Available online at www.sciencedirect.com
SciVerse ScienceDirect
Computer Speech and Language 28 (2014) 467–482
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

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

Catherine Middag [a,*], Renee Clapham [b,c], Rob van Son [c,b], Jean-Pierre Martens [a]

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

Received 5 March 2012; received in revised form 10 October 2012; accepted 29 October 2012
Available online 9 November 2012

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.

Keywords: Intelligibility; Pathological speech; Text-independent; Phonological features; Automatic intelligibility assessment

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).

* Corresponding author. Tel.: +32 9 264 33 97; fax: +32 9 264 35 94.
E-mail addresses: catherine.middag@elis.ugent.be (C. Middag), r.p.clapham@uva.nl (R. Clapham), r.v.son@nki.nl (R. van Son), martens@elis.ugent.be (J.-P. Martens).

0885-2308/$ – see front matter
http://dx.doi.org/10.1016/j.csl.2012.10.007

In a test on 10 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 (Hosom et al., 2004; Gu et al., 2005; Falk et al., 2011), 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, 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 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.

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 (Van Nuffelen, 2009; Kent et al., 1989). 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.

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

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