Parkinson's disease detection using voice signal decomposition

Evaldas Vaiciukynas[a,b,*], Antanas Verikas[a,c], Adas Gelzinis[a], Marija Bacauskiene[a], Aivaras Simulis[b]

[a] *Department of Electrical Power Systems,*
Kaunas University of Technology, Studentu 50, LT-51368 Kaunas, Lithuania
[b] *Department of Information Systems,*
Kaunas University of Technology, Studentu 50, LT-51368 Kaunas, Lithuania
[c] *Centre for Applied Intelligent Systems Research,*
Halmstad University, Kristian IV:s väg 3, PO Box 823, S-301 18 Halmstad, Sweden

Abstract

Accurate detection of Parkinson's disease by acoustic analysis of sustained phonation is the goal of this research. Phonation corresponds to the vowel /a/ voicing task and was recorded through two channels simultaneously, namely, acoustic cardioid (AC) and smart phone (SP) microphones. Signal decomposition into intrinsic mode functions (IMFs) is explored in a novel way to create an expert system for medical decision support. Decomposition approaches considered are empirical mode decomposition (EMD) and variational mode decomposition (VMD). Several frequency signatures, – Bark frequency cepstral coefficients (BFCC), Mel frequency cepstral coefficients (MFCC) and perceptual linear predictive cepstral coefficients (PLPCC), – are tested as audio descriptors to characterize each extracted IMF as well as the original signal. Random forest (RF) classifier is used both as a base-learner and as a meta-learner for the decision-level fusion. Baseline solution by summarizing cepstral coefficients from all frames of a voice recording through statistical functionals was compared to the proposed solution of using EMD and VMD of a single center frame or three evenly-spaced frames and summarizing base-learner's decisions instead. Experiments indicate that the voice signal decomposition followed by the decision-level fusion is capable to improve the detection performance and achieve the out-of-bag equal error rate of ∼1% for AC and ∼12% for the SP channel.

Keywords: Parkinson's disease, Voice analysis, Empirical mode decomposition, Variational mode decomposition, Random forest, Medical decision support

1. Introduction

Parkison's disease (PD) is the second most common neurodegenerative disease after Alzheimer's [1] and the prevalence of PD is expected to increase due to population ageing.

[*] Corresponding author. Tel.: +370-37-453078

Email addresses: evaldas.vaiciukynas@ktu.lt (Evaldas Vaiciukynas), antanas.verikas@hh.se (Antanas Verikas), adas.gelzinis@ktu.lt (Adas Gelzinis), marija.bacauskiene@ktu.lt (Marija Bacauskiene), aivaras.simulis@outlook.com (Aivaras Simulis)

The huge loss, amounting at up to half of all dopaminergic neurons, can be witnessed at the time of clinical diagnosis [2] with the rapid increase in the number of dead neurons happening until the fourth year after diagnosis [3]. Future neuroprotective strategies could be too late to effectively slow down this neurodegenerative process. Therefore, early objective diagnostic markers are critically needed and automated acoustic analysis can be considered as a potential and convenient non-invasive tool in PD screening.

Recent computational and digital advancements have made it possible to explore such ambitious concepts as smart homes or personalized medicine and to bring ambient intelligence to our day-to-day environment [4]. Ambient intelligence could provide low-cost healthcare monitoring in an unobtrusive way and enhance the healthcare domain, especially in the preventive healthcare direction. Usage of hand-held devices, such as smart-phones, for non-invasive measurements is getting popular among researchers. The most notable examples in this direction with respect to PD screening are Johns Hopkins [5] and the mPower [6] studies.

The empirical mode decomposition (EMD) has been used to detect PD from brain images in [7], which entails an invasive approach. EMD-based non-invasive approaches assessed tremor by using gyroscope sensors around a joint in [8] or by accelerometers attached to a finger in [9]. Application of EMD to a voice signal is reported in [10], where it is concluded that the EMD-based features can significantly improve PD detection. Those EMD-based features are among a set of various dysphonia measures in the `Voice Analysis Toolbox` [11], introduced by [12].

The main emphasis of the related work remains on the extraction of various audio features to characterize a voice or speech signal. Some researchers use large sets of audio features with an aim to comprehensively characterize recordings [13], including the renowned cepstral coefficients such as MFCCs or PLPCCs, while others adopt only "clinically useful" measures or apply feature selection [12] to derive a compact set of audio descriptors. For a comprehensive review of the related work see [14].

In our previous work [15] several well-known collections of audio descriptors were tested for the PD detection task, resulting in the equal error rate of 19.3% for the AC and 23% for the SP channel after the RF-based decision-level fusion of the phonation and speech modalities. Application of the convolutional neural network on segments of speech recording in [16] resulted in the equal error rate of 14.1% for the AC channel. In the current study we consider only phonation recordings. Noting that a short-term frame from the voice signal exhibits rather stationary properties, we've performed a deeper analysis by extracting IMFs of a frame through the empirical mode decomposition (EMD) and the variational mode decomposition (VMD). The spectral chracteristics of the extracted modes are summarized by cepstral coefficients to build a base-learner. Base-learner's decisions for a single recording are compressed into a fixed-length meta-vector through simple statistics and a meta-learner is built, resulting in an excellent PD detection performance.

2. Voice database

Recordings were done in a sound-proof booth at Lithuanian University of Health Sciences using two channels simultaneously – an acoustic cardioid (AKG Perception 220 with frequency range 20 – 20000 Hz) and a smart phone (an internal microphone of Samsung Galaxy Note 3). Both microphones were located at ∼10 cm distance from the lips. The phonation task corresponded to a sustained voicing of vowel $/a/$ vocalized at a comfortable pitch and the loudness level for at least 5 s. The task was repeated 3 times, resulting in ∼3 recordings per subject. The format of recording was mono PCM wav with resolution of 16 bits at a 44.1 kHz sampling rate. A mixed gender database, identical to the one used in [15], contains 99 subjects with both AC and SP recordings, see Table 1.

Table 1: Summary of the database. Note: numbers correspond to the count of subjects (recordings).

Microphone channel	**Healthy control subjects**			**Parkinson's disease patients**			**Overall**
	Male	Female	Total	Male	Female	Total	
AC - acoustic cardioid	11 (33)	24 (72)	35 (105)	30 (89)	34 (101)	64 (190)	99 (295)
SP - smart phone	11 (33)	24 (72)	35 (105)	30 (90)	34 (102)	64 (192)	99 (297)

3. Methodology

The baseline solution uses all short-term frames, characterized by the cepstral coefficients, and summarizes them by statistical functionals. The proposed solution extracts a single central frame and, after the mode decomposition, characterizes the spectral content of the extracted components and the original signal as well by the cepstral coefficients and builds a base-learner. In case of several frames, sampled at specific locations, separate base-learners are built. Then decisions from a base-learner for each recording are compressed by the same statistical functionals as in the baseline case, to obtain a feature vector for the meta-learner.

The proposed solution can be summarized in the following steps:

1. Extract a short-term frame from the sustained phonation recording and filter with the Hanning window, this results in fade-in and fade-out for the cut-out frame.
2. Characterize an original signal segment by cepstral coefficients.
3. Perform signal decomposition using the EMD and VMD approaches.
4. Characterize each IMF, after the EMD (or VMD), by cepstral coefficients.
5. Concatenate cepstral coefficients (original with IMF) to create a feature vector.
6. Build a base-RF for EMD and VMD using the feature vectors of all IMFs and recordings.

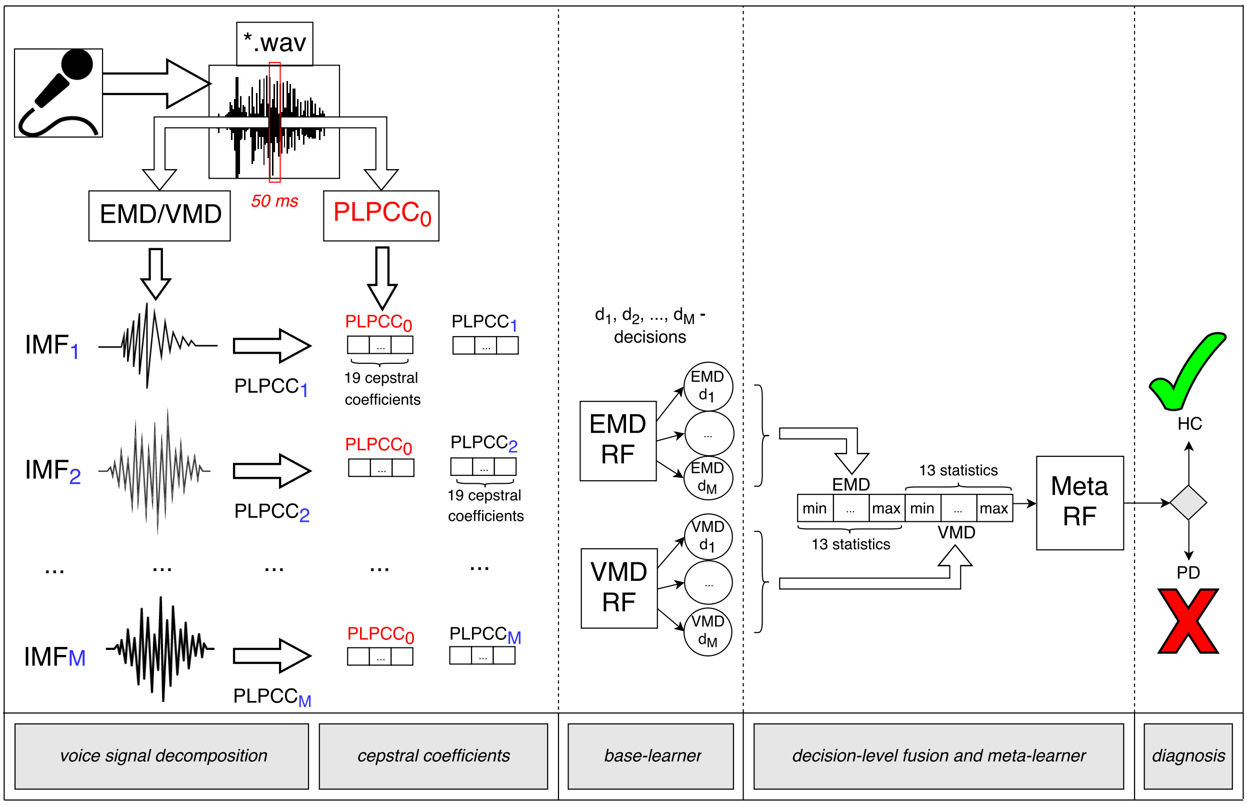

Figure 1: The proposed decomposition-based solution: a uni-frame variant with a central window of sustained phonation extracted for the decomposition and calculation of the perceptual linear predictive cepstral coefficients. Statistical functionals compress M number of base-learner's decisions into meta-feature vector.

7. Obtain a base-RF score for each IMF using the class probabilities from the out-of-bag votes.

8. Summarize the base-RF scores for all IMFs of a single recording through the statistical functionals as meta-features.

9. Build a meta-RF after concatenating the meta-features from the EMD-RF and VMD-RF.

3.1. Feature extraction

Information, contained in an audio recording of a voice signal, can be extracted using various signal analysis methods. Resulting measures are commonly known as audio features. Depending on the amount of signal used the features can be categorized into:

- global features: long-term or recording-based (high-level descriptors);
- local features: short-term or frame-based (low-level descriptors).

The short-term parametrization is performed by dividing a recording into short and usually overlapping segments (frames or windows) and applying an algorithm that computes a respective local feature for each segment. In the baseline solution, local features are compressed into global features by computing various statistical functionals. In the proposed approach, local features are used in the base-learner directly and all recording-wise decisions are summarized into meta-features by statistical functionals.

The statistical functionals in this work encompass the following 13 characteristics: minimum, maximum, mean, median, lower quartile (Q_{lo}), upper quartile (Q_{hi}), trimean ($\frac{2 \cdot median + Q_{lo} + Q_{hi}}{4}$), standard deviation, inter-quartile range ($Q_{hi} - Q_{hi}$), lower range ($median - Q_{lo}$), upper range ($Q_{hi} - median$), skewness, and kurtosis.

3.1.1. Signal decomposition

Empirical mode decomposition (EMD) is an adaptive technique for signal analysis, which decomposes a signal into additive components – functions of time, known as intrinsic mode functions (IMFs). Each IMF can be viewed as a sub-band of the signal and represents fast to slow oscillations. The main iterative process behind EMD is known as sifting. The sifting process is continued until the final residue is a constant, a monotonic function or a function with one maxima and one minima from which no more IMFs can be derived.

After the sifting process, the unidimensional signal $x(t)$ is represented as

$$x(t) = \sum_{c=1}^{M} IMF_c(t) + r_M(t) \tag{1}$$

where M is the number of IMFs and $r_M(t)$ is the final residue.

The EMD algorithm variant used in this work is an improved complete ensemble EMD with adaptive noise (CEEMDAN) [17], which is capable of automatically finding an optimal number of IMFs. The number of IMFs set for the variational mode decomposition (VMD) [18] algorithm was equal to the extracted number of IMFs in EMD. The VMD decomposes the signal into additive modes using the calculus of variation. Each mode of the signal is assumed to have a compact frequency support around a central frequency. The VMD tries to find out these central frequencies and IMFs centered on the frequencies concurrently using optimization by the alternate direction method of multipliers (ADMM13).

Parameters for EMD were set as follows: noise standard deviation $Nstd = 0.1$, number of realizations $NR = 100$, maximum number of sifting iterations allowed $MaxIter = 1000$, increasing signal-to-noise ratio setting $SNRFlag = 1$. Parameters for VMD were set as follows: the balancing parameter of the data-fidelity constraint $alpha = 1000$, time-step of the dual ascent $tau = 0.1$, the first mode is kept at zero frequency $DC = 1$, all omegas start uniformly distributed $init = 1$, tolerance of convergence criterion $tol = 0.0000001$.

3.1.2. Cepstral coefficients

Mel-frequency cepstral coefficients (MFCCs) is one of the most popular sets of audio features to characterize a voice signal. Difference between MFCCs and the Bark-frequency cepstral coefficients (BFCCs) is only in the type of the filter-bank applied. The signal is windowed in the time domain into 50 ms length frames and converted into the frequency domain by the non-parametric discrete fast Fourier transform (FFT), which gives the amount of energy present within a particular frequency range for each of 4096 bins. The full frequency range in our experiments was constrained to the 0 – 8000 Hz range. Then 27 triangular Mel-frequency or trapezoid-like Bark-frequency filters are applied to compress the frequency-based information by summing the filtered FFT bin values, to get the filter-bank outputs (or energies). Application of the Mel- or Bark-scaling provides higher resolution at low frequencies and lower resolution at high frequencies, which is motivated the human perception, where a relationship between the real frequency scale (Hz) and the perceptual frequency scale (Mel or Bark) is logarithmic above 1000 Hz and linear below. Finally, MFCCs are obtained by applying the discrete cosine transform (DCT of type 2) to the logarithm of the Mel or Bark filter-bank outputs. The DCT represents the signal in terms of the first basis function (constant component) and the remaining basis functions (components of successively increasing frequency), which are uncorrelated. In this work, first 19 components after the DCT represent a compacted MFCC or BFCC vector of the corresponding frame. The algorithm for extracting MFCCs is illustrated by Fig. 2.

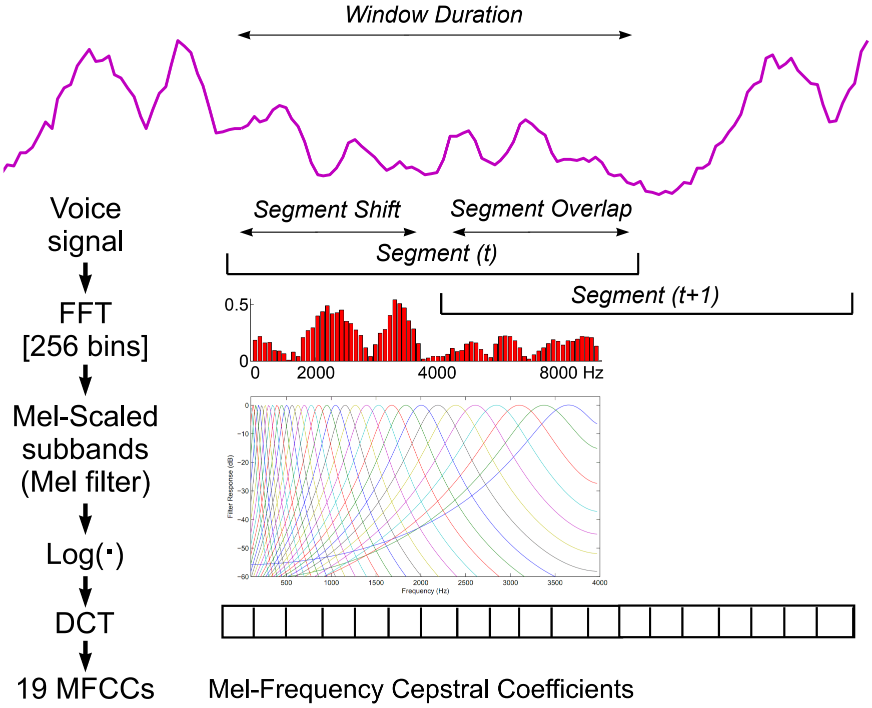

Figure 2: Extraction of the Mel-frequency cepstral coefficients from a voice signal.

The perceptual linear predictive analysis (PLP), introduced in [19], combines spectral analysis with linear prediction analysis and contains the following steps:

1. Signal windowing. Speech samples are weighted using 50 ms Hamming window (Eq. 1 in [19]) with 25 ms overlap. Due to the 44.1 kHz sampling rate, a 50 ms frame corresponds to approximately 2200 samples.

2. Power spectrum. After windowing, each obtained frame is transformed into the frequency domain by the Cooley-Tukey variant [20] of the FFT using 512 bins. The real and imaginary components of the FFT result are squared and added to get the power spectrum (Eq. 2 in [19]).

3. Bark spectrum. The power spectrum is warped into the Bark scale by (Eq. 3 in [19]):

$$F_{Bark} = 6 \cdot sinh^{-1}(\frac{F_{Hz}}{600}) = 6 \cdot ln(\frac{F_{Hz}}{600} + \sqrt{(\frac{F_{Hz}}{600})^2 + 1}) \qquad (2)$$

   where F_{Bark} is frequency in Barks, F_{Hz} is frequency in Hertz, $sinh^{-1}$ denotes the inverse function of the hyperbolic sine, and ln means the natural logarithm.

4. Critical-band spectral resolution. Energy in the FFT bins is collected through the 1-Bark wide overlapping triangle filter-banks, equally spaced by a 1-Bark interval. The triangles appear increasingly spaced in the Hertz scale, therefore, more filters are allocated for lower frequencies where human hearing is considered more sensitive. Each triangle of the filter-bank was additionally convolved with a simulated critical-band masking curve (Eq. 4 in [19]). After this convolution (Eq. 5 in [19]) filters become trapezoid-like with flat tops, covering an increasing bandwidth and significantly reducing the spectral resolution.

5. Loudness equalization. The equal loudness pre-emphasis (Eq. 7 in [19]) weights the filter-bank to approximate the sensitivity of the human ear to certain frequencies at roughly 40 dB. Frequencies in the area of human speech are among those that the ear shows heightened sensitivity.

6. Intensity-loudness conversion. Further modification to approximate human hearing is implemented by the amplitude compression, also known as the cubic root compression, since it simply corresponds to taking a cubic root from the result of each filter (Eq. 8 in [19]). The resulting spectrum could be regarded as perceived loudness, measured in Son units.

7. Autoregressive modeling. In the final step, the perceived loudness spectrum is approximated by the spectrum of all-pole spectral modeling, which can be summarized as follows: the inverse discreet Fourier transform is applied to the perceived loudness spectrum (of Step 6); from the resulting autocorrelation function the first $M + 1$ values are used to solve the Yule-Walker equations for the autoregressive coefficients of the Mth-order all-pole model, we used $M = 14$ in our experiments. The autoregressive coefficients are further transformed into the cepstral coefficients.

The psychophysical aspects of human hearing, which was the main motivation behind the PLPCC, encompass Steps 3–6. The Matlab code to extract MFCC, BFCC and PLPCC features was obtained from [21]. Signal pre-emphasis and post-processing of coefficients by liftering was not used.

3.2. Machine learning

With respect to the statistical learning framework we consider a learning set $\mathbf{Z} = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$ consisting of n observations, where each observation is represented by a vector $\mathbf{x} = (x^1, \ldots, x^p)$ of p features (also referred to as explanatory variables or predictors), say $\mathbf{x} \in \mathbb{R}^p$, and a corresponding class label $y \in \mathcal{Y}$, where $\mathcal{Y}$ denotes a set of possible class labels. The aim of a machine learning model is to obtain the mapping function between $\mathbf{x}$ and y, which is not only as accurate as possible, but also performs well on unseen data coming from the same distribution.

3.2.1. Random forest classifier

RF is a committee of many (B in total) unprunned CART (classification and regression tree) models (see Fig. 3), built on different bootstrap samples of the original dataset X and a random subset (of predetermined size q) of features $x^1, \ldots, x^p$. For our experiments B was set to 10000 and after testing several specific values of q ($\sqrt{p}$, $2 \cdot \sqrt{p}$, $\frac{1}{2} \cdot p$), the value providing the best out-of-bag performance was chosen.

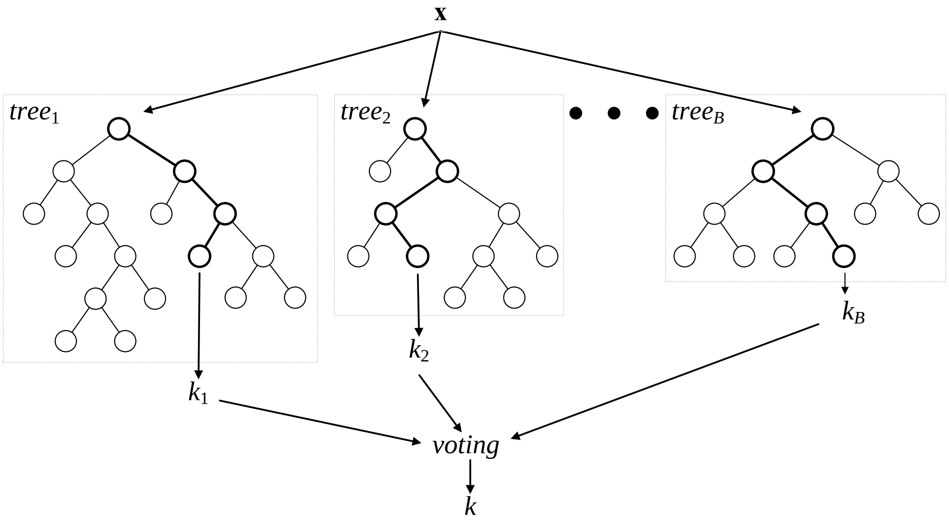

Figure 3: Architecture of RF. For classification, a final decision is obtained by majority voting.

RF is famous as being robust against over-fitting and, when the number of trees increases, the generalization error converges to a limit [22]. Low bias and low correlation between trees are essential for good generalization of the final ensemble. Therefore, to achieve low bias, trees are unpruned (grown to the maximum depth) and to achieve the low correlation, randomization is applied when constructing an individual tree.

RF is constructed according to the following steps:

1. Choose the forest size B as a number of trees to grow and the subspace size $q \leq p$ as a number of features to provide for each tree node.

2. Draw a bootstrap sample (random sample with replacement) of the dataset, which generally results in $\sim \frac{2}{3} \cdot n$ unique observations for training, thus leaving $\sim \frac{1}{3} \cdot n$ for testing as the out-of-bag (OOB) dataset for that particular tree, where n is the number of observations in the dataset.

3. Grow an unpruned tree using the bootstrap sample. When growing a tree, at each node, q variables are randomly selected out of the p available.

4. Repeat steps 2 and 3, until the size of the forest reaches B.

The generalization performance of RF was estimated using the internal out-of-bag (OOB) validation, meaning that an observation is classified only by the trees constructed without using that observation. It is known that the OOB validation provides an unbiased estimate of a test set error, similar to the leave-one-out procedure. Due to the "repeated measures" aspect, when each subject is represented by several recordings of voiced phonation, the sampling part of the Matlab implementation [23] had to be modified to ensure that all data of each subject are included either in a bootstrap sample or left aside as OOB. The modification conforms to the leave-one-subject-out approach and prevents pathology detection intermingling with speaker detection. Additionally, the RF setting of stratified sampling was used to preserve the class and gender balance of the full dataset in each bootstrap sample.

3.2.2. Decision-level fusion

A base-RF was built for each type of decomposition (EMD and VMD) and each sampled frame (1 central or 3 equidistant). Outputs from the first stage base-RF were converted into the difference between class posteriori probabilities. Given a trained base-learner, this difference is estimated as:

$$d(\{t_1, ..., t_b\}, \mathbf{x}) = \frac{\sum_{i=1}^{b} f(t_i, \mathbf{x}, c=2)}{b} - \frac{\sum_{i=1}^{b} f(t_i, \mathbf{x}, c=1)}{b} \tag{3}$$

where $\mathbf{x}$ is the object being classified, b is the number of trees $t_1, ..., t_b$ in the RF, for which the observation $\mathbf{x}$ is OOB, c is a class label (1 corresponds to HC, 2 to PD), and $f(t_i, \mathbf{x}, c)$ stands for the c-th class frequency in the leaf node, into which $\mathbf{x}$ falls in the i-th tree t_i of the forest:

$$f(t_i, \mathbf{x}, c) = \frac{n(t_i, \mathbf{x}, c)}{\sum_{j=1}^{C} n(t_i, \mathbf{x}, c_j)} \tag{4}$$

where C is the number of classes and $n(t_i, \mathbf{x}, c)$ is the number of training data from the class c falling into the same leaf node of t_i as $\mathbf{x}$.

After having the base-RF score as a difference between the class probabilities, all scores for a single voice recording were summarized by applying the statistical functionals. The number of scores summarized corresponds to the number of IMFs extracted. Since this number is not fixed and dependent upon the frame analysed, the statistical functionals effectively convert a decisions sequence of varied length into a meta-vector of fixed length. Creating the meta-features by 13 statistical functionals for the two types of decomposition results in a meta-feature vector of 26 elements for the uni-frame and 78 elements for the tri-frame case.

3.2.3. Assessing detection

The output scores obtained for OOB data from the base-RF's (the baseline case) and the meta-RF's (the proposed case) were used to evaluate the goodness of detection. RF votes were converted to a proper score by dividing the votes for a specific class by the total number of times the case was OOB, as in formula (3). A soft decision (score) instead of hard decision (predicted class) makes evaluation more precise by enabling visual summary of detection performance through the detection error trade-off (DET) curve, as recommended by [24]. Detectors with different DET curves can be conveniently compared by the equal error rate (EER) – the equilibrium point where the curve intersects the diagonal [25] and the false positive rate (miss rate) becomes equal to the false negative rate (false alarm rate) or the true positive rate (sensitivity) becomes equal to the true negative rate (specificity). The minimum cost of the log-likelihood-ratio ($\mathrm{C_{llr}}$) is a comprehensive detection metric used here as the main criterion for RF model selection. The log-likelihood-ratio is the logarithm of the ratio between the likelihood that the target (PD subject) produced the signal and the likelihood that a non-target (HC subject) produced the phonation signal. The DET curve, EER and $\mathrm{C_{llr}}$ measures were computed by the ROC convex hull method using the BOSARIS toolkit [26]. A well-calibrated and useful detector should provide $\mathrm{C_{llr}} < 1$ and $\mathrm{EER} < 50\%$, the lower the better.

The detection task was repeated several times and the OOB performance by $\mathrm{C_{llr}}$ and EER was estimated each time. The number of repetitions was set to 9 for the baseline and 27 for the decomposition-based solution – 3 feature extractions × 3 base-learners × 3 meta-learner runs. The mean and the 95% confidence interval were calculated, to assess the performance. A medoid, a median detection result, was obtained using the k-medoids clustering with $k = 1$ in the 2D space, formed by the z-scored $\mathrm{C_{llr}}$ and EER. Transformation of EER and $\mathrm{C_{llr}}$ metrics through z-score (by subtracting the mean and dividing by standard deviation) before clustering helps to avoid influence of different scales. The medoid result is reported in tables and used to summarize the detection performance by the DET curve. Multiple comparisons using the non-parametric Kruskal-Wallis test were done to compare the decomposition-based detection variants.

4. Experimental results

The number of IMFs extracted slightly varied, but was in the range from 8 to 12. The OOB detection performance is summarized in Table 2 and Table 3. The baseline solution with the EER of 35.48% for the AC and 37.36% for the SP microphone was clearly outperformed by the decomposition-based variants, achieving the EER of 0.95% for the AC and 12.2% for the SP microphones. It can be noted from the confidence invervals that the detection performance results are more stable in the baseline than in the proposed solution. Interestingly, when switching from the uni-frame to the tri-frame variant, the variance of the detection performance for the SP channel noticeably increased.

With regard to the choice of cepstral coefficients, the PLPCC appear to be the best in the baseline and the uni-frame cases, whereas in the tri-frame case, either MFCC for the AC or BFCC for the SP channel appear to perform slightly better. According to $\mathrm{C_{llr}}$

Table 2: Detection results by C_{llr}. Note: 95CI stands for the 95% confidence interval.

Method	**AC microphone**		**SP microphone**	
	Mean ± 95CI	**Medoid**	**Mean ± 95CI**	**Medoid**
BFCC baseline	0.921 ± 0.003	0.918	0.897 ± 0.004	0.896
MFCC baseline	0.940 ± 0.003	0.939	0.913 ± 0.002	0.914
PLPCC baseline	0.868 ± 0.003	0.870	0.891 ± 0.003	0.892
BFCC uni-frame	0.315 ± 0.013	0.324	0.509 ± 0.019	0.513
MFCC uni-frame	0.073 ± 0.037	0.077	0.598 ± 0.008	0.598
PLPCC uni-frame	0.257 ± 0.016	0.256	0.503 ± 0.007	0.507
BFCC tri-frame	0.067 ± 0.025	0.028	0.345 ± 0.039	0.329
MFCC tri-frame	0.034 ± 0.017	0.029	0.469 ± 0.035	0.466
PLPCC tri-frame	0.141 ± 0.017	0.143	0.366 ± 0.032	0.377

Table 3: Detection results by EER (in %). Note: 95CI stands for the 95% confidence interval.

Method	**AC microphone**		**SP microphone**	
	Mean ± 95CI	**Medoid**	**Mean ± 95CI**	**Medoid**
BFCC baseline	37.74 ± 0.53	37.64	39.08 ± 0.39	39.10
MFCC baseline	41.97 ± 0.50	41.95	39.05 ± 0.29	39.17
PLPCC baseline	35.62 ± 0.40	35.48	37.46 ± 0.43	37.36
BFCC uni-frame	10.31 ± 0.60	10.17	18.42 ± 0.98	18.76
MFCC uni-frame	2.07 ± 0.99	2.41	25.56 ± 0.70	26.62
PLPCC uni-frame	8.05 ± 0.65	7.89	18.18 ± 0.61	17.91
BFCC tri-frame	2.38 ± 0.84	1.03	12.71 ± 1.47	12.20
MFCC tri-frame	1.11 ± 0.56	0.95	18.26 ± 1.69	18.03
PLPCC tri-frame	4.70 ± 0.57	4.75	13.33 ± 1.33	12.79

and EER, the tri-frame has a significant advantage over the uni-frame variant for the SP channel, based on BFCC and PLPCC results in the bottom part of Fig. 6. Meanwhile, no significant difference between tri-frame and uni-frame variants can be noticed for the AC channel, based on MFCC results in the top part of Fig. 6. Due to logarithmic axis in DET plot, the difference appears to be more noticeable in Fig. 4 due to smaller error values than in Fig. 5.

5. Discussion and conclusions

The fact that the baseline results are more stable (exhibit smaller variance), could be due to the bias-variance trade-off, well known in machine learning. The tri-frame variant, even if results are more volatile for the SP channel, could be recommended over the uni-frame variant. Although, the difference between variants was not significant when using MFCC features for the AC channel. Volatility in detection results is the main limitation of the proposed solution. Aiming to reduce volatility, more frames, for example, at random locations and/or choice of an optimal frame size could constitute a potential direction for

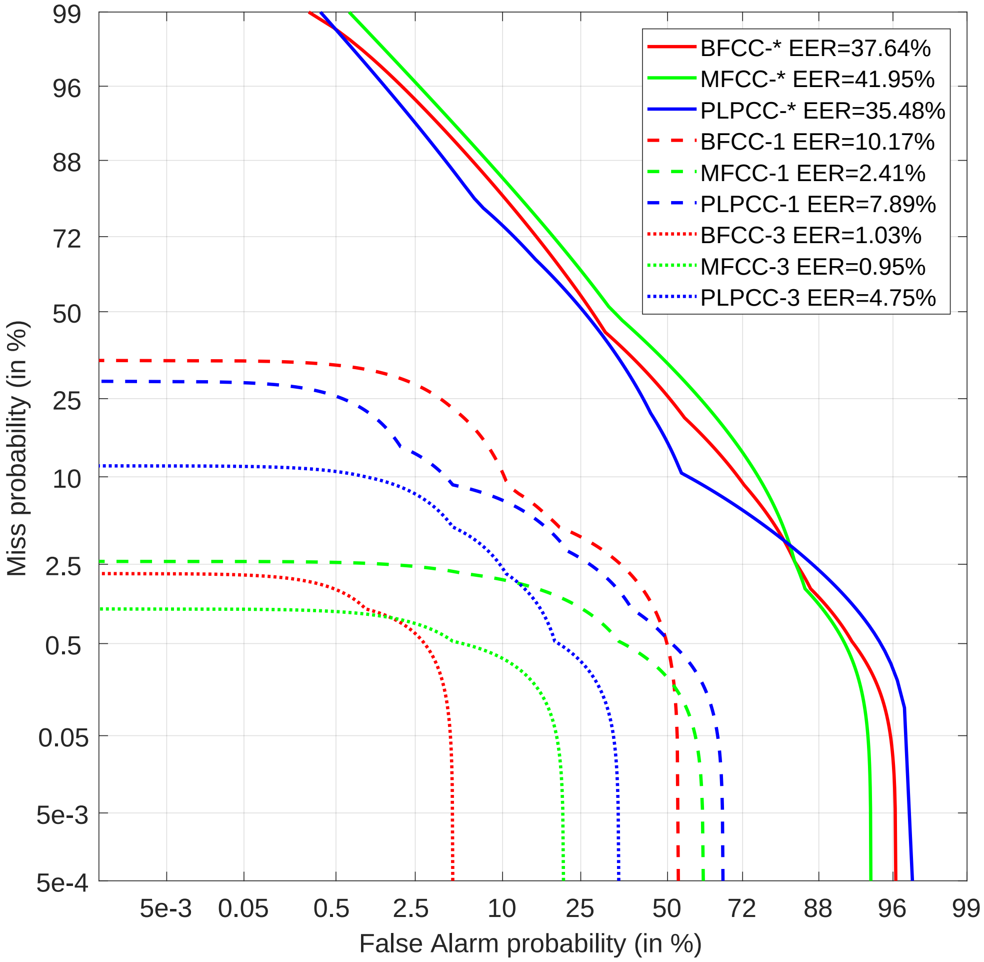

Figure 4: The OOB detection performance by the DET curves for the acoustic cardioid microphone. Notes: the baseline (CC-*) and the proposed uni-frame (CC-1) and tri-frame variants (CC-3).

the future.

Experiments using the same database in [16] have also indicated that the SP channel is more challenging. With respect to the results published in [16], which indicated the phonation modality as inferior to speech, the decomposition-based solution proposed here improved the EER in the voice-based PD detection from 20.78% (using the Essentia descriptors [27]) to 0.95% (using the MFCC after the tri-frame EMD/VMD) for the AC channel and from 29.02% (using Tsanas features [11]) to 12.2% (using the BFCC after the tri-frame EMD/VMD) for the SP channel. Moreover, the current results indicate that using just a few frames from sustained phonation can outperform the decision-level fusion of various audio feature sets extracted from both phonation and speech modalities in [16].

To summarize spectral characteristics of the extracted IMFs in decomposition-based approach, MFCC features could be recommended for the AC and BFCC (or PLPCC) features for the SP microphone. The introduced application of the EMD/VMD has potential to achieve an accurate voice-based PD screening and could be useful as an expert system for

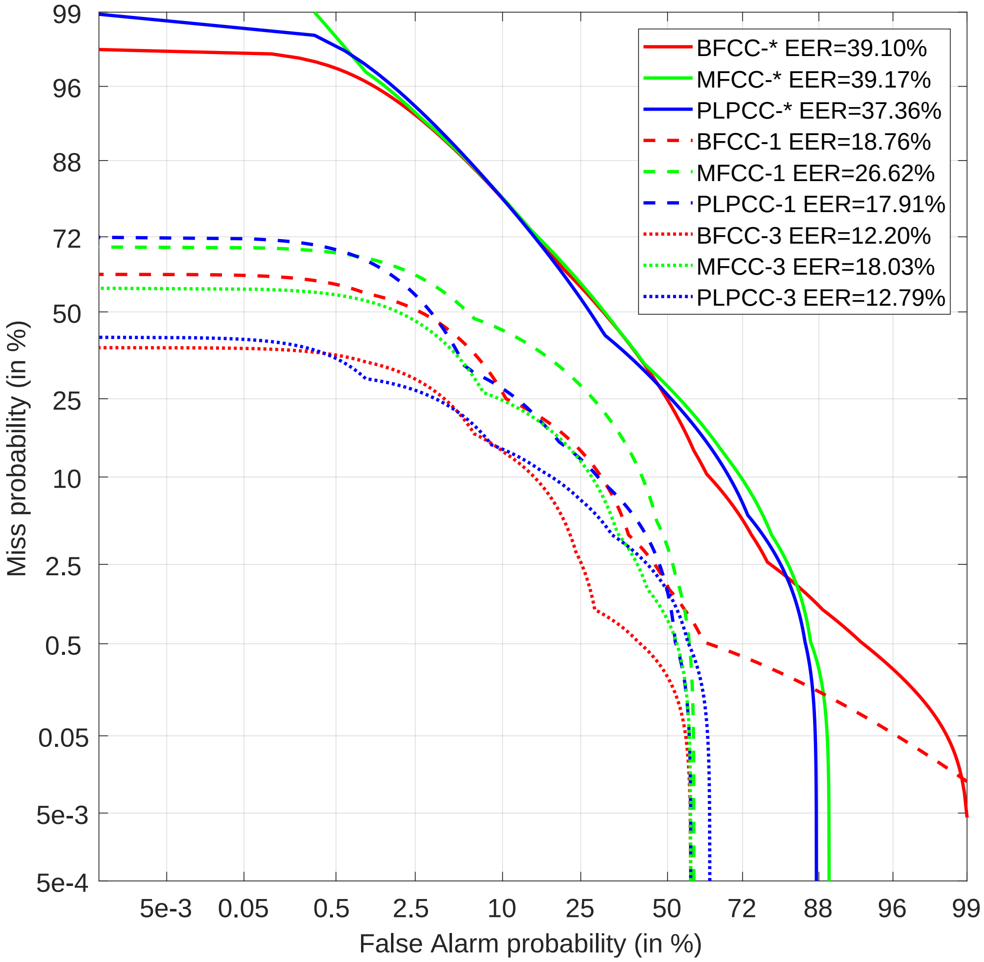

Figure 5: The OOB detection performance by the DET curves for the smart phone microphone. Notes: the baseline (CC-*) and the proposed uni-frame (CC-1) and tri-frame variants (CC-3).

medical decision support.

Acknowledgments

This research was funded by a grant (No. MIP-075/2015) from the Research Council of Lithuania. Initial findings were presented at the IVUS 2017 conference [28].

References

[1] M. C. de Rijk, L. J. Launer, K. Berger, M. M. B. Breteler, J.-F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. M. Martínez-Lage, C. Trenkwalder, A. Hofman, Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts, Neurology 54 (11 Suppl 5) (2000) S21–S23, Neurologic Diseases in the Elderly Research Group. doi:10.1212/WNL.54.11.21A.

[2] J. M. Fearnley, A. J. Lees, Ageing and Parkinson's disease: substantia nigra regional selectivity, Brain 114 (5) (1991) 2283–2301. doi:10.1093/brain/114.5.2283.

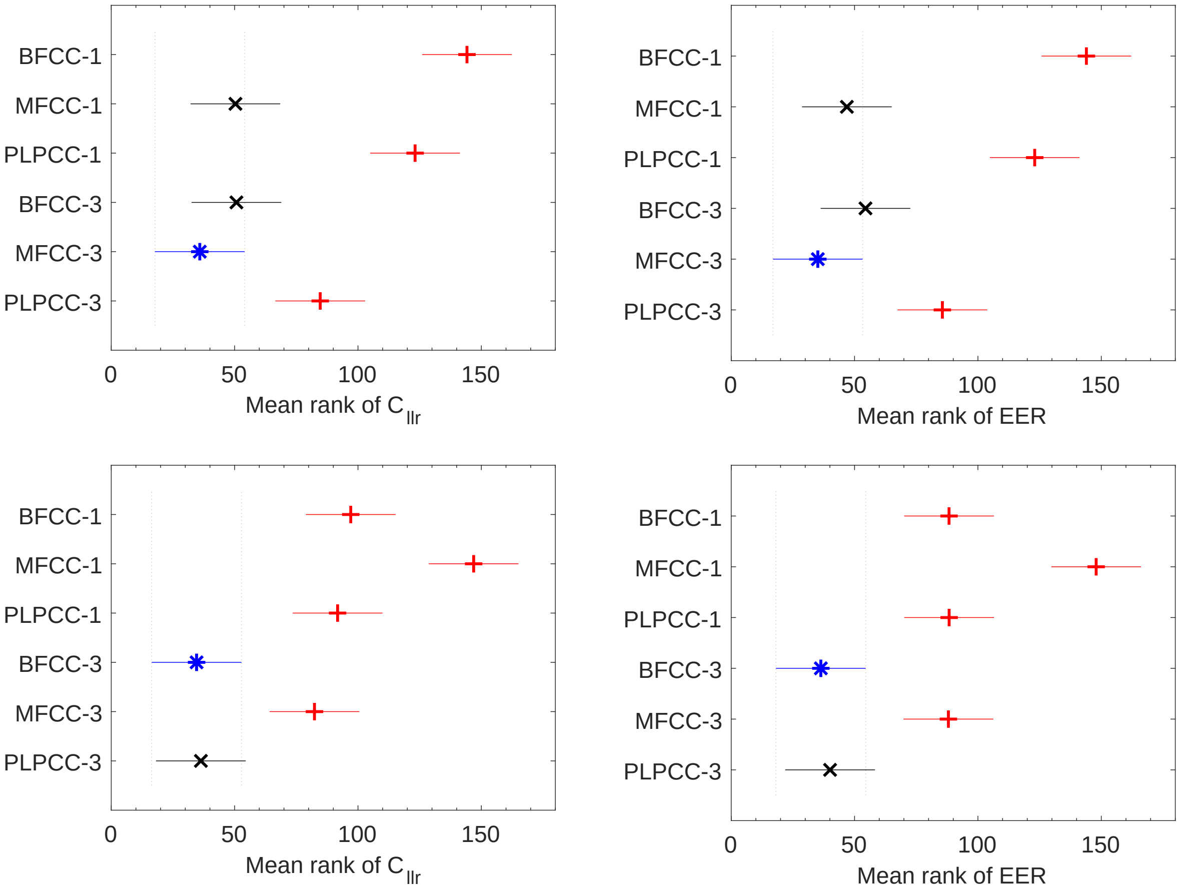

Figure 6: The multiple comparisons procedure of the detection performance by the non-parametric Kruskal-Wallis test (with Tukey's HSD criterion and the 95% confidence). Performance metrics: C_{llr} (**left**) and EER (**right**). Microphone channels: AC (**top**) and SP (**bottom**). The best result is denoted by an asterisk (*), results similar to the best by a cross (×), and statistically significantly worse results are denoted by a plus (+) sign. Detection variants compared: uni-frame (CC-1) and tri-frame (CC-3).

[3] J. H. Kordower, C. W. Olanow, H. B. Dodiya, Y. Chu, T. G. Beach, C. H. Adler, G. M. Halliday, R. T. Bartus, Disease duration and the integrity of the nigrostriatal system in Parkinson's disease, Brain 136 (8) (2013) 2419. `doi:10.1093/brain/awt192`.

[4] G. Acampora, D. J. Cook, P. Rashidi, A. V. Vasilakos, A survey on ambient intelligence in healthcare, Proceedings of the IEEE 101 (12) (2013) 2470–2494. `doi:10.1109/JPROC.2013.2262913`.

[5] S. Arora, V. Venkataraman, A. Zhan, S. J. Donohue, K. M. Biglan, E. R. Dorsey, M. A. Little, Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study, Parkinsonism and Related Disorders 21 (6) (2015) 650–653. `doi:10.1016/j.parkreldis.2015.02.026`.

[6] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, A. D. Trister, The mPower study, Parkinson disease mobile data collected using ResearchKit, Scientific Data 3 (2016) 160011. `doi:10.1038/sdata.2016.11`.

[7] A. Rojas, J. Górriz, J. Ramírez, I. Illán, F. Martínez-Murcia, A. Ortiz, M. G. Río, M. Moreno-Caballero, Application of empirical mode decomposition (emd) on datscan spect images to explore parkinson disease, Expert Systems with Applications 40 (7) (2013) 2756–2766. `doi:10.1016/j.eswa.2012.11.017`.

[8] E. R. de Lima, A. O. Andrade, J. L. Pons, P. Kyberd, S. J. Nasuto, Empirical mode decomposition: a novel technique for the study of tremor time series, Medical and Biological Engineering and Computing 44 (7) (2006) 569–582. doi:10.1007/s11517-006-0065-x.

[9] L. Ai, J. Wang, R. Yao, Classification of parkinsonian and essential tremor using empirical mode decomposition and support vector machine, Digital Signal Processing 21 (4) (2011) 543 – 550. doi:10.1016/j.dsp.2011.01.010.

[10] Z. Smekal, J. Mekyska, Z. Galaz, Z. Mzourek, I. Rektorova, M. Faundez-Zanuy, Analysis of phonation in patients with Parkinson's disease using empirical mode decomposition, in: 2015 International Symposium on Signals, Circuits and Systems (ISSCS), 2015, pp. 1–4. doi:10.1109/ISSCS.2015.7203931.

[11] A. Tsanas, Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning, Ph.D. thesis, Oxford Centre for Industrial and Applied Mathematics, University of Oxford, Oxford, United Kingdom, Supervisors: M. A. Little and P. E. McSharry. (June 2012).
URL http://people.maths.ox.ac.uk/tsanas/software.html

[12] A. Tsanas, M. A. Little, P. E. McSharry, J. L. Spielman, L. O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Transactions on Biomedical Engineering 59 (5) (2012) 1264–1271. doi:10.1109/TBME.2012.2183367.

[13] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. D. Arias-Londoño, E. Nöth, Automatic detection of Parkinson's disease from continuous speech recorded in non-controlled noise conditions, in: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), International Speech Communication Association, Dresden, Germany, 2015, pp. 105–109.

[14] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, E. Nöth, Automatic detection of Parkinson's disease in running speech spoken in three different languages, The Journal of the Acoustical Society of America 139 (1) (2016) 481–500. doi:10.1121/1.4939739.

[15] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, Detecting Parkinson's disease from sustained phonation and speech signals, PLOS ONE 12 (10) (2017) 1–16. doi:10.1371/journal.pone.0185613.

[16] E. Vaiciukynas, A. Gelzinis, A. Verikas, M. Bacauskiene, Parkinson's disease detection from speech using convolutional neural networks, in: B. Guidi, L. Ricci, C. Calafate, O. Gaggi, J. Marquez-Barja (Eds.), Smart Objects and Technologies for Social Good, Springer International Publishing, 2018, pp. 206–215. doi:10.1007/978-3-319-76111-4_21.

[17] M. A. Colominas, G. Schlotthauer, M. E. Torres, Improved complete ensemble EMD: A suitable tool for biomedical signal processing, Biomedical Signal Processing and Control 14 (0) (2014) 19–29. doi:10.1016/j.bspc.2014.06.009.

[18] K. Dragomiretskiy, D. Zosso, Variational mode decomposition, IEEE Transactions on Signal Processing 62 (3) (2014) 531–544. doi:10.1109/TSP.2013.2288675.

[19] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, Journal of the Acoustical Society of America 87 (4) (1990) 1738–1752.

[20] J. Cooley, J. Tukey, An algorithm for the machine calculation of complex Fourier series, Mathematics of Computation 19 (90) (1965) 297–301. doi:10.2307/2003354.

[21] D. P. W. Ellis, PLP and RASTA (and MFCC, and inversion) in Matlab (2005).
URL http://labrosa.ee.columbia.edu/matlab/rastamat/

[22] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[23] A. Jaiantilal, Random forest (regression, classification and clustering) implementation for Matlab (and standalone) (2012).
URL https://github.com/jrderuiter/randomforest-matlab

[24] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, P. Gómez-Vilda, Methodological issues in the development of automatic systems for voice pathology detection, Biomedical Signal Processing and Control 1 (2) (2006) 120–128, Voice Models and Analysis for Biomedical Applications. doi:10.1016/j.bspc.2006.06.003.

[25] M. Faundez-Zanuy, E. Monte-Moreno, State-of-the-art in speaker recognition, IEEE Aerospace and Electronic Systems Magazine 20 (5) (2005) 7–12. doi:10.1109/MAES.2005.1432568.

[26] N. Brümmer, E. de Villiers, The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF, arXiv 1304 (2865v1) (2013) 1–23, Presented at the NIST SRE'11 Analysis Workshop, Atlanta, December 2011. arXiv:1304.2865v1.
URL http://sites.google.com/site/bosaristoolkit

[27] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, X. Serra, Essentia: an audio analysis library for music information retrieval, in: International Society for Music Information Retrieval Conference (ISMIR), Curitiba, Brazil, 2013, pp. 493–498.
URL http://essentia.upf.edu

[28] A. Šimulis, E. Vaičiukynas, Exploiting voice signal decomposition in expert system for Parkinson's disease detection, in: R. Damaševičius, T. Krilavičius, A. Lopata, C. Napoli, M. Woźniak (Eds.), Proceedings of the IVUS International Conference on Information Technology, Vol. 1856 of CEUR Workshop Proceedings, Kaunas, Lithuania, 2017, pp. 49–54.
URL http://ceur-ws.org/Vol-1856/p10.pdf