On the design of automatic voice condition analysis systems. Part I: review of concepts and an insight to the state of the art.

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 first of a two-part series devoted to review the current state of the art of automatic voice condition analysis systems. The goal of this paper is to provide to the scientific community and to newly comers to the field of automatic voice condition analysis a resource that presents introductory concepts, a categorisation of different aspects of voice pathology and a systematic literature review that describes the methodologies and methods that are mostly used in these systems. To this end, the phenomena of pathological voice is firstly described in terms of perceptual characteristics and its relationship with physiological phenomena. Then, a prototypical automatic voice condition analysis system is described, discussing each one of its constituting parts and presenting an in-depth literature review about the methodologies that are typically employed. Finally, a discussion about some variability factors that affect the performance of these systems is presented.

Keywords: Automatic voice condition analysis, voice pathology detection, extralinguistic aspects of the speech, voice quality.

1. Introduction

Speech is accomplished through complex articulatory movements that mould the vocal excitation source in order to convey spoken sounds. In this process, three components can be identified: The *excitation source* (be it voiced, unvoiced, a mixture of both or its absence -such as in a pause-) providing the driving force for the speech production process, the *articulation* defined by the movements of the speech articulators moulding the production of a certain sound, and the *fluency* defining the rate at which the speech is generated. Despite the main objective of speech is transmitting information by means of sounds that encode linguistic content, the inherent intricacy of the production process embeds a substantial amount of non-linguistic information that is often described in terms of dimensions [1–3]. In this regard, the *paralinguistic* dimension of speech conveys information about the affective, attitudinal or emotional state of the speaker; the *extralinguistic* dimension informs about the speaker's identity and state (with *traits* such as age, sex, condition, etc.); whereas the *linguistic* dimension is related to the message, variations in language, dialect, sociolect, idiolect and speech style of the speaker. There is often described a fourth *transmittal* dimension that tells nothing about the speaker but about its physical location.

*Corresponding author

By virtue of the valuable information contained in speech and the inexpensiveness and easiness of capturing speech signals in a non-invasive manner, a great deal of interest has arisen in designing systems capable of isolating a certain dimension (or trait within a dimension) for automatic analysis purposes. As a matter of example, the literature presents systems that have been focused on extracting extralinguistic information to automatically determine the identity of speakers [4], age [5] or sex [6]. In the same manner, paralinguistic information has been extracted to identify the speaker's emotions [7] or level of interest [8]. Likewise, the linguistic dimension has been studied to recognize the accent, dialect [9] or the message itself (speech recognition) [10].

One application that has been gaining popularity during the last years is in the *analysis of speakers' condition* using voice recordings. In this respect, the clinical evaluation of voice disorders often relies on an instrumental examination and a perceptual analysis of the speech. The *instrumental* medical examination focuses on a primary aetiological diagnosis through the investigation of acoustic, aerodynamic, electroglottographic, videolaryngostroboscopic and/or the exploration of other types of biosignals; whereas the *perceptual* examination extracts multidimensional information that is not quantifiable instrumentally, by means of a qualitative description of the perceived degree of dysphonia that is present in the voice [11–13]. This information might be complemented by an interview where the patient states his/her symptoms, the examination of his/her medical records, evaluation of other body functions and systems, and exploration of the laryngeal structures and their function. These procedures should lead the medical expert to a diagnosis about the condition of the patient. The diagnostic process is differential, i. e., all possible causes of a problem are considered, and then the available information is matched against each one of the hypothesis explaining the disorder in the search for a match [14]. In some severe pathological cases a decision about the condition of the patient is straightforward, but in others, it would probably be conditioned by subjective factors or the observations and hypothesis made by the clinician. The increasing need of improving the diagnosis of voice pathologies has given rise to an emerging field called *Automatic Voice Condition Analysis* (AVCA), that aims at analysing, classifying and quantifying the degree to which a patient is affected by a voice disorder. This analysis is performed using automatic systems that provide objective measurements of the patient's vocal condition, exploiting the close relationship that exists between acoustic features extracted from the speech and voice pathology [15]. This reduces the evaluation time and the cost of diagnosis and treatment, providing extra advantages such as the avoidance of invasive procedures thanks to the employment of speech signals which are easily recorded by inexpensive means [16].

With these precedents in mind, the aim of this paper is to provide a review of AVCA systems, introducing key concepts related to vocal pathologies and their acoustic consequences in voice signals. A typical AVCA system is also described, detailing each one of its constituting blocks while providing a thorough literature review to portray the most used methodologies. Finally, some confounding factors that affect AVCA systems are discussed as well. It is worth noting that the main interest of this paper is related to the automatic *detection* (classification of control vs. pathological) and *identification* of voice disorders (multiclass classification of the actual disorder affecting speech), rather than the *assessment or grading* of voice signals. Indeed, we consider that the assessment of voice pathologies deserves of a separate paper to handle the particularities of this classification task.

This paper is organised as follow: section 2 describes introductory concepts related to voice pathologies; section 3 introduces AVCA systems, whereas section 4 describes its constituting blocks, presenting a review of the most typical methodologies employed in literature. Then, section 5 describes some variability factors affecting this type of systems. Finally, section 6 presents some discussions and concluding remarks.

2. Voice pathologies

Following the description presented in the introduction, a *speech disorder* can be defined as an impairment of the articulation of speech sounds, fluency and/or voice [17, 18]. It is worth noting that from all these elements, *this paper is only focused in those pathologies affecting voice*, and therefore the main interest of this paper is in the study of phonatory aspects of the speech. Articulatory, prosodic or language disorders are by themselves topics that should be handled separately.

To address the concepts of *voice condition* it is firstly necessary to describe the properties of a "normal" voice. This, however, poses numerous difficulties since there exist several definitions of "normality", and the distinction from what can be considered healthy or abnormal relies on subjective perceptual judgements made by listeners or by the speaker itself [19]. Indeed, a singer who uses a deviant voice as a trademark, might acknowledge his/her voice as normal, but this can be perceived otherwise by some listeners. By contrast, a high-pitched voice which in different circumstances would be considered normal if uttered by a child, might be deemed as pathological if uttered by an adult. In spite of that, there are certain common characteristics that can be regarded as normative, and thus, can be utilised as synonyms of *non-pathological voice condition*. Literature presents definitions which differ in terms of what (and how) can be categorised as normal or **normophonic**. In this paper, we adopt the *perceptual* definition presented in [20], on which a normophonic voice is described in terms of the following properties: (i) a pleasant quality, with an absence of noise, inappropriate breaks, perturbations or atonality; (ii) pitch in accordance to the age and sex of the speaker; (iii) loudness that is appropriate to the communication event; (iv) pitch and loudness variations that are available to express emphasis, meaning or subtleties indicating individual feelings and semantic differences; (v) sustainability to meet social and occupational needs.

Abnormal voices do not posses any, a combination, or all of the above properties. Typically, three types of aberrant voices are usually identified: aphonia, dysphonia and muteness [20]. **Aphonia** is characterised by the absence of vibration of the vocal folds -but not of sound- resulting in a voice that is perceptually described as extremely breathy. Similarly, **muteness** is referred to the absence of vocal folds vibration, accompanied by the inability to produce audible sounds. Finally, **dysphonia** is described by the absence of vocal quality, pitch, loudness, and/or variability which is inappropriate for an individual's age and/or sex [17, 20].

From the perspective of AVCA systems, muteness has to be discarded from the study due to the unavailability of audible outputs for automatic analysis purposes using voice registers. Similarly, the perceptual consequences of aphonia are so notorious that an automatic analysis to detect or assess the impairment is seldom considered. Consequently, and to the authors' knowledge, there is not a single work in literature dealing with automatic analysis of aphonic voices. At the end, only dysphonic and normophonic voices are examined by AVCA systems for labours of identification, detection or grading of pathological states.

Revisiting the definitions of dysphonia and normophonia it can be observed that 4 elements are identified, i. e., loudness, pitch, quality and variability. In this regard, *loudness* is defined as a perceptual correlate of the intensity of the sound pressure created by the release of air through the glottis. Disorders affecting loudness occur when the voice is louder or softer in concordance to the speaker's context. Loudness impairments are often indicators of personality disorders (overly aggressive, shy, or socially insecure behaviour), or are the consequence of certain pathologies such as Parkinson's disease or paresis. The second perceptual trait is *pitch*, which is the correlate of the frequency of vibration of the vocal folds, i. e., the fundamental frequency (f_0). The rate of vibration is determined by the physical characteristics of the vocal folds such as the mass, elasticity or length. Impairments affecting pitch include those where voice is tremulous or abnormal in concordance to the speakers' context. Examples include mutational falsetto (abnormally high-pitched voice uttered by an adult not correlating with his age or sex) or ventricular phonation (abnormally low-pitched voice product of the vibration of the false vocal folds). The third perceptual trait is *quality*, which is a correlate of the vibrational patterns of the vocal folds and resonant characteristics of the vocal tract. This is in turn composed of other traits describing very differentiated physiological phenomena as described next [18, 20]: (i) *strain*, related to disturbances in the vibratory patterns of the vocal folds due to an excessive tension in the larynx which may result in over adduction of vocal folds; (ii) *breathiness*, related to turbulent air streams released during incomplete vocal closure; (iii) *roughness (or harshness)*, related to irregularities or vibration defects of the vocal folds; and (iv) *resonance*, caused by abnormalities present in the vocal tract, such as defects in the closure of the velopharyngeal port. Despite there is a fourth trait related to *variability*, examining the flexibility of voice in relation to variations of pitch, loudness and quality in spoken contexts, we consider that -at least in AVCA methodologies- these variations can be included directly into their respective descriptors. We also believe that is possible to consider a superclass that embeds pitch and loudness into a sphere of *vocal aspects* as both examine the concordance of the voice respecting the speaker or its context. Therefore, we propose to analyse voices using automatic systems on the basis of two spheres, one related to the *vocal aspects* on which the adequacy of the pitch, loudness and its variations are examined in relation to the context and to normative values of an average speaker in the same population group. And a sphere related to *vocal quality* on which the resonant characteristics and the vibrational patterns of the vocal folds are examined.

The sphere of vocal aspects has been seldom studied in literature, with a vast majority of papers published in topics related to the analysis of vocal quality. In addition, and despite resonant patterns make up for voice quality, the most important descriptor often found in the state of the art is referred to a quantification of the *vocal aperiodicity* describing the manner of vibration of the vocal folds. Due to their importance in the design of AVCA systems, the mechanisms of vocal aperiodicity are described next.

It has been stated that three[1] types of vibrational patterns are often encountered in voice signals [22]: (i) *type I* voices, characterised by a nearly periodic behaviour; (ii) *type II* signals, which contain bifurcations, sub-harmonics or modulating frequencies; and (iii) *type III* voices, which are characterised by an aperiodic behaviour. In accordance with such distinction, normophonic voices are usually enclosed into the Type I typology, whereas pathological voices are embodied into the Type II and III categories [22, 23]. As a matter of example, Figure 1 illustrates some cases of voice signals following the above-mentioned typology.

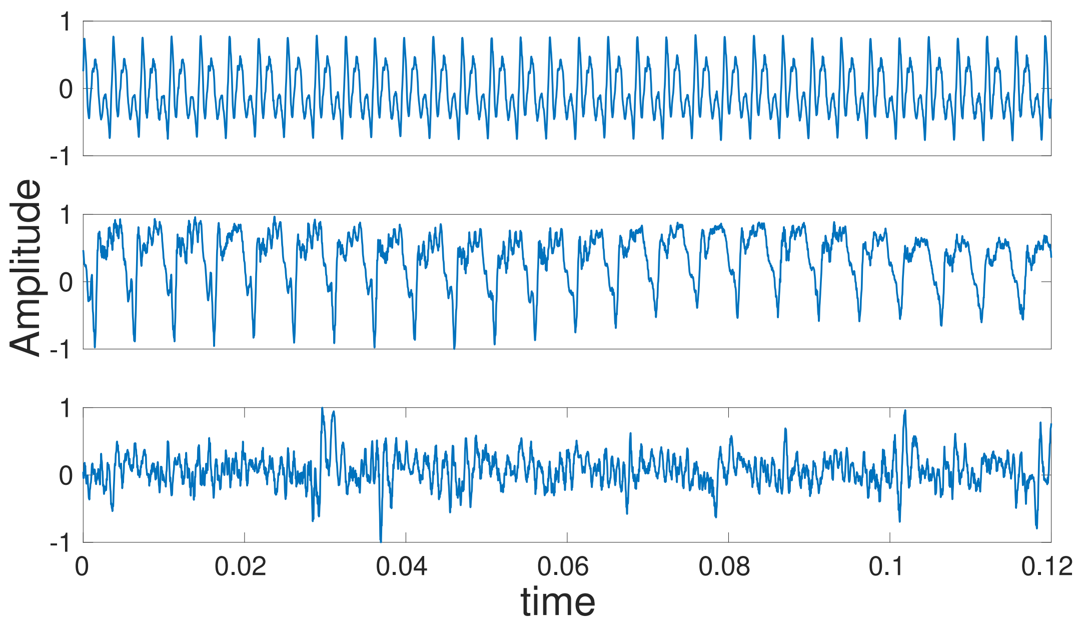

Figure 1. Typology of voice signals according to [22]. *Top panel:* normophonic type I signal, characterised by a periodic behaviour; *middle panel:* pathological type II signal having modulating frequencies; *bottom panel:* pathological type III signal characterised by an aperiodic behaviour.

Vocal aperiodicity is explained as the result of some very distinctive processes occurring during the voice production process such as [24–27]: (i) *irregular dynamics* of the vocal folds and involuntary transients between dynamic regimes (distinguishing features on very specific voice impairments such as diplophonia or biphonation); (ii) *modulation noise* owing to extrinsic perturbations in amplitude and frequency of the glottal cycle and which if often associated to roughness; and (iii) *additive noise* owing to turbulent airflow and which is correlated to breathy vocal quality.

To summarise some of the concepts introduced in this section, Figure 2 is presented.

[1]Despite there are authors defining a fourth category [21], a vast majority of works still employ the most classical definition.

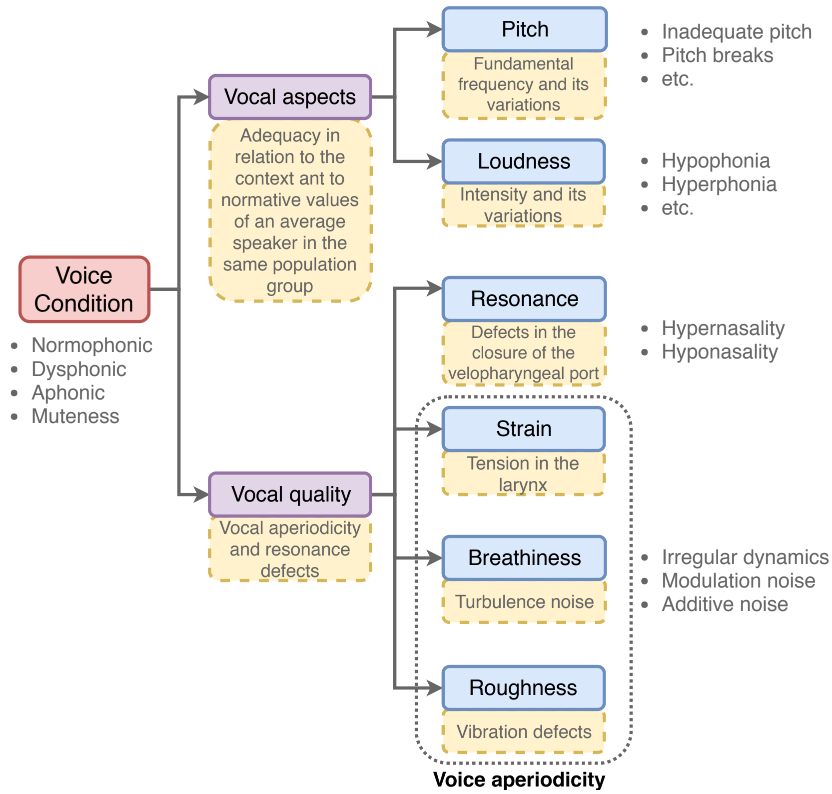

Figure 2. Graphical representation of the concepts introduced in the section. The dashed lines are referred to the physical phenomena whereas the box in blue is referred to the perceptual trait.

3. Automatic voice condition analysis systems

Generally, AVCA systems follow a pattern recognition-like structure on which characteristics are extracted from the acoustic signal in the form of a set of features to accomplish a further decision making task. An example of a typical AVCA scheme is presented in Figure 3.

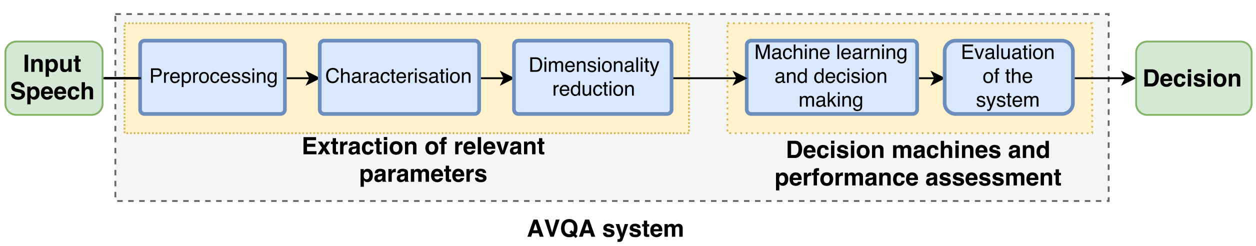

Figure 3. Depiction of a typical AVCA system.

Before going deeper into each one of the bulding stages of AVCA systems, two initial considerations -referred to the *input speech* and *decision* blocks in the depiction- are to be addressed first in the following subsections to respond to the questions: (i) what type of speech task is to be used for the design of the system; (ii) what type of decision should the system provide.

3.1. The input speech

The interest of considering the acoustic material in the design of AVCA systems arises from the fact that depending on the type of utterance, different configurations of the speech production subsystems arise, permitting the analysis of certain aspects of speech or others. Indeed, some pathologies are more likely to be identified when examining determined units of speech. For instance, resonance impairments are more easily perceived in utterances containing /m/ or /n/ prompts.

In general, two types of *speech production tasks* are employed for the evaluation of voice condition: sustained phonation of vowels and running speech. On one hand, *sustained phonation* is the result of the production of voiced sounds due to the vibration of the vocal folds, as when a vowel is uttered and maintained during a certain amount of time. Some advantages of using sustained phonations in AVCA systems include [23, 28]: (i) the facility to be analysed by automatic tools; (ii) the production of vowels is straightforward; (iii) vowels are not affected by paralinguistic or extralinguistic characteristics such as speaking rate, dialect, intonation, and idiosyncratic articulatory behaviour; (iv) vowels often generate simpler acoustic structures that might lead to consistent and reliable perceptual judgements of voice quality; (v) vowels do not depend on extra processing stages (such as voiced/unvoiced detectors) for the design of AVCA systems.

The selection of the vowel to be uttered is also a relevant matter. It has been stated that the type of vowel -along with the vocal effort and the muscle tension in the larynx- influences the degree of vocal folds approximation, affecting the perception of voice quality [29]. For this reason certain open vowels, such as /a/, are often employed in AVCA systems since they are produced with a relatively open tract allowing the examination of the entire vocal tract apparatus. By contrast vowels like /i/ and /u/, may not allow this examination due to the separation between the front and back cavities of the mouth during its production [30].

On the other hand, *running (or connected) speech* is the result of the source signal (either voiced, unvoiced, a mixture of both or its absence) being modulated by the articulatory subsystems, as when uttering a certain word or pronouncing sentences. Despite this speech production task is not as widely popular in AVCA systems as the one based on sustained vowels, there exist strong arguments favouring its use. Indeed, one interesting property of running speech comes from a phenomenon called *coarticulation*, which is related to the influence of the preceding and succeeding acoustic unit on the current unit under analysis. The dynamical effects introduced by coarticulation might be relevant for certain applications. Besides that, it has been stated that the impressions of certain characteristics of vocal quality are more easily perceived on vowels generated in a voiced context, vowels after a glottal closure, or during the production of strained vowels [31]. Some additional advantages reported in literature in favour of analysing connected speech include [15, 25]: (i) it requires switching on and off the vibration of the vocal folds continuously, or maintaning the voicing while the supraglottal apparatus changes, facilitating the exploration of certain dynamic aspects of the speech production; (ii) speakers are less likely to compensate for voice problems while producing connected speech than while phonating sounds; (iii) running speech provides a more realistic scenario since sustained phonations are more characteristic of singing rather than speaking; (iv) running speech contains fluctuations of vocal characteristics in relation to voice onsets, terminations and breaks; (v) in certain voice disorders (e. g., spasmodic dysphonia), the production of sustained vowels is less symptomatic than in connected speech, which may lead to an underestimation of the impairment; (vi) running speech contains variations in pitch and loudness, parameters that are important in the analysis of abnormal voice quality.

3.2. Automatic decision tasks

Three fundamental tasks may be considered in an AVCA system: voice pathology detection, voice pathology identification and voice pathology assessment.

On one hand, *voice pathology detection* is a two-classes decision making process aiming to decide whether a given speech register is normophonic or pathological (dysphonic or aphonic). On the other hand, *voice pathology identification* is a multi-class decision making process on which the goal is to assign a category to the input speech. The identification task is typically made in terms of the actual pathology (nodules, Reinke's oedema, etc.), the aetiology (organic, functional, etc.) or any other categorisation that groups general aspects of the analysed speech. From a practical point of view, identification is more challenging than detection, because of the multiple-class scenario on which it is defined. Nonetheless, both tasks are intricately complex due to several factors, such as the wide range of profiles that are found for normophonic voices, the documented overlap between normophonic and pathological states [32], the close relationship between disease and certain quality factors associated to normal processes such as ageing [33], the simultaneous presence of pathologies of different aetiologies in the same patient, etc.

By contrast, *voice pathology assessment* is aimed at grading the level of pathology that is perceived in a given speech signal. This is of great relevance since it is not possible to instrumentally delimit a phonation behaviour categorically. An useful descriptor of dysphonia is the **hoarseness**, which portraits the noisy, atonal and/or odd vocal resonance patterns encountered in voices [20]. The hoarseness is widely employed in literature, as perceptually, it is described as a superclass that contains roughness and breathiness - the two most reliable traits describing vocal quality- [31]. The assessment task is generally performed in concordance to a *perceptual rating scale* that evaluates voice quality and provides information about the level of impairment. Some popular perceptual evaluation scales include the GRBAS [12, 31], Voice Handicap Index [34] and CAPE-V [35, 36] scales. Even though the perceptual scales have been designed to evaluate every aspect that is relevant to voice quality, the reliability of the ratings is conditioned by the multidimensional aspects of voice quality, the intrinsic variability of speech, the subjectivity of perception [37], and the nonlinear relationship between pathology and measured or perceived voice quality [32].

4. Prototypical AVCA systems: an insight to the state of the art

The present section describes each one of the building blocks of Figure 3, providing a review of some relevant techniques often employed. Along with the description, a literature review is provided to identifiy the techiques which have been used before by other authors.

The literature review is performed using the web search of Scopus® and Google scholar®. The terms that are employed include: "dysphonia", "pathology", "automatic", "voice", "quality", "classification", "detection", "identification" and combinations and derivations of them. Pathological states including "Parkinson's disease", "Alzheimer", "Obstructive Sleep Apnoea", "Nodules", etc. are also used in the web search. The review is limited to those papers published after the year 2000, focusing on journal papers (although some documents in reputable conferences are also included) listed in the Journal Citation Reports® or the Scimago Journal Rank®. The list of predatory journals (`https://predatoryjournals.com/journals/`) is also consulted to discard papers published in journals engaged in predatory practices.

4.1. Input speech

The collection of exemplar recordings describing the classes under study conforms a *dataset* or *corpus* of speakers which are typically used to train and test an AVCA system. Although the term *database* is often found in the literature, we discourage its use to avoid the technical connotations that it has in computer science.

The data acquisition process should follow certain guidelines to prevent the introduction of unexpected variability, including the avoidance of external sources of noise or the preservation of similar acoustic and instrumental conditions during the recording process. Some recommendations for the acquisition of voice signals for acoustic voice analysis has been presented in [38], advising -among others- the use of professional condenser microphones with a minimum sensitivity of -60 dB, constant mouth-to-microphone distances less than 10 cm, sampling frequencies between 20 to 100 kHz, and sound-treated rooms with ambient noise lower to 50 dB. Other considerations in terms of technical characteristics of microphones have been described in [39], where a flat frequency response microphone is recommended, within the frequency of lowest fundamental frequency and highest spectral component, equivalent noise level at least 15 dB lower than the sound level of the softest phonation, etc. In the same way, the recorded corpus should be large enough to contain all possible variations within the class, while being balanced in terms of age, sex, etc. , properly representing variations in speech due to accent, dialect, socialect, etc. Likewise, conditions such as smoking or professional voice should be accounted. The management of the corpus for the storage and accessibility of recordings from a medical perspective should also be considered, as this permits the creation of synergies towards certain tasks such as the diagnosis of the pathologies, or the assessment from a perceptual point of view as given by different evaluators. Some considerations referred to the management in a clinical setting for a large corpus of dysphonic and dysarthric speakers are discussed in [40].

Literature reports the existence of several public and privative datasets that have been recorded for the purposes of detection, identification or assessment of voice pathologies. Regarding *public* datasets, the *Massachussets Ear and Eye Infirmary* (`MEEI`) [41] is probably the most widely employed corpus, being for years, the sole resource that was available for the study of pathological speech. `MEEI` contains approximately 700 registers of the vowel /`a`/ and the first sentence of *"the rainbow passage"* text, recorded at varying sampling frequencies (25 kHz to 50 kHz). A subset of the corpus -chosen to ensure a balance in age, sex and pathologies- has been published in [42], becoming a standard partition for comparisons among different works in literature. Despite its popularity, `MEEI` suffers from well-known problems which might bias outcomes, like the different recording conditions of normophonic and pathological speakers [43, 44]. Recently, another speech pathology corpus has been made accessible publicly: the `SVD` dataset [45]. This partition was recorded by the Institut für Phonetik at Saarland University and the Phoniatry Section of the Caritas Clinic St. Theresia in Saarbrücken, Germany. It contains more than 2000 acoustic and *electroglottographic* (EGG) registers of the vowels /`a`/, /`i`/ and /`u`/ phonated at normal, high, low, and rising-falling pitch; as well as registers of the sentence *"Guten Morgen, wie geht es Ihnen?"* (*Good morning, how are you?*), recorded at 50 kHz and 16 bits of resolution.

Some of the most well-known *privative* datasets include the *Hospital Príncipe de Asturias* (`HUPA`) corpus [46] which contains registers of the sustained vowel /`a`/ of 366 adult Spanish speakers (169 pathological and 197 normophonic); or the *Arabic Voice pathology Dataset* (`AVD`) [47, 48] which is composed of registers of the vowel/`a`/ and running speech of 188 normophonic and 178 pathological Arabic speakers. Another notable corpora, which is perhaps one of the largest in terms of number of patients, is the one recorded in hospitals in Marseilles and Aix-en-Provence in France [40]. It is composed by registers of *sound-pressure level* (SPL), oral airflow, and subglottal air pressure of more than 2500 dysphonic, dysarthric and normophonic speakers. There exist other privative corpora, exhibiting a large variety of characteristics respecting the acoustic conditions followed during the recording process, the instrumentation, the type of speech material that is elicited, the type of disorders, etc. Indeed, most of the privative datasets contain microphonic recordings of the sustained phonation of vowel /`a`/ [49–67] or a combination of several vowels [68–72]. There are some datasets with registers of running speech for different languages, which in text-dependent scenarios employ isolated words [73, 74], reading of phrases such as *"the rainbow passage"* [75–77], *"the north wind and the sun"* [78, 79], *"the story of Arthur the rat"* [80], or other texts [73, 81–86]. In the text-independent case they employ conversational speech or other types of elicitation tasks [84]. There exist some other datasets that include other type of complementary biosignals besides the acoustic recording. Namely, some contain EGG recordings of the vowel /`i`/ [87, 88], /`a`/ [51, 66]; while others complement the acoustic registers with questionnaire data [89, 90] or laryngoscopy information [90].

Regarding the languages that are reported, literature indicates datasets uttered by Russian [57, 74], Korean [91], Spanish [46], Colombian [92], Arabic [48], German [32, 93], Czech [94], Dutch [85, 86], Chinese [84], Brazilian [69], French [40] or Lithuanian [95, 96] speakers. Likewise, some include a broad range of voice pathologies [41, 45, 46], whereas other are concentrated in certain disorders such as nodules [97], polyps [59, 98], larynx cancer [87, 88, 99, 100], hypofunctional voices [11], diplophonia [101], spasmodic disphonia and muscle tension dysphonia [102], unilateral laryngeal paralysis [50, 103], obstructive sleep apnea [104, 105], hypernasality [106], Parkinson's disease [92, 107, 108], dysphagia [109, 110], lupus [111], etc.

4.2. Preprocessing

Since speech is intrinsically non-stationary, some preprocessing methods are often employed before the utilisation of conventional signal processing techniques which rely on stationary requirements. One common procedure is the *short-time analysis*, which decomposes the input signal into a series of equal-length chunks of speech, called *frames*, permitting the treatment of each individual chunk as a stationary or quasi-stationary fragment. This procedure is composed of two operations: *framing*, which divides the signal into frames (typically overlapped); and *windowing*, which tapers the beginning and ending of the frames, through the product with a window function to improve spectral properties. The window function should be selected to provide a frequency response with a narrow bandwidth in the main lobe and large attenuation in the side-lobes. Popular choices include the triangular, Hanning or Hamming windows, whereas the window length varies depending on the application. Typically for applications using sustained phonation, the duration of the window is set in between 20-40 ms. The upper limit ensures that frames are not that large to make the quasi-stationarity assumption void, whereas the lower limit is set to make the analysis independent of the location of pitch pulses within the segment, while ensuring at least two to three times pitch periods (since the typical range of pitch frequency is between $80 - 500$ Hz, a pitch pulse is expected every $12 - 2$ ms [112]). For applications using running speech, window lengths are typically set in the order of $20 - 30$ ms to conserve the quasi-stationary assumptions [113].

From those works reporting the type of window that is employed, literature indicates the popularity of Hamming [11, 23, 28, 37, 43, 44, 57, 63, 74, 74, 81, 89, 93, 94, 114–120, 120–132, 132, 133, 133–143] or Hanning [27, 88, 91, 96, 138, 144–150] windows. The length of the window varies depending on the application, type of speech task and characteristics that are utilised. Popular values found in literature include 10 ms [65, 91, 139, 144], 16 ms [151], 20 ms [48, 81, 121, 131–133, 152–154], 25 ms [74, 128, 155], 30 ms [68, 114, 120, 130, 143], 40 ms [43, 129, 134, 142, 146, 147, 156–163], 50 ms [61, 77, 88, 96, 118, 149] or 55 ms [141, 156]. Notwithstanding, for certain types of features they can be as large as 80 ms [164], 100 ms [165], 200 ms, [23, 166], 262 ms [119], 400 ms [167], or 800 ms [168].

Other preprocessing techniques often found are voice/unvoiced and endpoint detectors [37, 43, 66, 146, 147, 162], which ensure that only segments generated during the vibration of the vocal folds are employed; or silence detectors [169] that eliminate utterances not containing speech. Similarly, and to remove the influences of the vocal tract in the speech signal, inverse filtering is often employed [52, 69, 120, 170–174]. Likewise, the use of pre-emphasis filtering to accentuate the high frequency content of speech [74, 89, 121, 130] has been employed, although it has been reported that it does not improve detection results in AVCA systems [175].

4.3. Characterisation

The characterisation stage has the goal of extracting features capable of portraying the properties of the classes under analysis. The idea is to extract a d-dimensional vector of characteristics, $\vec{x} = \{x[1], \cdots, x[d]\}$, describing d properties of the input speech. Usually this vector is associated to a label ℓ indicating the membership of the utterance to a certain class, although this is not necessary. The features can be extracted either in a *short-time basis* (as introduced in the previous section) having as many vectors of features as frames of speech, or in a *long-term basis* calculating a single vector of characteristics per audio register.

Finding characteristics that effectively describe the presence of voice impairment is difficult, specially since some phenomena associated to voice disorders (such as aperiodicity) are present even in non-pathological states due to perturbations inherent to the phonation process [176]. As a result, there is no single feature, in the context of screening, that perfectly differentiates between normophonia and pathology, or that biunivocally correlates acoustic measurements and voice quality [157]. A common approach to counteract this, consists on studying different types of features, in the hope of finding combinations of characteristics that complement with each other. In this sense, the best characteristics would be those with the lowest correlation with the others but capable to provide the best discrimination capabilities [138]. Although multidimensional studies have reported good performance in screening tasks, this type of analysis is usually carried out by complex pattern recognition techniques, which makes difficult the interpretation of results from the perspective of a human evaluator [138].

The vast majority of descriptors of voice condition seek to compute metrics of vocal quality due to their close relationship to features extracted from voice signals. Literature reports diverse characterisation schemes which have been found to perform differently according to the pathologies under study or the dataset that has been employed. In general, some popular features -to be described in the following sections- include those based on *temporal and acoustic analysis*, *perturbation and fluctuation*, *spectral-cepstral*, *complexity*, *3-dimensional representations* and *other types of features* not fitting in the above categories.

4.3.1. Temporal and acoustical analysis

To the best of the authors' knowledge, there are only a fwe papers accounting for the *vocal aspects* of speakers (see Figure 2). In the first two, authors seek to monitor aberrant patterns of f_0 and SPL of hyperfunctional speakers and employ them for the distinction between dysphonic and normophonic speakers [77, 177]. Likewise, in [102], the degree of voice breaks are used to distinguish spasmodic dysphonia speakers from others suffering from muscle tension dysphonia and a control population. There are not automatic systems that seek to correlate SPL and f_0 to the contextual and personal traits of the speaker under analysis.

By contrast, most of literature is referred to the analysis of descriptors of vocal quality. With regards to the analysis of *irregular dynamics* in pathologies such as diplophonia, literature reports the computation of the degree of subharmonics and the diplophonia diagram [101, 178]. To capture *modulation* and *additive noise*, some approaches are based on tracking f_0 and deriving low-order statistics to track disturbances in the normal vibration patterns of the vocal folds [32, 59, 62, 91, 95, 99, 124, 131, 179]. A system for the personalised computation of f_0 according to the sex and age of the speaker has been presented in [180], using this value for the discrimination of normophonic and dysphonic voices. A different approach consists on measuring the vocal function through the quantification of the energy contained in the signals. Since this quantity is dependent on the distance between mouth and microphone, measures of SPL are preferred instead. These can be achieved by means of intraoral or subglottal pressure apparatuses, or indirectly computed by using accelerometers placed on the neck [77, 177]. Measurements of the vocal level at diverse frequency ranges define voice range profiles (or phonetograms), which have been employed for voice pathology detection [181, 182]. An extension to the method relies on the characterisation of a speech range profile, that has also been used in AVCA systems [78, 183]. Finally, it is possible to extract features from the glottal signal, characterising the opening and closing phases of the glottal waveform as in [52], or via the residue obtained after inverse filtering, with measures such as the mean square residue or the excess coefficient (kurtosis of the magnitude distribution) [69]. Other acoustic characteristics measure include, for instance, the number and degree of voice breaks and unvoiced frames in speech [184].

4.3.2. Perturbation and fluctuation analysis

Perturbations are minor disturbances or temporary changes that deviate from an expected behaviour [22]. Perturbation parameters have been frequently used to analyse *vocal aperiodicities* that are the product of modulation or additive noise. The most popular *modulation noise quantifiers* include families of parameters based on jitter and shimmer.

On one hand, *jitter* is a measure of the short-term (cycle-to-cycle) perturbation of f_0, with popular examples including [185]: (i) *jitter relative*, which is the average absolute difference between consecutive periods, divided by the average period; (ii) *jitter Relative Average Perturbation* (RAP), which is the average absolute difference between a period and the average of this and its two neighbours, divided by the average period; and *jitter 5-point Period Perturbation Quotient* (PPQ5), which is the average absolute difference between a period and the average of this and its four closest neighbours, divided by the average period. Jitter has been extensively used in AVCA systems, with some relevant examples in [52, 54, 66, 69, 85, 91, 95, 98, 99, 103, 131, 179, 184, 186–188]. On the other hand, *shimmer* measures short-term (cycle-to-cycle) amplitude perturbations, having popular examples in [185]: (i) *shimmer absolute*, which is the variability of the peak-to-peak amplitude in dB; (ii) *shimmer relative*, which is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude; and (iii) *shimmer 3-point Amplitude Perturbation Quotient* (APQ3), which is the average absolute difference between the amplitude of a period and the average of the amplitudes of this and its neighbours, divided by the average amplitude. Literature reports examples of shimmer in AVCA systems in [52, 85, 91, 95, 98, 99, 103, 131, 179, 184, 186–188]. Despite both families of features have been of considerable utility for describing type I signals, its validity for type II and III typologies has been put into question due to the need of an accurate identification of f_0 [189]. Some methods have been devised to calculate jitter using spectral techniques, avoiding the need of a precise f_0 computation, having accomplished favourable results in pathology detection tasks [27, 148], whereas others have derived shimmer and jitter through autoregressive decomposition and pole tracking [190]. Another perturbation measure proposed for the characterisation of Parkinson's disease is the *Pitch Period Entropy* [107, 191, 192] which takes into account the smooth vibrato and microtremor in normophonic voices, and the logarithmic nature of speech.

Popular *additive noise quantifiers* compute the relationship between the harmonics and background noise contained in speech, with notable examples including the *Signal-to-Noise Ratio* (SNR) [82, 186–188], *Harmonics-to-Noise Ratio* (HNR) [85, 91, 131, 187, 193] (and its variation Cepstral HNR [194, 195]), *Normalized Noise Energy* (NNE) [42, 95, 187, 196] (and its variation adaptive-NNE [135]), and *Glottal-to-Noise Excitation Ratio* (GNE) [138, 145, 194]. These features have being extensively applied in the evaluation of voice quality correlating positively with many speech disorders and perceptual ratings [37, 138]. Other noise quantifiers include the empirical mode decomposition excitations ratios, the vocal fold excitation ratios or the glottis quotient, which have been employed for the characterisation of Parkinson's disease [191, 192].

By contrast, *fluctuation analysis* is referred to the study of severe disturbances in the dynamics of the vocal folds behaviour, reflecting the inherent instability of the system [22]. *Tremor* is one prominent characteristic that is often studied, referred to low-frequency fluctuation in amplitude and/or frequency (modulation noise), related to pathologies of neurologic origin [22]. Some popular tremor estimates include *amplitude and frequency tremor* [179, 197, 198], or the *turbulent noise index* [58, 122].

4.3.3. Spectral-Cepstral analysis

Measures derived from the acoustic spectrum/cepstrum have been widely used in the study of pathological phonation, to mostly characterise *vocal quality*. Indeed, spectral and cepstral features have demonstrated high correlation with the perceptual assessment of dysphonia, providing large sensitivity when used in classification tasks [11]. The analysis with spectral/cepstral features also presents several advantageous properties, including its appropriateness for analysing both sustained vowels and running speech with no extra procedures, and the ability to characterise speech signals without depending of the estimation of f_0 [11, 37].

Spectral measures derived directly from the speech spectrum include diverse *Long-Term Average Spectrum* (LTAS) characteristics [23, 28, 76, 85], where the spectral tilt [23] -which indicates the degree of energy concentrated in low- vs. high-frequency areas of the spectrum- is a noteworthy example. Other popular sets of features are based on the estimation of the spectral envolope with a noteworthy example in the *Linear Predictive Coding Coefficients* (LPC) [126, 153, 199–202], which can also be used to decompose the speech signal into its residual and vocal tract components and hence derive parameters for characterisation [153]. Other features extracted from the residual signal include the *pitch amplitude* [23, 28, 42], which measures the dominant peak in the auto-correlation function of the residual signal; and the *spectral flatness ratio* [28, 42, 69], which measures the residual flatness. Some variations of LPC include cepstral transformation like the *Cepstral Linear Predictive Coding Coefficients* (LPCC) [48, 121, 141, 201–203], or a mel-transformation of LPC called Mel-line spectral frequencies [169].

Literature also reports several approaches relying on filter-banks to decompose voice signals into different sub-bands. For instance in [73], correlation functions measure the relationship between the bands of an octave filter-bank for the detection of pathologies. Another example is in [117], where HNR is computed on different frequency bands along with the energy. It is also possible to employ filter-banks that rely on psychoacoustic criteria to condense spectral information into perceptually-relevant frequency bands. Within this category, *Mel-Frequency Cepstral Coefficients* (MFCC) are probably the most popular features in speech-related applications, exploiting auditory principles and the decorrelating property of the cepstrum to characterize speech signals. MFCC have been used extensively in several works [43, 57, 115, 121, 130, 133, 140, 141, 146, 152, 156, 169, 201, 203–205], and often in combination with its first derivative (Δ) representing velocity [81, 150, 206], and/or its second derivative ($\Delta\Delta$) representing acceleration [116, 128, 129, 139, 147, 163]. Works related to MFCC characterisation include the study of the influence of the tapering window that is employed for the spectrum estimation during the MFCC processing [114], or the derivation of an index based on MFCC: the *pathological likelihood index* [157]. Other psychoacoustic characteristics include *Perceptual Linear Predictive Coding* (PLP), which have been designed in accordance to a scale modelling the human auditory system [207–209]. This has been used in AVCA systems in [48, 163, 210], next to its bandpass filtered variation RASTA-PLP [48, 141, 163, 201, 209, 210]. Other characteristics include the band power decorrelation time obtained through the Meddis and O'Mard filterbank model [37]. The filterbanks can also be employed to perform time-frequency decompositions, through wavelets transformations [50, 57, 60, 122, 123, 134, 165, 199, 211–215] -which are often accompanied by the calculation of energy and/or entropies of the sub-band decompositions- or adaptive time-frequency distribution [164]. An interesting review about these time-frequency decompositions is presented in [216].

Finally, measures based on the cepstral prominence of the harmonics are often utilised, by means of features such as *Cepstral peak prominence* (CPP) (next to its variation called smoothed CPP), which is a normalised measure of the cepstral peak amplitude, comparing the level of harmonic organisation of the speech to the cepstral background noise resulting from aspiration [217, 218]. This measure has been reported as one of the strongest correlates of breathiness [11, 15, 37]. Several studies have indicated that cepstral measures may be supplemented by other acoustic quantifiers such as the *Low-to-High Harmonic Ratio* (LHr) which measures the spectral tilt of the spectrum above and below 4 kHz [15]. Relevant examples employing these measures include [11, 25, 76, 80, 85, 97, 186]. A derived index that incorporates the information provided by CPP, LHr, and standard deviation of LHr is the cepstral spectral index of dysphonia presented in [75].

4.3.4. Complexity analysis

Complexity is a controversial term that has been classically linked to randomness and mistakenly associated to information measures such the algorithmic complexity [219]. However, it is more appropriately related to a "*meaningful structural richness*" [220] or to *fractal* behaviour rather than to randomness. Complex behaviour is typically observed in biological systems that manifest at least one of the following dynamical properties [221]: (i) nonlinearity; (ii) nonstationarity; (iii) time irreversibility (or asymmetry); or (iv) multiscale variability. One of the most popular approaches to investigate the complexity of a system is through *Nonlinear Dynamics Analysis* (NDA). Nonlinear phenomena arises naturally in physiological systems, and voice production is not an exception to this regard. Indeed, supporting findings of nonlinearity in phonation include nonlinear pressure-flow relations in the glottis, the delayed feedback from the mucosal wave, the nonlinear stress-strain curves of vocal fold tissues, nonlinearities associated with vocal fold collision [222], or asymmetries between the right and left vocal folds vibrations [223]. In addition, nonlinear dynamical behaviour of models of the vocal folds such as period-doubling (subharmonics), bifurcations, and transitions to irregular vibration have been observed in experiments with excised larynges; whereas period-doubling bifurcations and chaos-like features have been observed in signals from patients with organic and functional dysphonia [224]. Aforementioned facts suggests the appropriateness of using NDA to characterise the dynamics of voice production even in pathological scenarios, as voice pathologies can be considered disorders of glottal dynamics [225]. Indeed, the complexity features attempt to measure *vocal aperiodicity* phenomena in AVCA systems. The usual approximation to NDA relies on a reconstruction, termed *embedding*, to reveal the system dynamics in a space called *state space*. The most popular indices calculated using NDA, compute the dimensionality of the reconstructed state space. They have been used for discrimination of pathological and normophonic states, being popular examples the fractal dimension [72] or the *correlation dimension* (D2) [56, 67, 98, 103, 107, 156, 167, 186, 188, 226]. Other measures are based on the rate of divergence of trajectories in the state space. This has been explored to differentiate normophonic and dysphonic voices through the computation of *Largest Lyapunov Exponent* (LLE) [54, 71, 94, 141, 156, 167, 226, 227] or the Lyapunov spectrum [166]. Measures of entropy (the rate of information gain) have also been employed in AVCA systems, by means of the first and/or second order entropies [68, 98, 188], the relative entropy [61], pseudo-estimators such as the Ziv-Lempel complexity [71, 141, 226], or entropy metrics based on *Hidden Markov Models* (HMM) [156, 228]. A concept related to entropy is the *regularity*, which measures the predictability of the time series. The most popular regularity estimator used in pathological voice analysis is the ApEn [87, 88, 149, 226]. Other ApEn-derived metrics used in AVCA systems include the SampEn [51], GSampEn and FuzzyEn [71, 156, 228].

Likewise, time-frequency decomposition have been employed to explore the fractal properties of speech [176, 229] or to characterise complexity (using ApEn) on each decomposed sub-band [212]. Other measures explore the use of nonlinear prediction [70, 227]; measure the self-similarity of the voice by means of the DFA [107, 141, 191, 192, 224], the Hurst exponent [71, 94, 141, 176, 212]; or characterise properties of recurrence to compute the effects of modulation noise using the *Recurrence Period Density Entropy* (RPDE) [107, 141, 191, 192, 224, 230]. A discussion about characterisation using nonlinear dynamics can be found in [231].

4.3.5. 3-dimensional representations

A popular approach that has been gathering attention lately is based on the multidimensional representations of speech, and the employment of image processing techniques or matrix tools for the extraction of characteristics. In this regard, one approach is based on *Modulation Spectra* (MS), which characterises the modulation and frequency components of speech. MS produces a 3-dimensional representation that has been employed for the detection of voice pathologies [119, 140, 232].

Following the idea of 3-dimensional representations, a mel-spectrogram is characterised by means of the recurrence texture plots and the local binary pattern operator in [233]. Matrices can also be formed using time-frequency decomposition as in [132], where features are extracted through the employment of a multidirectional regression, the use of interlaced derivative patterns of the glottal source excitation signal as in [120]. Similarly, authors in [160] extract features from co-occurrence matrices formed after using filter-banks on the input speech or octave-spectrogram [142].

4.3.6. Other types of features

There exist some other approaches that do not fit into the above categorisation. For instance, the multidimensional *acoustic voice quality index* is a metric based on weighted multivariate regression of 6 temporal, spectral and cepstral characteristics that has been used for voice pathology assessment and detection [234]. The use of variograms and the characterisation with the signal-to-dysperiodicity ratio has been explored in [25, 116]. Likewise, decompositions of speech based on non-negative matrix factorisation [164] or empirical mode decomposition [168] have also been reported. The hoarseness diagram has been proposed to visualise additive and modulation noise components [53, 145, 235]. The utilisation of higher order statistics for the characterisation of dysphonic voices is discussed in [137], whereas the spectral properties of centralized auto-correntropy is used to detect and classify vocal pathologies in [236].

Some multidimensional approaches consider a combination of several features of different type, having some notable examples in [28, 55, 65, 96, 144, 156, 171, 172, 237, 238]. Other multidimensional approaches consider metrics derived from the `MPEG-7` standard as described in [128, 154, 162]. Another type of characterisations are based on biomechanical models to describe the behaviour of the glottal and mucosal waveforms [172, 237]. Finally, literature also reports the employment of MFCC features for the construction of a phonological model of 14 features (voicing, place of articulation, turbulence, nasality, etc.) [86].

4.4. Dimensionality reduction

Dimensionality reduction is aimed to decrease the size of the feature space in order to remove redundant or irrelevant features that might affect performance. Two major types of techniques can be defined: those based on feature extraction, which employ a transformation of the input space; or those based on feature selection, not relying on any transformation. *Feature extraction* techniques include the classical approaches of singular value decomposition [134], *Linear Discriminant Analysis* (LDA) [123, 161, 166, 199, 211, 237, 239], and *Principal Components Analysis* (PCA) [25, 57, 123, 124, 127, 136, 145, 153, 161, 172, 227, 237–239], or extensions of PCA such as kernel-PCA [127, 238], neural-networks PCA [127], or dynamic feature extraction using PCA [150]. Other type of transformations include those based on HMM [71, 136], clustering-based feature weighting methods [134, 161] or others based on multiple regression analysis [93]

Within *feature selection*, two types of methods arise. In one hand, the *wrapper feature selection* ties the selection of features to the maximisation of a performance metric obtained with a classifier/regressor. Some notable examples in AVCA include the use of genetic algorithms to select the best set of features for recognition purposes [50, 57, 90, 166, 212], binary logistic regression analyses with stepwise variable selection [78], sequential backward and forward feature selection [65, 239] or angle modulated differential evolution [63]. On the other hand, *filter feature selection* employs correlation and information approaches to find the most pertaining sets of features. In this respect, literature reports the use of the mutual information [145, 171], correlation analysis [145], Fisher discrimination ratio [47, 129, 129, 154, 226, 233, 237], Fisher discriminant analysis [144], or the Davies-Bouldin index [212].

4.5. Machine learning and decision making

The machine learning procedure receives different names depending on the type of decision making task that is involved. If given a set of observations $\boldsymbol{\mathcal{X}} = \{\vec{x}_1, \cdots, \vec{x}_n, \cdots, \vec{x}_N\}$, where each $\vec{x}_n$ is associated to a label $\vec{\ell} = \{\ell_1, \cdots, \ell_n, \cdots, \ell_N\}$, the aim of the procedure is to learn a mapping from the input set of observations to the labels. This type of task is known as *supervised learning*; in opposition to *unsupervised learning*, which is related to the discovery of structure in the data in the absence of labels.

To the author's knowledge all of the machine learning methodologies presented in AVCA systems are based on supervised learning. Within this category, the most widely employed decision machines include the *Support Vector Machines* (SVM) or *Gaussian Mixture Models* (GMM). SVM is a discriminative classifier constructed from sums of kernel functions which has been used in AVCA systems in [47, 50, 52, 60, 62, 63, 71, 73, 86, 90, 91, 94, 96, 116, 119, 120, 123, 140, 141, 150, 153, 154, 166, 169, 199, 201, 210, 212, 226, 229, 233, 239–241]. By contrast, GMM is a type of generative classifier that has provided excellent results in diverse speech-related applications. Its popularity arises from the modelling capability they offer and the probability framework on which they stand. The use of GMM in AVCA systems is reported in [114, 121, 126, 129, 132, 139, 140, 150, 156, 169, 200, 201, 204], next to variations such as GMM-Universal Background models [81, 163], GMM-SVM [130, 133, 201], i-Vectors [163].

Other popular decision machines include *Artificial Neural Networks* (ANN) [49, 57, 68, 123, 137, 143, 146, 152, 165, 166, 186, 211, 239], *Deep Neural Networks* (DNN) [142, 151, 155], HMM [74, 115], random forests [63, 65, 65, 94, 187], LDA [23, 25, 42, 122, 131, 139, 168, 169, 179, 203, 213, 214] quadratic-LDA [224], Hidden Markov Models [52, 71, 115, 121, 136, 139, 201, 228], *k-nearest neighbours* (KNN) [96, 117, 203, 212] and the Bayes classifier [150]. The use of regression techniques has also been reported in [28, 75, 78, 85, 234].

4.6. Evaluation of the system

A common approach followed in machine learning applications to generalise results and provide valid measures about the actual efficiency of the systems consists in the use of validation techniques. The basis of these methodologies is the decomposition of the available dataset into subsets which are used independently for training and testing purposes (and often for parameter tuning). On one hand, the *training partitions* are used to estimate a mapping between observations and labels in the supervised machine learning algorithms. On the other, the *testing partitions* are employed to assess the performance of the system. A third partition often arise for the purposes of parameter tuning of the algorithms.

The most straightforward approach for validation is the *split sample* method, which consist on using a percentage of the dataset to conform exclusive training and testing partitions. The problem of split sample is the lack of generalisation of results -specially if the data is scarce- as the randomly sampled partition might not be representative of the data under analysis, and the reported results will probably depend on the partitions that have been chosen. The use of this methodology has been reported in [52, 123, 137, 147, 169, 184, 201, 206, 239].

A different approach -which is one of the most popular validation techniques- is the *k-folds cross-validation*, which produces k disjoint sets of size N/k, called folds, with N representing the number of observations. In total k iterations are performed, using in each case a different subset for testing purposes, and the remaining $k-1$ for training algorithms. The measures of performance are then evaluated as the mean value calculated among iterations. Works using cross-validation techniques include [43, 47–49, 55, 57, 63, 63, 68, 71, 73, 91, 119–121, 124, 126, 127, 130–133, 136, 140, 146, 150–156, 160, 161, 168, 199, 200, 204, 211, 229, 233, 236, 240–242].

Another popular evaluation methodology is the *leave-one-out validation* which arises in the limit $k = N$ in a k-folds cross-validation. In this case, only one observation is used for testing and the remaining registers are employed for training, repeating this procedure N times. Leave-one-out validation is usually preferred to cross-validation when the dataset sizes are small as it allows to maximise the size of the training partition. Notable examples of leave-one-out validation are reported in [23, 49, 81, 96, 176, 213, 214].

In the same way, *boothstrapping* consists on randomly selecting a number of points from the training partition, with replacement, to train machine learning models and then calculating performance on a testing partition. This process is repeated k times, thus generating k different models. At the end, the performance is computed as the mean performance obtained in the testing partition. Some examples reporting the use of boothstrapping in AVCA are presented in [65, 75, 89, 101, 136, 224].

In addition to those validation methodologies which are trained and tested in the same corpus, other manners to validate performance are based on *cross-dataset validation* on which the training set corresponds to a particular dataset, whereas the testing partition corresponds to another different one. The advantage of such an approach is in the possibility of testing the robustness of the AVCA system in a more realistic scenario with an increased variability. This is methodology has been used in works such as those presented in [47, 120, 148, 151, 204, 243].

With regards to metrics of performance, the simplest approach consists on computing measures that compare the predicted labels given by the decision machines to the actual labels of the dataset. In this regard the most commonly used metric is the *accuracy* (ACC) -which has been used in almost all the reviewed papers- representing the rate of the correctly identified labels in comparison to the total number of instances. Another manner to analyse the performance of binary detection systems, is by means of *Receiver-Operating Curve* (ROC) [244] and *Detection Error Tradeoff* (DET) curves [245]. An additional measure derived from these curves is the *Area Under ROC Curve* (AUC), which is a value ranging between 0 and 1, that is obtained after integrating the ROC curve. A number of reasons favour the use of this metric instead of other classical measures such as the ACC, including [246]: (i) a standard error that decreases as both AUC and the number of test samples increase; (ii) decision threshold independence; (iii) and invariance to a-priori class probabilities. Some examples demonstrating the use of ROC curves and AUC are included in [23, 129, 144, 164, 166, 201]

Other types of performance evaluation techniques include the cost of log-likelihood-ratio [63, 89, 158], the DET curves [43, 65, 89, 91, 119, 129, 133, 138, 140, 146, 150, 150, 152, 174], or sensitivity versus specificity plots [75]. Statistical analysis based on the Mann-whitney U-test [69, 98, 103, 126, 167, 179, 188] or the t-test [47, 51, 57, 59, 64, 73, 80, 146, 187] have also been reported to compare the means of different populations.

4.7. Applications of AVCA systems

There exist a variety of applications of AVCA methodologies to characterise a wide variety of voice impairments. Without being extensive, the following will introduce some relevant applications of AVCA systems. In this manner, most of the works in literature are related to the analysis of laryngeal pathologies such as nodules [97], polyps [59, 98], larynx cancer [87, 88, 99, 149], diplophonia [101, 178], spasmodic disphonia [102], unilateral laryngeal paralysis [50, 103], laringectomised patients using oesophageal voice [70] . Notwithstanding, there exist several works focusing on other disorders such obstructive sleep apnea [104, 105, 247–249], hypernasality [106, 141, 250], Parkinson's disease [92, 107, 163, 192, 209, 210], Alzheimer's disease [251, 252], dysphagia [109, 110], lupus [111].

5. Aspects affecting AVCA systems

The variability embedded in speech has long been recognised as a major source of errors in automatic classification systems based on speech. For instance, several variability factors identified in the design of speaker recognition systems are described in [253]. Translated to terms of AVCA systems, these variability factors are the following: (i) *Peculiar intra-class variability:* manner of speaking, age, sex, inter-session variability, dialectal variations, emotional condition, etc. (ii) *Forced intra-class variability:* Lombard effect, external-influenced stress, cocktail-party effect, etc. (iii) *Channel-dependent external influences:* type of microphone, bandwidth and dynamic range reduction, electrical and acoustical noise, reverberation, distortion, etc. It is worth noting that the forced intra-class variability is more common of unsupervised recording environments rather than from controlled clinical settings. That does not imply, though, that their effects should be disregarded. For instance, they are of significant importance in telemedicine scenarios where the recording conditions might vary widely. Despite that, and for the purpose of simplicity, the forced intra-class variability is to be omitted from further discussions, and the term *intra-class variability is to be referred to the peculiar intra-class variability only.* Moreover, the intra-class and channel-dependent factors can be further associated to the linguistic, paralinguistic, transmittal and extralinguistic spheres.

Having this in mind, the current section introduces some factors that are of interest in the design of AVCA systems. It is worth noting that this list is not exhaustive but it only include what could be considered as the most important factors affecting these systems.

5.1. Intra-class variability

One important aspect to be outlined in any speech-related system is the effect of the intra-class or *intersession* variability. In a speaker recognition system, aiming at recognising the identity of a target speaker, the intersession effect might be described by the differences arising between recordings of the target speaker due to vocal effort, physical or emotional condition, etc. In the AVCA field, the intersession variability might be explained by the acoustic diversity among different pathologies, the sex or age of the speaker, or the spurious information introduced by other linguistic, paralinguistic or extra-linguistic effects. In this regard, several intra-class factors that might affect the performance of voice pathology classification systems include *linguistic* aspects such as the speech production task, the dialect and accent of the speaker; *paralinguistic* events such as the emotion or the vocal effort; or *extralinguistic* effects such as the sex or age of the speaker.

5.1.1. Dialects and accents

Dialects are the result of systematic, internal linguistic changes that occur within a language, reflected in the form of structural alterations in phonology, morphology, syntax, lexicon or semantic [33]. Dialects have been identified as an important aspect when defining communication disorders. Undoubtedly, not accounting for dialect features may result in the misdiagnosis of communication disorders [33]. For instance, several key features of African-American English phonology have been found to overlap with identifiers of speech sound delay/disorder in the phonology of general American English, making the distinction from normal and disordered states problematic in African-American speakers [254]. This phenomenon has been found in other contexts where non-prestige social dialects are often incorrectly associated to disordered speech [33]. *Accents*, on the other hand, are linguistic changes within a language that occur mainly at the phonological level. It has been long identified as an important confounding source in speech-related applications. For instance, accent is described as the most important source of variability between speakers in speech recognition systems in [255]. A further study presented in [256], confirmed that accent degrades classification rates, with errors increasing around 40-50% in cross-accent speech recognition scenarios. In general, it has been found that performance degrades when recognizing accented from non-native speech [257].

5.1.2. Vocal effort

Vocal effort is a subjective physiological interpretation of the voice level, as given by judges, or by the speaker itself to adapt speech to the demands of communication [258]. There exist evidences indicating that the vocal effort alter perceptual and acoustic parameters extracted from speech, and therefore might impact AVCA systems. Phenomenologically, variations in vocal effort affect the shape of the glottal pulses, changing the closing velocity waveform and affecting the relative duration of the closed interval [258]. In voices produced with increased vocal effort there have also been found significantly greater values in parameters such as subglottal pressure, translaryngeal airflow, and maximum flow declination [259]. Similarly, it has been stated that the medial compression of the vocal folds is enhanced when vocal effort is augmented, which results in an improved glottal closure, enlarged vocal intensity, and increased f_0 and amplitude [260]. Vocal effort also alters the duration of vowels, consonants, and the pausing behaviour during speech production [258]. In terms of quality, voices produced at excessive vocal effort are perceptually described as creaky [258] or strained [259]. Not surprisingly, this has consequences on parameters extracted from speech. For instance, jitter, shimmer, NNE and two EGG parameters have been found to vary significantly among vowels produced at three vocal effort levels (low, normal, high) [261].

Similarly, it has been reported that jitter and shimmer significantly increased their value with decreasing voice intensity [262], being also identified as one of the most important factors influencing the computation of these perturbation parameters, alongside with the sex of the speaker and the type of uttered vowel [263]. Other sets of parameters which are affected include cepstral features, which have been reported to differ substantially at diverse effort levels [260]. This has been confirmed in [259], where significant differences arose in 4 aerodynamic and 2 cepstral measures when comparing phonation at different effort levels. Authors in [264] have investigated the effects of increased vocal effort in pathological phonation. Results indicate that louder voicing reduces the values of perturbation parameters in normophonic speakers or in superficial vocal fold pathologies, while in cancer or vocal fold paralysis, louder phonation significantly enhances the irregularity of vocal folds vibration.

5.1.3. Emotion

The study of the emotional content embodied on speech has garnered a lot of attention within the speech research community. Indeed, the term *affective computing* has being coined to describe the automatic sensing, recognition and synthesis of human emotions from any biological modality such as speech or facial expressions [265]. There exist some studies considering the effect of emotions in speaker recognition systems. For instance, emotions are regarded as a factor affecting automatic recognition of children's speech in [266]. In [267], the effect of an emotions recogniser previous to a speech classification process is investigated. Authors report that affective speech downgrades the system performance, and that a cascading scheme is highly effective in improving recognition rates. Despite these facts, little is known about the influence of emotion in AVCA schemes, but according to the evidence found in the field of speaker recognition, it might be hypothesized that affective speech is a confounding factor that should be taken into consideration.

5.1.4. Sex

The variability introduced by the sex of the speaker remains as a major concern in the design of speech-based systems. Indeed, authors in [255] reported that this factor accounted for the second most prominent source of variability -after accent- in speech recognition systems. Certainly, literature states that the performance of speech recognition, identification or verification systems improves by employing a-priori information about the sex of the speaker [268]. For instance, authors in [269] obtained a 2% of accuracy improvement in a speaker recognition system when using sex-specific models.

The nature of the variability introduced by the sex of the speaker stands on physiological, acoustic, and psychophysical factors [6]. Regarding *physiological* differences, the human laryngeal anatomy differs between sexes at a variety of levels. Particularly, males tend to have a more acute thyroid angle; thicker vocal folds; a longer vocal tract; a larger pharyngeal-oral apparatus, thyroid lamina and skull compared to that of females [270, 271]. Studies of excised human larynges have shown that anteroposterior dimensions of the glottis are 1.5 times larger in men than in women [272]. Besides that, the female pharynx has been found to be shorter than of males during the production of the three cardinal vowels. This may be a key factor in distinguishing between male and female voice qualities during speech production [271]. In addition, the observation of the glottis during phonation has suggested the presence of a posterior glottal opening that persists throughout a vibratory cycle and which is common for female speakers, but occurs much less frequently among male speakers [273]. Indeed, about 80% of females and 20% of males have a visible posterior glottal aperture during the closed portion of a vocal period [274]. Regarding *perceptual* differences, parameters such as effort, pitch, stress, nasality, melodic patterns of intonation and coarticulation are used for characterising female voices, while male voices are judged on the basis of effort, pitch and hoarseness [275]. It is also argued that female voices possess a "breathier" quality than male voices [274]. The pitch is the most known trait differentiating sexes [275], with females' pitch higher than of males' by as much as an octave [276]. This pitch difference might influence the perception of dysphonic voices since lower pitch is perceived as rougher [31]. In addition to the pitch, literature reports significant differences between male and female speakers' formants (f_1, f_2, f_3, f_4) [6]. This is because the vocal tract length for males is longer than that of females, producing on average formant patterns scaled upward in frequency by about 20% [275]. There are also several important acoustic consequences of the posterior glottal opening during the closed phase of phonation, which is more frequent in females. A first consequence is a breathier voice quality which is the result of a larger amount of air passing through the vocal tract [270] and that affects the relative amplitude of the first harmonic of the speech spectrum [272, 277]. A second consequence is the widening of the f_1 bandwidth, which is the result of the glottal aperture that produces energy losses particularly at low frequencies [273, 277]. A third acoustic consequence is the generation of turbulence in the vicinity of the glottis [277], perceived with a high level of aspiration noise in the spectral regions corresponding to f_3, and contributing to a breathier voice quality [276]. A final consequence is a lower spectral tilt due to the presence of aspiration noise [276], which turns out to be a significant parameter for differentiating between male and female speech samples [273].

In addition to the acoustic differences reported from the study of the raw speech, there are some differences in the glottal components among sexes. On one hand, the female glottal waveform tends to have a shorter period, lower peaks and peak-to-peak flow amplitudes than that of males [278]. Likewise, the derivative of the glottal waveform does not present an abrupt discontinuity during the closing time due to the incomplete closure of the vocal folds [6]. In general, it is stated that female glottal components are symmetric, with opening and closing portions of the waveform tending towards equal duration [279]. Conversely, and regarding the glottal waveform of male speakers, it is found that the open quotient is smaller and the maximum flow declination rate is greater than of females [273]. Moreover, the closing portion of the waveform generally occupies $20 - 40\%$ of the total period and it might not exist an easily identifiable closed period [275]. In general, it is stated that male glottal waveforms are asymmetrical and present a hump in the opening phase. Finally, it is worth noting that sex differences are found to be age and hormone-dependent, and thus is of great importance considering the effect of age when studying male or female voices.

5.1.5. Age

According to the male-female coalescence model of ageing voice, hormone-related factors cause changes in voice production systems. In this manner, hormones during puberty are responsible for the differences between males and females in adolescence, but these changes are counteracted to some degree by hormone related factors associated with menopause and ageing [271]. During males' puberty, the thyroid cartilage develops the Adam's apple, the muscular and mucosal layers of the vocal folds thicken, the vocal folds lengthen and widen, the cricothyroid membrane widens, and the corresponding muscle becomes more powerful [270, 280]. As a result, the fundamental frequency decreases an octave compared to that of a child [280]. During females' puberty, there is little development of the thyroid cartilage or of the cricothyroid membrane, and the vocal muscle thickens slightly but remains supple and narrow. As a result, the female's f_0 becomes one third lower than that of a child [280]. The age effects on the larynx tend to be more significant in men than in women. In this manner, males experience an increasing of the fundamental frequency as a result of muscle atrophy, thinning of the lamina propria, general loss of mass and ossification and calcification of the larynx that starts during the third decade of life [271]. In females, ossification and calcification starts in the fourth decade of life, and in some cases never completely ossify. However, due to menopause effects, a lowering of fundamental frequency prior to senescence occurs [271].

The effects of age in AVCA systems have not been studied in depth. One of the few works that accounts for its effects in AVCA is introduced in [75], where this trait has been used as a predictor in a binary logistic regressor for the prediction of dysphonia, having found a marginal but statistical significant increment in performance when age is introduced in the model.

5.2. Channel-dependent external influences

This dimension includes all the effects that aggregate variability to speech registers because of the mere act of recording. This is a well-known problem that has long been identified in speech and speaker recognition systems [113]. Several aspects affect the recording process, including the instrumentation (type of microphone, analogue-to-digital converter, etc.), the acoustic environment (office, recoding studio, etc.) and the transmission means (land-line, cellular, etc.) [281]. Similarly, background noises, noises made by speakers (such as lip smacks), noise in the input device itself, etc., are recognised as sources that impair performance of speech recognition systems [113]. Another problem that might arise, is the variability introduced because of the mismatch in the recording conditions, between the registers employed for training the models and those used for testing purposes. For instance, as when a certain microphone is used for recording the training utterances, but the model is verified with a different equipment. Indeed, the microphone is expected to modify the speech spectrum, and anything that modifies the spectrum may cause difficulties in recognition tasks [281]. In this regard, the study of [38] demonstrated that the type of microphone has effects in the computation of perturbation parameters, providing evidence favouring the use of condenser cardioid microphones instead of dynamic or omnidirectional microphones. This study also showed that sensitivity and microphone-to-mouth distance have the largest effect on perturbation measures, whereas the angle had little effect for short distances, but a greater effect for longer distances. The study presented in [282] showed that a signal-to-noise ratio of 42 dB is needed to provide reliable estimations of perturbations measures, whereas values less than 30 dB have been shown to impact negatively in their computation. In [283], the effect of background noise, reverberation, clipping and speech compression on the calculation of MFCC features was tested out, demonstrating significant (but predictable effects) in MFCC computations.

6. Discussions and conclusions

This paper has presented concepts in relation to voice impairments and AVCA systems. In this manner, a categorisation of diverse aspects of voice conditions in terms of perceptual and physiological phenomena has been proposed, as well as a description of a prototypical AVCA system along with each one of its constituting parts. With relation to the latter, a systematic review has been carried out to overview the methodologies that are more often employed in AVCA systems.

Regarding the categorisation presented in section 3 and according to the systematic review of section 4, some inferences can be made. Firstly, a large number of papers still employ the `MEEI` corpus despite its well-known limitations. Notwithstanding, the field has received the advent of novel public datasets such as the `SVD` or privative corpora shared among different research groups, which has permitted the reproducibility of results and has opened up the possibility to carry out comparisons among methodologies in other datasets apart from `MEEI`. Despite that, there is room for improvement, as there is still necessary to record larger datasets, more balanced in terms of pathologies, age or sex, and containing a larger variety of acoustic material based either in sustained vowels, isolated words or running speech.

Literature has also revealed that most of the works employ sustained phonation due to its simplicity, despite the potential that running speech presents. In this regard more investigation on novel features and methodologies that employ this type of speech task is required. Similarly, it has been found that the effects of extralinguistics or paralinguistics (such as age, sex, accent, etc.) have been seldom considered in the vast majority of systems reported in literature, despite their relevance as confounding factors on this type of systems. Accounting for this variability factors should be a relevant matter to study in the future.

Regarding the characterisation techniques, most of the reviewed papers report the employment of descriptors to quantify vocal quality. However, vocal aspects describing variations in intensity and f_0 are also important aspects to consider as they might serve to characterise other phenomena such as hypophonia or inadequate pitches in speakers. Quantifying both intensity and f_0 might complement the information obtained with descriptors of voice quality, with potential -for instance- to improve results of differential analysis in identification tasks. This comes with the added cost, though, of having to record other variables besides audio, such as the SPL or EGG (or to employ robust f_0 estimators such as those based on inverse filtering). In the same manner, the systematic review also indicates that the increasing need of novel biomarkers for the early diagnosis, differential analysis or assessment of disorders such as Parkinson, Alzheimer or obstructive sleep apnea, has generated an emerging interest on quantifying dysphonic conditions. The analysis of these disorders have also brought new features and processing techniques which have indeed enriched the field.

Regarding the machine learning and decision making methodologies, there is a large amount of methodologies related to supervised learning. The typical approach followed in AVCA systems is based on bottom-up schemes on which the voice disorders phenomena is firstly studied, to build up systems from the inferences obtained in the previous analysis stage. Notwithstanding, other related fields (such as speech recognition) have experienced an increased interest in up-bottom schemes through unsupervised methodologies for the purposes of pattern discovery or data mining. This same path could be followed in AVCA systems as well.

The literature review also served to reveal the existence of certain methodological issues that might compromise the interpretation and validity of results in certain works. In an attempt to provide general recommendations, some practical considerations are indicated next. One concern that is found in a few papers is due to the addition of registers of other corpora besides the one under study, for the purposes of balancing patients in terms of pathology, sex, age, etc.; or simply to increase the size of the studied corpus. The effects of following such an approach are certainly important and might bias or affect the validity of results due to the channel divergences between the datasets. Likewise, and despite there is an trend on employing more robust validation techniques, there are certain issues that should be taken into account as well. For instance, there are some works that suggest having included registers of the same speaker on both training and testing partitions. This seems to be the results of having datasets on which speakers record audio in different sessions, but not accounting for the possibility of including audio of the same speaker in a training or a testing partition. Following this approach might introduce speaker information that might bias the machine learning algorithms. Besides that, one problem that is common in a variety of papers is to report results with confidence intervals which are larger than the range of measurement. For instance, there are certain works reporting accuracy values of the type $99 \pm 1.5\%$. It is recommended the use more robust estimators of confidence not to allow values larger than the range of acceptable values. A few papers have reported the employment of energy measurements to characterise normophonic or dysphonic conditions without having used SPL or a normalisation procedure to account for differences in the recording condition of the different registers (due to divergences in the mouth-to-microphone distance between recordings for instance). Another comment should be made on the importance of using an appropriate feature space in concordance to the size of the dataset. There are some papers reporting a large number of features but using a small dataset. This is certainly discouraged.

A final comment should be made in regards to the clinical assessment of the systems presented in literature. The systematic review has demonstrated a lack of validation of the proposed systems in clinical settings where these automatic tools have served to guide or improve the diagnosis of voice impairments. Indeed, most of these systems have been tested under very restricted settings, depending on the dataset that has been used for training, the recording conditions, etc.; factors which might hinder the generalisability of the results. It is necessary to test out the validity of the AVCA methodology in more realistic scenarios, where its ability to contribute to the diagnosis of voice pathologies is tested.

To finalise this paper, Table 1 presents, in the authors' opinion, a list of some interesting works presenting AVCA systems, in the hope that they might result useful for new comers to the field. In the second part of this series, entitled "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" we will introduce a series of experiments following the methodologies described in this paper, using diverse corpora and analysing the effects of certain variability factors in the design of AVCA systems.

Table 1. Some relevant works in literature according to the taxonomy presented in Figure 3. RS.: running speech; AC.: acoustic and temporal features; SC.: spectral/cepstral features; Pn.: perturbation features; Cx.: complexity features; Ot.: other type of features.

Authors	Material	Database	Characterization					Dimensionality reduction	Decision making	Validation	Results
			AC.	SC.	Pn.	Cx.	Ot.				
[156]	/a/	MEEI						–	GMM	Crossvalidation	ACC=98%
[75]	RS.	privative						–	Regressor	Boothstrapping	AUC=0.85
[96]	/a/	privative						–	SVM; KNN; commitee	Leave-one-out	ACC=95%(detection); ACC=85%(identification)
[146]	/a/+RS.	MEEI						–	ANN	Crossvalidation	ACC=94%(vowel); ACC=96%(RS.)
[155]	/a/	3 databases						–	XGBoost; ANN; isolation forest	Crossvalidation	ACC = 62-73%
[224]	/a/	MEEI						–	QDA	Bootstrapping	ACC=92%
[119]	/a/	MEEI						Max-relevance	SVM	Crossvalidation	ACC=94%(detection); ACC=85-92%(identification)
[94]	/a/	MEEI; SVD; privative						Mann-Whitney U test	SVM; random forest		ACC=68-100%
[23]	/a/+RS.	MEEI						–	LDA	Leave-one-out	ACC=96%(vowel); ACC=96%(RS.)
[43]	/a/	MEEI						–	ANN	Crossvalidation	ACC=90%
[213]	/a/	MEEI						–	LDA	Leave-one-out	ACC=96%
[63]	/a/	privative						wrapper-based feature selection	SVM; random forests	Cross-validation	ACC=87%
[90]	/a/	privative						GA	SVM	Random-split	ACC=98%(multimodal)
[191, 192]	/a/	privative						4 filter-based	SVM; random forests	Cross-validation	ACC=98%

Acknowledgment

This work was supported by the Ministry of Economy and Competitiveness of Spain under grant DPI2017-83405-R1.

References

[1] J. Laver, Principles of Phonetics, Cambridge University Press, Cambridge, 1994.

[2] H. Traunmüller, Conventional, Biological and Environmental Factors in Speech Communication: A Modulation Theory, Phonetica 51 (1-3) (1994) 170–183.

[3] H. Traunmüller, Evidence for demodulation in speech perception, 2000, pp. 790–793.

[4] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: From features to supervectors, Speech Communication 52 (1) (2010) 12–40.

[5] M. H. Bahari, M. McLaren, H. Van hamme, D. a. Van Leeuwen, Speaker age estimation using i-vectors, Engineering Applications of Artificial Intelligence 34 (2014) 99–108.

[6] D. Childers, K. Wu, Gender recognition from speech. Part II: Fine analysis., The Journal of the Acoustical Society of America 90 (4 Pt 1) (1991) 1841–56.

[7] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (3) (2011) 572–587.

[8] B. Schuller, G. Rigoll, Recognising Interest in Conversational Speech âĂŞ Comparing Bag of Frames and Supra-segmental Features, in: INTERSPEECH, 2009, pp. 1999–2002.

[9] F. Biadsy, Automatic Dialect and Accent Recognition and its Application to Speech Recognition (2011).

[10] J. Benesty, M. M. Sondhi, Y. Huang, Springer Handbook of Speech Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[11] C. R. Watts, S. N. Awan, Use of Spectral/Cepstral Analyses for Differentiating Normal From Hypofunctional Voices in Sustained Vowel and Continuous Speech Contexts, Journal of Speech, Language, and Hearing Research 54 (December) (2011) 1525–1538.

[12] M. Anniko, M. Bernal-Sprekelsen, V. Bonkowsky, P. Bradley, S. Iurato, Otorhinolaryngology, Head and Neck Surgery, Springer, 2010.

[13] S. N. Awan, N. Roy, S. M. Cohen, Exploring the relationship between spectral and cepstral measures of voice and the voice handicap index (VHI), Journal of Voice 28 (4) (2014) 430–439.

[14] J. K. Casper, R. Leonard, Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment, illustrate Edition, Lippincott Williams & Wilkins, 2006.

[15] 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.

[16] 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.

[17] A. S.-L.-H. Association, Definitions of Communication Disorders and Variations, Tech. rep. (1993).

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

[19] N. B. Anderson, G. H. Shames, Human Communication Disorders: An Introduction, 8th Edition, Pearson, 2010.

[20] A. E. Aronson, D. Bless, Clinical Voice Disorders, 4th Edition, Thieme, 2009.

[21] A. Sprecher, A. Olszewski, J. J. Jiang, Y. Zhang, Updating signal typing in voice: addition of type 4 signals., The Journal of the Acoustical Society of America 127 (6) (2010) 3710–16.

[22] I. Titze, Workshop on Acoustic Voice Analysis, National Centre for Voice and Speech, America (1994) 1–36.

[23] 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.

[24] J. Schoentgen, Spectral models of additive and modulation noise in speech and phonatory excitation signals, The Journal of the Acoustical Society of America 113 (1) (2003) 553.

[25] A. Alpan, Y. Maryn, A. Kacha, F. Grenez, J. Schoentgen, Multi-band dysperiodicity analyses of disordered connected speech, Speech Communication 53 (1) (2011) 131–141.

[26] J. Hanquinet, F. Grenez, J. Schoentgen, Synthesis of Disordered Voices, in: International Conference on Non-Linear Speech Processing, NOLISP 2005, 2006, pp. 231–241.

[27] M. Vasilakis, Y. Stylianou, Spectral jitter modeling and estimation, Biomedical Signal Processing and Control 4 (3) (2009) 183–193.

[28] T. L. Eadie, P. C. Doyle, Classification of Dysphonic Voice: Acoustic and Auditory-Perceptual Measures, Journal of Voice 19 (1) (2005) 1–14.

[29] L. Baghai-Ravary, S. W. Beet, Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders, Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders (Fuchs 2005) (2013) 7–12.

[30] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, K. Johnson, Formants of children, women, and men: the effects of vocal intensity variation., The Journal of the Acoustical Society of America 106 (3 Pt 1) (1999) 1532–42.

[31] C. Moers, B. Möbius, F. Rosanowski, E. Nöth, U. Eysholdt, T. Haderlein, Vowel- and text-based cepstral analysis of chronic hoarseness, Journal of Voice 26 (4) (2012) 416–424.

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

[33] R. D. Kent, The MIT Encyclopedia of Communication Disorders, MIT Press, 2004.

[34] B. H. Jacobson, A. Johnson, C. Grywalski, A. Silbergleit, G. Jacobson, M. S. Benninger, C. W. Newman, The voice handicap index (VHI), American Journal of Speech-Language Pathology 6 (3) (1997) 66.

[35] G. B. Kempster, B. R. Gerratt, K. Verdolini Abbott, J. Barkmeier-Kraemer, R. E. Hillman, Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol, American Journal of Speech-Language Pathology 18 (2) (2009) 124.

[36] A. S.-L.-H. Association, [Online; accessed 18-August-2016].

[37] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, J. M. Gutiérrez-Arriola, Characterization of Dysphonic Voices by Means of a Filterbank-Based Spectral Analysis: Sustained Vowels and Running Speech, Journal of Voice 27 (1) (2013) 11–23.

[38] I. R. Titze, W. S. Winholtz, Effect of microphone type and placement on voice perturbation measurements, Journal of Speech & Hearing Research 36 (6) (1993) 1177–1190.

[39] J. G. Svec, S. Granqvist, Guidelines for Selecting Microphones for Human Voice Production Research, American Journal of Speech-Language Pathology 19 (4) (2010) 356–368.

[40] A. Ghio, G. Pouchoulin, B. Teston, S. Pinto, C. Fredouille, C. De Looze, D. Robert, F. Viallet, A. Giovanni, How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers?, Speech Communication 54 (5) (2012) 664–679.

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

[42] V. Parsa, D. G. Jamieson, Identification of pathological voices using glottal noise measures, Journal of speech, language, and hearing research 43 (2) (2000) 469.

[43] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, P. Gómez-Vilda, Methodological issues in the development of automatic systems for voice pathology detection, Biomedical Signal Processing and Control 1 (2) (2006) 120–128.

[44] N. Malyska, T. Quatieri, D. Sturim, Automatic Dysphonia Recognition using Biologically-Inspired Amplitude-Modulation Features, in: Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., Vol. 1, IEEE, 2005, pp. 873–876.

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

[46] 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.

[47] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T. a. Mesallam, M. Farahat, K. H. Malki, M. A. Bencherif, An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification, Journal of Voice 31 (1) (2017) 113.e9–113.e18.

[48] T. A. Mesallam, M. Farahat, K. H. Malki, M. Alsulaiman, Z. Ali, A. Al-nasheri, G. Muhammad, Development of the Arabic Voice Pathology Database and Its Evaluation by Using Speech Features and Machine Learning Algorithms, Journal of Healthcare Engineering 2017 (2017) 1–13.

[49] C. D. P. Crovato, A. Schuck, The Use of Wavelet Packet Transform and Artificial Neural Networks in Analysis and Classification of Dysphonic Voices, IEEE Transactions on Biomedical Engineering 54 (10) (2007) 1898–1900.

[50] R. Behroozmand, F. Almasganj, Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis, Computers in Biology and Medicine 37 (4) (2007) 474–485.

[51] C. Fabris, W. De Colle, G. Sparacino, Voice disorders assessed by (cross-) Sample Entropy of electroglottogram and microphone signals, Biomedical Signal Processing and Control 8 (6) (2013) 920–926.

[52] L. A. Forero M., M. Kohler, M. M. Vellasco, E. Cataldo, Analysis and Classification of Voice Pathologies Using Glottal Signal Parameters, Journal of Voice 30 (5) (2016) 549–556.

[53] M. Fröhlich, D. Michaelis, H. W. Strube, E. Kruse, Acoustic voice analysis by means of the hoarseness diagram., Journal of speech, language, and hearing research : JSLHR 43 (3) (2000) 706–20.

[54] A. Giovanni, M. Ouaknine, J. M. Triglia, Determination of largest Lyapunov exponents of vocal signal: application to unilateral laryngeal paralysis., Journal of voice : official journal of the Voice Foundation 13 (3) (1999) 341–54.

[55] S. Hadjitodorov, B. Boyanov, B. Teston, Laryngeal pathology detection by means of class-specific neural maps, IEEE Transactions on Information Technology in Biomedicine 4 (1) (2000) 68–73.

[56] J. Jiang, Y. Zhang, Nonlinear dynamic analysis of speech from pathological subjects, Electronics Letters 38 (6) (2002) 294–295.

[57] A novel hybrid of genetic algorithm and ANN for developing a high efficient method for vocal fold pathology diagnosis, EURASIP Journal on Audio, Speech, and Music Processing 2015 (1) (2015) 3.

[58] P. Mitev, S. Hadjitodorov, A method for turbulent noise estimation in voiced signals, Med. Biol. Eng. Comput 38 (2000) 625–631.

[59] M. Petrović-Lazić, S. Babac, M. Vuković, R. Kosanović, Z. Ivanković, Acoustic Voice Analysis of Patients With Vocal Fold Polyp, Journal of Voice 25 (1) (2011) 94–97.

[60] P. Saidi, F. Almasganj, Voice Disorder Signal Classification Using M-Band Wavelets and Support Vector Machine, Circuits, Systems, and Signal Processing 34 (2015) 2727–2738.

[61] P. R. Scalassara, M. E. Dajer, C. D. Maciel, R. C. Guido, J. C. Pereira, Relative entropy measures applied to healthy and pathological voice characterization, Applied Mathematics and Computation 207 (1) (2009) 95–108.

[62] V. Uloza, A. Verikas, M. Bacauskiene, A. Gelzinis, R. Pribuisiene, M. Kaseta, V. Saferis, Categorizing normal and pathological voices: Automated and perceptual categorization, Journal of Voice 25 (6) (2011) 700–708.

[63] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, Z. Kons, A. Satt, R. Hoory, Fusion of voice signal information for detection of mild laryngeal pathology, Applied Soft Computing 18 (2014) 91–103.

[64] J. H. Van Stan, D. D. Mehta, S. M. Zeitels, J. A. Burns, A. M. Barbu, R. E. Hillman, Average ambulatory measures of sound pressure level, fundamental frequency, and vocal dose do not differ between adult females with phonotraumatic lesions and matched control subjects, Annals of Otology, Rhinology and Laryngology 124 (11) (2015) 864–874.

[65] A. Verikas, A. Gelzinis, E. Vaiciukynas, M. Bacauskiene, J. Minelga, M. Hållander, V. Uloza, E. Padervinskis, Data dependent random forest applied to screening for laryngeal disorders through analysis of sustained phonation: Acoustic versus contact microphone, Medical Engineering & Physics 37 (2) (2015) 210–218.

[66] M. N. Vieira, F. R. McInnes, M. a. Jack, On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures., The Journal of the Acoustical Society of America 111 (2) (2002) 1045–1055.

[67] Y. Zhang, J. J. Jiang, Nonlinear dynamic analysis in signal typing of pathological human voices, Electronics Letters 39 (13) (2003) 1021.

[68] P. Henriquez, J. B. Alonso-Hernandez, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, F. Diaz-de Maria, Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics, IEEE Transactions on Audio, Speech, and Language Processing 17 (6) (2009) 1186–1195.

[69] M. de Oliveira Rosa, J. Pereira, M. Grellet, Adaptive estimation of residue signal for voice pathology diagnosis, IEEE Transactions on Biomedical Engineering 47 (1) (2000) 96–104.

[70] L. Landini, C. Manfredi, V. Positano, M. Santarelli, N. Vanello, Non-linear prediction for oesophageal voice analysis, Medical Engineering & Physics 24 (7-8) (2002) 529–533.

[71] C. M. Travieso, J. B. Alonso, J. Orozco-Arroyave, J. Vargas-Bonilla, E. Nöth, A. G. Ravelo-García, Detection of different voice diseases based on the nonlinear characterization of speech signals, Expert Systems with Applications 82 (2017) 184–195.

[72] V. Péan, M. Ouayoun, B. C. Fugain, A Fractal Approach to Normal and Pathological Voices, Acta Oto-Laryngologica 120 (2) (2000) 222–224.

[73] A. Al-nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions, Journal of Voice 31 (1) (2017) 3–15.

[74] V. Majidnezhad, A HTK-based Method for Detecting Vocal Fold Pathology, Acta Informatica Medica 22 (4) (2014) 246.

[75] S. N. Awan, N. Roy, D. Zhang, S. M. Cohen, Validation of the Cepstral Spectral Index of Dysphonia (CSID) as a Screening Tool for Voice Disorders: Development of Clinical Cutoff Scores, Journal of Voice 30 (2) (2016) 130–144.

[76] S. Y. Lowell, R. H. Colton, R. T. Kelley, Y. C. Hahn, Spectral- and Cepstral-Based Measures During Continuous Speech: Capacity to Distinguish Dysphonia and Consistency Within a Speaker, Journal of Voice 25 (5) (2010) e223–e232.

[77] D. D. Mehta, J. H. Van Stan, M. Zanartu, M. Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Cortes, H. A. n. Cheyne, R. E. Hillman, Using Ambulatory Voice Monitoring to Investigate Common Voice Disorders: Research Update., Frontiers in bioengineering and biotechnology 3 (October) (2015) 155.

[78] E. Ma, J. Robertson, C. Radford, S. Vagne, R. El-Halabi, E. Yiu, Reliability of Speaking and Maximum Voice Range Measures in Screening for Dysphonia, Journal of Voice 21 (4) (2007) 397–406.

[79] M. A. Little, D. A. Costello, M. L. Harries, Objective Dysphonia Quantification in Vocal Fold Paralysis: Comparing Nonlinear With Classical Measures, Journal of Voice 25 (1) (2011) 21–31.

[80] L. F. Brinca, A. P. F. Batista, A. I. Tavares, I. C. Gonçalves, M. L. Moreno, Use of Cepstral Analyses for Differentiating Normal From Dysphonic Voices: A Comparative Study of Connected Speech Versus Sustained Vowel in European Portuguese Female Speakers, Journal of Voice 28 (3) (2014) 282–286.

[81] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, A. Ghio, Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia), in: Proceedings of European Conference on Speech Communication and Technology (Eurospeech 2005), 2005, pp. 149–152.

[82] F. Klingholtz, Acoustic recognition of voice disorders: a comparative study of running speech versus sustained vowels., The Journal of the Acoustical Society of America 87 (5) (1990) 2218–2224.

[83] J.-W. Lee, S. Kim, H.-G. Kang, Detecting pathological speech using contour modeling of harmonic-to-noise ratio, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 5969–5973.

[84] T. Lee, Y. Liu, P.-W. Huang, J.-T. Chien, W. K. Lam, Y. T. Yeung, T. K. T. Law, K. Y. S. Lee, A. P.-H. Kong, S.-P. Law, Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 2016-May, IEEE, 2016, pp. 6475–6479.

[85] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, M. De Bodt, Toward Improved Ecological Validity in the Acoustic Measurement of Overall Voice Quality: Combining Continuous Speech and Sustained Vowels, Journal of Voice 24 (5) (2010) 540–555.

[86] C. Middag, Y. Saeys, J.-p. Martens, Towards an ASR-free objective analysis of pathological speech, in: Interspeech, no. September, 2010, pp. 294–297.

[87] K. Manickam, C. Moore, T. Willard, N. Slevin, Quantifying aberrant phonation using approximate entropy in electrolaryngography, Speech Communication 47 (3) (2005) 312–321.

[88] C. Moore, K. Manickam, T. Willard, S. Jones, N. Slevin, S. Shalet, Spectral pattern complexity analysis and the quantification of voice normality in healthy and radiotherapy patient groups, Medical Engineering & Physics 26 (4) (2004) 291–301.

[89] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, J. Minelga, M. Hållander, E. Padervinskis, V. Uloza, Fusing voice and query data for non-invasive detection of laryngeal disorders, Expert Systems with Applications 42 (22) (2015) 8445–8453.

[90] A. Verikas, A. Gelzinis, M. Bacauskiene, M. Hållander, V. Uloza, M. Kaseta, Combining image, voice, and the patient's questionnaire data to categorize laryngeal disorders, Artificial Intelligence in Medicine 49 (1) (2010) 43–50.

[91] J. W. Lee, H.-G. Kang, J.-Y. Choi, Y.-I. Son, An Investigation of Vocal Tract Characteristics for Acoustic Discrimination of Pathological Voices, BioMed Research International 2013 (2013) 1–11.

[92] J. R. Orozco, J. D. Arias, J. F. Vargas, M. C. González Rátiva, E. Nöth, New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease, LREC 2014. Proceedings of the Ninth International Conference on Language Resources and Evaluation (2014) 342–347.

[93] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, E. Nöth, PEAKS - A system for the automatic evaluation of voice and speech disorders, Speech Communication 51 (5) (2009) 425–437.

[94] J. Mekyska, E. Janousova, P. Gómez-Vilda, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, J. B. Alonso-Hernandez, M. Faundez-Zanuy, K. López-de Ipiña, Robust and complex approach of pathological speech signal analysis, Neurocomputing 167 (2015) 94–111.

[95] V. Uloza, V. Saferis, I. Uloziene, Perceptual and Acoustic Assessment of Voice Pathology and the Efficacy of Endolaryngeal Phonomicrosurgery, Journal of Voice 19 (1) (2005) 138–145.

[96] A. Gelzinis, A. Verikas, M. Bacauskiene, Automated speech analysis applied to laryngeal disease categorization, Computer Methods and Programs in Biomedicine 91 (1) (2008) 36–47.

[97] B. Radish Kumar, J. S. Bhat, N. Prasad, Cepstral analysis of voice in persons with vocal nodules, Journal of Voice 24 (6) (2010) 651–653.

[98] Y. Zhang, C. McGilligan, L. Zhou, M. Vig, J. J. Jiang, Nonlinear dynamic analysis of voices before and after surgical excision of vocal polyps., The Journal of the Acoustical Society of America 115 (5 Pt 1) (2004) 2270–7.

[99] A. Rovirosa, E. Martíez-Celdrán, A. Ortega, C. Ascaso, R. Abellana, M. Velasco, M. Bonet, C. Herrera, F. Casas, R. M. Francisco, M. Arenas, V. Hernández, A. Sánchez-Reyes, C. León, J. Traserra, A. Biete, Acoustic analysis after radiotherapy in T1 vocal cord carcinoma: a new approach to the analysis of voice quality, International Journal of Radiation Oncology*Biology*Physics 47 (1) (2000) 73–79.

[100] T. Ritchings, M. McGillion, C. Moore, Pathological voice quality assessment using artificial neural networks, Medical Engineering & Physics 24 (7-8) (2002) 561–564.

[101] P. Aichinger, I. Roesner, M. Leonhard, B. Schneider-Stickler, D.-M. Denk-Linnert, W. Bigenzahn, A. K. Fuchs, M. Hagmüller, G. Kubin, Comparison of an audio-based and a video-based approach for detecting diplophonia, Biomedical Signal Processing and Control 31 (31) (2017) 576–585.

[102] G. Schlotthauer, M. E. Torres, M. C. Jackson-Menaldi, A Pattern Recognition Approach to Spasmodic Dysphonia and Muscle Tension Dysphonia Automatic Classification, Journal of Voice 24 (3) (2010) 346–353.

[103] Y. Zhang, J. J. Jiang, L. Biazzo, M. Jorgensen, Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis, Journal of Voice 19 (4) (2005) 519–528.

[104] E. Goldshtein, A. Tarasiuk, Y. Zigel, Automatic detection of obstructive sleep apnea using speech signals., IEEE transactions on bio-medical engineering 58 (5) (2011) 1373–82.

[105] J. L. Blanco-Murillo, L. A. Hernández, R. Fernández-Pozo, D. Ramos, Improving Automatic Detection of Obstructive Sleep Apnea Through Nonlinear Analysis of Sustained Speech, Cognitive Computation.

[106] P. Vijayalakshmi, M. R. Reddy, D. O'Shaughnessy, Acoustic Analysis and Detection of Hypernasality Using a Group Delay Function, IEEE Transactions on Biomedical Engineering 54 (4) (2007) 621–629.

[107] M. Little, P. McSharry, E. Hunter, J. Spielman, L. Ramig, Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease, IEEE Transactions on Biomedical Engineering 56 (4) (2009) 1015–1022.

[108] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, E. Ruzicka, Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task., The Journal of the Acoustical Society of America 134 (3) (2013) 2171–81.

[109] K. López-de Ipiña, P. Calvo, M. Faundez-Zanuy, P. Clavé, W. Nascimento, U. Martinez de Lizarduy, D. Alvarez, V. Arreola, O. Ortega, J. Mekyska, P. Sanz-Cartagena, Automatic voice analysis for dysphagia detection, Speech, Language and Hearing 21 (2) (2018) 86–89.

[110] J. S. Ryu, S. R. Park, K. H. Choi, Prediction of laryngeal aspiration using voice analysis, American Journal of Physical Medicine and Rehabilitation 83 (10) (2004) 753–757.

[111] M. S. F. C. de Macedo, K. M. Costa, M. da Silva Filho, Voice disorder in systemic lupus erythematosus, PLOS ONE 12 (4) (2017) e0175893.

[112] K. Paliwal, K. Wojcicki, Effect of analysis window duration on speech intelligibility, IEEE Signal Processing Letters 15 (2008) 785–788.

[113] X. Huang, A. Acero, H. W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall PTR, 2001.

[114] Ö. Eskidere, A. GürhanlÄś, Voice Disorder Classification Based on Multitaper Mel Frequency Cepstral Coefficients Features, Computational and Mathematical Methods in Medicine 2015 (2015) 1–12.

[115] A. A. Dibazar, T. W. Berger, S. S. Narayanan, Pathological Voice Assessment, in: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2006, pp. 1669–1673.

[116] A. Alpan, J. Schoentgen, Y. Maryn, F. Grenez, Automatic perceptual categorization of disordered connected speech, in: INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, 2010, pp. 2574–2577.

[117] K. Shama, A. Krishna, N. U. Cholayya, Study of Harmonics-to-Noise Ratio and Critical-Band Energy Spectrum of Speech as Acoustic Indicators of Laryngeal and Voice Pathology, EURASIP Journal on Advances in Signal Processing 2007 (1) (2007) 085286.

[118] C. Manfredi, M. D'Aniello, P. Bruscaglioni, A. Ismaelli, A comparative analysis of fundamental frequency estimation methods with application to pathological voices, Medical Engineering & Physics 22 (2) (2000) 135–147.

[119] M. Markaki, Y. Stylianou, Voice Pathology Detection and Discrimination Based on Modulation Spectral Features, IEEE Transactions on Audio, Speech, and Language Processing 19 (7) (2011) 1938–1948.

[120] G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, A. Al-nasheri, M. A. Bencherif, Voice pathology detection using interlaced derivative pattern on glottal source excitation, Biomedical Signal Processing and Control 31 (2017) 156–164.

[121] S. Jothilakshmi, Automatic system to detect the type of voice pathology, Applied Soft Computing 21 (2014) 244–249.

[122] S. Hadjitodorov, P. Mitev, A computer system for acoustic analysis of pathological voices and laryngeal diseases screening, Medical Engineering & Physics 24 (6) (2002) 419–429.

[123] M. K. Arjmandi, M. Pooyan, An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine, Biomedical Signal Processing and Control 7 (1) (2012) 3–19.

[124] D. Panek, A. Skalski, J. Gajda, Quantification of Linear and Non-linear Acoustic Analysis Applied to Voice Pathology Detection, in: E. Piętka, J. Kawa, W. Wieclawek (Eds.), Information Technologies in Biomedicine, Vol. 284 of Advances in Intelligent Systems and Computing, Springer International Publishing, Cham, 2014.

[125] Y. Qi, R. E. Hillman, C. Milstein, The estimation of signal-to-noise ratio in continuous speech for disordered voices, Journal of the Acoustical Society of America 105 (4) (1999) 2532–2535.

[126] Z. Ali, I. Elamvazuthi, M. Alsulaiman, G. Muhammad, Automatic Voice Pathology Detection With Running Speech by Using Estimation of Auditory Spectrum and Cepstral Coefficients Based on the All-Pole Model, Journal of Voice 30 (6) (2016) 757.e7–757.e19.

[127] D. Panek, A. Skalski, J. Gajda, R. Tadeusiewicz, Acoustic Analysis Assessment in Speech Pathology Detection, International Journal of Applied Mathematics and Computer Science 25 (2015) 631–643.

[128] M. S. Hossain, Cloud-Supported CyberâĂŞPhysical Localization Framework for Patients Monitoring, IEEE Systems Journal 11 (1) (2017) 118–127.

[129] 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.

[130] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, V. Uloza, Exploring similarity-based classification of larynx disorders from human voice, Speech Communication 54 (5) (2012) 601–610.

[131] R. Moran, R. Reilly, P. de Chazal, P. Lacy, Telephony-Based Voice Pathology Assessment Using Automated Speech Analysis, IEEE Transactions on Biomedical Engineering 53 (3) (2006) 468–477.

[132] G. Muhammad, T. a. Mesallam, K. H. Malki, M. Farahat, A. Mahmood, M. Alsulaiman, Multidirectional regression (MDR)-based features for automatic voice disorder detection, Journal of Voice 26 (6) (2012) 817.e19–817.e27.

[133] X. Wang, J. Zhang, Y. Yan, Discrimination Between Pathological and Normal Voices Using GMM-SVM Approach, Journal of Voice 25 (1) (2011) 38–43.

[134] M. Hariharan, K. Polat, S. Yaacob, A new feature constituting approach to detection of vocal fold pathology, International Journal of Systems Science 45 (8) (2013) 1622–1634.

[135] C. Manfredi, Adaptive noise energy estimation in pathological speech signals, IEEE Transactions on Biomedical Engineering 47 (11) (2000) 1538–1543.

[136] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, G. Castellanos-Domínguez, An improved method for voice pathology detection by means of a HMM-based feature space transformation, Pattern Recognition 43 (9) (2010) 3100–3112.

[137] 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.

[138] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, F. Cruz-Roldán, The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders, Journal of Voice 24 (1) (2010) 47–56.

[139] A. Dibazar, S. Narayanan, T. Berger, Feature analysis for automatic detection of pathological speech, in: Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology, Vol. 1, IEEE, 2002, pp. 182–183.

[140] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, Y. Stylianou, On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices., Logopedics, phoniatrics, vocology 36 (2) (2011) 60–69.

[141] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolaños, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Hönig, E. Nöth, Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases, IEEE Journal of Biomedical and Health Informatics 19 (6) (2015) 1820–1828.

[142] G. Muhammad, M. F. Alhamid, M. Alsulaiman, B. Gupta, Edge Computing with Cloud for Voice Disorder Assessment and Treatment, IEEE Communications Magazine 56 (4) (2018) 60–65.

[143] G. Muhammad, S. M. M. Rahman, A. Alelaiwi, A. Alamri, Smart Health Solution Integrating IoT and Cloud: A Case Study of Voice Pathology Monitoring, IEEE Communications Magazine 55 (1) (2017) 69–73.

[144] T. Dubuisson, T. Dutoit, B. Gosselin, M. Remacle, On the Use of the Correlation between Acoustic Descriptors for the Normal/Pathological Voices Discrimination, EURASIP Journal on Advances in Signal Processing 2009 (1) (2009) 173967.

[145] D. Michaelis, M. Fröhlich, H. W. Strube, Selection and combination of acoustic features for the description of pathologic voices. 103 (3) (1998) 1628–1639.

[146] J. I. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, P. Gómez-Vilda, Automatic detection of voice impairments from text-dependent running speech, Biomedical Signal Processing and Control 4 (3) (2009) 176–182.

[147] J. I. Godino-Llorente, P. Gomez-Vilda, Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors, IEEE Transactions on Biomedical Engineering 51 (2) (2004) 380–384.

[148] M. Vasilakis, Y. Stylianou, Voice pathology detection based on short-term jitter estimations in running speech, Folia Phoniatrica et Logopaedica 61 (3) (2009) 153–170.

[149] C. Moore, K. Manickam, N. Slevin, Collective spectral pattern complexity analysis of voicing in normal males and larynx cancer patients following radiotherapy, Biomedical Signal Processing and Control 1 (2) (2006) 113–119.

[150] G. Daza-Santacoloma, J. D. Arias-Londono, J. I. Godino-Llorente, N. Saenz-Lechon, V. Osma-Ruiz, G. Castellanos-Dominguez, Dynamic Feature Extraction: an Application To Voice Pathology Detection, Intelligent Automation and Soft Computing 15 (4) (2009) 667–682.

[151] S. H. Fang, Y. Tsao, M. J. Hsiao, J. Y. Chen, Y. H. Lai, F. C. Lin, C. T. Wang, Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach, Journal of Voice-doi:10.1016/j.jvoice.2018.02.003.

[152] 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.

[153] G. Muhammad, G. Altuwaijri, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, A. Alnasheri, Automatic voice pathology detection and classification using vocal tract area irregularity, Biocybernetics and Biomedical Engineering 36 (2) (2016) 309–317.

[154] G. Muhammad, M. Melhem, Pathological voice detection and binary classification using MPEG-7 audio features, Biomedical Signal Processing and Control 11 (1) (2014) 1–9.

[155] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, Z. Smekal, Towards robust voice pathology detection, Neural Computing and Applications (2018) 1–11.

[156] 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.

[157] J. I. Godino-Llorente, P. Gómez-Vilda, F. Cruz-Roldán, M. Blanco-Velasco, R. Fraile, Pathological Likelihood Index as a Measurement of the Degree of Voice Normality and Perceived Hoarseness, Journal of Voice 24 (6) (2010) 667–677.

[158] D. Martínez González, E. Lleida, A. Ortega, A. Miguel, Score level versus audio level fusion for voice pathology detection on the Saarbrücken Voice Database, Communications in Computer and Information Science 328 CCIS (2012) 110–120.

[159] D. Martínez González, E. Lleida, A. Ortega, A. Miguel, J. Villalba, Voice pathology detection on the Saarbrücken Voice Database with calibration and fusion of scores using multifocal toolkit, Communications in Computer and Information Science 328 CCIS (2012) 99–109.

[160] G. Muhammad, M. F. Alhamid, M. S. Hossain, A. S. Almogren, A. V. Vasilakos, Enhanced Living by Assessing Voice Pathology Using a Co-Occurrence Matrix, Sensors 17 (2) (2017) 267.

[161] M. Hariharan, K. Polat, R. Sindhu, S. Yaacob, A hybrid expert system approach for telemonitoring of vocal fold pathology, Applied Soft Computing 13 (10) (2013) 4148–4161.

[162] M. S. Hossain, G. Muhammad, Healthcare Big Data Voice Pathology Assessment Framework, IEEE Access 4 (2016) 7806–7815.

[163] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, N. Dehak, Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease, Applied Soft Computing 62 (2018) 649–666.

[164] B. Ghoraani, S. Krishnan, A Joint Time-Frequency and Matrix Decomposition Feature Extraction Methodology for Pathological Voice Classification, EURASIP Journal on Advances in Signal Processing 2009 (2009) 1–12.

[165] J. Nayak, P. Bhat, R. Acharya, U. Aithal, Classification and analysis of speech abnormalities, ITBM-RBM 26 (5-6) (2005) 319–327.

[166] H. Ghasemzadeh, M. Tajik Khass, M. Khalil Arjmandi, M. Pooyan, Detection of vocal disorders based on phase space parameters and Lyapunov spectrum, Biomedical Signal Processing and Control 22 (2015) 135–145.

[167] J. J. Jiang, Y. Zhang, C. McGilligan, Chaos in voice, from modeling to measurement, Journal of Voice 20 (1) (2006) 2–17.

[168] M. Kaleem, B. Ghoraani, A. Guergachi, S. Krishnan, Pathological speech signal analysis and classification using empirical mode decomposition., Medical & biological engineering & computing 51 (2013) 811–21.

[169] H. Cordeiro, J. Fonseca, I. Guimarães, C. Meneses, Hierarchical Classification and System Combination for Automatically Identifying Physiological and Neuromuscular Laryngeal Pathologies, Journal of Voice.

[170] T. Drugman, T. Dubuisson, T. Dutoit, Phase-based information for voice pathology detection, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2011, pp. 4612–4615.

[171] T. Drugman, T. Dubuisson, T. Dutoit, On the mutual information between source and filter contributions for voice pathology detection, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2009, pp. 1463–1466.

[172] P. Gómez-Vilda, R. Fernández-Baillo, A. Nieto, F. Díaz, F. Fernández-Camacho, V. Rodellar, A. Álvarez, R. Martínez-Olalla, Evaluation of Voice Pathology Based on the Estimation of Vocal Fold Biomechanical Parameters, Journal of Voice 21 (4) (2007) 450–476.

[173] C. Carmona-Duarte, R. Plamondon, P. Gómez-Vilda, M. A. Ferrer, J. B. Alonso, A. R. M. Londral, Application of the Lognormal Model to the Vocal Tract Movement to Detect Neurological Diseases in Voice, in: Smart Innovation, Systems and Technologies, Vol. 60, 2016, pp. 25–35.

[174] P. Gómez-Vilda, R. Fernández-Baillo, V. Rodellar-Biarge, V. N. Lluis, A. Álvarez-Marquina, L. M. Mazaira-Fernández, R. Martínez-Olalla, J. I. Godino-Llorente, Glottal Source biometrical signature for voice pathology detection, Speech Communication 51 (9) (2009) 759–781.

[175] J. Godino-Llorente, S. Aguilera-Navarro, P. Gomez-Vilda, Automatic detection of voice impairments due to vocal misuse by means of Gaussian mixture models, in: 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2001, pp. 1723–1726.

[176] R. F. Leonarduzzi, G. a. Alzamendi, G. Schlotthauer, M. E. Torres, Wavelet leader multifractal analysis of period and amplitude sequences from sustained vowels, Speech Communication 72 (2015) 1–12.

[177] D. D. Mehta, M. Zañartu, S. W. Feng, H. A. Cheyne, R. E. Hillman, Mobile Voice Health Monitoring Using a Wearable Accelerometer Sensor and a Smartphone Platform, IEEE Transactions on Biomedical Engineering 59 (11) (2012) 3090–3096.

[178] P. Aichinger, I. Roesner, B. Schneider-Stickler, M. Leonhard, D.-M. Denk-Linnert, W. Bigenzahn, A. K. Fuchs, M. Hagmüller, G. Kubin, Towards Objective Voice Assessment: The Diplophonia Diagram, Journal of Voice 31 (2) (2017) 253.e17–253.e26.

[179] M. Döllinger, M. Kunduk, M. Kaltenbacher, S. Vondenhoff, A. Ziethe, U. Eysholdt, C. Bohr, Analysis of vocal fold function from acoustic data simultaneously recorded with high-speed endoscopy, Journal of Voice 26 (6) (2012) 726–733.

[180] L. Verde, G. De Pietro, G. Sannino, A methodology for voice classification based on the personalized fundamental frequency estimation, Biomedical Signal Processing and Control 42 (2018) 134–144.

[181] a. M. Sulter, H. P. Wit, H. K. Schutte, D. G. Miller, A structured approach to voice range profile (phonetogram) analysis., Journal of speech and hearing research 37 (5) (1994) 1076–85.

[182] A. Behrman, C. J. Agresti, E. Blumstein, G. Sharma, Meaningful features of voice range profiles from patients with organic vocal fold pathology: A preliminary study, Journal of Voice 10 (3) (1996) 269–283.

[183] A. E. Hallin, K. Fröst, E. B. Holmberg, M. Södersten, Voice and speech range profiles and Voice Handicap Index for males âĂŤ methodological issues and data, Logopedics Phoniatrics Vocology 37 (2) (2012) 47–61.

[184] J. Goddard, G. Schlotthauer, M. Torres, H. Rufiner, Dimensionality reduction for visualization of normal and pathological speech data, Biomedical Signal Processing and Control 4 (3) (2009) 194–201.

[185] M. FarruÌĄs, J. Hernando, Using Jitter and Shimmer in speaker verification, IET Signal Processing 3 (November 2008) (2009) 247.

[186] M. Shu, J. J. Jiang, M. Willey, The Effect of Moving Window on Acoustic Analysis, Journal of Voice 30 (1) (2016) 5–10.

[187] V. Uloza, E. Padervinskis, A. Vegiene, R. Pribuisiene, V. Saferis, E. Vaiciukynas, A. Gelzinis, A. Verikas, Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening, European Archives of Oto-Rhino-Laryngology 272 (11) (2015) 3391–3399.

[188] 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.

[189] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, G. S. Berke, Comparison of voice analysis systems for perturbation measurement, Journal of Speech & Hearing Research 39 (1) (1996) 126–134.

[190] P. R. Scalassara, C. D. Maciel, R. C. Guido, J. C. Pereira, E. S. Fonseca, A. N. Montagnoli, S. B. Júnior, L. S. Vieira, F. L. Sanchez, Autoregressive decomposition and pole tracking applied to vocal fold nodule signals, Pattern Recognition Letters 28 (11) (2007) 1360–1367.

[191] A. Tsanas, M. a. Little, P. E. McSharry, J. Spielman, L. O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease, IEEE Transactions on Biomedical Engineering 59 (5) (2012) 1264–1271.

[192] A. Tsanas, M. A. Little, P. E. McSharry, L. O. Ramig, Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity, Journal of The Royal Society Interface 8 (59) (2011) 842–855.

[193] E. Yumoto, W. J. Gould, T. Baer, Harmonics to Noise Ratio as hoarseness index of degree of hoarseness, Journal of the Acoustical Society of America 71 (6) (1982) 1544–1549.

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

[195] 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.

[196] 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.

[197] C. Peng, W. Chen, X. Zhu, B. Wan, D. Wei, B. Engineering, Pathological Voice Classification Based on a Single Vowel ' s Acoustic Features, IEEE Seventh International Conference on Computer and Information Technology (2007) 1106–1110.

[198] W. S. Winholtz, L. O. Ramig, Vocal tremor analysis with the Vocal Demodulator., Journal of speech and hearing research 35 (3) (1992) 562–73.

[199] A. Akbari, M. K. Arjmandi, Employing linear prediction residual signal of wavelet sub-bands in automatic detection of laryngeal pathology, Biomedical Signal Processing and Control 18 (2015) 293–302.

[200] Z. Ali, G. Muhammad, M. F. Alhamid, An Automatic Health Monitoring System for Patients Suffering from Voice Complications in Smart Cities, IEEE Access 5.

[201] M. Alsulaiman, Voice pathology assessment systems for dysphonic patients: Detection, classification, and speech recognition, IETE Journal of Research 60 (2) (2014) 156–167.

[202] J. I. Godino-Llorente, S. Aguilera-Navarro, P. Gómez-Vilda, PC, LPCC and MFCC Parameterisation Applied to the Detection of Voice Impairments, in: Sixth International Conference on Spoken Language Processing, 2000.

[203] J. C. Saldanha, T. Ananthakrishna, R. Pinto, Vocal Fold Pathology Assessment Using Mel-Frequency Cepstral Coefficients and Linear Predictive Cepstral Coefficients Features, Journal of Medical Imaging and Health Informatics 4 (2) (2014) 168–173.

[204] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, A. Al-nasheri, T. A. Mesallam, M. Farahat, K. H. Malki, Intra- and Inter-database Study for Arabic, English, and German Databases: Do Conventional Speech Features Detect Voice Pathology?, Journal of Voice 31 (3) (2017) 386.e1–386.e8.

[205] M. S. Hossain, G. Muhammad, A. Alamri, Smart healthcare monitoring: a voice pathology detection paradigm for smart cities, Multimedia Systems (2017) 1–11.

[206] F. Amara, M. Fezari, H. Bourouba, An Improved GMM-SVM System based on Distance Metric for Voice Pathology Detection, Appl. Math. Inf. Sci 10 (3) (2016) 1061–1070.

[207] H. Florian, G. Stemmer, C. Hacker, F. Brugnara, Revising Perceptual Linear Prediction (PLP), Interspeech 2005 (2005) 2997–3000.

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

[209] A. Benba, A. Jilbab, A. Hammouch, Discriminating Between Patients With Parkinson's and Neurological Diseases Using Cepstral Analysis, IEEE Transactions on Neural Systems and Rehabilitation Engineering 24 (10) (2016) 1100–1108.

[210] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, E. Nöth, Spectral and cepstral analyses for Parkinson's disease detection in Spanish vowels and words, Expert Systems 32 (6) (2015) 688–697.

[211] A. Akbari, M. K. Arjmandi, An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features, Biomedical Signal Processing and Control 10 (1) (2014) 209–223.

[212] H. Khadivi Heris, B. Seyed Aghazadeh, M. Nikkhah-Bahrami, Optimal feature selection for the assessment of vocal fold disorders, Computers in Biology and Medicine 39 (10) (2009) 860–868.

[213] K. Umapathy, S. Krishnan, Feature analysis of pathological speech signals using local discriminant bases technique., Medical & biological engineering & computing 43 (4) (2005) 457–64.

[214] K. Umapathy, S. Krishnan, V. Parsa, D. G. Jamieson, Discrimination of pathological voices using a time-frequency approach., IEEE transactions on bio-medical engineering 52 (3) (2005) 421–30.

[215] N. Erfanian Saeedi, F. Almasganj, F. Torabinejad, Support vector wavelet adaptation for pathological voice assessment, Computers in Biology and Medicine 41 (9) (2011) 822–828.

[216] B. Ghoraani, K. Umapathy, L. Sugavaneswaran, S. Krishnan, Pathological Speech Signal Analysis Using Time-Frequency Approaches, Critical Reviews in Biomedical Engineering 40 (1) (2012) 63–95.

[217] J. Hillenbrand, R. A. Cleveland, R. L. Erickson, Acoustic Correlates of Breathy Vocal Quality, Journal of Speech Language and Hearing Research 37 (4) (1994) 769–778.

[218] 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.

[219] P. Grassberger, Randomness, Information, and Complexity, Complexity (1) (2012) 1–20.

[220] M. Costa, A. Goldberger, C. Peng, Multiscale entropy analysis of biological signals, Physical Review E 71 (2) (2005) 1–18.

[221] M. D. Costa, C.-K. Peng, A. L. Goldberger, Multiscale Analysis of Heart Rate Dynamics: Entropy and Time Irreversibility Measures, Cardiovascular Engineering 8 (2) (2008) 88–93.

[222] I. R. Titze, The Myoelastic Aerodynamic Theory of Phonation, 1st Edition, National Center for Voice and Speech, 2006.

[223] I. Steinecke, H. Herzel, Bifurcations in an asymmetric vocal-fold model., The Journal of the Acoustical Society of America 97 (3) (1995) 1874–84.

[224] 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.

[225] R. T. Sataloff, Laryngology (Sataloff's Comprehensive Textbook of Otolaryngology: Head & Neck Surgery), Jaypee Brothers, Medical Publishers Pvt. Limited, 2015.

[226] G. Vaziri, F. Almasganj, R. Behroozmand, Pathological assessment of patients' speech signals using nonlinear dynamical analysis, Computers in Biology and Medicine 40 (1) (2010) 54–63.

[227] L. Matassini, R. Hegger, H. Kantz, C. Manfredi, Analysis of vocal disorders in a feature space, Medical Engineering & Physics 22 (6) (2000) 413–418.

[228] J. Arias-LondoÃśo, J. Godino-Llorente, G. Castellanos-DomÃŋnguez, N. SÃąenz-Lechon, V. Osma-Ruiz, Complexity analysis of pathological voices by means of hidden markov entropy measurements, in: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2009, pp. 2248–2251.

[229] Z. Ali, I. Elamvazuthi, M. Alsulaiman, G. Muhammad, Detection of Voice Pathology using Fractal Dimension in a Multiresolution Analysis of Normal and Disordered Speech Signals, Journal of medical systems 40 (1) (2016) 20.

[230] J. A. Gómez-García, J. I. Godino-Llorente, G. Castellanos-Dominguez, Non uniform Embedding based on Relevance Analysis with reduced computational complexity: Application to the detection of pathologies from biosignal recordings, Neurocomputing 132 (0) (2014) 148–158.

[231] J.-R. Orozco Arroyave, J.-F. Vargas Bonilla, E. Delgado Trejos, Acoustic Analysis and Non Linear Dynamics Applied to Voice Pathology Detection: A Review, Recent Patents on Signal Processing 2 (2) (2012) 96–107.

[232] 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).

[233] Z. Ali, M. Talha, M. Alsulaiman, A Practical Approach: Design and Implementation of a Healthcare Software for Screening of Dysphonic Patients, IEEE Access 3536 (c) (2017) 1–1.

[234] Y. Maryn, M. De Bodt, N. Roy, The Acoustic Voice Quality Index: Toward improved treatment outcomes assessment in voice disorders, Journal of Communication Disorders 43 (3) (2010) 161–174.

[235] M. Fröhlich, D. Michaelis, H. W. Strube, E. Kruse, Acoustic voice quality description : Case studies for different regions of the hoarseness diagram, Advances in Quantitative Laryngoscopy, 2nd âĂŸRound Table (1997) 143–150.

[236] A. Fontes, P. Souza, A. Neto, A. Martins, L. Silveira, Classification System of Pathological Voices Using Correntropy, Mathematical Problems in Engineering 2014 (2014) 1–7.

[237] P. Gómez-Vilda, E. S. Segundo, L. M. Mazaira-Fernández, A. Álvarez-Marquina, V. Rodellar-Biarge, Using Dysphonic Voice to Characterize Speaker's Biometry, Language and Law / Linguagem e Direito 1 (2) (2014) 42–66.

[238] D. Hemmerling, A. Skalski, J. Gajda, Voice data mining for laryngeal pathology assessment, Computers in Biology and Medicine 69 (2016) 270–276.

[239] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, A. Moqarehzadeh, Identification of voice disorders using long-time features and support vector machine with different feature reduction methods, Journal of Voice 25 (6) (2011) e275–e289.

[240] Z. Ali, M. Alsulaiman, I. Elamvazuthi, G. Muhammad, T. A. Mesallam, M. Farahat, K. H. Malki, Voice pathology detection based on the modified voice contour and SVM, Biologically Inspired Cognitive Architectures 15 (2016) 10–18.

[241] S. Shilaskar, A. Ghatol, P. Chatur, Medical decision support system for extremely imbalanced datasets, Information Sciences 384 (2017) 205–219.

[242] N. Erfanian Saeedi, F. Almasganj, Wavelet adaptation for automatic voice disorders sorting, Computers in Biology and Medicine 43 (6) (2013) 699–704.

[243] M. Markaki, Y. Stylianou, Normalized modulation spectral features for cross-database voice pathology detection, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2009, pp. 935–938.

[244] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.

[245] T. Dunstone, N. Yager (Eds.), Biometric System and Data Analysis, Springer US, Boston, MA, 2009.

[246] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145–1159.

[247] R. Fernández-Pozo, J. L. Blanco-Murillo, L. Hernández Gómez, E. López, J. Alcázar-Ramírez, D. Torre-Toledano, Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker Recognition Techniques, EURASIP Journal on Advances in Signal Processing 2009 (1) (2009) 1–12.

[248] A. Montero Benavides, J. L. Blanco Murillo, R. Fernández Pozo, F. Espinoza Cuadros, D. Torre Toledano, J. D. Alcázar-Ramírez, L. A. Hernández Gómez, Formant Frequencies and Bandwidths in Relation to Clinical Variables in an Obstructive Sleep Apnea Population, Journal of Voice 30 (1) (2016) 21–29.

[249] J. Alcázar-Ramírez, R. Fernández-Pozo, J. L. Blanco-Murillo, L. Hernández, L. López, F. Linde, D. Torre-Toledano, Automatic Speaker Recognition Techniques: A new Tool for Sleep Apnoea Diagnosis, Am. J. Respir. Crit. Care Med.

[250] J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. D. Arias-Londoño, S. Murillo-Rendón, G. Castellanos-Domínguez, J. F. Garcés, Nonlinear Dynamics for Hypernasality Detection in Spanish Vowels and Words, Cognitive Computation 5 (4) (2013) 448–457.

[251] P. Gómez-Vilda, V. Rodellar-Biarge, V. Nieto-Lluis, K. L. de Ipiña, A. Álvarez-Marquina, R. Martínez-Olalla, M. Ecay-Torres, P. Martínez-Lage, Phonation biomechanic analysis of Alzheimer's Disease cases, Neurocomputing 167 (2015) 83–93.

[252] K. López-de Ipiña, J.-B. Alonso, C. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage, U. Lizardui, On the Selection of Non-Invasive Methods Based on Speech Analysis Oriented to Automatic Alzheimer Disease Diagnosis, Sensors 13 (5) (2013) 6730–6745.

[253] J. Ortega-Garcia, J. Gonzalez-Rodriguez, V. Marrero-Aguiar, AHUMADA: A large speech corpus in spanish for speaker characterization and identification, Speech Communication 31 (2-3) (2000) 255–264.

[254] S. L. Velleman, B. Z. Pearson, Differentiating Speech Sound Disorders From Phonological Dialect Differences: Implications for Assessment and Intervention, Top Language Disorders 30 (3) (2010) 176–188.

[255] C. Huang, T. Chen, S. Z. Li, E. Chang, J.-L. Zhou, Analysis of speaker variability., in: Proc. INTERSPEECH, no. 49, 2001, pp. 1377–1380.

[256] C. Huang, T. Chen, E. Chang, Accent Issues in Large Vocabulary Continuous Speech Recognition, International Journal of Speech Technology 7 (2/3) (2004) 141–153.

[257] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, a. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens, Automatic speech recognition and speech variability: A review, Speech Communication 49 (10-11) (2007) 763–786.

[258] H. Traunmüller, A. Eriksson, Acoustic effects of variation in vocal effort by men, women, and children, The Journal of the Acoustical Society of America 107 (6) (2000) 3438.

[259] A. L. Rosenthal, S. Y. Lowell, R. H. Colton, Aerodynamic and Acoustic Features of Vocal Effort, Journal of Voice 28 (2) (2014) 144–153.

[260] S. N. Awan, A. Giovinco, J. Owens, Effects of Vocal Intensity and Vowel Type on Cepstral Analysis of Voice, Journal of Voice 26 (5) (2012) 670.e15–670.e20.

[261] D. Z. Huang, F. D. Minifie, H. Kasuya, S. X. Lin, Measures of vocal function during changes in vocal effort level, Journal of Voice 9 (4) (1995) 429–438.

[262] M. Brockmann, C. Storck, P. N. Carding, M. J. Drinnan, Voice Loudness and Gender Effects on Jitter and Shimmer in Healthy Adults, Journal of Speech, Language, and Hearing Research 51 (5) (2008) 1152–1160.

[263] M. Brockmann, M. J. Drinnan, C. Storck, 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, Journal of Voice 25 (1) (2011) 44–53.

[264] P. H. Dejonckere, Effect of louder voicing on acoustical measurements in dysphonic patients, Logopedics Phoniatrics Vocology 23 (2) (1998) 79–84.

[265] L. Chen, X. Mao, Y. Xue, L. L. Cheng, Speech emotion recognition: Features and classification models, Digital Signal Processing 22 (6) (2012) 1154–1160.

[266] S. Steidl, A. Batliner, D. Seppi, B. Schuller, On the Impact of Children's Emotional Speech on Acoustic and Language Models, EURASIP Journal on Audio, Speech, and Music Processing 2010 (14) (2010) 1–14.

[267] B. Schuller, J. Stadermann, G. Rigoll, in: Speech Prosody, 2006.

[268] D. Childers, K. Wu, Automatic recognition of gender by voice, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1988, pp. 603–606.

[269] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, J. Hernández-Cordero, Gender-dependent phonetic refraction for speaker recognition, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, IEEE, 2002, pp. I–149.

[270] T. J. Hixon, G. Weismer, J. D. Hoit, Preclinical Speech Science: Anatomy, Physiology, Acoustics, and Perception, 2nd Edition, Plural Publishing, 2013.

[271] A. Behrman, Speech and Voice Science, Plural Publishing Inc, 2007.

[272] M. Södersten, S. Hertegå rd, B. Hammarberg, Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women., Journal of Voice 9 (2) (1995) 182–197.

[273] H. M. Hanson, E. S. Chuang, Glottal characteristics of male speakers: acoustic correlates and comparison with female data., The Journal of the Acoustical Society of America 106 (2) (1999) 1064–77.

[274] D. H. Klatt, L. C. Klatt, Analysis, synthesis, and perception of voice quality variations among female and male talkers, The Journal of the Acoustical Society of America 87 (2) (1990) 820–857.

[275] K. Wu, D. G. Childers, Gender recognition from speech. Part I: Coarse analysis., The Journal of the Acoustical Society of America 90 (4 Pt 1) (1991) 1828–40.

[276] E. Mendoza, N. Valencia, J. Muñoz, H. Trujillo, Differences in voice quality between men and women: use of the long-term average spectrum (LTAS)., Journal of Voice 10 (1) (1996) 59–66.

[277] H. M. Hanson, Glottal characteristics of female speakers: acoustic correlates., The Journal of the Acoustical Society of America 101 (1) (1997) 466–81.

[278] E. Holmberg, R. Hillman, J. Perkell, Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice, The Journal of the Acoustical Society of America 84 (2) (1988) 511–529.

[279] R. B. Monsen, a. M. Engebretson, Study of variations in the male and female glottal wave., The Journal of the Acoustical Society of America 62 (4) (1977) 981–93.

[280] J. Abitbol, P. Abitbol, B. Abitbol, Sex hormones and the female voice., Journal of Voice 13 (3) (1999) 424–46.

[281] B. Gold, N. Morgan, D. Ellis, Speech and Audio Signal Processing: Processing and Perception of Speech and Music, Wiley-Interscience, 2011.

[282] D. D. Deliyski, H. S. Shaw, M. K. Evans, Adverse Effects of Environmental Noise on Acoustic Voice Quality Measurements, Journal of Voice 19 (1) (2005) 15–28. doi:10.1016/j.jvoice.2004.07.003.

[283] A. H. Poorjam, J. R. Jensen, M. A. Little, M. G. Christensen, Dominant Distortion Classification for Pre-Processing of Vowels in Remote Biomedical Voice Analysis, in: Interspeech 2017, Vol. 2017-Augus, 2017, pp. 289–293.