Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 61–62 (2014) 17–32

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Measurement of signal-to-noise ratio in dysphonic voices by image processing of spectrograms

Maurílio N. Vieira [a,*], João Pedro H. Sansão [b,c], Hani C. Yehia [a]

[a] Departamento de Engenharia Eletrônica, Universidade Federal de Minas Gerais, Avenida Antônio Carlos, 6627, CEP 31.270-010 Belo Horizonte, MG, Brazil

[b] Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais, Avenida Antônio Carlos, 6627, CEP 31.270-901 Belo Horizonte, MG, Brazil

[c] Departamento das Engenharias de Telecomunicações e Mecatrônica, Universidade Federal de São João del-Rei (Campus Alto-Paraopeba), Rodovia MG 443, km 7, CEP 36420-000 Ouro Branco, MG, Brazil

Received 8 May 2013; received in revised form 5 February 2014; accepted 2 April 2014
Available online 18 April 2014

Abstract

The measurement of glottal noise was investigated in human and synthesized dysphonic voices by means of two-dimensional (2D) speech processing. A prime objective was the reduction of measurement sensitivities to fundamental frequency (f_0) tracking errors and phonatory aperiodicities. An available fingerprint image enhancement algorithm was used for signal-to-noise measurement in narrow band spectrographic images. This spectrographic signal-to-noise ratio estimation method (S^2NR) creates binary masks, mainly based on the orientation field of the partials, to separate energy in regions with strong harmonics from energy in noisy areas. Synthesized vowels with additive noise were used to calibrate the algorithm, validate the calibration, and systematically evaluate its dependence on f_0, shimmer (cycle-to-cycle amplitude perturbation), and jitter (cycle-to-cycle f_0 perturbation). In synthesized voices with known signal-to-noise ratios in the 5–40 dB range, S^2NR estimates were, on average, accurate within ±3.2 dB and robust to variations in f_0 (120 Hz or 220 Hz), jitter (0–3%), and shimmer (0–30%). In human /a/ produced by dysphonic speakers, S^2NR values and perceptual ratings of breathiness revealed a non-linear but monotonic decay of S^2NR with increased breathiness. Comparison between S^2NR and related acoustic measurements indicated similar behaviors regarding the relationship with breathiness and immunity to shimmer, but the other methods had marked influence of jitter. Overall, the S^2NR method did not rely on accurate f_0 estimation, was robust to vocal perturbations and largely independent of vowel type, having also potential application in running speech.

Keywords: Signal-to-noise ratio; Breathiness; Dysphonic voice; 2D speech processing

1. Introduction

There is a vast literature on acoustic analysis of dysphonic voices. In general, investigations on the clinical use of acoustic measurements have described objective parameters that (i) *discriminate dysphonic from non-dysphonic speakers* (e.g., Lieberman, 1963; Schoentgen, 1982; Hirano et al., 1988; Zhang et al., 2005); (ii) *correlate with perceptual ratings of vocal quality* (e.g., Kane and Wellen, 1985; Eskenazi et al., 1990; Feijoo and Hernandez, 1990; Martin et al., 1995); or (iii) *permit the longitudinal monitoring of patients* (e.g., Muta et al., 1988; Dejonckere and Wieneke, 1994). These studies used mostly sustained vowels and measurements of jitter (cycle-to-cycle fundamental frequency perturbations), shimmer (cycle-to-cycle

* Corresponding author. Tel.: +55 31 3409 3456; fax: +55 31 3409 4850.
E-mail addresses: maurilionunesv@cpdee.ufmg.br (M.N. Vieira), jsansao@gmail.com (João Pedro H. Sansão), hani@cpdee.ufmg.br (H.C. Yehia).

http://dx.doi.org/10.1016/j.specom.2014.04.001
0167-6393/

amplitude perturbations), and signal- or harmonic-to-noise ratios (*SNR* or *HNR*).

Jitter and shimmer contribute auditorially to the perceptual attribute usually called harshness, while glottal noise, caused by the turbulent passage of air through incomplete glottal closure, relates to the perception of breathiness (Laver et al., 1992; Klatt and Klatt, 1990; Chen et al., 2013). Despite their wide use, acoustic measurements of phonatory dysfunctions have a paradoxical behavior (Bielamowicz et al., 1996), that is, they lack reliability with increasing levels of the quantity to be measured. This may not be a problem for the differentiation between dysphonic and non-dysphonic voices, but the automatic evaluation of voice quality or the longitudinal monitoring of patients require acoustic parameters reliably related to the degree of vocal dysfunction. Moreover, the measurement of a certain parameter should not be affected by the coexistence of other signal perturbations. Jitter and shimmer, for example, contaminate most *SNR* measurements, while glottal noise inflates measurements of cycle-to-cycle perturbations (e.g., Hillenbrand, 1987; Muta et al., 1988; Qi, 1992; de Krom, 1993; Qi et al., 1995; Murphy, 1999; Vieira et al., 2002).

The literature on the quantification of vocal perturbations (see Murphy, 1999 for a comprehensive review) reports a number of *SNR* estimation methods usually based on signal analysis in *time* (e.g., Yumoto et al., 1982; Kasuya et al., 1986a), *frequency* (e.g., Hiraoka et al., 1984; Kasuya et al., 1986b), or *cepstrum domains* (e.g., de Krom, 1993; Murphy, 1999; Murphy and Akande, 2007). Yumoto et al. (1982) were the precursors of *SNR* measurements in dysphonic voices. In their method, 50 successive glottal cycles are pitch-synchronously averaged to yield the periodic component, while glottal noise is estimated by subtracting each cycle from the mean cycle. Variations in the duration of the cycles are dealt with by truncation or zero-padding but this strategy inflates the measurements. Kasuya et al. (1986a) described an alternative pitch-synchronous algorithm based on adaptive comb-filters. The filters, tuned by measurements of the fundamental period, extract the periodic component of the signal, while noise is obtained by subtracting the filtered signal from the original signal. The unreliable demarcation of glottal cycles in dysphonic voices is a major limitation of time-domain methods. Kasuya et al. (1986b) also described a frequency-domain method where the spectrum is computed from 7 successive glottal cycles. In this approach, the noise component is estimated from the inter-harmonic energy while the periodic component is the energy around the spectral peaks. The correct estimation of noise in the spectral valleys is a central problem in spectral domain methods. In the cepstrum-based method originally proposed by de Krom (1993), glottal noise is estimated after the cepstral peaks (i.e., periodic components) are removed. Other methods for *SNR* estimation in dysphonic voices are considered in Section 3.5.

It is known that reported *SNR* estimation methods can perform well over a wide range of additive noise but are sensitive to *jitter and shimmer* (Yumoto et al., 1982; Hillenbrand, 1987; Muta et al., 1988; Qi, 1992; de Krom, 1993; Qi et al., 1995; Murphy, 1999, 2000), *recording apparatus* (Titze and Winholtz, 1993), or *vowel formant structure* (Cox et al., 1989). Although strategies for reducing such artifacts have been described (e.g., Hillenbrand, 1987; Qi, 1992; Qi et al., 1995; Murphy, 1999), estimated *SNR* values are not necessarily corroborative evidence of glottal noise. On the other hand, clinicians have long relied on visual inspections of narrow-band spectrograms to predict the severity of voice disorders (Yanagihara, 1967; Wolfe and Steinfatt, 1987) because such relevant features as voice harmonic structure, interharmonic noise, phonatory breaks, or subharmonics can be promptly seen in the images. The reliability of the information in spectrographic images of speech is a major motivation for the present study.

This paper describes a *SNR* measurement method based on spectrographic image processing. Image processing of speech have been used for speech modification and resynthesis (Horn, 1998), speech enhancement (Soon and Koh, 2003; Ding et al., 2009), formant and fundamental frequency estimation (Ezzat et al., 2007), and speech coding (Jellyman et al., 2009). The application described here is based on an algorithm originally developed for fingerprint image enhancement (Hong et al., 1998; Bazen, 2002). Similarly to spectrograms, fingerprints have nearly parallel line structures (formed by skin ridges and valleys), noise (caused by data collection artifacts, aberrant formations, or occupational marks), and peculiar minutiae (mainly ridge endings and bifurcations). The robust tracking of fingerprint line structures by this algorithm motivated its use in the identification of harmonic structures in spectrograms. In the present study, Kovesis' (2005) implementation of the algorithm (Hong et al., 1998) was applied to *SNR* estimation in narrowband spectrograms of synthesized and human dysphonic voices. The aim is a voice parameter reliably related to the level of glottal noise and independent of vowel formant patterns and vibratory aperiodicities.

The paper is organized as follows. Section 2 reviews the fingerprint enhancement algorithm and describes its use for *SNR* measurement. Section 3 deals with the synthesized and natural voice stimuli used to evaluate the algorithm. Section 4 presents and discusses the results of the calibration and systematic evaluations. The concluding remarks are presented in Section 5.

2. Spectrographic signal-to-noise ratio

Fig. 1 gives a schematic view of the method, henceforth spectrographic signal-to-noise ratio (S^2NR) algorithm. The various blocks and variables mentioned in the diagram are discussed throughout this section. Briefly, the binary mask M_{signal} (black = 1, white = 0) and its complement, M_{noise},