IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 4, 2020, accepted December 13, 2020, date of publication December 23, 2020, date of current version January 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046873

Multicondition Training for Noise-Robust Detection of Benign Vocal Fold Lesions From Recorded Speech

MARIO MADRUGA[1], YOLANDA CAMPOS-ROCA[2], AND CARLOS J. PÉREZ[1]

[1]Departamento de Matemáticas, Universidad de Extremadura, 10003 Cáceres, Spain
[2]Departamento de Tecnología de los Computadores y las Comunicaciones, Universidad de Extremadura, 10003 Cáceres, Spain

Corresponding author: Mario Madruga (mariome@unex.es)

This work was supported in part by the Ministerio de Ciencia, Innovación y Universidades, under Project MTM2017-86875-C3-2-R; in part by the Junta de Extremadura/European Regional Development Funds, EU, under Project IB16054, Project GR18108, and Project GR18055; and in part by the Ministerio de Ciencia, Innovación y Universidades, under Grant FPU18/03274.

ABSTRACT This study evaluates the effects of Multicondition Training (MCT) on computer aided diagnosis systems for voice quality assessment associated to exudative lesions of Reinke's space. This technique adds various noise conditions to the speech recordings in order to recreate realistic acoustic environments. Four different databases (Massachussets Eye and Ear Infirmary, UEX-Voice, Saarbrücken, and Hospital Universitario Príncipe de Asturias) recorded in very different acoustic environments are used. We compare the outcomes of random forest classifier models comprising feature selection, hyperparameter tuning, and cross-validation attending the specific MCT schema used to separate healthy from pathological subjects for three diseases (nodules, polyps, and Reinke's edema). Apart from the clean case baseline, an asymmetric (one subject recording is affected only by one noise recording) and two symmetric (one subject recording is affected by all the noise recordings) noise-based MCT scenarios are considered. These scenarios are created by adding realistic acoustic noise of different types to the sustained /a/ vowel recordings. The symmetric approaches are affected by methodological concerns and are tested with a comparative purpose, to emphasize these issues. Experimental results highlight the drawbacks of symmetric MCTs and exclude these techniques as a viable option. In contrast, asymmetric MCT is proven to be a suitable noise-robust approach to build a diagnosis system for exudative lesions of Reinke's space, as performance obtained with the resulting classifiers is not far from the performance obtained for clean training.

INDEX TERMS Acoustic features, computer aided diagnosis (CAD), machine learning, multicondition training (MCT), nodules, polyps, Reinke's edema.

I. INTRODUCTION

Human voice production can be affected by a wide range of conditions, either vocal specific like nodules, polyps, cleft lip and palate, or by other disorders which affect motor control like neurodegenerative diseases. Either way, voice quality assessment is a reliable source of information for physicians and patients for diagnosis and monitoring of the underlying disease.

Nodules, polyps, and Reinke's edema are the main lesions that occur in Reinke's space [1]. Although their etiologic factors are different, their pathologic features are quite similar and diagnosis usually relies on the clinical description of the patient. Classical voice quality assessment relies on cumbersome techniques such as videostroboscopy or laryngoscopy, procedures which are highly invasive and uncomfortable for patients, and require expensive equipment and expert practitioners. It is for such that Computer Aided Diagnosis (CAD) tools are of great interest since they can help diagnosis procedures by using voice recordings as a non-invasive biomarker. They are non-intrusive as they only perform signal processing of voice samples [2].

Different signal sources have been taken into consideration, being the most usual vocal production recordings and electroglottography (EGG) [3]. Both techniques have their pros and cons: whereas the latter one needs of specific equipment like electrodes and laryngograph, voice analysis only needs common recording equipment like microphones and

The associate editor coordinating the review of this manuscript and approving it for publication was Jiri Mekyska.

sound interfaces, being high quality devices widely available even in portable format like modern smartphones. However, such vocal recording devices are prone to be affected by interferences like environmental and electronic noise or reverberation, whereas EGG, measuring the glottal activity, is affected only by noise induced in the equipment electronics. Furthermore, the need of specific devices makes EGG less common and available. Vocal recordings will be, therefore, the subject of this study.

Research conducted in order to find reliable automatic voice quality assessment systems has considered different approaches [4]. One of them is by looking for new meaningful features, using a well known classifier. In that regard multiple research lines have been proposed, from pitch related features [5], cepstral analysis [6]–[8], non-linear analysis [9], [10] or wavelet transformation [11]. Other common route is researching a good new classifier which improves the already known ones, since new machine learning techniques are being constantly researched, and many of them have been applied to this particular field using already known features [12]. Examples are hidden Markov models (HMM) [13], gaussian mixture models (GMM) [14], support vector machines [15], random forests [16] or more recently artificial neural networks [7] and deep neural networks [17] among others. Even data augmentation techniques have been proposed, creating synthetic feature values in order to supply data for the classifiers due to the lack of pathological recordings [18], or new selection techniques, like paraconsistent machines [19].

Most of these systems are developed on voice databases collected in the best recording conditions available. The most common database is the Massachusets Eye and Ear Infirmary (MEEI) database [20], available since 1994, but nowadays some other databases have been created, like Hospital Universitario Príncipe de Asturias (HUPA), spanish database [21], Saarbrücken Voice Database (SVD), german database [22], or the Arabic Voice Pathology Database (AVPD), arabic database [23]. All of them were recorded in sound proofed rooms and even use KayPENTAX Computerized Speech Lab. However, those controlled acoustical and technical conditions can not be replicated in a real clinical environment, or from the opposite side, realistic noise conditions are not represented in the databases.

Multicondition Training (MCT) alleviates such underrepresentation by artificially adding noise to selected samples from the voices database prior any processing. That technique has been used in other application fields [24], [25] but, to the best of the author's knowledge, it has never been applied to voice quality assessment. The field of pathological voice detection represents a new challenge since the noise components caused by the pathology have to be discriminated within a noisy environment. In the present study we build MCT systems and evaluate their effects on the ability of the resulting classifier to distinguish between healthy and pathological voices affected by Reinke's space diseases such as nodules, polyps, and Reinke's edema.

II. VOICE DATABASES

We use four voice databases recorded in different environments: MEEI, well known and widely used as a research dataset, recorded in the most favorable conditions; a dataset collected at Universidad de Extremadura (UEX-Voice), recorded at a more realistic environment; SVD collected by at Institut für Phonetik, Universität des Saarlandes; and HUPA database, recorded by Universidad Politécnica de Madrid.

A. PARTICIPANTS

Details of the participants taken into consideration can be found below. All of the databases were previously sanitized in order to avoid undesired issues, as some databases lack information like some subjects' age at the time of recording, others include more than one recording for a given subject and health status, and there are even cases where a subject has samples in both healthy and pathological groups in the same database.

MEEI database, commercialized by KayPentax Corp, compiles recordings of voices affected by a wide variety of diseases along with a control group of healthy recordings as well. 53 healthy people are present, and nodules, polyps, and Reinke's edema have a representation of 18, 20, and 25 subjects, respectively.

UEX-Voice database recordings were performed in a diagnosis room at Hospital San Pedro de Alcántara (HSPdA), Cáceres [26], with no special sound isolation from aisles and surroundings (street noise, waiting rooms…). Those recordings include 24 nodules, 30 polyps, and 30 Reinke's edema samples. 30 healthy subjects were recruited among administration staff volunteers from Universidad de Extremadura during an annual health check-up, where an otorhinolaryngologist performed an evaluation and assessed a good vocal health status. All of the volunteers signed an informed consent concerning subsequent studies using the collected information.

SVD database [22] is a vast collection of recordings compiled by Institut für Phonetik at Universität des Saarlandes and the Phoniatry Section of the Caritas Clinit St. Theresia in Saarbrücken. It contains 869 healthy recordings, 17 nodules, 40 polyps, and 51 Reinke's edema samples. This huge imbalance in number had to be addressed by making a selection of healthy subjects: We tried to match the numbers of female and male subjects while keeping the average and standard deviation of the age as even as possible by matching each of the pathological utterances with a healthy one of the same sex and closest age possible, without repetitions.

HUPA database [27] was recorded by Universidad Politécnica de Madrid in Hospital Universitario Príncipe de Asturias. It contains 239 healthy, 29 nodules, 28 polyps, and 28 Reinke's edema utterances. As for SVD database, the imbalance was addressed by picking healthy subjects which matched the sex and age distribution of each of the diseases being considered, again matching healthy sex-age samples with each pathological recording without repetitions.

Table 1 shows sex and age distribution for each combination of database and disease after balancing SVD and HUPA databases.

B. RECORDING EQUIPMENT

MEEI database was recorded in a most optimal environment using KayPENTAX Computerized Speech Lab, a state-of-the-art equipment purposely designed for voice disease research, including features like professional grade audio capture or calibrated input [28]. Although recording conditions were strictly controlled, they vary among pathological and healthy voices, with different sampling rates, 50 kHz for normal vs. 25 kHz for pathological, with normal and pathological voices also recorded in different locations, which are not described but assumed to be acoustically identical [28].

Regarding UEX-Voice database, it was compiled using an AKG 520 head-worn condenser cardioid microphone attached to a TASCAM US322 interface using Audacity 2.0.5 recording software, with no special sound isolation from aisles and surroundings. The sampling rate was 44.1 kHz, and the resolution was 16 bits per sample.

SVD recordings were collected using a headset condenser microphone fed directly into a Kay elemetrics Computerized Speech Lab (CSL) station model 4300B, and recorded at 50 kHz sample rate and a bit depth of 16 bits inside a sound-treated room [29].

Finally, for HUPA database recordings were performed with the CSL 4300B equipment of Kay Elemetrics, using a condender microphone as input device, sampling both signals with a frequency of 50 kHz and 16 bits of quantization. All the recordings were taken under the same conditions and recording parameters, and were collected in a soundproof room [27].

C. VOCAL TASK

In MEEI database each subject was asked to perform a sustained phonation at a comfortable pitch and level for at least 3 seconds of the /a/ vowel, repeating the process 3 times, after which an expert speech pathologist chose the best sample for the database [28]. That sample was also trimmed down to 1 second looking for the stable part of the phonation before including it into the database.

In the case of UEX-Voice, the phonation of the /a/ vowel was kept up for at least 5 seconds in a single breath. Laryngological evaluation was performed by an otorhinolaryngologist using videostroboscopy. The leading and trailing segments of the recording were discarded prior to storing the utterance in the database. The depicted recording and research protocol was approved by the bioethics committees from both UEX and HSPdA.

SVD subjects on their side had to perform a phonation of the /a/ vowel, among other tasks which are not of interest for this study. A mid-section of the phonation was stored in the database, avoiding onset and offset segments.

TABLE 1. Age distribution by database, disease, health status, and sex.

Database	Disease	Health	Sex	Nº	Mean	Std
MEEI	Nodules	Normal	M	21	38.81	8.49
			F	32	34.16	7.87
			T	53	36.00	8.36
		Pathologic	M	1	47.00	0.00
			F	17	28.05	10.08
			T	18	29.11	10.75
	Polyps	Normal	M	21	38.81	8.49
			F	32	34.16	7.87
			T	53	36.00	8.36
		Pathologic	M	12	37.83	15.63
			F	8	55.00	14.91
			T	20	44.7	16.82
	Reinke	Normal	M	21	38.81	8.49
			F	32	34.16	7.87
			T	53	36.00	8.36
		Pathologic	M	5	50.6	14.72
			F	20	47.4	11.87
			T	25	48.04	12.22
UEX-Voice	Nodules	Normal	M	4	39.00	14.17
			F	26	41.04	11.18
			T	30	40.42	11.58
		Pathologic	M	1	64.00	0.00
			F	23	39.39	10.66
			T	24	40.42	11.58
	Polyps	Normal	M	4	39.00	14.17
			F	26	41.04	11.18
			T	30	40.42	11.58
		Pathologic	M	6	43.33	13.26
			F	24	46.21	11.83
			T	30	45.63	11.95
	Reinke	Normal	M	4	39.00	14.17
			F	26	41.04	11.18
			T	30	40.42	11.58
		Pathologic	M	3	35.67	22.19
			F	27	51.29	8.38
			T	30	47.97	11.97
SVD	Nodules	Normal	M	4	41.75	19.63
			F	13	31.92	10.87
			T	17	34.24	13.40
		Pathologic	M	4	40.25	23.10
			F	13	31.92	10.87
			T	17	33.88	14.21
	Polyps	Normal	M	23	52.00	14.49
			F	17	54.35	15.24
			T	40	53.00	14.67
		Pathologic	M	23	51.04	12.93
			F	17	54.64	15.94
			T	40	52.57	14.21
	Reinke	Normal	M	7	60.29	5.77
			F	44	53.57	11.57
			T	51	54.49	11.16
		Pathologic	M	7	60.14	5.08
			F	44	51.5	11.36
			T	51	52.69	11.07
HUPA	Nodules	Normal	M	1	18.00	0.00
			F	28	27.5	9.39
			T	29	27.17	9.39
		Pathologic	M	1	11.00	0.00
			F	28	27.46	9.79
			T	29	26.90	9.82
	Polyps	Normal	M	14	37.28	8.68
			F	14	40.29	8.17
			T	28	38.78	8.41
		Pathologic	M	14	37.07	8.30
			F	14	40.29	8.17
			T	28	38.68	8.24
	Reinke	Normal	M	12	54.00	11.09
			F	16	46.18	8.61
			T	28	49.53	10.33
		Pathologic	M	12	53.83	11.17
			F	16	46.25	8.68
			T	28	49.5	10.36

Patients in HUPA database had to perform a sustained phonation of the /a/ vowel. The resulting recording was later trimmed, discarding the first 500 ms and the last part of the utterances to avoid onset and offset issues, storing a midvowel segment of about 3 seconds length for each utterance.

III. CORRUPTION METHODOLOGY

The main problem we find in moving from a research context to a clinical one is the difference in environmental conditions. Most diagnosis rooms are much more noise affected than the labs where research recordings are usually taken. This is especially true in the case of MEEI database, where not only recording conditions are strictly controlled, but recordings are also screened in order to obtain the best examples of each disease. Therefore we have created a series of noise corruption schemata that try to replicate some of the most usual noises that could happen inside or in the surroundings of a typical diagnosis room.

A. NOISE DATABASE

There are many resources available on the Internet, with repositories containing sound samples from different sources, some of them oriented to other fields such as speech recognition. However, we have not found any published noise database for voice corruption in a CAD setting. Specifically, we were looking for sounds that meet the following requirements:

- The noise source would be common in a clinical environment.
- The recording is clean, containing one kind of noise.
- The noise is recognizable when listening, so the recording contains mostly noise from the source and not static noise.

The most suitable alternative we found is the MUSAN dataset [30], included in the OpenSLR repository.[1] It contains recordings of a variety of sounds, from which we extracted a subgroup which fulfills the aforementioned conditions. We selected 31 noise files which contain 7 different noise types. Table 2 shows the distribution of recordings and noise types present in the database. Noise classes include: indistinct voices, keyboard typing, doors (opening, closing, and squeaking), paper flicking, phone buzzers, meteorological conditions, and people walking around.

TABLE 2. Types of noises considered and number of recordings present in the corruption schemata.

Noise type	# of recordings
Babble	3
Keyboards	7
Doors	3
Paper manipulation	2
Phone buzzes	6
Rain	4
Steps	6

[1] www.openslr.org

MUSAN dataset recording characteristics remain unknown, as it is a compilation of different sources and recording situations. However, all the noise samples contained are available at a sampling rate of 16 kHz and a resolution of 16 bits per sample, with a highly variable recording length.

B. SPEECH CORRUPTION

Voice samples from both databases are intended to be affected by selected noise samples in a realistic way. In this case, noise is added to the recordings making sure that the Signal-to-Noise Ratio (SNR) does not exceed a given threshold to be configured at corruption time. We consider that noise is usually produced at a low enough level to be unnoticed by the patient or the practitioner at recording time, so the maximum noise level should remain below the desired threshold at all times during the voice recording.

In order to mimic such level, we perform the corruption by applying some gain to the additive noise in order to limit the effects of residual noise present in the voice recording, as we only have control over the former. Even though voice samples are intended to be recorded so that their signal power remains constant, one of the effects of voice diseases is the inability to control a steady output level. The same happens for noise recordings since no considered noise is stationary. Therefore, we have to ensure that the minimum signal and noise difference stays in a predefined range. Consequently, a Welch's periodogram is computed both on the voice and noise samples using a sliding window 10 ms long and a stride of 5 ms for power calculation; the window with least difference between signal and noise power is used to calculate the noise gain in order to get the desired SNR. We decided to add noise using SNRs of 20 and 30 dB, as lower levels would probably be noticed at recording time.

Noise recordings usually exceed voice recording length, so a random segment of the noise waveform is selected each time corruption was performed, adding some variability as noise samples will not be repeated in any iteration.

C. MCT APPROACHES

MCT requires a variety of conditions in the development dataset but, from that starting point, there are different ways to confront such task, which are shown in Fig. 1 and explained below.

The first one is asymmetric MCT, where the development dataset equals the size of the original dataset, but noise is added proportionally to the number of noise types present in the noise database, plus clean condition where no noise is added. In our experiments there are 7 types of noise, so for each classifier trained, 1/8 random subset of the original dataset is affected by each type of noise and the rest remain intact. As we have different number of noise recordings for each type of noise, random selection of noise recording is performed prior noise addition and subsequently we pick a random one-second clip for noise addition.

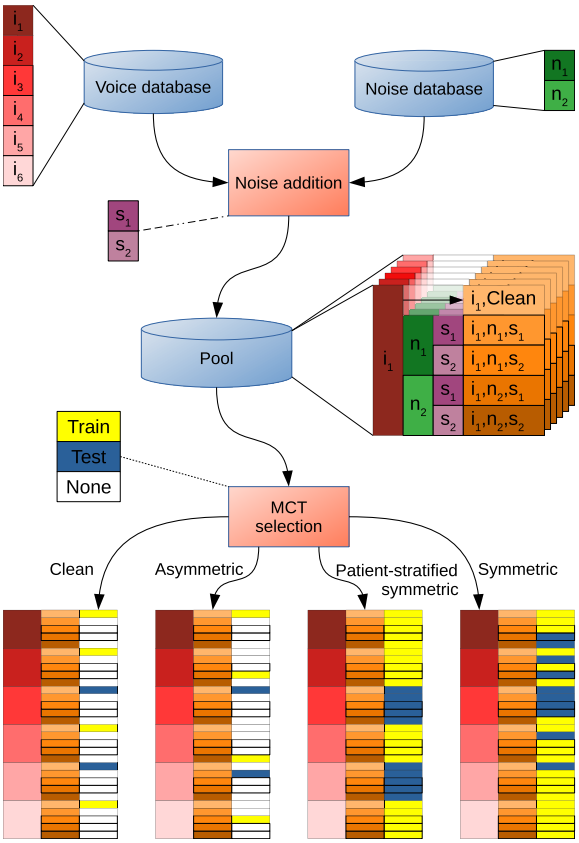

FIGURE 1. MCT selection process. The diagram shows an example: color coded are i=6 subjects, n=2 noises, and s=2 SNR levels in a 2/3 train - 1/3 test split. In our case i is the database size as shown in Table 1, n=31, s=2.

Another approach is symmetric MCT, where data augmentation is performed. This MCT technique takes the original recordings database and increases its size by adding all the different noise conditions being considered. In our case, 31 noise recordings were chosen, so the final dataset is 32 times the size of the original one (31 different noise affected datasets plus clean recordings). It is important to note that with symmetric MCT one subject appears in the dataset as many times as corruption conditions are present. This leads to two different approaches: the first one treats each recording as an independent instance and, when splitting into training and test, recordings from the same subject can lie in both subsets. It is also possible to add a stratification level in which training and test sets are not built with recordings but subjects, thus assigning all the recordings from a subject to the randomly chosen subset, either training or test, so a subject never has representation in both of them.

However, symmetric MCT methodology raises major concerns. Data augmentation can lead to good classification metrics, but constraints the generalization of the system, and performance when assessing new unknown recordings usually suffers. This is especially true in the case of symmetric MCT, since all of the individuals present in the development dataset can have representation in both training and test sets. In any case, although symmetric multicondition appears to be flawed by design, we are including the experiments and results obtained in order to further emphasize the concerns this approach rises.

Figure 1 shows the process followed to implement the aforementioned strategies. We start with one of the voice databases containing recordings for i individuals and the noise database of n noise recordings. We perform noise addition by adding the noises to the voice utterances using s different SNRs and create a pool of recordings available for MCT selection. That pool contains a total of $i+(i\times n\times s)$ utterances, i for the database size, $i\times n\times s$ for all the combinations with noises. For visual simplicity, in Fig. 1 $i=6$, $n=2$, $s=2$.

From that pool, the MCT selection schema can be clean, where only clean utterances are selected; asymmetric, where each noise type is present proportionally, including clean recordings, and only one recording per individual; patient-stratified symmetric, where all of a individual recordings lay either in train or test set; symmetric, where train and test utterances are selected randomly.

IV. CAD SYSTEM

The process followed to build a CAD system for each pathology is described next, specifically, feature extraction, feature selection and classification, and cross-validation methods.

A. FEATURE EXTRACTION

An initial number of 94 features was originally considered, from which 2 are sex and age, and the rest are described next. That set includes linear and non-linear features, all of them used in previous work either for functional voice disease diagnosis or other biomedical signal analysis. Extraction methods are coded in Python either using free implementations available in public repositories or translating code from other implementations. Analysis is performed in a long term basis since all recordings have been pre-processed to match some standard parameters as shown in section II-C.

Linear features include Cepstral Peak Prominence (CPP) [6], [31], Glottal-to-Noise Excitation ratio (GNE, 4 features: mean, standard deviation, Teager Kaiser energy Operator and squared energy operator) [32], [33], Glottal Quotient (GQ, 3 features) [33], [34], Harmonic-to-Noise Ratio (HNR) [33], [35], Jitter (22 features) [33], [36], Shimmer (22 features) [31], [33], Mel Frequency Cepstral Coefficients (MFCC, 13 features) [7].

In the nonlinear subset we consider correlation dimension (D2) [37], [38], First Minimum in Mutual Information (FMMI) [38], [39], First Zero in Correlation Function (FZCF) [38], [39], Hurst's exponent (HURST) [37], [40], MultiFractal Spectrum Width (MFSW) [40], and Zero Crossing Rate [38] (ZCR).

Finally, a set of entropies and complexities was computed, including permutation entropy (PERMUTATION) [41], Pitch Period Entropy (PPE) [33], [34], Recurrence Period Density Entropy (RPDE) [42], Shannon's entropy (SHANNON) [39], [43] and Lempel-Ziv complexity (LZ, 16 features attending to different quantization bin size) [44], [45].

B. FEATURE SELECTION AND CLASSIFICATION

The number of features extracted is very high, and comparable to the development set size for each disease. One desirable characteristic in CAD systems is simplicity, as it would not only solve the problem but also provide some insight in the possible causes of the disease and why the system assigns a label to a given sample. In classification tasks using acoustic features, complexity grows as we increase the number of features considered in the solution: a low number of features can be interpretable as it is possible to discern which conditions cause abnormal values.

Moreover, big feature vectors imply the possibility of overfitting. In our case that risk is evident since the initial number of features being considered outnumbers the size of the databases used as seen in Table 1, where the sum of pathological and normal individuals is lower than the number of features for all but one database-disease combination (SVD-Reinke's edema).

Given that we do not know the optimal number of features, our approach mixes feature selection an classification techniques in order to obtain optimal, small feature subsets: The first step is getting rid of redundant information considering pairwise correlation, reducing all the feature pairs that have a high Pearson coefficient to a single representative, repeating the process for every feature pair until no high correlation pairs are present. This step is performed once and applied for all the experiments proposed, as correlation only depends on feature extraction step.

From the low correlation feature set we select, in each case, a subset making use of Recursive Feature Elimination with cross-validation (RFECV). A significant number of RFECV repetitions with random cross-validation sampling are made and the selected features of each one are collected. Then, we created an optimum subset by counting the number of times each feature is selected and choosing only the ones which exceed the median number of repetitions.

Once we have a unique feature set for each training schema we proceed to apply random forest classifiers. Prior to any training we obtain an idea of the best hyperparameters by means of a grid search over each MCT strategy-dataset combination.

Finally, making use of the selected features and hyperparameters in each combination of database, disease and corruption schema we train a set of classifiers: Starting with the most repeated single feature in the RFECV step, a random forest is trained and its performance measured. The process is repeated adding features following the number of selections order obtained by the RFECV process, until all features are used, collecting the results for every feature set size. These steps are repeated, all classifier outcomes are collected, performance metrics are averaged and accuracy rate is used as performance measurement.

Confusion matrices containing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are collected. Average results for accuracy rate $((TP+TN)/(TP+TN+FP+FN))$, specificity $(TN/(TN+FP))$, sensitivity or recall $(TP/(TP+FN))$, precision $(TP/(TP+FP))$, and area under the curve - receiver operating characteristic (AUC-ROC) are collected as well as their coefficient of variation $(s/\bar{x} \times 100)$, where s is the standard deviation, and $\bar{x}$ is the arithmetic mean.

Though the number of features selected by RFECV is much lower than the original feature set size, we consider that it is still high since accuracy usually reaches a plateau or decays due to overfitting, so we chose to set a limit in the number of features used by taking the lowest subset whose mean accuracy reaches a certain threshold with respect to the maximum accuracy obtained.

C. CROSS-VALIDATION

Training a classifier and thus creating a model is a process driven by chance. The outcome is highly dependent on the selection of training and test sets, especially when the development set is small. In an ideal situation any combination of training and test sets would yield equivalent models of nearly identical performance. However, real life systems do not fulfil this requirement, so we need a reliable method to check a system performance. Cross-validation replicates an experiment multiple times with different test-train splits and averages their results, thus obtaining closer to the ideal situation metrics. Two steps in the pipeline require of cross-validation, and each one is performed in a different way: RFECV and classifier training.

In feature selection, RFECV uses K-Fold cross-validation in each step to select the least relevant features and discard them. K-Fold is designed making sure to keep the subject stratification correct, meaning that we take special care in the patient-stratified symmetric case for which, instead of splitting by recording, we pick subgroups by patient, and all recordings from a given patient lay in one of the folds.

To check the possible performance impact of MCT schemata, we perform cross-validation using a stratified shuffle split strategy, where in each iteration we randomly choose a portion of healthy and sick patients for the training set and the rest for test set. In the cases of clean and asymmetric MCT that task is trivial since pathological voice stratification is enough, keeping the normophonic-pathological proportion constant in training and test sets. However, in the case of symmetric corruption the multiplicity of recordings from each patient needs a closer look.

Two options arise, and both of them are tested: firstly, a simple shuffle and splitting technique on the recordings is performed, so we do not care if a patient had recordings in both training and test sets; secondly, a patient-stratified shuffling and splitting is performed along the usual pathological stratification, ensuring that all the recordings from a given individual lay in either training or test sets while maintaining the normophonic-pathological proportion in each one.

V. RESULTS

A. EXPERIMENTAL SETTINGS

We performed the steps detailed in Section IV: feature extraction, feature selection, classification, and cross-validation as follows, repeating the experiment several times and averaging the results. We have taken into consideration all 4 different scenarios depicted:

- Clean recordings: Using the original datasets without further manipulation.
- Asymmetric MCT: Partitioning the datasets into not overlapping equal size subsets and adding one kind of noise to each subset choosing a different noise sample for each recording. We also kept one of the partitions untouched.
- Symmetric MCT: Adding every sample from all of the noise types to the whole recording set of each dataset, thus working with an augmented database. Two different approaches were taken in this case regarding patients:
  - Patient stratified: Data manipulation in CAD training is aware of the patient, and it is taken into consideration when splitting the dataset (patient-stratified symmetric).
  - Raw datasets: Every recording is considered as an independent event (symmetric).

All vocal recordings were processed in the same way: First, all samples were trimmed down to 1 second length in order to ensure homogeneous length across databases; later, all of them were downsampled to 16 kHz prior corruption in order to match noise files sampling rate; after that, noise was added from all sources at all proposed SNRs; preprocessing was applied to the sound files prior feature extraction, normalizing amplitude to range $[-1, 1]$; and lastly, feature extraction was performed for each recording.

Highly correlated features were discarded when the Pearson coefficient exceeded 0.8. After feature discarding, most of the *feature families* such as jitter or shimmer were stripped down to one representative feature. We finally worked with the following 34 features: SEX, CPP, D2, FMMI, FZCF, GNE mean value (GNE_mean), GNE standard devition (GNE_std), GNE Teager Kaiser energy operator (GNE_SNR_TKEO), GNE squared energy operator (GNE_SNR_SEO), GQ pecentiles 5-95 (GQ_prc5-95), HNR, HURST, JITTER absolute difference (JITTER_abs_diff), LZ2, MFCC (MFCC_1-13), MFSW, PERMUTATION, PPE, RPDE, SHANNON, SHIMMER absolute difference (SHIMMER_abs_diff), and ZCR.

RFECV was performed following a 2-Fold cross-validation strategy, which consequently uses a 50/50 training/test splitting, computing 500 iterations during feature selection stage, using a random forest classifier with default parameters. For the classification task, 1000 shuffle and split repetitions were made using a 2/3 to 1/3 train/test proportion, and each train-test pair was used to train classifiers using an increasing number of features following the number of times each feature was selected in the RFECV selection step until all features were used, and their confusion matrices were collected. The threshold in accuracy for the final feature selection step was 0.975 times the maximum mean accuracy rate.

We will now detail the results obtained after training classifiers for the studied diseases: nodules, polyps, and Reinke's edema, and will compare the outcomes of using the original voice recordings, and the noise corruption scenarios proposed. Different scenarios will make use of different feature sets, which will be detailed and compared. Average results for accuracy, specificity, sensitivity, precision, and AUC-ROC will be displayed as well as their coefficient of variation.

B. NODULES

Metrics (Table 4) reveal that classifiers trained using MEEI database recordings are much more capable of a correct classification than the classifiers trained using any other database by a huge margin of more than 25% in accuracy rate: the almost perfect MEEI recordings easily achieve accuracies over 0.9 for all the experiments, no matter the corruption method, whereas the more realistic recordings of UEX-Voice, SVD, and HUPA do not get over 0.71 of accuracy, with the exception of symmetric corruption.

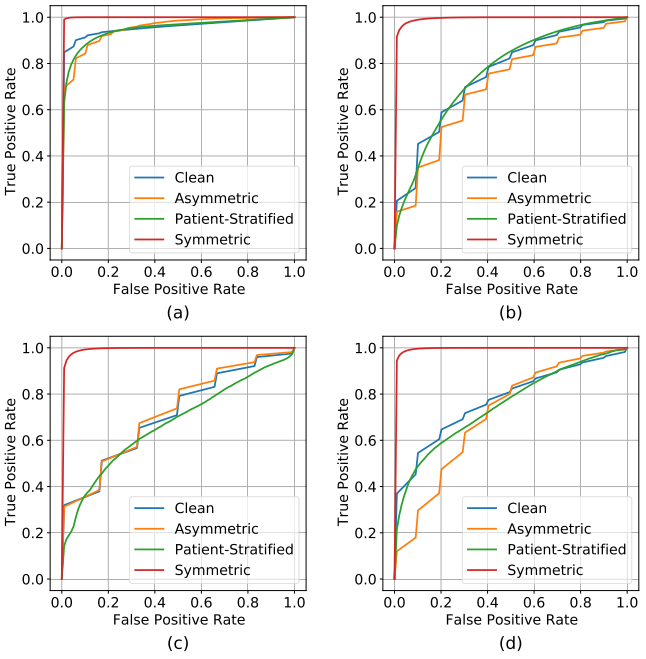

FIGURE 2. Mean ROC curves for nodules disease experiments. (a) MEEI database, (b) UEX-Voice database, (c) SVD database, (d) HUPA database.

Furthermore, the behavior of specificity, sensitivity, precision, and AUC-ROC appears to follow that of accuracy rate as a general rule, decreasing in a similar way as noise is introduced, so the system tends to maintain its ability throughout all the patients for a given database. AUC-ROC (curves on Fig. 2) values under clean conditions indicate a moderate ability to discern healthy from pathological voices for any disease. However, specificity shows a sub-par

TABLE 3. Features selected for nodules disease. Corruption cases are: Clean, Asymmetric, Patient-stratified symmetric, Symmetric.

	MEEI				UEX-Voice				SVD				HUPA			
	C	A	P	S	C	A	P	S	C	A	P	S	C	A	P	S
CPP	■			■	■	■	■						■		■	■
D2	■			■									■	■		■
FMMI																
FZCF			■	■			■									
GNE_SNR_SEO																
GNE_SNR_TKEO																
GNE_mean				■					■	■	■	■				■
GNE_std																
GQ_prc5_95																
HNR																
HURST	■	■	■					■		■	■	■				
JITTER_abs_dif									■			■				
LZ2													■			
MALE																
MFCC01	■	■	■													
MFCC02														■		
MFCC03					■	■	■	■								
MFCC04						■										
MFCC05																■
MFCC06					■		■	■								
MFCC07								■								
MFCC08								■								
MFCC09								■	■		■	■		■		
MFCC10								■								
MFCC11							■		■	■	■	■				
MFCC12													■		■	■
MFCC13															■	■
MFSW		■	■	■												
PERMUTATION						■							■	■	■	■
PPE																
RPDE													■			
SHANNON					■								■			■
SHIMMER_abs_dif		■		■					■	■						
ZCR												■				
TOTAL	4	4	4	6	4	4	5	7	5	4	4	6	7	4	4	8

performance for clean, asymmetric, and patient-stratified symmetric MCTs for all but SVD, showing that the classifier struggles to correctly classify healthy utterances, which is interesting as MEEI and UEX-Voice databases healthy group outnumbers pathological groups.

Coefficient of variation provides a deeper insight in the different performances. In MEEI database, while accuracy, sensitivity, and AUC-ROC variation tend to stay low, specificity and precision variation coefficient is three times as high. UEX-Voice, SVD, and HUPA on the other hand show a lower performance, not only in the mean values, but also in variability, with extreme cases like sensitivity for SVD database, asymmetric case, where we find that the coefficient of variation reaches 34%.

Differences in performance as we change corruption are remarkable: as we introduce noise, in the asymmetric case, performance decays slightly for MEEI and HUPA databases, but for UEX-Voice and SVD database accuracy remains almost equal, and even some variation coefficients are better. Looking at the symmetric corruption schema performance, it is very interesting to compare the results when performing two different data augmentation strategies: not taking care of patients when dividing the dataset, and splitting the training and test sets attending to the patient. In the former case performance levels rise to almost perfect classifiers with accuracy, specificity, sensitivity and precision levels between 0.94 and 0.99. For the latter case results are quite interesting: the levels achieved are generally lower than the clean and asymmetric counterparts.

Table 3 shows features selected for nodules disease when using the different database-MCT schema combinations. When taking apart MEEI database which is not realistic, and both symmetric MCT schemata because of their metholodogical issues, the only feature selected more than once under good methodological and environmental conditions is PERMUTATION.

C. POLYPS

Table 6 shows that for MEEI database, the baseline of clean case is quite good, with high accuracy, specificity, sensitivity, precision, and AUC-ROC mean levels, being specificity the worst and also the most affected by corruption, with a 13.4% performance dropping in the case of asymmetric corruption and even more for patient-stratified symmetric corruption.

Meanwhile, UEX-Voice database shows more homogeneous values: for clean, asymmetric, and patient-stratified

TABLE 4. Mean and coefficient of variation (CV) for accuracy, specificity, and sensitivity obtained for nodules disease.

		Clean		Asymmetric		Patient		Symmetric	
		Mean	CV	Mean	CV	Mean	CV	Mean	CV
MEEI	Accuracy	0,95	4,43	0,92	5,40	0,90	4,15	0,98	0,37
	Specificity	0,87	15,13	0,79	20,78	0,75	18,87	0,94	1,24
	Sensitivity	0,98	4,04	0,96	5,44	0,96	3,14	0,99	0,32
	Precision	0,93	11,01	0,88	15,38	0,84	11,35	0,97	0,92
	AUC-ROC	0,95	4,53	0,95	4,71	0,95	4,28	0,99	0,02
UEX-Voice	Accuracy	0,68	13,02	0,69	11,92	0,67	8,31	0,96	0,31
	Specificity	0,54	32,16	0,54	30,55	0,50	23,03	0,94	0,57
	Sensitivity	0,79	16,26	0,81	16,06	0,81	9,22	0,97	0,39
	Precision	0,70	23,19	0,62	21,64	0,67	15,18	0,96	0,99
	AUC-ROC	0,75	13,39	0,71	13,54	0,76	11,04	0,99	0,13
SVD	Accuracy	0,62	19,68	0,62	19,97	0,57	16,43	0,95	0,86
	Specificity	0,55	37,78	0,62	32,60	0,59	25,40	0,95	1,39
	Sensitivity	0,69	30,52	0,62	34,07	0,56	27,29	0,95	1,32
	Precision	0,67	27,33	0,64	25,02	0,58	17,95	0,95	1,24
	AUC-ROC	0,71	18,37	0,72	17,54	0,67	13,77	0,99	0,15
HUPA	Accuracy	0,71	13,35	0,65	14,74	0,65	10,31	0,95	0,70
	Specificity	0,68	20,40	0,63	27,55	0,54	22,43	0,95	1,07
	Sensitivity	0,75	19,02	0,67	25,04	0,75	13,60	0,95	1,15
	Precision	0,74	16,54	0,67	18,26	0,69	12,79	0,95	1,07
	AUC-ROC	0,77	11,93	0,71	13,77	0,75	10,48	0,99	0,08

symmetric cases, we obtain less than 5% difference in mean accuracy. SVD and HUPA databases yield worse results: whereas in the former asymmetric corruption only drops 3% and patient-stratified MCT drops 10%, the latter decays about 12% for asymmetric MCT and 15% for patient-stratified MCT. It is also remarkable the surprisingly low values obtained in the symmetric case for SVD database, which are good in comparison within the dataset, but quite low for a MCT-based comparison. Apart from that exception, symmetric corruption on its side gets overoptimistic results between 0.95 and 0.98 values for all the databases.

Once again, specificity, sensitivity, precision, and AUC-ROC follow the values obtained for accuracy, although in this case, unlike with nodules disease, specificity does not show the same weakness, with the exception of MEEI database. In this case, area under ROC curves, shown in Fig. 3 is quite good, reaching values over 0.80 for UEX-voice and HUPA databases which makes the system a fairly good detector in both cases. We can see in the flatter curves of subfigure 3(c) the difficulties with SVD database.

However, corruption affects differently all datasets: MEEI and HUPA corruption tends to be more noticeable with worse outcomes as we introduce noise, whereas UEX-Voice and SVD mean levels usually remain closer to clean condition for asymmetric and patient-stratified symmetric corruption schemata. Coefficient of variation follows the same trend: whereas asymmetric corruption in MEEI affects more negatively than in the other three databases, patient-stratified symmetric levels are better and, in some cases, even outperform the clean case with less variation for mean values in the same range.

In Table 5 we can see that in this case CPP stands as a good predictor under all circumstances. Furthermore, if we restrict the selection to realistic conditions (UEX-Voice, SVD, HUPA databases, and clean or asymmetric MCT), CPP is the only common feature selected.

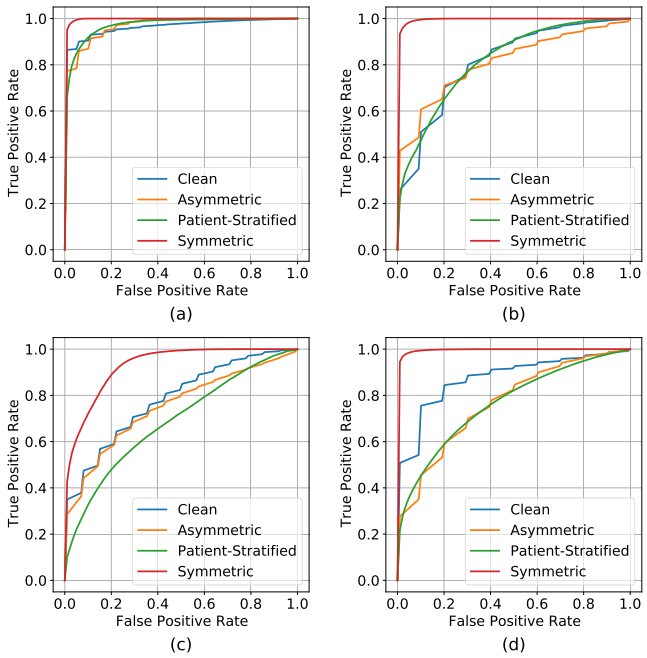

FIGURE 3. Mean ROC curves for polyps disease experiments. (a) MEEI database, (b) UEX-Voice database, (c) SVD database, (d) HUPA database.

D. REINKE's EDEMA

Once again, performances obtained, shown in Table 8, are great for MEEI database, with all metrics over 0.9 under clean training conditions, and accuracy, sensitivity, precision, and AUC-ROC above 0.96. Asymmetric and patient-stratified symmetric accuracy stay in the same range, with a penalty

TABLE 5. Features selected for polyps disease. Corruption cases are: Clean, Asymmetric, Patient-stratified symmetric, Symmetric.

	MEEI				UEX-Voice				SVD				HUPA			
	C	A	P	S	C	A	P	S	C	A	P	S	C	A	P	S
CPP			■		■	■	■	■	■	■	■		■		■	
D2	■												■	■		
FMMI																
FZCF			■					■				■			■	
GNE_SNR_SEO																
GNE_SNR_TKEO												■			■	■
GNE_mean		■	■	■				■								
GNE_std						■										
GQ_prc5_95																
HNR																
HURST	■		■				■	■			■					
JITTER_abs_dif						■				■	■	■				
LZ2									■				■			
MALE																
MFCC01																
MFCC02																
MFCC03	■	■	■	■	■		■		■		■	■				
MFCC04																
MFCC05																■
MFCC06					■	■	■	■								
MFCC07								■								■
MFCC08								■							■	■
MFCC09												■	■	■	■	
MFCC10								■				■				
MFCC11					■							■				■
MFCC12												■		■	■	■
MFCC13																
MFSW		■	■	■						■						
PERMUTATION		■	■							■		■				
PPE													■	■		
RPDE									■							
SHANNON	■															
SHIMMER_abs_dif				■									■			
ZCR								■								
TOTAL	4	4	7	4	4	4	4	9	4	4	4	9	6	4	6	6

of 4-5%, and sensitivity stays above 0.96, while specificity suffers a significant drop of 11% for asymmetric MCT, and 14% for patient-stratified MCT.

UEX-Voice, on its side, reaches good accuracy, specificity, and sensitivity levels, all over 0.72, for clean and asymmetric schemata, and results are also good for patient-stratified symmetric corruption, which gets the best accuracy and sensitivity results within the database. The same is true for HUPA database, with very similar to those of UEX-Voice mean levels for all metrics in all clean, asymmetric, and patient-stratified schemas. SVD on its side yields worse accuracy results. While UEX-Voice and HUPA performance drop with respect to MEEI database is 21%, in the case of SVD it goes further, up to 26%. Once again, symmetric MCT yields almost 1 accuracy values for every database.

Specificity, sensitivity, precision, and AUC-ROC easily follow accuracy in both, values and trend, as we introduce corrupted recordings, which shows the classifiers consistency for both healthy and pathological samples, although it is worth mentioning that for MEEI database, specificity drop is more noticeable than in any other database. AUC-ROC is remarkably good for HUPA and UEX-Voice databases, with values over 0.85. Once again, the flatter curves for SVD database shown in Fig. 4 show the difficulties the system finds in detecting diseases within this dataset.

In this case, Table 7 shows that GNE_mean is a great predictor since it is selected by 10 out of 12 database-MCT schema combinations. If we restrict ourselves to UEX-Voice, SVD, and HUPA databases, and clean and asymmetric MCT, GNE_mean is also the only common selected feature.

VI. DISCUSSION

We have studied the effects of three MCT strategies over three diseases and four databases. Results show a clear influence of the MCT strategy on the outcomes. Symmetric MCT is noteworthy as it gets very good results in every database-disease combination, not only in mean values, but also in relative dispersion. Under this type of corruption method, all considered noises are added to every utterance in the database. The result is striking, especially comparing it with patient-stratified symmetric corruption, for which the performance is assimilable to the one obtained with clean recordings and asymmetric corruption.

Athough addressed for other non physiological diseases, voice replication and data augmentation techniques are a major concern in the field of diagnosis using vocal

TABLE 6. Mean and coefficient of variation (CV) for accuracy, specificity, and sensitivity obtained for polyps disease.

		Clean		Asymmetric		Patient		Symmetric	
		Mean	CV	Mean	CV	Mean	CV	Mean	CV
MEEI	Accuracy	0,93	4,98	0,88	6,45	0,89	4,36	0,98	0,31
	Specificity	0,82	17,35	0,71	24,45	0,66	21,64	0,97	0,92
	Sensitivity	0,97	4,07	0,95	5,67	0,98	1,41	0,99	0,32
	Precision	0,93	10,01	0,88	14,67	0,90	7,50	0,97	0,90
	AUC-ROC	0,96	0,04	0,97	2,94	0,97	2,45	0,99	0,06
UEX-Voice	Accuracy	0,72	13,01	0,69	13,68	0,70	7,25	0,95	0,35
	Specificity	0,67	23,23	0,64	25,15	0,63	16,79	0,95	0,50
	Sensitivity	0,78	16,40	0,74	19,78	0,77	8,83	0,94	0,57
	Precision	0,78	13,82	0,77	14,03	0,72	11,24	0,95	0,88
	AUC-ROC	0,81	9,95	0,81	9,39	0,82	9,36	0,99	0,08
SVD	Accuracy	0,70	9,44	0,68	10,39	0,63	8,90	0,77	1,83
	Specificity	0,67	17,92	0,67	20,77	0,60	17,87	0,72	4,42
	Sensitivity	0,72	16,00	0,69	19,27	0,65	17,97	0,81	2,60
	Precision	0,72	11,80	0,69	12,91	0,64	10,70	0,80	2,02
	AUC-ROC	0,78	8,55	0,75	9,70	0,69	9,41	0,93	0,86
HUPA	Accuracy	0,79	10,43	0,69	13,52	0,67	10,62	0,97	0,52
	Specificity	0,80	17,82	0,63	26,85	0,58	21,84	0,96	0,80
	Sensitivity	0,78	16,71	0,74	20,49	0,75	14,54	0,97	0,71
	Precision	0,80	12,62	0,73	17,53	0,71	13,26	0,97	0,70
	AUC-ROC	0,87	8,41	0,76	13,20	0,76	0,08	0,99	0,11

TABLE 7. Features selected for Reinke's edema. Corruption cases are: Clean, Asymmetric, Patient-stratified symmetric, Symmetric.

	MEEI				UEX-Voice				SVD				HUPA			
	C	A	P	S	C	A	P	S	C	A	P	S	C	A	P	S
CPP				■	■		■	■		■	■	■				■
D2		■											■	■		
FMMI																
FZCF	■	■	■	■												
GNE_SNR_SEO																
GNE_SNR_TKEO						■		■								
GNE_mean		■	■	■	■	■	■	■	■	■	■	■		■	■	■
GNE_std																
GQ_prc5_95																
HNR	■								■	■						
HURST	■	■						■			■	■				■
JITTER_abs_dif													■	■		■
LZ2									■							
MALE																
MFCC01	■															
MFCC02										■						
MFCC03		■			■	■	■								■	■
MFCC04																
MFCC05																
MFCC06							■	■								
MFCC07			■	■				■								
MFCC08																
MFCC09																■
MFCC10								■								
MFCC11													■		■	
MFCC12									■	■	■	■				
MFCC13						■	■					■				■
MFSW		■	■	■								■				
PERMUTATION		■			■					■		■			■	■
PPE																
RPDE									■							
SHANNON	■											■	■		■	■
SHIMMER_abs_dif													■	■		
ZCR								■								
TOTAL	5	7	4	5	4	4	5	8	5	6	4	8	5	4	5	9

recordings [46]. The overoptimistic performance of symmetric MCT shows the methodological failure and origin of the great difference between symmetric and the rest of MCT schemata: the same subject can have, and in fact has, recordings both in training and testing sets. The presence of subjects in both sets helps the classifier, which learns

TABLE 8. Mean and coefficient of variation (CV) for accuracy, specificity, and sensitivity obtained for Reinke's disease.

		Clean		Asymmetric		Patient		Symmetric	
		Mean	CV	Mean	CV	Mean	CV	Mean	CV
MEEI	Accuracy	0,96	4,43	0,91	5,33	0,92	4,11	0,98	0,58
	Specificity	0,91	11,37	0,81	15,50	0,78	14,97	0,96	1,87
	Sensitivity	0,98	4,03	0,96	5,35	0,99	1,25	0,99	0,40
	Precision	0,97	5,81	0,92	9,64	0,98	2,54	0,98	0,70
	AUC-ROC	0,99	0,71	0,98	2,25	0,98	1,28	0,99	0,03
UEX-Voice	Accuracy	0,76	10,37	0,72	12,09	0,77	6,31	0,97	0,26
	Specificity	0,75	20,56	0,73	21,54	0,71	12,22	0,97	0,34
	Sensitivity	0,77	16,55	0,72	18,80	0,82	8,56	0,96	0,45
	Precision	0,78	12,76	0,79	14,32	0,72	11,60	0,96	0,79
	AUC-ROC	0,87	7,52	0,85	8,53	0,82	7,91	0,99	0,07
SVD	Accuracy	0,71	8,71	0,68	10,07	0,63	7,52	0,95	0,46
	Specificity	0,67	18,08	0,63	17,90	0,58	20,48	0,95	0,78
	Sensitivity	0,75	15,20	0,73	16,21	0,68	18,19	0,96	0,74
	Precision	0,73	11,48	0,71	12,53	0,65	10,34	0,96	0,70
	AUC-ROC	0,79	7,53	0,77	9,33	0,73	7,22	0,99	0,07
HUPA	Accuracy	0,76	11,54	0,74	10,68	0,70	10,11	0,94	0,83
	Specificity	0,78	18,54	0,73	19,05	0,67	17,61	0,96	0,99
	Sensitivity	0,75	19,72	0,76	19,89	0,74	15,51	0,91	1,51
	Precision	0,77	14,03	0,77	14,34	0,72	11,68	0,92	1,30
	AUC-ROC	0,85	8,83	0,84	8,85	0,79	9,62	0,99	0,06

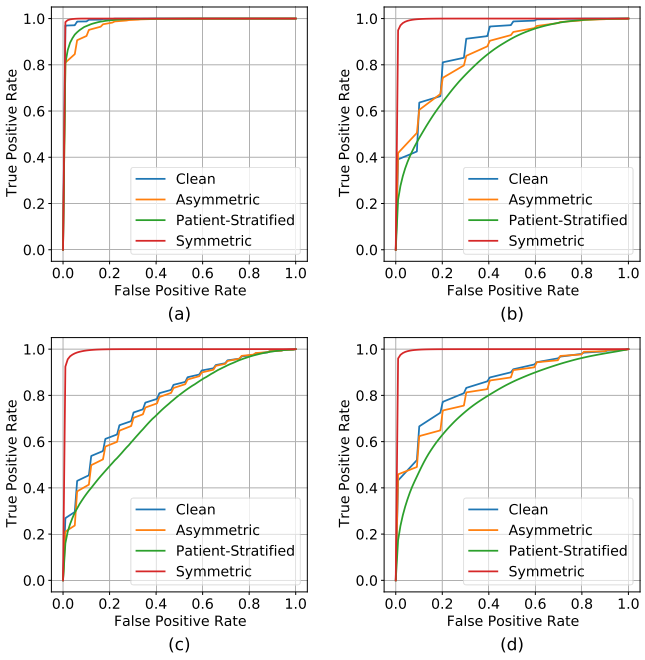

FIGURE 4. Mean ROC curves for Reinke's disease experiments. (a) MEEI database, (b) UEX-Voice database, (c) SVD database, (d) HUPA database.

to distinguish not only disease from normal recordings, but subjects themselves. This fact has a great influence in the outcome but raises strong methodological concerns.

Besides, symmetric MCT shows another weak spot in the number of features selected in each case. We can see that for most database-disease combinations, the number of features required to achieve its results is higher than any other combination. Typical numbers range around 8 selected features with sporadic cases where only 4 or 5 features are needed. On the contrary, the rest of MCT schemata behave the opposite, usually selecting 4 or 5 features with sporadic cases where up to 7 features are needed.

Although less evident than in the symmetric case, patient-stratified symmetry still involves methodological concerns. The lack of presence of subjects in both training and testing sets prevents the results to be overoptimistic, but the sample database size is still artificially increased. Asymmetric corruption and patient-stratified symmetric corruption perform similarly, but a closer look reveals that whereas asymmetry tends to yield better mean metrics, coefficient of variation is usually better in the patient-stratified symmetric case, so there appears to be a trade-off. This can be explained by the multiple repetitions of a subject within training or testing sets, which lowers speaker variability.

The results obtained with asymmetric MCT indicate that this strategy is effective to achieve noise-robustness, since the maximum degradation in mean accuracy across the twelve cases with respect to the clean case is 12.6% for HUPA-polyps combination, followed by HUPA-nodules with 8.45% and most differences below 6%. Furthermore, the results shown when using patient-stratified symmetric MCT approach do not support a performance improvement. Therefore, asymmetric MCT is proposed as the most suitable strategy to follow, being also methodologically rigorous since it does not artificially increase the sample size.

Selected features and their significance play an important role in the outcomes of the experiments. Every feature and *feature family* considered in section IV-A has its own peculiarities, strengths and weaknesses. Some of them depend on non-acoustical characteristics present in the signal, like its length in the case of entropies. This question is solved by maintaining as much homogeneity as possible across recordings in all their "physical" aspects like length, sample rate or bit depth. Moreover, nonlinear analysis requires of a careful selection of hyperparameters in order to obtain

significant results which is addressed making use of some simple strategies found in literature [38], [39], [47].

An analysis of selected features from the experiments based on the two noise conditions that do not increase the sample size, clean and asymmetric MCT, and the three realistic databases, UEX_Voice, SVD, and HUPA, reveals which features are more reliable. Table 9 summarizes those features.

TABLE 9. Most selected features by subgroups.

Subgroup	Features	# experiments	# selections
Clean	CPP	9	6
	LZ2	9	4
	MFCC03	9	4
Asymmetric	CPP	9	4
	GNE_mean	9	4
	PERMUTATION	9	4
UEX_Voice	CPP	6	5
	MFCC03	6	5
SVD	CPP	6	3
	GNE_mean	6	4
HUPA	D2	6	6
Nodules	CPP	6	3
	PERMUTATION	6	3
Polyps	CPP	6	5
Reinke	GNE_mean	6	5

Subgroups identify which parameter is fixed and its value. For noise conditions we fix values clean and asymmetric and for each one of them we iterate over database (UEX_Voice, SVD, HUPA) and disease (nodules, polyps, Reinke). If we look at databases fixed values are UEX_Voice, SVD, HUPA and the counting is carried on noise condition (clean, asymmetric) and disease (nodules, polyps, Reinke). Finally, if we focus on diseases, fixing nodules, polyps, and Reinke's disease, we iterate over noise condition (Clean, asymmetric) and database (UEX_Voice, SVD, HUPA)

Although there is a variety of highlighted features, there are some common features being selected, which are, therefore, the most robust ones as they are valid in a wide range of conditions. Cepstral analysis seems to be very useful as it includes two features: CPP, which seems to be the most reliable, and MFCC03. Glottal-to-Noise excitation also appears in every situation (fixing noise condition, database, and disease). Nonlinear features are also present with PERMUTATION, D2, and LZ2, although the latter one only appears in clean cases.

Obtained from the cepstrum of a sound, CPP has been considered the most successful acoustic feature for vocal quality assessment [48]. High CPP values correspond to a well-defined harmonic structure, whereas periodicity perturbations (commonly present due to the considered pathologies) decrease their values. Being selected in 4 out of 9 experiments under asymmetric MCT, CPP feature seems to be still reliable under noisy conditions. It does not dominate the classification processes as in the clean case, but it is as important as the other two most repeated features (PERMUTATION and GNE_mean).

GNE estimates the excitation due to vocal fold oscillations versus the excitation created by turbulent noise. It uses the correlation of Hilbert envelopes of frequency channels uniformly distributed along the spectrum, and detects turbulent noise as narrow band noise. As our noise is not bandwidth limited, such detection can be performed efficiently. Furthermore, [49] considers GNE calculation robust because it does not require estimations of the fundamental frequency, which is a complicated task, encumbered by the pathological voice, and even more difficult to perform in the presence of environmental noise.

The capability of PERMUTATION to model the characteristics of a biological system even when there is contamination by noise is known from other biomedical applications, such as studies related to brain or heart activity [50]. Its robustness for the detection of benign vocal fold lesions under noisy conditions is demonstrated with this work as it mostly appears under asymmetric MCT.

Despite the fact they are not vocal source-related features, previous scientific work has considered the use of MFCCs for the detection of laryngeal pathologies. In [51] the authors report a lower first formant frequency of vocal polyp patients based on a higher tongue position during phonation, compared to healthy subjects. This means that, for subjects with a laryngeal disorder, also the shape of the vocal tract is changed during phonation.

The system does not select common features for a given disease for different MCT schema, as neither does MCT schemata comparison for different disease as well, if we do not take account of the database. MEEI database seems to prefer nonlinear characteristics, with MFSW or FZCF unlike the other databases, where they have a low number of appearances. UEX-Voice seems to prefer cepstral analysis with a great number of MFCCs, being selected, specifically MFCC3, on the top selected features. SVD concentrates a great number of selected features around Glottal-To-Noise Excitation ratio. For HUPA database entropies seem to be the best predictor. Therefore, apart from the fact that MEEI database metrics are better than those of any other one, database has a greater impact on selected features than disease or corruption.

Table 1 shows sex and age distribution for each combination of database and disease after subject selection process to create a balanced experiment, described before. Age usually does not constitute a problem as it is relatively easy to find pathological voices for each disease in a wide range of ages, as is shown by average and standard deviation values on table 1.

Gender on its side has shown to be a more important issue in voice pathology. Women are more prone to suffer from vocal fold diseases like Reinke's edema because of their vocal fold structure [52], but gender aspects also influence the acoustical feature values obtained in signal analysis [53]. This might explain the differences in feature selection among databases: although sex is never selected as a good predictor, the proportion of male/female subjects in both healthy and pathological recordings varies throughout databases. Table 1 shows an obvious female prevalence in all diseases, and

a female/male proportion that does not match for different database-disease combinations. The effect on acoustical features, feature selection, and therefore in classification task, although interesting, can not be usually addressed due to the imbalance [12].

There is a lack of comparable results due to the novelty of applying MCT to the specific field of voice diagnosis. Moreover, robustness assessment has not been thoroughly discussed beyond some specific pitch related features [54]. However, we can check our results against those obtained in studies that overlap in the use of similar parameters (database, disease, features and/or classifier) as a baseline.

Clean results from our classifiers stand a comparison with previous related work. Accuracy, specificity, sensitivity, precision, and AUC-ROC for MEEI database are shown in Table 4 for nodules (0.95, 0.87, 0.98, 0.93, 0.95), Table 6 for polyps (0.93, 0.82, 0.97, 0.93, 0.96), and Table 8 for Reinke's edema (0.96, 0.91, 0.98, 0.97, 0.99), and establish the baseline to which corruption results will be compared. This baseline is in the vicinity of results obtained in other MEEI research studies: [13] reaches accuracies between 0.91 and 0.97, specificities between 0.73 and 0.90, sensitivities between 0.94 and 0.98, and AUC-ROC between 0.89 and 0.98 diagnosing pathological voices with a feature set consisting of MFCCs, Energy, HNR, NNE, and GNE; [55] achieves 0.95 accuracy, 0.94 specificity, 0.95 sensitivity, and 0.99 ROC using HNR, Normalized Noise Energy (NNE), GNE, and 12 MFCCs with a GMM detector, and accuracies ranging 0.88 - 0.96, specificities ranging 0.87 - 0.98, sensitivities ranging 0.88 - 0.97, and AUC-ROC ranging 0.94 - 0.99, using other feature sets; [15] achieves 0.94 accuracy discriminating nodules and polyps among others. These results, although not directly comparable because of the discrepancies on methodology since they mix diseases, use other features or build a different classifier, consolidate our clean results as a good enough baseline to which compare MCT performance.

SVD and HUPA databases have been available for a shorter period of time, thus they have not been used as thoroughly as MEEI in research, making it more difficult to find comparable studies. However, some results match our accuracy levels. Reference [56] uses different combinations of features, including glottal source features, spectral and cepstral analysis (using MFCCs) to achieve 0.78 accuracy, 0.80 sensitivity, and 0.77 specificity when classifying healthy and pathological voices from HUPA, whereas for SVD yields 0.74 accuracy, 0.75 sensitivity, and 0.71 specificity. Reference [13] uses HMM to detect the pathological voices present in the dataset with accuracy, specificity, sensitivity, and AUC-ROC ranging 0.68 - 0.82, 0.53 - 0.83, 0.78 - 0.86 and 0.72 - 0.83 respectively, whereas we detect one pathology each time against a balanced normomorphic subset and our results coherently range 0.71 - 0.79, 0.68 - 0.80, 0.75 - 0.78, and 0.77 - 0.87.

Metrics analysis confirms the different performance of MEEI in relation to the other three databases. MEEI mean levels for clean baseline are all in the same range for every disease, with great accuracy, as expected, and sensitivity levels, and very good specificity. This is due to the different recording conditions for normal and pathological speakers, and subsequent selection of disease affected utterances.

Observed performance difference when applying the methodology to any other database comes undoubtedly from the recording conditions. An inter-database MCT analysis is interesting, as we can see how the performance for asymmetric training in MEEI is not comparable to clean training with UEX-Voice, SVD, and HUPA databases, what tells us that MEEI database collecting methodology, including strictly controlled environment along with screening and selection of the included recordings, makes pathological voices easily discernible. This is an issue that has already been addressed, and as such, should be only used as starting point, and for research where classification accuracy is not the main goal [57], which is the case.

MCT applied to speaker and speech recognition, fields where this work is inspired from, gets results that support the use of this technique in this scenario. Word accuracy in [58] drops approximately 1% when a MCT with a SNR of 20 dB is applied to the word recognition problem. Those results encourage us to further study the capabilities of this technique.

VII. CONCLUSION

We have studied the effects of MCT approach in voice disease detection from sustained vowel recordings. We made use of MEEI, UEX-Voice, SVD, and HUPA databases healthy samples and nodules, polyps, and Reinke's edema affected recordings. For every database-disease combination a set of random forest classifiers was trained under four conditions: clean, asymmetric corruption, symmetric corruption, and patient-stratified symmetric corruption and their ability to discriminate between healthy and pathological samples using a set of acoustic features extracted from each condition set was tested.

The noise used in the corruption strategies (asymmetric, symmetric, and patient-stratified symmetric) was chosen and added in a way that it accurately replicates the acoustical conditions that could be found in a typical clinical environment, either in its nature, selecting the appropriate sources, and in its relative level with respect to the specific recording.

Symmetric corruption adds all considered noises to every utterance in the database, performing also a data augmentation schema. That augmentation has a great influence in the outcome, leads to overoptimistic results if no further subject-stratification is performed, and raises methodological concerns due to the artificial increase of dataset size. If the classifier is trained using a patient-stratified schema, accuracy, specificity, and sensitivity values align with those obtained using clean and asymmetric strategies, though variance is usually lower.

Asymmetric corruption, which adds noise to randomly chosen samples from the database, causes only a small degradation in accuracy, specificity, and sensitivity in every case.

However, such degradation is small enough to consider the accuracy-robustness trade-off beneficial. Furthermore, preserving the dataset size makes this strategy the only one that does not raise any concerns about its validity. We strongly advise on using it as the methodology to be used in future research.

The effects mentioned in the two previous paragraphs have been observed in all databases, what allows us to consider that the results can be extrapolated to new unknown inputs. Further work would be necessary to check the consistency of results using larger voice datasets (number of samples per disease and multi-class classification) and increasing the number of noise conditions (noise types and amount of samples per type present in the noise database).

ACKNOWLEDGMENT

The authors would like to thank Dr. Moreno for his medical advising, Sandra Paniagua and Esther de la O. for their work recording the UEX-Voice database in HSPdA, and all the voluntary individuals, patients, and healthy subjects.

REFERENCES

[1] A. Hantzakos, M. Remacle, F. Dikkers, J.-C. Degols, M. Delos, G. Friedrich, A. Giovanni, and N. Rasmussen, "Exudative lesions of Reinke's space: A terminology proposal," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 266, no. 6, p. 869, 2009.

[2] L. Baghai-Ravary and S. W. Beet, *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*. New York, NY, USA: Springer, 2012.

[3] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art," *Biomed. Signal Process. Control*, vol. 51, pp. 181–199, May 2019.

[4] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *J. Voice*, vol. 33, no. 6, pp. 947.e11–947.e33, 2019.

[5] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis—Jitter, shimmer and HNR parameters," *Procedia Technol.*, vol. 9, pp. 1112–1122, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212017313002788, doi: 10.1016/j.protcy.2013.12.124.

[6] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomed. Signal Process. Control*, vol. 14, pp. 42–54, Nov. 2014.

[7] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb. 2004.

[8] Y. D. Heman-Ackah, R. T. Sataloff, G. Laureyns, D. Lurie, D. D. Michael, R. Heuer, A. Rubin, R. Eller, S. Chandran, M. Abaza, K. Lyons, V. Divi, J. Lott, J. Johnson, and J. Hillenbrand, "Quantifying the cepstral peak prominence, a measure of dysphonia," *J. Voice*, vol. 28, no. 6, pp. 783–788, Nov. 2014.

[9] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. Farahat Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2018.

[10] J. R. O. Arroyave, J. F. V. Bonilla, and E. D. Trejos, "Acoustic analysis and non linear dynamics applied to voice pathology detection: A review," *Recent Patents Signal Process.*, vol. 2, no. 2, pp. 96–107, Jul. 2012.

[11] E. S. Fonseca, R. C. Guido, P. R. Scalassara, C. D. Maciel, and J. C. Pereira, "Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders," *Comput. Biol. Med.*, vol. 37, no. 4, pp. 571–578, Apr. 2007.

[12] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal, "Towards robust voice pathology detection," *Neural Comput. Appl.*, vol. 32, pp. 15747–15757, Apr. 2018.

[13] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "An improved method for voice pathology detection by means of a HMM-based feature space transformation," *Pattern Recognit.*, vol. 43, no. 9, pp. 3100–3112, Sep. 2010.

[14] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, Oct. 2006.

[15] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 1938–1948, Sep. 2011.

[16] D. Hemmerling, A. Skalski, and J. Gajda, "Voice data mining for laryngeal pathology assessment," *Comput. Biol. Med.*, vol. 69, pp. 270–276, Feb. 2016.

[17] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41034–41041, 2018.

[18] A. Ben Aicha, "Contribution of data augmentation for the prenventive detection of vocal fold precancerous lesions," *Procedia Comput. Sci.*, vol. 159, pp. 212–220, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050919313559, doi: 10.1016/j.procs.2019.09.176.

[19] E. S. Fonseca, R. C. Guido, S. B. Junior, H. Dezani, R. R. Gati, and D. C. M. Pereira, "Acoustic investigation of speech pathologies based on the discriminative paraconsistent machine (DPM)," *Biomed. Signal Process. Control*, vol. 55, Jan. 2020, Art. no. 101615.

[20] M. Eye and E. Infirmary, *Voice Disorders Database, Version. 1.03 (CD-ROM)*. Lincoln Park, NJ, USA: Kay Elemetrics Corporation, 1994.

[21] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocol.*, vol. 36, no. 2, pp. 60–69, Jul. 2011.

[22] *Saarbrücken Voice Database*. Accessed: May 27, 2019. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de

[23] T. A. Mesallam, M. Farahat, K. H. Malki, M. Alsulaiman, Z. Ali, A. Al-Nasheri, and G. Muhammad, "Development of the Arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *J. Healthcare Eng.*, vol. 2017, pp. 1–13, Oct. 2017.

[24] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4257–4260.

[25] Y. Huang, M. Slaney, M. L. Seltzer, and Y. Gong, "Towards better performance with heterogeneous training data in acoustic modeling using deep neural networks," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1–5.

[26] M. S. Paniagua, C. J. Pérez, F. Calle-Alonso, and C. Salazar, "An Acoustic-Signal-Based preventive program for university Lecturers' vocal health," *J. Voice*, vol. 34, no. 1, pp. 88–99, Jan. 2020.

[27] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, and C. Ramírez-Calvo, "Acoustic analysis of voice using WPCVox: A comparative study with multi dimensional voice program," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 265, no. 4, pp. 465–476, Apr. 2008.

[28] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomed. Signal Process. Control*, vol. 1, no. 2, pp. 120–128, Apr. 2006.

[29] M. Pützer and W. J. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements," *Clin. Linguistics Phonetics*, vol. 22, no. 6, pp. 407–420, Jan. 2008.

[30] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, *arXiv:1510.08484*. [Online]. Available: http://arxiv.org/abs/1510.08484

[31] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *J. Speech, Lang., Hearing Res.*, vol. 37, no. 4, pp. 769–778, Aug. 1994.

[32] D. Michaelis, M. Fröhlich, and H. W. Strube, "Selection and combination of acoustic features for the description of pathologic voices," *J. Acoust. Soc. Amer.*, vol. 103, no. 3, pp. 1628–1639, Mar. 1998.

[33] A. Tsanas, "Acoustic analysis toolkit for biomedical speech signal processing: Concepts and algorithms," in *Models and Analysis of Vocal Emissions for Biomedical Applications*, vol. 2. Firenze, Italy: Firenze Univ. Press, 2013, pp. 37–40.
[34] A. Tsanas and P. Gómez-Vilda, "Novel robust decision support tool assisting early diagnosis of pathological voices using acoustic analysis of sustained vowels," in *Proc. Multidisciplinary Conf. Users Voice, Speech Sing.*, 2013, pp. 3–12.
[35] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phonetic Sci.*, vol. 17, no. 1193. Amsterdam, The Netherlands, 1993, pp. 97–110.
[36] P. Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," *J. Acoust. Soc. Amer.*, vol. 35, no. 3, pp. 344–353, Mar. 1963.
[37] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolanos, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Honig, and E. Noth, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 6, pp. 1820–1828, Nov. 2015.
[38] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, vol. 7. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[39] P. Henriquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Diaz-de-Maria, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1186–1195, Aug. 2009.
[40] E. A. F. Ihlen, "Introduction to multifractal detrended fluctuation analysis in MATLAB," *Frontiers Physiol.*, vol. 3, p. 141, Jun. 2012.
[41] M. Riedl, A. Müller, and N. Wessel, "Practical considerations of permutation entropy," *Eur. Phys. J. Special Topics*, vol. 222, no. 2, pp. 249–262, Jun. 2013.
[42] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomed. Eng. OnLine*, vol. 6, no. 1, p. 23, 2007.
[43] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.
[44] J. R. Orozco, J. F. Vargas, J. B. Alonso, M. A. Ferrer, C. M. Travieso, and P. Henriquez, "Voice pathology detection in continuous speech using nonlinear dynamics," in *Proc. 11th Int. Conf. Inf. Sci., Signal Process. Appl. (ISSPA)*, Jul. 2012, pp. 1030–1033.
[45] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 75–81, Jan. 1976.
[46] L. Naranjo, C. J. Pérez, Y. Campos-Roca, and J. Martín, "Addressing voice recording replications for Parkinson's disease detection," *Expert Syst. Appl.*, vol. 46, pp. 286–292, Mar. 2016.
[47] C. M. Travieso, J. B. Alonso, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, E. Nöth, and A. G. Ravelo-García, "Detection of different voice diseases based on the nonlinear characterization of speech signals," *Expert Syst. Appl.*, vol. 82, pp. 184–195, Oct. 2017.
[48] C. A. Ferrer Riesgo and E. Nöth, "What makes the cepstral peak prominence different to other acoustic correlates of vocal quality?" *J. Voice*, vol. 34, no. 5, pp. 806.e1–806.e6, Sep. 2020.
[49] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *J. Voice*, vol. 24, no. 1, pp. 47–56, Jan. 2010.
[50] M. Zanin, L. Zunino, O. A. Rosso, and D. Papo, "Permutation entropy and its main biomedical and econophysics applications: A review," *Entropy*, vol. 14, no. 8, pp. 1553–1577, Aug. 2012.
[51] J.-W. Lee, H.-G. Kang, J.-Y. Choi, and Y.-I. Son, "An investigation of vocal tract characteristics for acoustic discrimination of pathological voices," *BioMed Res. Int.*, vol. 2013, pp. 1–11, Jul. 2013.
[52] N. Çomunoğlu, C. S. Batur, and A. M. Önenerk, "Pathology of nonneoplastic lesions of the vocal folds," in *Voice and Swallowing Disorders*. Rijeka, Croatia: IntechOpen, 2019.
[53] M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding, "Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task," *J. Voice*, vol. 25, no. 1, pp. 44–53, Jan. 2011.
[54] D. D. Deliyski, H. S. Shaw, and M. K. Evans, "Adverse effects of environmental noise on acoustic voice quality measurements," *J. Voice*, vol. 19, no. 1, pp. 15–28, Mar. 2005.
[55] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 370–379, Feb. 2011.
[56] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 367–379, Feb. 2020.
[57] K. Daoudi and B. Bertrac, "On classification between normal and pathological voices using the MEEI-KayPENTAX database: Issues and consequences," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1–6.
[58] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR-Autom. Speech Recognit., challenges New Millenium ISCA Tutorial Res. Workshop*, 2000, pp. 1–8.

MARIO MADRUGA received the degree in computing engineering with a capstone project involving machine learning and automatic classification, in 2010, and the degree in electrical engineering with great interest in acoustics, in 2017. He is currently pursuing the Ph.D. degree with the Universidad de Extremadura. He joined the Department of Mathematics as a Research Assistant, where he started his Ph.D. degree in researching speech processing for biomedical applications. In his final and following years, he worked in several companies as a Developer and a Systems Administrator.

YOLANDA CAMPOS-ROCA received the M.S. and Ph.D. degrees in telecommunication engineering from the Universidade de Vigo, Spain, in 1994 and 2000, respectively. From 1996 to 2000, she performed research stays that accumulate almost three years at the Fraunhofer Institute for Applied Solid State Physics, Freiburg, Germany, where she has been a Guest Researcher from the University of Vigo or a Staff Member. In 2000, she joined the School of Technology, Universidad de Extremadura, Cáceres, Spain, as an Assistant Professor and becoming an Associate Professor in 2002. Her current research interests include circuit design in the microwave and millimeter-wave range and speech processing for biomedical applications.

CARLOS J. PÉREZ received the master's degree in mathematical science in 1996 and the Ph.D. degree in mathematics in 2003. He is currently a Full Professor of statistics with the Department of Mathematics, University of Extremadura, Spain. His main research interest includes the area of Bayesian statistical inference and classification. He has authored or coauthored more than 60 JCR-indexed journal articles about statistical methodology and applications in diverse knowledge fields, including computer aided diagnosis systems based on acoustic features extracted from voice recordings. He has participated in many research projects from competitive calls and contracts. He also has been a Reviewer for some journals as *Expert Systems with Applications*, *Reliability Engineering and Safety Systems*, *Journal of Applied Statistics*, or *Annals of Applied Statistics*.