



# Robust and complex approach of pathological speech signal analysis



Jiri Mekyska<sup>a,b</sup>, Eva Janousova<sup>c</sup>, Pedro Gomez-Vilda<sup>d</sup>, Zdenek Smekal<sup>a,b</sup>,  
Irena Rektorova<sup>e,f,\*</sup>, Ilona Eliasova<sup>e,f</sup>, Milena Kostalova<sup>g,f</sup>, Martina Mrackova<sup>e,f</sup>,  
Jesus B. Alonso-Hernandez<sup>h</sup>, Marcos Faundez-Zanuy<sup>i</sup>, Karmele López-de-Ipiña<sup>j</sup>

<sup>a</sup> Department of Telecommunications, Brno University of Technology, Technická 10, 61600 Brno, Czech Republic

<sup>b</sup> SIX Research Centre, Technická 10, 61600 Brno, Czech Republic

<sup>c</sup> Institute of Biostatistics and Analyses, Masaryk University, Kamenice 3, 62500 Brno, Czech Republic

<sup>d</sup> Facultad de Informatica, Universidad Politecnica de Madrid, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain

<sup>e</sup> First Department of Neurology, St. Anne's University Hospital, Pekarska 53, 65691 Brno, Czech Republic

<sup>f</sup> Applied Neuroscience Research Group, Central European Institute of Technology, Masaryk University, Komenského nám. 2, 60200 Brno, Czech Republic

<sup>g</sup> Department of Neurology, Faculty Hospital and Masaryk University, Jihlavská 20, 63900 Brno, Czech Republic

<sup>h</sup> Institute for Technological Development and Innovation in Communications (IDeTIC), University of Las Palmas de Gran Canaria, 35001 Las Palmas de Gran Canaria, Spain

<sup>i</sup> Escola Universitaria Politecnica de Mataro, Tecnocampus, Avda. Ernest Lluch 32, 08302 Mataro, Barcelona, Spain

<sup>j</sup> Department of Systems Engineering and Automation, University of the Basque Country UPV/EHU, Av de Tolosa 54, 20018 Donostia, Spain

## ARTICLE INFO

### Article history:

Received 1 October 2014

Received in revised form

12 January 2015

Accepted 9 February 2015

Communicated by Carlos Manuel Travieso

González

Available online 21 May 2015

### Keywords:

Pathological speech

Disordered voice

Dysarthria

Speech processing

Bicepstrum

Non-linear dynamic features

## ABSTRACT

This paper presents a study of the approaches in the state-of-the-art in the field of pathological speech signal analysis with a special focus on parametrization techniques. It provides a description of 92 speech features where some of them are already widely used in this field of science and some of them have not been tried yet (they come from different areas of speech signal processing like speech recognition or coding). As an original contribution, this work introduces 36 completely new pathological voice measures based on modulation spectra, inferior colliculus coefficients, bicepstrum, sample and approximate entropy and empirical mode decomposition. The significance of these features was tested on 3 (English, Spanish and Czech) pathological voice databases with respect to classification accuracy, sensitivity and specificity. To our best knowledge the introduced approach based on complex feature extraction and robust testing outperformed all works that have been published already in this field. The results (accuracy, sensitivity and specificity equal to  $100.0 \pm 0.0\%$ ) are discussable in the case of Massachusetts Eye and Ear Infirmary (MEEI) database because of its limitation related to a length of sustained vowels, however in the case of Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid database we made improvements in classification accuracy ( $82.1 \pm 3.3\%$ ) and specificity ( $83.8 \pm 5.1\%$ ) when considering a single-classifier approach. Hopefully, large improvements may be achieved in the case of Czech Parkinsonian Speech Database (PARCZ), which are discussed in this work as well. All the features introduced in this work were identified by Mann–Whitney  $U$  test as significant ( $p < 0.05$ ) when processing at least one of the mentioned databases. The largest discriminative power from these proposed features has a cepstral peak prominence extracted from the first intrinsic mode function ( $p = 6.9443 \times 10^{-32}$ ) which means, that among all newly designed features those that quantify especially hoarseness or breathiness are good candidates for pathological speech identification. The paper also mentions some ideas for the future work in the field of pathological speech signal analysis that can be valuable especially under the clinical point of view.

© 2015 Elsevier B.V. All rights reserved.

## 1. Introduction

Voice Pathology description and characterization has always demanded attention from the physiological and medical fields [1],

as well as from the voice function point of view [2]. The term voice is described by [3] both in a broad and a narrow sense. In the broad sense voice may be taken as synonymous of speech, therefore terms as Voice over IP (VoIP) can be found in the literature and media with the meaning of speech data on internet. In the narrow sense voice refers to the vibration of the vocal folds. Speech sounds resulting from the interaction of this vibration with the Oro-Naso-Pharyngeal Tract (ONFT) are referred to, as voiced. Speech sounds produced by turbulent flow within the ONFT are

\* Corresponding author at: First Department of Neurology, St. Anne's University Hospital, Pekarska 53, 65691 Brno, Czech Republic. Tel.: +420 549 497 825.

E-mail address: [irena.rektorova@fnusa.cz](mailto:irena.rektorova@fnusa.cz) (I. Rektorova).

termed voiceless. Phonation is the recommended term to refer to vocal fold vibration. A speaker showing an anomalous vocal fold vibration pattern is referred as dysphonic, and aphonic if there is no vocal fold vibration at all. If no anomalies are present the speaker is referred as normophonic.

Dysphonic voice is a perceptual and subjective term associated to Voice Pathology. Dysphonia is the perceptual quality of voice signaling that something wrong is happening in the phonation organs (mainly the larynx and its associated structures). Voice Pathologies and Dysphonic Voice are thus intrinsically related. The classification of Voice Pathologies or Disorders is described in [3] as tissue infection (e.g. laryngitis, bronchitis, croup, ...), systemic changes (e.g. dehydration, pharmacological and drug effects, hormonal changes, ...), mechanical stress (e.g. vocal nodules, polyps, ulcers, granulomae, laryngocele, hemorrhage, ...), surface irritation (e.g. laryngitis, leukoplakia, gastroesophageal reflux, ...), tissue changes (e.g. laryngeal carcinoma, keratosis, papillomas, cysts, ...), neurological and muscular changes (e.g. bilateral and unilateral vocal fold paralysis, Parkinson's Disease (PD), Amyotrophic Lateral Sclerosis (ALS), myotonic dystrophy, Huntington's Chorea, myasthenia gravis, ...), and abnormal muscle patterns (e.g. conversion aphonia or dysphonia, spasmodic dysphonia, mutational dysphonia, ventricular phonation, ...). In [4] a description of vocal pathologies can be found. A last important group of neurological diseases which leave a correlate in voice and speech is that of cognitive origin, Alzheimer's Disease (AD) being the most relevant one for their impact in well-being and in aging specialized-attention demand. Some references on the influence of AD in speech and voice can be found in [5–8]. Going a step further, emotional alterations (either temporary or persistent) leave also correlates in the speech and phonation signature, and may be subjects of further study by acoustic analysis [9,10].

The relationship between acoustic correlates and voice pathology has been clinically established in the last decades, subjectively and quantitatively [11,12]. Acoustic Voice Quality Analysis (AVQA) is a wide term for a set of different methodologies designed to quantify acoustic correlates giving a definition of the quality of phonation or speech production. Therefore AVQA would be the procedural way to objectively quantify anomalies manifested in Dysphonic Voice. Modern signal processing technologies provide estimates of voice and speech correlates in time and frequency [13], allowing to better visualize and quantify phonation patterns. Spectral techniques facilitate the study of pathologic phonation, establishing relations between harmonic–harmonic and harmonic–formant ratios, which were found as important correlates to organic voice pathology [14]. Similarly, harmonic–noise ratios were found significant in characterizing certain types of dysphonic pathologies [15,16]. Time domain estimates, as jitter, shimmer and open-closed phase quotients are also used in describing dysphonic voice [17,18]. These descriptions gave rise to AVQA as a specific field [19].

Under the point of view of AVQA the following objectives can be established in order of difficulty: dysphonic voice detection, dysphonic voice grading, and dysphonic voice classification according to etiology. Dysphonic voice detection would be the task of assigning normophonic (normative) or dysphonic (non-normative) labels to a given phonation produced by a specific speaker. The determination of the dysphonic grade is traditionally carried out by independent referees according to a subjective criterion on a given scale. One of the most popular is GRBAS (grade, roughness, breathiness, asthenia and strain) [11]. The relative dependence of the assigned grade to the referee's subjective opinion results in wide grading differences among referees. To overcome this problem requires the design of clinical assessment methodologies [20]. The task of dysphonic voice classification according to etiology is far more difficult, as a given

acoustic correlate may be attributed to different pathologies. If this problem is stated in terms of associating acoustic correlates to specific pathologies, the potential risk is that it will remain unsolved for long, because it is an ill-posed problem (a many-to-one subjective mapping). A preliminary step to be covered first is to define the implications of different pathologies in the vocal function, especially at the level of the larynx. Under the functional point of view, the following main behaviors may be observed in the abnormal operation of the vocal folds: asymmetric vibration, contact defects and dystonia (hypo-, hyper-tension and tremor). Most organic larynx pathologies reproduce either one or another behavior, or all of them. Asymmetric vocal fold vibration is to be expected in pathologies as polyps, cysts, carcinomae, vocal fold paralysis, ulcers, cysts, and papillomae, and produces acoustic correlates as jitter, shimmer, poor harmonic–noise ratios, unbalance, and sub- and inter-harmonics. On its turn, contact defects are to be expected in pathologies as polyps, nodules, edemae, and cysts, where full closure of the glottal gap is not granted by vocal fold adduction and produce acoustic correlates as open and close phase perturbations, recovery phase attenuation, and harmonic display reduction. Other pathologies, especially those of neurological origin produce deviations in biomechanical parameters, as vocal fold tension (hypo- and hyper-tonic) and tremor, which can also be manifested as modulations in amplitude or frequency, and as changes in the vocal fold tension. As many organic pathologies induce a hyper-tonic behavior, the main problem when dealing with correlates produced by neurological pathology is to differentiate their origin from that of organic origin. Contact defect pathologies can also show asymmetric vocal fold vibration, and this may also be the case in aging voice (presbyphonia). To establish differentiation criterion, for using acoustic correlates when dealing with organic, neurologic or aging-induced perturbations, is a major issue in AVQA [21].

Another problem regarding AVQA is the lack of good reference baselines to establish the methodologies for voice pathology classification from acoustic analysis. Rigorous databases, acquired to represent each of the different pathologies under well-defined sample population size and recording conditions are scant. Many times the problem comes from the simultaneous presence of different larynx pathologies in the same patient, either related or unrelated (e.g. it is common to find a counter-lateral lesion as a consequence of a unilateral polyp). The problem then is how to decide if a specific acoustic correlate is produced by one cause or another. Most of the times acoustic databases are produced by laryngological services as part of examination protocols [22,23], but this is not always the case, as many laryngologists prefer to depend on visual exploration, neglecting the possibilities offered by AVQA for different reasons [24]. Speech therapists rely more on acoustic exploration, but many times it is mainly restricted to measurements of vocal effort, long term frequency analysis, respiratory efficiency, or distortion measures. Therefore voice records from speakers ranked by etiology and severity index, using compatible standards (digitalization, channel, microphones) are scant. Most of the available databases are either incomplete, inconsistent (made up of recordings taken under different conditions) or deficient (many non-frequent pathologies are not well represented in sample size). Besides, there is a lack of good normative databases, as most of the records produced by medical services contain information from dysphonic voice, but normative speakers are absent or poorly represented (this is the case of the most widely used database [25]). This is a severe limitation for systematic AVQA. Another problem is cross-lingual representation. As far as sustained vowels are concerned, this would not be a problem, but it becomes a major obstacle when segmental parameters are involved, as in the use of passages (either read or spontaneous).

After all these considerations, it may be said that the aim of AVQA is to design the best methodology to use Voice Correlates in Voice Pathology detection, grading and classification. Several steps are to be assumed and methodologically formulated for such. Voice Pathologies and Correlates have to be well connected by adequate physiological and biomechanical modeling of larynx dynamics to understand how acoustic correlates and organic dysfunctions could be associated [26]. This requirement is of vital importance in meeting the challenge of pathology classification from acoustic correlates. On its turn, inverse methods to obtain robust and significant estimates of voice correlates are of great relevance [27]. These are to be combined with other parameterization methods based in direct time and frequency domain estimates to produce rich feature sets [28]. Finally, good Statistical Pattern Matching and Machine Learning methodologies have to be used to reduce redundancy and to make a selection by relevance within the feature set, to determine the best combinations for each specific task (detect, grade, classify), and procure specific results under quality criteria (sensitivity, specificity and accuracy) [29]. These methods have to be contrasted on generally-accepted benchmark databases [30].

Having in mind all these conditions the present study is aimed to describe a wide set of tests on different databases, using an exhaustive set of acoustic features, by means of well-known classifiers to estimate the performance of the methods and features regarding specificity, sensitivity and accuracy in dysphonic voice detection tasks. There are already many publications describing the methods of pathological speech detection. Usually the authors use just a limited set of speech features and most of the publications present the detection accuracies tested only on one monolingual database [31–37]. Probably the widest range of features describing different aspects of speech was tested in a work of Tsanas et al. [34]. Regarding the databases there are just a few publications considering 2 different data sets [38–41].

To sum up the state-of-the-art approaches in the field of pathological voice analysis there is still a lack of publications providing a complex overview of features quantifying pathological speech and providing strong conclusions supported by a robust testing. Therefore this work has 4 main goals:

1. According to complex parameterization and consequent robust testing identify features that have the largest discriminative power in the field of pathological speech analysis.
2. Design new features that can quantify hoarseness, breathiness and non-linearities in pathological speech signals.
3. Prove that the proposed large set parameterization approach can provide better classification results (with respect to classification accuracy, sensitivity and specificity) than those published in the field of pathological speech analysis by the other researchers.
4. Select a database that has high potential for the future, especially in terms of speech features design, tuning and testing.

The paper is organized as follows: the classically used features as well as the newly designed ones are discussed in Section 2, Section 3 describes the 3 databases, Section 4 describes the testing procedure, Section 5 is devoted to the experimental results and discussion. The conclusions are given in Section 6.

## 2. Features

The aim of this work was to explore significance of the mostly used speech features when focusing on the ability of differentiation between healthy and pathological speech. The features that

are usually known in different fields of speech signal processing (speech recognition, enhancement, denoising, and identification) and newly designed features originally introduced in this paper will be investigated as well.

Due to the limited size of this paper the features that were not originally introduced in this work will be mentioned without their deeper description by an algorithm. However each parameter will be accompanied by a reference where the reader can find further information.

### 2.1. Features describing phonation

Probably the most popular features describing pathological voice are fundamental frequency  $F_0$  and parameters describing its variability in time (jitter): PPQ5 (five-point Pitch Perturbation Quotient), RAP (Relative Average Perturbation),  $jitt_{loc}$  (average absolute difference between consecutive periods, divided by the average period),  $jitt_{abs}$  (average absolute difference between consecutive periods),  $jitt_{ddp}$  (average absolute difference between consecutive differences on neighbor glottal periods, divided by the average period) [39,41–43]. These features are good candidates especially for quantification of voice tremor [44].

A disadvantage of previously mentioned measurements is that the values of the features are highly dependent on gender and variable acoustic environment. To overcome this disadvantage Little et al. proposed the PPE (Pitch Period Entropy) [45]. During the calculation of PPE, logarithmic semitone scale, inverse filtering and entropy estimation are incorporated. Another method which is close to jitter is a measure of standard deviation (std) of the time that vocal folds are apart ( $GQ_{open}$ ) and in collisions respectively ( $GQ_{closed}$ ) [34].

To effectively describe hypophonia or intensity perturbations, short-time energy  $E$  or pitch-level variations (shimmer) can be used [34,42,46]. In this work 6 kinds of shimmer will be considered (they are calculated similarly to jitter but the intensity is used): APQ3 (three-point Amplitude Perturbation Quotient), APQ5, APQ11,  $shimm_{loc}$ ,  $shimm_{ddp}$ ,  $shimm_{dB}$  (average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods) [43,47].

Another feature describing speech intensity is TKEO (Teager-Kaiser Energy Operator) [48]. The advantage over simple  $E$  is that it takes into account also signal frequency. It has been shown that speech contains dominant modulation frequencies in a range 2–20 Hz with maximum at approximately 4 Hz [49]. The 4 Hz modulation energy (ME) was selected as a feature which is related to a measure describing energy distribution in power spectra. Similarly MPSD (Median of Power Spectral Density), usually called median frequency, was selected [50]. The last feature in this category is LSTER (Low Short-Time Energy Ratio) [51]. This feature is usually used for differentiation between speech and music signals because speech exhibits higher variations in energy per 10–30 ms frames. However this feature was selected for analysis of pathological speech as well. It is considered that patients will reach higher values of this feature in the case of maintained vowels due to the inability to sustain the same amount of airflow during the whole phonation.

### 2.2. Features describing tongue movement

Frequencies of first three formants  $F_1$ ,  $F_2$ ,  $F_3$  and their bandwidths  $B_1$ ,  $B_2$ ,  $B_3$  are related to volumes of vocal tract cavities. Especially the volume of throat and oral cavity is modified by the tongue position. According to this fact it is possible to consider formants as a measure of tongue movement [52,53].

### 2.3. Features describing speech quality

Signs of vocal fold dysfunctions are usually associated with breathiness or hoarseness [32]. From signal theory point of view these dysfunctions are characteristically decreasing voice quality, under a simplified consideration, by additive noise. This implies that methods based on speech quality measurements can suitably quantify vocal folds impairment and describe its progress. Besides breathiness and hoarseness these methods can be also used for analysis of hypernasality caused by an improper work of soft palate [54].

Probably the simplest quality measure feature is ZCR (Zero-Crossing Rate) and its modification HZCRR (High Zero-Crossing Rate Ratio) which takes into account a variation of ZCR in time [51]. Another simple measure is FLUF (Fraction of Locally Unvoiced Frames) which can describe an impossibility of carrying out periodical glottal closure [55].

The next three measures are based on the variation of spectrum values between adjacent frames. Specifically these are SF (Spectral Flux) [56], SDBM (Spectral Distance Based on Module) and SDBP (Spectral Distance Based on Phase) [55].

We have also used several features based on cepstral analysis. It was shown that cepstrum and its harmonics correlate well with the perception of breathiness [57]. Moreover from the nature of cepstrum it can be said that it is a kind of periodicity measure and therefore it should also predict roughness. The most famous feature based on the real cepstrum is CPP (Cepstral Peak Prominence) originally introduced by Hillenbrand et al. [57]. Besides this feature, PECM (Pitch Energy Cepstral Measure) [55] and VR (Variation in Ratio between the second/first harmonic within the derived cepstral domain) were also used [55].

The other quality measure features are based on an estimation of the level of noise present. We will test the significance of HNR (Harmonic-to-Noise Ratio) [37,38,54,58], NHR (Noise-to-Harmonic Ratio) [37,59], NNE (Normalized Noise Energy) [60], GNE (Glottal-to-Noise Excitation ratio) [58], SPI (Soft Phonation Index) [59] and VTI (Voice Turbulence Index) [59]. The methods differ especially in the estimation of noise.

The last feature in this category is SSD (Segmental Signal-to-Dysperiodicity ratio). It was shown that this feature correlates strongly with the perceived degree of the speaker's hoarseness [38].

### 2.4. Segmental features

Although some of the features mentioned in the other categories can be denoted as segmental as well (they are calculated from 10–30 ms segments), here the segmental features are considered as matrices (not only vectors) calculated from the whole signal. Probably the most popular segmental features in the field of speech signal analysis are MFCC (Mel Frequency Cepstral Coefficients) [33,36,42,54]. The advantage of these coefficients is that they can indirectly detect slight misplacements of articulators [34]. In this work 20 MFCC coefficients are extracted. The coefficient number zero is replaced by an estimate of log-energy.

Although MFCC are generally used, there is a lack of publications that compare these features to other segmental ones for the purpose of pathological speech analysis. Therefore we decided to include other segmental features. We extracted 20 mel frequency cepstral coefficients but in this case the bank of triangular filters was adjusted to the equal loudness curve [61]. We call these features MFCCE. The other two sets of features derived from MFCC are LFCC (Linear Frequency Cepstral Coefficients) and CMS (Cepstral Mean Subtraction coefficients). In the case of LFCC the bank of triangular filters is equidistantly spread in the frequency scale [62]. CMS is a kind of standardized z-score of MFCC (subtraction of mean and division by std over the time).

MSC (Modulation Spectra Coefficients) can provide information complementary to MFCC [36,63]. These features can capture a class of source mechanism characteristics related to voice quality [33].

In the next step features based on linear prediction were extracted. LPC (Linear Predictive Coefficients) [42], PLP (Perceptual Linear Predictive coefficients) [64], LPCC (Linear Predictive Cepstral Coefficients) [65], LPCT (Linear Predictive Cosine Transform coefficients) [42] and ACW (Adaptive Component Weighted coefficients) [65] were tested. The advantage of PLP over simple LPC or MFCC is that it also takes into account an adjustment to the equal loudness curve and intensity-loudness power law [64]. The advantage of LPCC and LPCT over “classic” LPC is that a transformation is used into the cepstral domain and thus the values do not correlate much. The advantage of ACW is that these coefficients are less sensitive to channel distortion [65].

The last segmental features used in this work are parameters that analyze amplitude modulations in voice using a biologically-inspired model of the inferior colliculus [66]. These features are called ICC (Inferior Colliculus Coefficients).

Segmental features are sometimes extended by their 1st and 2nd order regression coefficients ( $\Delta$  and  $\Delta\Delta$  respectively). In this work we used the  $\Delta$  coefficients.

### 2.5. Features based on bispectrum

Alonso et al. proved that a greater presence of quadratic coupling is observed in healthy voice when comparing it to the pathological one [55]. It is probably due to a fact that healthy voice is characterized by a vocal tract which is more non-linear than in the case of pathological voices. This quadratic coupling can be appropriately described by bispectrum and features derived from this 2D signal.

There were proposed measures such as BII (Bicoherence Index Interference), HFEB (High Frequency Energy of one-dimensional Bicoherence), LFEB (low Frequency Energy of one-dimensional Bicoherence), BMII (Bispectrum Module Interference Index) and BPII (Bispectrum Phase Interference Index) [55].

### 2.6. Features based on wavelet decomposition

The wavelet transform is widely used especially in the field of coding and speech denoising. However its application can be found in the field of pathological speech analysis too [67]. Detail coefficients after the decomposition can be used to estimate the present noise and consequently it is possible to calculate SNR (Signal-to-Noise ratio). In fact this is just another method of voice quality measurement.

According to some measurements we have empirically selected 7 wavelets: 10th, 15th, 20th-order Daubechies wavelets; 10th, 15th, 20th-order symlet wavelets and 5th-order coiflet wavelet. We will mark the features derived from the detail coefficients of the wavelet transform as SNRW(*wvl*), where *wvl* corresponds to the specific wavelet (e.g. daub15).

### 2.7. Features based on empirical mode decomposition

Recently, in speech processing new methods based on EMD (Empirical Mode Decomposition) have been used. Using EMD it is possible to decompose the arbitrary non-linear and time-varying signal into countable and usually a small number of IMF (Intrinsic Mode Functions). These functions are modulated in amplitude and frequency and their sum gives the original signal.

Tsanas et al. proposed several measures of SNR and NSR based on the IMFs [34]. The time-varying high frequency components are present in the first few IMFs. Therefore these first few IMFs can be used to represent the noise in the signal and the rest of IMFs can be



used for a representation of the useful information. According to this consideration features like IMF–SNR<sub>TKEO</sub> (based on Teager-Kaiser Energy Operator), IMF–SNR<sub>SEO</sub> (based on Squared Energy Operator), IMF–SNR<sub>SE</sub> (based on Shannon Entropy), IMF–NSR<sub>TKEO</sub>, IMF–NSR<sub>SEO</sub> and IMF–NSR<sub>SE</sub> have been introduced.

## 2.8. Non-linear dynamic features

Tension on the vocal folds can significantly differ in the case of pathological speech. Voice becomes aperiodic, noisy-like, and it is very difficult to find any regularities in the signal. There is a frequent presence of sub-harmonics and chaos which can lead to a failure of conventional techniques of speech signal analysis. However it was shown that these kinds of signals can be sufficiently described by non-linear dynamical analysis [32,37,40,54].

The first representative in this category is CD (Correlation Dimension) which statistically measures attractor geometry in the phase space. This measure is related to a number of independent variables necessary for generating the attractor [32,45,47]. Another dimension measure FD (Fractal Dimension) is based on a number of basic building blocks that form a pattern [32,68]. We will also use complexity measures like ZL (Ziv-Lempel complexity) which quantify the regularity embedded in a time series [69]. Possible long-term dependencies in the speech signal will be described by HE (Hurst Exponent) [54].

Another set of measures is based on entropy. We will use SHE (Shannon Entropy), RE (second-order Rényi Entropy) [70], CE (Correlation Entropy) [40,70], RBE1 (first-order Rényi Block Entropy) [40], RBE2 (second-order Rényi Block Entropy) [40], AE (Approximate Entropy) [32,71,72], SE (Sample Entropy) [72] and RPDE (Recurrence Probability Density Entropy) [37]. Generally the entropy is a measure of uncertainty and it can be used to quantify the complexity of a system. Rényi entropies quantify the loss of information in time in a dynamic system [40], correlation entropy gives an indication of the predictability of the nonlinear time series [70] and RPDE represents the uncertainty in the measurement of the pitch period [37]. The only difference between AE and SE is that SE does not evaluate a comparison of embedding vectors with themselves.

Another measure we have considered in this work is FMMI (First Minimum of Mutual Information function) which was found by Henriquez et al. as a feature that better discriminates among the different voice qualities of the multiquality database [40]. To include also a measure of sensitivity to an initial condition, the LLE (Largest Lyapunov Exponent) was selected [32,54]. We also used detrended fluctuation analysis (DFA) to characterize the self-similarity of the graph of a signal from a stochastic process [34,37]. In this field NSE (Normalized Scaling Exponent) and FA (Fluctuation Amplitudes) were evaluated.

## 2.9. High-level features

The features that are calculated for each speech segment separately form a vector or a matrix at the output of the parametrization process. This representation must be then transformed to a scalar value to be able to carry out the next processing like statistical analysis and classification. This is usually done by an extraction of some kind of statistics. These statistics are called high-level features while parameters extracted directly from the speech signal are called local features. If the local feature is represented by a matrix, then the high-level feature is calculated for each row separately (this monitors feature changes in time). We have extracted 60 high-level features:

- range, relative range, interquartile range, rel. interquartile range, interdecile range, rel. interdecile range, interpercentile range, rel. interpercentile range, studentized range.
- mean, geometric mean, harmonic mean, mean excluding 10%, 20%, 30%, 40% and 50%, of outliers, median, mode.
- var, std, mean absolute deviation, median absolute dev., geometric standard dev., coefficient of variation, index of dispersion.
- 3rd, 4th, 5th and 6th moment, kurtosis, skewness, Pearson's 1st and 2nd skewness coefficient
- 1st, 5th, 10th, 20th, 30th, 40th, 60th, 70th 80th, 90th 95th and 99th percentile, 1st and 3rd quartile.
- Slope, offset and error of linear regression.
- Modulation, Shannon entropy, second-order Rényi entropy.

## 2.10. Newly designed features

In Sections 2.1–2.8 we have presented 92 local features that are already used in the field of speech signal processing. However during the experiments these parameters will be extended by the other 36 features that are originally introduced in this work.

### 2.10.1. Features based on modulation spectra

An initial step of modulation spectra calculation employs a short-time Fourier transform (STFT) of the discrete input speech signal  $s[n]$  with length  $N$ :

$$S[k, m] = \sum_{n=0}^{N-1} s[n]w[n-mL]e^{-jk(2\pi/N)n},$$

$$k = 0, 1, \dots, N-1,$$

$$m = 0, 1, \dots, M-1, \quad (1)$$

where  $w[n]$  is a window function (in our case a Hamming window) with a step of  $L$  samples and  $M$  number of speech segments. Consequently the power spectra  $|S[k, m]|^2$  is extracted and filtered by a bank of  $P$  triangular filters equidistantly spaced out in the mel scale. This procedure forms a matrix  $X[p, m]$  with subbands  $p = 1, 2, \dots, P$ . A distribution of amplitudes of each subband envelope  $X[p, m]$  of the voiced speech signal has a strong exponential component which is suppressed by logarithmization and mean subtraction [33]

$$\hat{X}[p, m] = \ln(X[p, m]) - \overline{\ln(X[p, m])}, \quad (2)$$

where  $\overline{\cdot}$  corresponds to the average operator over  $m$ . Next the frequency analysis of subband envelopes is performed using the discrete Fourier transform (DFT)

$$\Psi[p, l] = \sum_{m=0}^{M-1} \hat{X}[p, m]e^{-jl(2\pi/M)m},$$

$$l = 0, 1, \dots, M-1, \quad (3)$$

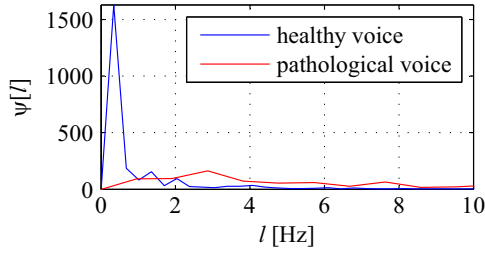
where  $p$  and  $l$  denote the acoustic and modulation frequency respectively. In the last step of modulation spectra extraction the second power of each subband is taken and normalized which partially suppresses the effect of training and testing conditions mismatch

$$\Psi_n[p, l] = \frac{\Psi[p, l]}{\sum_l \Psi[p, l]}. \quad (4)$$

The features we are proposing are based on a function  $\psi[l]$  which is extracted from the normalized modulation spectra according to

$$\psi[l] = \sum_{p=1}^P \Psi_n[p, l]. \quad (5)$$

- max, min, position of max, position of min, relative pos. of max, relative pos. of min.



**Fig. 1.** Function  $\psi[l]$  calculated for healthy and pathological female vowels [a] obtained from the database PdA [73,74].

Due to the instability of vocal fold vibrations pathological speech exhibits larger energy spread on higher modulation frequencies. This fact can be sufficiently expressed in function  $\psi[l]$ . Fig. 1 illustrates function  $\psi[l]$  calculated for healthy and pathological female vowels [a] obtained from the database Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid [73,74]. As it can be seen the peak of function  $\psi[l]$  in the case of healthy voice is higher, narrower and more shifted to lower modulation frequencies than in the case of the pathological one.

The proposed features derived from  $\psi[l]$  are MSER (Modulation Spectra Energy Ratio), MFP (Modulation Frequency of Peak) and RPHM (Relative Peak Height of Modulation spectra). MSER is defined as

$$\text{MSER} = \frac{\sum_{l=l_5 \text{ Hz}}^{l_5 \text{ Hz}} \psi[l]}{\sum_{l=l_5 \text{ Hz}+1}^{M-1} \psi[l]}, \quad (6)$$

where  $l_5 \text{ Hz}$  is a sample corresponding to 5 Hz modulation frequency (this limit was empirically found). MFP is defined as

$$\text{MFP} = \arg \max_l (\psi[l]), \quad (7)$$

where MFP is given in Hz. Finally RPHM can be calculated according to the following expression:

$$\text{RPHM} = \frac{\psi[i] - r[i]}{\psi[i]}, \quad (8)$$

$$i = \arg \max_l (\psi[l]),$$

where  $r$  is a linear regression line of  $\psi[l]$  for  $l = l_5 \text{ Hz}, \dots, M-1$  and  $i$  is the index at which the function  $\psi[l]$  reaches maximum value.

### 2.10.2. Features based on inferior colliculus coefficients

Using inferior colliculus coefficients (ICC) it is possible to extract the frequency content of the modulation envelopes applied to different bands of an auditory stimuli [66]. Similarly to modulation spectra, the ICC extraction process employs at the beginning STFT and consequent calculation of power spectra  $|S[k, m]|^2$ . However in the next step instead of the bank of the triangular filters, the bank of  $P$  mel-spaced gammatone filters are applied. We used filters defined by the impulse response

$$g[n] = \left(\frac{n}{f_s}\right)^{o-1} \cdot \cos\left(\frac{2\pi f_c n}{f_s}\right) \cdot e^{-2\pi b n / f_s}, \quad (9)$$

$$b = 24.7 \left(4.37^{-3} f_c + 1\right),$$

where  $f_c$  is the center frequency in Hz,  $o$  the filter order (in our case 4) and  $f_s$  the sampling frequency in Hz. The DFT on the subband envelopes was applied next, this time without prior to normalization

$$T[p, l] = \sum_{m=-\infty}^{\infty} X[p, m] e^{-j l (2\pi / M) m}. \quad (10)$$

Finally the magnitude spectrum of each envelope  $|T[p, l]|$  is filtered by a bank of  $Q=13$  resonance filters with exponentially spaced

frequencies from 12 to 107 Hz. In this work we used resonance filters defined by the following transfer function:

$$H(z) = \frac{0.1z^2 - 0.09}{z^2 - 1.8 \cos(2\pi f_c / f_s)z + 0.81}. \quad (11)$$

At the output of this process a matrix  $\Xi[p, q]$  ( $q = 1, 2, \dots, Q$ ) of ICC for the input signal  $s[n]$  is extracted.

The proposed features are based on a function  $\xi[p]$  which is extracted from  $\Xi[p, q]$  according to

$$\xi[p] = \sum_{q=1}^Q \ln(\Xi[p, q]). \quad (12)$$

Contrary to  $\psi[l]$  this function reflects the misplacement of articulators better than the instability of vocal fold vibrations. In Fig. 2 it is possible to see an example of  $\xi[p]$  calculated for the pathological and healthy voice obtained from the PdA database. The features derived from  $\xi[p]$  are ICER (Inferior Colliculus Energy Ratio) and RPHIC (Relative Peak Height of Inferior Colliculus). The ratio ICER is defined by the following expression:

$$\text{ICER} = \frac{\sum_{p=1}^{12} \xi[p]}{\sum_{p=13}^{20} \xi[p]}. \quad (13)$$

Similarly to RPHM the RPHIC is extracted using the regression line but in this case  $r[p]$  is calculated for  $p = 13, \dots, 20$

$$\text{RPHIC} = \frac{\xi[i] - r[i]}{\xi[i]}, \quad (14)$$

$$i = \arg \max_p (\xi[p]).$$

### 2.10.3. Features based on bicepstrum

We will use a definition of the real bicepstrum  $c[n_1, n_2]$  similar to the definition of the real cepstrum

$$c[n_1, n_2] = \frac{1}{N^2} \sum_{k_1, k_2=0}^{N-1} \hat{B}[k_1, k_2] e^{j \frac{2\pi}{N} (k_1 n_1 + k_2 n_2)}, \quad (15)$$

$$\hat{B}[k_1, k_2] = \ln(|B[k_1, k_2]|), \quad (16)$$

where the bispectrum  $B[k_1, k_2]$  is calculated as the DFT of triple correlation or circular triple correlation  $\gamma[n_1, n_2]$  [75]. In our work we used the second approach based on  $\gamma[n_1, n_2]$

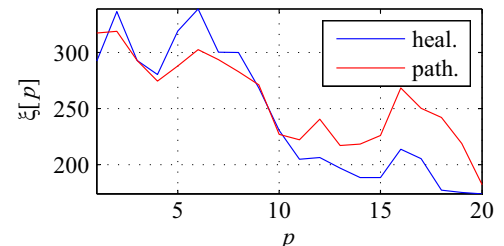
$$B[k_1, k_2] = \sum_{n_1, n_2=0}^{N-1} \gamma[n_1, n_2] e^{-j(2\pi/N)(k_1 n_1 + k_2 n_2)},$$

$$\gamma[n_1, n_2] = \frac{1}{N} \sum_{n=0}^{N-1} \delta[n] \delta[n+n_1] \delta[n+n_2],$$

$$\delta[i] = s[(n+i) \bmod N]. \quad (17)$$

The first feature we are proposing in the present work is BCII (BiCepstral Index Interference)

$$\text{BCII} = \frac{1}{N^2 - 1} \frac{1}{\max(b[n_1, n_2])}.$$



**Fig. 2.** Function  $\xi[p]$  calculated for the healthy and pathological female vowels [a] obtained from the database PdA [73,74].

$$\sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-2} |b[n_1, n_2+1] - b[n_1, n_2]|, \quad (18)$$

$$b[n_1, n_2] = \frac{\left| \sum_{m=0}^{M-1} c_m[n_1, n_2] \right|}{\sum_{m=0}^{M-1} |c_m[n_1, n_2]|}, \quad (19)$$

where  $c_m[n_1, n_2]$  is the bicepstrum calculated from the  $m$ th speech segment. The two features introduced next are HFEBC (High Frequency Energy of one-dimensional BiCepstral index) and LFEBC (Low Frequency Energy of one-dimensional BiCepstral index)

$$\text{LFEBC} = \frac{\sum_{n=0}^L \rho[n]}{\sum_{n=0}^{N-1} \rho[n]}, \quad (20)$$

$$\text{HFEBC} = \frac{\sum_{n=L+1}^{N-1} \rho[n]}{\sum_{n=0}^{N-1} \rho[n]}, \quad (21)$$

$$\begin{aligned} \rho[n] &= c[n, n], \\ L &= \frac{f_s}{f_{\max}}, \end{aligned} \quad (22)$$

where  $\rho[n]$  (for  $n = 0, 1, \dots, N-1$ ) is the one-dimensional bicepstral index and  $f_{\max}$  the maximum expected fundamental frequency (in our case  $f_{\max} = 350$  Hz).

It is supposed that pathological voice contains much more white noise (symmetrically distributed) than healthy voice. This is especially due to incorrect glottal closure. This kind of noise disappears in the real cepstrum estimated by  $\rho[n]$  (see Fig. 3), therefore a difference between the real cepstrum  $c[n]$  (estimated using DFT) and the one-dimensional bicepstral index can estimate noise components of the analyzed signal. According to this idea BCMII (BiCepstrum Module Interference Index) and BCPPI (BiCepstrum Phase Interference Index) are proposed

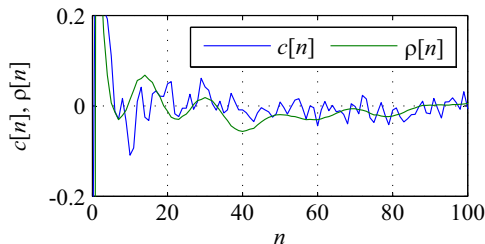
$$\begin{aligned} \text{BCMII} &= \frac{1}{N^2 - 1} \frac{1}{\max(\eta_m[m])} \\ &\cdot \sum_{m=0}^{M-2} |\eta_m[m+1] - \eta_m[m]|, \end{aligned} \quad (23)$$

$$\begin{aligned} \text{BPMII} &= \frac{1}{N^2 - 1} \frac{1}{\max(\eta_p[m])} \\ &\cdot \sum_{m=0}^{M-2} |\eta_p[m+1] - \eta_p[m]|, \end{aligned} \quad (24)$$

$$\eta_m[m] = \sum_{n=0}^{N-1} (|c_m[n]| - |\rho_m[n]|)^2, \quad (25)$$

$$\eta_p[m] = \sum_{n=0}^{N-1} (\text{ang}(\tilde{c}_m[n]) - \text{ang}(\tilde{\rho}_m[n]))^2, \quad (26)$$

where  $c_m[n]$  and  $\rho_m[n]$  are the  $m$ th frame real cepstrum and the one-dimensional bicepstral index respectively.  $\tilde{c}_m[n]$  is the



**Fig. 3.** Comparison of the one-dimensional bicepstral index  $\rho[n]$  and the real cepstrum  $c[n]$  calculated for a maintained vocal [a] uttered by a healthy speaker. Only first 101 samples are displayed.

complex cepstrum and  $\tilde{\rho}_m[n]$  is the one-dimensional bicepstral index where the absolute value in Eq. (16) was not taken.

The other features, based on  $c[n]$  and  $\rho[n]$  are LCBCER (Low Cepstra/BiCepstra Energy Ratio) and HCBCER (High Cepstra/BiCepstra Energy Ratio)

$$\text{LCBCER} = \frac{\sum_{n=0}^L |c[n]|}{\sum_{n=0}^L |\rho[n]|}, \quad (27)$$

$$\text{HCBCER} = \frac{\sum_{n=L+1}^{N-1} |c[n]|}{\sum_{n=L+1}^{N-1} |\rho[n]|}. \quad (28)$$

According to these equations we also propose the features LSBER (Low Spectra/Bispectra Energy Ratio) and HSBER (High Spectra/Bispectra Energy Ratio), however the ratios of spectrum  $S[k]$  and one-dimensional bispectral index  $\vartheta[k]$  for  $L = \lfloor N/2 \rfloor$  are calculated in this case.

The other two features BCMD (BiCepstral Module Distance) and BCPD (BiCepstral Phase Distance) are based on distance measures. The values of these features for the  $m$ th speech segment are

$$\begin{aligned} \text{BCMD} &= \sum_{n_1, n_2=0}^{N-1} \left| |\tilde{c}_{m+1}[n_1, n_2]| \right. \\ &\quad \left. - |\tilde{c}_m[n_1, n_2]| \right|, \end{aligned} \quad (29)$$

$$\begin{aligned} \text{BCPD} &= \sum_{n_1, n_2=0}^{N-1} \left| \text{ang}(\tilde{c}_{m+1}[n_1, n_2]) \right. \\ &\quad \left. - \text{ang}(\tilde{c}_m[n_1, n_2]) \right|, \end{aligned} \quad (30)$$

where  $\tilde{c}_m[n_1, n_2]$  is the complex bicepstrum calculated without the absolute value in Eq. (16). Similarly to Eqs. (29) and (30) features BMD (Bispectral Module Distance) and BPD (Bispectral Phase Distance) are also extracted, where  $B_m[k_1, k_2]$  is used instead of  $\tilde{c}_m[n_1, n_2]$ .

#### 2.10.4. Different kernel based approximate and sample entropy

Approximate entropy (AE) is a measure of regularities inside the analyzed time series. The advantage of AE is that it can robustly estimate the system complexity using just a limited number of samples (100–5000) [76]. To define AE we need to firstly reconstruct a state space using Takens' embedding theorem [77]

$$\begin{aligned} x[n] &= [s[n], s[n+\tau], \dots, s[n+(m-1)\tau]], \\ n &= 0, 1, \dots, N-1-(m-1)\tau, \end{aligned} \quad (31)$$

where  $x[n]$  is an embedding vector,  $m$  is an embedding dimension and  $\tau$  is a time delay ( $\tau = 1$  for AE). Next, the regularity quantity of a particular pattern  $C[i, m, r]$  is defined as

$$C[i, m, r] = \frac{1}{N-m} \sum_{j=0}^{N-m} \kappa(i, j, r). \quad (32)$$

Finally AE can be calculated according to expression [32]

$$\text{AE} = \Phi[m, r] - \Phi[m+1, r], \quad (33)$$

$$\Phi[m, r] = \frac{1}{N-m} \sum_{i=0}^{N-m} \ln(C[i, m, r]). \quad (34)$$

A disadvantage of AE is its dependence on the signal length due to the self comparison of points in the attractor. This fact can be avoided using the sample entropy (SE) which does not evaluate the comparison of embedding vectors among themselves [78]

$$\text{SE} = \Gamma[m, r] - \Gamma[m+1, r], \quad (35)$$

$$\Gamma[m, r] = \frac{1}{N-m} \sum_{i=0}^{N-m} \ln(C_x[i, m, r]), \quad (36)$$

$$C_x[i, m, r] = \frac{1}{N-m} \sum_{j=0, i \neq j}^{N-m} \kappa(i, j, r). \quad (37)$$

The usual values for embedding dimension are  $m=1, 2$ . In this work we used  $m=2$  which provides more detailed reconstruction. Finally, a function must be defined  $\kappa(i, j, r)$  we call kernel function. Originally the AE or SE are based on

$$\kappa(i, j, r) = \Theta\{r - d(x[i], x[j])\}, \quad (38)$$

$$d(x[i], x[j]) = \max_k |s[i+k] - s[j+k]|, \quad (39)$$

$$k = 0, 1, \dots, m-1,$$

where  $\Theta$  is the Heaviside function and  $r$  a radius in our case calculated according to

$$r = 0.2 \text{std}(s[n]). \quad (40)$$

We will denote these entropies as AE (Heaviside) and SE (Heaviside) respectively. Orozco-Arroyave et al. proposed AE (Gaussian) and SE (Gaussian) based on the Gaussian kernel [78]

$$\kappa(i, j, r) = \exp\left(-\frac{\|x[i] - x[j]\|^2}{10r^2}\right). \quad (41)$$

In this work we propose AE and SE based on the other 6 kernels: exponential kernel

$$\kappa(i, j, r) = \exp\left(-\frac{\|x[i] - x[j]\|}{2r^2}\right); \quad (42)$$

Laplacian kernel,

$$\kappa(i, j, r) = \exp\left(-\frac{\|x[i] - x[j]\|}{r}\right); \quad (43)$$

circular kernel,

$$\kappa(i, j, r) = \frac{2}{\pi} \arccos\left(-\frac{\|x[i] - x[j]\|}{r}\right) - \frac{2}{\pi} \frac{\|x[i] - x[j]\|}{r} \sqrt{1 - \left(\frac{\|x[i] - x[j]\|}{r}\right)^2}, \quad (44)$$

for  $\|x[i] - x[j]\| < r$ , zero otherwise; spherical kernel,

$$\kappa(i, j, r) = 1 - \frac{3}{2} \frac{\|x[i] - x[j]\|}{r} + \frac{1}{2} \left(\frac{\|x[i] - x[j]\|}{r}\right)^3, \quad (45)$$

for  $\|x[i] - x[j]\| < r$ , zero otherwise; Cauchy kernel,

$$\kappa(i, j, r) = \frac{1}{1 + (\|x[i] - x[j]\|^2/r^2)}, \quad (46)$$

for  $\|x[i] - x[j]\| < r$ , zero otherwise and triangular kernel,

$$\kappa(i, j, r) = 1 - \frac{|x[i] - x[j]|}{r}, \quad (47)$$

for  $|x[i] - x[j]| < r$ , zero otherwise.

#### 2.10.5. Features based on empirical mode decomposition

As was mentioned in Section 2.7 Tsanas et al. proposed several measures of SNR and NSR based on EMD [34]

$$\text{IMF-SNR} = \frac{\sum_{i=1}^I 4\mu_i}{\sum_{i=1}^3 \mu_i}, \quad (48)$$

$$\text{IMF-NSR} = \frac{\sum_{i=1}^2 \hat{\mu}_i}{\sum_{i=3}^I \hat{\mu}_i}, \quad (49)$$

where  $\mu_i$  is a parameter, or mean sequence value, calculated from the original  $i$ th IMF and  $I$  is the total number of the IMFs. In the case of  $\hat{\mu}_i$  the  $i$ th IMF was logarithmized before the consequent parametrization. To extract the sequence from IMF, Tsanas et al.

used SEO (Squared Energy Operator) and TKEO. As a parameter they also used SHE.

We have extended this idea on these parameters: IMF-SNR<sub>RE</sub> (based on second-order Rényi Entropy), IMF-SNR<sub>ZCR</sub> (based on Zero-Crossing Rate) and IMF-NSR<sub>RE</sub>. In the case of IMF-SNR<sub>RE</sub>  $\mu_i$  is defined as

$$\mu_i = -\log_2 \left( \sum_{j=1}^J p^2(x_j^i) \right), \quad (50)$$

where  $p^2(x_j^i)$  is the probability  $P(\text{IMF}_i = x_j^i)$  and  $\{x_1^i, x_2^i, \dots, x_J^i\}$  are the possible values of the  $i$ th IMF. In the case of IMF-SNR<sub>ZCR</sub>  $\mu_i$  is calculated according to

$$\mu_i = \frac{1}{N} \sum_{n=1}^{N-1} |\text{sgn}(f_i[n]) - \text{sgn}(f_i[n-1])|, \quad (51)$$

where  $f_i[n]$  is the  $i$ th IMF.

The time-varying high frequency components present in the 1st IMF represent the noise part of the speech signal. We propose a new feature IMF-FD which is based on the fractal dimension calculated from the 1st IMF. This complexity measure appropriately quantifies the amount of noise present in the signal and indirectly describes hoarseness or breathiness. The feature is defined as

$$\text{IMF-FD} = \frac{\log_{10} N}{\log_{10} N + \log_{10} (N/N + 0.4N_{\text{ch}})}$$

$$N_{\text{ch}} = \sum_{n=1}^{N-1} |\text{sgn}(f_1[n]) - \text{sgn}(f_1[n-1])|. \quad (52)$$

Another feature based on the first intrinsic mode function is IMF-CPP (Cepstral Peak Prominence extracted from the 1st IMF). It can be calculated as

$$\text{IMF-CPP} = \frac{c[i] - r[i]}{c[i]},$$

$$i = \arg \max_n (c[n]),$$

$$n = \frac{f_s}{f_{\text{max}}}, \dots, N-1, \quad (53)$$

where  $c[n]$  is real cepstrum calculated from  $f_1[n]$  and  $r[i]$  is a linear regression line of  $c[n]$  for  $n = f_s/f_{\text{max}}, \dots, N-1$ . We consider  $f_{\text{max}} = 350$  Hz.

The last feature proposed in this work is IMF-GNE (Glottal-to-Noise Excitation ratio based on the 1st IMF). The whole procedure of IMF-GNE extraction can be described in following steps:

1. Segment the speech signal and extract the 1st IMF for each frame.
2. Repeat step 3–6 for each segment.
3. Do an inverse filtering of the 1st IMF.
4. For each band of 1000 Hz with 1000 Hz step get the Hilbert envelope (the absolute value of analytical signal).
5. Calculate cross correlation functions for all possible combinations of Hilbert envelopes and pick the maximum of each function.
6. Pick the maximum from all the maxima in 5.

### 3. Databases

To provide robust results 3 (English, Spanish and Czech) databases have been used during the testing procedure. Each database represents a different language group (Germanic, Romanic and Slavic). This approach is advantageous from the cultural difference point of view. Speakers of different languages exhibit especially different prosodic characteristics. The aim of this work



is to find features significant for the particular language, but we have focused on a selection of features that are language-independent as well.

### 3.1. MEEI disordered voice database

The Massachusetts Eye and Ear Infirmary (MEEI) database [25] has been for many years used as a benchmark in the field of pathological speech analysis. This commercially available database consists of 53 healthy and 657 pathological speakers with different pathologies (e.g. adductor spasmodic dysphonia, conversion dysphonia, erythema, and hyperfunction). The data of each speaker contain 12 s of a standard text “The Rainbow Passage” [79] and a sustained phonation of the vowel [a] pronounced as in the word “father”. The recordings are sampled at  $f_s = 50$  kHz or  $f_s = 25$  kHz. For our purpose only the vowels [a] are used.

Although MEEI consists of approximately 700 speakers in total, the age distributions between normal and pathological speakers are not matched. Therefore the amount of pathological speakers is usually limited to 173 according to criteria published by Parsa and Jamieson (we call this “limited version of MEEI database”) [15]. The statistical characteristics of MEEI database used in this work can be found in Table 1.

Although this database is very popular and very often used, its size and data are considered insufficient. Table 2 shows detection results (obtained on the MEEI database) of several works. The highest accuracy reached by Henriquez et al. is 99.69% [40]. However using the same features and same experimental setup the authors achieved an accuracy of 82.47% using the “Multi-quality” database [40]. This fact shows that despite the popularity of the MEEI database there is a need to introduce new, larger and more complex (when considering the speech tasks) benchmark databases in order to provide more reliable results and conclusions. A discussion on the reliability of results mentioned in Table 2 can be found in Section 3.4.

### 3.2. PdA database

The second database we have used is Príncipe de Asturias (PdA) database [73,74]. This database consists of 239 healthy and 200 pathological speakers with different organic pathologies (e.g. nodules, polyps, oedemas, and carcinomas). Every speaker uttered a sustained Spanish vowel [a]. The recordings are sampled at  $f_s = 25$  kHz. The statistical characteristics of this database can be found in Table 3. Table 4 shows detection results (obtained on the PdA database) of two available works. As can be seen, the accuracies are not so high as in the case of MEEI database. Moreover, the PdA consists of more speakers than the limited version of MEEI. All these facts highlight the high potential of the PdA for future use.

### 3.3. PARCZ database

The last database we have included in our test is the Czech Parkinsonian Speech Database (PARCZ) recorded at St. Anne's University Hospital in the Czech Republic. This database consists

**Table 1**  
Statistical characteristics of the MEEI database used in this work.

Speakers	Number		Mean age		Age range		std of age	
	Male	Female	Male	Female	Male	Female	Male	Female
Healthy	21	32	38.81	34.16	26–59	22–52	8.49	7.87
Pathological	70	103	41.7	37.59	26–58	21–51	9.40	8.19

**Table 2**  
Summary of pathological speech detection results obtained on the MEEI database.

Reference	Accuracy (%)
Henriquez et al. [40]	99.69
Diabazar et al. [80]	99.44
Parsa and Jamieson [15]	98.70
Alpan et al. [81]	98.70
Hariharan et al. [82]	98.45
Arias-Londono et al. [83]	98.23

**Table 3**  
Statistical characteristics of the PdA database used in this work.

Speakers	Number		Mean age		Age range		Std of age	
	Male	Female	Male	Female	Male	Female	Male	Female
Healthy	101	138	34.44	35.29	18–78	8–71	16.24	14.73
Pathological	74	126	48.05	36.71	11–76	9–72	13.89	13.14

**Table 4**  
Summary of pathological speech detection results obtained on the PdA database.

Reference	Accuracy (%)
Arias-Londono et al. [74]	84.15
Vasilakis and Stylianou [39]	77.68

**Table 5**  
Statistical characteristics of the PARCZ database used in this work.

Speakers	Number		Mean age		Age range		Std of age	
	Male	Female	Male	Female	Male	Female	Male	Female
Healthy	26	26	65.65	62.15	49–83	45–87	9.02	9.50
Pathological	36	21	66.22	68.81	46–87	49–86	9.20	9.00

of 52 healthy speakers and 57 speakers with Parkinson's disease (PD) who suffer from hypokinetic dysarthria [84]. This database contains 91 speech tasks (free speech, reading text, maintained vowels, and diadochokinetic tasks) which are used for an analysis of speech dysfunctions that usually accompany PD. However for our purpose only the sustained Czech vowel [a] is used. The recordings are sampled at  $f_s = 48$  kHz. The statistical characteristics of the PARCZ database can be found in Table 5. As it can be seen, contrary to the MEEI or PdA, PARCZ is more focused on elder people.

### 3.4. The rule of 30

To decide whether the corpus size is sufficient for the robust conclusions Doddington et al. introduced “the rule of 30” which comes directly from the binomial distribution, assuming independent trials [85]. The rule is “To be 90% confident that the true error rate is within  $\pm 30\%$  of the observed error rate, there must be at least 30 errors.”

If we apply this rule to the original MEEI database (710 speakers), then the observed error rates below 4.25% cannot be considered as reliable. Moreover if the MEEI database is limited to 226 speakers, then threshold reaches 13.27%. In other words, although the results in Table 2 predict promising approaches for

pathologic speech detection results, we must be critic about these values.

If we apply this rule to the PdA (436 speakers) or PARCZ (109 speakers) databases, we get the thresholds of 6.88% and 27.52% respectively. According to this values it can be said that the results obtained with the PdA are more reliable (contrary to PARCZ).

#### 4. Experiments

All the databases were resampled to  $f_s = 16$  kHz and in dependence on the next processing the data have been divided into 9 groups. Two approaches have been considered: gender-dependent and gender independent. Each database is randomly divided into 75% and 25% training and testing subsets respectively. The classifier is evaluated consequently. This procedure (data split, classifier tuning and evaluation) is repeated 100 times. The resulting accuracy (ACC), sensitivity (SEN) and specificity (SPE) are calculated according to

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100 [\%], \quad (54)$$

$$SEN = \frac{TP}{TP + FN} \cdot 100 [\%], \quad (55)$$

$$SPE = \frac{TN}{TN + FP} \cdot 100 [\%], \quad (56)$$

where  $TP$  (True Positive) and  $FP$  (False Positive) represent the number of correctly identified pathological speakers and number of speakers diagnosed as pathological, but being healthy. Similarly,  $TN$  (True Negative) and  $FN$  (False Negative) represent the total number of correctly identified healthy speakers, and pathological speakers evaluated as healthy controls.

Before classification the training data are z-score normalized. The testing data are normalized by subtracting the training set mean and dividing by the training set standard deviation for each feature.

##### 4.1. Parameterization programs

Several toolboxes and programs have been used for the purpose of feature extraction. Features based on the detrended fluctuation analysis have been calculated using FastDFA [86]. The glottal quotients ( $GQ_{open}$  and  $GQ_{closed}$ ) were extracted using the algorithm DYPSA implemented in VOICEBOX [87]. To estimate the largest Lyapunov exponent (LLE) TSTOOL has been used [88]. The software Praat has been used to estimate the fundamental frequency  $F_0$ , all kinds of jitter and shimmer, harmonic-to-noise ratio (HNR) and formant frequencies ( $F_1, F_2, F_3$ ) [89]. The rest of features (108 in total), including the features introduced in this work, have

been implemented in the Neurological Disorder Analysis Tool (NDAT) developed at the Brno University of Technology [90].

##### 4.2. Feature selection

Considering all possible local and high-level feature combinations the parameterization process extracts approximately 28,000 features for each speaker. This feature space reduced for each scenario using the filtering feature selection approach based on the non-parametric Mann–Whitney  $U$  test. The significance level was set to  $\alpha = 0.05$ . This feature selection method is relatively simple, but there are several serious studies indicating that there are scenarios where simple univariate methods perform similarly or even better than more complex methods. For instance Haury et al. show that a simple Student's  $t$ -test provides better results than SVM-RFE (Support Vector Machine Recursive Feature Elimination), GFS (Greedy Forward Selection), LASSO (Least Absolute Shrinkage and Selection Operator) or elastic net [91].

After the evaluation a list of the ten most significant features is drawn up. Finally the density estimation plots (computed using kernel density estimation with Gaussian kernels) of the most significant features in all scenarios are given.

##### 4.3. Classification methods

In our test we have used 2 classifiers: SVM (Support Vector Machine with a radial kernel) and RF (Random Forest). Regarding the SVM, the parameter kernel gamma  $\gamma$  and penalty parameter  $C$  were optimized using a grid search for possible values.

#### 5. Results

A summary of pathological speech detection results expressed by accuracy, sensitivity and specificity can be found in Table 6. This table also provides some statistics related to the number of features selected in each scenario. In the case of MEEI database the accuracy, sensitivity and specificity were equal to  $100.0 \pm 0.0\%$  in all scenarios (considering both genders together and separately) when using an RF classifier. A discussion focused on the credibility of these results is given below. In comparison to RF the SVM classifier provided slightly worse results.

In the case of the PdA database the best results were found when classifying by RF too. The accuracy ( $80.9 \pm 5.1\%$ ) and specificity ( $85.4 \pm 6.7\%$ ) are larger for male speakers while the sensitivity ( $77.2 \pm 7.7\%$ ) is larger for the female ones. When considering both genders together the accuracy ( $82.1 \pm 3.3\%$ ) and sensitivity ( $80.0 \pm 5.9\%$ ) reach the best values in the frame of all PdA scenarios, however the specificity is approximately 1.6% lower than in the case of the male-only scenario. In comparison to the best accuracies published by Arias-Londono et al. our approach

**Table 6**

Summary of pathological speech detection results represented as mean  $\pm$  std (%) (SVM – Support Vector Machine with a radial kernel, RF – Random Forest, F – female, M – male, MF – all genders).

ID	Scenario		No. of sel. features (Mean $\pm$ std)	Accuracy		Sensitivity		Specificity	
	Dataset	Gender		SVM	RF	SVM	RF	SVM	RF
M1	MEEI	F	13,996 $\pm$ 288	99.5 $\pm$ 1.5	<b>100.0 <math>\pm</math> 0.0</b>	99.3 $\pm$ 2.0	<b>100.0 <math>\pm</math> 0.0</b>	<b>100.0 <math>\pm</math> 0.0</b>	<b>100.0 <math>\pm</math> 0.0</b>
M2	MEEI	M	13,561 $\pm$ 398	99.2 $\pm$ 1.7	<b>100.0 <math>\pm</math> 0.0</b>	99.1 $\pm$ 2.1	<b>100.0 <math>\pm</math> 0.0</b>	99.3 $\pm$ 3.3	<b>100.0 <math>\pm</math> 0.0</b>
M3	MEEI	MF	15,521 $\pm$ 231	99.9 $\pm$ 0.4	<b>100.0 <math>\pm</math> 0.0</b>	99.8 $\pm$ 0.5	<b>100.0 <math>\pm</math> 0.0</b>	99.9 $\pm$ 0.7	<b>100.0 <math>\pm</math> 0.0</b>
P1	PdA	F	9,726 $\pm$ 447	75.7 $\pm$ 4.3	<b>78.5 <math>\pm</math> 4.9</b>	72.8 $\pm$ 6.5	<b>77.2 <math>\pm</math> 7.7</b>	78.4 $\pm$ 6.8	<b>79.6 <math>\pm</math> 7.3</b>
P2	PdA	M	9,721 $\pm$ 458	78.6 $\pm$ 5.1	<b>80.9 <math>\pm</math> 5.1</b>	71.0 $\pm$ 10.1	<b>74.7 <math>\pm</math> 9.8</b>	84.2 $\pm$ 6.1	<b>85.4 <math>\pm</math> 6.7</b>
P3	PdA	MF	11,540 $\pm$ 419	77.7 $\pm$ 3.2	<b>82.1 <math>\pm</math> 3.3</b>	74.9 $\pm$ 5.3	<b>80.0 <math>\pm</math> 5.9</b>	80.1 $\pm$ 5.0	<b>83.8 <math>\pm</math> 5.1</b>
C1	PARCZ	F	2,141 $\pm$ 415	65.9 $\pm$ 11.9	<b>67.1 <math>\pm</math> 8.3</b>	10.3 $\pm$ 16.9	<b>35.3 <math>\pm</math> 30.3</b>	79.0 $\pm$ 15.3	<b>91.4 <math>\pm</math> 11.3</b>
C2	PARCZ	M	2,121 $\pm$ 489	<b>67.3 <math>\pm</math> 10.7</b>	66.5 $\pm$ 10.3	42.6 $\pm$ 21.6	<b>50.4 <math>\pm</math> 20.2</b>	79.4 $\pm$ 14.5	<b>83.6 <math>\pm</math> 16.0</b>
C3	PARCZ	MF	1,750 $\pm$ 331	65.4 $\pm$ 7.6	<b>67.9 <math>\pm</math> 6.0</b>	31.0 $\pm$ 14.6	<b>39.3 <math>\pm</math> 14.9</b>	79.3 $\pm$ 10.0	<b>87.5 <math>\pm</math> 8.5</b>

provides a lower accuracy by 2.05%, however it should be highlighted that the authors used classifier score fusion, which is not considered in this work [74]. When looking at their single-classifier solution, the authors reported these values: 81.70% (accuracy), 80.50% (sensitivity), 82.91% (specificity). Comparing these values to our results, we can say that our approach provides better accuracy (by 0.40%) and specificity (by 0.89%), but worse sensitivity (by 0.50%). Moreover, Arias-Londono et al. used in their work an older version of PdA database which has only 199 healthy speakers, while our version has 239. Therefore our results should be more trustworthy. In comparison to the work of Vasilakis et al. our approach outperformed the classification results provided by the authors [39].

Regarding the PARCZ database the results are much worse than in the case of MEEI or PdA. In scenario C1 (female speakers) the best accuracy was  $67.1 \pm 8.3\%$  (classification by RF), sensitivity  $35.3 \pm 30.3\%$  (SVM) and specificity  $91.4 \pm 11.3\%$  (RF). The accuracy ( $67.3 \pm 10.7\%$ ) and sensitivity ( $50.4 \pm 20.2\%$ ), both obtained using the SVM, were slightly better in the case of male speakers. Vice versa, the specificity ( $83.6 \pm 16.0\%$ ) was worse. Finally in scenario C3 (both genders) the best accuracy was  $67.9 \pm 6.0\%$  (RF), sensitivity  $39.3 \pm 14.9\%$  (SVM) and specificity  $87.5 \pm 8.5\%$  (RF). It should be also mentioned that in comparison to MEEI or PdA the PARCZ database exhibits much larger standard deviations.

The poor classification results are caused probably by the fact that PD patients were in different progression stages of hypokinetic dysarthria, from first to more advanced ones (this fact most likely explains the large standard deviations that were obtained). It means that we were performing a two-class classification over the data that can be split into approximately 4 classes (1 healthy and 3 levels of dysarthria: mild, moderate and severe). It can be an issue for future work to develop a system that would not only identify the presence of pathological speech, but also estimate the level of voice pathology. The PARCZ database is a good candidate for a development of such a system. The system can be interesting especially if we are able to detect the first stages of different disorders so that the doctors can start the treatment early and slow down the progress. One of the works that deals with this issue has been published by Henriquez et al. [40].

When looking into a number of significant features selected by the Mann–Whitney *U* test, it can be concluded that this number positively correlates with the classification accuracy. For example in the case of both genders the number of selected features is  $15,521 \pm 231$  (MEEI),  $11,540 \pm 419$  (PdA) and for PARCZ only  $1,750 \pm 331$ .

The ten most significant features selected by Mann–Whitney *U* test in all considered scenarios can be found in Tables 7–15. The density estimation plots (computed using kernel density estimation with Gaussian kernels) of the most significant features in these scenarios can be seen in Fig. 4. In all MEEI scenarios (M1–3) the 10 most significant features produce equivalent *p* values and they are sorted alphabetically. Regarding scenarios M1 (females) and M3 (both genders) the most discriminative features are those derived from MSC (Modulation Spectra Coefficients) while in the case of M2 (males) features based on ACW (Adaptive Component Weighted coefficients) and FADFA (Fluctuation Amplitudes of Detrended Fluctuation Analysis) are mainly selected. Looking at Fig. 4(a)–(c) it can be concluded that the relative interpercentile range of first modulation spectra coefficients and the third moment of the first adaptive component weighted cepstral coefficients show always a single value for pathological speech. Moreover, in the case of male speakers the single value also represents the healthy one. In other words, regarding the MEEI database we can use a very simple classifier (theoretically a decision tree with only one node) to differentiate between healthy and pathological speech. This fact supports our classification accuracies equal to

**Table 7**

10 most significant features selected by Mann–Whitney *U* test in scenario M1: MEEI, females (MSC – Modulation Spectra Coefficients).

Feature	<i>p</i> Value
Relative interpercentile range of 1st MSC	$2.8610 \times 10^{-30}$
Relative interpercentile range of 10th MSC	$2.8610 \times 10^{-30}$
Relative interpercentile range of 11th MSC	$2.8610 \times 10^{-30}$
Relative interpercentile range of 12th MSC	$2.8610 \times 10^{-30}$
Relative interpercentile range of 13th MSC	$2.8610 \times 10^{-30}$
Relative interpercentile range of 14th MSC	$2.8610 \times 10^{-30}$
Relative interpercentile range of 15th MSC	$2.8610 \times 10^{-30}$
Relative interpercentile range of 16th MSC	$2.8610 \times 10^{-30}$
Relative interpercentile range of 17th MSC	$2.8610 \times 10^{-30}$
Relative interpercentile range of 18th MSC	$2.8610 \times 10^{-30}$

**Table 8**

10 most significant features selected by Mann–Whitney *U* test in scenario M2: MEEI, males (ACW – Adaptive Component Weighted coefficients, FADFA – Fluctuation Amplitudes of Detrended Fluctuation Analysis).

Feature	<i>p</i> Value
3rd moment of 1st ACW	$2.5336 \times 10^{-21}$
4th moment of 1st ACW	$2.5336 \times 10^{-21}$
5th moment of 1st ACW	$2.5336 \times 10^{-21}$
6th moment of 1st ACW	$2.5336 \times 10^{-21}$
Mean absolute deviation of 1st ACW	$2.5336 \times 10^{-21}$
Mean excluding 50% outliers of 1st ACW	$2.5336 \times 10^{-21}$
Mean of 1st ACW	$2.5336 \times 10^{-21}$
Offset of linear regression of 1st ACW	$2.5336 \times 10^{-21}$
Position of max. of FADFA	$2.5336 \times 10^{-21}$
Relative position of min. of FADFA	$2.5336 \times 10^{-21}$

**Table 9**

10 most significant features selected by Mann–Whitney *U* test in scenario M3: MEEI, all (MSC – Modulation Spectra Coefficients).

Feature	<i>p</i> Value
Relative interpercentile range of 1st MSC	$1.0560 \times 10^{-49}$
Relative interpercentile range of 10th MSC	$1.0560 \times 10^{-49}$
Relative interpercentile range of 11th MSC	$1.0560 \times 10^{-49}$
Relative interpercentile range of 12th MSC	$1.0560 \times 10^{-49}$
Relative interpercentile range of 13th MSC	$1.0560 \times 10^{-49}$
Relative interpercentile range of 14th MSC	$1.0560 \times 10^{-49}$
Relative interpercentile range of 15th MSC	$1.0560 \times 10^{-49}$
Relative interpercentile range of 16th MSC	$1.0560 \times 10^{-49}$
Relative interpercentile range of 17th MSC	$1.0560 \times 10^{-49}$
Relative interpercentile range of 18th MSC	$1.0560 \times 10^{-49}$

**Table 10**

10 most significant features selected by Mann–Whitney *U* test in scenario P1: PdA, females (UCPP – Unsmooth Cepstral Peak Prominence).

Feature	<i>p</i> Value
Mode of UCPP	$1.6224 \times 10^{-19}$
Mean of UCPP	$2.9605 \times 10^{-19}$
Mean excluding 10% outliers of UCPP	$3.6871 \times 10^{-19}$
60th percentile of UCPP	$4.2979 \times 10^{-19}$
Mean excluding 20% outliers of UCPP	$4.5230 \times 10^{-19}$
Mean excluding 40% outliers of UCPP	$4.9362 \times 10^{-19}$
Median of UCPP	$4.9719 \times 10^{-19}$
Mean excluding 30% outliers of UCPP	$5.0820 \times 10^{-19}$
Mean excluding 50% outliers of UCPP	$5.0820 \times 10^{-19}$
90th percentile of UCPP	$6.6980 \times 10^{-19}$

$100.0 \pm 0.0\%$ . But the question is: Are these results trustworthy enough?

The MEEI database itself introduces many issues related to the credibility of the results. The main problem is that both the

**Table 11**

10 most significant features selected by Mann–Whitney  $U$  test in scenario P2: PdA, males (IMF-CPP – Cepstral Peak Prominence of first IMF).

Feature	$p$ Value
Median absolute deviation of IMF-CPP	$6.4974 \times 10^{-17}$
Interquartile range of IMF-CPP	$1.2600 \times 10^{-16}$
60th Percentile of IMF-CPP	$8.6433 \times 10^{-16}$
3rd Quartile of IMF-CPP	$1.1901 \times 10^{-15}$
Error of linear regression of IMF-CPP	$1.3454 \times 10^{-15}$
Median of IMF-CPP	$1.4129 \times 10^{-15}$
Mean absolute deviation of IMF-CPP	$2.0881 \times 10^{-15}$
Mean excluding 50% outliers of IMF-CPP	$2.2461 \times 10^{-15}$
70th Percentile of IMF-CPP	$2.3014 \times 10^{-15}$
Mean excluding 40% outliers of IMF-CPP	$3.1541 \times 10^{-15}$

**Table 12**

10 most significant features selected by Mann–Whitney  $U$  test in scenario P3: PdA, all (IMF-CPP – Cepstral Peak Prominence of first IMF).

Feature	$p$ Value
Error of linear regression of IMF-CPP	$6.9443 \times 10^{-32}$
Median absolute deviation of IMF-CPP	$1.3082 \times 10^{-31}$
Mean absolute deviation of IMF-CPP	$1.8834 \times 10^{-31}$
80th Percentile of IMF-CPP	$4.5625 \times 10^{-31}$
Interquartile range of IMF-CPP	$5.3469 \times 10^{-31}$
Std. of IMF-CPP	$5.5387 \times 10^{-31}$
Var. of IMF-CPP	$5.5387 \times 10^{-31}$
3rd Quartile of IMF-CPP	$5.7880 \times 10^{-31}$
70th Percentile of IMF-CPP	$7.0861 \times 10^{-31}$
Interdecile range of IMF-CPP	$7.4045 \times 10^{-31}$

**Table 13**

10 most significant features selected by Mann–Whitney  $U$  test in scenario C1: PARCZ, females (LPCC – Linear Predictive Cepstral Coefficients, GNE – Glottal-to-Noise Excitation ratio, MPSD – Median of Power Spectral Density,  $GQ_{open}$  – Qlotal Quotient (vocal folds are apart)).

Feature	$p$ Value
Coeff. of var. of 11th LPCC	$1.0832 \times 10^{-04}$
Position of max. of GNE	$1.3509 \times 10^{-04}$
Index of dispersion of 11th LPCC	$1.6228 \times 10^{-04}$
Min. of MPSD	$3.0797 \times 10^{-04}$
1st Percentile of MPSD	$3.1013 \times 10^{-04}$
Modulation of MPSD	$3.1013 \times 10^{-04}$
Relative range of MPSD	$3.1013 \times 10^{-04}$
Harmonic mean of MPSD	$3.1085 \times 10^{-04}$
Modulation of $GQ_{open}$	$4.0151 \times 10^{-04}$
Relative range of $GQ_{open}$	$4.0151 \times 10^{-04}$

**Table 14**

10 most significant features selected by Mann–Whitney  $U$  test in scenario C2: PARCZ, males (UCPP – Unsmooth Cepstral Peak Prominence, LPCT – Linear Predictive Cosine Transform coefficients, CMS – Cepstral Mean Subtraction coefficients, IMF-NSR<sub>RE</sub> – Noise-to-Signal Ratio derived from IMF based on second-order Rényi Entropy, AE (Laplacian) – Approximate Entropy based on Laplacian kernel).

Feature	$p$ Value
Slope of UCPP	$3.4500 \times 10^{-05}$
40th Percentile of 2nd LPCT	$4.2438 \times 10^{-05}$
Offset of linear regression of 5th CMS	$4.2438 \times 10^{-05}$
1st Quartile of 2nd LPCT	$5.2089 \times 10^{-05}$
Offset of linear regression of IMF-NSR <sub>RE</sub>	$6.3797 \times 10^{-05}$
Mean excluding 40% outliers of 2nd LPCT	$7.0546 \times 10^{-05}$
Mean excluding 50% outliers of 2nd LPCT	$7.0546 \times 10^{-05}$
Offset of linear regression of AE (Laplacian)	$7.0546 \times 10^{-05}$
Mean excluding 10% outliers of 2nd LPCT	$7.7966 \times 10^{-05}$
Mean excluding 20% outliers of 2nd LPCT	$7.7966 \times 10^{-05}$

**Table 15**

10 most significant features selected by Mann–Whitney  $U$  test in scenario C3: PARCZ, all (LFCC – Linear Frequency Cepstral Coefficients, MPSD – Median of Power Spectral Density, FADFA – Fluctuation Amplitudes of Detrended Fluctuation Analysis, ICC – Inferior Colliculus Coefficients).

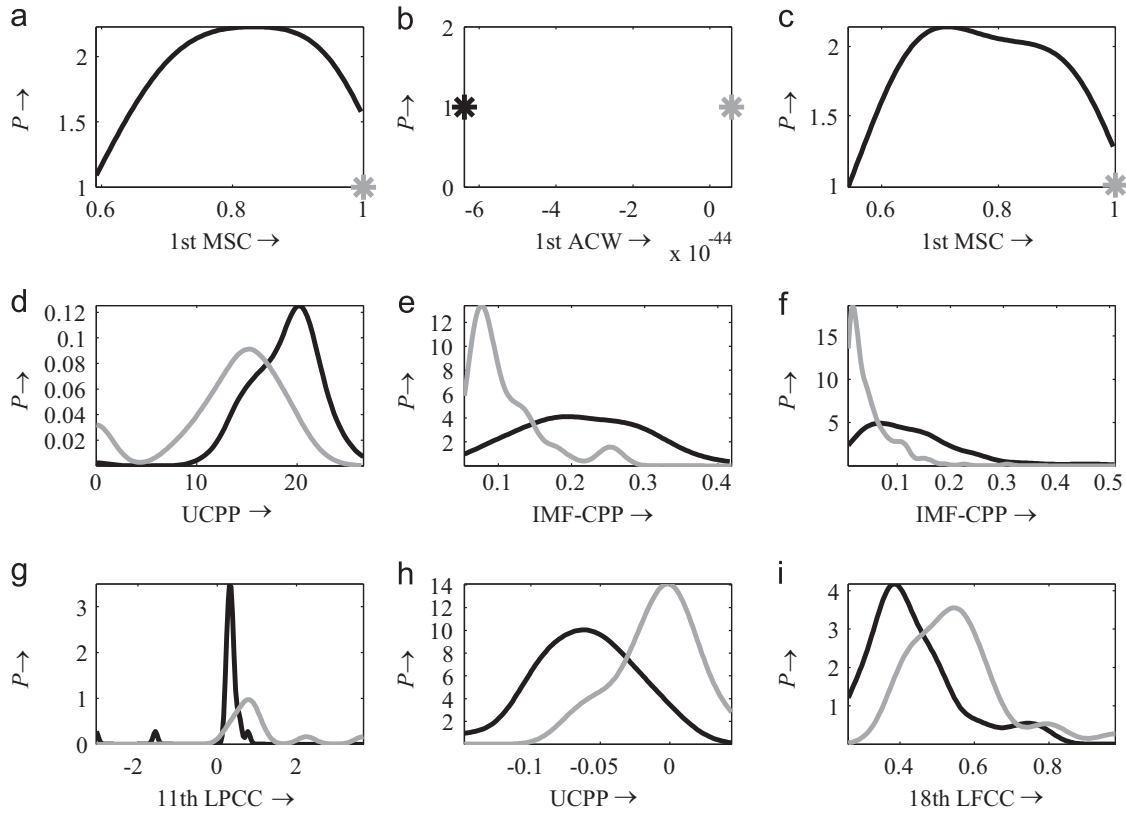
Feature	$p$ Value
Interquartile range of 18th LFCC	$1.8665 \times 10^{-05}$
Median absolute deviation of 18th LFCC	$1.9509 \times 10^{-05}$
Min. of MPSD	$1.9970 \times 10^{-05}$
Modulation of MPSD	$2.0055 \times 10^{-05}$
Relative range of MPSD	$2.0055 \times 10^{-05}$
Harmonic mean of MPSD	$2.1543 \times 10^{-05}$
40th percentile of FADFA	$3.7451 \times 10^{-05}$
Median absolute deviation of 2nd ICC	$5.7179 \times 10^{-05}$
Mean excluding 10% outliers of 2nd ICC	$5.9621 \times 10^{-05}$
Mean excluding 20% outliers of 2nd ICC	$5.9621 \times 10^{-05}$

healthy and disordered speech were recorded in a different way. First of all the healthy one was sampled at  $f_s = 50$  kHz and the disordered one at  $f_s = 25$  kHz. In our experiment all databases were resampled to  $f_s = 16$  kHz but some small differences will remain among signals. But probably the most serious problem is that the duration of a sustained vowel is 3 s and 1 s for healthy and disordered voice respectively. Therefore all high-level or local features that somehow reflect the signal length (entropy, range, std, and duration) can provide very good discriminative results. This can also be a case of the relative interpercentile range mentioned in Tables 7 and 9. Although it is a relative measure it is still dependent on the length of the input vector. We made a small experiment repetitively and randomly generating vectors with length  $N = 10^i$  for  $i = 1, 2, 3, 4$  and with a normal distribution. After that we calculated the relative interpercentile range of these vectors and found out that the value of this measure is increasing with the decreasing vector length.

To avoid the problem of different vowel durations one can skip features dependent on this property but we would lose parameters that can be potentially very good candidates for pathological speech identification. The other approach is to take only a one-second segment from the healthy speech but in our opinion this is not a good solution. The first second segment could be selected but then we are losing the sustained part of the signal (used for estimating jitter) and phonation trail. If we take only the sustained part, then we will lose information about phonation onset and offset. It has been shown that an analysis of these parts is useful for example for the dysarthric speech description [90,92]. Therefore there is no elegant solution that would not distort the results. Moreover, Malysla et al. mention that some speakers were recorded in different sites and over potentially different channels [66]. In conclusion, although the MEEI database is a very popular benchmark in the field of pathological speech analysis, the results obtained using this database should be taken very carefully. An introduction of a new English disordered voice database is thus highly important for the future evaluation of new state-of-the-art speech signal processing techniques.

In scenario P1 (PdA, males) all the most significant features are based on UCPP (Unsmooth Cepstral Peak Prominence). On the other hand, in the case of scenarios P2 (females) and P3 (both genders) all the selected features are based on IMF-CPP (Cepstral Peak Prominence of first Intrinsic Mode Function) which is originally introduced in this work. However looking at Tables 10–12 it is evident that the features listed here correlate significantly (e.g. mean and median, std and var). We have not carried out an analysis of correlation in this work due to a large number of features, but some statistics related to the most popular ones have been published, for example by Tsanas et al. [29].





**Fig. 4.** Density estimation plots (computed using kernel density estimation with Gaussian kernels) of the most significant features in scenarios M1 – 3, P1 – 3 and C1 – 3. Black color represents healthy speech and grey the pathological one. (a) Scenario M1 (MEEI, females): relative interpercentile range of 1st modulation spectra coefficients. (b) Scenario M2 (MEEI, males): 3rd moment of 1st adaptive component weighted cepstral coefficients. (c) Scenario M3 (MEEI, all): relative interpercentile range of 1st modulation spectra coefficients. (d) Scenario P1 (PdA, females): mode of unsmooth cepstral peak prominence. (e) Scenario P2 (PdA, males): median absolute deviation of cepstral peak prominence of first intrinsic mode function. (f) Scenario P3 (PdA, all): error of linear regression of cepstral peak prominence of first intrinsic mode function. (g) Scenario C1 (PARCZ, females): coefficient of variance of 11th linear predictive cepstral coefficients. (h) Scenario C2 (PARCZ, males): slope of unsmooth cepstral peak prominence. (i) Scenario C3 (PARCZ, all): interquartile range of 18th linear frequency cepstral coefficients.

In the case of scenario C1 (PARCZ, females) there were selected features based on LPCC (Linear Predictive Cepstral Coefficients), GNE (Glottal-to-Noise Excitation ratio), MPSD (Median of Power Spectral Density) and  $GQ_{open}$  (Qlottis Quotient–vocal folds are apart). Regarding scenario C2 (males), the 10 most significant features are based on UCPP, LPCT (Linear Predictive Cosine Transform coefficients), CMS (Cepstral Mean Subtraction coefficients), IMF–NSR<sub>RE</sub> (Noise-to-Signal Ratio derived from IMF based on second-order Rényi Entropy) and AE (Laplacian: Approximate Entropy based on Laplacian kernel). And finally, when considering both genders, there were selected features based on LFCC (Linear Frequency Cepstral Coefficients), MPSD, FADFA and ICC (Inferior Colliculus Coefficients). Generally  $p$  values are much lower than in the case of MEEI or PdA databases. This is also closely related to the poor classification results mentioned in Table 6.

To summarize the discussion about the databases we can say that the MEEI database should no longer be used as a benchmark. The more challenging one is PARCZ database, however it has some disadvantages from the scientific point of view: it is in Czech language, which is not a widespread language in the world; it is focused only on PD people with hypokinetic dysarthria; and it is not really suitable for a binary classification (at least 4 classes should be considered). Probably the most suitable database for the evaluation of pathological speech identification methods is PdA. The classification accuracies are still challenging and they can be significantly improved. Another advantage of this database is that it is freely available for research purposes. The only disadvantage is the limitation to Spanish language. This should not be such a big

problem in the case of sustained vowel [a], but there will certainly be cultural differences when dealing with the spoken text analysis.

Another question is, whether analysis of vowel [a] is really the best way to identify pathological speech, at least in the field of vowels analysis (not considering the other speech tasks like read/repeated/spontaneous words, and sentences). Although most of researchers automatically use sustained vowel [a], just a few publications report classification accuracies based on analysis of the other vowels. For this purpose Henriquez et al. made an experiment where they tried to identify pathological speech based on analysis of 5 Spanish vowels separately ([a], [e], [i], [o], [u]) [40]. They observed that in comparison to the other vowels the classification based on vowel [a] provides slightly better results. Probably the choice of vowel [a] is good when classifying the pathological speech generally (but still this should be proved by robust testing in future). However, as soon as we focus on a specific pathology, we can get better results when analysing another vowel. For instance Orozco-Arroyave et al. classified hypokinetic dysarthria in patients with Parkinson's disease and found out that it is more advantageous to analyze vowel [o] [78].

Our initial idea was also to try the inter-database classification (training the classifier on one database and testing it using the other one). However the classification results were very poor. It was caused mainly by these two facts: (1) When we trained the classifier on MEEI database, the features reflecting the signal length were selected. But these features were not so significant when testing them on PdA or PARCZ databases. (2) The PARCZ database is very different from MEEI or PdA, because it contains

only one specific speech disorder (hypokinetic dysarthria), while the other two databases contain many different voice pathologies.

From the feature selection point of view it is interesting to point out that many segmental parameters have been found significant. Some of them (MSC, ACW, LPCC, LPCT, CMS and ICC) were also selected as the 10 most significant ones. Although the segmental features are not very frequent when analysing pathological speech (except MFCC, MSC and ICC), their potential seems to be high. In fact, to our best knowledge, features like MFCC, LFCC, CMS or ACW were used for this purpose for the first time in the present research.

In this work we have introduced 36 new speech features. Just one (IMF-CPP) has been mentioned among the 10 most significant ones (Tables 11 and 12), however the rest of them are significant ( $p < 0.05$ ) as well, see Table 16. In the case of MEEI database we firstly checked if high-level features do not reflect the length of signal. If not we kept the feature in the table. If yes we checked the next most significant variant of the local parameter. At the top positions of this table we can see mainly features based on modulation spectra, bicepstrum and approximate entropy.

## 6. Conclusions

This work provides an insight into the robust and complex approach of pathological speech analysis. To our best knowledge this is the first contribution providing a complex evaluation of feature significance from different fields of speech signal processing (e.g. speech analysis, recognition, coding, and enhancement). It is also the first contribution deriving conclusions according to

robust tests where 3 (English, Spanish, Czech) databases were used. These languages belong to 3 different language groups (Germanic, Romanic, Slavic). In general, the work has 4 goals, yet each of them has its conclusion.

(1) *According to complex parameterization and consequent robust testing identify features that have the largest discriminative power in the field of pathological speech analysis.* Unfortunately most of the works published in the field of pathological speech analysis provides conclusions based on a limited set of parameterization methods. In other words there is still a lack of publications providing a complex overview of features quantifying pathological speech and providing strong conclusions supported by a robust testing.

Our work is unique in this way, because together it provides testing based on 128 local features. We made the wide overview of all these features used for pathological speech quantification so that the researchers can find out what exactly the specific feature quantifies and how it can be implemented. We used features describing phonation, tongue movement, speech quality (including non-linear dynamic features and features based on bispectrum/bicepstrum, empirical mode decomposition, wavelet decomposition) and segmental features.

Using the non-parametric Mann–Whitney  $U$  test we observed that among all parameterization techniques those based on segmental features provide the best classification results. To our best knowledge this is the first work that tested the significance of 11 segmental features. The largest discriminative power can be obtained thanks to especially segmental features like modulation spectra coefficients, adaptive component weighted coefficients and linear predictive cepstral coefficients. Although the segmental

**Table 16**

Best significance levels (computed using the Mann–Whitney  $U$  test) selected for all 36 features originally introduced in this work (F – female, M – male, MF – all genders).

Local feature	High-level feature	$p$ Value	Sc. ID	Dataset	Gender
IMF-CPP (Cepstral Peak Prominence extracted from the 1st IMF)	Error of linear regression	$6.9443 \times 10^{-32}$	P3	PdA	MF
MFP (Modulation Frequency of Peak)	–	$3.2291 \times 10^{-28}$	M3	MEEI	MF
RPHM (Relative Peak Height of Modulation spectra)	–	$1.5614 \times 10^{-27}$	M3	MEEI	MF
MSER (Modulation Spectra Energy Ratio)	–	$2.9477 \times 10^{-26}$	M3	MEEI	MF
BCPD (BiCepstral Phase Distance)	Harmonic mean	$4.9663 \times 10^{-19}$	M3	MEEI	MF
HFEBC (High Frequency Energy of one-dimensional BiCepstral index)	–	$2.5381 \times 10^{-15}$	M3	MEEI	MF
LFEBBC (Low Frequency Energy of one-dimensional BiCepstral index)	–	$2.5381 \times 10^{-15}$	M3	MEEI	MF
AE (triangular kernel)	Offset of linear regression	$7.7587 \times 10^{-15}$	P3	PdA	MF
BCII (BiCepstral Index Interference)	–	$9.1669 \times 10^{-15}$	P3	PdA	MF
AE (exponential kernel)	1st quartile	$1.4641 \times 10^{-14}$	P3	PdA	MF
SE (exponential kernel)	Median	$5.5444 \times 10^{-14}$	P3	PdA	MF
IMF-GNE (Glottal-to-Noise Excitation ratio based on the 1st IMF)	Mean	$6.0985 \times 10^{-14}$	M3	MEEI	MF
AE (Cauchy kernel)	Mean	$1.3344 \times 10^{-13}$	M3	MEEI	MF
AE (spherical kernel)	Offset of linear regression	$1.4599 \times 10^{-13}$	P3	PdA	MF
SE (spherical kernel)	Mean excluding 40% outliers	$2.5551 \times 10^{-13}$	P3	PdA	MF
IMF-SNR <sub>RE</sub> (based on second-order Rényi Entropy)	Mean	$3.1582 \times 10^{-12}$	P3	MEEI	MF
LCBCER (Low Cepstra/BiCepstra Energy Ratio)	Mean	$5.6242 \times 10^{-12}$	M3	MEEI	MF
IMF-FD (based on Fractal Dimension)	Median	$1.4092 \times 10^{-11}$	M3	MEEI	MF
ICER (Inferior Colliculus Energy Ratio)	–	$2.8611 \times 10^{-11}$	M3	MEEI	MF
BMD (Bispectral Module Distance)	Median	$5.3843 \times 10^{-11}$	M3	MEEI	MF
AE (circular kernel)	4th moment	$7.3659 \times 10^{-10}$	P3	PdA	MF
SE (circular kernel)	Std	$1.3186 \times 10^{-09}$	P3	PdA	MF
SE (triangular kernel)	Mean	$1.3839 \times 10^{-09}$	P2	PdA	M
BPD (Bispectral Phase Distance)	Mean	$1.0753 \times 10^{-08}$	P3	PdA	MF
IMF-NSR <sub>RE</sub> (based on second-order Rényi Entropy)	90th percentile	$1.1298 \times 10^{-08}$	P2	PdA	M
IMF-SNR <sub>ZCR</sub> (based on Zero-Crossing Rate)	Mean	$1.5834 \times 10^{-08}$	M3	MEEI	MF
HCBCEER (High Cepstra/BiCepstra Energy Ratio)	1st percentile	$4.7190 \times 10^{-08}$	P3	PdA	MF
SE (Cauchy kernel)	Harmonic mean	$5.1617 \times 10^{-08}$	M1	MEEI	F
BCPII (BiCepstrum Phase Interference Index)	–	$1.7589 \times 10^{-07}$	M1	MEEI	F
RPHIC (Relative Peak Height of Inferior Colliculus)	–	$2.2653 \times 10^{-07}$	M2	MEEI	M
BCMD (BiCepstral Module Distance)	Harmonic mean	$1.7894 \times 10^{-06}$	M3	MEEI	MF
HSBER (High Spectra/Bispectra Energy Ratio)	Mean	$3.2552 \times 10^{-06}$	M1	MEEI	F
AE (Laplacian kernel)	Offset of linear regression	$4.2554 \times 10^{-06}$	P2	PdA	M
SE (Laplacian kernel)	80th percentile	$7.5456 \times 10^{-06}$	P2	PdA	M
BCMII (BiCepstrum Module Interference Index)	–	$1.1755 \times 10^{-04}$	P3	PdA	MF
LSBER (Low Spectra/Bispectra Energy Ratio)	Mean excluding 20% outliers	$3.7662 \times 10^{-04}$	M2	MEEI	M

features are not very frequent when analysing pathological speech, their potential seems to be high. However, their disadvantage is that they are usually difficult to be interpreted clinically. This is probably the reason they are not frequently used. Although they can provide good classification results they do not say much about specific pathology or speech dysfunction.

(2) *Design new features that can quantify hoarseness, breathiness and non-linearities in pathological speech signals.* Clinical signs of vocal fold dysfunctions are usually associated with breathiness or hoarseness. Moreover voice can become aperiodic, noisy-like, and it is very difficult to find any regularities in the signal. Sometimes there is a frequent presence of sub-harmonics and chaos, which can lead to a failure of conventional techniques of speech signal analysis and which requires new parameterization methods developed specifically for pathological speech description.

We introduced 36 new measures based on modulation spectra, inferior colliculus coefficients, bicepstrum, sample and approximate entropy and empirical mode decomposition. Features based on modulation spectra quantify instability of vocal fold vibrations and complements features based on inferior colliculus coefficients that reflect the misplacement of articulators. Due to incorrect glottal closure the pathological voice contains much more white noise that can be effectively quantified by proposed features derived from bicepstrum. New features based on empirical mode decomposition are able to describe the noise component of analyzed signal as well. Finally we proposed different kernel-based approximate and sample entropies to measure regularities inside the signal.

These novel features were statistically processed by the non-parametric Mann–Whitney *U* test. All of them have been identified as significant and they have passed the feature selection process in at least one database. Moreover, some of them were listed among top ten significant features selected in specific scenario (they outperformed the other conventional features). In other words they helped to improve classification results in terms of accuracy, sensitivity and specificity and due to their effective quantification abilities they have high impact on the future work.

(3) *Prove that the proposed large set parameterization approach can provide better classification results (with respect to classification accuracy, sensitivity and specificity) than those published in the field of pathological speech analysis by the other researchers.* We tested the significance of all the mentioned parameters on 3 (English, Spanish, Czech) databases. In the case of the Massachusetts Eye and Ear Infirmary (MEEI) database we get accuracy, sensitivity and specificity equal to  $100.0 \pm 0.0\%$ , which are the best results that have been published in the frame of this database (Henriquez et al. reached  $99.69\%$  [40]). However, we are very critical with these results. Therefore we discussed their trustability and the viability of using the MEEI database as a benchmark for pathological speech signal analysis.

The results obtained with the PdA database are more challenging. When we considered the single-classifier approach, we reached the accuracy  $82.1 \pm 3.3\%$ , which is the best among the published numbers (Arias-Londono et al. published  $81.7\%$  [74]).

Regarding the last Czech Parkinsonian Speech Database (PARCZ), we obtained poor accuracy ( $67.9 \pm 6.0\%$ ), however, this is probably due to the fact that we used a binary classifier for a multi-class database. We would like to do more experiments with this database in a near future splitting the data into healthy speech and mild, moderate and severe dysarthria.

(4) *Select a database that has high potential for the future, especially in terms of speech features design, tuning and testing.* Due to some issues related to vowels' length, different sampling frequencies, and different recording conditions, the MEEI database should no longer be used as a benchmark. Results obtained using this database are not trustable. The more challenging one is PARCZ

database. Unfortunately it is focused only on parkinsonic people with hypokinetic dysarthria and it is not really suitable for a binary classification (at least 4 classes should be considered).

Therefore the most suitable database for the evaluation of pathological speech identification methods is PdA, where the classification accuracies are still challenging and they can be significantly improved. Moreover this database is freely available for research purposes.

There are many works that deal with the development of pathological voice identification methods and there is still a lot that can be improved in this field of science. However, the researchers should go further and focus not only on the pathological speech identification, but also on more sophisticated analysis that would be more helpful for doctors and that can make the treatment or diagnosis more effective. Probably the most important challenges to face in the next decade are:

1. *Identification of particular voice pathology.* Identification of pathological speech itself is not so interesting. There are many pathologies (adductor spasmodic dysphonia, erythema, and hypokinetic dysarthria) and the issue is to classify them individually. This can be very problematic, therefore we propose to do some kind of clustering and split this big set into subsets, for example according to the way they are reflected in speech (problems with tongue movement, improper work of soft palate, and disordered vocal folds).
2. *Identification of voice pathology in its first stage or estimation of its progress.* This would enable doctors to start the treatment very early and slow down the progress.

We are going to deal with these issues in future works. However this research is very dependent on good databases and especially in the case of voice pathology identification, as in its first stage, there is still a lack of suitable training data.

## Acknowledgments

This work was supported by project NT13499 (Speech, its impairment and cognitive performance in Parkinson's disease), GACR 102/12/1104, COST IC1206, LO1401, project "CEITEC, Central European Institute of Technology": (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund, FEDER and Ministerio de Economía y Competitividad TEC2012-38630-C04-03 and -04 (Kingdom of Spain). The described research was performed in laboratories supported by the SIX project; the registration number CZ.1.05/2.1.00/03.0072, the operational program Research and Development for Innovation.

## References

- [1] P.H. Dejonckere, Assessment of voice and respiratory function, in: M. Remacle, H.E. Eckel (Eds.), *Surgery of Larynx and Trachea*, Springer, Berlin, Heidelberg, 2010, pp. 11–26.
- [2] J.G. Svec, J. Sundberg, S. Hertegard, Three registers in an untrained female singer analyzed by videokymography, strobolaryngoscopy and sound spectrography, *J. Acoust. Soc. Am.* 123 (1) (2008) 347–353.
- [3] I.R. Titze, *Principles of Voice Production*, Prentice Hall, New Jersey, 1994.
- [4] J.P. Dworkin, R.J. Meleca, *Vocal Pathologies: Diagnosis, Treatment, and Case Studies*, Singular Publishing Group, San Diego, 1997.
- [5] J. Illes, Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, Huntington's, and Parkinson's, *Brain Lang.* 37 (4) (1989) 628–642.
- [6] A. Habash, C. Guinn, D. Kline, L. Patterson, Language analysis of speakers with dementia of the Alzheimer's type, *Ann. Master Sci. Comput. Sci. Inf. Syst. UNC Wilmington* 6 (1) (2012) 8–13.
- [7] K. Lopez-de Ipina, J.-B. Alonso, C.M. Travieso, J. Sole-Casals, H. Egrau, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage U.M.d. Lizardui, On the selection of non-invasive methods based on speech



- analysis oriented to automatic Alzheimer disease diagnosis, *Sensors* 13 (5) (2013) 6730–6745.
- [8] K. Horley, A. Reid, D. Burnham, Emotional prosody perception and production in dementia of the Alzheimer's type, *J. Speech Lang. Hearing Res.* 53 (5) (2010) 1132–1146.
  - [9] R.S. Bucks, S.A. Radford, Emotion processing in Alzheimer's disease, *Aging Ment. Health* 8 (3) (2004) 222–232.
  - [10] C. Gobl, A.N. Chasaide, The role of voice quality in communicating emotion mood and attitude, *Speech Commun.* 40 (1–2) (2003) 189–212.
  - [11] M. Hirano, *Psycho-Acoustic Evaluation of Voice*, Springer-Verlag, New York, 1981.
  - [12] R.J. Baken, R.F. Orlikoff, *Clinical Measurement of Speech and Voice*, Singular Thomson Learning, San Diego, 2000.
  - [13] J.R. Deller, J.G. Proakis, J.H. Hansen, *Discrete Time Processing of Speech Signals*, Prentice Hall PTR, New York, 1993.
  - [14] J. Kuo, E.B. Holmberg, R.E. Hillman, Discriminating speakers with vocal nodules using aerodynamic and acoustic features, in: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1999, pp. 77–80.
  - [15] V. Parsa, D.G. Jamieson, Identification of pathological voices using glottal noise measures, *J. Speech Lang. Hearing Res.* 43 (2) (2000) 469–485.
  - [16] P.J. Murphy, O.O. Akande, Noise estimation in voice signals using short-term cepstral analysis, *J. Acoust. Soc. Am.* 121 (3) (2007) 1679–1690.
  - [17] P. Alku, Parameterization methods of the glottal flow estimated by inverse filtering, in: *Voice Quality: Functions, Analysis and Synthesis*, 2003, pp. 81–87.
  - [18] R. Orr, B. Cranen, F.I.D. Jong, An investigation of the parameters derived from the inverse filtering of flow and microphone signals, in: *Voice Quality: Functions, Analysis and Synthesis*, 2003, pp. 35–40.
  - [19] J. Godino-Llorente, P. Gomez-Vilda, T. Lee, Analysis and signal processing of oesophageal and pathological voices, *EURASIP J. Adv. Signal Process.* 2009 (1) (2009) 1–4.
  - [20] N. Roy, J. Barkmeier-Kraemer, T. Eadie, M.P. Sivasankar, D. Mehta, D. Paul, R. Hillman, Evidence-based clinical voice assessment: a systematic review, *Am. J. Speech Lang. Pathol.* 22 (2) (2013) 212–226.
  - [21] P. Gomez-Vilda, V. Rodellar-Biarge, V. Nieto-Lluis, C. Munoz-Mulas, L. Mazaira-Fernandez, R. Martinez-Olalla, A. Alvarez-Marquina, C. Ramirez-Calvo, M. Fernandez-Fernandez, Characterizing neurological disease from voice quality biomechanical analysis, *Cognitive Comput.* 5 (4) (2013) 399–425.
  - [22] Saarbrücken voice database, ([http://www.stimmdatenbank.coli.uni-saarland.de/help\\_en.php4](http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4)), June 2014.
  - [23] Speech and language data repository, (<http://crdo.up.univ-aix.fr/>), June 2014.
  - [24] M.M. Hakkesteght, M.P. Brocaar, M.H. Wieringa, The applicability of the dysphonia severity index and the voice handicap index in evaluating effects of voice therapy and phonosurgery, *J. Voice* 24 (2) (2010) 199–205.
  - [25] Massachusetts Eye and Ear Infirmary, *Voice Disorders Database*, version 1.03, CD-ROM, Kay Elemetrics Corp., Lincoln Park, NJ, 1994.
  - [26] I.R. Titze, B.H. Story, Rules for controlling low-dimensional vocal fold models with muscle activation, *J. Acoust. Soc. Am.* 112 (3) (2002) 1064–1076.
  - [27] P. Alku, Glottal inverse filtering analysis of human voice production – a review of estimation and parameterization methods of the glottal excitation and their applications, *Sadhana* 36 (5) (2011) 623–650.
  - [28] J.I. Godino-Llorente, P. Gomez-Vilda, M. Blanco-Velasco, Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters, *IEEE Trans. Bio-Med. Eng.* 53 (10) (2006) 1943–1953.
  - [29] A. Tsanas, M. Little, P. McSharry, L. Ramig, Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests, *IEEE Trans. Bio-Med. Eng.* 57 (4) (2010) 884–893.
  - [30] A. Ghio, G. Pouchoulin, B. Teston, S. Pinto, C. Fredouille, C.D. Looze, D. Robert, F. Viallet, A. Giovanni, How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Commun.* 54 (5) (2012) 664–679.
  - [31] J. Lee, A two-stage approach using Gaussian mixture models and higher-order statistics for a classification of normal and pathological voices, *EURASIP J. Adv. Signal Process.* 2012 (252) (2012) 1–8.
  - [32] G. Vaziri, F. Almasganj, R. Behroozmand, Pathological assessment of patients' speech signals using nonlinear dynamical analysis, *Comput. Biol. Med.* 40 (1) (2010) 54–63.
  - [33] M. Markaki, Y. Stylianou, J. Arias-Londono, J. Godino-Llorente, Dysphonia detection based on modulation spectral features and cepstral coefficients, in: *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5162–5165.
  - [34] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity, *J. R. Soc. Interface* 8 (59) (2010) 842–855.
  - [35] C. Fredouille, G. Pouchoulin, A. Ghio, J. Revis, J.-F. Bonastre, A. Giovanni, Back-and-forth methodology for objective voice quality assessment: from/to expert knowledge to/from automatic classification of dysphonia, *EURASIP J. Adv. Signal Process.* 2009 (1) (2009) 1–13.
  - [36] M. Markaki, Y. Stylianou, Using modulation spectra for voice pathology detection and classification, in: *Annual International Conference of the IEEE, Engineering in Medicine and Biology Society*, 2009. EMBC 2009, pp. 2514–2517.
  - [37] M.A. Little, P.E. McSharry, S.J. Roberts, D.A.E. Costello, I.M. Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, *Biomed. Eng. Online* 6 (2007) 23.
  - [38] A. Alpan, Y. Maryn, A. Kacha, F. Grenet, J. Schoentgen, Multi-band dysperiodicity analyses of disordered connected speech, *Speech Commun.* 53 (1) (2011) 131–141.
  - [39] M. Vasilakis, Y. Stylianou, Voice pathology detection based on short-term jitter estimations in running speech, *Folia Phoniatr. Logop.* 61 (3) (2009) 153–170.
  - [40] P. Henriquez, J. Alonso, M. Ferrer, C. Travieso, J. Godino-Llorente, F. Diaz-de-Maria, Characterization of healthy and pathological voice through measures based on nonlinear dynamics, *IEEE Trans. Audio Speech* 17 (6) (2009) 1186–1195.
  - [41] D.G. Silva, L.C. Oliveira, M. Andrea, Jitter estimation algorithms for detection of pathological voices, *EURASIP J. Adv. Signal Process.* 2009 (2009) 1–9.
  - [42] A. Gelzinis, A. Verikas, M. Bacauskiene, Automated speech analysis applied to laryngeal disease categorization, *Comput. Methods Progr. Biol.* 91 (1) (2008) 36–47.
  - [43] C. Moers, B. Möbius, F. Rosanowski, E. Nöth, U. Eysholdt, T. Haderlein, Vowel- and text-based cepstral analysis of chronic hoarseness, *J. Voice* 26 (4) (2012) 416–424.
  - [44] S. Skodda, W. Visser, U. Schlegel, Short- and long-term dopaminergic effects on dysarthria in early Parkinson's disease, *J. Neural Transm.* 117 (2010) 197–205.
  - [45] M. Little, P. McSharry, E. Hunter, J. Spielman, L. Ramig, Suitability of dysphonia measurements for telemonitoring of Parkinson's disease, *IEEE Trans. Bio-Med. Eng.* 56 (4) (2009) 1015–1022.
  - [46] I. Rektorova, J. Barrett, M. Mikl, I. Rektor, T. Paus, Functional abnormalities in the primary orofacial sensorimotor cortex during speech in Parkinson's disease, *Mov. Disord.* 22 (14) (2007) 2043–2051.
  - [47] J. Shao, J.K. MacCallum, Y. Zhang, A. Sprecher, J.J. Jiang, Acoustic analysis of the tremulous voice: assessing the utility of the correlation dimension and perturbation parameters, *J. Commun. Disord.* 43 (2010) 35–44.
  - [48] D. Dimitriadis, A. Potamianos, P. Maragos, A comparison of the squared energy and Teager-Kaiser operators for short-term energy estimation in additive noise, *IEEE Trans. Signal Process.* 57 (7) (2009) 2569–2581.
  - [49] T.H. Falk, W.-Y. Chan, F. Shein, Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility, *Speech Commun.* 54 (5) (2012) 622–631.
  - [50] M. Gonzalez-Izal, I. Rodriguez-Carreño, A. Malanda, F. Mallor-Gimenez, I. Navarro-Amezqueta, E. Gorostiaga, M. Izquierdo, sEMG wavelet-based indices predicts muscle power loss during dynamic contractions, *J. Electromyogr. Kinesiol.* 20 (6) (2010) 1097–1106.
  - [51] Y. Song, W.-H. Wang, F.-J. Guo, Feature extraction and classification for audio information in news video, in: *International Conference on Wavelet Analysis and Pattern Recognition*, 2009, ICWAPR 2009, 2009, pp. 43–46.
  - [52] G. Weismer, J.Y. Jeng, J.S. Laures, R.D. Kent, J.F. Kent, Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders, *Folia Phoniatr. Logop.* 53 (1) (2001) 1–18.
  - [53] J. Mekyska, I. Rektorova, Z. Smekal, Selection of optimal parameters for automatic analysis of speech disorders in Parkinson's disease, in: *2011 34th International Conference on Telecommunications and Signal Processing (TSP)*, 2011, pp. 408–412.
  - [54] J.R.O. Arroyave, S.M. Rendon, A.M. Alvarez-Meza, J.D. Arias-Londono, E. Delgado-Trejos, J.F.V. Bonilla, C.G. Castellanos-Dominguez, Automatic selection of acoustic and non-linear dynamic features in voice signals for hypernasality detection, in: *INTERSPEECH'11*, 2011, pp. 529–532.
  - [55] J.B. Alonso, J. de Leon, I. Alonso, M.A. Ferrer, Automatic detection of pathologies in the voice by hos based parameters, *EURASIP J. Adv. Signal Process.* 2001 (4) (2001) 275–284.
  - [56] S.K. Banchhor, Discrimination between speech and music signal, *Int. J. Soft Comput. Eng.* 2 (3) (2012) 28–31.
  - [57] J. Hillenbrand, R.A. Houde, Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech, *J. Speech Hearing Res.* 39 (2) (1996) 311–321.
  - [58] D. Michaelis, T. Grams, H.W. Strube, Glottal-to-noise excitation ratio – a new measure for describing pathological voices, *Acta Acust. United Acust.* 83 (4) (1997) 700–706.
  - [59] D.D. Deliyski, Acoustic model and evaluation of pathological voice production, in: *The Third Conference on Speech Communication and Technology EURO-SPEECH'93*, 1993, pp. 1969–1972.
  - [60] H. Kasuya, S. Ogawa, K. Mashima, S. Ebihara, Normalized noise energy as an acoustic measure to evaluate pathologic voice, *J. Acoust. Soc. Am.* 80 (5) (1986) 1329–1334.
  - [61] J. Makhoul, L. Cosell, LPCW: An LPC vocoder with linear predictive spectral warping, in: *IEEE International Conference on ICASSP '76 Acoustics, Speech, and Signal Processing*, vol. 1, 1976, pp. 466–469.
  - [62] H. Atassi, A. Esposito, Z. Smekal, Analysis of high-level features for vocal emotion recognition, in: *2011 34th International Conference on Telecommunications and Signal Processing (TSP)*, 2011, pp. 361–366.
  - [63] L. Atlas, S.A. Shamma, Joint acoustic and modulation frequency, *EURASIP J. Adv. Signal Process.* 2003 (7) (2003) 668–675.
  - [64] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, *J. Acoust. Soc. Am.* 87 (4) (1990) 1738–1752.
  - [65] R. Mammone, X. Zhang, R. Ramachandran, Robust speaker recognition: a feature-based approach, *IEEE Signal Process. Mag.* 13 (5) (1996) 58–71.
  - [66] N. Malyska, T. Quatieri, D. Sturim, Automatic dysphonia recognition using biologically-inspired amplitude-modulation features, in: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, (ICASSP'05), vol. 1, 2005, pp. 873–876.
  - [67] P. Hosseini, F. Almasganj, T. Emami, R. Behroozmand, S. Gharibzade, F. Torabinezhad, Local discriminant wavelet packet basis for voice pathology classification, in: *The Second International Conference on Bioinformatics and Biomedical Engineering*, 2008, ICBBE 2008, pp. 2052–2055.



- [68] R. Esteller, G. Vachtsevanos, J. Echaz, B. Litt, A comparison of waveform fractal dimension algorithms, *IEEE Trans. Circuits Syst. I: Fundam. Theory Appl.* 48 (2) (2011) 177–183.
- [69] M. Aboy, R. Hornero, D. Abasolo, D. Alvarez, Interpretation of the Lempel-Ziv complexity measure in the context of biomedical signal analysis, *IEEE Trans. Bio-Med. Eng.* 53 (11) (2006) 2282–2288.
- [70] A.W. Jayawardena, P. Xu, W.K. Li, Modified correlation entropy estimation for a noisy chaotic time series, *Chaos* 20 (2) (2010) 1–11.
- [71] H.K. Heris, B.S. Aghazadeh, M. Nikkhah-Bahrami, Optimal feature selection for the assessment of vocal fold disorders, *Comput. Biol. Med.* 39 (10) (2009) 860–868.
- [72] J.M. Yentes, N. Hunt, K.K. Schmid, J.P. Kaipust, D. McGrath, N. Stergiou, The appropriate use of approximate entropy and sample entropy with short data sets, *Ann. Biomed. Eng.* 41 (2) (2013) 349–365.
- [73] J.I. Godino-Llorente, P. Gomez-Vilda, F. Cruz-Roldan, M. Blanco-Velasco, R. Fraile, Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness, *J. Voice* 24 (6) (2010) 667–677.
- [74] J.D. Arias-Londono, J.I. Godino-Llorente, M. Markaki, Y. Stylianou, On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices, *Logop. Phoniatr. Vocol.* 36 (2) (2011) 60–69.
- [75] M.G. Kang, K.-T. Lay, A.K. Katsaggelos, Phase estimation using the bispectrum and its application to image restoration, *Opt. Eng.* 30 (7) (1991) 976–985.
- [76] W.-t. Chen, Z.-z. Wang, X.-m. Ren, Characterization of surface EMG signals using improved approximate entropy, *J. Zhejiang Univ. Sci. B* 7 (10) (2006) 844–848.
- [77] F. Takens, Detecting strange attractors in turbulence, in: *Dynamical Systems and Turbulence*, Warwick 1980, Lecture Notes in Mathematics, vol. 898, Springer, Berlin, Heidelberg, 1981, pp. 366–381.
- [78] J.R.O. Arroyave, J.D. Arias-Londono, J.F.V. Bonilla, E. Noth, Analysis of speech from people with Parkinson's disease through nonlinear dynamics, in: T. Drugman, T. Dutoit (Eds.), *Advances in Nonlinear Speech Processing*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2013, pp. 112–119.
- [79] G. Fairbanks, *Voice and Articulation Drillbook*, 2nd edition, Harper and Row, New York, 1960.
- [80] A. Dibazar, S. Narayanan, T. Berger, Feature analysis for automatic detection of pathological speech, in: 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference on Engineering in Medicine and Biology, 2002, Proceedings of the Second Joint, 2002, vol. 1, 2002, pp. 182–183.
- [81] A. Alpan, J. Schoentgen, Y. Maryn, F. Grenez, Automatic perceptual categorization of disordered connected speech, in: *INTERSPEECH*, 2010, pp. 2574–2577.
- [82] M. Hariharan, M.P. Paulraj, S. Yaacob, Time-domain features and probabilistic neural network for the detection of vocal fold pathology, *Malays. J. Comput. Sci.* 23 (1) (2010) 60–67.
- [83] J. Arias-Londono, J. Godino-Llorente, N. Saenz-Lechon, V. Osmar-Ruiz, G. Castellanos-Dominguez, Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients, *IEEE Trans. Bio-Med. Eng.* 58 (2) (2011) 370–379.
- [84] J. Mekyska, Z. Smekal, M. Kostalova, M. Mrackova, S. Skutilova, I. Rektorova, Motor aspects of speech impairment in Parkinson's disease and their assessment, *Ces. Slov. Neurol. N* 74 (6) (2011) 662–668.
- [85] G.R. Doddington, M.A. Przybicki, A.F. Martin, D.A. Reynolds, The NIST speaker recognition evaluation—overview, methodology systems results perspective, *Speech Commun.* 31 (2–3) (2000) 225–254.
- [86] M. Little, P. McSharry, I. Moroz, S. Roberts, Nonlinear, biophysically-informed speech pathology detection, in: 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings, vol. 2, 2006, pp. II-1080–II-1083.
- [87] M. Brookes, Voicebox: Speech processing toolbox for matlab, (<http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>), October 2011.
- [88] Tstool version 1.2, (<http://www.physik3.gwdg.de/tstool/>), February 2009.
- [89] P. Boersma, D. Weenink, Praat: doing phonetics by computer, (<http://www.fon.hum.uva.nl/ptraat/>), May 2013.
- [90] I. Eliasova, J. Mekyska, M. Kostalova, R. Marecek, Z. Smekal, I. Rektorova, Acoustic evaluation of short-term effects of repetitive transcranial magnetic stimulation on motor aspects of speech in Parkinson's disease, *J. Neural Transm.* 120 (4) (2013) 597–605.
- [91] A.-C. Haury, P. Gestraud, J.-P. Vert, The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures, *PLoS ONE* 6 (12) (2011) 1–12.
- [92] A.M. Goberman, M. Blomgren, Fundamental frequency change during offset and onset of voicing in individuals with Parkinson disease, *J. Voice* 22 (2) (2008) 178–191.



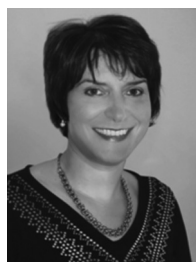
**Eva Janousova** was born in Brno, Czech Republic. She graduated at Masaryk University (MU) in Brno in 2010 with M.Sc. degree with honors in Computational Biology. Currently, she is a Ph.D. student in Neuroscience at the Faculty of Medicine, MU. She has been working as a data analyst at the Institute of Biostatistics and Analyses, MU, since 2009; and teaching courses on biostatistics and data analysis at MU since 2012. Her research interests are focused on medical data analysis, image data classification and analysis, and pattern recognition.



**Pedro Gomez-Vilda** was born in Burgo de Osma, Spain. He received the M.Sc. degree in Communications Engineering in 1978 and the Ph.D. degree in Computer Science from the Universidad Politecnica de Madrid, Madrid, Spain, in 1983. He is a Professor in the Computer Science and Engineering Department, at Universidad Politecnica de Madrid, since 1988. His current research interests are biomedical signal processing, speaker identification, robust speech recognition, and genetic signal processing. He is a member of the IEEE, ISCA and EURASIP and the LSSA-TC Biochip Group.



**Zdenek Smekal** is a full-time Professor at the Faculty of Electrical Engineering and Communication, Brno University of Technology. Since 1998 he has been a member of IEEE (Signal Processing, and Communications Societies), since 2006 a Senior Member. His research interests are discrete and digital processing of one- and multi-dimensional signals with application in digital signal processors, emotional speech and image analysis and recognition, and also biometrics system identification. He is an author of three monographs about digital signal processing and digital signal processors.



**Irena Rektorova** is a Professor of Neurology at the Department of Neurology, School of Medicine, Masaryk University and St. Anne's Hospital, Brno, Czech Republic. She is a Secretary of the International Association of Parkinsonism and Related Disorders. She has been appointed as the Head of the Centre for Cognitive Impairment of the Department of Neurology and Head of the Applied Neuroscience Research Group, Central European Institute of Technology, CEITEC MU, Brno. She has published over 100 papers in peer-reviewed journals and has over 570 SCI citations. Her clinical and research interests are in movement disorders, cognitive impairment and dementia, brain imaging and transcranial magnetic stimulation.



**Ilona Eliasova** is an independently working Consultant and Physician for outpatients at the Department of Neurology, School of Medicine, Masaryk University and St. Anne's Hospital, Brno, Czech Republic. She is also a member of the Centre for Cognitive Impairment at the Department of Neurology and a member of the Applied Neuroscience Research Group, Central European Institute of Technology, CEITEC MU, Brno. She is a Ph.D. student in Neurology at the Faculty of Medicine with particular interests in transcranial magnetic stimulation, motor aspects of speech in Parkinson's disease and cognitive impairment.



**Milena Kostalova** works as a Speech and Language Therapist and as an Instructor at the Department of Neurology, Faculty of Medicine, Masaryk University in Brno, Czech Republic. She is a Ph.D. student in Neuroscience at the Faculty of Medicine, MU. She is a Lecturer teaching health care courses in postgraduate education. Her clinical and research interests lie in aphasiology and development of tools of speech, language and related disorders diagnosis.



**Jiri Mekyska** was born in Holesov, Czech Republic. He received the Ph.D. degree from Telecommunications, in 2014, at BUT (Brno University of Technology). Currently he is a researcher and member of SPlab (Signal Processing Laboratory) at the Department of Telecommunications at BUT. His research interest lies in the field of speech signal analysis and its application in biomedicine. He deals with the hand-written text analysis, thermal image processing and information theory as well.



**Martina Mrackova** graduated at Masaryk University at the Faculty of Medicine in 2009. She has been working as a Physician at the Department of Neurology, St. Anne's Hospital, Brno, Czech Republic. Currently she is a Ph.D. student in Neurology, Junior Researcher and member of Applied Neuroscience group, Central European Institute of Technology, CEITEC MU, Brno. Her clinical and research interests are focused on movement disorders, cognitive impairment and functional Magnetic Resonance Imaging.



**Jesus B. Alonso-Hernandez** received the Telecommunication Engineer degree in 2001 and the Ph.D. degree in 2006 from University of Las Palmas de Gran Canaria (ULPGC-Spain) where he has been an Associate Professor in the Department of Signal and Communications since 2002. He has researched in different International and Spanish Research Projects. He has published numerous papers in international journals and international conferences. He has been a reviewer in different international journals and conferences since 2003. His research interests include signal processing in biocomputing, biometrics, nonlinear signal processing, recognition systems, audio characterization and data mining. He was a guest editor of special issues for Springer in Cognitive Computation and for Elsevier Neurocomputing. He was the Head of excellent network in biomedical engineering in ULPGC. He was the Vice-Dean from 2009 to 2013 in School of Telecommunications Engineering in ULPGC.



**Marcos Faundez-Zanuy** was born in Barcelona, Spain. He received the B.Sc. degree in Telecommunication in 1993 and the Ph.D. degree in 1998, both from the Polytechnic University of Catalunya.

He has been a full Professor at EUP Mataro since 2010 and heads the Signal Processing Group there. Since 2009 he has been the Dean of Escola Universitaria Politècnica de Mataro (Polytechnic University of Catalunya) and is responsible of Tecnocampus Research. His research interests lie in the fields of biometrics and speech coding. He is the author of more than 30 papers indexed in ISI Journal citation report, more than 100 conference papers, around 10 books, and responsible for 10 National and European research projects.



**Karmele López-de-Ipiña** received the Ph.D. degree in computer science in 2003 and the B.Sc. degree in physics in 1990, at the Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU). She is an IEEE member. She worked for enterprises until 1995 when she joined the Department of Systems Engineering and Automation of the University of the Basque Country. She was the Director at the UPV/EHU (2004–2009). She is currently the Head of the Engineering and Society research group. Her research interests are in bioengineering and biomedical engineering, pattern recognition, signal processing, ambient intelligence and robotics.