Biomedical Signal Processing and Control 47 (2019) 350–357

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Effect of vowel context in cepstral and entropy analysis of pathological voices

Andreas Selamtzis [a,*], Antonella Castellana [b], Giampiero Salvi [a], Alessio Carullo [b], Arianna Astolfi [c]

[a] Department of Speech, Music, and Hearing (TMH), KTH Royal Institute of Technology, Stockholm, Sweden
[b] Department of Electronics and Telecommunications, Politecnico di Torino, Italy
[c] Department of Energy, Politecnico di Torino, Italy

ARTICLE INFO

Article history:
Received 26 December 2017
Received in revised form 28 June 2018
Accepted 20 August 2018

Keywords:
Dysphonia
Voice analysis
Cepstral peak prominence
Sample entropy
Vowel context

ABSTRACT

This study investigates the effect of vowel context (excerpted from speech versus sustained) on two voice quality measures: the cepstral peak prominence smoothed (CPPS) and sample entropy (SampEn). Thirty-one dysphonic subjects with different types of organic dysphonia and thirty-one controls read a phonetically balanced text and phonated sustained [a:] vowels in comfortable pitch and loudness. All the [a:] vowels of the read text were excerpted by automatic speech recognition and phonetic (forced) alignment. CPPS and SampEn were calculated for all excerpted vowels of each subject, forming one distribution of CPPS and SampEn values per subject. The sustained vowels were analyzed using a 41 ms window, forming another distribution of CPPS and SampEn values per subject. Two speech-language pathologists performed a perceptual evaluation of the dysphonic subjects' voice quality from the recorded text. The power of discriminating the dysphonic group from the controls for SampEn and CPPS was assessed for the excerpted and sustained vowels with the Receiver-Operator Characteristic (ROC) analysis. The best discrimination in terms of Area Under Curve (AUC) for CPPS occurred using the mean of the excerpted vowel distributions (AUC=0.85) and for SampEn using the 95th percentile of the sustained vowel distributions (AUC=0.84). CPPS and SampEn were found to be negatively correlated, and the largest correlation was found between the corresponding 95th percentiles of their distributions (Pearson, r=−0.83, $p < 10^{-3}$). A strong correlation was also found between the 95th percentile of SampEn distributions and the perceptual quality of breathiness (Pearson, r=0.83, $p < 10^{-3}$). The results suggest that depending on the acoustic voice quality measure, sustained vowels can be more effective than excerpted vowels for detecting dysphonia. Additionally, when using CPPS or SampEn there is an advantage of using the measures' distributions rather than their average values.

1. Introduction

Laryngeal pathologies often result in irregularities and noise in the voice signal, such as aperiodicity, breathiness, and fundamental frequency breaks. There is great potential in using objective acoustic measures for quantifying voice quality in clinical practice. Such measures can be used to support the diagnostic process, as well as the monitoring of the post-therapy (or -surgery) progress of a vocal patient. When there is lack of periodicity, conventional metrics of voice quality such as *jitter* and *shimmer* are difficult, or meaningless to compute for disordered voice signals [1]. Therefore, in analyzing pathological voices, it is advantageous to use measures that do not depend on detecting glottal cycle boundaries.

Sustained vowels at comfortable pitch and loudness are often used in the clinic for endoscopic examination, perceptual evaluation, and acoustic quantification of voice quality. However, sustained vowels do not constitute an appreciable part of everyday voice use, at least for non-singers [2]. Running speech on the other hand commonly occurs in real life situations, i.e., it is a natural and ecologically valid signal that could serve as a basis for perceptual assessment and acoustic analysis [2]. Using running speech though is not as straightforward as using sustained vowels, since the voiced parts of speech are rather short, and the phonetic context of the vowels can affect objective voice quality measures [3].

Several earlier studies have investigated how different perceptual or acoustic measures depend on the vowel context [3–6].

* Corresponding author.
E-mail address: selamt@kth.se (A. Selamtzis).

https://doi.org/10.1016/j.bspc.2018.08.021
1746-8094/

Gerratt et al. [3] concluded that when analyzing or evaluating perceptually either sustained vowels or vowels excerpted from continuous speech, the information on deviation from normal voice quality was the same. The aim of the present study is to investigate how vowel context (sustained versus excerpted) affects the predictive power for dysphonia of two objective voice quality measures, i.e., the cepstral peak prominence smoothed (CPPS) and the sample entropy (SampEn).

The cepstral peak prominence smoothed (CPPS) [7], is a measure based on the cepstrum [8] that has been used as an indicator of voice quality. The computation of the cepstrum of digitized signals relies on the Discrete Fourier Transform, and does not require any detection of glottal cycles. CPP is known to be affected by amplitude and frequency perturbations of the analyzed signal, as well as the presence of aerodynamic noise [9]. The smoothed version of CPP (CPPS), has been found to correlate with breathiness, i.e., the perception of aerodynamic noise in the voice signal [7]. A low value for CPPS signifies a lower prominence of the cepstral peak, which correlates with degraded voice quality. Previous studies have established that CPPS correlates with perceptual measures of the GRBAS (Grade, Roughness, Breathiness, Asthenia, Strain) scale in acoustic material from text readings [10,11] or sentences [12]. Specifically, Brinca et al. [10] found that in text readings CPPS correlated with breathiness (Spearman ρ=−0.43), but none of the other perceptual measures. Jannetts et al. [11] used text readings and obtained the highest correlation with asthenia (Pearson, r=−0.47), followed by r=−0.38 for breathiness, and r=−0.35 for roughness. Heman-Ackah et al. [12] limited their investigation to a sentence considering only breathiness and roughness; they found that both perceptual qualities correlated with CPPS, with a coefficient of Pearson's r=−0.71 for breathiness and r=−0.50 for roughness.

Signals originating from disordered biological systems are likely to present irregularities. These irregularities can be quantified using time-domain based entropy measures, such as *sample entropy* (SampEn) and *approximate entropy* (ApEn). SampEn was introduced by Richman and Moorman [13] as an improved version of Pincus' approximate entropy (ApEn) [14,15]. SampEn and ApEn have been extensively used in biomedical signal processing, in a variety of contexts, such as heart rate variability [13,16], brain activity in newborns [17], and postural sway [18]. A signal that is completely predictable and regular exhibits a lower SampEn value than an irregular signal that contains random occurrence of noise bursts, or stationary noise [19]. Few studies have explored the utility of ApEn and SampEn for pathological voice analysis. ApEn has been used for analyzing electroglottographic signals [20–22], and SampEn for both electroglottograms and acoustic signals [23–25]. Occurrence of noise and other irregularities in pathological voices are expected to be reflected in higher SampEn values, as compared to normal voices. Fabris et al. [23] computed SampEn for one second long sustained [a:] vowels, and found that SampEn differed significantly in pathological voices compared to controls. Londoño et al. [24] computed SampEn from sustained [a:] vowels using windows of 200 ms, and used it as input feature to a pre-trained Gaussian Mixture Model–based classifier. They also reported higher mean SampEn for the pathological group compared to controls. Their SampEn-based classifier discriminated pathological from normal voices with an accuracy of 87% (sensitivity 94%, specificity 87%).

Despite the use of SampEn for quantifying irregularity in voices, its relationship to perceptual ratings of voice quality has not been studied before. In addition, it is not clear how phonetic context of vowels may affect SampEn and CPPS for their ability to discriminate between healthy and pathological subjects. In a previous study [25] it was found that for excerpted [a:] vowels, the mean of CPPS distributions had a greater predictive power for dysphonia over mean SampEn, and that mean CPPS was significantly correlated with mean SampEn (Spearman, ρ = −0.6). The aim of the present study is to investigate the effect of vowel context on the predictive power of CPPS and SampEn, using both excerpted and sustained vowels. As in earlier studies [25–27], the individual distributions of the two metrics (CPPS and SampEn) are taken into account and their statistics are evaluated as potentially more effective descriptors of vocal health than mean values. Additionally, correlations of CPPS and SampEn distributions with perceptual assessment of voice quality are presented.

Table 1
Diagnoses for the patient group.

Organic dysphonia	Number
Cyst	5
Edema	9
Sulcus vocalis	3
Polyp	4
Chronic laryngitis	2
Vocal fold hyposthenia	3
Vocal fold paresis	2
Vocal fold nodule	1
Post-surgery dysphonia	2
Overall	*31*

2. Materials and methods

2.1. Data acquisition and perceptual evaluation

The data comprised voice samples from 31 voluntary patients (24 females and 7 males) and 31 controls (17 females and 14 males). All speakers were native Italian speakers. All patients were diagnosed by two otolaryngologists with some form of organic dysphonia, as shown in Table 1.

Two tasks were performed by both the patient and the control groups:

(a) The reading of a standardized phonetically balanced Italian text of 300 words length [28].
(b) The production of the sustained vowel [a:], at comfortable pitch and loudness.

The acoustic signal was recorded using an omnidirectional head mounted microphone (model MU-55HN, Mipro Electronics, Chiayi, Taiwan) with an approximate distance of 2.5 cm from the speaker's mouth, slightly to the side at about 20°–45° horizontally, depending on the subject's face shape. The microphone was connected to a bodypack transmitter (model ACT-30 T, Mipro Electronics, Chiayi, Taiwan), which transmitted the signal to a wireless system (model ACT 311, Mipro Electronics, Chiayi, Taiwan). The signal was recorded using a portable recorder (model H1 "Handy Recorder", Zoom Corp., Tokyo, Japan) with a sampling rate of 44.1 kHz and 16 bit resolution. All voice signals were recorded in a quiet room with an A-weighted equivalent background noise level of 50.0 dB (std = 2 dB), measured with a sound level meter (model XL2, NTi Audio AG, Schaan, Liechtenstein), over a period of 5 min for each recording session. According to Šrámková et al. [29], the softest vowel sounds produced by healthy males and females had A-weighted levels of 39 dB (60 dB) and 44 dB (65 dB) respectively at 30 cm (2.5 cm). This suggests that the background noise level of 50 dB should guarantee at least a 10 dB signal-to-noise ratio or more, since the subjects were instructed to read aloud.

Two expert speech-language pathologists rated the recordings of the text reading of each patient. Ratings were discussed, and consensus was reached using the perceptual Stockholm Voice Evaluation Approach (SVEA) visual analogue scale [30] with ratings for the qualities of *aphonia, breathiness, hyperfunction, hypofunction, vocal fry or creaky, roughness, high pitch roughness, instability, voice*