in a narrow-band spectrogram (window length .1 s, time step .001 s, frequency step 10 Hz, maximum frequency 2 kHz) and raters were unable to play the sound file. Using the TEVA tool, each rater visually identified the most stable segment of the spectrogram and then classified this segment according to signal type. The raters were blind to speaker gender, speaker age and prosthesis type. After individual classification, the raters came together and agreed upon the 1.75-second segment to be evaluated and the signal type of this segment.

Rater reliability and agreement Table 6.2 lists the inter-rater agreement and disagreement grouped according to consensus signal type. Raters agreed on signal type categorization in 50 cases (57%; permutation average 29 % and sd 4 %) and were in close agreement for the remaining 31 cases (36 %; permutation average 38%, sd 5%). The kappa for inter-rater agreement was statistically significant (weighted kappa: $\kappa = .55$, $p < .001$, weights set at 0, .33, .66, 1.0). There was a statistically significant correlation between the two rater's evaluations (tau = .63, $p < .00$) and there was acceptable reliability between the raters (single-measure ICC (consistency) using a two-way model: ICC = .73, 95% CI .62-.82).

		(Dis)Agreement		
Variable	n	Exact (%)	Close (%)	Disagree (%)
AST				
I	14	6 (43)	6 (43)	2 (14)
II	44	25 (57)	18 (41)	1 (2)
III	12	4 (33)	5 (42)	3 (25)
IV	17	15 (88)	2 (12)	0 (0)
	87	50 (57)	31 (36)	6 (7)
Voice Quality				
Good	15	9 (60)	4 (27)	2 (13)
Fair	30	13 (43)	8 (27)	9 (30)
Moderate	23	7 (30)	7 (30)	9 (39)
Poor	19	7 (37)	6 (32)	6 (32)
	87	36 (41)	25 (29)	26 (30)

Table 6.2: Inter-rater agreement and disagreement for (i) acoustic signal type (AST) grouped according to consensus signal type and (ii) voice quality scores grouped according to consensus voice quality scores (converted into ordinal categories). Note: AST agreement divided into exact agreement (same category selected by raters), close agreement (categories differ by one type) and disagreement (categories differ by two types). Voice quality agreement divided into exact agreement (two scores ±125), close agreement (two scores ±250) and disagreement (two scores differ by > 250).

6.2.3 Auditory perceptual evaluation

Procedure Three months after performing signal typing evaluation, the same raters completed the auditory-perceptual evaluation task. The perceptual variables were based on scales used for the auditory-perceptual evaluation of running speech (van As-Brooks et al., 2006) and those developed for the INFVo (Moerman et al., 2006). The raters discussed and adjusted scale definitions during the evaluation of 12 practice items. Although several additional parameters were included in the data collection, we restrict our analysis to the parameter 'overall voice quality'.

The raters were blind to all speaker information, including signal type data. Stimuli were presented in a random order via an online self-paced experiment and raters listened to recordings via a headset. Stimuli were not re-presented. Raters recorded their evaluations on a computerised visual analogue scale built within the TEVA tool. The response scale contained textual anchors at both extreme and did not display tick marks. Raters moved the cursor along the line to the desired location between the two anchors and the cursor location was then saved as a value between 0 ("least similar to normal") and 1000 ("most similar to normal").

Scores that differed between the raters by more than 125 points were discussed and re-scored in the consensus round. When scores were within the range of agreement, the mean score was considered the consensus score and these cases were not discussed. The value ±125 is derived from dividing the scale into four intervals, which corresponds with a four-point ordinal equal appearing interval scale. To aid scoring in the consensus round, major and minor tick marks were placed at every 10% and 5% scale distances, respectively. Numeric anchors were displayed at major tick marks.

Rater reliability and agreement Table 6.2 lists the inter-rater agreement and disagreement grouped according to consensus voice quality scores (converted into four ordinal categories). The two rater's scores were in exact agreement (difference equal to or less than 125 points) in 36 cases (41%) and were in close agreement (difference equal to or less than 250 points) in the remaining 25 cases (29 %). The strength of the correlation between the two raters' individual judgements was statistically significant (tau = .43, $p < .001$) and the reliability between the raters was acceptable (single-measure ICC (consistency) using a two-way model: ICC = .63, 95% CI .49-.74).

6.2.4 Statistical analysis

All statistical analyses were completed with the statistics program R (R Core Team, 1998–2012) and p-value was set to $p < 0.05$ for testing main effects.

A Bonferroni correction was applied for post-hoc comparisons. Although the evaluation task (i.e. voice quality vs. signal type) and stimuli (i.e. auditory-perceptual vs. visual) differed between the two measurements, where possible we used statistical tests for dependent samples as the stimuli were derived from the same recordings and the raters were the same for each task.

Relationship between the two variables The Chi-square linear-by-linear test of association for ordinal data was used to test the association between consensus signal type categories and consensus voice quality categories. To do this, the visual analogue scale was divided into four equal parts and the consensus scores were coded into one of four ordinal categories: 'good' ($>$ 750.75), 'fair' ($>$ 500.5 & $\leq$ 750.75), 'moderate' ($>$ 250.25 & $\leq$ 500.5) and 'poor' ($\leq$ 250.25). To further understand the relationship between the two variables, a non-parametric Anova (Kruskal-Wallis test) with Mann-Whitney test for post-hoc comparisons was used and we evaluated whether voice quality was a significant predictor of signal type using linear regression.

Comparing proportions of agreement To compare proportions of inter-rater agreement between the two measures, we used McNemar's non-parametric test for paired samples. That is, we completed two analyses: (1) signal type exact agreement with voice quality exact agreement and (2) signal type agreement (close + exact) with voice quality agreement (close + exact). We used a permutation method (data resampling without replacement, N=100,000) to calculate the level of chance agreement within the data.

6.3 Results

6.3.1 Relationship between signal type and voice quality

Ordinal scores of voice quality Consensus voice quality scores were converted into 4-point ordinal scale by dividing the visual analogue scale into four equal parts and labelled "good", "fair", "moderate" and "poor". The largest category was for 'fair' (30 cases) and the category with the least number of cases was for 'good' (15 cases) (see Table 6.2). The relationship between consensus signal type and consensus ordinal voice quality scores is presented in Figure 6.1. Results of the kappa statistic indicate low, but statistically significant agreement between the two measures (κ = 0.22, p = .004).

A test of the linear-by-linear association for ordinal variables indicates the association between the two variables was significant (X^2 (1, N = 87) = 29.71, $p <$ .001). If we consider signal type (I-IV) as a predictor of voice quality category (good to poor), signal type correctly predicts 38 cases (44%) when the perceptual scale is divided into four equal categories.

Continuous scores of voice quality To further investigate the relationship between the consensus scores, we performed a non-parametric test of the effect of signal type (ordinal data) on the perceptual scores (continuous data). Kruskal-Wallis test shows there is a significant main effect of signal type on perceptual scores of voice quality ($X^2 = 31.4$, $p < .05$). Mann-Whitney tests (p set to $< .0083$ for multiple comparisons) indicate significant differences in median voice quality scores for signal type categories I-IV, I-III and II-IV. If signal

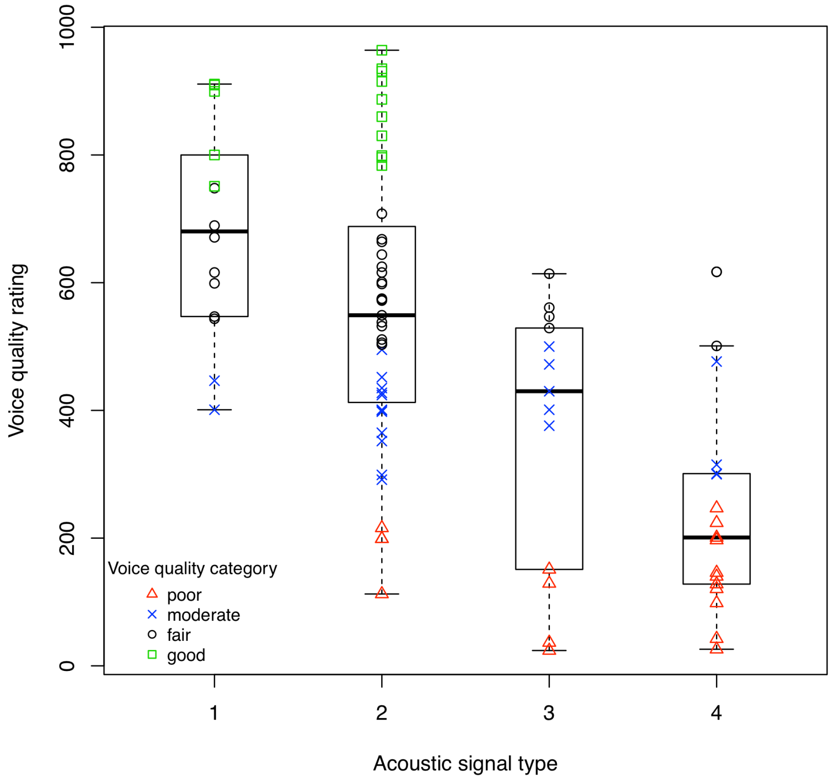


Figure 6.1: Voice quality scores by acoustic signal type. Voice quality data points overlay the boxplot and are coded according to boundaries for converting the continuous scores into categorical data ("good" n=15 (17%): AST I n=5, II n=10; "Fair" n=30 (34%): AST I n=7, II n=17, III n=4, IV n=2; "Moderate" n=23 (28%): AST I n=2, II n=12, III n=5, IV n=4; "Poor" n=19 (22%): AST II n=4, III n=4, IV n=11).

type is considered pseudo-continuous data, a linear regression analysis indicated that voice quality score significantly predicts acoustic signal types ($p < .001$) and explains a statistically significant proportion of the variation (multiple R^2 = .37, F(1,85) = 49.8, $p < .001$).

6.3.2 Comparing proportions of rater agreement

McNemar's test revealed no statistically significant difference in proportion of exact inter-rater agreement for perceptual voice quality scores (41%) and acoustic signal type (58%) (McNemar's X^2 (1, N = 87) = 3.84, p = .050). The difference in proportion of exact/close inter-rater agreement between voice quality measures (70%) and signal type (94%) was significant (McNemar's X^2 (1, N = 87) = 12.89, $p < .001$).

6.4 Discussion

Our primary research line was to investigate the association between consensus judgments of signal type and consensus judgments of voice quality for segments of sustained vowel /a/. In terms of data distribution, over half the stimuli (51 %) was classified signal type II and the least frequent classification was for signal type III (14 %). This distribution pattern is unlike that reported in van As-Brooks et al. (2006) (signal type IV was the most frequent and type I was the least frequent) and overlaps somewhat with the distribution pattern reported D'Alatri et al. (2011) (signal type I was the most frequent and type III was the least frequently occurring category).

To allow comparison between signal type (ordinal data) and voice quality (continuous data), the visual analogue scale was divided into four equal parts. Over 60 percent of stimuli fell in the central "fair" and "moderate" range with the most frequent category being "fair" (34 %) and least frequent category being "good" (17 %). Direct comparison of the category frequencies with van As-Brooks et al. (2006) and D'Alatri et al. (2011) is not possible as both these studies used a 3-point ordinal rating scale. We elected to convert the visual analogue scale into four parts as opposed to three in an attempt to maintain the sensitivity of the scale. Converting scores made on a continuous scale into an interval scale is a technique used in researcher (e.g., Eadie and Kapsner-Smith, 2011; Kreiman and Gerratt, 1998; Peterson et al., 2013; Wuyts et al., 1999).

Figure 6.1 displays the relationship between signal type and ordinal voice quality scores. Stimuli with signal types III (unstable / some harmonics) and IV (no harmonics / mostly without harmonics) were never rated as having 'good' voice quality. Likewise, stimuli with signal type I (stable with clear harmonics) were never rated as having 'poor' voice quality. This pattern is similar to that reported by van As-Brooks et al. (2006) where the two extreme signal types never co-occurred with the opposite perceptual extreme when a 3-point ordinal scale was used. In line with Van As' study comparing signal type of vowels and voice quality of running speech, we also found a statistically significant linear association between signal type and voice quality (ordinal scores) for the same stimuli. The strength of the agreement between the two variables, however, was low. It is clear from Figure 6.1 that signal type II co-occurs with a broad range of the quality scale (predominately 'fair' to 'poor'). Excluding stimuli with signal type II from the kappa analysis results in increased agreement (from $\kappa = 0.22$ to 0.31).

Less than half of the ordinal voice quality scores can be correctly predicted by signal type. This highlights that our division of the perceptual scale into four equally spaced intervals may be too simplistic and an alternative division with unequal intervals may more accurately reflect severity (e.g., Lopes et al., 2012; Yu et al., 2001) and increase the strength of the agreement between the two scales. However, we also completed exploratory analyses of signal type on continuous voice quality data and found a statistically significant main effect. The post-hoc analysis revealed that voice quality median scores differed for three of the signal type comparisons: only quality scores for signal type I could be differentiated from the other signal types. As far as we are aware, only the study by D'Alatri et al. (2011) has found significant differences between adjoining signal types and that was only for types III and IV. We hypothesize that the broad definition of signal type II and the "and/or" criteria for signal type III results in high levels of variability in the data. This is in line with the result from our previous study in which signal types II and III were the most difficult to predict using acoustic measures (Clapham et al., 2013).

Our secondary research line was to compare the proportion of inter-rater agreement and the reliability of signal type evaluations with voice quality evaluations. Our primary argument for using the proportion of exact and proportion of close agreement as indices of agreement is that these measures can be applied to both continuous and ordinal data and allow us to directly compare proportions between the two scales. The drawback of this measurement method is that it does not take chance agreement between the two raters into consideration. We decided against converting individual voice quality scores into ordinal scores as this would not account for the situation where scores differ by a few points but fall either side of a cut-off point.

Before discussing the comparison results, we discuss the inter-rater agreement data for first signal type then voice quality. For signal type, the inter-rater disagreements were between signal types I-III (n=4) and II-IV (n=2). In no case was the disagreement larger than two categories (only possible for signal types I or IV). In no cases did the two raters disagree on signal type IV stimuli (see

Table 6.2 for details). In terms of patterns of agreement, agreement was largest for signal types II and IV. This is most likely a reflection of the number of signal type II stimuli and that signal type IV is an easily identified category. However, due to the procedure used for data collection, we are unable to state whether the disagreement occurred because of differences in categorisation (i.e. identification of signal type) or because of differences in segment selection. In a future experiment these two aspects might be separated by asking the raters to agree which 1.75 segment should be evaluated for signal type and only then do the individual ratings of signal type.

For voice quality, 61 (70%) of the rating pairs were in exact or close agreement. Scores in the centre of the scale had higher counts of disagreement than scores at the extremes of the scale (see Table 6.2). The strength of the association between the two raters' evaluations was statistically significant. Although the inter-rater agreement results indicated statistically significant levels of agreement and that the evaluations were made above chance level agreement, the results highlight that similar to perceptual evaluation, signal typing remains a subjective task and hence why consensus evaluations should be used in the clinical and research setting (for all subjective tasks) where possible.

Concerning differences in the proportion of inter-rater agreement between the two measures, although the proportions of agreement were higher for judgements of signal type than voice quality (exact agreement: 58% and 41% and close agreement: 94% and 70%, respectively), this difference was statistically significant for measures of close agreement. For measures of exact agreement, the results were just beyond the set level of statistical significance. That the proportions of agreement are larger for signal typing data is not an unexpected result; the signal typing task requires each rater to select one of four described categories (i.e. 25% agreement due to chance) whereas in the voice quality task, the scale does not force the rater to select a category and textual anchors are only provided at the scale extremes. Although the proportion of close agreement on signal type was significantly higher than for voice quality, because of differences in the scales the distances are not equal between the two variables and as such are difficult to compare directly. That is, close signal type agreement means that the scores differ by a maximum of $\frac{1}{2}$ the 'scale' whereas close voice quality agreement means that the scores differ by a maximum of $\frac{1}{4}$ of the scale.

Regarding the inter-rater reliability for the two measures, the reliability for both variables was significant but stronger correlations were found for signal type measures than voice quality measures. This is not surprising as the signal type variable requires the rater to make a forced choice from four options with each option having some criteria whereas the voice quality scale is on a visual analogue scale without textual anchors over the continuum of the scale. Compared to other studies of perceptual voice quality employing ordinal scoring systems, the signal type results are similar to the average correlation value reported in Shrivastav et al. (2005) for evaluations of breathiness on a 5-point scale (average Tau .64) and are lower than the coefficient reported in Karnell et al. (2007) for the Grade scale on a 4-point scale (Spearman = .85).

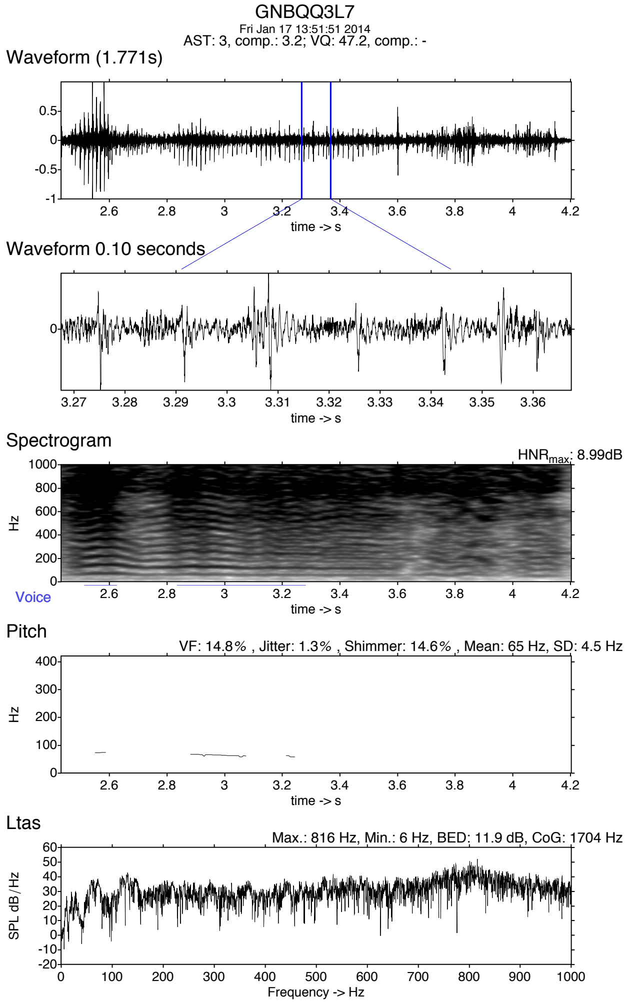


Figure 6.2: Concept version of the voicegram. The print displays (from top down) speaker code, date of print, observed signal type and voice quality score, computed signal type and voice quality score, waveform (box 1) and central 10 ms from waveform (box 2), spectrogram of predetermined segment used for signal typing and perceptual evaluation (box 3), pitch contour (box 4), and long-term average spectrum (Ltas) (box 5).

The ICC values for the two variables are stronger for signal type variable than the voice quality variable (.73 and .63, respectively). Compared to other studies, the reliability results are lower than that reported in Nemr et al. (2012) for a 3-point scale to evaluate Grade from the GRBAS for healthy control and speakers with dysphonia (ICC = .88) and that reported in Zraick et al. (2011) for the Grade part of the GRBAS (ICC = .66). Although agreement and reliability data are low, the procedure used to collect data (i.e. consensus scores) is a technique that can be used in the clinical situation.

The results suggest that signal typing in its current form can be used as part of a multi-dimensional assessment of voice quality predominately as a way to categorize voice quality and serve as a decision making tool on acoustic analysis. We anticipate that future work will consider updating the signal type definitions by including signal sub-types for types II and III (e.g. differentiation between types that contain continuous, flat harmonics and types that contain continuous, fluctuating harmonics). Part of this difficulty may be due to the variability in TE speech, that is, type III instability can be cause by hypertonicity or hypotonicity which both sound very different to a listener. However, in terms of signal typing serving as a basis for further acoustic analysis, type III would indicate that there is not stable fundamental frequency and this should be considered when acoustic analyses are carried out.

This current study is part of our efforts to automate subjective evaluation of speech and voice quality so they can complement a clinician's evaluation. To this end, we have already begun work on automating signal type based on acoustic information (Clapham et al., 2013). We envisage that signal typing could be a useful component in the multidimensional evaluation of voice quality and when paired with automatic acoustic data, predicted perceptual scores and observed perceptual scores, a clinician can have a "voicegram" of the speaker that can be printed and kept in a patient's file for comparison with other patients and assessment of treatment results. We are currently developing a function within the TEVA application to produce a "voicegram" of a speaker which contains several automated acoustic measures and can display the predicted acoustic signal type (see Figure 6.2 for a concept voicegram).

6.5 Conclusions

The results support the use of signal typing as part of a multi-dimensional evaluation of functional voice assessment. There is a statistically significant relationship between the two measures but signal typing in its current form provides limited predictive information on voice quality. The two extreme signal type categories are clear but there is a large overlap between signal types II and III and perceptual categories. However, signal typing can serve as a basis for determining further acoustic analysis (e.g. type III would indicate that there is not stable fundamental frequency and fundamental frequency-based acoustic measures should be avoided. Our results have confirmed that while signal typing is a useful approach to evaluating voice quality, the definitions of the four existing signal types and inclusion of subtypes warrants further investigation.

Acknowledgements

Part of this research was funded by unrestricted research grants from Atos Medical (Horby, Sweden) and the Verwelius Foundation (Naarden, the Netherlands). The authors wish to thank all the researchers from the NKI/AVL who collected the speech recordings.

References

Paul Boersma and David Weenink. Praat: doing phonetics by computer. Computer program: http://www.Praat.org/, 2009.

Renee Peje Clapham, Corina J Van As-Brooks, Michiel WM Van den Brekel, Frans JM Hilgers, and RJJH Van Son. Automatic tracheoesophageal voice typing using acoustic parameters. In *INTERSPEECH*, pages 2162–2166, 2013.

L. D'Alatri, F. Bussu, E. Scarano, G. Paludetti, and M.R. Marchese. Objective and subjective assessment of tracheoesophageal prosthesis voice outcome. *Journal of Voice*, pages 607–613, 2011.

Philippe H Dejonckere. Assessment of voice and respiratory function. In *Surgery of larynx and trachea*, pages 11–26. Springer, 2010.

Tanya L Eadie and Mara Kapsner-Smith. The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research*, 54(2):430–447, 2011.

Michael P Karnell, Sarah D Melton, Jana M Childes, Todd C Coleman, Scott A Dailey, and Henry T Hoffman. Reliability of clinician-based (grbas and cape-v) and patient-based (v-rqol and ipvi) documentation of voice disorders. *Journal of Voice*, 21(5):576–590, 2007.

Jody Kreiman and Bruce R Gerratt. Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America*, 104(3):1598–1608, 1998.

Georges Lawson, Jacques Jamart, and Marc Remacle. Improving the functional outcome of tucker's reconstructive laryngectomy. *Head & neck*, 23(10):871–878, 2001.

Leonardo Wanderley Lopes, Ivonaldo Leidson Barbosa Lima, Larissa Nadjara Alves Almeida, Débora Pontes Cavalcante, and Anna Alice Figueirêdo de Almeida. Severity of voice disorders in children: correlations between perceptual and acoustic data. *Journal of Voice*, 26(6):819–e7, 2012.

M. Moerman, Jean Pierre Martens, M. Van der Borgt, M. Peleman, M. Gillis, and P. Dejonckere. Perceptual evaluation of substitution voices: development and evaluation of the (i)infvo rating scale. *European Archives of Oto-Rhino-Laryngology*, 263(2):183–187, 2 2006.

Katia Nemr, Marcia Simões-Zenari, Gislaine Ferro Cordeiro, Domingos Tsuji, Allex Itar Ogawa, Maysa Tibério Ubrig, and Márcia Helena Moreira Menezes. Grbas and cape-v scales: high reliability and consensus when applied at different times. *Journal of Voice*, 26(6):812–e17, 2012.

Elizabeth A Peterson, Nelson Roy, Shaheen N Awan, Ray M Merrill, Russell Banks, and Kristine Tanner. Toward validation of the cepstral spectral index of dysphonia (csid) as an objective treatment outcomes measure. *Journal of Voice*, 27(4):401–410, 2013.

R Core Team. The R project for statistical computing, version 2.15.2 (2012-10-26). Computer program : http://www.r-project.org/, 1998–2012.

Rahul Shrivastav, Christine M Sapienza, and Vuday Nandur. Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language, and Hearing Research*, 48(2):323–335, 2005.

A. Sprecher, A. Olszewski, J.J. Jiang, and Y. Zhang. Updating signal typing in voice: Addition of type 4 signals. *The Journal of the Acoustical Society of America*, 127(6):3710–3716, 2010.

I.R. Titze. Workshop on Acoustic Voice Analysis: Summary Statement. pages 1–36. Denver, CO: National Center for Voice and Speech, 1995.

C. J. van As-Brooks, F.J. Koopmans-van Beinum, L.C.W. Pols, and F.J.M. Hilgers. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *Journal of Voice*, 20(3):355–368, 2006.

C.D.L. Van Gogh, J.M. Festen, I.M. Verdonck-de Leeuw, A.J. Parker, L. Traissac, A.D. Cheesman, and H.F. Mahieu. Acoustical analysis of tracheoesophageal voice. *Speech Communication*, 47(1):160–168, 2005.

R. J. J. H. van Son. NKI TE-VOICE Analysis tool (TEVA). Computer program: http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/NKIcorpora/NKI_TEVA/, 2012.

Floris L Wuyts, Marc S De Bodt, and Paul H Van de Heyning. Is the reliability of a visual analog scale higher than an ordinal scale? an experiment with the grbas scale for the perceptual evaluation of dysphonia. *Journal of voice*, 13 (4):508–517, 1999.

Ping Yu, Joana Revis, Floris L Wuyts, Michel Zanaret, and Antoine Giovanni. Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics (IALP)*, 54(6):271–281, 2001.

Richard I Zraick, Gail B Kempster, Nadine P Connor, Susan Thibeault, Bernice K Klaben, Zoran Bursac, Carol R Thrush, and Leslie E Glaze. Establishing validity of the consensus auditory-perceptual evaluation of voice (cape-v). *American Journal of Speech-Language Pathology*, 20(1):14–22, 2011.

7

Computing scores of voice quality and speech intelligibility in tracheoesophageal speech for speech stimuli of varying lengths [0]

Abstract

In this paper, automatic assessment models are developed for two perceptual variables: speech intelligibility and voice quality. The models are developed and tested on a corpus of Dutch tracheoesophageal (TE) speakers. In this corpus, each speaker read a text passage of approximately 300 syllables and two speech therapists provided consensus scores for the two perceptual variables. Model accuracy and stability are investigated as a function of the amount of speech that is made available for speaker assessment (clinical setting). Five sets of automatically generated acoustic-phonetic speaker features are employed as model inputs. In Part I, models taking complete feature sets as inputs are compared to models taking only the features which are expected to have sufficient support in the speech available for assessment. In Part II, the impact of phonetic content and stimulus length on the computer-generated scores is investigated. Our general finding is that a text encompassing circa 100 syllables is long enough to achieve close to asymptotic accuracy.

[0] R.P. Clapham, JP. Martens, R.J.J.H. van Son, F.J.M. Hilgers, M.W.M. van den Brekel, C. Middag. *Computer Speech & Language, 37, 2016, p1-10.*

7.1 Introduction

The ability to generate automatically computed scores for perceptual variables such as speech intelligibility and voice quality is a relatively recent development in the area of automatic speech and voice evaluation. An advantage of computer-generated scores is that they are not susceptible to extraneous factors, such as listener familiarity with the speaker and differences in internal anchors. In the clinical setting, computer-generated scores can be a valuable adjunct to subjective methods of assessment, especially if the evaluation is part of a therapy outcome measurement. In fact, prior knowledge of whether a recording is pre-therapy or post-therapy does not influence computed scores as it does with listeners (Ghio et al., 2013) and there is no inter-rater variation for computed scores as there is when perceptual scores are provided by different clinicians.

Computer-generated scores of perceptual variables have predominately been limited to research studies with a focus on developing assessment models, but the methodology is slowly making its way to evaluation studies as a dependent variable (Mayr et al., 2010; Stelzle et al., 2011; Windrich et al., 2008). In most cases, researchers have used speech recordings from existing databases that encompass readings of phonetically balanced texts (e.g. German *Der Nordwind und die Sonne* used in Mayr et al. 2010 and Windrich et al. 2008). In perceptual evaluation of speech intelligibility, some assessments have been developed so that the phonetic material reflects the phoneme frequencies one would expect to measure in long texts from the target language (see review article by Miller, 2013). To our knowledge, the effects of speech stimulus length and phonetic composition on the computed scores has not yet been investigated. There is, however, evidence that improved automatic binary classification (healthy control speakers vs speakers with dysarthria) benefits from more speech material (Bocklet et al., 2013).

The stimulus length varies between research institutes and hospitals as a result of differences in protocol, speaker characteristics (e.g. patient is unable to read the entire text due to reading skills, fatigue or underlying pathology) or both protocol and speaker characteristics. The speech material used across studies within the same institute can also vary and developing distinct assessment models for the various speech materials available is not possible. This motivated us to investigate the impact of phonetic variety and stimulus length on the outputs of automatic assessment models.

The present paper extends our previous work on assessment models for speech and voice quality for speakers treated for head and neck cancer (Clapham et al., 2014; Middag et al., 2014). Where the focus of our previous work was on developing models that perform at a level comparable to that of a human listener when given a sufficiently large amount of speech, the focus of the present work is on developing models that also offer reliable and stable results in a clinical setting where considerably less speech material per subject is available. The main goals are thus (1) to establish strategies for creating more robust models and (2) to offer insight into the minimum amount of speech material needed to attain accurate and stable computer-generated scores with these robust models.

In Section 7.2.1 we present the audio stimuli and perceptual evaluation data and describe how the various assessment models were created. We also discuss the methodology used to investigate phonemic variation and model robustness (Part I) and the influence of stimulus length and phonetic composition (Part II). Results from the two experiments are separately listed in the Results section and are discussed as a whole in Section 7.4.

7.2 Method

7.2.1 Audio stimuli

All audio recordings were collected at the Netherlands Cancer Institute (Amsterdam, the Netherlands) as part of previous research studies. As the recordings were gathered over a period of more than 10 years, the recording conditions are partly unknown and most likely differed across studies. Digital audio recordings were re-converted into digital form and all recordings were then standardized (sampling frequency of 44.1 kHz, 16-bit linear PCM).

There were recordings of 81 Dutch TE speakers (70 males, 11 females) and all speakers provided informed consent at the time of recording, allowing the recordings to be used for research purposes. Although multiple recordings existed for many speakers, only one recording per speaker (the earliest one) is included in the present study.

All speakers used indwelling voice prostheses (Provox) and read a Dutch text (*80 dappere fietsers*) of neutral content, meaning that the text did not evoke any emotions. The text was divided into six sentences and the average sentence length is 25 words (SD = 12, range 13-47) or 47 syllables (SD = 23, range 28-88). The text is not phonetically balanced because the recordings stem from research studies which did not require such a balance.

Since we want to study the effects of stimulus length (in syllables), we decided to divide the text into text fragments of almost equal lengths. In a first step we subdivided the longest sentences into two parts by cutting them at a position where a prosodic boundary can be expected. This way we got nine text parts some of which were still too short. In a second step, we therefore merged two short parts into one text fragment. The end result was a set of six text fragments of approximately 50 syllables each (mean 47, range 35-54). The upper part of Table 7.1 illustrates how these text fragments relate to the original sentences. Of the 40 Dutch phonemes, 27 appear in all six fragments.

7.2.2 Perceptual evaluation

Stimuli

For the auditory-perceptual experiment, we manually extracted the second sentence from the read passage (16 words and 31 syllables) of all speakers as experimental items. We extracted the fifth sentence of 12 randomly chosen speakers as practice items, intended to acquaint the listener with the experimental procedure.

Experimental set-up

Each experimental item was scored by two speech language pathologists, both with extensive clinical experience in the area of TE speech. They individually evaluated the overall voice quality (with descriptors 'least similar to normal' and 'most similar to normal') and speech intelligibility (descriptors 'poor' and 'good') on a computerised version of a visual analogue scale in an online self-paced listening experiment. No tick marks were observable during the individual evaluation; the rater moved the cursor along the scale and the final cursor location was saved as a value from 0-1000. The latter led to a quantization step that is much smaller than the distinction a listener can make in a statistically confident way, and thus, small enough to justify an interpretation of the discrete value as a continuous score. No speaker information (i.e. gender, age, prosthesis type) was available to the raters.

Scores that differed by more than 125 points between raters were discussed and re-evaluated in a consensus round. In cases where individual scores were within 125 points (corresponds to scores being the same if the scale was converted into a 4-point ordinal scale) the mean score was considered as the consensus score. To aid scoring in the consensus round, major (10% of scale distance) and minor (5% scale distance) tick marks were shown on the scale together with the two individual scores.

Rater agreement and reliability

Before the consensus round, 32 (38%) of the voice quality scores and 46 (54%) of the speech intelligibility scores were within 125 points of each other. The strength of the correlation between the rater's individual evaluations was significant but low for voice quality (Pearson's correlation coefficient, PCC = .47, $p < .001$) and adequate for speech intelligibility (PCC = .62, $p < .001$). The intra-class correlation coefficient (for a two-way consistency model) for the variables was fair for voice quality (ICC = .42, $p < .001$) and good for speech intelligibility (ICC = .62, $p < .001$) (Cicchetti, 1994).

7.2.3 Automatic evaluation tools

Automatic evaluation involves three stages of processing: (1) an acoustic front-end analysis describing the energy and shape of the spectrum of Hamming windowed segments of 25 ms shifted over 10 ms time steps, (2) an analysis of the acoustic information generating global acoustic-phonetic features that characterise the speaker (termed 'speaker features') and (3) a prediction of the perceptual variable by means of a regression model. As in our previous work (Middag et al., 2014), we employ an ensemble linear regression model per perceptual variable. Such a model computes the mean of 50 scores generated by 50 small linear regression models. Each small linear model computes the weighted sum of a couple of selected input features and a bias. The features to select and the weights to use are learned on a small randomly selected subset of the training samples. Since an exhaustive search for the best feature subset is computationally prohibitive, the selection strategy works as follows: (a) retain the feature triplets offering the highest model accuracies by means of an exhaustive search, (b) extend each retained feature set by adding the feature inducing the largest gain in accuracy and (c) repeat this until the accuracy of the best feature set saturates or starts to degrade. The model accuracies follow from cross-validation tests on the training subsets.

In the present study, the model inputs are speaker features and these features are organized into five feature sets. We now briefly introduce these feature sets and refer to our previous publications for more details.

Phonological and phonemic features

To derive these features, an automatic speech recognizer matches the acoustic information with the phonetic transcription of the speech via a process of forced speech-to-text alignment. Two types of speaker features can be extracted: phonological features (PLFs) and phonemic features (PMFs). We employ 24 binary phonological properties reflecting manner of articulation (e.g. "burst"), place of articulation (e.g. "bilabial") and voicing (e.g. "voiced"). Each property is either present or absent in the signal, meaning that there are 48 PLFs available to characterize a Dutch speaker: 24 positive and 24 negative features. A high positive PLF indicates that a particular property was present in the intervals it should have been present. A low negative PLF indicates that a particular property was not present in the intervals where it was not supposed to be present.

The PMFs reflect how well phonemes such as /s/, /z/ or /A/ are realized by the speaker. From the likelihoods of the different phonemes in a particular frame one can estimate the posterior probabilities of these phonemes in that frame. The mean of the posterior probabilities of a particular phoneme in the frames aligned with that phoneme is a positive PMF corresponding to that phoneme. A particular PMF thus reflects how well the acoustic properties of a particular phoneme are found in the intervals where that phoneme was uttered. Dutch has 40 phonemes, and therefore, there are 40 PMFs available for characterizing a Dutch speaker.

Alignment-free phonological and alignment-free phonemic features

It is also possible to analyze speech without considering its phonetic transcription. Such an analysis does not involve any speech-to-text alignment and is, therefore, termed 'alignment free'. Alignment-free phonological features (ALF.PLFs) provide information about the phonological properties of the speech signal. As explained in Middag et al. (2010), we use a slightly different phonological feature extractor with 25 instead of 24 phonological outputs in this stage. These phonological outputs are individually analyzed as a function of time (for details see Middag et al., 2010). Per property, this analysis yields 12 features such as "mean value", "mean value of the peaks" and "steepness of the peak onsets". This way, 300 ALF.PLFs (= 25 * 12) are created to characterize the speaker.

In a similar vein, one can also compute ALF.PMFs which provide information about the phonetic properties of the speech signal. Here, a distinction is made between 55 phones (the 40 traditional phonemes plus 6 closures, 6 bursts, 1 glottis and 2 silence symbols) and per phone, analyzing its posterior probability as a function of time now generates six statistical measurements, so that in total 330 ALF.PMFs (6 * 55 = 330) are created to characterize the speaker.

Pitch and voicing related features (AMPEX)

The AMPEX feature extractor generates eight acoustic parameters by means of a built-in auditory model developed by Van Immerseel and Martens (1996). It extracts both voicing-related features (e.g. proportion of voiced frames) and pitch-related features (e.g. jitter). The created features have already been proven successful for the assessment of pathological speech (Clapham et al., 2014; Moerman et al., 2004, 2015). The AMPEX feature extractor is freely available and can be downloaded from the website of the ELIS department at Ghent University.

7.2.4 Part I: Phonemic variation and model robustness

We compare performances of voice quality and speech intelligibility models that have access to an individual feature sets or combinations of feature sets.

Speech material and sampling procedure

The models are trained and validated using a 5-fold cross validation strategy. A difference with our previous studies (Clapham et al., 2014; Middag et al., 2014) is that we now set aside 1/5 of the stimuli (16 speakers) as test data to be used in Part II of our present study. The remaining stimuli (64 speakers) are divided into five folds: four of which are designated as training data and the remaining one as validation data.

Model inputs

We investigate the performances of full-set models that have access to one or more (up to three) complete speaker feature sets as model inputs and reduced-set models that have only access to preselected features from these feature sets.

As each phonetic feature corresponds to a particular sound (phone or phoneme) and each phonological feature to a set of sounds (both the positive and the negative feature of a sound set is supposed to correspond to the same sound set) it can be characterized by a frequency of occurrence of this sound (set). If a certain sound is not uttered in the analyzed text, the acoustic analysis cannot come up with values for the features corresponding to that sound. In the case of full-set models, we then replace such features by their mean value observed in a big sample of normal speech.

In the case of reduced-set models we only consider features whose expected frequency of occurrence in Dutch (as derived from the phoneme frequencies found in spoken Dutch as reported in Luyckx et al., 2007) exceeds a threshold of 5%. This meant that

1. we retained only six PMFs (the average posterior probabilities of /@/, /A/, /d/, /n/, /r/ and /t/) and the ALF.PMFs derived from these six PMFs (e.g. "percentage of frames in which /n/ is the phoneme with the highest posterior probability", "standard deviation of the posterior probability of /n/"),

2. we kept all PLFs except the ten (5 positive and 5 negative) corresponding to the properties 'approximant', 'lateral', 'labio-dental', 'glottal' and 'high', and

3. we expelled the ALF.PLFs that were derived from the phonological properties ‘nasal vowel’, ‘labio-dental’, ‘glottal’ and ‘palatal’.

None of the AMPEX features were expelled as voicing and pitch features are always supported by a sufficient amount of speech.

Performance measures

The primary measure of model performance is the root mean squared error (RMSE). It is defined as the square root of the mean of the squared differences between the predicted (computed) scores and the consensus (perceptual) scores. The goal is to attain a low RMSE. The Pearson Correlation Coefficient (PCC) between the two scores is used as a secondary performance measure, and the goal is to achieve a high PCC.

The Wilcoxon Signed Ranks test is used to establish whether one model significantly outperforms a competitor model. Here it is used to investigate whether the baseline full-set model performance differs significantly from that of the best reduced-set model. A Bonferroni correction for multiple comparisons was used and a conservative p-value ($p < .005$) was deemed statistically significant.

7.2.5 Part II: Influence of stimulus length and composition

In order to understand why the model scores can be sensitive to the length and the composition of the test stimulus, we recall that many of the speaker features (e.g. components of PMF and PLF) are of the type “average posterior probability of a particular phonological class (either a phone such as /s/ or a phonological class such as ‘plosive’) in the speech intervals realizing a phone of that class". During model development, we usually employ as much speech material per speaker as possible and means measured on that material are thus bound to approximate the true means for that speaker. During test, however, we want to use short stimuli to reduce the measurement time. In that case, a mean over the test material can significantly deviate from the “true" mean. Moreover, since the relative frequencies of occurrence of the infrequent phonemes of the language strongly depend on the choice of the text, it follows that the number of intervals supporting a particular feature also strongly depends on the text.

In principle, the same reasoning also applies to the alignment-free features, that is, there are two phenomena that can affect the computer-generated scores: variations in the phonetic composition of the text and variations in the length of the text (in phonemes or syllables). If we only consider randomly chosen text fragments, the effects of both phenomena will be strongly correlated as they both give rise to an effect that is expected to be inversely proportional to the square root of the text length. The two experiments we conceived attempt to isolate the two phenomena as much as possible.

Models and speech material

We predict consensus scores of voice quality and speech intelligibility with the best-performing models identified in Part I. These models were developed using the readings of complete paragraphs, but are here used to compute scores from speech samples of different lengths.

The test material in this Part is the material that was set aside in Part I. As stated in Section 7.2.1, the paragraph was divided into six text fragments (F1-F6) of approximately 50 syllables each. Per speaker, we compute scores for all possible stimuli we can construct by combining one up to five text fragments: 6, 15, 20, 15 and 6 stimuli containing 1, 2, 3, 4, and 5 text fragments respectively. The number of fragments in a stimulus is referred to as the stimulus length (in text fragments). Table 7.1 lists the individual fragments and examples of fragment combinations of different lengths.

Text	
Parts:	P1 P2 P3 P4 P5 P6 P7 P8 P9

Fragments:	F1(P1); F2(P2+P3); F3(P4+P8); F4(P5);
	F5(P6); F6(P7+P9)

Combining Fragments	
Single fragment (n=6)	F1; F2; . . . ; F6
2 fragments (n=15)	F1F2; . . . ; F5F6
3 fragments (n = 20)	F1F2F3; . . . ;
4 fragments (n = 15)	F1F2F3F4; . . . ;
5 fragments (n = 6)	F1F2F3F4F5; . . . ;

Table 7.1: Illustration of how the nine text parts were recombined into six text fragments of comparable lengths (upper part of table) and of how the six text fragments can be employed to create stimuli of varying lengths (lower part of table).

Score processing

Per speaker and per stimulus length, we measure the standard deviation (*SD*) of the scores of all stimuli of that length. The SDs reveal the effect of the phonetic composition on the computed scores for a stimulus of that length. Obviously, we would not have been able to estimate the SD for a stimulus length of six as there is only one stimulus consisting of all six text fragments. This explains why this stimulus length was not included in the experiment. It is important to note here that stimuli of length 2 or larger are not independent of each other as they always share at least one text fragment with another stimulus. However, in spite of this, the measured SDs are bound to provide useful information on the effect of phonetic composition as a function of stimulus length.

In order to assess the effect of the length under the assumption that phonetic variation could be eliminated per length (e.g., by carefully choosing texts for which the phoneme frequencies are identical), we consider the mean of the scores of all stimuli of a particular length provided by a particular speaker as the computer generated speaker score. Using these scores we then compute the RMSE and PCC for that length.

7.3 Results

7.3.1 Part I: Phonemic variation and model robustness

Voice quality

Table 7.2 lists the performances of the 10 best full-set models and the corresponding reduced-set models when applied to full paragraphs. The full-set models PMF+AMPEX (RMSE = 122.2) and PMF+AMPEX+ALF.PLF (RMSE = 122.2) attain the highest accuracy, but the differences across models are small: only 2 of the 10 models of the same type (full-set/reduced-set) perform significantly worse (Wilcoxon, $p < .005$) than the best model of that type. There are no statistically significant differences between the full-set and the reduced-set models of a particular combination of feature sets.

Observe that a combination of three feature sets (e.g. PMF+AMPEX+PLF) does not necessarily lead to a better model than a combination of only two of these feature sets (e.g. PMF+AMPEX), despite the fact that the model found in the latter case is an example of an eligible model that could have been found in the first case. This observation reveals the sub-optimality of the feature selection process incorporated in the regression model training. The degree of sub-optimality is expected to increase with the number of features.

Speech intelligibility

For speech intelligibility, the full-set PMF+AMPEX+PLF model is the strongest performing model (RMSE = 97.4) but the strongest performing reduced-set PMF model (RMSE = 98.6) is not far behind (see Table 7.3). Nevertheless, since the PMF+AMPEX model comes very close to the best model in the two conditions and since it was also the chosen model for voice quality, we will consider PMF+AMPEX as the baseline feature set combination in Part II. As seen for voice quality, only 2 of the 10 competitors of the same type perform statistically worse than the baseline and there are no statistically significant differences between the reduced-set and the full-set models for a given feature set combination.

Feature sets	Full set		Reduced set	
	RMSE	PCC	RMSE	PCC
PMF	122.7	0.66	126.3	0.61
PMF+AMPEX	122.2	0.66	123.5	0.64
PLF+PMF	125.2	0.64	130.0	0.58
PMF+ALF.PLF	127.6	0.63	124.8	0.64
PMF+ALF.PMF	129.1	0.58	130.8	0.58
PLF+ALF.PLF	138.9 **	0.53	137.3 **	0.54
PMF+AMPEX+ALF.PLF	122.2	0.66	124.1	0.65
PMF+AMPEX+PLF	124.5	0.64	128.9	0.59
PMF+AMPEX+ALF.PMF	127.3	0.63	129.9	0.58
PLF+ALF.PLF+AMPEX	138.9 **	0.53	136.8 **	0.54

*Table 7.2: Performances (PCC reported to two decimal places and RMSE reported to one decimal place) of the best 10 full-set models and the corresponding reduced-set models for voice quality. Also indicated is whether a result is is statistically worse (** $p < 0.005$) than the best result in the column.*

Feature sets	Full set		Reduced set	
	RMSE	PCC	RMSE	PCC
PMF	98.3	0.67	98.6	0.66
PLF+PMF	97.8	0.67	100.9	0.64
PMF+AMPEX	98.8	0.67	100.0	0.67
PMF+ALF.PLF	100.9	0.66	114.4	0.62
PMF+ALF.PMF	101.3	0.65	102.4	0.63
PLF+AMPEX	117.3 **	0.44	116.1 **	0.45
PLF+ALF.PMF	117.3 **	0.45	123.6 **	0.40
PMF+AMPEX+PLF	97.4	0.67	100.2	0.65
PMF+AMPEX+ALF.PLF	98.8	0.67	110.7	0.63
PMF+AMPEX+ALF.PMF	100.8	0.66	102.5	0.63

*Table 7.3: Performances (PCC reported to two decimal places and RMSE reported to one decimal place) of the best 10 full-set models and the corresponding reduced-set models for speech intelligibility. Also indicated is whether a result is is statistically worse (** $p < 0.005$) than the best result in the column.*

7.3.2 Part II: Effects of stimulus composition and length

Influence of phonetic composition

To investigate the influence of phonetic composition on the computed scores, we considered the reduced-set PMF+AMPEX model. Per perceptual variable, per speaker and per stimulus length, we record the SD of the scores across stimuli. The statistics of these SDs across speakers are listed in Table 7.4. The mean SD clearly decreases with an increasing stimulus length, but the SD-range only seems to decrease for speech intelligibility and not for voice quality.

Influence of stimulus length

To isolate the influence of stimulus length on the reliability of the speaker features, we consider, per variable, speaker and stimulus length, the mean score found across stimuli. Figure 7.1 shows the RMSE and PCC obtained by comparing these means to the consensus scores. For both perceptual variables, the accuracy improves (RMSE decreases and PCC increases) when the test stimulus is longer. The improvement is significant and close to 10% when going from 47 syllables (1 text fragment) to 94 syllables (2 text fragments). The improvement caused by adding a third text fragment is not statistically significant anymore.

We also inspected the speaker scores generated for single text fragments. We found that for both perceptual variables, fragment F4 consistently gave rise to low RMSE and high PCC values (RMSE = 144.7 for voicing and RMSE = 106.19 for speech intelligibility). When two fragments are considered, the best combination is F2F4 for both variables (RMSE = 134 for voice quality and RMSE = 96 for speech intelligibility). It happens that F4 comprises a high number of distinct phonemes and syllables and a number of shorter phrases (Dutch: "... en hielp gedurende vijf dagen mee bij het plakken van banden, het maken van gebroken kettingen, het verzorgen van slaapgelegenheid en het op-

Fragment combination	n	Voice quality		Speech intelligibility	
		Mean	Range	Mean	Range
Single fragment	6	62	16 - 176	56	18 - 145
2 fragments	15	39	10 - 182	37	14 - 104
3 fragments	20	30	7 - 225	27	9 - 90
4 fragments	15	23	5 - 261	20	8 - 73
5 fragments	6	15	2 - 198	12	3 - 41

Table 7.4: Mean SD and SD-range per perceptual variable of the speaker scores per combination of text fragments.

sporen van verkeerd gereden deelnemers"; English: "... and helped for five days repairing tubes, fixing broken chains, organising accomodation and tracking down lost participants"). From a modeling perspective, high phonetic variety may lead to good model accuracy. From a speaker perspective, shorter phrases may work to the advantage of TE speakers as the syntactic structure allows inhalation at appropriate boundaries. In other words, possible phonetic variety is easier to realize.

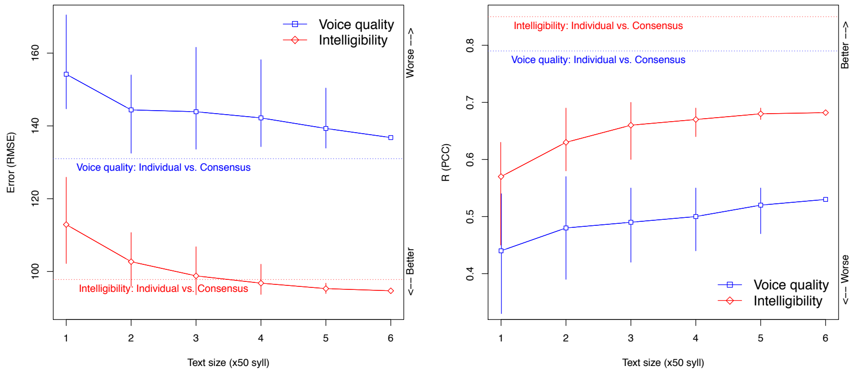


Figure 7.1: Accuracy of the mean score (squares/diamonds) and range (vertical lines) across groups (RMSE and PCC) for voice quality and speech intelligibility as a function of the number of fragment combinations (= text size). For comparison, we include the average RMSE and PCC values for the individual SLPs versus the consensus scores (horizontal lines).

7.4 Discussion

The first objective of this study was to investigate if and how assessment models designed to predict human ratings of voice quality and speech intelligibility of tracheoesophageal (TE) speech degrade when less speech material is available for making the predictions. This question addresses the boundary conditions under which current technologies can be applied in clinical practice (cf., the studies by Clapham et al., 2014, 2015; Mayr et al., 2010; Middag et al., 2014; Stelzle et al., 2011; Windrich et al., 2008). The second objective was to check whether ignoring input features that are insufficiently supported by the speech material leads to models that are less sensitive to variations in the phonetic content of that material.

We investigated the second objective first. In fact, if we could show that ignoring input features does not degrade model performance in the case of sufficient speech material, we could restrict the study of the first objective to an analysis of the scores emerging from the best reduced-set model. Based on our earlier conjecture that the observed frequencies of infrequent phonemes of a language may differ significantly between texts of the same length, we investigated whether it is possible to reduce the sensitivity to that source of variation by prohibiting the model training to access speaker features derived from utterances of such infrequent phonemes.

To begin with, we created five sets of speaker features that can serve as inputs to the envisioned assessment models. Using the different feature sets we then trained assessment models towards consensus ratings of two perceptual variables. Recall that these ratings can take values from 0 to 1000. We trained models that had access to one, two or three feature sets because previous studies (Clapham et al., 2014; Middag et al., 2014) had shown that combining feature sets generally results in stronger models.

We considered two conditions for the training. In the full-set condition, the models had access to all features of a feature set while in the reduced-set condition, they only had access to features referring to a sound (set) with a sufficiently high frequency of occurrence in Dutch. Note that the linear regression model training automatically determines which and how many eligible features it incorporates. Consequently, the number of model parameters is not necessarily proportional to the number of eligible features.

Comparing corresponding full-set and reduced-set models, led to the conclusion that the performance differences between both model types are not statistically significant. This means that expelling features does not hurt even when the test material (all text fragments of the speaker in this case) is long enough and matched to the length and phonetic content of the training material. In both conditions, the PMF+AMPEX and the PMF+AMPEX+ALF.PLF models attained the best voice quality models. For speech intelligibility, PMF+AMPEX+PLF model was the best (RMSE = 97.4) in the full-set condition whereas in the reduced-set condition, it was the PMF model (RMSE = 98.6). However, for both perceptual variables the best models were not statistically better than most other models. Taking all results into account, we selected the reduced-set models built on the PMF+AMPEX feature set combination as the baseline models for investigating our first objective. Note that the PMF features alone suffice to create good models, but the AMPEX features focusing on voicing and pitch stability do seem to offer a small improvement which is not so surprising given that TE speakers have difficulties in this respect (Clapham et al., 2015).

Clearly, we expect that longer test stimuli give rise to more reliable model predictions. To assess how much the phonetic **composition** of the text influences the scores we measured the SD of the scores emerging from different stimuli of a given length provided by the same speaker. To assess how stimulus

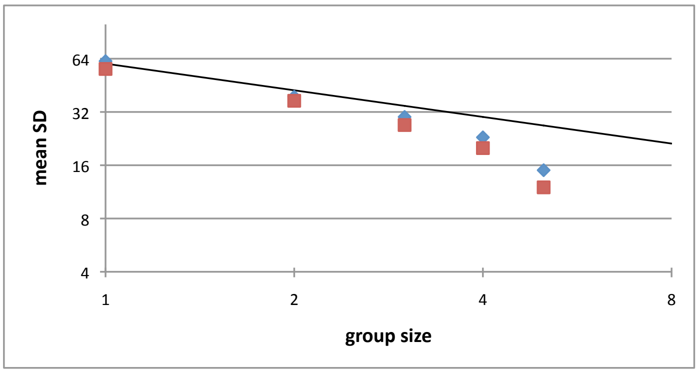


Figure 7.2: Mean SD of the speaker scores as a function of the stimulus length (in text fragments). The results for voicing (diamonds) and speech intelligibility (squares) are depicted together with the trend line SD = 60/$\sqrt{length}$

length influences the scores, we compared the mean of these scores with the consensus score for the speaker.

The first result we can derive from Tables 7.2 - 7.4 is that for a stimulus length of about 50 syllables, the impact of the phonetic composition is substantial. The mean SD is equal to about 50 - 60% of the expected error made by the model (compare an SD of 62 to an RMSE of 122.2 and an SD of 56 to an RMSE of 98.8).

Plotting the mean SD against the stimulus length in a log-log-plot (see Figure 7.2) reveals that in the beginning, the descent follows the trend that SD is inversely proportional to the square root of the stimulus length (trend line in the figure) whereas it is larger for larger stimulus lengths. As mentioned before, two stimuli composed of multiple text fragments share at least one text fragment. In fact, the longer the stimulus (in fragments) the larger the percentage of text they are sharing and the more the observed SD is an under-estimation of the SD one would have obtained with measurements on independent stimuli of the same length. The latter explains the larger descent for larger stimulus lengths. Taking everything into account, we conjecture as a second result that the impact of the phonetic composition is bound to be inversely proportional to the square root of the number of syllables in the text: it would take 200 syllables to reduce the relative impact to 25-30% of the asymptotic RMSE (obtainable with a very long text).

From Figure 7.1 we conclude that the impact of the stimulus length under the assumption of equal phonetic composition is not that large: less than 15% relative for a length of 50 syllables (compare an RMSE of 154.2 to the asymptotic value of 136.8 and an RMSE of 112.9 to 94.7). As expected, this impact also appears to be inversely proportional to the square root of the stimulus length, as one can verify by plotting the RMSE as a function of the stimulus length in a log-log-plot.

In conclusion, voice quality and intelligibility prediction models that only have access to acoustic-based speaker features describing sufficiently frequent sounds or sound classes are robust with respect to the length of the speech material they are being tested on. When there is enough speech material available, such models are as accurate as similar models that have access to all speaker features, and, as a rule of thumb, they continue to yield accurate and stable scores for as long as the test material encompasses at least 100 syllables.

Acknowledgements

This research was supported in part by an unrestricted research grant from Atos Medical (Horby, Sweden), the Verwelius Foundation Huizen, the Netherlands) and 'Kom op tegen Kanker' the campaign of the Flemish League against Cancer foundation. The authors wish to thank Anna Kornman and Merel Latenstein for their assistance with the data collection.

References

Tobias Bocklet, Stefan Steidl, Elmar Nöth, and Sabine Skodda. Automatic evaluation of parkinson's speech - acoustic, prosodic and voice related cues. In *Interspeech*, pages 1149–1153, 2013.

D.V. Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290, 1994.

R Clapham, C Middag, F Hilgers, J-P Martens, M van den Brekel, and R van Son. Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer. *Speech Communication*, 59:44–54, 2014.

Renee P Clapham, Corina J van As-Brooks, Rob JJH van Son, Frans JM Hilgers, and Michiel WM van den Brekel. The relationship between acoustic signal typing and perceptual evaluation of tracheoesophageal voice quality for sustained vowels. *Journal of Voice*, 29(4):517.e23–517.e29, 2015. URL http://dx.doi.org/10.1016/j.jvoice.2014.10.002.

A Ghio, J Revis, S Merienne, and A Giovanni. Top-down mechanisms in dysphonia perception the need for blind tests. *Journal of Voice*, 27(4):481–485, 2013.

K Luyckx, H Kloots, E Cousse, and S Gillis. Klankfrequenties in het nederlands. In D Sandra, editor, *Tussen Taal, Spelling en Onderwijs. Essays bij het emeritaat van Frans Daems*, pages 145–154. Gent: Academia Press, 2007.

S Mayr, K Burkhardt, M Schuster, K Rogler, A Maier, and H Iro. The use of automatic speech recognition showing the influence of nasality on speech intelligibility. *Eur Arch Otorhinolaryngol*, 267(11):1719–1725, 2010.

C Middag, Y Saeys, and J-P Martens. Towards an asr-free objective analysis of pathological speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 294–297, Tokio, Japan, 2010.

C Middag, R P Clapham, R van Son, and J-P Martens. Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. *Computer Speech and Language*, 28(2):467–482, 2014.

Nick Miller. Measuring up to speech intelligibility. *Int J Lang Commun Disord*, 48(6):601–612, Nov-Dec 2013. ISSN 1460-6984 (Electronic); 1368-2822 (Linking). doi: 10.1111/1460-6984.12061.

Mieke Moerman, Glenn Pieters, Jean-Pierre Martens, Marie-Jeanne Van der Borgt, and Phillippe Dejonckere. Objective evaluation of the quality of substitution voices. *Eur Arch Otorhinolaryngol and Head & Neck*, 261(10): 541–547, 2004.

Mieke Moerman, Jean-Pierre Martens, and Philippe Dejonckere. Multidimensional assessment of strongly irregular voices such as in substitution voicing and spasmodic dysphonia: A compilation of own research. *Logopedics Phoniatrics Vocology*, 40(01):24–29, 2015.

F Stelzle, A Maier, E Nöth, T Bocklet, C Knipfer, M Schuster, FW Neukam, and E Nkenke. Automatic quantification of speech intelligibility in patients after treatment for oral squamous cell carcinoma. *J Oral Maxillofac Surg*, 69 (5):1493–1500, May 2011.

Luc Van Immerseel and Jean-Pierre Martens. Pitch and voiced/unvoiced determination with an auditory model. *J Acoust Soc Am*, 91(6):3511–3526, 1996.

M Windrich, A Maier, R Kohler, E Noth, E Nkenke, U Eysholdt, and M Schuster. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. *Folia Phoniatr Logop*, 60(3):151–6, 2008. doi: 10.1159/000121004.

8 General discussion

Abstract

This dissertation is the result of a desire for an automated, objective method to support a Speech Pathologist's auditory-perceptual evaluation of voice and speech. We investigated the use of automatic tools to predict speech intelligibility and voice quality scores for two cohorts of speakers treated for head and neck cancer. At the project outset, two tools that use speech technology to model auditory-perceptual scores were identified. Extending one of these tools, we have developed speech intelligibility and voice quality prediction models for the two speaker cohorts. Overall, the models attained performance levels similar to that of a group of listeners. The results highlight that with refinement of the models, the clinical application of the tool to *support* a clinician's perceptual evaluation is possible. In this chapter we summarise our findings and discuss what we have learnt about developing prediction models within the context of the speech pathology practice.

8.1 Study objective

Auditory-perceptual evaluation is a frequently used method of evaluation in the Speech Pathology clinic despite its drawbacks in terms of listener variability and listener bias (Kent, 1996). Unlike perceptual evaluation, computer-generated perceptual scores should only vary upon re-analysis if the system's underlying acoustic models or prediction models are changed. Computer-generated evaluation has the potential to be performed with a similar level of ease as perceptual evaluation. We believe there is a role for speech technology in the clinical setting and this belief has shaped our vision of computer-generated "perceptual" scores being used alongside a clinician's auditory-perceptual assessment. This thesis explores whether, and how, existing automatic evaluation tools could be adapted to predict voice quality and speech intelligibility of Dutch speakers with head and neck cancer.

Our literature review in **Chapter 1** highlighted that strong correlations had been reported between automatically derived scores (e.g. word accuracy rate) and observed scores and that promising prediction models had been developed. We hypothesised that a tool already developed for Flemish speakers to predict transcription-based speech intelligibility scores for consonant-vowel-consonant stimuli could be adapted for our study objective. This required expanding the handling capability of the tool on a speech stimuli level (running speech/connected words), type of predicted variable (scale based) and development of prediction models for new variables (e.g., voice quality). It was also unknown how the Flemish-trained system would perform if provided Dutch speech material.

8.1.1 Study aims

The primary aim was to adapt existing tools for our outcome variables (speech intelligibility and voice quality), speaker groups (two groups of speakers treated for head and neck cancer) and the recorded material available (Dutch passage read by Dutch speakers). Our hope was to create models that could predict these variables as a function of various input features (e.g., phonological features) and attain model performance levels comparable to an average listener. In this way, computer-derived scores could reflect a rating provided by an additional listener. A secondary aim was to explore the relationship among acoustic measures, categorizations of acoustic information according to signal type and evaluations of voice quality for vowel level stimuli. See Table 8.1 for a summary of the speaker groups and investigations.

8.1.2 Research strategy

For our two speaker groups we investigated how various systems could be used individually or combined to achieve reliable prediction models for running speech. We explored the use of up to five feature sets as inputs for prediction models. These five feature sets can be categorised according to alignment strategy (forced alignment or alignment free and the type of information (monophone, phonological, acoustic). Figure 8.1 summarizes the five

Stimuli	Variable	Speaker group CCRT	 TL
Running speech	Speech intelligibility	X	X
	Voice quality	X	X
	Articulation	X	
	Accent	X	
Sustained vowel	Voice quality		X
	Acoustic signal typing		X

CCRT: concomitant chemoradiotherapy; TL: total laryngectomy

Table 8.1: Speaker groups and investigated variables

feature sets. The four alignment free and forced alignment feature sets were extended from the existing Flemish tool.

Our primary variables were speech intelligibility and voice quality but we extended our models to include articulation and accent given the strong link between articulation and speech intelligibility and the accent variations present in the Netherlands.

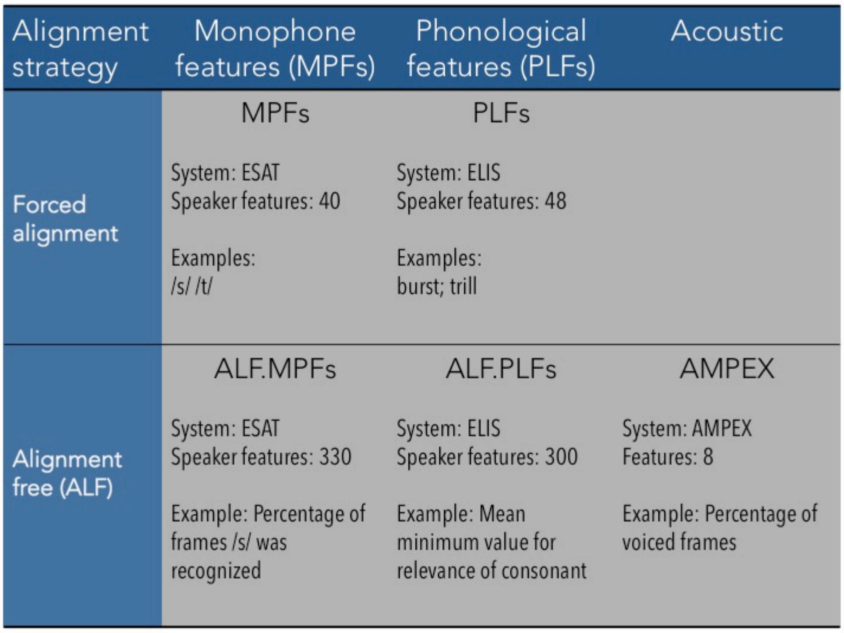
For sustained vowel stimuli, we completed exploratory work into the relationships between acoustic measures, signal typing and voice quality for speakers after total laryngectomy.

8.2 Chapter summary by speaker group

8.2.1 CCRT speaker group (Part I)

Chapter 2 In this chapter we described the recordings, perceptual data and annotations used in our studies. The speech recordings were collected by van der Molen et al. (2012) as part of a clinical trial into preventative rehabilitation on speech and swallowing after concomitant chemoradiotherapy (for full details on speakers, treatment and interventions see van der Molen et al., 2012). Part of this speech material included recordings of a 189-word passage from a Dutch fairy tale read by speakers before treatment (n=54) and at 10-weeks and 12-months post treatment (n=48 and n=39, respectively).

To develop a group-specific prediction model, we collected scale-based ratings for variables speech intelligibility (7-point scale) and phonation (5-point scale) as well as several additional aspects (e.g., articulation and accent, both on a 5-point scale). A group of 13 semi-professional listeners evaluated two fragments of the recorded text. There was no effect of text fragment on speech intelligibility score. At a group level there was no effect of evaluation moment on intelligibility score.



Alignment strategy	Monophone features (MPFs)	Phonological features (PLFs)	Acoustic
Forced alignment	MPFs System: ESAT Speaker features: 40 Examples: /s/ /t/	PLFs System: ELIS Speaker features: 48 Examples: burst; trill	
Alignment free (ALF)	ALF.MPFs System: ESAT Speaker features: 330 Example: Percentage of frames /s/ was recognized	ALF.PLFs System: ELIS Speaker features: 300 Example: Mean minimum value for relevance of consonant	AMPEX System: AMPEX Features: 8 Example: Percentage of voiced frames

Figure 8.1: Overview of feature sets used in studies with running speech stimuli. Sets are listed by alignment strategy and feature type. Auditory model-based pitch extractor (AMPEX)

Chapter 3 Using the speech intelligibility perceptual data presented in Chapter 2, we started from the work by Middag et al. (2008, 2010) to investigate two questions:

1. Was a new alignment-free feature set less text dependent and less language specific than alignment-based feature sets?
2. Does the type of acoustic model (Flemish versus Dutch) in the underlying system impact prediction model accuracy?

Answering these questions was necessary to extend the tool to analysis of running speech of Dutch speakers. We also reported whether the top-performing model was able to track changes in speech intelligibility over time.

Comparing results from intelligibility prediction models built on Flemish or on Dutch acoustic models, we established that in general, the strongest performance occurs when the underlying acoustic models match that of the speaker. The alignment-based methods are sensitive to language whereas alignment-free methods attain comparable results regardless whether the underlying acoustic model was Flemish or Dutch. Comparing results emerging from prediction models trained on differing fragments of the text, we discovered that all feature sets are largely text-independent, at least in the absence of reading errors.

A model combining forced-alignment and alignment-free information attained stronger performance than performance with individual features and was capable of detecting progress/deterioration to a similar extent as a group of raters. On the basis of the results from this study, our future experiments only considered alignment-free and forced-alignment feature sets supported with underlying Dutch acoustic models.

Chapter 4 In a similar vein to the previous chapter, we developed prediction models for the perceptual variables articulation and phonation/voice quality. Our reasoning was that if these could be achieved, combining these models with the speech intelligibility model from Chapter 3 would mean an automatic tool capable of multidimensional 'perceptual-like' evaluation.

There is considerable articulatory-acoustic variation in the Netherlands due to aspects such as regional variations and language background. For the event that degree of accent could influence model performance for speech-related perceptual variables, we also explored a prediction model for the perceptual variable accent (perception of degree of accent compared to 'standard' Dutch). By including this perceptual variable, we envisage that clinicians can consider the computed accent score into account when interpreting computer-derived scores of speech intelligibility or articulation.

Our models could make use of monophone and phonological feature sets derived from forced-alignment or alignment-free processes and pitch and voicing outputs from the AMPEX pitch and voicing extractor (Figure 8.1). We anticipated that pitch and voicing features would be particularly suitable for the phonation/voice quality model. The tested models included single-feature models (e.g., only monophone features) and multiple-feature models that used two or three feature sets (e.g., monophone features + phonological features). Once the strongest articulation and phonation models were developed, we investigated the prediction model's success in identifying change in perceptual scores over time. We did this by considering the direction of change in perceptual score between the three evaluation momement, that is, recordings made before cancer treatment, at short-term follow-up and at long-term follow-up.

The results highlight that speaker features emerging from a forced alignment between the speech and the text as well as speaker features emerging from a plain analysis of the temporal evaluation of acoustic model outputs (i.e., alignment-free method) give rise to good assessment models of comparable accuracies. The models attain varying levels of success performing trend classification but there were no instances where a positive change in perceptual scores was classified as a negative change in computed scores, and vice versa.

8.2.2 Total laryngectomy speaker group (Part II)

Tracheoesophageal speech is a method of voice restoration after total laryngectomy. Tracheoesophageal speech involves a prosthetic device being placed in a surgically created fistula between the trachea and esophagus. When a speaker occludes the tracheostoma after inhalation, pulmonary air is redirected through the prosthesis where it passes and causes a new voicing source, called the neoglottis, to vibrate and create sound (see 1.1.2 for details). Many people develop functional alaryngeal speech after total laryngectomy, but voice quality is variable.

The audio stimuli used in Part II of this thesis were collected at the Netherlands Cancer Institute as part of various research studies over a 10-year period. The speech material includes recordings of the sustained vowel /a/ and a read text approximately 300-syllables/150 words long from 87 tracheoesophageal speakers. We collected scale-based ratings for variables voice quality (vowel and text stimuli) and speech intelligibility (text stimuli) and acoustic signal type categorizations according to the visual characteristics of the vowel spectograms. Computerized visual analogue scales and consensus-derived scores were used in these studies as opposed to equal appearing interval scale and averaged scores used in the CCRT studies.

Compared to Part I of this thesis, Part II is more exploratory in nature. In Chapter 5 and 6 we look at the use of acoustic signal typing as a correlate of voice quality and the association between acoustic measures and acoustic signal typing. In the remaining chapter we consider prediction models using the same speaker features discussed in Part I from the CCRT speaker group, only we focus on the effects of phonetic variety and the length of the spoken material rather than identifying which combination of speaker features produce the strongest prediction models.

Chapter 5 We took the opportunity to investigate the extent acoustic variables for vowels, such as harmonic-to-noise ratio, can be used to predict the consensus-derived categorization of the same vowel into one of four acoustic signal types. Although classification into signal type may be less subjective than ratings of voice quality because of the use of visual-based criteria, there remains a subjective component in the task. Our research interest led to the development of the TEVA computer program (van Son, 2012) which runs as an extension to the speech analysis program Praat (Boersma and Weenink, 2009). In this way, categorization of the spectrogram into signal type and extraction of acoustic measurements are achievable within a single computer application.

Although several acoustic measures are reported in the literature to correlate with acoustic signal type (see Table 5.1), we found measures reflecting the presence and duration of voicing (voicing fraction; maximum voicing duration) and measures reflecting the amount of noise (harmonic-to-noise ratio) were the most salient acoustic measures. On their own, voicing fraction and maximum voicing duration supply enough information to achieve a classification accuracy of a vowel into its acoustic signal type of above 60%. Including acoustic measures of the harmonic-to-noise ratio improved the classification rate to above 70%.

Chapter 6 Previous research reports a relationship between acoustic signal type and perceptual ratings of voice quality for two different speech materials, namely vowels and running speech for signal type and voice quality, respectively (D'Alatri et al., 2011; van As-Brooks et al., 2006). In this chapter we investigated whether this relationship held when only the sustained vowel was considered and what the rater-reliability was between signal type and perceptual scores.

We made use of two types of perceptual scores: continuous scores made on a computerized visual analogue scale and these scores converted to a 4-point ordinal score. The agreement between the two raters was higher for categorization into acoustic signal type than for ordinal scale scores of voice quality. Although there was a statistical relationship between signal type and voice quality, each signal type co-occured with a range of voice quality scores.

Chapter 7 In this chapter we turned our attention to the impact of the speech material, that is, the phonetic composition and length of the material, on prediction models. This is an important consideration as it could influence the selection of test material for future use of automatic predicted scores. We first considered model performance under two conditions: model development with access to all speaker-features as model inputs or model development with access to a sub-set of speaker-features as model inputs. The sub-set was restricted to features that one could expect to occur in Dutch. Results indicated no statistically significant difference in model performance between the conditions.

In the second part of this chapter we investigated how predicted scores vary depending on the length of the stimulus and its composition. We found that model stability generally improves as more speech material is available and performance is close to what is attainable when the speech material contains approximately 100 syllables.

8.3 Key findings

Our results can be summarised in eight key findings:

1 Prediction models generally achieve stronger performance when the system's underlying acoustic models match the speaker group (e.g. model for Dutch speakers utilizing acoustic models trained on Dutch speech) [Ch.2];

II Dual-feature models generally attain stronger performance results than single-feature models, however these differences may not be statistically significant [Ch.3, Ch.4, Ch.5, Ch.7];

III Combining alignment-free and alignment-based features generally leads to optimal model configuration [Ch.3, Ch.4] (nb: speech intelligibility model for speakers post total laryngectomy attains best performance with two forced-alignment speaker features;

IV Feature sets are largely text independent with performance close to what is attainable when the speech material contains approximately 100 syllables [Ch.2, Ch.7];

V Low perceptual scores, in particular for the perceptual variable voice quality, are difficult to model (Ch.3, Ch.4);

VI The models attain varying levels of success performing trend classifications, however there are no instances where a clear positive trend was classified as a clear negative trend [Ch.3, Ch.4];

VII Including pitch and voicing information in voice quality prediction models leads to improved model performance [Ch.4, Ch.7];

VIII In its current form, acoustic signal typing of the vowel /a/ provides limited information on the perception of voice quality for the same vowel [Ch.5, Ch.6].

8.4 Methodological considerations

In the sections below we outline some of the methodological considerations we encountered in our studies. The topics are discussed under three themes of auditory-perceptual, speech stimuli and automatic evaluation considerations. There is, however, considerable overlap among the themes.

8.4.1 Auditory-perceptual considerations

To develop prediction models for the two speaker cohorts, we considered it necessary to (i) have auditory-perceptual measures that allowed us to compare any speaker in a group to another and (ii) have auditory-perceptual scores for a corpus that were from the same group of listeners. To achieve this, the recordings from each corpus were re-evaluated by groups of listeners as the original scores for the CCRT cohort were derived from paired-comparisons of single speakers and the scores for the TL corpus were from varied listener protocols and evaluation methods.

8.4.1.1 Perceptual scale

There is little consistency across published studies regarding how quality of speech and voice is measured. The scoring methods vary according to the type of stimuli evaluated (e.g. sustained vowel, single words, running speech), the method of evaluation (e.g. transcription, scale), the scale type (e.g. visual analogue scale, Likert scale) and the features used on a scale (e.g. tick marks, anchors, polarity) (see reviews by Barreto and Ortiz, 2008; Kent, 1996; Miller, 2013).

Our motivation for using scaling measurement techniques was because it is a frequently used scale in auditory-perceptual studies (see reviews by Barreto and Ortiz, 2008; Miller, 2013), is often used to derive the target perceptual score in other prediction studies (e.g., Bocklet et al., 2012; Haderlein et al., 2007; Hattori et al., 2010). At the time the speech recordings were collected, standardized sentence intelligibility tests or established sets of semantically unpredictable sentences for evaluation by means of transcription were not available for Dutch speakers.

Although transcription-based scores are considered a closer representation of the construct speech intelligibility, transcription accuracy can increase as the length of the speech stimulus increases as the listener can make use of linguistic and prosodic information if the content is not controlled. To negate this, all listeners had access to the written transcripts of the connected speech samples. In this way, we endeavoured to collect speech intelligibility scores that represent an estimate of the listener's ease decoding the auditory signal and voice quality scores that reflect the listener's judgement of phonation quality.

Our investigation began with the CCRT corpus and we elected to use an ordinal scale (7-point scale for speech intelligibility and 5-point scale for other variables). Ordinal scales are a common method of data collection in computer modeling studies and in general auditory-perceptual studies. In the second part of our study, we changed our evaluation protocol to include perceptual ratings on a digitalized version of a visual analogue scale. This change was driven by a desire for continuous rather than quasi-continuous perceptual scores (i.e. mean score derived from ordinal ratings) and to maximize the sensitivity of observed scores as a rater was not limited to a set number of ordinal categories.

Compared to ordinal scales, visual analogue scales are less frequently used in auditory perceptual studies. This is likely because the original analogue scales required manual calculation (i.e. an evaluator manually measuring the distance with a ruler). Our computerized version required no manual measurement. A criticism of the visual analogue scale is that listeners tend to avoid using scale extremes and the mid-section of the scale then becomes over-used (Cowley and Youngblood, 2009; Eadie and Kapsner-Smith, 2011). Despite this drawback, auditory-perceptual data collected on a visual analogue scale is an established method of data collection (Karnell et al., 2007; Nemr et al., 2012; Wuyts et al., 1999; Zraick et al., 2011). Subsequent researchers in our research group have also continued to use this scale (Kraaijenga et al., 2016).

8.4.1.2 Observed score

Regardless of the scale used for data collection, it is common practice to use the average score from a group of listeners within the area of modeling speech and voice quality (Bocklet et al., 2012; Haderlein et al., 2011; Schuster and Stelzle, 2012). Likewise, this is the technique we utilized in our initial studies. Although the listeners who evaluated the CCRT recordings were, as a whole, reliable (evidenced by significant ICC), we considered it likely that variability of lower-range perceptual scores was impacting on model performance. This is because if there is more variation among the scores of the listeners for lower quality speakers, this will be reflected in variation in the mean scores, which in turn, will result in lower performance accuracy for prediction models.

In the second part of our study, rather than use mean scores we moved to consensus-derived scores for the total laryngectomy speaker group. This method is less frequently used in research studies (see examples by de Bruijn de Bruijn et al. (2011a,b) and De Bodt et al. (2002)). Our rational for using consensus data was to maximize the sensitivity of the scale and obtain scores closer to clinical practice, for example, a clinician consulting a colleague for his/her opinion with the outcome being an agreed rating. A criticism of this method is, however, that the consensus-derived rating may not be independent of the other rater as a listener may allow the other rater to influence his/her opinion on the 'severity' of a recording.

We used consensus-derived scores for our studies investigating acoustic signal type categorization of vowels (categorical data) and the auditory-perceptual evaluation of vowels and running speech made on a visual analogue scale (0-1000 scale points) when individual ratings differed by more than 125 scale points. We believe consensus-derived scores are particularly important in our study given the low, yet significant, levels of inter-rater reliability before consensus round and the percentages of agreement for auditory-perceptual scores (TE sustained vowel voice quality 41%; running speech voice quality 38% and speech intelligibility 54%). Note that the consensus methodology was established before perceptual-data collection occurred. Without a consensus round, observed scores would have had a smaller scale distribution.

In our study exploring the relationship between acoustic signal type (categorical four types of signals) and auditory-perceptual evaluation of voice quality, we utilized both consensus continuous quality scores and ordinal quality scores (derived by dividing the visual analogue scale into four equal parts and recoding the continuous consensus scores according to these intervals). The low level of agreement between signal type and recoded interval scores, yet the significant main effect of signal type on the visual analogue data highlights that scale division into equal intervals may be too simplistic a division. An alternative division with unequal intervals may more accurately reflect the way listeners partition the scale into severity regions (e.g., Lopes et al., 2012; Yu et al., 2001).

8.4.2 Speech stimuli considerations

The speech recordings used in this thesis were collected as part of various research projects at the Netherlands Cancer Institute over a period of a decade. The running speech recordings are not the same for the two speaker groups and neither text is phonetically balanced. The recorded texts were similar in that they were both of neutral content and did not require high-level literacy skills. Ideally the read text would be the same for both speaker groups, however this is the reality in the clinical situation; a clinician will provide a text to elicit the type of speech needing to be evaluated (e.g., text with considerable nasal consonants or certain prosody requirements) and this text may not be a standardised text.

It is difficult to assess the impact of having different text-level stimuli for the two corpora in order to compare model performance for the two speaker groups, however our results indicate that text differences are not detrimental to model performance. The corpus for the CCRT speaker group contained a read paragraph made at three evaluation moments (before medical treatment, short-term follow-up and long-term follow-up) and for each recording there was a perceptual score for fragment A and fragment B. The two fragments were not identical, but were similar in terms of number of total and unique words, average word length and phoneme balance (see page 57 for details). Despite not being identical, our results showed that the prediction models achieved similar performance results regardless whether the model was trained on one fragment and tested on the other.

The results in our final chapter also support our prediction models being robust against differences in phonetic context, on the condition that there is phonemic variety and the analysed text is approximately 100 syllables in length. This finding may also explain the initial result that there was no performance difference between a prediction model built for Fragment A and tested on Fragment B as both fragments were approximately 100 syllables in length. The implication is that we can, with some degree of confidence, compare model performances across speaker groups and across speech stimuli so long as the studies share methodologies in terms of automatic evaluation protocols (i.e., sampling strategies, performance measures).

8.4.3 Automatic evaluation

8.4.3.1 Measuring performance

As highlighted in the introduction chapter (see Section 1.2, pg. 9), performance is often reported as the relationship between the predicted, computer-derived scores and the observed perceptual scores. Frequently used measures are the Spearman rank order correlation coefficient, Pearson correlation coefficient and the root mean square error (RMSE).

We elected to use both the RMSE and Pearson correlation coefficient in our studies. We used the RMSE to guide selecting speaker features to include in a model and we used the RMSE complemented with Pearson correlation coefficient to evaluate model performance. The RMSE reflects the distance between the predicted and observed data points and, as Middag et al. (2009) stated, compared to other measures the RMSE is more stable when a model is developed on a large data set and evaluated on a smaller data set. Incorporating the Pearson correlation coefficient provided information on the strength of the relationship between computed and observed and this measure continues to be frequently used (e.g., Bocklet et al., 2012; Haderlein et al., 2011, 2012, 2014).

Although it is tempting to use the Pearson correlation coefficient as a manner of directly comparing prediction models among studies, differences in systems, cross-validation strategies, and perceptual scales inhibit interpreting results in this fashion. We have applied the PCC for quasi-continuous data, including it as a secondary measure allows us to us to compare performance results with those from other research groups.

In terms of identifying the target performance level for a prediction model, our first study (Chapter 2) used the average Pearson correlation coefficient of each rater against the mean of the group of raters to set performance target. In subsequent studies we also calculated the RMSE to reflect this measure being our primary performance measure.

The inclusion of the RMSE and the Pearson correlation coefficient allowed us to identify models that achieved target performance levels (i.e., RMSE below target performance and Pearson correlation coefficient above target performance) and models that were competitive (i.e., only one measure achieved target performance).

8.4.3.2 Tracking trends

The ability to accurately identify change or no change is a key requirement of a prediction model if automatic perceptual evaluation is to be incorporated into clinical evaluation protocols. One of our study aims was to look at whether automatically derived scores could track change (increase/decrease, no difference) in speech intelligibility and voice quality between evaluation moments. Unlike the TL corpus, the CCRT corpus contained longitudinal speech recordings and allowed us to investigate the ability of the best-performing speech intelligibility model to identify change between evaluation moments.

Although our preliminary analysis of the perceptual data for this group of speakers (see Chapter 2) indicated that there was no significant difference in mean perceptual scores over time, we were interested in ability of the best-performing prediction models to identify change between evaluation moments at a speaker level for the cases where listeners agreed on a change in quality. The listeners as a whole disagreed on the direction of change for much of the speech recordings. For example, of the 245 speech intelligibility comparisons available[1], a clear change (increase/decrease) was observed in 62 pairs (see Table 3.5, pg. 72).

The reasons for this bias are likely a combination of connected factors. One explanation may be that parts of the underlying system fail or underestimate values for speech samples with low perceptual scores. As far as we are aware, systematic evaluation of the performance of the phonological features and monophone features either via forced-alignment or alignment-free has not been investigated.

A second explanation may be the increased variation in the perceptual scores at low speech qualities, means that the prediction model will have lower accuracy for scores in this range. In other words, if low perceptual scores are less reliable and difficult to predict, the trends derived thereof are also bound to

[1] T1-T0 93 pairs, T3-T1 74 pairs, T3-T0 78 pairs

be unreliable and inaccurate. The plots of predicted against observed scores consistently highlight that prediction models are less accurate for lower mean perceptual scores. Whether this is because of system failure, fewer data points in this range, increased variation in observed scores in this range or a combination of all these points is unclear.

8.5 Clinical application

Within the area of head and neck cancer and outside our studies, prediction models have been developed for speakers treated surgically for oral cancer who have undergone total laryngectomy and speak using tracheoesophageal speech and have undergone partial laryngectomy The majority of studies have modelled speech intelligibility.

Model performance varies among the studies and most models achieve performance levels that are similar to or that exceed the average listener performance. Directly comparing the performance results among studies is difficult due to differences in speaker characteristics, measuring scales, calculation methods and sampling strategies.

8.5.1 Beyond predicting perceptual scores?

It is interesting to consider the features selected as assessment model inputs and the similarities with the known speech and voice characteristics of a speaker group. Within the CCRT speaker cohort, speaker features selected as model inputs for the articulation and speech intelligibility models may reflect tongue movement in the diagonal of the vowel trapezium and production of anterior lingual consonants (e.g., /l/, /s/). Acoustic studies by Jacobi et al. (2010, 2013) and Kraaijenga et al. (2015) on the same speaker cohort discussed in this thesis, reported significant relationships between tumor location and acoustic measures related to tongue movement, tongue precision, velum movement and velum control.

Similar results have been reported for other speaker cohorts treated for head and neck cancer (de Bruijn et al., 2009; Whitehill et al., 2006) and tongue motility is regarded as a strong predictor of speech outcome after cancer treatment (Schuster and Stelzle, 2012, p.295). Features selected as inputs for the CCRT phonation/voice quality model may reflect the properties at the level of the vocal-source (e.g., presence and amount of voicing) and resonance characteristics (e.g., nasality).

Unfortunately, we have not yet reported the speaker features in the running speech prediction models for the total laryngectomy speaker group due to time constraints. The acoustic variables used in the exploration of sustained vowels and signal typing showed that the presence and duration of voicing were salient features related for the four acoustic signal types. The predictors of voice quality for the CCRT phonation model and the acoustic predictors of signal type for the speakers post total laryngectomy use input reflecting the presence and duration of voicing.

Although the model inputs (i.e., the various individual speaker features) can be linked to known speech or voice characteristics of the speaker group, the selection of a feature does not necessarily equate with a deficit compared to control speakers. The underlying acoustic models are developed on speech materials from control speakers but the prediction models takes input features based on a feature's discriminate or predictive power for a specific group of speakers. The goal of a prediction model is to find the smallest number of features that provide the greatest model accuracy.

The papers contained in this thesis have focused on the development of prediction models using various speaker-feature information as model inputs. We did not investigate whether speaker feature information provided clinically relevant information. There is a trend for studies to consider whether this information could guide clinical intervention (see 8.6). In its current form, the strength of speaker features lies in their power as model inputs and speaker profiling is only recommended as a descriptor.

8.6 New developments

Since the initial literature search discussed in Chapter 1, research groups have continued developing prediction models (Haderlein et al., 2011, 2012; Mayr et al., 2010) and have investigated how sensitive models are to language (Haderlein et al., 2014). Over the last five years, papers have been published in which computer-derived ratings are used as a dependent variable.

Repeating this initial search strategy with updated search dates, several studies were published in which computer-based speech intelligibility scores were the only speech-related dependent variable for three speaker groups: speakers treated for oral cancer (Stelzle et al., 2011, 2013), speakers with cleft lip and palate (Schuster et al., 2012) and speakers with dental prostheses (Knipfer et al., 2012, 2014)[2].

8.6.1 Tracking change

Reports indicate that automatically-derived speech intelligibility scores can be sensitive to changes in nasality post nasal surgery (Mayr et al., 2010), medical

[2] search dates 1 January 2011 to 1 August 2016

treatment for oral cancer (Stelzle et al., 2011, 2013) and changes in dentition (Knipfer et al., 2012, 2014).

The largest of these studies used the PEAKS system to measure the word recognition rate of speakers before treatment for oral cancer and at four moments after treatment (12-24 days; 3, 6 and 12 months) (Stelzle et al., 2013). There authors reported a significant influence of tumor localization, resection volume and radiotherapy on recognition scores post-treatment. Although word recognition rate is not a direct substitute for speech intelligibility scores, as only correlation with auditory-perceptual ratings of speech intelligibility have been established, these results are promising.

8.6.2 Language-/system independence

Haderlein et al. (2014) reported that AMPEX features combined with alignment-free phonological features (ALF.PLFs with Flemish acoustic models) could be used to predict the speech intelligibility scores of German speakers with dysphonia to a level close to target performance. This could indicate that although phonological features are trained on Flemish speakers, the broad categories relating to manner, place and voicing are robust enough that when combined with AMPEX acoustic information, model performance is similar to the average of a group of listeners.

Our findings in Chapter 3 indicated that in general models attained stronger performance levels when the underlying acoustic models matched those of the speaker, however alignment-free features were more consistent when a mismatch was present. In view of the data presented in Chapter 4 where we observed that AMPEX features were include in top-performing speech intelligibility models when cancer involved the larynx, it is understandable that AMPEX features be selected for modelling speech intelligibility of speakers with dysphonia.

8.6.3 ASISTO

The speaker features discussed in this thesis have been included in a prototype system ASISTO (Automatic speech analysis during speech therapy in oncology[3]). This system is designed to offer speech pathologists an opportunity to access automatic computer-derived perceptual scores of speech and voice quality.

Kraaijenga et al. (2016) recently applied the ASISTO prototype system to new speech recordings from a subset of the OC speaker cohort discussed in this thesis. These 22 new speech recordings were obtained 10 or more years post CCRT. The authors also collected auditory-perceptual information on several

[3] https://asisto.elis.ugent.be

aspects quality, including speech intelligibility. Although a strong correlation was reported between ASISTO speech intelligibility scores and mean perceptual speech intelligibility scores, ASISTO scores did not indicate an effect of treatment modality (conventional radiotherapy vs. intensity modulated radiotherapy) on intelligibility scores whereas auditory-perceptual scores significantly differed between the two groups.

It is unclear based on the information presented in Kraaijenga et al. (2016) how the authors computed the speech intelligibility scores as outputs are only reported for text-aligned and alignment-free data obtained using the ELIS system (the ELIS system provides phonological information, referred to in this thesis as PLFs). This initial step in the development of ASISTO indicates a development from research-based use of automatic evaluation to the clinic-based implementation of automatic evaluation.

8.7 Conclusion

The studies reported in this thesis showed that automatic *perceptual-like* evaluation of speech intelligibility and voice quality for speakers treated for head and neck cancer is attainable with the current speech technologies. The prediction models we have developed for running speech are created to support a clinician's auditory-perceptual evaluation; models have not been created with the intention of substitution. Our studies have also shown that acoustic signal typing in its current form for sustained vowels provides limited information on the auditory-perceptual evaluation of the same stimuli.

The ASISTO project was undertaken to move the technology developed and presented in this thesis closer to the clinical setting. The project aims to produce a therapy and evaluation tool that can be used by both clinicians and patients for evaluation purposes and as a feedback tool.

References

SS Barreto and KZ Ortiz. Intelligibility measurements in speech disorders: a critical review of the literature. *Pró Fono*, 20(3):201–6, 2008.

T Bocklet, T Riedhammer, E Nöth, U Eysholdt, and T Haderlein. Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling. *Journal of Voice*, 26(3):390–397, 2012.

P Boersma and D Weenink. Praat: doing phonetics by computer. Computer program: www.Praat.org, 2009.

JA Cowley and H Youngblood. Subjective response differences between visual analogue, ordinal and hybrid response scales. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 53, pages 1883–1887, October 2009. doi: doi:10.1177/154193120905302506.

L D'Alatri, F Bussu, E Scarano, G Paludetti, and MR Marchese. Objective and subjective assessment of tracheoesophageal prosthesis voice outcome. *Journal of Voice*, pages 607–613, 2011.

Marc S De Bodt, Huici Maria E Hernández-Díaz, and Paul H Van De Heyning. Intelligibility as a linear combination of dimensions in dysarthric speech. *J Commun Disord*, 35(3):283–92, 2002.

M. J. de Bruijn, L. ten Bosch, D. J. Kuik, H. Quené, J. A. Langendijk, and C .R. Leemans I. M. Verdonck-de Leeuw. Objective Acoustic-Phonetic Speech Analysis in Patients Treated for Oral or Oropharyngeal Cancer. *Folia Phoniatrica et Logopaedica*, 61(3):180–187, 2009.

M. J. de Bruijn, L. ten Bosch, D. J. Kuik, B. Witte, J. A. Langendijk, and C. R. Leemans I. M. Verdonck-de Leeuw. 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*, 54(5):632–640, 2011a.

Marieke J de Bruijn, Louis ten Bosch, Dirk J Kuik, Johannes A Langendijk, C René Leemans, and Irma M Verdonck-de Leeuw. Artificial neural network analysis to assess hypernasality in patients treated for oral or oropharyngeal cancer. *Logopedics Phoniatrics Vocology*, 36:168–174, 2011b.

TL Eadie and M Kapsner-Smith. The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research*, 54(2):430–447, 2011.

T Haderlein, E Nöth, T Hikmet, A Batliner, M Schuster, U Eysholdt, J Hornegger, and F Rosanowski. Automatic evaluation of prosodic features of tracheoesophageal substitute voice. *European Archives of Oto-Rhino-Laryngology*, 264(11):1315–1321, 2007. doi: 10.1007/s00405-007-0363-4.

T Haderlein, E Nöth, A Batliner, U Eysholdt, and F Rosanowski. Automatic intelligibility assessment of pathological speech over the telephone. *Logopedics Phoniatrics Vocology*, 36:175–181, 2011.

T Haderlein, C Moers, B Möbius, and E Nöth. Automatic rating of hoarseness by text-based cepstral and prosodic evaluation. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings*, pages 573–580. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-32790-2. doi: 10.1007/978-3-642-32790-2_70.

T Haderlein, C Middag, J-P Martens, M Döllinger, and E Nöth. Language-independent automatic evaluation of intelligibility of chronically hoarse persons. *Folia Phoniatrica et Logopaedica*, 66(6):219–226, 2014. doi: 10.1159/000365969.

M Hattori, Y Sumita, and H Taniguchi. Application of an automatic conversation intelligibility test system using computerized speech recognition technique. *Journal of Prosthodontic Research*, 54:7–13, 2010.

I Jacobi, L van der Molen, M van Rossum, and FJM Hilgers. Pre- and short-term posttreatment vocal functioning in patients with advanced head and neck cancer treated with concomitant chemoradiotherapy. In *Proceedings of Interspeech*, pages 2582–2585. ISCA, 2010.

Irene Jacobi, Maya A van Rossum, Lisette van der Molen, Frans JM Hilgers, and Michiel WM van den Brekel. Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy. *Annals of Otology, Rhinology & Laryngology*, 122(12): 754–762, 2013.

MP Karnell, SD Melton, JM Childes, TC Coleman, SA Dailey, and HT Hoffman. Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice*, 21(5): 576–590, 2007.

RD Kent. Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology Speech Lang Pathol*, 5(3):7–23, 1996.

C Knipfer, T Bocklet, E Nöth, M Schuster, B Sokol, S Eitner, E Nkenke, and F Stelzle. Speech intelligibility enhancement through maxillary dental rehabilitation with telescopic prostheses and complete dentures: a prospective study using automatic, computer-based speech analysis. *International Journal of Prosthodontics*, 25(1):24–32, Jan-Feb 2012. ISSN 0893-2174 (Print); 0893-2174 (Linking).

C Knipfer, M Riemann, T Bocklet, E Nöth, M Schuster, B Sokol, S Eitner, E Nkenke, and F Stelzle. Speech intelligibility enhancement after maxillary denture treatment and its impact on quality of life. *International Journal of Prosthodontics*, 27(1):61–69, Jan-Feb 2014. ISSN 0893-2174 (Print); 0893-2174 (Linking).

SAC Kraaijenga, Lisette van der Molen, Irene Jacobi, Olga Hamming-Vrieze, Frans J. M. Hilgers, and Michiel W. M. van den Brekel. Prospective clinical study on long-term swallowing function and voice quality in advanced head and neck cancer patients treated with concurrent chemoradiotherapy and preventive swallowing exercises. *European Archives of Oto-Rhino-Laryngology*, 272(11):3521–3531, 2015. ISSN 1434-4726. doi: 10.1007/s00405-014-3379-6. URL http://dx.doi.org/10.1007/s00405-014-3379-6.

SAC Kraaijenga, IM Oskam, RJJH van Son, O Hamming-Vrieze, FJM Hilgers, MWM van den Brekel, and L van der Molen. Assessment of voice, speech, and related quality of life in advanced head and neck cancer patients 10-years+ after chemoradiotherapy. *Oral Oncology*, 55:24–30, 2016. doi: 10.1016/j.oraloncology.2016.02.001.

L.W. Lopes, I.L.B Lima, D.P. Cavalcante, and A.A.F de Almeida. Severity of voice disorders in children: correlations between perceptual and acoustic data. *Journal of Voice*, 26(6):819.e7 – 819.e12, 2012. ISSN 0892-1997. doi: http://dx.doi.org/10.1016/j.jvoice.2012.05.008. URL http://www.sciencedirect.com/science/article/pii/S089219971200077X.

S Mayr, K Burkhardt, M Schuster, K Rogler, A Maier, and H Iro. The use of automatic speech recognition showing the influence of nasality on speech intelligibility. *Eur Arch Otorhinolaryngol*, 267(11):1719–1725, 2010.

C Middag, G van Nuffelen, J-P Martens, and Marc S De Bodt. Objective intelligibility assessment of pathological speakers. In *Interspeech*, pages 1745–1748. Interspeech, 2008.

C Middag, JP Martens, G van Nuffelen, and M de Bodt. DIA: a tool for objective intelligibility assessment of pathological speech. In *Proceedings of*

Models and Analysis of Vocal Emissions for Biomedical Applications, 6th International workshop,, pages 165–167, 2009.

C Middag, Y Saeys, and JP Martens. Towards an ASR-free objective analysis of pathological speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 294–297, Tokio, Japan, 2010.

N Miller. Measuring up to speech intelligibility. *International Journal of Language and Communication Disorders*, 48(6):601–612, 2013. doi: 10.1111/1460-6984.12061.

K Nemr, M Simões-Zenari, GF Cordeiro, D Tsuji, AI Ogawa, MT Ubrig, and MHM Menezes. GRBAS and CAPE-V scales: high reliability and consensus when applied at different times. *Journal of Voice*, 26(6):812–e17, 2012.

M. Schuster and F. Stelzle. Outcome measurements after oral cancer treatment: speech and speech-related aspects - an overview. *Oral Maxillofac Surg*, 16: 291–298, 2012.

M Schuster, A Maier, T Bocklet, E Nkenke, A Holst, U Eysholdt, and F Stelzle. Automatically evaluated degree of intelligibility of children with different cleft type from preschool and elementary school measured by automatic speech recognition. *International Journal of Pediatric Otorhinolaryngology*, 76(3): 362–369, Mar 2012. doi: 10.1016/j.ijporl.2011.12.010.

F Stelzle, A Maier, E Noth, T Bocklet, C Knipfer, M Schuster, FW Neukam, and E Nkenke. Automatic quantification of speech intelligibility in patients after treatment for oral squamous cell carcinoma. *J Oral Maxillofac Surg*, 69 (5):1493–1500, May 2011.

F Stelzle, C Knipfer, M Schuster, T Bocklet, E Noth, W Adler, L Schempf, P Vieler, M Riemann, F W Neukam, and E Nkenke. Factors influencing relative speech intelligibility in patients with oral squamous cell carcinoma: a prospective study using automatic, computer-based speech analysis. *International Journal of Oral and Maxillofacial Surgery*, 42(11):1377–1384, Nov 2013. doi: 10.1016/j.ijom.2013.05.021.

CJ van As-Brooks, FJ Koopmans-van Beinum, LCW Pols, and FJM Hilgers. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *Journal of Voice*, 20(3):355–368, 2006.

L. van der Molen, M A. van Rossum, I. Jacobi, R. van Son, Ludi E Smeele, Coen R N Rasch, and FJM. Hilgers. Pre- and posttreatment voice and speech outcomes in patients with advanced head and neck cancer treated with chemoradiotherapy: expert listeners' and patient's perception. *Journal of Voice*, 26:664:e25–33, 2012.

RJJH van Son. Nederlands Kanker Instituut Tracheoesophageal Voice Analysis Tool (TEVA). Computer program: www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/NKIcorpora/NKI_TEVA/, 2012.

Tara L Whitehill, Valter Ciocca, Judy C-T Chan, and Nabil Samman. Acoustic analysis of vowels following glossectomy. *Clin Linguist Phon*, 20(2-3):135–40, 2006.

FL Wuyts, MS de Bodt, and PH van de Heyning. Is the reliability of a visual analog scale higher than an ordinal scale? an experiment with the grbas scale for the perceptual evaluation of dysphonia. *Journal of Voice*, 13(4):508–517, 1999.

P Yu, J Revis, FL Wuyts, M Zanaret, and A Giovanni. Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatrica et Logopaedica*, 54(6):271–281, 2001.

RI Zraick, GB Kempster, NP Connor, S Thibeault, BK Klaben, Z Bursac, CR Thrush, and LE Glaze. Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology*, 20(1):14–22, 2011.

Samenvatting: Automatische beoordeling van stem en spraakverstaanbaarheid na behandeling voor hoofd-halskanker

Achtergronden Kanker in het hoofd-halsgebied en de behandeling daarvan kunnen een negatief effect hebben op iemands stem en spraak. Voor de logopedist vormt het beoordelen van de verstaanbaarheid en kwaliteit van de spraak een belangrijk onderdeel van de behandeling van de patiënt omdat het een indicatie geeft van de ernst van de pathologie en de vooruitgang die de patiënt reeds maakte. Het objectief beoordelen van iemands spraak of stem is echter een subjectief gegeven dat wordt beïnvloed door tal van factoren, zoals familiariteit van de logopedist met de patiënt en de test items. Luisteraars beoordelen vaak ook op een inconsistente manier. Een computer is van nature objectief en wel consistent in het uitvoeren van deze taak.

In dit onderzoek werden automatische voorspellingsmodellen ontwikkeld voor de beoordeling van spraakverstaanbaarheid en stemkwaliteit van twee groepen van sprekers die behandeld werden voor hoofd-halskanker. De eerste groep, besproken in deel I van dit proefschrift, betreft patiënten met voortgeschreden kwaadaardige tumoren in het hoofd-halsgebied. Deze patiënten kregen een niet-chirurgische kankerbehandeling, namelijk concomitante chemoradiotherapie (CCRT). Bij deze behandeling worden de chemotherapie en bestraling gelijktijdig toegediend. Een dergelijke behandeling kan iemands stem en spraak nadelig beïnvloeden. De tweede groep wordt gevormd door patiënten die werden behandeld voor voortgeschreden kwaadaardige tumoren van het strottenhoofd. Hiervoor ondergingen deze patiënten een totale laryngectomie (TL), een operatie waarbij het strottenhoofd en dus ook de stembanden worden verwijderd. Na een TL is spreken weer mogelijk met behulp van een stemprothese. Dit hulpmiddel bevat een éénrichtingsklep, waardoor weer stemvorming mogelijk is omdat de uitademingslucht langs trillende structuren in de slokdarm naar de mondholte kan stromen. Deze zogenaamde tracheoesofageale spraak klinkt duidelijk anders dan de spraak vóór de operatie.

Overzicht van het proefschrift Het Nederlands Kanker Instituut heeft twee grote collecties spraakopnames van deze twee groepen sprekers, waarin ze een kort verhaal voorlezen. Deze collecties worden beschreven in hoofdstuk 2 (CCRT) en hoofdstuk 5 (TL). Al deze spraakopnames zijn door logopedisch geschoolde luisteraars beoordeeld op stemkwaliteit en spraakverstaanbaarheid. In hoofdstukken 3, 4 en 6 presenteren we een aantal manieren om automatisch te voorspellen hoe een luisteraar de opnames zal beoordelen. Daarvoor gebruiken we verschillende vormen van spraaktechnologie die zijn gebaseerd op het feit dat de computer herkent wat de spreker gezegd heeft in termen van individuele klanken, op de kenmerken van het spraakgeluid, en/of op basis van akoestische informatie.

Met behulp van computers kunnen we voorspellingsmodellen ontwikkelen op basis van informatie uit de spraakopnames zelf en/of door het vergelijken van wat er gezegd zou moeten zijn met wat de computer 'gehoord' heeft. In een van de studies onderzochten we ook hoe goed de voorspellingsmodellen waren wanneer er een mismatch was tussen de taal waarmee ze waren getraind en de taal van de spreker. In ons geval was de computer getraind om Vlaamse spraakopnames te analyseren, maar voerden wij Nederlandse opnames in. We vonden dat modellen, die wat er daadwerkelijk gezegd was vergeleken met wat had moeten worden gezegd, in deze situatie minder goed presteerden dan modellen, die het door de computer "gehoorde" niet hoefden te vergelijken met de doeltekst maar gewoon de variatie van het spraakgeluid analyseerden.

In onze zoektocht naar de optimale modellering van de voorspellingen, vonden we dat wanneer het model meer en verschillende soorten sprekerkenmerken ter beschikking heeft, het verschil tussen de beoordelingen door de computer en de luisteraar kleiner wordt. Sommige van de beste modellen presteren op een niveau dat vergelijkbaar is met dat van een gemiddelde luisteraar.

In de hoofdstukken 4 en 7 konden we laten zien dat vergelijkbare resultaten kunnen worden verkregen voor het voorspellen van beoordelingen van *Articulatiekwaliteit* en *Accent* (zie hoofdstuk 4) en dat de modellen ook gebruikt kunnen worden voor TL-sprekers (zie hoofdstuk 7).

Een belangrijk aspect van computermodellen is uiteraard de betrouwbaarheid ervan. In hoofdstuk 7 vonden we dat de prestaties van het voorspellingsmodel betrouwbaar worden wanneer de computer voldoende voorbeelden van elke klank ziet. We vonden dat aan dit criterium voldaan is wanneer het spraakmateriaal tenminste 100 lettergrepen lang is.

Veel akoestische metingen correleren, vaak sterk, met Acoustic Signal Typing (*AST*), een classificatie van de spraak van *TL*-patiënten. In hoofdstuk 5 laten we zien dat slechts twee soorten akoestische informatie nodig zijn om een stem te kunnen indelen in één van de vier AST groepen. De eerste is de **stemhebbendheid**, gerepresenteerd door de *Voiced Fraction* (*VF*) en/of de maximale fonatieduur (*Maximum Voicing Duration*; *MVD*). De tweede is de **ruis** in de stem, gerepresenteerd door de *Harmonics to Noise Ratio* (*HNR*). Dit betekent dat de stemhebbendheid, fonatieduur en hoeveelheid ruis de belangrijkste aspecten zijn van de AST.

De relatie tussen de AST classificatie en beoordelingen van spraakkwaliteit door luisteraars is onderzocht in hoofdstuk 6. De resultaten ondersteunen het gebruik van AST als onderdeel van een ruimere evaluatie van de stem. Hoewel er een statistisch significant verband tussen de twee maten, levert de AST in zijn huidige vorm echter slechts in beperkte mate prognostische informatie over spraakkwaliteit.

Conclusies De resultaten beschreven in dit proefschrift tonen aan dat het goed mogelijk is om computer-gebaseerd automatische beoordelingen van spraakverstaanbaarheid en stemkwaliteit te gebruiken bij sprekers die behandeld zijn voor hoofd-halskanker en dat computer-gebaseerde automatische beoordelingen een clinicus bij zijn/haar evaluaties kunnen helpen. Let wel dat deze modellen niet ontwikkeld zijn met als doel de clinicus te vervangen. De clinicus blijft de belangrijkste beoordelaar van hoe een spreker praat en klinkt, maar deze computermodellen kunnen gezien worden als een handige, objectieve en accurate toevoeging aan het oordeel van de expert.

Summary:
Automatic evaluation of voice and speech intelligibility after treatment of head and neck cancer

Background Cancer of the head and neck and its medical treatment and management, can have a negative impact on how a person sounds and talks. For the speech pathologist, rating a person's speech intelligibility and voice quality is an important part of patient management. Rating someone's speech or voice, however, can be difficult task to perform objectively as a listener's ratings are often inconsistent. Computerized ratings, on the other hand, are consistent.

This thesis has focused on developing automatic prediction models for speech intelligibility and voice quality assessment for two groups of speakers treated for head and neck cancer. The first group discussed in Part I of this thesis are people with advanced tumours in the head and neck. These people received a type of non-surgical cancer treatment, called concurrent chemoradiotherapy (CCRT). This type of treatment can affect a person's voice and speech. The second group of people were treated for advanced tumors in the larynx. These people underwent a total laryngectomy (TL), in which the larynx (also known as the 'voice box') is removed. After a TL, speaking is possible with the aid of a valve that redirects air past vibrating structures in the neck towards the mouth. This type of speech is called tracheoesophageal speech and it sounds very different to how a person sounded before the surgery.

Thesis overview The Netherlands Cancer Institute has a large collection of speech recordings for the two speaker groups reading a short story. These collections are described in Chapter 2 (CCRT) and Chapter 5 (TL). For every recording of a person talking, we have asked people to rate the voice quality and speech intelligibility. In the chapters we present ways to automatically predict how a person rates the recordings by using different forms of speech technology. Some of these ways are based on the computer recognising what the speaker

said in terms of some of the sounds or the characteristics of the sounds or based on acoustic information.

The computers are able to develop prediction models using information contained in the speech recordings and sometimes by comparing what the computer 'heard' with what should have been said. In one of the studies we compared the how well the prediction models performed when there was a mismatch between the language it was trained to listen to and the language of the speaker. In our case, if the computer had been trained to analyse Flemish speech recordings but we gave it Dutch speech recordings. We found that models that want to match-up what was said with what should have been said, did not perform as well as models that did not need to compare what it heard to a target text.

In our experiments investigating how we can combine computer information to create strong models, we found that when the model can access more, and different types of information, the difference between the computer rating and the real, listener rating, becomes smaller. Some of the best models performed at a level similar to an average listener.

In the other chapters we were able to show that similar results could also be found for predicting ratings of *Articulation quality* and *Accent* (see Chapter 4) and could also apply to TL speakers (see Chapter 7). In Chapter 7, we also found that the performance of the prediction model is close to its maximum when the speech material is at least 100 syllables long.

Many acoustic measurements correlate, often strongly, with the Acoustic Signal Typing (*AST*) classification of speech from *TL* patients. However, in Chapter 5 we find that only two types of acoustic information can can be used to classify a recording into one of the four ASTs: voice detection using the *Voiced Fraction* (*VF*) or *Maximum Voicing Duration* (*MVD*) and the *harmonics-to-noise ratio* (*HNR*). This indicates that the presence and duration of voice and the amount of noise are important aspects of AST.

The relationship between listeners' judgments of AST and judgments of voice quality were investigated in Chapter 6. The results support the use of AST as part of a larger evaluation of a person's voice. Although there is a statistically significant relationship between the two measures, AST in its current form provides limited predictive information on voice quality.

Conclusion In short, the results discussed in this thesis showed that computer-based ratings of speech intelligibility and voice quality for speakers treated for head and neck cancer is possible and the computer ratings can help a clinician with his/her evaluations. None of the computer models were created with the intention of replacing the clinician's judgment of how a speaker talks and sounds.

A

Description of recording databases

A.1 Introduction

The individual chapters of this thesis contain meta-data statistics of the patients and recordings used in these studies. To complete the description of the patient recordings used in the current thesis, we list an anonymised comprehensive extract from the corpus meta-data for all the recordings used. In these tables, patients are identified using a unique 3 letter abbreviation of the 8 letter patient identifier that has been used throughout the speech corpora.

A.2 CCRT Speakers and evaluation results

Table A.1: CCRT speaker characteristics and average perceptual scores on 5 point scales. L1: + native, – non-native speaker of Dutch (perceptual evaluation). Articulation, Intelligibility, Voice, Dialect: average scores (1-5). T0: Pre treatment, T1: 10 weeks, T3: 12 months after treatment. Note that speaker ANU also had TLE.

					Intelligibility			Articulation			Voice			Dialect		
	ID	Sex	Age	L1	T0	T1	T3	T0	T1	T3	T0	T1	T3	T0	T1	T3
1	0CH	M	58	+	4.73	4.54	4.62	4.73	4.54	4.62	4.08	4.12	3.92	4.62	4.62	4.81
2	0EH	M	63	+	4.23	4.19	4.46	4.23	4.19	4.46	3.69	2.19	2.54	3.88	3.88	4.15
3	0IH	M	51	+	4.73	4.73	4.58	4.73	4.73	4.58	4.38	4.00	4.08	4.88	4.69	4.38
4	0RG	M	63	+	4.23	4.08	4.12	4.23	4.08	4.12	3.88	4.00	3.35	3.69	3.50	3.62
5	0SJ	M	36	+	2.81	3.23	–	2.81	3.23	–	2.19	1.38	–	4.23	4.46	–
6	182	M	71	–	2.58	2.12	2.46	2.58	2.12	2.46	3.50	3.65	3.58	1.62	1.46	1.76
7	1CX	M	64	+	4.38	4.19	4.42	4.38	4.19	4.42	3.92	3.88	4.23	3.96	4.12	4.00
8	1JB	F	62	–	3.27	3.08	3.08	3.27	3.08	3.08	4.15	3.77	4.08	2.12	1.81	1.96
9	1PL	M	52	+	4.42	–	–	4.42	–	–	4.15	–	–	4.35	–	–
10	259	M	54	+	4.62	4.88	4.73	4.62	4.88	4.73	4.04	3.54	3.50	4.54	4.88	4.62
11	28C	F	55	+	4.42	–	4.15	4.42	–	4.15	1.46	–	1.65	4.65	–	4.50
12	31P	M	63	+	3.54	3.72	3.72	3.54	3.72	3.72	3.50	3.35	3.15	3.38	3.62	3.50
13	31T	M	68	+	3.50	–	–	3.50	–	–	2.58	–	–	4.00	–	–
14	ANU	M	63	+	4.04	4.00	4.04	4.04	4.00	4.04	3.69	3.58	3.27	4.08	4.31	4.23
15	BK0	M	55	+	4.35	4.31	4.46	4.35	4.31	4.46	2.15	3.69	2.27	4.42	4.46	4.50
16	C6R	M	44	+	4.04	4.27	4.04	4.04	4.27	4.04	4.38	4.12	3.96	3.77	3.54	3.81
17	CGO	M	62	+	2.96	–	–	2.96	–	–	4.04	–	–	4.38	–	–
18	CHZ	M	53	+	4.62	4.19	4.65	4.62	4.19	4.65	4.69	4.42	4.42	4.65	4.42	4.50

Continued on next page

Table A.1: CCRT speaker characteristics and average perceptual scores. Continued from previous page

					Intelligibility			Articulation			Voice			Dialect		
	ID	Sex	Age	L1	T0	T1	T3	T0	T1	T3	T0	T1	T3	T0	T1	T3
19	CIP	M	50	+	3.69	–	–	3.69	–	–	4.00	–	–	4.42	–	–
20	CPO	M	60	+	3.38	4.04	3.81	3.38	4.04	3.81	3.81	3.65	3.85	2.96	3.65	3.23
21	DXL	M	77	+	4.35	3.88	4.08	4.35	3.88	4.08	3.92	3.85	3.92	4.46	4.12	4.35
22	EJ3	M	60	+	3.73	3.31	–	3.73	3.31	–	4.12	3.81	–	3.88	3.77	–
23	ERE	M	50	+	4.42	4.35	–	4.42	4.35	–	3.81	3.54	–	4.19	4.19	–
24	EV0	M	66	+	4.38	4.31	4.42	4.38	4.31	4.42	4.31	4.12	4.23	4.73	4.58	4.54
25	FY0	F	62	+	4.38	3.73	4.15	4.38	3.73	4.15	4.23	4.04	4.19	3.58	3.92	3.69
26	HCZ	F	55	+	4.31	3.92	4.08	4.31	3.92	4.08	4.00	3.15	3.42	4.12	4.24	4.42
27	I8P	M	58	+	4.62	4.81	4.77	4.62	4.81	4.77	4.58	4.38	4.46	4.31	4.69	4.46
28	K3M	M	66	+	4.69	4.58	4.77	4.69	4.58	4.77	4.23	4.31	4.08	4.54	4.69	4.73
29	KZC	M	49	–	2.85	3.00	2.81	2.85	3.00	2.81	4.54	4.50	4.62	1.96	2.00	1.96
30	L8U	M	62	+	3.27	3.73	4.04	3.27	3.73	4.04	2.38	3.85	3.96	4.15	3.88	3.73
31	MB7	M	79	+	2.96	–	–	2.96	–	–	4.15	–	–	4.42	–	–
32	NZB	M	53	+	2.12	–	–	2.12	–	–	3.62	–	–	4.69	–	–
33	OEL	M	78	+	4.46	4.23	–	4.46	4.23	–	4.08	3.46	–	3.69	3.69	–
34	P4U	M	57	+	3.85	4.19	4.23	3.85	4.19	4.23	3.23	4.15	4.15	4.50	4.31	4.15
35	PNM	M	48	+	4.65	4.65	4.19	4.65	4.65	4.19	3.96	3.88	4.27	4.04	3.88	3.88
36	PTC	F	46	–	4.19	4.19	3.92	4.19	4.19	3.92	4.46	4.73	4.62	3.27	3.04	3.15
37	PTO	M	73	+	4.46	4.12	4.19	4.46	4.12	4.19	3.96	3.12	3.58	4.12	3.88	4.12
38	R83	M	61	+	–	4.42	4.42	–	4.42	4.42	–	4.15	3.96	–	3.85	3.96
39	STF	M	57	–	2.50	2.65	–	2.50	2.65	–	4.12	4.50	–	2.12	1.73	–
40	T08	F	75	+	4.50	3.65	4.19	4.50	3.65	4.19	4.35	4.19	4.23	4.12	4.15	3.88

Continued on next page

Table A.1: CCRT speaker characteristics and average perceptual scores. Continued from previous page

					Intelligibility			Articulation			Voice			Dialect		
	ID	Sex	Age	L1	T0	T1	T3	T0	T1	T3	T0	T1	T3	T0	T1	T3
41	TH6	M	57	+	4.85	4.77	4.54	4.85	4.77	4.54	3.81	3.50	3.38	4.73	4.77	4.77
42	U81	M	52	+	4.27	3.27	4.40	4.27	3.27	4.40	4.19	3.54	3.77	3.85	4.23	3.96
43	UU8	M	60	+	3.96	4.42	4.46	3.96	4.42	4.46	3.92	4.12	4.31	4.35	4.15	4.19
44	UUC	F	64	+	3.50	4.15	4.19	3.50	4.15	4.19	3.81	3.85	4.23	4.46	3.92	4.04
45	UX8	M	54	+	4.23	4.23	–	4.23	4.23	–	4.15	4.27	–	3.46	3.42	–
46	V3C	M	42	+	4.00	4.31	–	4.00	4.31	–	3.85	4.38	–	3.69	3.58	–
47	VAY	F	45	+	3.00	2.69	4.12	3.00	2.69	4.12	4.58	3.62	4.69	4.58	4.62	4.04
48	W8S	M	32	–	3.12	3.04	3.12	3.12	3.04	3.12	4.62	4.65	4.54	1.96	1.85	2.23
49	X54	M	51	–	3.46	3.12	3.12	3.46	3.12	3.12	4.31	4.31	3.96	2.85	2.58	3.00
50	X9U	M	46	+	3.96	3.62	4.23	3.96	3.62	4.23	4.00	3.88	4.04	4.46	4.38	4.35
51	XFP	M	57	+	4.62	4.46	–	4.62	4.46	–	4.27	4.12	–	4.35	3.92	–
52	Y7H	M	48	–	2.69	2.65	2.58	2.69	2.65	2.58	3.54	3.62	3.46	1.88	1.92	1.92
53	Y8B	F	54	+	2.46	2.96	–	2.46	2.96	–	4.12	4.08	–	4.65	4.69	–
54	YBY	F	39	+	4.62	4.58	4.58	4.62	4.58	4.58	4.31	3.92	4.15	4.42	4.54	4.38
55	YKH	M	52	+	4.58	4.23	–	4.58	4.23	–	3.62	3.77	–	4.04	3.96	–

Table A.2: CCRT speaker characteristics and normalized perceptual scores on 5 point scales. L1: + native, – non-native speaker of Dutch (perceptual evaluation). Articulation, Intelligibility, Voice, Dialect: average scores normalized per listener (Z values). T0: Pre treatment, T1: 10 weeks, T3: 12 months after treatment. Note that speaker ANU also had TLE.

					Intelligibility			Articulation			Voice			Dialect		
	ID	Sex	Age	L1	T0	T1	T3	T0	T1	T3	T0	T1	T3	T0	T1	T3
1	0CH	M	58	+	0.91	0.74	0.70	0.84	0.75	0.81	0.22	0.24	0.04	0.73	0.70	0.92
2	0EH	M	63	+	0.42	-0.31	-0.01	0.37	0.29	0.64	-0.19	-1.88	-1.49	-0.01	0.11	0.25
3	0IH	M	51	+	0.77	0.91	0.83	0.84	0.90	0.69	0.62	0.17	0.27	0.98	0.82	0.49
4	0RG	M	63	+	0.46	0.04	0.12	0.31	0.17	0.21	0.04	0.19	-0.6	-0.06	-0.36	-0.24
5	0SJ	M	36	+	-1.51	-2.25	–	-1.2	-0.65	–	-2.00	-2.82	–	0.34	0.67	–
6	182	M	71	–	-1.78	-2.39	-1.96	-1.56	-2.04	-1.67	-0.37	-0.2	-0.39	-2.08	-2.26	-1.98
7	1CX	M	64	+	0.68	0.50	0.56	0.40	0.26	0.55	0.05	-0.03	0.41	0.17	0.26	0.05
8	1JB	F	62	–	-0.82	-0.82	-0.76	-0.79	-0.91	-0.89	0.39	-0.17	0.18	-1.78	-1.91	-1.8
9	1PL	M	52	+	0.69	–	–	0.57	–	–	0.18	–	–	0.47	–	–
10	259	M	54	+	0.76	0.66	0.65	0.69	1.01	0.80	0.20	-0.36	-0.43	0.63	0.97	0.72
11	28C	F	55	+	-0.81	-2.31	-0.46	0.62	-0.56	0.34	-2.74	-3.06	-2.51	0.72	0.69	0.59
12	31P	M	63	+	-0.3	-0.28	-0.17	-0.46	-0.23	-0.21	-0.55	-0.68	-0.85	-0.42	-0.21	-0.31
13	31T	M	68	+	-0.55	–	–	-0.51	–	–	-1.46	–	–	0.10	–	–
14	ANU	M	63	+	0.22	0.30	0.12	0.26	0.10	0.18	-0.2	-0.3	-0.8	0.32	0.42	0.39
15	BK0	M	55	+	-0.16	0.41	-0.16	0.50	0.44	0.57	-1.94	-0.22	-1.72	0.51	0.58	0.58
16	C6R	M	44	+	0.17	0.42	0.35	0.09	0.37	0.12	0.66	0.25	0.15	-0.09	-0.27	-0.06
17	CGO	M	62	+	-0.21	–	–	-0.99	–	–	0.22	–	–	0.52	–	–
18	CHZ	M	53	+	0.80	0.41	0.75	0.70	0.26	0.68	0.89	0.70	0.67	0.83	0.59	0.68
19	CIP	M	50	+	-0.11	–	–	-0.26	–	–	0.18	–	–	0.57	–	–
20	CPO	M	60	+	-0.51	0.06	-0.01	-0.55	0.10	-0.08	-0.07	-0.18	0.06	-0.82	-0.15	-0.45

Continued on next page

Table A.2: CCRT speaker characteristics and normalized perceptual scores. Continued from previous page

					Intelligibility			Articulation			Voice			Dialect		
	ID	Sex	Age	L1	T0	T1	T3	T0	T1	T3	T0	T1	T3	T0	T1	T3
21	DXL	M	77	+	0.24	-0.18	0.23	0.50	-0.03	0.21	0.09	-0.03	0.09	0.54	0.27	0.44
22	EJ3	M	60	+	0.10	-0.37	–	-0.17	-0.71	–	0.35	0.02	–	0.09	-0.08	–
23	ERE	M	50	+	0.50	0.36	–	0.57	0.53	–	-0.03	-0.38	–	0.34	0.33	–
24	EV0	M	66	+	0.78	0.61	0.60	0.55	0.40	0.53	0.49	0.24	0.44	0.81	0.68	0.66
25	FY0	F	62	+	0.49	-0.05	0.24	0.47	-0.25	0.29	0.47	0.32	0.42	-0.28	0.00	-0.12
26	HCZ	F	55	+	0.61	-0.05	0.30	0.48	0.10	0.16	0.18	-0.84	-0.46	0.27	0.33	0.49
27	I8P	M	58	+	0.90	0.99	0.94	0.72	0.98	0.83	0.85	0.62	0.72	0.40	0.74	0.47
28	K3M	M	66	+	0.86	0.83	0.93	0.84	0.72	0.94	0.50	0.53	0.27	0.70	0.79	0.85
29	KZC	M	49	–	-1.22	-1.16	-1.18	-1.17	-1.05	-1.31	0.78	0.72	0.86	-1.76	-1.67	-1.79
30	L8U	M	62	+	-0.82	-0.04	0.16	-0.76	-0.02	0.18	-1.77	0.03	0.13	0.26	0.09	-0.08
31	MB7	M	79	+	-0.73	–	–	-1.08	–	–	0.40	–	–	0.55	–	–
32	NZB	M	53	+	-1.66	–	–	-1.77	–	–	-0.24	–	–	0.86	–	–
33	OEL	M	78	+	0.45	0.17	–	0.55	0.35	–	0.22	-0.59	–	-0.17	-0.08	–
34	P4U	M	57	+	-0.04	0.59	0.59	0.00	0.35	0.42	-0.63	0.40	0.36	0.64	0.46	0.36
35	PNM	M	48	+	0.62	0.26	0.49	0.72	0.77	0.35	0.17	0.04	0.51	0.21	0.04	0.00
36	PTC	F	46	–	0.28	0.29	0.28	0.31	0.33	0.04	0.72	1.02	0.91	-0.56	-0.74	-0.64
37	PTO	M	73	+	0.58	0.23	0.38	0.62	0.25	0.34	0.06	-0.84	-0.32	0.31	0.03	0.33
38	R83	M	61	+	–	0.56	0.56	–	0.55	0.59	–	0.40	0.16	–	0.06	0.13
39	STF	M	57	–	-2.17	-1.72	–	-1.79	-1.46	–	0.31	0.76	–	-1.76	-2.00	–
40	T08	F	75	+	0.62	0.01	0.37	0.72	-0.35	0.36	0.57	0.37	0.48	0.30	0.39	0.04
41	TH6	M	57	+	0.95	0.75	0.71	1.05	1.02	0.75	-0.12	-0.45	-0.53	0.85	0.83	0.89
42	U81	M	52	+	0.49	-0.24	0.29	0.44	-0.74	0.50	0.37	-0.37	-0.15	-0.01	0.46	0.15

Continued on next page

Table A.2: CCRT speaker characteristics and normalized perceptual scores. Continued from previous page

					Intelligibility			Articulation			Voice			Dialect		
	ID	Sex	Age	L1	T0	T1	T3	T0	T1	T3	T0	T1	T3	T0	T1	T3
43	UU8	M	60	+	0.45	0.63	0.70	0.09	0.60	0.65	0.08	0.33	0.56	0.44	0.37	0.33
44	UUC	F	64	+	-0.26	0.47	0.44	-0.51	0.21	0.25	-0.04	-0.02	0.42	0.49	0.13	0.28
45	UX8	M	54	+	0.22	0.36	–	0.31	0.32	–	0.48	0.44	–	-0.34	-0.46	–
46	V3C	M	42	+	0.27	0.45	–	0.22	0.38	–	-0.02	0.62	–	-0.07	-0.23	–
47	VAY	F	45	+	-0.19	-0.65	0.30	-0.99	-1.36	0.10	0.81	-0.26	0.98	0.70	0.72	0.28
48	W8S	M	32	–	-0.88	-0.89	-0.9	-0.85	-1.00	-0.87	0.88	0.90	0.78	-1.75	-1.89	-1.57
49	X54	M	51	–	-0.46	-0.65	-0.65	-0.52	-0.82	-0.92	0.51	0.48	0.11	-0.99	-1.15	-0.85
50	X9U	M	46	+	0.04	0.06	0.46	0.02	-0.31	0.32	0.12	0.01	0.26	0.63	0.54	0.50
51	XFP	M	57	+	0.67	0.59	–	0.74	0.59	–	0.52	0.30	–	0.51	0.10	–
52	Y7H	M	48	–	-1.42	-1.38	-1.74	-1.39	-1.45	-1.55	-0.34	-0.25	-0.41	-1.84	-1.75	-1.79
53	Y8B	F	54	+	-1.07	-0.51	–	-1.59	-0.96	–	0.31	0.27	–	0.76	0.76	–
54	YBY	F	39	+	0.86	0.70	0.38	0.70	0.71	0.72	0.53	0.01	0.35	0.55	0.66	0.42
55	YKH	M	52	+	0.62	0.38	–	0.70	0.33	–	-0.26	-0.06	–	0.19	0.10	–

A.3 TE Speakers and evaluation results

*Table A.3: Tracheoesophageal (TE) speakers and perceptual consensus scores. Age - Age at laryngectomy; Age(2) - Age at recording; * - Age at recording ± 1 year; Chapter - Chapter in this thesis where data were used; AST - Acoustic Signal Typing; VQ - Voice quality; Int - IINFVo Intelligibility rating; Imp - IINFVo Impression rating. AST is on a 4 point scale, the others are VAS ratings converted to a 1000 point scale. Speakers not used in studies did not complete the recordings as planned or there were technical difficulties with the recordings. Note that speaker ANU also had CCRT.*

	ID	Sex	Age	Age(2)	Chapter	AST	VQ	Int	Imp
1	02R	M	73	79*	5,6,7	2	796	939	882
2	0EO	M	64	64	5,6,7	4	617	695	640
3	0T3	M	62	62	–	–	–	–	–
4	123	M	72	76*	5,6,7	2	435	631	564
5	1BX	F	49	56	5,6,7	2	668	798	680
6	23K	M	66	75*	5,6,7	3	500	555	569
7	2B0	M	58	61	5,6,7	2	664	665	652
8	33Q	M	71	71	5,6,7	4	128	749	580
9	3E6	M	41	41	5,6,7	2	424	385	355
10	A58	M	51	52	–	–	–	–	–
11	A6P	M	44	54*	5,6,7	2	506	675	448
12	ANU	M	50	68*	5,6,7	3	36.5	568	250
13	AU0	M	64	66	5,6,7	2	598	640	395
14	B23	M	44	45	–	–	–	–	–
15	B2N	M	62	72*	5,6,7	4	197	600	397
16	B85	M	46	50*	5,6,7	2	511	702	500
17	BI1	F	71	73	5,6,7	4	98.5	655	330
18	BOG	M	62	69	5,6,7	4	42.5	530	425
19	BPN	M	75	75	–	–	–	–	–
20	BQK	M	51	52*	5,6,7	4	247	765	452
21	BTB	M	52	57*	5,6,7	3	430	680	650
22	C0V	M	78	81	5,6,7	3	376	360	215
23	C1K	M	46	48	5,6,7	2	935	351	399
24	C1S	M	50	55*	5,6,7	3	129	500	400
25	CTH	M	69	70	–	–	–	–	–
26	D0U	M	66	76	5,6,7	2	401	701	400
27	DCA	M	73	67	–	–	–	–	–
28	DCB	M	69	72	5,6,7	2	616	598	402
29	DM8	F	59	61	5,6,7	1	599	790	600
30	EIQ	M	51	67*	5,6,7	3	614	755	608
31	EUZ	M	52	56*	5,6,7	2	572	751	600
32	F6Z	M	51	75*	5	2	427.5	–	–

Continued on next page

Table A.3: Tracheoesophageal (TE) speakers and perceptual consensus scores. Continued from previous page

	ID	Sex	Age	Age(2)	Chapter	AST	VQ	Int	Imp
33	FC1	M	75	77*	5,6,7	3	547	600	485
34	FKW	M	60	60	5,6,7	2	112.5	680	571
35	GNB	M	72	74*	5,6,7	3	472	685	561
36	H2N	F	75	77	5,6	2	199	125	70
37	HC3	F	66	76	5,6,7	1	401	630	520
38	HHV	F	54	62	5,6,7	2	352	585	386
39	HNB	M	78	80*	5,6,7	3	24	625	418
40	I0F	F	62	70	5	2	601	–	–
41	IHF	F	63	64	5,6,7	2	400	751	749
42	IZ9	M	54	57	5,6,7	1	616	580	552
43	J73	F	53	66	5,6,7	2	365	751	465
44	J8W	F	54	61	5,6,7	4	26	254	140
45	JTZ	M	50	59*	5,6,7	2	575	749	750
46	K2J	M	45	46	5,6,7	2	495	725	490
47	K9S	M	59	80*	5,6,7	1	911	655	751
48	KF0	M	67	70*	5,6	2	799	214	399
49	KRH	M	44	64*	5,6,7	2	964	702	580
50	L5Y	M	54	66	5,6,7	1	748	690	490
51	LBD	F	53	56	5,6,7	2	538	682	551
52	LIW	F	44	48	5,6,7	4	120.5	435	114
53	LMN	M	56	61	5,6,7	2	574	715	750
54	LR1	M	59	66	5,6,7	3	561	625	420
55	LS2	M	67	70	5,6,7	2	216	660	440
56	M4I	M	57	75	5,6,7	1	800	655	580
57	M5J	M	68	70	–	–	–	–	–
58	M6S	M	63	71	5,6,7	1	547	700	515
59	M8D	M	45	46	5,6,7	2	931	805	751
60	MHQ	M	71	72	5,6,7	2	199	450	515
61	MLF	M	72	72	5,6,7	2	887	501	499
62	MNE	M	69	78*	5,6,7	1	446.5	595	450
63	MUI	F	49	49	5,6,7	3	151	809	651
64	N00	M	41	41*	5,6,7	1	689.5	692	623
65	N5H	M	61	85*	5,6,7	2	291.5	695	530
66	N7Y	M	57	68	5,6,7	3	529	452	280
67	NPJ	M	53	55	5,6,7	1	751	675	645
68	O08	M	85	87*	5,6,7	2	830	801	702
69	PU1	M	58	75	6	–	–	730	698
70	Q9V	M	44	58*	5,6,7	1	671	730	451
71	QKM	M	72	72	5,6,7	2	915	500	400
72	QPR	M	52	56	5,6,7	1	909.5	782	662

Continued on next page

Table A.3: Tracheoesophageal (TE) speakers and perceptual consensus scores. Continued from previous page

	ID	Sex	Age	Age(2)	Chapter	AST	VQ	Int	Imp
73	QXM	M	44	55*	5,6,7	1	543.5	350	551
74	R49	M	60	67*	5,6,7	4	146	565	348
75	RX1	M	67	69	5,6,7	4	301	731	452
76	SAB	M	73	81*	5,6,7	4	501	685	585
77	SOZ	M	82	84*	5,6,7	2	549	715	549
78	SS6	M	56	56	5,6,7	4	224	605	320
79	T9B	M	45	48*	5,6,7	2	532	727	681
80	TLO	M	62	67	–	–	–	–	–
81	TMF	M	59	60	5,6,7	2	783.5	801	698
82	UCX	M	46	47*	5,6	4	315	100	350
83	USG	M	45	54	5,6,7	2	398	902	902
84	VEH	M	56	59*	5,6,7	2	625	655	410
85	VH8	M	57	71	5,6,7	4	201	550	201
86	VNR	M	48	51*	5,6,7	4	476.5	598	420
87	W7I	M	45	47*	5	2	503	–	–
88	WJ2	M	55	64*	5,6,7	3	401	650	551
89	WNV	M	74	81	–	–	–	–	–
90	WSY	M	75	77	5,6,7	2	452	585	395
91	WWL	M	74	75	5,6,7	2	860	749	751
92	YCF	M	70	73	5,6,7	2	299	350	209
93	YCL	M	64	65	5,6,7	4	140	705	451
94	Z6J	M	51	66	5,6,7	1	899.5	630	600
95	ZK9	M	53	53	5,6,7	4	300	610	470
96	ZOU	M	46	46	5,6,7	2	708	799	652
97	ZQ1	M	56	59*	5,6,7	2	644	848	801