Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 54 (2012) 622–631

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility

Tiago H. Falk [a,*], Wai-Yip Chan [b], Fraser Shein [c,d]

[a] *Institut National de la Recherche Scientifique, INRS-EMT, Montréal, Canada*
[b] *Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada*
[c] *Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, Canada*
[d] *Department of Computer Science, University of Toronto, Toronto, Canada*

Available online 7 April 2011

Abstract

Objective measurement of dysarthric speech intelligibility can assist clinicians in the diagnosis of speech disorder severity as well as in the evaluation of dysarthria treatments. In this paper, several objective measures are proposed and tested as correlates of subjective intelligibility. More specifically, the kurtosis of the linear prediction residual is proposed as a measure of vocal source excitation oddity. Additionally, temporal perturbations resultant from imprecise articulation and atypical speech rates are characterized by short- and long-term temporal dynamics measures, which in turn, are based on log-energy dynamics and on an auditory-inspired modulation spectral signal representation, respectively. Motivated by recent insights in the communication disorders literature, a composite measure is developed based on linearly combining a salient subset of the proposed measures with conventional prosodic parameters. Experiments with the publicly-available 'Universal Access' database of spastic dysarthric speech (10 patient speakers; 300 words spoken in isolation, per speaker) show that the proposed composite measure can achieve correlation with subjective intelligibility ratings as high as 0.97; thus the measure can serve as an accurate indicator of dysarthric speech intelligibility.

Keywords: Dysarthria; Vocal source excitation; Temporal dynamics; Intelligibility; Linear prediction

1. Introduction

Dysarthria comprises a group of motor speech disorders resultant from damage to the central and/or peripheral nervous systems (Doyle et al., 1997). Dysarthric speech is often associated with excessive nasalization, disordered speech prosody, imprecise articulation, and variable speech rate (Doyle et al., 1997) – factors that often render speech unintelligible. One of the most common subtypes of dysarthria is termed "spastic dysarthria" with symptoms that can range from strained phonation, imprecise placement of articulators, incomplete consonant closure, monotone speech, and reduced voice onset time distinctions between voiced and unvoiced stops (Duffy, 2005). Spastic dysarthria is most commonly associated with cerebral palsy and traumatic brain injury (Duffy, 2005).

Currently, speech-language pathologists mainly rely on subjective intelligibility assessment tests to characterize the severity of speech disorders, as well as to monitor, plan treatment, and document changes in intelligibility over time (Klopfenstein, 2009). Subjective intelligibility tests, however, are costly, laborious, and subject to many intrinsic variables and biases due to e.g., familiarity with the patients and their speech pathologies (De Bodt et al., 2002; Van Nuffelen et al., 2009). Objective measurement, on the other hand, is economical and reliable (repeatable) and can assist in surgical and/or pharmacological treatment evaluation as well as in remote patient rehabilitation monitoring

[*] Corresponding author. Tel.: +1 514 875 1266x3066; fax: +1 514 875 0344.
E-mail address: tiago.falk@ieee.org (T.H. Falk).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.03.007

(Constantinescu et al., 2010). In fact, there is growing evidence suggesting that clinicians are becoming more receptive to automated machine-based systems that assist in treatment decisions (e.g., Hill et al., 2006; Maier et al., 2009).

In the past, a handful of objective intelligibility measures have been proposed for dysarthric speech. The system proposed by Middag et al. (2009) used phonemic and phonological features that were force-aligned to the acoustic-phonetic transcription of the target word. Alignment was achieved by means of an automatic speech alignment algorithm trained on acoustic models of "healthy" speech. Features were then mapped to an intelligibility score using a linear regression function. Additionally, the work described by Gu et al. (2005) computed distance measures (e.g., Itakura–Saito distortion) between the produced disordered speech utterance and the same utterance spoken by a healthy individual. To account for differences in utterance durations, dynamic time warping was applied.

Today, automatic speech recognition (ASR) has become a popular method of objectively quantifying dysarthric speech intelligibility for speakers with mild or moderate dysarthria (e.g., Doyle et al., 1997; Ferrier et al., 1995; Maier et al., 2009; Sharma et al., 2009); technological advances, however, are still needed before ASR is used for severe dysarthric speakers (Middag et al., 2009; Rudzicz, 2007). Major limiting factors in the widespread use of ASR, however, include limited vocabulary sizes ranging from 10–70 words (Doyle et al., 1997), the need for speaker-dependent (or adaptive) acoustic models (Raghavendra et al., 2001; Rudzicz, 2007), and the sparseness of available data needed to accurately train such models (Green et al., 2003).

The methods mentioned above require *a priori* information, such as the signal or feature prototypes of the target word being uttered. In many practical applications, however, such information may not be available and "blind" measures are more convenient. The majority of existing blind methods rely on prosodic measurements, such as fundamental frequency ($f0$) variation, tone unit duration, and second-formant slope transitions (Bunton et al., 2000; Schlenck et al., 1993; Kent et al., 1989), which have been shown to be useful indicators of dysarthric speech intelligibility (Klopfenstein, 2009). Recently, the power spectrum of the envelope of the speech signal, or modulation spectrum, was used to characterize rhythmic disturbances in dysarthric speech. The study suggested that the perturbations of speech temporal patterns associated with dysarthria played an important role in intelligibility (LeGendre et al., 2009).

Subjective listening tests of dysarthric speech suggest that intelligibility can be expressed as a weighted linear combination of different perceptual dimensions, such as articulation, vocal harshness, prosody, and nasality (De Bodt et al., 2002). In this paper, several parameters are proposed and tested as correlates of subjective intelligibility. The parameters measure abnormal behaviours found in dysarthric speech, such as vocal source excitation oddity, temporal dynamics perturbations, hypernasality, and disordered prosody. Our results (see Section 3.3) suggest that the measures are complementary and when linearly combined can serve as an accurate indicator of dysarthric speech intelligibility. Moreover, in comparison with ASR, the proposed method is considerably simpler to design and implement. The remainder of this paper is organized as follows: Section 2 describes the proposed measures; experimental setup and results are reported in Section 3; and discussion and conclusions are presented in Sections 4 and 5, respectively.

2. Objective measurement of spastic dysarthric word intelligibility

Several factors are known to adversely affect speech intelligibility for individuals with dysarthria (De Bodt et al., 2002). The most prominent are associated with atypical vocal source excitation (e.g., vocal harshness), temporal dynamics (e.g., unclear distinction between adjacent phonemes due to imprecise placement of articulators), hypernasality, and disordered prosody (e.g., monotonicity). In order to blindly assess speech intelligibility, measures need to be developed such that perturbations in typical vocal source excitation, temporal dynamics, nasality, and prosody can be characterized. In this paper, several such measures are proposed and tested as correlates of subjective intelligibility.

2.1. Separation of vocal source and vocal tract information

Linear prediction (LP) analysis has been widely used in speech applications to separate vocal source (glottal) excitation, $u(n)$, and vocal tract modulation, $h(n)$, from the produced speech signal, $s(n) = u(n) * h(n)$, where "*" indicates convolution (Benesty et al., 2008). Commonly, the vocal tract is modeled as a time-varying all-pole filter given by

$$H(z) = \frac{1}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{1}$$

Hence, the produced speech signal $s(n)$ can be approximated by

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n). \tag{2}$$

The coefficients a_k, $k = 1, \ldots, p$, of the all-pole filter depend on the shape and resonant characteristics of the vocal tract and determine the spectral characteristics of the particular sound being generated. The excitation signal, in turn, is approximated either as a quasi-periodic train of impulses for voiced speech segments, random noise for unvoiced segments, or a combination thereof for voiced fricatives (e.g., 'v') (Benesty et al., 2008); the multiplicative factor G is the gain applied to the excitation signal.

Linear prediction analysis assumes that the current signal sample can be predicted by a linear combination of p previous samples. The predicted sample $\hat{s}(n)$ is given by