Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 37 (2016) 1–10

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

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

Renee P. Clapham [a,b,*], Jean-Pierre Martens [c], Rob J.J.H. van Son [b,a], Frans J.M. Hilgers [b,a], Michiel M.W. van den Brekel [b,a], Catherine Middag [c]

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

Received 22 April 2015; received in revised form 19 August 2015; accepted 11 October 2015
Available online 10 November 2015

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.

Keywords: Laryngectomy; Tracheoesophageal speech; Automatic speech recognition; Speech intelligibility; Voice quality; AMPEX

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 listener 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.

☆ This paper has been recommended for acceptance by Tatsuya Kawahara.

* Corresponding author at: Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 210, 1012 VT Amsterdam, The Netherlands. Tel.: +31 0205253805.

E-mail addresses: r.p.clapham@uva.nl (R.P. Clapham), martens@elis.ugent.be (J.-P. Martens), r.v.son@nki.nl (R.J.J.H. van Son), f.hilgers@nki.nl (F.J.M. Hilgers), M.W.M.vandenBrekel@uva.nl (M.M.W. van den Brekel), Catherine.Middag@UGent.be (C. Middag).

http://dx.doi.org/10.1016/j.csl.2015.10.001
0885-2308/

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 2 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 4.

2. Method

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 tape 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