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

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

Andreas Selamtzis[1], Antonella Castellana[2], Giampiero Salvi[1], Alessio Carullo[2], Arianna Astolfi[2]

[1]Department of Speech, Music, and Hearing (TMH), KTH Royal Institute of Technology, Stockholm, Sweden

[2]Department of Electronics and Telecommunications, Politecnico di Torino, Italy

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.

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]. 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 of different lengths. Based on recent 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.

Materials and methods

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 documented in Table I.

Table I: 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*

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

(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-30T, 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 = 2dB), measured with a sound level meter (model XL2, NTi Audio AG, Schaan, Liechtenstein), over a period of 5 minutes 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 (2.5cm). 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 breaks, diplophonia*. An explanation of the voice quality parameters used for perceptual evaluation is adapted from Hammarberg [31] and presented in Table II.

Table II. Definition of voice quality parameters used in perceptual evaluation according to the Stockholm Voice Evaluation Approach (SVEA). Adapted from Hammarberg et al. [31].

Voice quality parameter	**Tentative definition**
Aphonic/intermittent aphonic	voice is constantly or intermittently lacking phonation, i.e. there are moments of whisper or loss of voice
Breathy	voice is produced with insufficient glottal closure, vocal folds are vibrating, but somewhat abducted, which creates an audible turbulent noise in the glottis
Hyperfunctional/tense	voice sounds strained, due to compression/constriction of vocal folds and larynx tube during phonation with insufficient airflow
Hypofunctional/lax	opposite to hyperfunctional, insufficient vocal fold tension and laryngeal muscle activity, resulting in a weak and slack voice
Vocal fry/creaky	low-frequency aperiodic/periodic vibration, vocal folds are very close together and only a section of them is free to vibrate; also known as pulse register
Rough	low-frequency aperiodicity, presumably related to some kind of irregular vocal fold vibrations
Gratings/‘scrapiness’	high-frequency aperiodicity, presumably related to some kind of irregular vocal fold vibrations
Unstable voice quality/pitch	voice is fluctuating in pitch or in voice quality over time
Voice breaks	intermittent breaks between modal and falsetto register
Diplophonic	two different pitches can be perceived simultaneously

Data processing

The following signals were analyzed for each subject: excerpted [a:] vowels from the text reading and one sustained, two-seconds-long [a:] vowel, at comfortable pitch and loudness. Only [a:] vowels were considered in order to minimize acoustic variability due to different articulation, and because SampEn values can depend on the shape of the waveform of the signal under study. Therefore, using the [a:] vowel allowed the analyzed pattern to remain as consistent as possible.

Vowel extraction and analysis

As described in an earlier study [25], the recordings of the read text were phonemically annotated using automatic speech recognition. In brief, the orthographic transcriptions (prompts) were obtained by means of forced alignment, and the speech recognizer was trained on the Italian SpeechDat corpus [32] using the Hidden Markov Model Toolkit (HTK) [33] and the RefRec scripts [34]. The transcriptions were used to extract all the [a:] vowels for computing CPPS and SampEn. Because only the excerpted vowels with a length of at least 50 ms were considered for analysis and the number of analyzed vowels was different for each subject, between 111 and 188 vowels. For one pathological subject the number of excerpted vowels was only 87, markedly less than average, therefore that particular subject was excluded from the analysis. All analyzed [a:] vowels were downsampled from 44.1 kHz to 25 kHz, and their middle 1024 samples (41 ms) were analyzed, to avoid onset and offset transients. For each excerpted [a:] vowel, a value for CPPS and a value for SampEn was calculated, resulting in one distribution for CPPS and one for SampEn per subject. For each of these distributions, descriptive statistics (mean, median, standard deviation [std], range, 5^{th} percentile, 95^{th} percentile), and the ROC curve with the corresponding AUC were computed.

For the sustained vowels, the middle 2 seconds were analyzed using a 1024 point (41 ms) sliding window with a step size of 2 ms. For each window, CPPS and SampEn were calculated, resulting in one distribution for CPPS and one for SampEn, for each subject. Similarly to the excerpted vowels, descriptive statistics were calculated and the ROC analysis was carried out. Figure 1 provides pseudocode to illustrate the calculation procedure for the CPPS and SampEn in sustained and excerpted vowels.

```
For each sustained vowel:
    Take the middle 2 seconds
    Compute CPPS using a window of 41 ms with a step of 2 ms
    Compute SampEn using a window of 41 ms with a step of 2 ms
    Obtain individual distributions Ds of CPPS and SampEn
    Evaluate the descriptive statistics of Ds
```

```
For each reading text:
    Perform the phonemic annotation using ASR
    Extract all the [a:] vowels with length of at least 50 ms
        For each [a:] vowel:
        Compute CPPS in the middle 41 ms
        Compute SampEn in the middle 41 ms
    Obtain individual distributions De of CPPS and SampEn
    Evaluate the descriptive statistics of De
```

Figure 1: Pseudocode describing the calculation procedure for CPPS and SampEn metrics per subject, in both sustained and excerpted vowels.

Cepstral peak prominence smoothed (CPPS)

The CPPS in sustained vowels was estimated as described in Castellana et al. [27], with two smoothing processes, i.e., a 7-frames smoothing in time and a 7-bin smoothing in quefrency.

The CPPS in excerpted vowels was calculated based on a modified version of the definition provided by Hillenbrand et al. [7]. Because the excerpted vowels were not long enough for time-smoothing the cepstra, that step was omitted. The algorithm for obtaining CPPS consisted of the following steps. First, each vowel was multiplied with a hamming window of 1024 points, and the Fast Fourier Transform was computed twice: once on the signal in time, and then on the log power spectrum, to obtain the cepstrum. The cepstrum was smoothed over quefrency with a seven-point averaging window. A regression line was calculated in the quefrency vs cepstral magnitude domain, from 1 ms to the maximum quefrency. One millisecond was taken as lower limit, because quefrencies below 1 ms are affected more by the spectral envelope than by the regularity of the harmonics [35]. CPPS was then calculated as the level difference (in dB) between the maximum cepstrum peak, and the value of the regression line at the same quefrency. The peak search was limited to the range from 3.3 ms to 16.7 ms, corresponding to fundamental frequencies of 300 Hz and 60 Hz, respectively. For the calculation of CPPS, a custom MATLAB (The MathWorks, Inc., Natick, MA) script was used.

Sample entropy

For a time series *y* of *N* samples, SampEn is calculated by using the values of the time series to define vectors of dimension of *m* and *m*+1, where *m* is a parameter known as template length. The components of these vectors are consecutive values of the time series, so that each value occupies a different dimension. The total number of defined vectors is *N*-*m*. Then, the conditional probability is calculated that the Chebyshev distance[1] between two vectors of length *m*+1 is less or equal than a matching tolerance *r*, given that the Chebyshev distance between two vectors of template length *m* is less or equal than the tolerance *r*. The probability is calculated by counting the respective pairs of vectors (excluding self-pairs) of dimension *m*, and dimension *m*+1 that satisfy the tolerance condition, and obtaining the ratio of their count. SampEn(*m*, *r*) is defined as the negative natural logarithm of this conditional probability as in Eq. (1), where $d(x_i, x_j)$ is the Chebyshev distance of a pair of vectors, and '#' symbolizes the number of vector pairs for $i \neq j$ with *d* less or equal than r.

$$\mathrm{SampEn}(y, m, r) = -\ln\left(\frac{\#\, d\left(x_i^{(m+1)}, x_j^{(m+1)}\right) \leq r}{\#\, d\left(x_i^{(m)}, x_j^{(m)}\right) \leq r}\right) \quad (1)$$

The template length *m* for the analyzed excerpted and sustained vowels was set to be the closest integer to $\log_{10}(N)$, where *N* is the window length used for analysis, as suggested by Fabris et al. [23] and Pincus [14,15]. In this study, the window was 1024 points long, which results in *m*=3. In the choice of the parameter *m* there is an implicit choice of the time-scale and frequency band for which SampEn is most effective in tracing irregularity. Templates of length *m*=3 have a length of τ=0.12 ms, and the corresponding frequency is $f_c=1/\tau$=8333 Hz. This high frequency belongs to a frequency band where additive noise, such as breathiness, can be stronger than the harmonic component of the voice source. Therefore, the choice of *m*=3 is expected to make SampEn sensitive to the presence of aerodynamic noise. SampEn($m=f_s/f_c$, *r*) is proportional to the inverse of the sampling frequency f_s; to compare SampEn values for signals acquired with different sampling frequencies, SampEn should be multiplied by f_s[2]. Here however, we will present the values unnormalized. The matching tolerance *r* was set equal to 0.1 times the standard deviation of the analyzed excerpted vowel, or window of sustained vowel [14,15,22]. A detailed explanation for the calculation algorithm of SampEn can be found in Fabris et al. [23]. SampEn was

[1] The Chebyshev distance between two *k*-dimensional vectors is defined as the largest of the absolute values of the *k* coordinate differences.

[2] If a sampling frequency f_s other than 25000 Hz is used, to make the SampEn values comparable with those reported in the present study, SampEn(f_s/f_c, *r*) should be multiplied by $f_s/25000$.

computed using a custom MATLAB (The MathWorks, Inc., Natick, MA) script, based on the implementation of the algorithm by Lake et al. [36].

Statistical analysis

The two-tailed Mann-Whitney U-test, a nonparametric test based on independent samples [37], was applied on each coupled list of descriptive statistics related to the group of patients and the control subjects. The null hypothesis (H0) states that MD = 0, where MD is the median of the population of the differences between the data sample for the two groups of patients and controls. If H0 is accepted, the two lists of values seem to come from the same population, i.e., data from healthy and unhealthy subjects is not significantly different, thus not allowing a possible discrimination. The one-sample Kolmogorov-Smirnov test verified that data in each list did not come from a normal distribution. The tests were performed using the MATLAB (The MathWorks, Inc., Natick, MA) environment.

The receiver-operator characteristic (ROC) analysis [38] was carried out for each descriptive statistic of CPPS and SampEn distributions from both excerpted and sustained vowels. ROC analysis is used to characterize the quality of the binary classification (healthy versus dysphonic) based on either CPPS or SampEn. Since distributions of the voice quality measures in healthy and dysphonic subjects typically overlap, one cannot simply select a threshold value to decide the presence or absence of dysphonia. Instead, the ROC curve is constructed, by varying the threshold value and plotting the resulting false positive rate (1-specificity) on the *x*-axis, versus the true positive rate (sensitivity) on the *y*-axis. The area under the ROC curve (AUC) is considered a metric of classification accuracy for the given voice quality measure. The AUC obtains values between 0.5 and 1, where values higher than 0.9 designate outstanding accuracy, values between 0.8 and 0.9 excellent accuracy, values between 0.7 and 0.8 acceptable accuracy, and values close to 0.5 imply poor accuracy.

To evaluate the accuracy of separating dysphonic subjects from controls, a leave-one-out validation scheme was computed using the statistical software R studio (version 3.5.0). The leave-one-out procedure is carried out by excluding one subject from data set, calculating a threshold based on the ROC curve from the data set of the remaining subjects, and classifying the excluded subject based on that threshold, to either the dysphonic or the control group. This scheme ensures that the classified data is different from the data that was used to establish a threshold belong to different groups. The procedure is repeated for all subjects and the percentage of correctly classified subjects is reported as the leave-one-out predictive accuracy.

Results

The *p*-values of the Mann-Whitney U-tests between the dysphonic and control group, for different descriptive statistics of the CPPS and SampEn distributions, as well as the corresponding AUC are reported in Table III. It can be seen that most descriptive statistics of CPPS and SampEn distributions from excerpted and sustained vowels, exhibit significant differences between the dysphonic and control groups. The smallest *p*-values occurred most commonly for the mean, median, 5^{th} percentile and 95^{th} percentile.

Table III. Results in terms of *p*-values of the Mann-Whitney test (U-test), AUC with its confidence intervals (CI), and leave-one out accuracy (Acc.) for different descriptive statistics. The *p*-values that are significant at the 1% level are designated with a star (*). The highest AUC and Acc. per measure and descriptive statistic are designated with boldface.

	Sustained						Excerpted					
Statistics	**CPPS**			**SampEn**			**CPPS**			**SampEn**		
	U-test	**AUC (CI)**	**Acc.**	**U-test**	**AUC (CI)**	**Acc.**	**U-test**	**AUC (CI)**	**Acc.**	**U-test**	**AUC (CI)**	**Acc.**
Mean	*<10^{-3}	0.78 (0.66-0.89)	66%	*<10^{-3}	0.80 (0.68-0.91)	76%	*<10^{-3}	**0.85 (0.75-0.95)**	**75%**	*0.003	0.72 (0.59-0.86)	66%
Median	*<10^{-3}	0.77 (0.65-0.88)	63%	*<10^{-3}	0.79 (0.66-0.90)	74%	*<10^{-3}	0.84 (0.73-0.94)	**75%**	*0.002	0.73 (0.60-0.86)	**67%**
Std	*0.012	0.68 (0.55-0.82)	65%	*<10^{-3}	0.81 (0.71-0.92)	68%	0.488	0.55 (0.40-0.70)	49%	0.090	0.63 (0.49-0.77)	57%
Range	0.037	0.66 (0.52-0.79)	57%	*<10^{-3}	0.83 (0.72-0.93)	71%	*<10^{-3}	0.78 (0.66-0.90)	71%	0.593	0.54 (0.39-0.69)	20%
5prc	*<10^{-3}	**0.80 (0.69-0.81)**	**68%**	*<10^{-3}	0.79 (0.67-0.91)	74%	*<10^{-3}	0.77 (0.65-0.89)	67%	*0.014	0.68 (0.55-0.82)	64%
95prc	*<10^{-3}	0.75 (0.63-0.87)	63%	*<10^{-3}	**0.84 (0.73-0.94)**	**77%**	*<10^{-3}	0.81 (0.70-0.93)	**75%**	*0.002	**0.73 (0.60-0.86)**	64%

As seen in Table III, the AUC of CPPS are larger for the excerpted vowels compared to sustained, for most descriptive statistics. The opposite trend is observed for the SampEn, with the AUC being larger in sustained compared to the excerpted vowels. For the sustained vowels, the highest AUC for CPPS occurs for the 5^{th} percentile (0.80), while for SampEn it occurs for the 95^{th} percentile (0.84). For the excerpted vowels, the highest AUC value for CPPS is seen for the mean (0.85), while SampEn exhibits the highest AUC for the median and the 95^{th} percentile (0.73). Figures 2 and 3 depict the ROC curves with the highest AUC for SampEn and CPPS in the case of sustained and excerpted vowels, respectively. The leave-one-out accuracy is also reported in Table III, and it can be seen that high AUC values

correspond to high accuracies, as expected. The highest accuracies are obtained in sustained vowels for the 95th percentile of SampEn, and in excerpted vowels for the mean, median and 95th percentiles of CPPS.

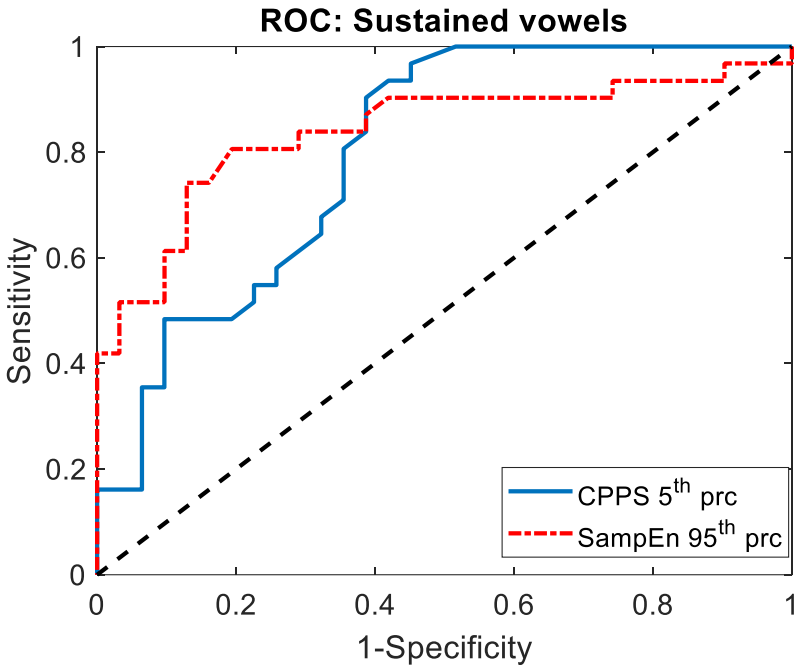


Figure 2: ROC curves for sustained vowels. The curves with the highest AUC are depicted, namely the ROC curve for the 5th percentile of CPPS and the 95th percentile of SampEn. The diagonal dashed line is a reference line.

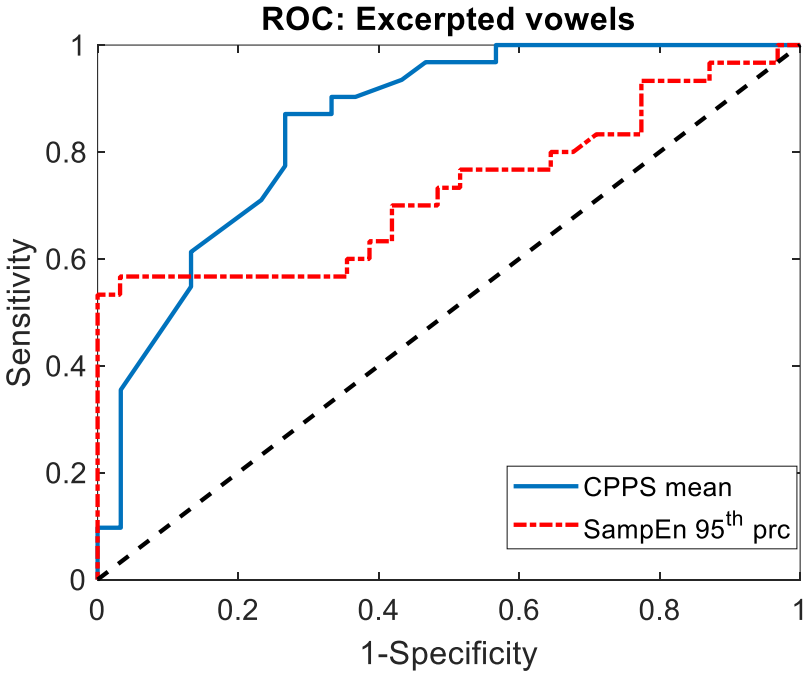


Figure 3: ROC curves for excerpted vowels. The curves with the highest AUC are depicted: the ROC curve for the mean CPPS and the 95th percentile of SampEn. The diagonal dashed line is a reference line.

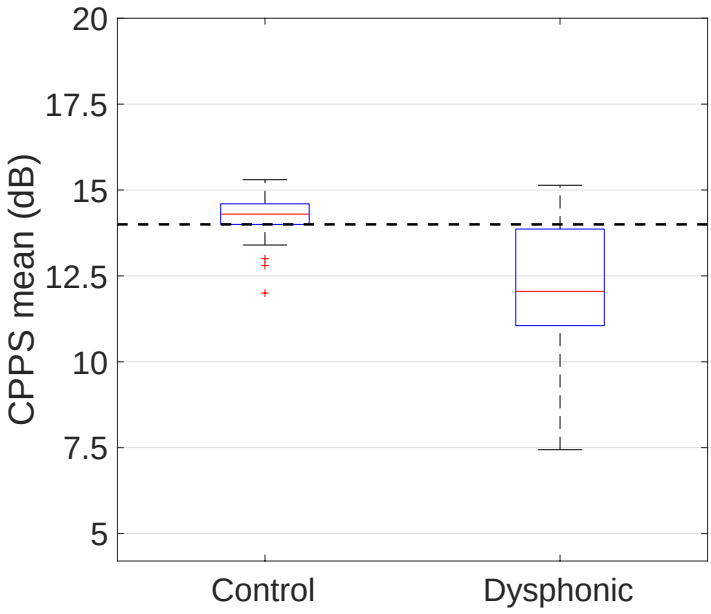
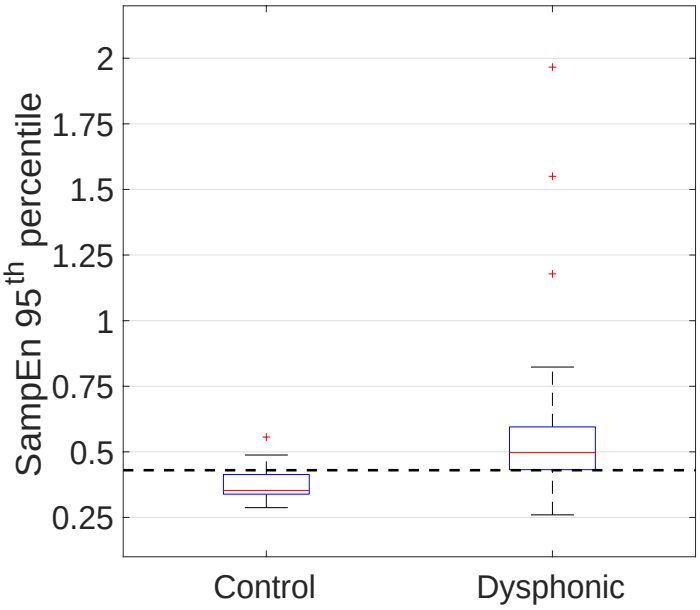


Figure 4: Boxplot of the mean CPPS (dB) from excerpted vowels for controls and dysphonic subjects. The thick horizontal dotted line marks the discrimination threshold (14 dB), for which sensitivity and specificity were equal to 78%.



Figures 4 and 5 show the boxplots of the distributions for the descriptive statistics that exhibit the highest AUC in CPPS and SampEn, respectively. Figure 4 depicts the

Figure 5: Boxplot of the 95th percentile of SampEn from sustained vowels for controls and dysphonic subjects. The thick horizontal dotted line marks the discrimination threshold (0.43), for which sensitivity and specificity were equal to 81%.

boxplot for mean CPPS from excerpted vowels in dysphonic subjects and controls. The thick horizontal dotted line denotes a discrimination threshold (14 dB) based on the ROC analysis with sensitivity and specificity equal to 78%. Figure 5 depicts the boxplot for the 95th percentile of SampEn in sustained vowels in dysphonic subjects and controls, and the thick horizontal dotted line marks the threshold from the ROC analysis (0.43), with sensitivity and specificity equal to 81%.

The set of recorded tokens of one subject corresponded to very breathy phonation with little voicing. These tokens exhibit the minimum mean CPPS (7.4 dB for excerpted, 4.90 dB for sustained) and the largest mean SampEn (1.50 for excerpted, 1.71 for sustained). The same tokens have the highest perceptual rating for aphonia (6.5/10) and breathiness (9.3/10). Despite the extremity of these tokens' values with respect to the rest of the data points, they were regarded as valid data points and were taken into account in the correlation analyses.

Table IV. Pearson's correlations between descriptive statistics of SampEn and CPPS distributions in excerpted vowels, with perceptual voice quality measures. Correlations with *p*-value smaller than 0.01 are denoted with a star. Boldface denotes the largest correlation among the descriptive statistics.

	Breathiness	Hyperfunction	High pitch roughness
CPPS			
Mean	-0.77 *	-0.13	-0.43
Median	-0.70 *	-0.14	-0.44
Std	-0.59 *	-0.02	-0.12
Range	-0.58 *	-0.15	-0.09
5prc	-0.72 *	-0.25	-0.38
95prc	**-0.83** *	-0.07	-0.32
SampEn			
Mean	0.79 *	0.11	0.15
Median	0.75 *	0.06	0.14
Std	0.72 *	0.19	0.00
Range	0.42	0.18	-0.12
5prc	0.78 *	0.10	0.25
95prc	**0.83 ***	0.18	0.14

Table IV shows Pearson's correlations between perceptual voice quality ratings and descriptive statistics of CPPS and SampEn from the excerpted vowels. Only those perceptual voice quality measures that had non-zero ratings for at least 15 subjects were considered; these were breathiness, hyperfunction, and high-pitch roughness. From those, only breathiness shows significant correlations with both CPPS and SampEn. The highest correlations with breathiness are observed for the 95th percentile

of CPPS (Pearson r = -0.83, $p < 10^{-3}$) and the 95th percentile of SampEn (Pearson r = 0.83, $p < 10^{-3}$).

Table V shows Pearson's correlation between CPPS and SampEn for different descriptive statistics, in sustained and excerpted vowels. For sustained vowels, the highest correlation is found for the mean of CPPS with the mean of SampEn (Pearson, r=-0.72, $p < 10^{-3}$). For excerpted vowels the highest correlation occurs for 95th percentile of CPPS and SampEn (r=-0.83 $p < 10^{-3}$).

Table V. Pearson's correlations between different descriptive statistics of SampEn and CPPS distributions in both sustained and excerpted vowels. Correlations with p-value smaller than 0.01 are denoted with a star. Boldface denotes the largest correlation among the descriptive statistics.

Pearson's r (CPPS, SampEn)	Sustained	Excerpted
Mean	**-0.72** *	-0.72 *
Median	-0.70 *	-0.67 *
Std	0.33 *	-0.31
Range	0.20	0.12
5prc	-0.67 *	-0.43 *
95prc	**-0.72 ***	**-0.83 ***

Discussion

The AUC from the ROC analysis for CPPS and SampEn indicates that both measures have good potential for discriminating dysphonic subjects from controls. As expected, degraded voice quality results in lower values for SampEn and higher values for CPPS, since the first quantifies the grade of disorder of the voice signal in the time domain, while the second quantifies the regularity of the harmonic components.

The highest AUCs were often obtained for descriptive statistics different than the mean, i.e., the 5th and 95th percentiles. In particular, the 5th percentile of individual distributions had the best discrimination power in the case of CPPS, since it corresponds to the lowest values of the distribution, which are associated with degraded voice quality. These extreme values are expected to be lower in the dysphonic subjects compared to controls. In the case of SampEn, it is the 95th percentile of individual distributions that had the best AUC, i.e., the highest values, which are associated with degraded voice quality. Therefore, these are expected to be higher in dysphonic subjects compared to controls. It follows from these results that the extremes of the distributions carry valuable information in discriminating between

pathological and normal voices. For that reason, it is useful to conduct distribution-based analysis, in order to capture a fuller picture of the analyzed voice signal.

These findings confirm and extend the results by Castellana et al. [27], who investigated descriptive statistics for CPPS distributions as possible indicators of vocal health status in 5-second long sustained vowels, which were acquired with the same experimental setup in the same clinic. The authors [27] found that the 5^{th} percentile was the best in discriminating between 41 patients and 35 controls, with an AUC of 0.95. Such outstanding discrimination power, which is higher than the one obtained in the present study for the 5^{th} percentile in sustained vowels may be linked to the use of longer vowels. To investigate whether longer vowels would show improved discriminatory power, we calculated CPPS and SampEn for four-second long vowels taken from the pathological and control groups. Three of the pathological subjects were not able to sustain a vowel for four seconds, and for that reason they were excluded from this analysis. The results in terms of accuracy showed an improvement of 2% for CPPS (5^{th} percentile) and a decline of 3% for SampEn (95^{th} percentile). The corresponding change in AUC for the four-second long compared to the two-seconds-long vowels was none for CPPS (5^{th} percentile) and 0.03 lower for SampEn (95^{th} percentile). Since this comparison between two-seconds and four-second long vowels showed relatively small differences, the improved performance in [27] is more likely to be due to the larger number of subjects and possibly higher grade of dysphonia of the pathological subjects.

The context of vowels was found to affect the discrimination power of both CPPS and SampEn, however in the opposite way. While the performance of CPPS was better using excerpted vowels, for SampEn sustained vowels worked best. It has been shown in earlier studies that perturbation and other voice quality measures such as jitter, shimmer, and harmonics-to-noise ratio (HNR) show improved values (smaller for jitter/shimmer, larger for HNR) with higher vocal intensity [38, 39]. For the analyzed dataset, the sustained vowels of both pathological subjects and controls had on average a higher intensity compared to the excerpted vowels by 2.9 dB (std 5 dB). That may explain why CPPS was not as successful in discriminating pathological subjects from controls for sustained vowels, since CPPS reflects to some extent amplitude and frequency perturbation as well as HNR [9]. On the other hand, SampEn is documented to give more reliable results with longer sequences [41], therefore sustained vowels may be more appropriate for discriminating regular from pathological voices.

As mentioned earlier in the Materials and Methods section, the algorithm used for calculating CPPS in sustained and excerpted vowels was not the same, since for the excerpted vowels it was not possible to apply smoothing in the time dimension. To ensure that the differences observed between excerpted and sustained vowels in CPPS were not due to the different algorithm used, we calculated CPPS for the sustained vowels using the exact same algorithm as for excerpted, i.e., omitting the smoothing-in-time step. The overall average difference between smoothed and non-smoothed (in time) CPPS was 0.5 dB, which indicates that the observed differences between excerpted and sustained vowels are not due to the variation of the used algorithm.

Descriptive statistics of CPPS and SampEn from excerpted vowels were both found to be correlated strongly with breathiness. This means that a large part of the irregularity that is reflected in higher SampEn values is due to presence of aerodynamic noise in the voice source. This result highlights the usability of SampEn as another strong correlate of breathiness. The descriptive statistic for both CPPS and SampEn with the highest correlation for breathiness was the 95^{th} percentile, which demonstrates again the importance of considering the extreme values of the distributions.

Concerning the relationship of CPPS to breathiness this work confirms and extends results from previous studies [10–12], which investigated the correlation between perceptual ratings and CPPS values obtained from the Hillenbrand software [7]. The correlation between CPPS and breathiness reported in these earlier studies ranged from -0.38 [11] to -0.71 [12], while in the present study for the mean CPPS the correlation was -0.77 and for the 95^{th} percentile of CPPS -0.83. Since both CPPS and SampEn seem to be affected by similar aspects of the voice signal, i.e., noise in the voice source and amplitude or frequency irregularity, it is expected that they will be strongly correlated with each other. In particular, all the distributional statistics of central tendency showed significant correlation, with the highest coefficient for 95^{th} percentile. The sign of the correlation is negative because as stated earlier, good voice quality corresponds to high CPPS but low SampEn. These results confirm and extend our preliminary study based on a limited excerpted vowels dataset [25].

In this study the perceptual evaluation was done based on the read text. Ideally, the evaluation should be done on the excerpted vowels, however due to the multitude of the stimuli this was judged impractical. In a future study there should also be perceptual evaluation of the sustained vowels, in order to examine if the perceptual measures correlate as well with CPPS and SampEn. Lastly, in order to investigate in greater detail how CPPS and SampEn depend on different perceptual aspects of

pathological voices, synthesized signals using a disordered voice synthesizer [42] could be analyzed.

Conclusion

The vowel context seems to affect the predictive performance for dysphonia of CPPS and SampEn in different ways; for CPPS, excerpted vowels showed better performance, while for SampEn sustained vowels worked best. CPPS and SampEn are strongly correlated with breathiness and among themselves.

Acknowledgements

The authors thank Massimo Spadola Bisetti and Jacopo Colombini for conducting the clinical examination of the subjects. Maria Södersten and Anna Lundblad contributed the perceptual evaluation of the readings. Sten Ternström provided editorial and scientific feedback. We thank Jean Schoentgen for the valuable comments on an earlier version of the manuscript. Andreas Selamtzis was supported by the Swedish Research Council (Vetenskapsrådet), Contract Nos. 2010-4565 and 2013-0632.

References

[1] S. Bielamowicz, J. Kreiman, B.R. Gerratt, M.S. Dauer, G.S. Berke, Comparison of Voice Analysis Systems for Perturbation Measurement, J. Speech Lang. Hear. Res. 39 (1996) 126. doi:10.1044/jshr.3901.126.

[2] S.Y. Lowell, R.H. Colton, R.T. Kelley, Y.C. Hahn, Spectral- and Cepstral-Based Measures During Continuous Speech: Capacity to Distinguish Dysphonia and Consistency Within a Speaker, J. Voice. (2010). doi:10.1016/j.jvoice.2010.06.007.

[3] B.R. Gerratt, J. Kreiman, M. Garellek, Comparing Measures of Voice Quality From Sustained Phonation and Continuous Speech, J. Speech Lang. Hear. Res. 59 (2016) 994. doi:10.1044/2016_JSLHR-S-15-0307.

[4] V. Parsa, D.G. Jamieson, Acoustic Discrimination of Pathological Voice, J. Speech Lang. Hear. Res. 44 (2001) 327. doi:10.1044/1092-4388(2001/027).

[5] A. Lederle, J. Barkmeier-Kraemer, E. Finnegan, Perception of Vocal Tremor During Sustained Phonation Compared With Sentence Context, J. Voice. 26 (2012) 668.e1-668.e9. doi:10.1016/j.jvoice.2011.11.001.

[6] K.R. Moon, S.M. Chung, H.S. Park, H.S. Kim, Materials of Acoustic Analysis: Sustained Vowel Versus Sentence, J. Voice. 26 (2012) 563–565. doi:10.1016/j.jvoice.2011.09.007.

[7] J. Hillenbrand, R.A. Houde, Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech, J. Speech Lang. Hear. Res. 39 (1996) 311. doi:10.1044/jshr.3902.311.

[8] B. P. Bogert, The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proc. Symposium on Time Series Analysis*, (1963). John Wiley & Sons.

[9] R. Fraile, J.I. Godino-Llorente, Cepstral peak prominence: A comprehensive analysis, Biomed. Signal Process. Control. 14 (2014) 42–54. doi:10.1016/j.bspc.2014.07.001.

[10] L.F. Brinca, A.P.F. Batista, A.I. Tavares, I.C. Gonçalves, M.L. Moreno, Use of Cepstral Analyses for Differentiating Normal From Dysphonic Voices: A Comparative Study of Connected Speech Versus Sustained Vowel in European Portuguese Female Speakers, J. Voice. 28 (2014) 282–286. doi:10.1016/j.jvoice.2013.10.001.

[11] S. Jannetts, A. Lowit, Cepstral Analysis of Hypokinetic and Ataxic Voices: Correlations With Perceptual and Other Acoustic Measures, J. Voice. 28

(2014) 673–680. doi:10.1016/j.jvoice.2014.01.013.

[12] Y.D. Heman-Ackah, D.D. Michael, G.S. Goding, The Relationship Between Cepstral Peak Prominence and Selected Parameters of Dysphonia, J. Voice. 16 (2002) 20–27. doi:10.1016/S0892-1997(02)00067-X.

[13] J.S. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy., Am. J. Physiol. Heart Circ. Physiol. 278 (2000) H2039-49. http://www.ncbi.nlm.nih.gov/pubmed/10843903.

[14] S. Pincus, Approximate entropy (ApEn) as a complexity measure, Chaos An Interdiscip. J. Nonlinear Sci. 5 (1995) 110–117. doi:10.1063/1.166092.

[15] S.M. Pincus, Approximate entropy as a measure of system complexity, Mathematics. 88 (1991) 2297–2301. doi:10.1073/pnas.88.6.2297.

[16] D.E. Lake, J.S. Richman, M.P. Griffin, J.R. Moorman, Sample entropy analysis of neonatal heart rate variability, Am. J. Physiol. Integr. Comp. Physiol. 283 (2002) R789-97. doi:10.1152/ajpregu.00069.2002.

[17] D. Zhang, H. Ding, Y. Liu, C. Zhou, H. Ding, D. Ye, Neurodevelopment in newborns: A sample entropy analysis of electroencephalogram, Physiol. Meas. 30 (2009) 491–504. doi:10.1088/0967-3334/30/5/006.

[18] S. Ramdani, B. Seigle, J. Lagarde, F. Bouchara, P.L. Bernard, On the use of sample entropy to analyze human postural sway data, Med. Eng. Phys. 31 (2009) 1023–1031. doi:10.1016/j.medengphy.2009.06.004.

[19] M. Aboy, D. Cuesta-Frau, D. Austin, P. Mico-Tormos, Characterization of sample entropy in the context of biomedical signal analysis., Conf. Proc. IEEE Eng. Med. Biol. Soc. 2007 (2007) 5943–6. doi:10.1109/IEMBS.2007.4353701.

[20] K. Manickam, C. Moore, T. Willard, N. Slevin, Quantifying aberrant phonation using approximate entropy in electrolaryngography, Speech Commun. 47 (2005) 312–321. doi:10.1016/j.specom.2005.02.008.

[21] C.M. Douglas, C. Moore, K. Manickam, L. Lee, A. Sykes, A. Carr, S. Jones, J. Jones, R. Swindell, J.J. Homer, N. Slevin, Electroglottogram approximate entropy: a novel single parameter for objective voice assessment, J. Laryngol. Otol. 124 (2010) 520. doi:10.1017/S0022215109992787.

[22] C. Moore, S. Shalet, K. Manickam, T. Willard, H. Maheshwari, G. Baumann, Voice abnormality in adults with congenital and adult-acquired growth hormone deficiency, J. Clin. Endocrinol. Metab. 90 (2005) 4128–4132. doi:10.1210/jc.2004-2558.

[23] C. Fabris, W. De Colle, G. Sparacino, Voice disorders assessed by (cross-) Sample Entropy of electroglottogram and microphone signals, Biomed. Signal

Process. Control. 8 (2013) 920–926. doi:10.1016/j.bspc.2013.08.010.

[24] J.D. Arias-Londoño, J.I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, G. Castellanos-Domínguez, Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients, IEEE Trans. Biomed. Eng. 58 (2011) 370–379. doi:10.1109/TBME.2010.2089052.

[25] A. Castellana, A. Selamtzis, G. Salvi, A. Carullo, A. Astolfi, Cepstral and Entropy Analyses in Vowels Excerpted from Continuous Speech of Dysphonic and Control Speakers, in: Interspeech 2017, ISCA, 2017: pp. 1814–1818. doi:10.21437/Interspeech.2017-335.

[26] A. Castellana, A. Carullo, S. Corbellini, A. Astolfi, M. Spadola Bisetti and J. Colombini, Cepstral Peak Prominence Smoothed distribution as discriminator of vocal health in sustained vowels, in Proceedings of 2017 I2MTC Conference, May 22–25, Torino, Italy.

[27] A. Castellana, A. Carullo, S. Corbellini, A. Astolfi, Discriminating pathological voice from healthy voice using Cepstral Peak Prominence Smoothed distribution in sustained vowels, IEEE Trans. Instrum. Meas. (accepted) (2018).

[28] I. Vernero, M. Gambino, A. Schindler, O. Schindler, Cartella logopedica: età evolutiva, Edizioni Omega, Torino, 2002.

[29] H. Šrámková, S. Granqvist, C.T. Herbst, J.G. Švec, The softest sound levels of the human voice in normal subjects, J. Acoust. Soc. Am. 137 (2015) 407–418. doi:10.1121/1.4904538.

[30] B. Hammarberg, J. Gauffin, Perceptual and Acoustic Characteristics of Quality Differences in Pathological Voices as Related to Physiological Aspects, in: O. Fujimura, M. Hirano (Eds.), Vocal Fold Physiol. Voice Qual. Control, Singular Publishing Group, San Diego, 1995: pp. 283–303.

[31] B. Hammarberg, Voice research and clinical needs, Folia Phoniatr. Logop. 52 (2000) 93–102. doi:10.1159/000021517.

[32] Italian SpeechDat (II) FDB-3000L, (n.d.). http://catalog.elra.info/product_info.php?products_id=631 (accessed December 1, 2017).

[33] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. (Andrew) Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book, Cambridge University Engineering Department, Cambridge, 2009.

[34] K. Elenius, G. Salvi, B. Lindberg, N. Warakagoda, F.T. Johansen, G. Lehtinen, Z.K. C, A. Zgank, A noise robust multilingual reference recogniser based on

SpeechDat(II), in: 6th Internetional Conf. Spok. Lang. Process. Vol. III, 2000: pp. 370–373.

[35] G. de Krom, A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals, J. Speech Lang. Hear. Res. 36 (1993) 254. doi:10.1044/jshr.3602.254.

[36] D.K. Lake, J.R. Moorman, C. Hanqing, SampEn for MATLAB 1.1-1, (2012). https://www.physionet.org/physiotools/sampen/matlab/1.1-1/.

[37] J.D. Gibbons, S. Chakraborti, Nonparametric Statistical Inference, Taylor & Francis, London, 2003.

[38] V. Bewick, L. Cheek, J. Ball, Statistics review 13: Receiver operating characteristic curves, Crit. Care. 8 (2004) 508. doi:10.1186/cc3000.

[39] M. Brockmann-Bauser, J.E. Bohlender, D.D. Mehta, Acoustic Perturbation Measures Improve with Increasing Vocal Intensity in Individuals With and Without Voice Disorders, J. Voice. (2017). doi:10.1016/j.jvoice.2017.04.008.

[40] P.H. Dejonckere, Effect of louder voicing on acoustical measurements in dysphonic patients, Logop. Phoniatr. Vocology. 23 (1998) 79–84. doi:10.1080/140154398434239.

[41] M. Costa, A.L. Goldberger, C.K. Peng, Multiscale entropy analysis of biological signals, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. 71 (2005) 1–18. doi:10.1103/PhysRevE.71.021906.

[42] S. Fraj, J. Schoentgen, F. Grenez, Development and perceptual assessment of a synthesizer of disordered voices, J. Acoust. Soc. Am. 132 (2012) 2603–2615. doi:10.1121/1.4751536.