On the design of automatic voice condition analysis systems. Part II: review of speaker recognition techniques and study on the effects of different variability factors.

J.A. Gómez-García[a,*], L. Moro-Velázquez[b,a], J.I. Godino-Llorente[a]

[a] *Universidad Politécnica de Madrid. Ctra. Valencia, km. 7, 28031. Madrid, Spain.*
[b] *Johns Hopkins University. Baltimore, Maryland 21218, USA.*

Abstract

This is the second of a two-part series devoted to the automatic voice condition analysis of voice pathologies, being a direct continuation to the paper "On the design of automatic voice condition analysis systems. Part I: review of concepts and an insight to the state of the art". The aim of this study is to examine several variability factors affecting the robustness of systems that automatically detect the presence of voice pathologies by means of audio registers. Multiple experiments are performed to test out the influence of the speech task, extralinguistic aspects (such as sex), the acoustic features and the classifiers in their performance. Some experiments are carried out using state-of-the-art classification methodologies often employed in speaker recognition. In order to evaluate the robustness of the methods, testing is repeated across several corpora with the aim to create a single system integrating the conclusions obtained previously. This system is later tested under cross-dataset scenarios in an attempt to obtain more realistic conclusions. Results identify a reduced subset of relevant features, which are used in a hierarchical-like scenario incorporating information of different speech tasks. In particular, for the experiments carried out using the Saarbrüecken voice dataset, the area under the ROC curve of the system reached 0.88 in an intra-dataset setting and ranged from 0.82 to 0.94 in cross-dataset scenarios. These results let us open a discussion about the suitability of these techniques to be transfered to the clinical setting.

Keywords: Robust Automatic Voice Condition Analysis, Universal Background Models, Extralinguistic Aspects of the Speech, Cross-Dataset Validation.

1. Introduction

Voice impairments arise due to misuse, infections, physiological or psychogenic causes, or due to the presence of other systematic disorders (including neurological), vocal abuse, surgery, trauma, congenital anomalies, irradiation, chemicals affecting vocal folds, etc. [1]. The classical approach to detect voice impairments consists on an instrumental (objective) and perceptual (subjective) evaluation, which are complemented by other types of examinations to determine the existence of a voice disorder and its grade of impairment. In order to assist medical specialists in the the diagnosis procedures, a field called *automatic*

*Corresponding author

voice condition analysis (AVCA) has arisen, providing advantages to traditional detection procedures such as objectiveness or non-invasiveness due to the use of speech signals.

During the first part of this review entitled *"On the design of automatic voice condition analysis systems. Part I: review of concepts and an insight to the state of the art"*, some relevant concepts regarding AVCA systems have been described, introducing the most widely employed methodologies that are found in the design of these automatic systems. This second paper explores the design of AVCA systems, carrying out a variety of tests with differing types of speech tasks, types of features and accounting for diverse variability factors. The aim is to design a single generalist AVCA system that is latter tested under cross-dataset scenarios. Some techniques that constitute state-of-the-art in speaker recognition systems and based on the idea of *Gaussian Mixture Models* (GMM) are also tested out. GMM are generative models that represent the probability density function of a training dataset by means of a linear combination of G multivariate Gaussian components. If the amount of training data is large, it is possible to accomplish a well-trained GMM representing the data; but when it is scarce, other approaches are preferred. In this respect, it is often useful to model, via GMM, a larger auxiliary dataset different to the training dataset. The resulting model is termed *Universal Background Model* (UBM) and serves as an initialisation which is then used to adapt specific -better trained and more generalist- models using the training data. These adapted models are termed GMM-UBM and have been widely employed in several speaker recognition tasks [2]. A variation to GMM-UBM is termed GMM-SVM, which is aimed at combining the discriminatory capabilities of SVM into the GMM framework [3]. Likewise, a further improvement to the GMM-UBM are the *i-Vectors* (IV) [4], which rely on the concept of GMM-UBM and factorial analysis for modelling the training dataset in a *total variability* space. IV are often accompanied by a *Probabilistic Linear Discriminative Analysis* (PLDA), which seeks to compensate for the effects of variability factors in the training data [5].

This paper is organised as follows: section 2 introduces the datasets and the methodological setup of the four major experiments followed in this paper; section 3 presents the obtained results; section 4 introduces some discussions, whereas section 5 presents some concluding remarks.

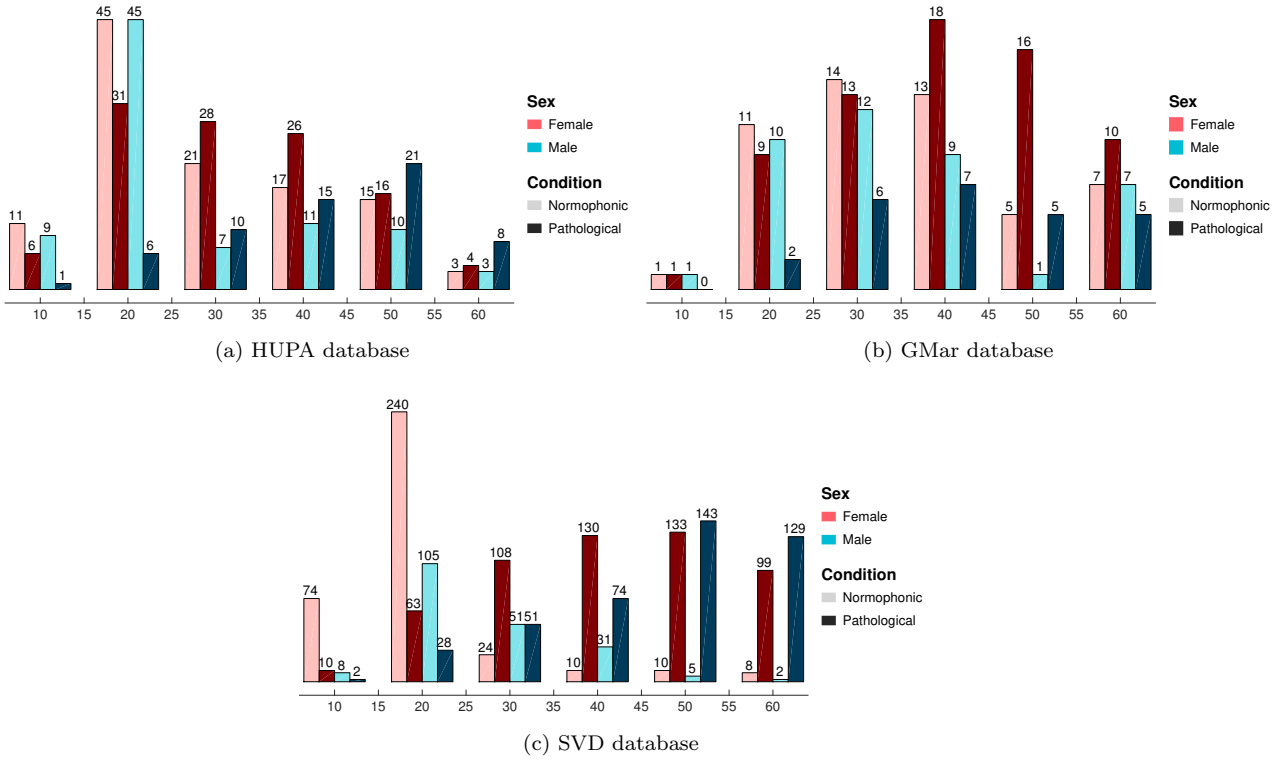
2. Experimental setup

This section presents the different experimental setups followed throughout the paper. The section begins with a description of the datasets used for training and testing the systems developed, and later presents four methodological frameworks used to test the influence of the speech task, the acoustic features, the classifiers and certain extralinguistic aspects.

2.1. Acoustic material

Three datasets containing normophonic and pathological recordings are used as training corpora: Hospital Universitario Príncipe de Asturias (`HUPA`), Hospital Gregorio Marañón (`GMar`) and Saarbrücken (`SVD`) voice disorders corpora. Similarly, four ancillary datasets are also utilised for the construction of the UBM. The

ancillary datasets include the EUROM and PhoneDat-I corpora which are composed of normophonic registers of speakers reading passages and pronouncing words, the well-known *Massachussets Ear and Eye Infirmary* (MEEI) partition of normophonic and dysphonic registers, and the Albayzin dataset which contains recordings of sentences uttered in Spanish. Additionally two extra corpora are used for cross-dataset trials: the *Hospital Doctor Negrín dataset* (DN) and the *Aplicación de las Tecnologías de la Información y las Comunicaciones* (ATIC) corpora. A brief description of each one of these datasets is presented next:



(a) HUPA database

(b) GMar database

(c) SVD database

Figure 1. Histograms representing the distribution of patients according to their sex and age for the (a) HUPA, (b) GMar and (c) SVD corpora.

2.1.1. HUPA dataset

Recorded at the Príncipe de Asturias hospital in Alcalá de Henares, Madrid, Spain [6]. The dataset contains the sustained phonation of the vowel /a/ of 366 adult Spanish speakers (169 pathological and 197 normophonic). Registers have been recorded using the *Kay Computerized Speech Lab Analysis station* 4300*B* with a sampling frequency of 50 kHz and 16 bits of resolution. Pathological voices contain a wide variety of organic pathologies including nodules, polyps, oedemas, carcinomas, etc. The distribution of registers according to the sex and age of the speaker is shown in Figure 1a.

2.1.2. GMar dataset

Composed of registers of Spanish speakers phonating the vowels /a/, /i/ and /u/. The dataset has been recorded at the Gregorio Marañón Hospital, Madrid, Spain, using the *MediVozCaptura* system [7] with a sampling frequency of 22050 Hz and 16 bits of resolution. The corpus is composed of 202 audio recordings, from which 95 are of normophonic and 107 of pathological speakers. The distribution of registers for the vowel /a/, according to the sex and age of the speakers is introduced in Figure 1b.

2.1.3. SVD dataset

It holds a collection of audio registers from more than 2000 normophonic and pathological German speakers [8, 9]. The dataset was recorded by the Institut für Phonetik at Saarland University and the Phoniatry Section of the Caritas Clinic St. Theresia in Saarbrücken, Germany. The corpus comprises recordings of the sustained phonation of vowels /a/, /i/ and /u/ uttered at normal, high and low pitch, as well as with rising-falling pitch. Besides, it incorporates recordings of the sentence *Guten Morgen, wie geht es Ihnen?* (*Good morning, how are you?*). Registers have been recorded using a sampling frequency of 50 kHz and 16 bits of resolution. For the purposes of this paper only the vowels /a/, /i/ and /u/ at normal pitch and the running speech recordings have been utilised, having defined a subset of the dataset after removing those registers with a low dynamic range or interferences. After this process, 1538 registers of speakers aged between 16 and 69 years are obtained (568 normophonic and 970 pathological). Figure 1c depicts the distribution according to the sex and age of the speakers in the dataset.

2.1.4. Ancillary datasets

As commented above, four ancillary datasets are used in this work. The first, the EUROM corpus, comprises recordings of 60 speakers in each of seven European languages: Danish, Dutch, British English, French, German, Norwegian and Swedish [10]. It has been explicitly designed to aid to the phonetic comparison of languages, with similar materials and recording protocols. In this manner, each corpus has been recorded at 20 kHz and 16 bit of resolution in an anechoic room and balancing the acoustic content among the different languages. In this paper only the German corpus is employed, for which a partition within the corpus called *Many Talker set* is used. It consists of 63 speakers which are asked to perform two tasks: (i) reading of 100 numbers in the range of 0 to 9999 grouped in 5 blocks of 20 numbers; (ii) reading of 5 sentences coming from 40 texts. The second, the Albayzin dataset, is a Spanish corpora designed for speech recognition purposes [11] and which is composed of three sets: phonetic, application and Lombard speech. Only the phonetic corpus is considered in this work. It contains 6800 recordings of 204 speakers uttering phonetically balanced phrases, which have been digitised using a sampling frequency of 16 kHz and 16 bits of resolution. The third, the PhoneDat-I dataset, contains 200 phonemically balanced artificial German sentences and two readings, namely *"the North wind"* fable and *"a Butter story"* [12]. The corpus contains around 20000 files uttered by 200 German speakers. All registers have been recorded at a sampling rate of 16 kHz and 16 bit of resolution. Finally, the MEEI voice disorders dataset contains 710 recordings of English speakers, phonating

the vowel /a/ and reading the first sentence of the *Rainbow Passage* [13]. The dataset has been recorded at sampling frequencies ranging from 10 to 50 kHz. To ensure a balance in the pathologies under study, a subset of the dataset is chosen in [14]. The resulting partition comprises 226 speakers: 173 pathological and 53 normophonic. The registers have been previously edited to remove the beginning and ending of each utterance, hence omitting the effects of vowel onsets and offsets.

2.1.5. Corpora for the cross-dataset trials

Two corpora are considered for cross-dataset trials. One one hand, the `DN` partition has been recorded by Hospital Dr. Negrín in Las Palmas de Gran Canaria, Spain. It contains 181 registers of Spanish speakers phonating the vowel /a/ [15]. The registers have been recorded at 22050 Hz and 16 bits of resolution. A partition of 130 registers has been randomly extracted and used for evaluation purposes.

On the other hand, the `ATIC` dataset contains recordings of 79 Spanish speakers (58 dysphonic and 21 normophonic) phonating the vowel /a/ [16]. Pathological voices have been obtained from public and private Otorhinolaryngology services in Málaga, Spain, whereas normophonic speakers are recorded from teachers and students recruited at Málaga University. Registers have been recorded in quiet rooms under controlled conditions, digitized at 16 bits and 44100 Hz.

2.2. Methodological setup

Four main experiments are carried out in this paper. The aim is to explore the influence of different variability factors that typically affect the performance of AVCA systems. Additionally, the paper studies the influence on the performance of different classification techniques often employed in the speaker recognition field.

2.2.1. Experiment 1: variability due to the acoustic material and the feature set

This experiment contains a series of trials using a variety of feature sets and acoustic material. This allow a direct comparison about the influence of the speech task in AVCA systems, as well as insight about the consistency of the considered features when the acoustic material varies. Likewise, and since no single feature can offer enough discrimination power to completely characterise the properties of dysphonia and normophonia, this allows to study the characteristics rendering the best results in subsequent classification labours. This might potentially permit the design of AVCA systems based on complementary features describing distinct properties of dysphonia.

Trials are performed using all the available acoustic material of the `HUPA`, `SVD` and `GMar` corpora. It is worth noting that not all the acoustic features, which will be mentioned next, can be extracted from *running speech*, as they rely on conditions -such as stationarity- that cannot be always met. For instance, certain perturbation features depend on the existence of a periodic glottal excitation, assumption that cannot be fulfilled on unvoiced segments of speech. As an alternative to directly analyse running speech, literature

often reports procedures based on retrieving voiced segments of speech using voiced-unvoiced algorithms [17, 18]. This methodology is considered as well in the following trials.

A total of 9 trials are carried out in *experiment 1*: 1 trial using the `HUPA` dataset (sustained phonation of the vowel /a/); 3 trials with `GMar` (sustained phonations of vowels /a/, /i/, /u/); and 5 with `SVD` (sustained phonations of vowels /a/, /i/, /u/, raw running speech, and voiced segments of the running speech task). The setup followed during each trial is presented in Fig. 2a and can be described as follows:

I. *Preprocessing:* all registers are downsampled to 20 kHz (the lowest sampling frequencies of all the available datasets) and *max-normalised* by dividing the signal by its absolute largest value. Then, a framing and a Hamming windowing procedure is followed, where the length of the window is determined depending on the feature sets that are extracted, and the overlap is varied to ensure that all the sets of characteristics contain the same number of frames.

II. *Characterisation:* this stage has the goal of extracting features capable of portraying the properties of normophonic and dysphonic conditions. The idea is to extract for each windowed frame, a d-dimensional vector of characteristics, $\vec{x} = \{x[1], \cdots, x[d]\}$ which is associated to a label ℓ, describing properties of the segment. The sets of characteristics that are considered in this paper are descriptors of vocal quality. These are illustrated in Table 2b and are described next:

- *Perturbation features* (*Pert* set): measure the presence of additive noise resulting from an incomplete glottal closure of the vocal folds, and the presence of modulation noise which is the result of irregularities in the movements of the vocal folds. Three perturbation metrics are employed: *Normalised Noise Entropy* (NNE) [19], *Cepstral Harmonics-to-Noise Ratio* (CHNR) [20] and *Glottal-to-Noise Excitation Ratio* (GNE) [21].
- *Spectral and cepstral features* (*SCs* set): measure the harmonic components of the voice. This set of features encloses *Perceptual Linear Prediction coefficients* (PLP) [22], *Mel-Frequency Cepstral Coefficients* (MFCC) [23], *Smoothed Cepstral Peak Prominence* (CPPS) [24] and *Low-to-High Frequency Spectral Energy Ratio* (LHr) [25]. The number of PLP and MFCC coefficients are varied in the range [10, 20] with steps of 2.
- *Features based on modulation spectrum* (MSs set): rely on the computation of the modulation spectrum, that characterises the modulation and acoustic frequencies of input voices [26]. The features considered in the set are: *Modulation Spectrum Homogeneity* (MSH), *Cumulative Intersection Point* (CIL), *Rate of Points above Linear Average* (RALA) and *Modulation Spectrum Percentile* (MSP)m, where the sub-index is referred to the percentile that is used, i.e. MSP_{25}, MSP_{75} and MSP_{95} [27, 28].
- *Complexity* (*Comp* set): this family of parameters characterises the dynamics of the system and its structure. Several features are extracted, which are further grouped according to the properties they measure. Hence: (i) the first subset comprises *dynamic invariants* (*Dyn* subset) extracted

from a reconstructed attractor such as the *Correlation Dimension* (D2), the *Largest Lyapunov Exponent* (LLE) [29], and the *Recurrence Period Density Entropy* (RPDE) [30]; (ii) the second subset contains features which measure *long-range correlations* (*LR* subset), such as *Hurst Exponent* (He) and *Detrended Fluctuation Analysis* (DFA) [31, 30]; (iii) the third subset includes *regularity estimators* (*Reg* subset) which are based on entropy-like quantifiers. It encompasses *Approximate Entropy* (ApEn) [32], *Sample Entropy* (SampEn) [33], *Modified Sample Entropy* (mSampEn) [34], *Gaussian Kernel Sample Entropy* (GSampEn) [35] and Fuzzy Entropy (FuzzyEn) [36]; (iv) finally, the fourth subset includes *entropy estimators* (*Ent* subset) such as *Permutation Entropy* (PE) [37, 38], *Rényi Hidden Markov Model Entropy* (rHMMEn) and Shannon Hidden Markov Model Entropy (sHMMEn) [23, 39].

For the trials based on *sustained phonation*, Hamming windows of 40 ms are employed for the Pert and SCs sets to ensure that each frame contains at least one pitch period, whereas windows of 55 ms length are used in the Comp sets as suggested in [23]. Likewise, for experiments in the MSs set, segments of 180 ms are utilised as in [27, 28].

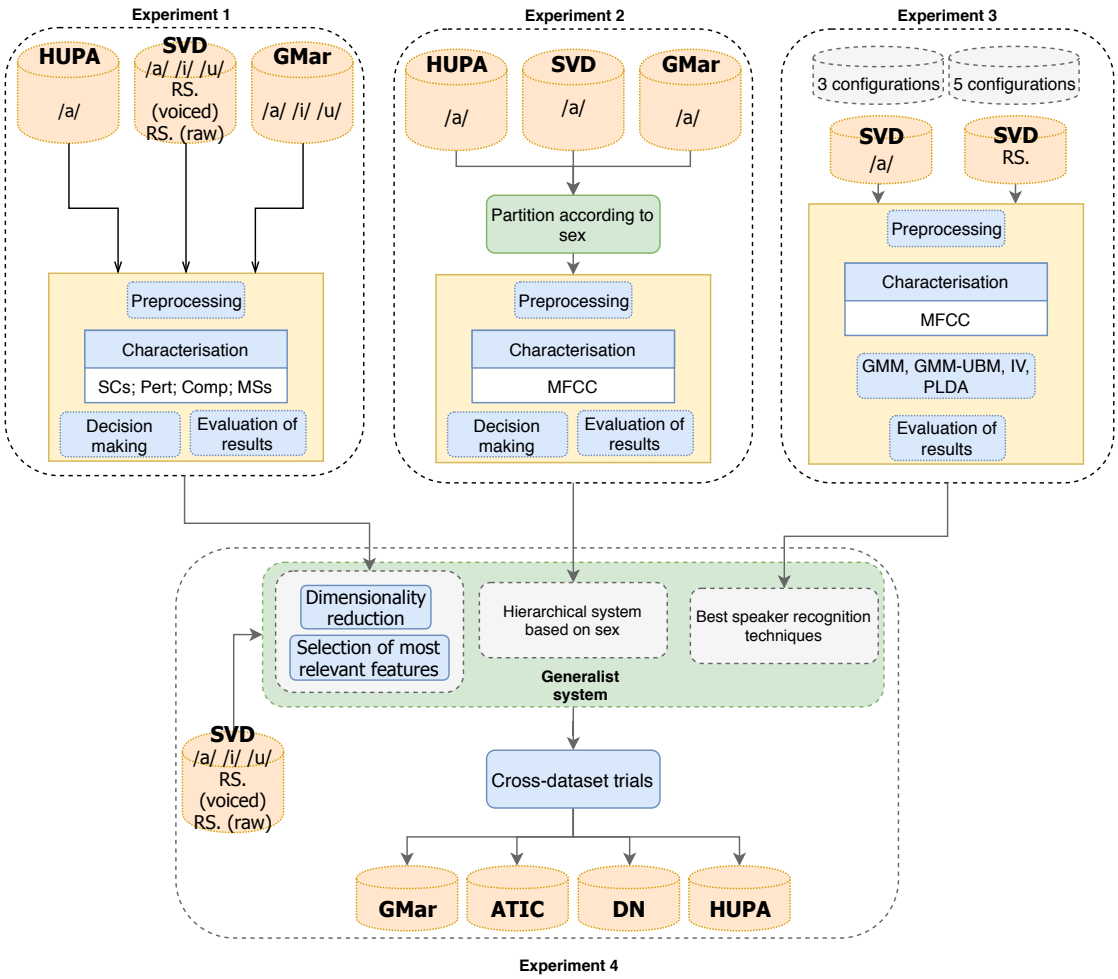
For those trials involving *running speech*, two different scenarios are considered. In the first, only segments representing voice are extracted from the speech. These segments are extracted automatically by means of the MAUS software, as described in [40]. The second makes use of the raw running speech registers, but using window lengths in the range [10, 30] ms with 5 ms steps. These values are selected to avoid violating the stationarity assumptions. A voice activity detector has been also employed to reject unwanted segments [41].

III. *Decision making:* Given a training set of observations $\boldsymbol{\mathcal{X}} = \{\vec{x}_1, \cdots, \vec{x}_n, \cdots, \vec{x}_N\}$, associated to a label vector $\vec{\ell} = \{\ell_1, \cdots, \ell_n, \cdots, \ell_N\}$, the aim of the decision making procedure is to learn a mapping from the input observations to the labels. GMM classifiers are used for modelling class membership before taking decisions using log-likelihood functions.

IV. *Evaluation of results: The Area Under the ROC curve* (AUC) is the preferred measure of performance evaluation. A k-folds cross-validation procedure has been employed setting the number of folds to $k = 11$. *Detection-Error Tradeoff* (DET) curves and standard performance metrics for a binary classifier are considered for each one of the aforementioned sets of features: *Accuracy* (ACC), *Sensitivity* (SE), *Specificity* (SP).

2.2.2. Experiment 2: design of a hierarchical system based on the sex of the speaker

The effects of the extralinguistic trait of sex in the design of AVCA systems is explored. The interest of this approach in AVCA is in the possibility of constructing hierarchical-like schemes taking into account the variability introduced by certain extralinguistic factors separately. The idea is to segment the population according to extralinguistic characteristics to train more accurate models. As a result, the complexity of the



(a) Methodological stages of the AVCA system.

Set	Subset			Features		
Pert	–	NNE	CHNR	GNE		
SCs	–	PLP	MFCC	CPPS	LHr	
MSs	–	MSH	CIL	RALA	MSP	
Comp	Dyn	D2	LLE	RPDE		
	LR	He	DFA			
	Reg	ApEn	SampEn	mSampEn	GSampEn	FuzzyEn
	Ent	rHMMEn	rHMMEn	FuzzyEn		

(b) Feature sets and subsets used for characterisation.

Figure 2. (a) methodology followed in this paper for the design of the AVQA system; (b) table containing all the considered sets of features.

voice pathology detection task is decreased, simpler signal processing models are generated, and a subsequent performance improvement is expected. To analyse the impact of the sex of the speakers in detection tasks,

a *sex-independent* and a *sex-dependent* system (an AVCA for female speakers and another for males) are designed.

Three trials are carried out using the partitions of the vowel /a/ belonging to the HUPA, SVD and GMar dataset. The setup followed for all the trials is presented in Figure 2a, whereas the main stages are described next:

1. *Preprocessing:* all registers are downsampled to 20 kHz and max-normalised. For the sex-dependent systems the datasets are decomposed according to sex of the speaker.
2. *Characterisation:* Only MFCC features are extracted in the current experiment. The number of MFCC coefficients is varied in the range [10, 20] with steps of 2. Hamming windows of 40 ms with a 50% of overlapping between consecutive frames are employed.
3. *Classification:* decision making is carried out using a GMM classifier, varying the number of Gaussian components as $\{2^i\} : i \in \mathbb{Z}; 1 \leq i \leq 9$.
4. *Evaluation of results:* the same metrics used in the experiment 1 are considered.

2.2.3. Experiment 3: testing out the performance of classification techniques used in speaker recognition

The performance of classifiers relying on the idea of UBMs in conjunction with MFCC features, a setup that has been proven useful in speaker recognition tasks -but that has not been widely explored in AVCA systems- is studied. In this experiment only the registers of the vowel /a/ and running speech belonging to the SVD partition are examined, thus defining two trials. In the one using sustained phonation, three configurations are defined in accordance to the auxiliary datasets that are used to train the UBM and compensation models: (i) *configuration* C_1 employs normophonic registers of HUPA corpora; (ii) *configuration* C_2 uses normophonic and dysphonic recordings of HUPA, MEEI and GMar (vowel /a/ only); (iii) *configuration* C_3 employs normophonic data of HUPA, MEEI, GMar and vowels extracted from EUROM to compare the influence of using acoustic material coming from different speech tasks (running speech, sustained vowels /a/, /i/ or /u/). With respect to EUROM, vowels are used to match as much as possible the acoustic content of the SVD partition. Since there is access to the phonological transcription of the audio files, these segments are extracted automatically by means of the MAUS software, as described in [40].

For the trial using running speech, diverse ancillary datasets are employed for training the UBM and compensation models: (i) *configuration* C_1: employs normophonic and dysphonic recordings of the vowels /a/, /i/, /u/ of the HUPA and GMar datasets, plus normophonic vowels extracted from the EUROM and Albayzin corpora; (ii) *configuration* C_2: considers normophonic-only recordings of the vowels /a/, /i/, /u/ of the HUPA and GMar datasets, plus normophonic vowels extracted from the EUROM and Albayzin corpora; (iii) *configuration* C_3: characterises the normophonic sentences of MEEI, EUROM and Albayzin; (iv) *configuration* C_4: characterises the normophonic sentences of EUROM; (v) *configuration* C_5: characterises the normophonic sentences of EUROM and PhoneDat-I.

Table 1 summarises the trials and the configurations in *experiment 3*, whereas the experimental setup is introduced graphically in Figure 2a and described as follows:

1. *Preprocessing:* for the trial involving *sustained phonation*, 40 ms Hamming windows with a 50% of overlapping between consecutive frames are employed, whereas for the running speech experiments, the length of the window is defined according to the best outcomes of *experiment 2* (where lengths were varied in the range [10, 30] ms with steps of 5 ms).
2. *Characterisation:* MFCC features are extracted from each one of the frames obtained in the preprocessing stage. The number of MFCC coefficients is varied between [10, 20] with steps of 2 coefficients.
3. *Decision machines:* GMM, GMM-UBM, IV, PLDA and GMM-SVM classifiers are employed. The number of Gaussian components is varied in such a manner that $\{2^i\} : i \in \mathbb{Z}; 1 \leq i \leq 9$.
4. *Evaluation of results:* the same procedure followed in previous experiments is carried out.

Table 1. Tested configurations for training the UBM models in *experiment 3*.

	Configuration	Datasets	Speech tasks	Content
Sustained phonation	C_1	HUPA	/a/	No.
	C_2	HUPA, MEEI, GMar	/a/	No. + Dy.
	C_3	HUPA, MEEI, GMar, EUROM	/a/+/i/+/u/+RSv.	No.
Running speech	C_1	HUPA, GMar, EUROM, Albayzin	/a/+/i/+/u/+RSv.	No. + Dy.
	C_2	HUPA, GMar, EUROM, Albayzin	/a/+/i/+/u/+RSv.	No.
	C_3	MEEI, EUROM, Albayzin	RSr.	No.
	C_4	EUROM	RSr.	No.
	C_5	EUROM, PhoneDat-I	RSr.	No.

RSr.: raw running speech; RSv.: vowels extracted from running speech; No.: normophonic; Dy.: dysphonic

2.2.4. Experiment 4: combination of the best systems.

This experiment is built around the lessons learnt during the first three experiments. Presumably it is to be of a hierarchical type, it will use a subset of informative features, and will employ all the available acoustic material to provide a single decision about the condition of speakers.

Two trials are considered in the current experiment: (i) One which provides a single decision about the condition of patients by combining the results of the systems based on the vowels /a/, /i/, /u/ and the running speech task of the SVD dataset; (ii) other designed to test the capabilities of the system in a cross-dataset scenario. In particular, the system is assessed using the HUPA, ATIC, DN and GMar corpora.

The methodological stages that are followed in both trials are presented in Figure 2a and are described next:

1. *Preprocessing:* all registers are downsampled to 20 kHz and max-normalised. for the trials involving

sustained phonation, 40 ms Hamming windows with a 50% of overlapping are used. For those using *running speech*, the length of the window is defined according to the outcomes of *experiment 1*.

2. *Characterisation:* two systems are considered. The first is designed with those features providing the best results in *experiment 1*. To this end, three dimensionality reduction methods are employed to rank features from the most to the least relevant, in the search for a decrease in the computational burden, an improvement in the accuracy of the analyses, and the avoidance of problems such as the curse of dimensionality [42]. This analysis is performed independently for the `HUPA`, `SVD` and `GMar` datasets by means of three selection techniques [43]: *Maximal Information Maximisation* (MIM), *Minimal Redundancy Maximal Relevance* (mRMR) and *Joint Mutual Information* (JMI).

   To generalise results across datasets, acoustic material and features, a scoring procedure is now performed. In this manner and with the results of the ranking techniques and for a certain dataset, the scoring procedure rewards the best features with a low score, while penalizing the worst with a large value. These scores are then summed up across datasets and feature selection techniques. At the end, the features with the lowest scores are regarded as the most informative and consistent and are employed for further testing.

   In addition to this system, another is designed with MFCC features as it is to be used for the speaker recognition classification techniques. If the hierarchical system is build up (in accordance to results of *experiment 2*), the number of MFCC coefficients is varied in the range [10, 20] with steps of 2, otherwise, the parameter that render the best results in *experiment 1* is used.

3. *Decision machines:* GMM classifiers are used to train the system employing the most consistent set of features. The MFCC system utilises the classifiers that render the highest efficiency as given by the results of *experiment 3*. The number of Gaussian components is varied in such a manner that $\{2^i\} : i \in \mathbb{Z}; 1 \leq i \leq 9$.

4. *Fusion of results:* logistic regression is employed to fuse the system using the most consistent features and a GMM classifier, and the system based on MFCC and classification based on UBM. A further fusion is considered to combine the information coming from diverse speech tasks, and thus, to provide a single decision about the condition of speakers.

5. *Evaluation of results:* the same evaluation procedures followed in previous experiments is carried out.

3. Results

3.1. Experiment 1: variability due to the acoustic material and the type of features

In total 9 trials are performed: one using the vowel /a/ in `HUPA`; three for the vowels /a/, /i/, /u/ in `GMar`; and three for the vowels /a/, /i/, /u/, and two using running speech in the `SVD` corpus.

Set	Subset	ACC	SP	SP	AUC
Pert	-	76.61 ± 4.30	0.77	0.77	0.85
MSs	-	71.77 ± 4.57	0.74	0.70	0.79
SCs	CPPS+LHr	62.10 ± 4.93	0.60	0.64	0.69
	MFCC	69.62 ± 4.67	0.66	0.74	0.79
	PLP	66.94 ± 4.78	0.58	0.77	0.80
Comp	LR	56.18 ± 5.04	0.51	0.62	0.60
	Dyn	65.59 ± 4.83	0.63	0.68	0.75
	Reg	69.89 ± 4.66	0.68	0.72	0.78
	Ent	75.00 ± 4.40	0.75	0.75	0.83

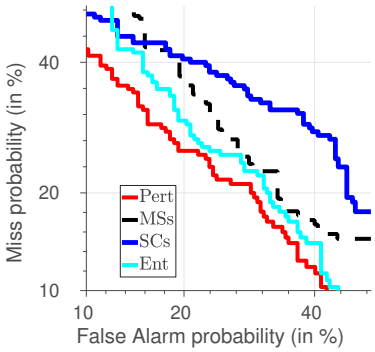


Figure 3. *Experiment 1*: results for the vowel /a/ in the HUPA database.

3.1.1. HUPA dataset:

The performance metrics of all the considered sets of features and the DET curves of the best performing subsets, using the vowel /a/ of the HUPA dataset are presented in Figure 3.

From the obtained outcomes it can be observed that the best results -in terms of AUC- are provided by the Pert set (0.85), whereas in the SCs set, MFCC and PLP achieve an almost equivalent performance (AUC of 0.80 and 0.79 respectively). Likewise, the subset Ent achieves the best results within the Comp set (0.83) while LR provides the worst (0.60).

3.1.2. GMar dataset:

The performance metrics of all the considered sets of features and the DET curves of the best performing subsets, for the vowels /a/, /i/ and /u/ in the GMar dataset are presented in Figure 4.

Respecting the vowel /a/, the best outcomes arise when using the Ent subset in Comp (AUC=0.83). Similarly, MSs, Pert and the whole SCs set provide acceptable outcomes, with AUC ranging from 0.73 to 0.77. By contrast, and for the vowel /i/, MSs achieves the best results (0.76), followed by the Pert set and the Reg subset, which accomplish in both cases an AUC of 0.73. Finally, and when the vowel /u/ is studied, the best results are given by the LR and the Ent subsets (AUC of 0.74 and 0.73 respectively). Notwithstanding, the PLP subset and the MSs set provide comparable results (0.73 and 0.72 respectively).

3.1.3. SVD dataset

The performance metrics of all the considered sets of features and the DET curves of the best performing subsets, for the vowels /a/, /i/ and /u/ are presented in Figure 5a.

Considering the vowel /a/, the best efficiency arises with the Pert set (AUC=0.78), followed by MFCC and PLP whose performance is alike (0.77 and 0.76 respectively). Within the Comp set, Ent and Reg provides the best performance (0.75 and 0.74 respectively). With regards to the vowel /i/, MFCC and PLP features accomplish the best results in the trial (AUC of 0.76 and 0.75 respectively), followed by other good performing sets such as Pert and MSs (AUC of 0.72 and 0.70 respectively). In a similar fashion, the results

		Vowel /a/				Vowel /i/				Vowel /u/			
Set	Subset	ACC	SP	SE	AUC	ACC	SP	SE	AUC	ACC	SP	SE	AUC
Pert	-	65.35 ± 6.56	0.65	0.65	0.77	66.32 ± 6.72	0.65	0.68	0.73	61.36 ± 7.19	0.64	0.59	0.70
MSs	-	67.82 ± 6.44	0.65	0.70	0.76	67.82 ± 6.44	0.65	0.70	0.76	63.64 ± 7.11	0.63	0.64	0.72
SCs	CPPS+LHr	66.83 ± 6.49	0.67	0.66	0.73	60.53 ± 6.95	0.62	0.59	0.67	59.09 ± 7.26	0.58	0.60	0.68
	MFCC	69.31 ± 6.36	0.69	0.69	0.77	60.53 ± 6.95	0.61	0.60	0.65	62.50 ± 7.15	0.62	0.63	0.68
	PLP	68.81 ± 6.39	0.67	0.70	0.76	58.42 ± 7.01	0.60	0.57	0.65	67.05 ± 6.94	0.70	0.64	0.73
Comp	LR	56.18 ± 5.04	0.51	0.62	0.60	61.05 ± 6.93	0.59	0.64	0.65	61.36 ± 7.19	0.64	0.59	0.74
	Dyn	65.59 ± 4.83	0.63	0.68	0.75	64.74 ± 6.79	0.63	0.67	0.70	58.52 ± 7.28	0.52	0.64	0.70
	Reg	69.89 ± 4.66	0.68	0.72	0.78	61.58 ± 6.92	0.48	0.75	0.73	63.07 ± 7.13	0.60	0.66	0.68
	Ent	75.00 ± 4.40	0.75	0.75	0.83	60.53 ± 6.95	0.61	0.60	0.68	63.07 ± 7.13	0.67	0.59	0.73

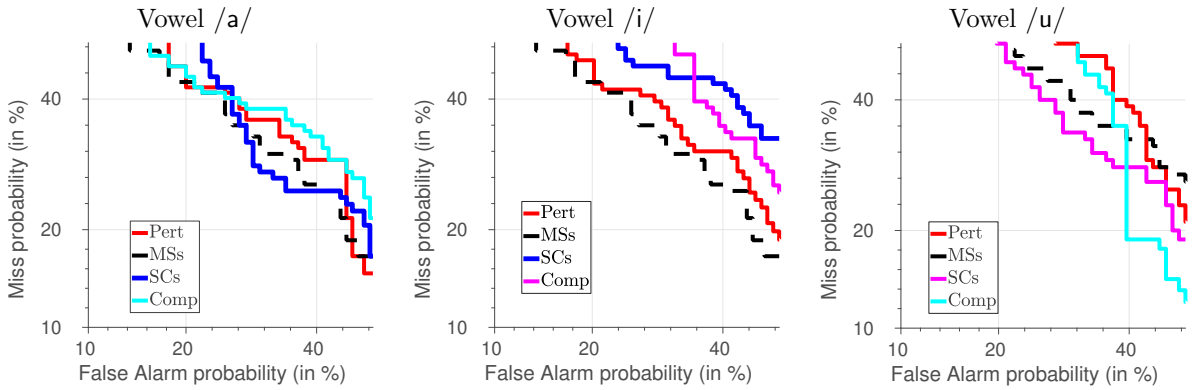


Figure 4. *Experiment 1:* results for the vowels /a/, /i/ and /u/ in the GMar database.

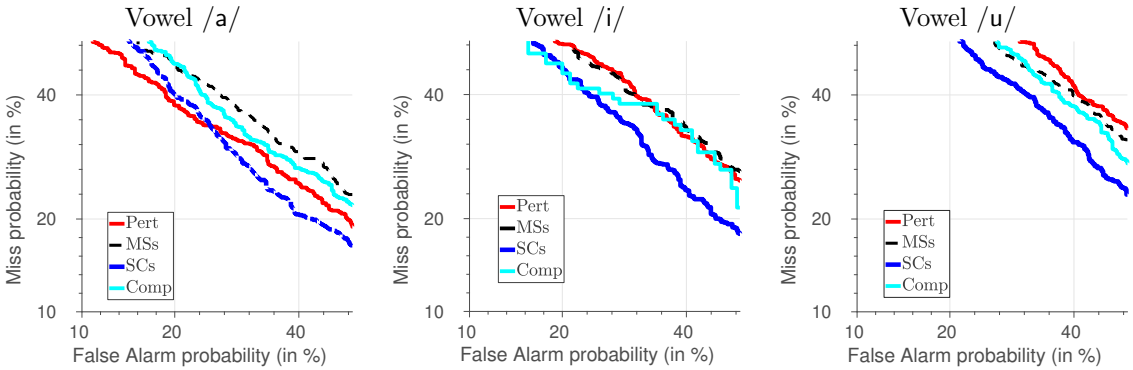
for the vowel /u/ crown MFCC and PLP as one of the best performing features when vowels are considered, but closely followed by the remaining.

Figure 5b presents the performance metrics of all the considered sets of features and the DET curves of the best performing subsets in the SVD dataset, for the running speech task using the raw acoustic material; whereas Figure 5c introduces the results when using vowels extracted from the running speech registers.

With regards to the trial using the raw speech, it is observed that both PLP and MFCC features perform almost equivalently (0.85 for PLP and 0.86 for MFCC). In a similar way, the window length that achieves the best results is in this case of 20 ms.

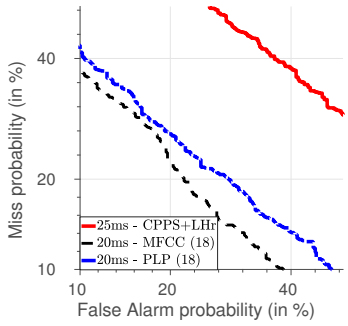
Respecting the trial using extracted vowels from running speech, it is worth noting that it was not possible to employ the same characterisation stage as in when using sustained phonation, as there are some practical reasons that difficult this type of analysis. In particular, the window length that is employed in some sets of features is too long to allow the extraction of characteristics. This is expected, as running speech is composed of fast transitions between silence, voiced or unvoiced sounds and intervals of sustained phonation are not produced frequently. Indeed, for characterisation with the MSs set it is necessary to include voiced segments with a window size of 180 ms. However, there are not vocal segments with such length in the

		Vowel /a/				Vowel /i/				Vowel /u/			
Set	Subset	ACC	SP	SE	AUC	ACC	SP	SE	AUC	ACC	SP	SE	AUC
Pert	-	68.79 ± 2.32	0.68	0.69	0.78	64.37 ± 2.39	0.64	0.65	0.72	59.88 ± 2.45	0.59	0.60	0.66
MSs	-	66.67 ± 2.35	0.60	0.70	0.73	63.20 ± 2.41	0.64	0.63	0.70	61.05 ± 2.44	0.53	0.66	0.67
SCs	CPPS+LHr	61.12 ± 2.44	0.61	0.61	0.68	57.87 ± 2.47	0.55	0.59	0.62	58.13 ± 2.47	0.55	0.60	0.63
	MFCC	70.48 ± 2.28	0.67	0.73	0.77	68.73 ± 2.32	0.67	0.69	0.76	65.41 ± 2.38	0.64	0.66	0.71
	PLP	71.07 ± 2.27	0.70	0.72	0.77	68.21 ± 2.33	0.67	0.69	0.75	64.69 ± 2.39	0.62	0.66	0.71
Comp	LR	61.31 ± 2.43	0.59	0.63	0.68	59.04 ± 2.46	0.57	0.60	0.61	58.06 ± 2.47	0.57	0.59	0.64
	Dyn	63.78 ± 2.40	0.62	0.65	0.73	62.29 ± 2.42	0.62	0.63	0.68	61.51 ± 2.43	0.58	0.64	0.68
	Reg	67.75 ± 2.34	0.55	0.75	0.74	63.46 ± 2.41	0.62	0.65	0.69	63.78 ± 2.40	0.50	0.72	0.67
	Ent	68.08 ± 2.33	0.68	0.68	0.75	61.25 ± 2.43	0.60	0.62	0.65	59.95 ± 2.45	0.57	0.62	0.66



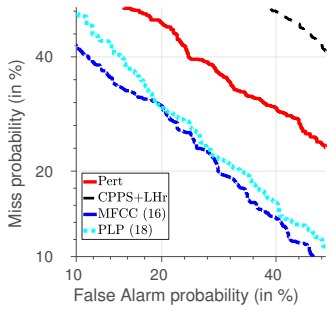
a. Results for vowels /a/, /i/ and /u/

		Running speech - raw speech			
Set	Length	ACC	SP	SE	AUC
CPPS+LHr	25 ms	62.26 ± 2.46	0.60	0.64	0.67
MFCC (18)	20 ms	80.32 ± 2.02	0.74	0.84	0.86
PLP (18)	20 ms	77.90 ± 2.11	0.65	0.85	0.85



b. Results using the raw running speech task.

	Running speech - extracted vowels			
Set	ACC	SP	SE	AUC
Pert	66.69 ± 2.39	0.67	0.67	0.74
CPPS+LHr	56.15 ± 2.52	0.52	0.58	0.58
MFCC (16)	76.96 ± 2.14	0.70	0.81	0.84
PLP (18)	75.55 ± 2.18	0.67	0.80	0.82



c. Vowels extracted from running speech.

Figure 5. *Experiment 1:* results for the SVD database using (a) the vowels /a/, /i/ and /u/; (b) the raw running speech task; (c) vowels extracted from the running speech task.

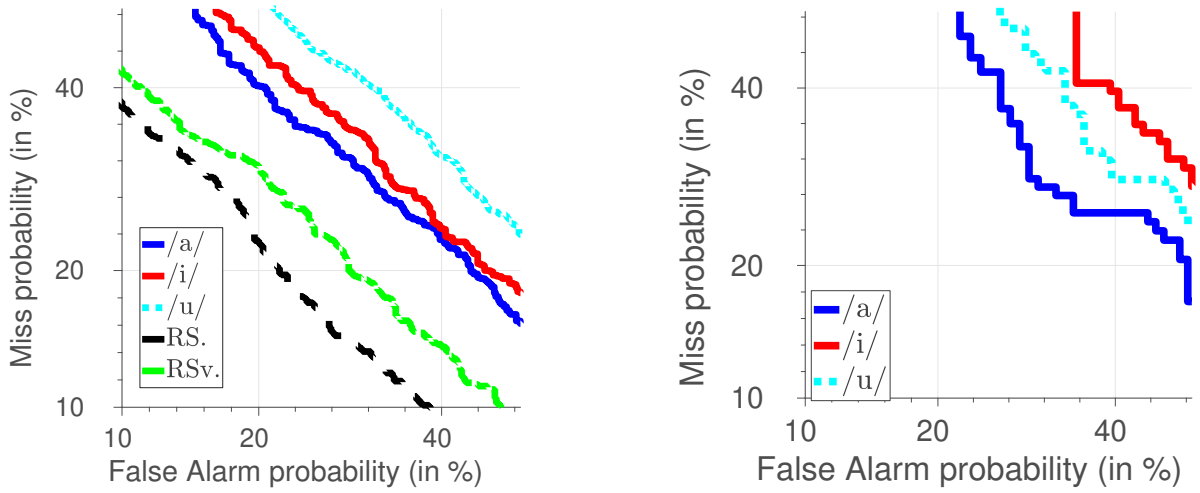
SVD dataset. This behaviour is also encountered when analysing complexity features, where the number of segments whose length is 55 ms is greatly restricted. For this reason, the reported results only employ sets using windows of 40 ms, i. e., Pert and SCs to characterise voiced segments of speech. The outcomes indicate that the MFCC features provide the best results (AUC = 0.86) next to PLP (AUC = 0.85). In comparison with the outcomes of the running speech trial, there is a light decrease of performance when using the vowels extracted from the speech.

3.1.4. General comments

During this experiment several sets of features have been employed for detection tasks, using different datasets of sustained vowels and running speech. By virtue of the obtained results, we can infer that there is no single feature or set of features that always perform better than the others. Indeed, depending on the type of acoustic material or the particularities of the corpus, a certain set of features outperforms the others. Despite that, there are some interesting observations indicating tendencies of good performance. For instance, Pert has proven to be useful for detection purposes as ascertained by the large AUC values that varied from 0.66 using the vowel /u/ and SVD, to 0.85 for the HUPA dataset (or to 0.86 if running speech is considered). In a similar fashion, the SCs set provides large AUC values that ranged from 0.71 for the vowel /u/ and the SVD dataset, to 0.8 for the HUPA dataset. In most of the cases (except for the vowel /i/ of the GMar dataset), either PLP or MFCC outperformed the results of CPPS+LHr. Regarding the MSs set, the performance of the whole subset remains among the highest compared to the remaining sets, with an AUC that varies from 0.67 for the vowel /u/ and the SVD corpus, to 0.79 using HUPA. Finally, and respecting the complexity features, the results indicate that the subset providing the worst general performance is LR as in almost all cases (but for GMar and vowel /u/), it is surpassed by the remaining subsets. Similarly, the performance of Reg and Ent tends to be superior to that of Dyn.

In reference to the acoustic material employed, some trends are observed as well. Revisiting the results of the trials using sustained phonation, it can be inferred that the vowel /a/ achieves, in all cases, better results than vowels /i/ and /u/, no matter the set that is analysed. When including running speech into the analysis, it is found that the spectral/cepstral features are in general better when the sentence is employed in comparison to the use of a sustained phonation. By contrast, for the Pert set, the vowel /a/ provides a better performance than when using running speech. In an attempt to compare speech material of different nature with the same set of features, Figure 6 presents the best DET curves for all the trials involving the SVD and GMar corpora. MFCC features are depicted as they are calculated in all tests, including those based on running speech. The outcomes suggest that -despite the difficulties of comparing results- there is a certain tendency to favour the employment of acoustic material based on running speech when spectral/cepstral analysis is considered. In addition, in both cases the vowel /a/ presents a better performance in comparison to using vowels /i/ or /u/.

It is important to highlight that a a fair comparison between speech tasks is difficult, as there are certain differences in the amount of acoustic material that is available when speech and voice are compared. As



(a) DET curve for the SVD dataset.

(b) DET curve for the GMar dataset.

Figure 6. *Experiment 1:* DET curves using MFCC features and speech material of different nature belonging to (a) SVD dataset; and (b) GMar dataset. RS.: running speech; RSv.: vowels extracted from running speech.

a matter of example, and taking the SVD dataset, the trial using the raw running speech material includes 219.000 frames into the analysis, whereas 36.691 voiced frames are encountered when including only vowels extracted from speech. By contrast, the analysis of the sustained phonation of the vowel /a/ is conformed by 95.422 frames.

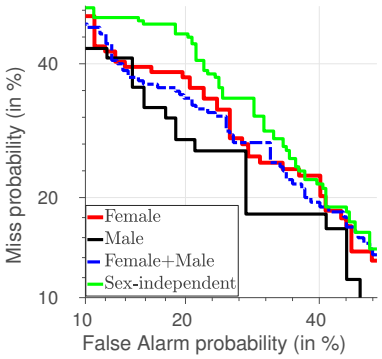
In a direct comparison of the two detection tasks using running speech, it is found that better results are obtained when no voiced detector is included. This might be simply a consequence of having used an automatic system for the segmentation of the voice recordings, as some errors might be introduced during this process, specially when pathological voice is considered. In any case, disregarding the use of voiced detectors comes with the added benefit of reducing the complexity of the AVCA system, at expenses of having to process irrelevant information.

3.2. Experiment 2: Effects of extralinguistic traits.

Figure 7 introduces the performance metrics of the sex-independent and sex-dependent systems trained for each one of the employed datasets.

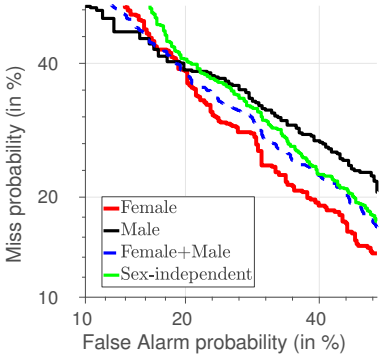
Results indicate that in the HUPA dataset, there is a tendency to a performance enhancement when the sex of the speaker is accounted in the classification (AUC=0.81 vs. AUC=0.79). This is expected since modelling female and male systems separately lead to a larger AUC (0.85 and 0.81 respectively) compared to a sex-independent scenario (0.79). The outcomes of the trial involving SVD also suggest that, in comparison to a sex-independent system, there is a certain tendency towards an efficiency improvement when the sex is deemed in the design of AVCA systems (AUC=0.78 when the sex is considered vs. AUC=0.77 when it is not). Unlike the previous trial, the female-only models achieve higher detection rates than the ones provided by the male-only models (AUC=0.79 vs. AUC=0.77). In the case of GMar, the same tendency of the previous

	Subtype	ACC	SP	SE	AUC
	Fe.+Ma.	73.66 ± 4.47	0.72	0.70	0.81
S.D.	*Fe.*: MFCC(20)	73.45 ± 5.76	0.71	0.75	0.81
	Ma.: MFCC(20)	73.97 ± 7.12	0.69	0.80	0.85
S.I.	MFCC(12)	69.62 ± 4.67	0.71	0.77	0.79



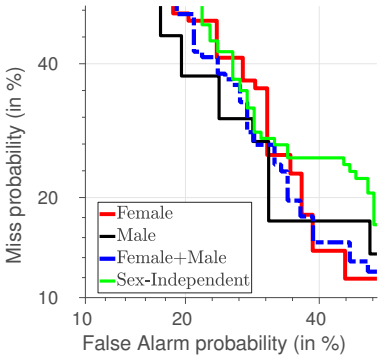
a. Results for the `HUPA` database.

Type	Subtype	ACC	SP	SE	AUC
	Fe. + Ma.	70.74 ± 2.27	0.68	0.73	0.78
S.D.	*Fe.*: MFCC(18)	72.17 ± 2.91	0.70	0.73	0.79
	Ma.: MFCC(16)	68.68 ± 3.62	0.63	0.71	0.77
S.I.	MFCC(18)	70.48 ± 2.28	0.67	0.73	0.77



b. Results for the `SVD` database.

Type	Subtype	ACC	SP	SE	AUC
	Fe.+Ma.	70.79 ± 6.27	0.72	0.70	0.78
S.D.	*Fe.*: MFCC(16)	70.45 ± 7.78	0.69	0.72	0.77
	Ma.: MFCC(10)	71.43 ± 10.58	0.76	0.66	0.82
S.I.	MFCC(20)	69.31 ± 6.36	0.69	0.69	0.77



c. Results for the `GMar` database.

Figure 7. *Experiment 2*: performance metrics and DET curves of the sex-dependent (*S.D.*) and sex-independent (*S.I.*) system using the (a) `HUPA`, (b) `SVD` and (c) `GMar` corpus. *Fe.*:Female, *Ma.*:Male.

two trials is observed, i. e., the model considering the sex of the speakers outperforms the sex-independent system (AUC=0.78 vs. AUC=0.77). Just as with the `HUPA` dataset, the results of the male-only models outperform those of the female-only ones (AUC=0.82 and AUC=0.77 respectively).

Despite the simplicity of the trials, some interesting observations arise. On one hand, and respecting

the inclusion of sex information, the outcomes indicate that accounting for the speaker's sex improves the efficiency in pathology detection tasks. Indeed, in the three datasets there is an absolute performance improvement which varied between 4% (using `HUPA`) to 0.3% (using `SVD`) compared to the sex-independent detector. It is worth noting that, in general, the best sex-independent and the best female/male detection systems establish a distinct number of MFCC coefficients. Ultimately, this may be a symptom of having decomposed the problem according to the speaker's sex that induce to different "optimal" operation points, i. e., the number of coefficients maximising performance for female models is not necessarily the same for male ones or when both sexes are considered in conjunction. This might be explained by the significant differences in the vocal tract and vocal folds of male and female speakers, which in turn has consequences on spectral characteristics of speech. Since MFCC characterise these spectral properties, it is reasonable to find these differing operation points for female and male models. Another interesting observation is that the systems trained with female voices generally performed worse than those trained with male data; occurring in two trials except when the `SVD` dataset was considered. This phenomenon has long been recognised in other speech-based applications, where a decreased performance is often achieved in systems based on spectral/cesptral characterisation and female speech [44]. Within the AVCA field, authors in [45], have also reported a reduced performance of systems based on female data. A possible explanation of this phenomenon, might be related to the differences in f_0 between male and females which has consequences in the the spectral/cepstral analyses. Similarly, the presence of a glottal gap which occurs more frequently for female speakers, and which is correlated to breathiness, might be a counfounding factor as well. Notwithstanding, it is difficult to conclude as there is not enough evidence supporting such behaviour and there is even a negative result about this particular.

3.3. Experiment 3: testing out the performance of classification techniques used in speaker recognition

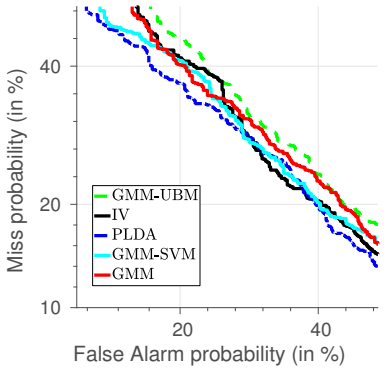
The current experiment explores the usefulness of classification techniques employed in speaker recognition in labours of pathological detection. Two trials are carried out using sustained phonation and running speech of the `SVD` dataset. Different configurations, presented in Table 1, are followed to train the UBM-based classifiers.

3.3.1. Trial using sustained phonation

Three configurations are tested out by varying the type of acoustic material that is employed for training the UBM and the compensation models in the IV and PLDA schemes, allowing the examination of interesting scenarios: (i) *configuration* C_1, a short amount of normophonic data is employed; (ii) *configuration* C_2 the amount of normophonic data is increased; (iii) *configuration* C_3 normophonic and dysphonic registers are utilised in conjunction. The best obtained results for each one of the tested configurations are depicted in Figure 8a.

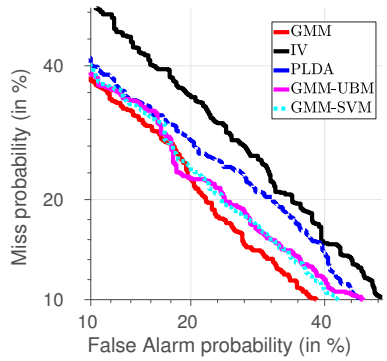
From the obtained outcomes it can be observed that a GMM-UBM model does not provide any further improvement with respect to the baseline GMM. However, when using more complex systems that stand

Configuration	Classifier	ACC	SP	SE	AUC
–	GMM	70.48 ± 2.28	0.67	0.73	0.77
C_1	GMM-UBM	69.27 ± 2.37	0.68	0.70	0.76
	IV	67.15 ± 2.41	0.66	0.68	0.75
	PLDA	71.66 ± 2.31	0.70	0.73	0.79
	GMM-SVM	71.18 ± 2.32	0.69	0.73	0.77
C_2	GMM-UBM	69.20 ± 2.37	0.69	0.69	0.76
	IV	68.38 ± 2.38	0.66	0.70	0.75
	PLDA	72.76 ± 2.28	0.72	0.73	0.79
	GMM-SVM	72.01 ± 2.30	0.69	0.74	0.77
C_3	GMM-UBM	68.24 ± 2.39	0.67	0.69	0.75
	IV	70.77 ± 2.33	0.71	0.71	0.78
	PLDA	71.32 ± 2.32	0.69	0.73	0.80
	GMM-SVM	71.73 ± 2.31	0.70	0.73	0.78



a. Trial using sustained phonation.

Configuration	Classifier	ACC	SP	SE	AUC
–	GMM	80.32 ± 2.02	0.74	0.84	0.86
C_1	GMM-UBM	76.70 ± 2.15	0.75	0.78	0.84
	IV	71.79 ± 2.29	0.70	0.73	0.79
	PLDA	74.01 ± 2.23	0.74	0.74	0.82
	GMM-SVM	77.03 ± 2.14	0.75	0.78	0.85
C_2	GMM-UBM	78.44 ± 2.09	0.72	0.82	0.86
	IV	73.00 ± 2.25	0.71	0.74	0.81
	PLDA	75.62 ± 2.18	0.75	0.76	0.85
	GMM-SVM	78.17 ± 2.10	0.76	0.80	0.85
C_3	GMM-UBM	78.44 ± 2.09	0.72	0.82	0.86
	IV	73.00 ± 2.25	0.71	0.74	0.81
	PLDA	75.62 ± 2.18	0.75	0.76	0.85
	GMM-SVM	78.71 ± 2.08	0.76	0.80	0.86
C_4	GMM-UBM	76.96 ± 2.14	0.74	0.79	0.84
	IV	72.73 ± 2.26	0.72	0.73	0.81
	PLDA	75.62 ± 2.18	0.74	0.76	0.84
	GMM-SVM	77.77 ± 2.11	0.71	0.81	0.86
C_5	GMM-UBM	78.64 ± 2.08	0.73	0.82	0.85
	IV	70.65 ± 2.31	0.69	0.71	0.79
	PLDA	76.09 ± 2.17	0.74	0.77	0.83
	GMM-SVM	79.25 ± 2.06	0.76	0.81	0.86



b. Trial using running speech.

Figure 8. *Experiment 3*: performance of the GMM-based classifiers in the detection of pathologies using the `SVD` dataset: (a) for the partition of vowel /a/ and the three tested conditions; (b) for the partitions using running speech.

on the idea of UBM models some subtle improvements are attained. In particular two considerations can be made. First, the PLDA provides the best efficiency in terms of AUC (0.80). Second, the best results almost always involve the use of normophonic-only registers for training UBM and compensation models (`MEEI`, `GMar`, `HUPA` and `EUROM`). Indeed, increasing the amount of normophonic material provides performance improvements as ascertained by comparing the results of *configuration* C_1 (which only employs one dataset of normophonic recordings) to *configuration* C_3 (which employs 4 dataset of normophonic registers).

3.3.2. Trial using running speech

The current trial employs the running speech partition of the `SVD` corpus, and different combinations of ancillary datasets that define five configurations that allow the examination of interesting scenarios: (i) *configuration* C_1, normophonic and dysphonic vowels plus voiced segments extracted from running speech are used; (ii) *configuration* C_2, normophonic data of sustained phonations and normophonic vowels extracted from running speecha are employed; (iii) *configuration* C_3, normophonic sentences are considered; (iv) *configuration* C_4, normophonic sentences uttered in the same language as `SVD` are utilised; (v) *configuration* C_5, similar to the latter but increasing the amount of registers used for training. The best results for each one of the configurations are presented in Figure 8b.

Despite the best absolute results -as ascertained by the DET curves and the AUC value- are given by the simple GMM classifier, it is worth comparing the influence of the ancillary datasets in the performance of speaker recognition classification techniques. In this respect, *configuration* C_3 provides the best performance, indicating that the inclusion of acoustic material based on sustained phonation does not contribute to any enhancement (see *configuration* C_2). By contrast, *configuration* C_1 attains, generally, the lowest AUC. This suggests that using normophonic and dysphonic registers for training the UBM and compensation models does not improve the results. *Configuration* C_4 is intended to test the behaviour of the system when the auxiliary dataset matches the language of the target partition, while *configuration* C_5 is similar to *configuration* C_4 but employing two auxiliary datasets. However, again, no further improvements in efficiency are found.

3.4. Experiment 4: Combination of the best systems.

The current experiment is aimed at designing AVCA systems with the insight obtained from the previous trials. The idea is to employ the best set of features in *experiment 1*, considering the hierarchical models of *experiment 2* and the speaker recognition techniques of *experiment 3*. The acoustic material of different sources is then fused by means of logistic regression.

Following this approach, a strategy based on feature selection is followed to rank the most consistent and generalist set via MIM, mRMR, JMI and a scoring procedure of the features analysed in *experiment 1*. The top-10 best characteristics for each dataset, and the top-10 global best set after having used the scoring procedure are presented in Table 2.

Table 2. *Experiment 4:* top-ranked features from each dataset.

HUPA	SVD	GMar	Global best
GNE	CHNR	RALA	GNE
PE	GNE	GNE	CHNR
RALA	RALA	CPPS	RALA
CHNR	PE	CHNR	PE
MSP_{25}	DFA	He	CPPS
LLE	He	LLE	LLE
CPPS	MSP_{95}	CIL	MSP_{95}
MSP_{95}	LLE	GSampEn	DFA
MSH	MSH	DFA	He
ApEn	GSampEn	PE	CIL

Results indicate that GNE, CHNR and RALA are the most consistent features amongst datasets and speech tasks. PE and CPPS are also proficient, but in the case of the first it is just ranked in 10-*th* position when using the GMar corpus, whereas the latter is not even included in the ranking of the SVD. As a result, an AVCA based solely on GNE, CHNR and RALA for characterisation and GMM classifiers to output decision scores is considered. In accordance to the results in *experiment 2*, the proposed systems should account for the sex of the speakers in a hierarchical-like categorisation procedure. Finally, the information of the most consistent features is fused at score level with the one provided by the UBM-based classifiers in *experiment 3*. Regarding the latter, and when using sustained phonation, PLDA systems trained with normophonic data are employed. For the case of running speech, the GMM algorithm is employed as it has been shown to provide better results than more complex schemes. The whole procedure is summarised graphically in Figure 9.

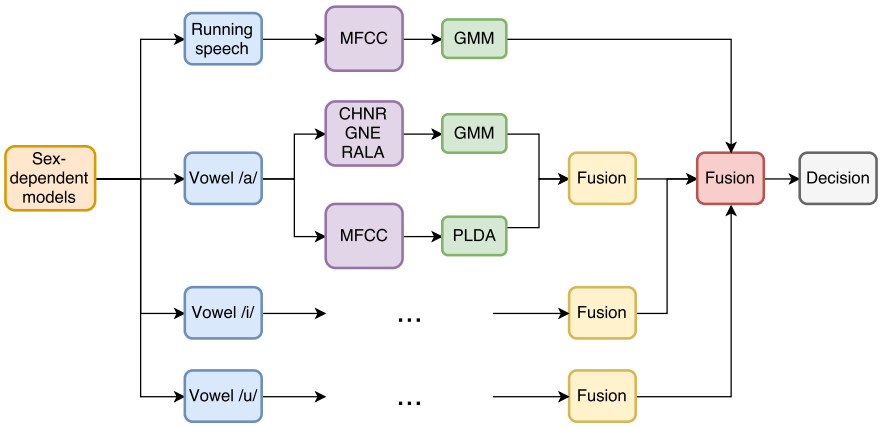
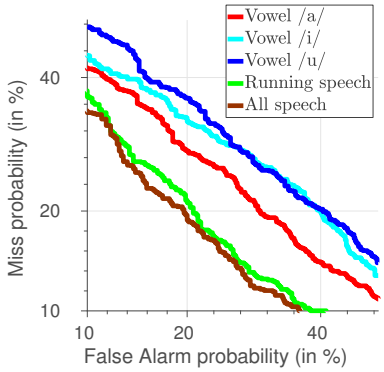


Figure 9. *Experiment 4:* resulting methodology after having considered the outcomes of *experiments 1*, *experiment 2* and *experiment 3*.

3.4.1. Trial in a intra-dataset scenario

The results after considering the methodology in Figure 9, using the vowels /a/, /i/, /u/, and the sentence in the SVD dataset are introduced in Figure 10a.

	ACC	SP	SE	AUC
Vowel /a/	75.42 ± 2.19	0.74	0.76	0.84
Vowel /i/	72.33 ± 2.27	0.70	0.73	0.81
Vowel /u/	72.20 ± 2.28	0.71	0.73	0.80
R.S.	80.66 ± 2.01	0.70	0.86	0.87
All speech	81.40 ± 1.98	0.71	0.87	0.88



a. Results of the experiment in an intra-dataset scenario.

Database	Vowel	ACC	SP	SE	AUC
GMar	/a/	72.00 ± 7.19	0.60	0.83	0.82
	/i/	62.67 ± 7.74	0.47	0.77	0.75
	/u/	72.67 ± 7.13	0.71	0.74	0.79
	All	74.00 ± 7.02	0.64	0.83	0.82
HUPA	/a/	74.46 ± 4.43	0.65	0.85	0.87
ATIC	/a/	78.21 ± 9.16	0.90	0.74	0.93
DN	/a/	82.87 ± 5.49	0.76	0.89	0.94

b. Results of the experiment in cross-dataset scenarios.

Figure 10. *Experiment 4*: Performance metrics for the trials in: (a) a intra-dataset scenario; (b) cross-dataset scenario.

The outcomes indicate that the fusion of different speech tasks increases performance, achieving in the best case scenario an AUC of 0.88 and a DET curve that is better than the remaining in all operation points.

3.4.2. Trial in a cross-dataset scenario

The current trial is focused on training an AVCA system with data of SVD and the methodology in Figure 9, but following a cross-dataset scenario tested in other four partitions: (i) vowels /a/, /i/ and /u/ of GMar; (ii) the vowel /a/ in HUPA; (iii) registers of the vowel /a/ in ATIC; (iv) and recordings of the vowel /a/ in DN. The best results for all the tested partitions are presented in Table 9b.

In general terms the results are acceptable, with an AUC varying between 0.75 to 0.94. When the vowel /a/ is considered AUC is always superior to 0.82. In particular for the GMar corpus, no performance enhancement is obtained when all the vowels are fused.

4. Discussion

Results indicate that a segmentation according to the sex of the speaker improves the performance, which in absolute terms varies between 0.3% to 4% depending on the dataset. These results are in line with those found in literature, where the classification accuracy of an automatic detector of pathology is lightly improved by using a manual segmentation of the dataset according to the sex of the speakers [46]. Results suggest that partitioning the dataset according to this extralinguistic criterion improves performance, indicating the usefulness of hierarchical systems that decompose the voice pathology detection problem into smaller sub-problems. Indeed, it has been found that the best operation points for female and male systems differ. This is expected as the vocal tract and vocal folds of both sexes differ, which in turn produces subsequent changes in spectral characteristics that consequently impacts the AVCA systems. Accounting for these differences separately simplifies the detection problem and provides performance enhancements.

With respect to the trials involving different sets of features, it has been demonstrated that, as expected, no single measurement characterises entirely the phenomena related to dysphonia. This suggests that multidimensional approaches should be followed to design AVCA systems. Outcomes indicate that measures based on perturbation and cepstrum produce positive results in detection of pathologies. Indeed, as it has been ascertained during *experiment 4*, the most consistent features are GNE, RALA and CHNR (along with CPPS and PE). The feasibility of these features has also been extensively demonstrated by the individual classification they provide, suggesting the feasibility of using this type of features for detection purposes. Interestingly, these characteristics come from different contexts. However, a performance improvement is usually obtained when they work in conjunction, evidencing its complementarity.

In reference to the *experiment 3* involving the speaker recognition classification techniques, it has been found that these classifiers have performed well when sustained phonation is used along with ancillary data of normophonic sustained phonation. By contrast, none of the datasets has contributed to enhance the performance when using running speech. In this regard, it is worth noting that the `SVD` dataset is composed of a single sentence that is uttered by German speakers, whereas `Albayzin` uses diverse sentences produced by Spanish speakers and `MEEI` contains registers of a text uttered in English. This mismatch might have affected the results as phonetics differs among languages. In an attempt to examine an scenario on which both ancillary and target coprora share the same language, *configuration* C_4 and *configuration* C_5 have been included. Notwithstanding, results indicate that not even under this setting the performance improved. In this particular case, we might attribute the diminished performance to the mismatch between the acoustic content of `SVD`, and that of `EUROM` and/or `PhoneDat-I`. A premise in the speaker recognition field, is that when the lexical content employed for the ancillary corpora matches with the one used for enrolment, a better performance is generally obtained [47]. When there is a mismatch, it is often necessary to include more data to generate more robust models that operate under a text-independent scenario. This is the reason for which the trial using sustained phonation of `SVD` reported positive outcomes as the ancillary dataset is composed of sustained vowels, matching with the lexical information of the `SVD` partition. Moreover, since

sustained phonation is relatively unaffected by language this effect is minimised. Finally, it is worth to remark that a common tendency that is observed in both trials is that better results are often obtained when normophonic data is used to construct models. A hypothesis explaining this, might be related to normophonic phonation having a more compact representation in the feature space (as there does not exist variability due to pathologies), which in turn permits a better adjustment of the initialisation models of the UBM systems. It is still necessary, though, to study the influence of dysphonic recordings in the training of compensation models. This also highlights the need of having more data available for the study of AVCA systems. Hence, translating methodologies used in speaker recognition (such as those based on UBM classifiers), should start by considering a translation of equivalent datasets to the AVCA field.

Outcomes also highlight the necessity of larger and better balanced corpora that fully describe variability factors such as age and sex, as these seem to affect AVCA systems. There is a large room for improving the systems presented in the current experiment, including a more thorough study of other types of features and classifiers correlating to ageing voices and/or to sexual conditions. In addition, other factors influencing performance should also be investigated, including accents, vocal effort, etc. An additional aspect that might be worth to consider, is in the study of methodologies that include more effectively extralinguistic information into the system, be it in the form of an a-priori probability or in a regression-like scheme.

5. Conclusions

The state of the art in AVCA systems reports lots of works evaluated under well controlled scenarios, providing results that suggest that the problem is almost solved. However, despite of these promising results, these systems are still far away from the clinical setting, because their accuracies are obtained in laboratory conditions and are quite dependent on the dataset used. With the aim to present more truthful results, this paper contributes with a more realistic baseline system, which can be used in the future for comparison purposes. To this respect, this paper analizes several variability factors affecting the robustness of these systems. Multiple experiments were performed to test out the influence on the performance of the speech task, extralinguistic aspects (such as sex), the acoustic features and the classifiers. The methodology followed has been developed to obtain dataset-independent results, so that they could be extrapolated to other corpora. In this sense, an analysis of several factors affecting the design of automatic voice pathology detection systems has been presented. Extralinguistics such as sex have been studied in hierarchical-like schemes, along with different sets of features in a variety of datasets containing speech material of diverse types. Moreover, speaker recognition classification techniques have been explored, fed with an acoustic material based on both sustained phonations and running speech.

Results demonstrate that including extralinguistic information regarding the sex of the speaker enhances the performance of an automatic detector of pathologies, suggesting that decomposing the pathology detection problem into sub-problems according to a certain extralinguistic criterion, decreases the complexity of the underliying models and therefore improves the efficiency of the system.

In reference to the types of speech tasks that have been evaluated, results suggest that the sustained vowel /a/ always achieve better results in comparison to other types of vowels. Notwithstanding, these results are outperformed when cepstral analysis and running speech is considered. Results also suggest that fusing the acoustic material of different speech tasks (vowels /a/, /i/, /u/, running speech) -via logistic regression- improves the performance of the system even further. This fusion stage has demonstrated its utility to combine information of heterogeneous but complementary systems, into a single decision machine that behaves better than the individual systems.

The speaker recognition classifiers have moderately improved results for those cases where sustained vowels are used in conjunction with ancillary corpora based on normophonic phonations. By contrast, nothing can be concluded regarding the use of running speech and these classifiers, as the phonetic mismatch between target and ancillary corpora has presumably affected the resulting UBM-based classifiers. Moreover, due to the unavailability of datasets containing dysphonic registers of running speech, an analysis on the effects of this acoustic material on the training process of UBM models remains an open issue. Despite that, it is safe to say that the results reported in this paper, involving the usage of UBM-based classifiers and running speech are not positive.

Besides, outcomes suggest that no single parameter is capable of completely characterising vocal pathology, and therefore multidimensional approaches are needed to enhance classification results. This fact has been reported through the analysis of several sets of features and speech tasks. The most coherent features, as ascertained in several trials, are two estimators of perturbation noise and one descriptor of dispersion based on the modulation spectrum; namely, GNE, CHNR and RALA. Interestingly, the latter is a novel characteristic that has been recently introduced in [27, 28].

Regarding the results in numeric terms, the system trained with the SVD dataset following the approach in *experiment 4* achieves an ACC of 81%, which to the best of the author's knowledge one of the best and more realistic results reported in the literature for this corpus. In a cross-dataset scenario the AUC of the system reached values ranging from 0.75 to 0.94, demonstrating that the procedures followed are robust to handle mismatches between training and testing conditions. It is worth noting that there exist in literature works reporting values superior to the ones in this paper, but many of those have only included a small portion of the dataset or have limited their analyses to a restricted number of voice pathologies. We expect that the results presented in this paper will help to establish a baseline for future comparisons following procedures that can be replicated by other researchers.

In general terms, it can be concluded that a methodology based in hierarchical detection, characterisation through a reduced set of consistent features and fusion of different speech tasks enhances the performance of the system. The cross-dataset trials have been demonstrated the robustness of the proposed AVQA system to mismatches in recording conditions, providing more realistic outcomes compared to intra-dataset experiments. As future work we plan to examine the feasibility of the system in a clinical setting, where the proposed system is evaluated and assessed as an assistive tool in the detection of pathologies. Similarly, we

plan to study the influence of other extralinguistics such as age, This include an in-depth analysis to correlate chronological to physiological age, and to establish frontiers to differentiate among age groups (young, elders, etc.). Similarly, paralinguistic information such as mood and accent are to be studied.

Acknowledgment

This work was supported by the Ministry of Economy and Competitiveness of Spain under grant DPI2017-83405-R1, and "Becas de Ayuda a la Movilidad" of the Universidad Politécnica de Madrid.

References

[1] J. B. Snow, J. J. Ballenger, Ballenger's Otorhinolaryngology Head and Neck Surgery, 2003.

[2] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing 10 (1-3) (2000) 19–41.

[3] W. Campbell, D. Sturim, D. A. Reynolds, Support vector machines using GMM supervectors for speaker verification, Signal Processing Letters, IEEE 13 (5) (2006) 308–311.

[4] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech and Language Processing 19 (4) (2011) 788–798.

[5] D. Garcia-Romero, C. Y. Espy-Wilson, Analysis of i-vector length normalization in speaker recognition systems, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (2011) 249–252.

[6] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, C. Ramírez-Calvo, Acoustic analysis of voice using WPCVox: a comparative study with Multi Dimensional Voice Program, European Archives of Oto-Rhino-Laryngology 265 (4) (2008) 465–476.

[7] J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, S. Aguilera-Navarro, P. Gómez-Vilda, An integrated tool for the diagnosis of voice disorders, Medical Engineering & Physics 28 (3) (2006) 276–289.

[8] M. Putzer, W. J. Barry, Instrumental dimensioning of normal and pathological phonation using acoustic measurements., Clinical linguistics & phonetics 22 (6) (2008) 407–20. doi:10.1080/02699200701830869.

[9] Saarbrüecken voice database.
URL `http://www.stimmdatenbank.coli.uni-saarland.de/index.php4`

[10] D. Chan, A. Fourcin, D. Gibbon, B. Grandstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. in'T Veld, J. Zeiliger, EUROM - A spoken language resource for the EU (1995).

[11] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mario, C. Nadeu, Albayzin speech database: design of the phonetic corpus., in: EUROSPEECH, 1993, pp. 175–179.

[12] C. Draxler, Introduction to the verbmobil-phondat database of spoken german, in: Proc. 3rd Int. Conf. Practical Application Prolog, 1995, pp. 201–212.

[13] Massachusetts Eye and Ear Infirmary, Voice disorders database, version.1.03 [cd-rom], Lincoln Park, NJ: Kay Elemetrics Corp (1994).

[14] V. Parsa, D. G. Jamieson, Acoustic discrimination of pathological voice: sustained vowels versus continuous speech, Journal of speech, language, and hearing research : JSLHR 44 (2) (2001) 327.

[15] J. B. Alonso-Hernandez, J. De Leon, I. Alonso, M. A. Ferrer, Automatic detection of pathologies in the voice by HOS based parameters, Eurasip Journal on Applied Signal Processing 2001 (4) (2001) 275–284. doi:10.1155/S1110865701000336.

[16] "aplicación de las tecnologías de la información y comunicaciones" database.
URL `http://www.atic.uma.es/index_atic.html`

[17] J. I. Godino-Llorente, P. Gómez-Vilda, M. Blanco-Velasco, Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters., IEEE transactions on bio-medical engineering 53 (10) (2006) 1943–1953.

[18] Y. Zhang, J. J. Jiang, Acoustic Analyses of Sustained and Running Voices From Patients With Laryngeal Pathologies, Journal of Voice 22 (1) (2008) 1–9.

[19] H. Kasuya, Normalized noise energy as an acoustic measure to evaluate pathologic voice, The Journal of the Acoustical Society of America 80 (5) (1986) 1329.

[20] G. de Krom, A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals., Journal of speech,language, and hearing research 36 (2) (1993) 254–266. doi:10.1044/jshr.3602.254.

[21] D. Michaelis, T. Gramss, H. W. Strube, Glottal-to-noise excitation ratioa new measure for describing pathological voices, Acta Acustica united with Acustica 83 (4) (1997) 700–706.

[22] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, The Journal of the Acoustical Society of America 87 (1990) 1738.

[23] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, G. Castellanos-Domínguez, Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients, IEEE Transactions on Biomedical Engineering 58 (2) (2011) 370–379. doi:10.1109/TBME.2010.2089052.

[24] J. Hillenbrand, R. A. Houde, Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech, Journal of Speech Language and Hearing Research 39 (2) (1996) 311.

[25] S. N. Awan, N. Roy, C. Dromey, Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model., Clinical linguistics & phonetics 23 (11) (2009) 825–841. doi:10.3109/02699200903242988.

[26] L. Atlas, S. A. Shamma, Joint acoustic and modulation frequency, EURASIP Journal on Advances in Signal Processing 2003 (7) (2003) 310290.

[27] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, G. Andrade-Miranda, Modulation Spectra Morphological Parameters: A New Method to Assess Voice Pathologies according to the GRBAS Scale, BioMed Research International 2015.

[28] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, Voice Pathology Detection Using Modulation Spectrum-Optimized Metrics, Frontiers in Bioengineering and Biotechnology 4 (1).

[29] H. Kantz, T. Schreiber, Nonlinear Time Series Analysis, 2nd Edition, Cambridge University Press, 2004.

[30] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, I. M. Moroz, Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection, BioMedical Engineering OnLine 6 (1) (2007) 23. doi:10.1186/1475-925X-6-23.

[31] C. Peng, S. Havlin, H. E. Stanley, A. L. Goldberger, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, Chaos: An Interdisciplinary Journal of Nonlinear Science 5 (1) (1995) 82–87.

[32] S. M. Pincus, Approximate entropy as a measure of system complexity., Proceedings of the National Academy of Sciences 88 (6) (1991) 2297–2301. doi:10.1073/pnas.88.6.2297.

[33] J. S. Richman, J. R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy., American journal of physiology. Heart and circulatory physiology 278 (6) (2000) H2039–49.

[34] H.-B. Xie, W.-X. He, H. Liu, Measuring time series regularity using nonlinear similarity-based sample entropy, Physics Letters A 372 (48) (2008) 7140–7146.

[35] L. Xu, K. Wang, L. Wang, Gaussian kernel approximate entropy algorithm for analyzing irregularity of time-series, in: Proceedings of 2005 International Conference on Machine Learning and Cybernetics., no. August, 2005, pp. 5605–5608.

[36] W. Chen, Z. Wang, H.-B. Xie, W. Yu, Characterization of surface EMG signal based on fuzzy entropy., IEEE transactions on neural systems and rehabilitation engineering 15 (2) (2007) 266–72. doi:10.1109/TNSRE.2007.897025.

[37] C. Bandt, Ordinal time series analysis, Ecological Modelling 182 (3-4) (2005) 229–238.

[38] M. Zanin, L. Zunino, O. A. Rosso, D. Papo, Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review, Entropy 14 (12) (2012) 1553–1577. doi:10.3390/e14081553.

[39] J. D. Arias-Londoño, J. I. Godino-Llorente, Entropies from Markov Models as Complexity Measures of Embedded Attractors, Entropy 17 (6) (2015) 3595–3620. doi:10.3390/e17063595.

[40] F. Schiel, Automatic Phonetic Transcription of Non-Prompted Speech, in: Proc. of the ICPhS, San Francisco, 1999, pp. 607–610.

[41] M. Brookes, VOICEBOX: Speech Processing Toolbox for MATLAB, Web page (2005).

[42] M. Cord, P. Cunningham (Eds.), Machine Learning Techniques for Multimedia, Cognitive Technologies, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-75171-7.

[43] G. Brown, A. Pocock, M.-J. Zhao, M. Luján, Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection, Journal Machine Learning Research 13 (2012) 27–66.

[44] J. Mason, J. Thompson, Gender effects in speaker recognition, Proc. ICSP-93, Beijing (1993) 733–736.

[45] R. Fraile, N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, C. Fredouille, Automatic Detection of Laryngeal Pathologies in Records of Sustained Vowels by Means of Mel-Frequency Cepstral Coefficient Parameters and Differentiation of Patients by Sex, Folia Phoniatrica et Logopaedica 61 (3) (2009) 146–152. doi:10.1159/000219950.

[46] J. A. Gómez-García, L. Moro-Velázquez, J. I. Godino-Llorente, C. G. Castellanos-Domínguez, An insight to the automatic categorization of speakers according to sex and its application to the detection of voice pathologies: A comparative study, Revista Facultad de Ingeniería (79) (2016) 50–62.

[47] A. Larcher, K. A. Lee, B. Ma, H. Li, Text-dependent speaker verification: Classifiers, databases and RSR2015, Speech Communication 60 (2014) 56–77.