Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 59 (2014) 44–54

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

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

Renee Clapham [a,b,*], Catherine Middag [c], Frans Hilgers [b,a], Jean-Pierre Martens [c], Michiel van den Brekel [b,a], Rob van Son [b,a]

[a] Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 210, 1012 VT Amsterdam, Netherlands
[b] Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, Netherlands
[c] Multimedia Lab ELIS, University of Gent, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium

Received 23 April 2013; received in revised form 18 December 2013; accepted 13 January 2014
Available online 23 January 2014

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).

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.

Keywords: Automatic evaluation; Head and neck cancer; Perceptual evaluation; Phonemic features; Phonological features; AMPEX

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

* Corresponding author at: Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 210, 1012 VT Amsterdam, Netherlands. Tel.: +31 205253805.
E-mail addresses: r.p.clapham@uva.nl (R. Clapham), Catherine.Middag@UGent.be (C. Middag), f.hilgers@nki.nl (F. Hilgers), martens@elis.ugent.be (J.-P. Martens), M.W.M.vandenBrekel@uva.nl (M. van den Brekel), r.v.son@nki.nl (R. van Son).

0167-6393/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2014.01.003

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; Middag et al., 2011; Middag et al., 2014; Van Nuffelen et al., 2009) and phonation (De Bruijn et al., 2009; De Bruijn et al., 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 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; Jacobi et al., 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).

2. General method

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.

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).

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).

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.

2.1.3.1. 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.