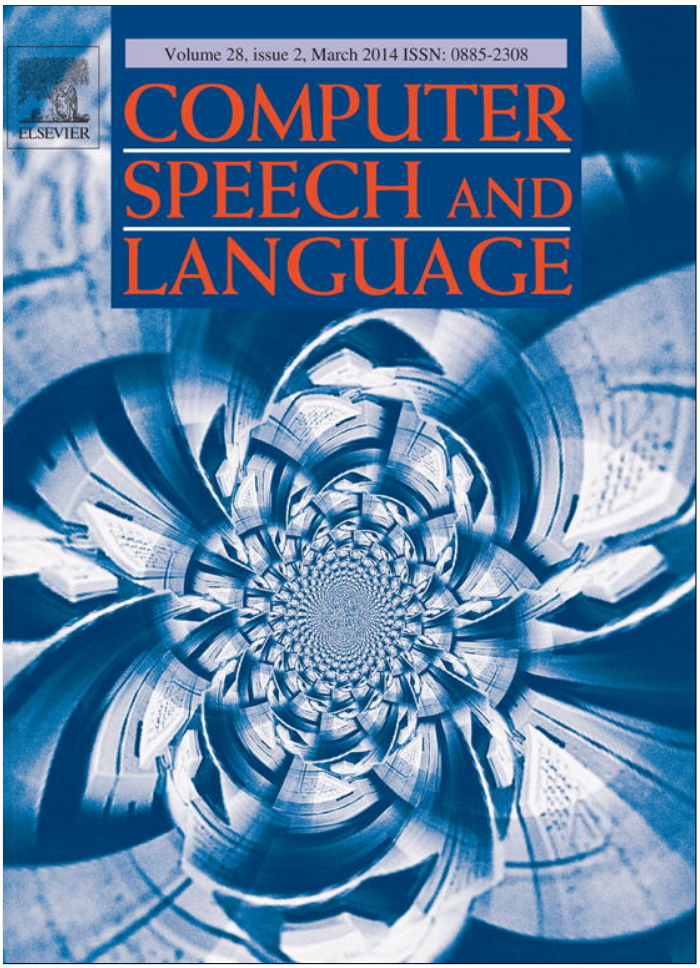

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 434–452

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

# Analysis of voice features related to obstructive sleep apnoea and their application in diagnosis support☆

Ana Montero Benavides[a], Rubén Fernández Pozo[a], Doroteo T. Toledano[b,*], José Luis Blanco Murillo[a], Eduardo López Gonzalo[a], Luis Hernández Gómez[a]

[a] *Signal, Systems and Radiocommunications Department, Universidad Politécnica de Madrid, Spain*
[b] *ATVS Biometric Recognition Group, Universidad Autónoma de Madrid, Spain*

Received 1 March 2012; received in revised form 31 July 2013; accepted 5 August 2013
Available online 22 August 2013

**Abstract**

*Obstructive sleep apnoea* (OSA) is a highly prevalent disease affecting an estimated 2–4% of the adult male population that is difficult and very costly to diagnose because symptoms can remain unnoticed for years. The reference diagnostic method, *Polysomnography* (PSG), requires the patient to spend a night at the hospital monitored by specialized equipment. Therefore fast and less costly screening techniques are normally applied for setting priorities to proceed to the polysomnography diagnosis. In this article the use of speech analysis is proposed as an alternative or complement to existing screening methods. A set of voice features that could be related to apnoea are defined, based on previous results from other authors and our own analysis. These features are analyzed first in isolation and then in combination to assess their discriminative power to classify voices as corresponding to apnoea patients and healthy subjects. This analysis is performed in a database containing three repetitions of four carefully designed sentences read by 40 healthy subjects and 42 subjects suffering from severe apnoea. As a result of the analysis, a *linear discriminant model* (LDA) was defined including a subset of eight features (signal-to-disperiodicity ratio, a nasality measure, harmonic-to-noise ratio, jitter, difference between third and second formants on a specific vowel, duration of two of the sentences and the percentage of silence in one of the sentences). This model was tested on a separate database containing 20 healthy and 20 apnoea subjects yielding a sensitivity of 85% and a specificity of 75%, with a F1-measure of 81%. These results indicate that the proposed method, only requiring a few minutes to record and analyze the patient's voice during the visit to the specialist, could help in the development of a non-intrusive, fast and convenient PSG-complementary screening technique for OSA.


*Keywords:* OSA screening tools; Apnoea discrimination; Voice features; Feature selection; Voice pathology

*Abbreviations:* OSA, obstructive sleep apnoea; PSG, polysomnography; AHI, apnoea–hypopnoea index; BMI, body mass index; CER, classification error rate; EER, equal error rate; LDA, linear discriminant analysis; MLR, multiple linear regression.

☆ This paper has been recommended for acceptance by 'Dr. Björn Schuller'.

* Corresponding author at: Av. Francisco Tomás y Valiente 11, 28049 Madrid, Spain. Tel.: +34 419 2217; fax: +34 419 2207.
*E-mail addresses:* ana.montero@gaps.ssr.upm.es (A. Montero Benavides), ruben@gaps.ssr.upm.es (R. Fernández Pozo), doroteo.torre@uam.es (D.T. Toledano), jlblanco@gaps.ssr.upm.es (J.L. Blanco Murillo), eduardo@gaps.ssr.upm.es (E. López Gonzalo), luisalfonso.hernandez@upm.es (L. Hernández Gómez).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.08.002

## 1. Introduction

*Obstructive sleep apnoea* (OSA) is a highly prevalent disease (Puertas et al., 2005), affecting an estimated 2–4% of the male population between the ages of 30 and 60. It is characterized by recurring episodes of sleep-related collapse of the upper airway at the level of the pharynx and it is usually associated with loud snoring and increased daytime sleepiness. Besides OSA, there are two other forms of sleep disease related to apnoeas: *central sleep apnoea* (CSA) and *combined* or *mixed sleep apnoea* that only constitute the 0.4% and 15% of cases respectively (Morgenthaler et al., 2006); therefore their relevance can be considered minor compared to OSA, and for this reason we will focus on the study of OSA. A common criterion for diagnosing OSA is having an *apnoea–hypopnoea index* (AHI) over 15. AHI represents the average number of apnoeas and hypopnoeas per hour of sleep. OSA is a serious threat to an individual's health if not treated. The condition is a risk factor for hypertension and cardiovascular diseases, it is usually related to traffic accidents caused by somnolent drivers, and it can lead to a poor quality of life and impaired work performance. At present, the most effective and widespread treatment for OSA is nasal CPAP (*continuous positive airway pressure*) which prevents apnoea episodes by providing a pneumatic splint to the airway.

OSA can be diagnosed on the basis of a physical examination and a medical history that includes questions about habits or the degree of snoring and the daytime sleepiness, although both screening characteristics by themselves are not always accurate discriminative features for OSA (Teculescu, 1998), and a full overnight sleep study is needed to confirm the disorder. The procedure is known as conventional *polysomnography* (PSG), which involves the recording of neuroelectrophysiological and cardiorespiratory variables. Nevertheless, this diagnostic procedure is expensive and time-consuming, and patients usually have to endure a waiting list of several years before the test is done, since recently the demand for diagnostic studies for OSA has recently increased (Puertas et al., 2005). There is, therefore, a strong need for alternatives methods to PSG of screening apnoea patients in order to detect patients with high risk of OSA and reduce these considerable delays. In the published literature, we found numerous efforts to devise PSG-complementary clinical methods of predicting OSA (Ramachandran and Josephs, 2009). These methods are broadly classified as questionnaires about sleepiness (Ahmadi et al., 2008) and clinical prediction models (Gurubhagavatula et al., 2004; Friedman et al., 2010) using anthropometric characteristics (i.e. BMI or morphology) and epidemiological and medical history parameters (i.e. snoring).

This need of searching for novel OSA screening tools gives a good reason for this contribution, which investigates the acoustical characteristics of the voice in patients with OSA for the purpose of learning whether OSA may be detected using speech analysis. The acoustic properties of voice from speakers suffering OSA are not well understood as not much research has been carried out in this area. However, some studies have suggested that certain abnormalities in voice *articulation*, *phonation* and *resonance* may be connected to the condition (Fox and Monoson, 1989). In order to have a controlled experimental framework to study apnoea voice characterization we have collected a speech database (Fernández et al., 2008) designed following linguistic and phonetic criteria derived from previous research in the field. Our work is focused on continuous speech rather than on sustained vowels, the latter being the standard approach in pathological voice analysis. This is justified because we are interested in the acoustic analysis of the characteristics of the speech signal in specific linguistic and phonetic contexts, and also in the analysis of other features not used previously in the characterization of apnoea voices such as the dynamics of speech, and in particular the patterns of pauses, which we hypothesized as valuable discriminative features. Together with these speech-dynamics and pause features, OSA-related voice characteristics associated to articulation, phonation and nasalization were evaluated over apnoea and non-apnoea voices to have a contrastive study on the acoustic discrimination attainable in our database using either individual features or a combination of some or all of the proposed features. In addition, and due to the crucial dependency of OSA disease with BMI (body mass index) and age factors, we finished with a study about relations between OSA-related voice features and these two important factors.

The rest of this article is organized as follows: Section 2 presents the main physiological characteristics of OSA patients and acoustic characteristics of their voices, as described in the previous literature. Section 3 describes the experimental setup, including the description, design and capture of the corpora used in this study, the definition and selection of features related to apnoea and the details of the experimental setup. Section 4 presents detailed results

comparing the features individually and in combination, as well as the final results obtained. Finally, discussion, conclusions and a brief outline of future research are given in Section 5.

## 2. Obstructive sleep apnoea: pathogenesis, physiological and acoustic characteristics

This section sets the theoretical background on *obstructive sleep apnoea*, describing its pathogenesis, its physiological characteristics and, more importantly, how these physiological characteristics can impact on the voice of OSA patients. This study then shows us which voice features can be used to discriminate between OSA voices and healthy voices as a first stage toward OSA screening based on speech analysis.

### 2.1. OSA pathogenesis

The pathogenesis of OSA defines a set of anatomic and physiological factors that constrict space for the soft tissues surrounding the pharynx and predispose to upper airway collapse during sleep (Ryan and Bradley, 2005). These pharyngeal anatomic factors obviously influence to the upper airway structure, so it seems reasonable to consider that these patients could have specific pathogenic features on their voices, whose elucidation facilitates the development of voice-related OSA screening techniques as presented in this work. Furthermore, it is worth noting here that OSA is an anatomic disease that may have been favored by the evolutionary adaptations in man's upper respiratory tract to facilitate speech, a phenomenon that Jared Diamond calls "The Great Leap Forward" (Davidson, 2003). That is, anatomic changes related to the development of voice, speech and language, may also have provided the structural basis for the occurrence of OSA.

### 2.2. Physiological features of OSA patients and acoustic characteristics of their voices

At present neither the physiological peculiarities nor the acoustic characteristics of speech in apnoea patients are well understood. However, as we have seen before, since the upper airways are affected by OSA, it is logical to assume the existence of specific characteristics of the speech signals of patients with OSA. Evidences on this hypothesis have been provided in a few remarkable references, but perhaps the most valuable information can be found in the work of Fox and Monoson (1989), a perceptual study in which skilled judges were asked to compare voices of apnoea patients with those of a healthy group (control group). Although acoustic cues for differences between both groups of speakers are somewhat contradictory and unclear, this study has pointed out several differences and has confirmed that the apnoea group voices' had certain anomalies that might be due to an altered structure or function of their upper airway. Consequently, the occurrence of a speech disorder in OSA population should be expected including anomalies in *articulation*, *phonation* and *resonance*. These anomalies can be described as follows:

- *Articulatory anomalies*: A neuromotor dysfunction could be found in apnoea population due to a "*lack of regulated innervations to the breathing musculature or upper airway muscle hypotonus.*" This dysfunction is normally related to speech disorders, especially *dysarthria*. There are several types of dysarthria, resulting in various different acoustic features, but all types affect the articulation of consonants and vowels causing the slurring of speech.
- *Phonation anomalies*: One of the conclusions of Fox and Monoson's work was that the most discriminative feature to distinguish apnoea subjects from healthy subjects was related to phonation anomalies. This may be surprising considering that phonatory organs, and in particular the larynx, should have a very limited role in the physiopathology of the disease. This study found certain roughness alterations in the speech of apnoea patients related to a lack of periodicity of the sounds. One of the hypotheses still to be confirmed is that these anomalies may be due to the characteristic pattern and heavy snoring of sleep apnoea patients, which can cause inflammation in vocal cords (Boyd et al., 2004) and also fatigue due to the extra respiratory effort in OSA patients (Payne et al., 2006).
- *Resonance anomalies*: OSA voices can also present abnormal resonances that might be due to an altered structure or function of the velopharyngeal mechanism, related to the coupling of the vocal tract with the nasal cavity. As a result, an abnormal vocal quality characteristic can be expected in the speech of these patients, that may be revealed through two features:
  - First, speakers with a defective velopharyngeal mechanism can produce speech with inappropriate nasal resonance. The term nasalization can refer to two different phenomena in the context of speech; *hyponasality* (no nasalization

is produced when the sound should be nasal) and *hypernasality* (nasalization during the production of non-nasal sounds). In Spanish, vowels only acquire either a nasal or a non-nasal quality depending on the presence or absence of adjacent nasal consonants. The work by Fox and Monoson (1989), concluded that these resonance abnormalities could not be perceived either as hyponasality or as hypernasality. Instead of this, OSA speakers exhibited smaller intra-speaker differences between non-nasal and nasal vowels due to this velopharyngeal dysfunction. The term applied to this speech disorder is "cul-de-sac" resonance that causes the sound to be perceived as if it were resonating in a blind chamber. This matches perfectly with the results of our previous work (Blanco et al., 2009) where it is confirmed that apnoea speakers showed smaller intra-speaker differences between non-nasal and nasal vowels than healthy speakers, measuring distances between statistical models.

- Secondly, due to this velopharyngeal dysfunction, different formant frequency values between apnoea and healthy voices can be expected. For instance, according to Hidalgo and Quilis (2002) the position of the third formant is related to the size of the velopharyngeal opening, and as a consequence, OSA-characteristic lowering of the velum might produce higher third formant frequencies in apnoea patient's voices. This performance was also studied by Robb et al. (1997), in which vocal tract acoustic resonance was evaluated in a group of OSA males. Statistically significant differences were found in formant frequency and bandwidth values between apnoea and healthy groups. In particular, first and second formant values among the OSA group were generally lower than those in the control group. The lower values were attributed to an OSA-specific altered structure of upper airway and a greater vocal tract length. This finding has been reported in cephalometric studies indicating that the distance from the hyoid bone to the mandibular plane is significantly longer in patients with OSA compared to non-OSA individuals (Maltais et al., 1991).

Despite the sparse literature focusing on the specific characteristics of OSA voices, there are other relevant references which have tackled this topic. The work by Fiz et al. (1993) is a good reference on vowel sounds analysis considering direct inspection on the spectral representation of the collected data. Our previous work (Fernández et al., 2009) reinforces the existence of differences between acoustic features of OSA and non-OSA speakers. In this piece of research, generative statistical modeling techniques were applied to describe the acoustic space of apnoea voices in a similar way to that being used in speaker recognition systems and using continuous speech, unlike most research in pathological voices analysis. Other related and relevant reference addressing the use of both connected and sustained speech for OSA detection can be found in our previous work (Blanco et al., 2011).

## 3. Acoustic analysis of apnoea voices

In this section we start by describing the speech corpora we have designed and collected to study voices of patients with OSA, then we discuss several specific acoustic features to characterize and detect apnoea voices, finally we present the experimental setup and tools that we have used through our study.

### *3.1. Speech apnoea corpus design*

The analysis of specific traces within speech signals that can be related to OSA requires a special effort to design and collect a consistent database. First, it should include recordings from speakers suffering from OSA and healthy ones, outlining a pair of reasonably homogeneous groups of speakers. Second, the recordings should be collected in the same acoustic conditions and following specifically designed criteria.

As we pointed out in the introduction, in this contribution we study how to apply speech processing techniques to automatically detect OSA-related traits in continuous speech. Thus, in the present paper we will not be concerned with sustained vowels, even though this has been the most common approach in the literature on pathological voice analysis with certain advantages related to a more time and speaker-style invariant voice parameters. Nevertheless, analyzing continuous speech may well give rise to other possibilities because certain traits of pathological voice patterns, and in particular those of OSA patients, could then be detected in different sound categories (i.e. nasals) and also in the co-articulation between adjacent sound units. Finally, other interesting possibility brought about by the use of continuous speech, and novel to our knowledge, is the analysis of speech dynamics, and in particular the pause patterns, to detect OSA.

The design of our corpus was based on covering all the relevant linguistic/phonetic contexts in which physiological OSA-related peculiarities derived from the acoustic as described in the previous work of Fox and Monoson (1989).

These peculiarities include the articulation, phonation and resonance anomalies revealed in the previous research review (Section 2.2), so we designed sentences that include instances of the following specific phonetic and linguistic contexts:

- In relation to *articulatory anomalies*, we collected voiced sounds affected by certain preceding phonemes that have their primary locus of articulation near the back of the oral cavity (specifically, guttural sounds as Spanish voiced and voiceless velar plosives/g/and/k/respectively). This anatomical region has been shown to exhibit physical anomalies in speakers suffering from OSA, so it is reasonable to suspect that different coarticulatory effects may occur in these phonemes with apnoea and healthy speakers.
- With regard to *phonation anomalies*, we included continuous voiced sounds to explore irregular phonation patterns (as the lack of periodicity) that could be related to fatigue and respiratory troubles in apnoea patients (Payne et al., 2006).
- Finally, to look at *resonance anomalies*, we designed sentences that allow intra-speaker variation measurements; that is, assessing differential voice features for each speaker, for instance to compare the degree of vowel nasalization in nasal and non-nasal contexts. Also we included specific allophones of the vowel phonemes where may occur some resonance anomalies referred to expected different formant values in OSA patients (Robb et al., 1997).

The speech corpora contain readings of four sentences in Spanish repeated three times by each speaker. The three repetitions of each sentence were used to calculate the averaged value for each voice feature. The sentences used were the following (including their phonetic transcription following International Phonetic Alphabet -IPA- and their English translation), with the different melodic groups underlined separately:

| | | | |
|---|---|---|---|
| 1. | Francia, Suiza y Hungría | ya hicieron causa común. | |
| | ‘fraN θja ‘suj θa i uŋ ‘gri a | ya j ‘θje roŋ ‘kaw sa ko ‘mun | |
| | France, Switzerland and Hungary | already made common cause | |
| 2. | Julián no vio la manga roja | que ellos buscan, | en ningún almacén. |
| | xu ‘ljan no ‘βjo la ‘maŋ ga ‘řo xa | ke ‘e ʎoz ‘βus kan | en niŋ ‘gun al ma ‘θen |
| | Julián did not see the red sleeve | they seek, | in any store |
| 3. | Juan no puso la taza rota | que tanto le gusta | en el aljibe. |
| | xwan no ‘pu so la ‘ta θa ‘řo ta | ke ‘taN to le ‘ɣus ta | en el al ‘xi βe |
| | Juan did not put the broken cup | that he likes so much | in the tank |
| 4. | Miguel y Manu llamarán entre ocho y nueve y media. | | |
| | mi ‘ɣel i ‘ma nu ʎa ma ‘ran ‘eN tre ‘o t∫o i ‘nwe βe i ‘me ðja | | |
| | Miguel and Manu will call between eight and half past nine | | |

The first sentence was taken from the ALBAYZIN database, a standard phonetically balanced speech database for Spanish (Moreno et al., 1993). It was chosen because it contains some interesting allophones of the vowel phoneme/i/([i] and [j] in IPA notation), where may be expected the different formant values for OSA patients. It also contains examples of guttural sounds as voiced and voiceless velar plosives/g/and/k/respectively.

The second and third sentences, both negative statements, have a similar grammatical and intonation structure. They are potentially useful for contrastive studies of vowels in similar positions in the sentence but in different linguistic contexts. Some examples of these contrastive pairs arise from comparing a nasal context, first/a/of “m**a**nga roja” (’maŋ ga ‘řo xa), with a neutral context, first/a/of “t**a**za rota” (’ta θa ‘řo ta). As we mentioned in the previous section, these contrastive analyses could be very helpful to confirm whether indeed the voices of speakers with apnoea have an abnormal nasality and display smaller intra-speaker differences between non-nasal and nasal vowels due to OSA-related velopharyngeal dysfunction (Fox and Monoson, 1989).

The fourth sentence has a single and relatively long melodic group containing mainly voiced sounds. The rationale for this fourth sentence is that apnoea speakers usually show fatigue in the upper airway muscles due to respiratory problems. Therefore, this sentence may be helpful to discover phonation anomalies during the sustained generation of voiced sounds, or the presence of differences in speech dynamics, and in particular pause patterns.

Given that the linguistic content of the recordings was known, we aligned the recordings with the phonetic content of the sentences using standard automatic speech recognition technology based on *hidden Markov models* (HMMs). The alignment of the recordings allowed us to extract information on particular phonemes and particular phonetic

contexts and also to analyze other features such as the distribution of pauses in pronunciations. More accurate phonetic alignment could be achieved using improved methods as that proposed by Toledano et al. (2003), but the analysis of the influence of precision in the segmentation on apnoea discrimination results have not yet been considered and remains as future work.

### 3.2. Selection and definition of voice features related to apnoea

We defined a set of voice features that will be analyzed on the speech corpus in order to determine its utility in sleep apnoea screening, both in isolation and combined. The definition of these features is based on:

- Other parameters traditionally used in the diagnosis of pathologic voices.
- Previous studies about the characteristics of the voice of apnoea patients.
- A preliminary perceptual informal analysis of our corpus.

The set of features is also directly related to the types of peculiarities that were detected in previous works for apnoea voices (Fox and Monoson, 1989), and henceforth we can organize each of the proposed features as features related to anomalies in articulation, phonation and resonance. Most features are clearly related to one of these abnormalities, although one of them ($F3 - F2\_i$) may be related to more than one of these types of anomalies.

#### 3.2.1. Difference between F3 and F2 in the phone/i/($F3 - F2\_i$)

An interesting conclusion from our initial perceptual contrastive study was that, when comparing the distance between the third (F3) and second formant (F2) for the vowel/i/, differences between the apnoea and control groups were found. For apnoea speakers this value was greater, and this was especially clear in diphthongs with/i/as the stressed vowel, as in the Spanish word "Suiza" ('suj θa) of the first sentence (see Fig. 1). This distance F3 − F2 is obtained as the average difference of the values for these two formants. We measured absolute distances in spite of the fact that the location of the formants is speaker-dependent. Nevertheless, we considered that normalization was not necessary because our database contains only male subjects with similar relevant physical characteristics, and the formants should lie roughly in the same regions for all of our speakers. We measured this feature in utterances of the first sentence in our corpus listed above, which contains good examples of stressed allophones for the vowel/i/.

This finding is in agreement with Robb's conclusion that the F2 formant value in the vowels produced by apnoea subjects is lower than healthy ones due to greater length of the vocal tract of OSA patients (Robb et al., 1997); but also, and perhaps more importantly, with the OSA-characteristically abnormal velopharyngeal opening and lowering of the velum which may cause higher third formant frequencies (Hidalgo and Quilis, 2002). Therefore, this is a voice feature related to articulatory and resonance OSA-related anomalies.

#### 3.2.2. Phonation features (HNR, Jitter, Shimmer, SDRseg)

Trying to characterize phonation abnormalities related to OSA patients, we will study several acoustic features traditionally used in the detection of pathologic voices: *harmonic-to-noise ratio* (HNR), *signal-to-disperiodicity ratio* (*SDRseg*) and features that measure the perturbation of fundamental frequency (*Jitter*) and energy (*Shimmer*). This type of features is generally measured on sustained vowels, although its use has been successfully extended to continuous speech (Parsa and Jamieson, 2001). A brief description of each of these features follows:

- *Harmonic-to-noise ratio* (HNR) is a phonation feature used to estimate the level of noise in human voice signals. HNR estimation is based on calculating the ratio of the energy of the harmonics to the noise energy (measured in dB) and can be accomplished in two ways: (1) on a time-domain basis; and (2) on a frequency-domain basis. HNR is a measurement of voice pureness where 'normal' voices will tend to have a higher HNR than a 'pathological' voice. In our case, we employ a time-domain approach (Boersma, 1993) that will be computed on voiced segments in connected speech.
- *Segmental signal-to-dysperiodicity ratio* (SDRseg). *Disperiodicity* is a common symptom of voice disorders that refers to anomalies in the glottal excitation signal generated by the vibrating vocal folds and the glottal airflow. Segmental signal-to-dysperiodicity ratio (SDRseg), measured in dB, is a parameter used to summarize these vocal dysperiodicities in connected speech. Many variants of the signal-to-dysperiodicity ratio are known to be good

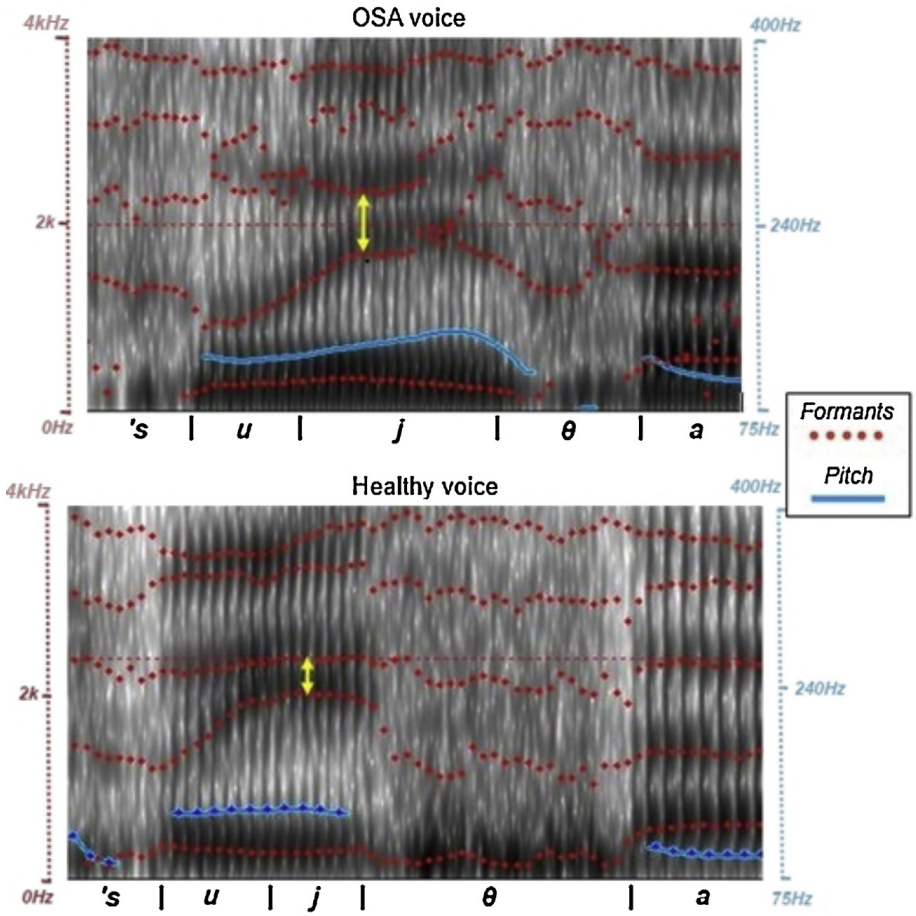


Fig. 1. Differences between third and second formant for the vowel/i/in the word "Suiza" ('suj θa), for an apnoea speaker (above) and a control group speaker (below).

estimators of perceived speech quality, among which we implemented a variant of the algorithm "*bidirectional distal linear predictive analysis*" presented in the work of Bettens et al. (2005). SDRSeg is used to predict the perceived degree of hoarseness of a voice.

- *Jitter* measures perturbations in voice production, especially in vowel phonation, due to small fluctuations in the opening and closing times of the vocal chords. Jitter can be used to discriminate between healthy and dysphonic speakers. A high degree of jitter occurs in a voice with roughness. In our case, we computed jitter on vocal segments of continuous speech (measured in seconds).
- Likewise *Shimmer* measures perturbations in glottal cycle lengths related to the fluctuations of amplitude (measured in dB). It is again calculated on vocal segments of continuous speech.

Specifically, these four features (HNR, SDRseg, Jitter and Shimmer) were computed only for the voice sounds of the fourth sentence of the corpus. In theory a lower HNR and SDRseg should be related to a more pathological voice, while a lower Jitter and Shimmer should correspond to a less pathological voice.

### *3.2.3. Nasality measurements (a1h1max800_Nasal, a1h1max800_Non-nasal, a1h1max800_Diff)*

In the work of Fox and Monoson (1989), the existence of certain irregular resonance characteristics and some abnormalities in nasalization on apnoea voices was hypothesized for the first time. The hypothesis to be confirmed is an unusual variation of nasalization between nasal and non-nasal contexts in apnoea subjects, exhibiting smaller intra-speaker differences between these contexts. In order to try to characterize these anomalies we used one of the nine best parameters for the measurement of nasal character of vowels proposed in the Ph.D. Thesis of Pruthi (2007). This feature is based on the difference between the amplitude of the first formant (A1) and the first harmonic (H1). The first formant is taken as the maximum in the 0–800 Hz frequency band (this explains the name of the feature

Table 1
OSA-related features under analysis.

| Feature # | Feature Name | Description | Related to abnormalities in |
|---|---|---|---|
| 1 | F3-F2_i | Difference between F3 and F2 in the phone/i/of the first sentence of the corpus (Hz) | Articulation and resonance |
| 2 | HNR | Harmonic to Noise Ratio (dB) measured on the fourth sentence | Phonation |
| 3 | Jitter | Jitter (s) measured on the fourth sentence | Phonation |
| 4 | SDRseg | Segmental Signal to Disperiodicity Ratio SDR (dB) measured on the fourth sentence | Phonation |
| 5 | Shimmer | Shimmer (dB) measured on the fourth sentence | Phonation |
| 6 | a1h1max800_Nasal | Nasality measure for vowels in nasal contexts (second sentence of corpus) | Resonance |
| 7 | a1h1max800_Non-nasal | Nasality measure for vowels in non-nasal contexts (third sentence of corpus) | Resonance |
| 8 | a1h1max800_Diff | Difference between a1h1max800_Nasal and a1h1max800_Non-nasal | Resonance |
| 9 | Duration1 | Average duration of sentence one of the corpus (s) | Articulation |
| 10 | Duration2 | Average duration of sentence two of the corpus (s) | Articulation |
| 11 | Duration3 | Average duration of sentence three of the corpus (s) | Articulation |
| 12 | Duration4 | Average duration of sentence four of the corpus (s) | Articulation |
| 13 | PercSil1 | Average Percentage of Silences in sentence one of the corpus (%) | Articulation |
| 14 | PercSil2 | Average Percentage of Silences in sentence two of the corpus (%) | Articulation |
| 15 | PercSil3 | Average Percentage of Silences in sentence three of the corpus (%) | Articulation |
| 16 | PercSil4 | Average Percentage of Silences in sentence four of the corpus (%) | Articulation |

*a1h1max800*) and the first harmonic is the maximum closest to 0 Hz with amplitude over 10 dB and wider that 80 Hz in bandwidth. These thresholds were defined in the latter reference. This measure is indicative of the vowel nasal character, being lower for nasal vocal segments as compared with non-nasal ones (i.e. it can be interpreted more as a measure of hyponasality than hypernasality since it is higher for less nasality).

In our speech corpus, this parameter was measured separately on vowels in nasal contexts (*a1h1max800_Nasal*) and vowels in non-nasal contexts (*a1h1max800_Non-nasal*), and was also computed the difference between both contexts (*a1h1max800_Diff*). In particular, in nasal contexts it was measured on the highlighted/a/phoneme of the words "m**a**nga" and "busc**a**n" (in the second sentence of the corpus); in non-nasal contexts it was measured on the highlighted/a/phoneme of the words "t**a**za" and "gust**a**" (in the third sentence of the corpus).

### *3.2.4. Sentence duration and silence features (Duration1-4, PercSil1-4)*

The last features we will analyze are those related to the duration of the pronunciation of the sentences and the percentage of silences in the sentence (considering only silences in the middle of the sentence, not before or after). These measurements, which can be categorized as suprasegmental features, have already been successfully applied in speaker recognition domain (Reynolds et al., 2003) and in the characterization of pathological voices (Zhang and Jiang, 2008). Our motivation to include these features in our study is based on the conclusions of the work of Fox and Monoson (1989), regarding the phonation anomalies of apnoea subjects. We already argued that these abnormalities in glottal activity could produce fatigue in the muscles of the larynx and respiratory troubles, which could result in difficulty of apnoea subjects to pronounce sentences without pauses (such as the fourth sentence in our corpus). Therefore we would expect more sentence duration and more silences in the apnoea patients than in the control subjects.

In order to assess the importance of the length and absence of pauses in the sentence we have measured these features for each of the sentences separately, and therefore we have four sentence durations and four percentages of silences per subject.

### *3.2.5. Summary of selected features to characterize apnoea speech*

The following table summarizes the 16 OSA-related features considered in this study that have been described in detail in Sections 3.2.1–3.2.4 (Table 1).

Table 2
Distribution of age and BMI in healthy and apnoea speakers in the databases.

| Database | Group | Number subjects | Age | | | BMI | | |
|---|---|---|---|---|---|---|---|---|
| | | | Mean | Std. dev. | *t*-test *p*-Value | Mean | Std. dev. | *t*-test *p*-Value |
| TRAIN | Control | 40 | 42.2 | 8.8 | **0.002*** | 26.2 | 3.9 | **0.000*** |
| | Apnoea | 42 | 49.5 | 10.8 | | 32.8 | 5.4 | |
| TEST | Control | 20 | 46.5 | 7.8 | 0.936 | 29.1 | 3.29 | 0.086 |
| | Apnoea | 20 | 46.3 | 7.7 | | 30.7 | 2.79 | |

* Statistically significant differences found at the 95% confidence level.

### 3.3. Experimental setup

#### 3.3.1. Subject selection

Speech material was recorded in the Respiratory Department at *Hospital Clínico Universitario of Málaga*, Spain. Two different databases were selected: a *training* database and a *test* database. The training database contains the readings of 82 male subjects; 42 of them suffer from severe sleep apnoea (AHI > 30), and the rest are either healthy subjects or only have mild OSA (AHI < 10). The test database contains readings of 40 different male subjects, 20 suffering from severe sleep apnoea (AHI > 30) and 20 that are either healthy subjects or only have mild OSA (AHI < 10). Difference in the severity of the condition between both groups is important since medical doctors expressed their interest in identifying subjects in major need for treatment. Such interest thus leads to focus on severe cases and the specific age ranges (over 40–45 years old) where OSA affects men mostly.

We have selected subjects in both groups trying to have physical characteristics as similar as possible. Table 2 shows age and body mass index (BMI) statistics for the two groups (*Apnoea* and *Control*) of the two databases. In order to study the statistical differences of means between the Apnoea and Control groups, a commonly *t*-test in both, train and test databases, has been carried out at the 95% confidence value. We have tested the null hypothesis of equal means between groups and found that there are statistical significant differences for age and BMI in the train database. On the other hand, for the test database, the conclusion is that, at the 95% confidence value, there are not statistical differences. This conclusions are consistent with Table 2 where we can grasp the train database has greater variability in both of the features age and BMI than the test database where those features are more controlled. This test group is more balanced in these features so, as will be discussed in section 4, this controlled environment will allow us to test the potential of speech features discrimination independently from age and BMI factors.

#### 3.3.2. Recording equipment

The recording equipment consisted of a standard laptop computer equipped with a *SP500 Plantronics* headset microphone with A/D conversion and digital data exchange through a USB-port. Speech was recorded using a sampling frequency of 16 kHz in an acoustically isolated room under the supervision of an expert. Subjects read the sentences listed above in a comfortable and natural mode, whereas the expert's role was to control the recording quality:

- By making sure there are no background noises, artifacts between or after the uttered sentence, and checking that there was neither distortion nor saturation in the speech signal.
- And assuring that speakers pronounce the correct words without making mistakes such as uttering different words.

The speech material for the apnoea group was recorded and collected in two different sessions: one just before being diagnosed (at the same time the control group was recorded) and the other after several months under CPAP treatment. This will allow the future study of the progress of apnoea voice characteristics for a particular patient before and after treatment. In this work we have only used the first session of the apnoea patients and the single session of the healthy subjects. The total length of the speech data is over 1.4 h for the train database and 45 min for the test database.

#### 3.3.3. Voice and data analysis tools

Regarding to voice analysis tools used in this study, we applied Praat software (Boersma and Weenink, 2006) to compute formant values, *harmonic-to-noise ratio* (HNR), *Jitter* and *Shimmer*. The *signal-to-dysperiodicity ratio*

(SDRseg) was implemented with Matlab, which was also used for feature selection and classification algorithms (described in the following section). Finally, the automatic phonetic segmentation of the recordings was obtained using standard speech recognition technology based on the HTK Toolkit (Young, 2002) and a set of acoustic models trained on the phonetic corpus of the ALBAYZIN database (Moreno et al., 1993).

## 4. Analysis and results

The main goals of our analysis are:

- *Analyze each one of the selected features* in order to determine their utility to discriminate voices of OSA patients from voices of healthy subjects.
- Propose *ways to combine those features* in order to obtain a classifier that could be helpful to assist in OSA screening.
- *Measure the performance of such classifiers* both in terms of classification error and in the characteristic metrics used in clinical practice.

After measuring the proposed voice features on our speech databases and organizing properly these measurements, we have structured our analysis and results section into three parts:

- Feature analysis and classification model training using the training database (Sections 4.1–4.3).
- Performance evaluation of the best system on the test database (Section 4.4).
- Comparison to screening based only on age and BMI factors (Section 4.5).

### 4.1. Analysis of individual features

Our first study is focused on the analysis of the individual features. This analysis has been performed by means of two different techniques: statistical and discriminant analysis.

#### 4.1.1. Statistical analysis of individual features using Mann–Whitney U test

First, using the training database, we want to determine whether the distributions of each selected feature present statistically significant differences between Apnoea and Control groups (and henceforth could potentially be used in the OSA screening). As our features are not normally distributed, we used the Mann–Whitney *U* test (UMW), as it is commonly used as an alternative non-parametric to a standard *t*-test (Mann and Whitney, 1947). This test is based on the median of the difference on ranked data and it shows us that some of the individual features present statistically significant differences at the 95% confidence value between both groups.

For each particular feature and group, Table 3 shows its median (as this is the statistical output of the UMW test), its standard deviation and the correspondent *p*-value provided by the test. The features with the lowest *p*-values are considered the most promising features in order to be used for discrimination. According to the results of UMW test the best features for discrimination are the following: *SDRseg, F3-F2_i*, *HNR*, *PercSil4, Duration4, a1h1max800_Diff Jitter and Duration1* while the features with the higher p-values: *Shimmer*, *a1h1max800_NonNasal*, *PercSil1-3*, *Duration2-3*, *a1h1max800_Nasal* would be considered the less promising features for which we do not find statistically significant differences.

Other interesting conclusions can be extracted from these results. It is remarkable that sentence duration and percentage of silence can be very discriminative features but only if the sentence pronounced is designed especially for this purpose (as fourth sentence of our speech corpus was). Otherwise, these suprasegmental features are not very discriminative. Regarding the nasal features, we defined these features based on the findings of Fox and Monoson (1989), in which most experts perceived some abnormalities in nasalization but they were not capable of describing them consistently (neither as hyponasality nor as hypernasality). Our experiments show that differences in nasalization in nasal (*a1h1max800_Nasal*) and non-nasal (*a1h1max800_Non-nasal*) contexts are not statistically significant at the 95% confidence value, which could explain why experts found difficult to find what the abnormality in nasalization was. We have found, however, that the difference in nasalization between nasal and non-nasal contexts (*a1h1max800_Diff*) is statistically significant at the 95% confidence value. Looking at the medians of the nasality measures it seems that apnoea patients have a smaller dynamic range of nasalization (i.e. they tend to pronounce with a more similar nasality

Table 3
Median, standard deviation and *p*-values for the proposed speech features computed on the training dataset. *p*-Values were obtained comparing apnoea/control median values using the MWU test. Results are sorted in decreasing order of *p*-value.

| Feature # | Feature name | Median (Std) Apnoea group | Median (Std) Control group | *p*-Value (95% confidence) |
|---|---|---|---|---|
| 4 | *SDRseg* | 30.9 (4) | 34 (3) | **<0.001*** |
| 1 | *F3-F2_i* | 614 (80) | 586.5 (66) | **<0.001*** |
| 2 | *HNR* | 10.5 (1.8) | 11.3 (1.6) | **0.003*** |
| 16 | *PercSil4* | 0.08 (0.08) | 0.06 (0.05) | **0.004*** |
| 12 | *Duration4* | 3.6 (0.9) | 3.2 (0.7) | **0.006*** |
| 8 | *a1h1max800_Diff* | −1.0 (1.8) | −2.5 (2.5) | **0.011*** |
| 3 | *Jitter* | 0.0177 (0.003) | 0.0170 (0.004) | **0.011*** |
| 9 | *Duration1* | 4.0 (0.8) | 3.9 (0.7) | **0.044*** |
| 11 | *Duration3* | 4.7 (1.2) | 4.6 (1.1) | 0.096 |
| 6 | *a1h1max800_Nasal* | 5.1 (2.2) | 3.9 (2.4) | 0.096 |
| 13 | *PercSil1* | 0.16 (0.07) | 0.15 (0.07) | 0.119 |
| 10 | *Duration2* | 5.0 (1.0) | 4.8 (1.1) | 0.205 |
| 15 | *PercSil3* | 0.20 (0.08) | 0.17 (0.1) | 0.264 |
| 14 | *PercSil2* | 0.18 (0.07) | 0.19 (0.09) | 0.315 |
| 7 | *a1h1max800_Non-nasal* | 6.2 (2.6) | 6.8 (3.0) | 0.328 |
| 5 | *Shimmer* | 0.11 (0.02) | 0.11 (0.02) | 0.458 |

* Statistically significant differences found at the 95% confidence level.

level in nasal and non-nasal contexts) than the healthy subjects that have more difference in nasalization between both contexts. This finding allows us to propose the new feature *a1h1max800_Diff* as a good candidate to be used in OSA detection, and is also consistent with both the hypothesis suggested by Fox and Monoson (1989), and the results of our previous work (Blanco et al., 2009), which shows us that OSA speakers exhibited smaller intra-speaker differences between non-nasal and nasal vowels.

#### 4.1.2. Discriminative analysis of individual features: EER

The analysis on the statistical differences of the distributions for each individual parameter presented before describes their underlying potential for OSA discrimination. In this section, their classification capability when used in isolation is measured through performance metrics relevant to their application for OSA diagnosis. For each feature, we have built a very simple classifier by means of setting a decision threshold in order to separate Apnoea from Control groups using the training database. Next we have computed the error rates obtained for both groups which correspond to the two types of errors that can be made in the OSA diagnosis: errors in the Apnoea group because apnoea subjects are incorrectly diagnosed as healthy, and errors in the Control group because healthy subjects are incorrectly detect as patients with OSA. The point at which those two error rate curves cross defines the equal error rate (EER), which is very interesting because it measures the discriminative power of the feature with a single number, hence facilitating comparison of features.

As we can see in Table 4, when analyzing EER per feature we see that the best individual feature is the *SDRseg* with an EER of 34%, followed by *a1h1max800_Diff* and *PercSil4* with EERs slightly below 38%. With lower discriminative power, *HNR* and *Duration4* have EERs a little below 40%, and the rest of the features provide EERs over 40%. *Shimmer*, the worst feature, still provides some discrimination since the EER is still below 48%. Although, similarities with the conclusions from the previous statistical analysis are evident, there are also notable differences. For instance the feature *F3-F2_i* seemed to be one of the two best features according to the statistical analysis, but it does not appear among the best features according to EER (the opposite applies to *a1h1max800_Diff* feature).

### 4.2. Analysis of features combination

In the previous section we have analyzed the behavior of the individual features proposed to discriminate voices of OSA patients from voices of healthy subjects. We have made this analysis from different points of view (i.e. statistical analysis and discriminative power) and have reached conclusions not totally equivalent (for instance regarding features *F3-F2_i* and *a1h1max800_Diff*). This makes it difficult to choose the best features to be used in our purpose of

Table 4
EER obtained with each one of the individual features on the training database. Features are sorted in increasing order of EER.

| Feature # | Feature name | EER (%) |
|---|---|---|
| 4 | *SDRseg* | 34.0 |
| 8 | *a1h1max800_Diff* | 37.6 |
| 16 | *PercSil4* | 37.8 |
| 2 | *HNR* | 39.7 |
| 12 | *Duration4* | 39.7 |
| 1 | *F3-F2_i* | 40.2 |
| 11 | *Duration3* | 40.3 |
| 10 | *Duration2* | 40.7 |
| 7 | *a1h1max800_Non-nasal* | 41.6 |
| 15 | *PercSil3* | 41.6 |
| 3 | *Jitter* | 41.7 |
| 9 | *Duration1* | 42.0 |
| 14 | *PercSil2* | 42.8 |
| 6 | *a1h1max800_Nasal* | 43.3 |
| 13 | *PercSil1* | 43.7 |
| 5 | *Shimmer* | 48.0 |

detecting OSA with speech signals. This selection is more difficult if we take into account that the features could be highly correlated between them, what could lead to a problem of multicollinearity that could worsen the detection model. In other words, once a feature is included in the set of best features adding another feature that individually seemed to provide a great deal of information does not necessarily add any new information.

For this reason, in this section we first analyze the correlation between different features, and then we present the outcomes of an incremental feature selection study for identifying the best feature subsets from 1 to 16 features.

### *4.2.1. Correlation analysis*

Aiming to study the possible linear dependencies between the selected features, we have computed a correlation matrix that is shown in Fig. 2. We can see that, in general, there is slight correlation between the features proposed: almost 60% of the correlation coefficients are below 0.2 in absolute value and almost 80% are below 0.5. There are only a few exceptions. Features 2 (HNR) and 4 (SDRseg) are highly correlated, which is reasonable because both measurements characterize disorders related to phonation anomalies. Features 6, 7 and 8 (*a1h1max800_Nasal*,

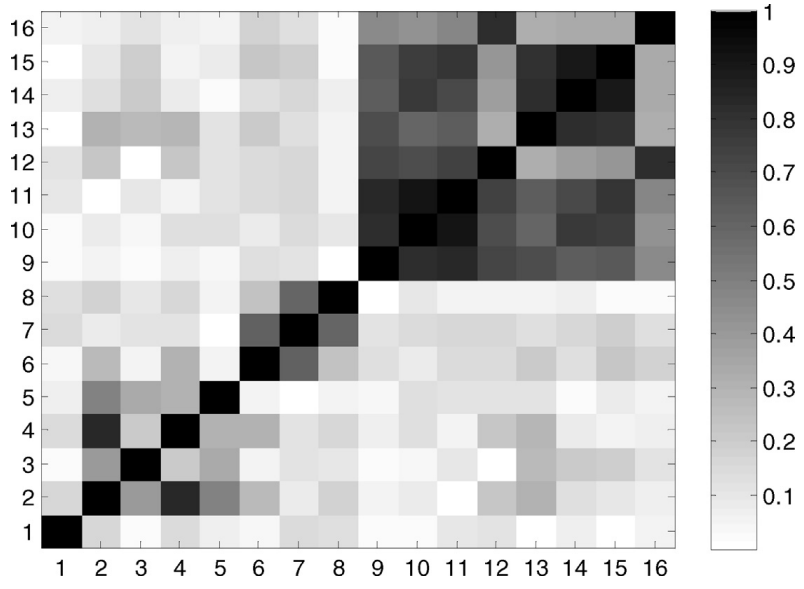


Fig. 2. Correlation Matrix (showing absolute values of correlation coefficients) computed on the training database for the 16 voice features proposed to characterize apnoea voices. Darker means stronger correlation.

*a1h1max800_Non-nasal* and *a1h1max800_Diff*) are also clearly correlated, what makes sense since *a1h1max800_Diff* is defined as the difference of the other two features. Finally, all features related to sentence durations and percentages of silence are highly correlated from sentence to sentence of the corpus. This fact is quite logical since a speaker that tends to read slower will do it on all sentences. The same happens with the silences. It is interesting to see that *Duration4* and *PercSil4* are less correlated with the equivalent measures on the other sentences. This may be due to the different nature of fourth sentence: one single melodic group that is asked to be pronounced without pauses.

#### 4.2.2. Analysis of incremental subsets of features

In order to determine the best possible combination of features, we have performed an incremental study in which we built models that start with one feature and are augmented with the next best feature remaining according to some feature comparison criteria.

We consider two different models for the combination of features: the *multiple linear regression* (MLR) model (Draper and Smith, 1998) and the *linear discriminant analysis* (LDA) model (Duda et al., 2001). These models need to be trained on a specific set and might probably produce overfitting to that set. To avoid overfitting, we have divided the training database (42 apnea and 40 healthy subjects) into two equal-size subsets containing the same number of apnoea (21) and healthy (20) subjects. One of these subsets will be used for training the model and selecting the best next feature to add (*train_train* subset) and the other subset only to evaluate each model selected on data not seen during training (*train_dev* subset). In fact, we evaluate the models on both subsets, so that we can study in more detail the problem of overfitting, but conclusions are drawn on results obtained on the *train_dev* subset since they are more representative of the situation of a real system that has to classify patterns not seen during training. To evaluate the models we used two different criteria:

- *Classification error rate* (CER) attained by the model. It is the ratio between the number of incorrectly classified subjects and the total number of subjects. It would be desirable to have as lower CER as possible (its minimum is 0% by definition). Measuring CER implies setting a threshold to take decisions. For LDA classification, this threshold has been adjusted based on the *train_train* subset, while for the MLR model it has been set to 0.5: the average between the 0 output value (i.e. dependent variable value) assigned to healthy subjects and the 1 output value assigned to apnoea subjects.
- $R^2$ *statistic* (Vittinghoff et al., 2004). The $R^2$ statistic is commonly used as a figure of merit of MLR models and can be interpreted as the amount of variance that the model explains. It is also the square of the Pearson's correlation coefficient between the values predicted by the model and the actual values. It would be desirable to have as higher $R^2$ as possible (its maximum is 1 by definition).

After establishing the classification models and performance criteria, the incremental study starts by considering all the individual features in isolation and choosing the best individual feature computed on the *train_train* subset according to the evaluation criteria. Although it would be natural to choose $R^2$ for the MLR model ($R^2$ is closely related to the regression residual, the parameter to optimize in MLR) and CER for the LDA model (CER is the parameter that LDA optimizes), we will take only $R^2$ as the evaluation criteria. The reason for this decision is that, after having split the training corpus into two subsets, the amount of subjects in each subset is only 41. This makes CER a very granular measurement because it only changes when at least one decision change. In our experiments this happened only a few times, and therefore the criterion had not enough resolution to choose among different features (hence the selection was many times random). In contrast, $R^2$ takes into account the predicted value and the actual value, and therefore is computed before a decision is taken, which allows finding differences in $R^2$ even when the decisions are exactly the same. Once the best individual feature is chosen according to $R^2$ computed on the *train_train* subset, all the remaining features are tested in combination with this best feature, and the next best feature is chosen, again according to $R^2$ of the predictions made by the model combining the two features on the *train_train* subset. By repeating this procedure we continue adding the best next feature until all features are contained in the models.

The outcomes of this procedure are a list of the features sorted in decreasing order of importance for each of the models, the 'best' models considering 1–16 features, and also the evaluation metrics ($R^2$ and CER) obtained by each of the $16 \times 2$ different models (LDA and MLR) on the *train_train* and *train_dev* subsets. All this information, and in particular results on the *train_dev* subset, will allow us to determine which features really provide important, discriminative and subset-independent (with good generalization capability) information, not previously provided

Table 5
Incremental feature combination using a MLR model and $R^2$ selection criterion. Best outcomes for each column are highlighted.

| Feature # | Feature name | *train_train* subset | | *train_dev* subset | |
|---|---|---|---|---|---|
| | | $R^2$ | CER | $R^2$ | CER |
| 4 | *SDRseg* | 0.12 | 39.0% | 0.17 | 26.8% |
| +8 | *+a1h1max800_Diff* | 0.23 | 26.8% | 0.20 | 36.6% |
| +2 | *+HNR* | 0.31 | 24.4% | 0.22 | 29.3% |
| +3 | *+Jitter* | 0.43 | 22.0% | 0.32 | 17.1% |
| +1 | *+F3-F2_i* | 0.45 | 19.5% | 0.37 | 22.0% |
| +12 | *+Duration4* | 0.47 | 19.5% | 0.45 | 14.6% |
| +13 | *+PercSil1* | 0.50 | 24.4% | 0.45 | 17.1% |
| +9 | *+Duration1* | 0.51 | 17.1% | **0.46** | 17.1% |
| +14 | *+PercSil2* | 0.53 | 19.5% | 0.43 | **12.2%** |
| +10 | *+Duration2* | 0.57 | 14.6% | 0.39 | 19.5% |
| +16 | *+PercSil4* | 0.59 | 9.8% | 0.27 | 26.8% |
| +15 | *+PercSil3* | 0.59 | 9.8% | 0.23 | 26.8% |
| +11 | *+Duration3* | 0.60 | 9.8% | 0.25 | 24.4% |
| +5 | *+Shimmer* | 0.60 | 9.8% | 0.26 | 24.4% |
| +6 | *+a1h1max800_Nasal* | 0.60 | **7.3%** | 0.26 | 24.4% |
| +7 | *+a1h1max800_Non-nasal* | **0.61** | 9.8% | 0.24 | 26.8% |

by other (more important) features, and which ones provide little discriminative information or poor generalization capabilities. Therefore this experiment will allow us to rank the analyzed features and define the subsets of the most important ones for OSA assessment.

Tables 5 and 6 show the incremental feature selection outcomes for the MLR and LDA model respectively. It can be seen that the order in which features are added is the same, irrespective of the model employed in the experiment. This consistency in the results of feature selection using two different models makes these results more reliable. Comparing the columns corresponding to results evaluated in the *train_train* subset and the *train_dev* subset, it can be observed that outcomes obtained on the *train_train* subset (in terms of $R^2$ and CER and for both MLR and LDA models) tend to improve as more and more features are included in the model. However, results for the *train_dev* subset improve clearly only with the first 6–9 features and from that point on results are clearly degraded. This is indicating an overfitting situation when more than 6–9 features are added to the models. This could be due to the nature of the features (the

Table 6
Incremental feature combination using a LDA model and $R^2$ selection criterion. Best outcomes for each column are highlighted.

| Feature # | Feature name | *train_train* subset | | *train_dev* subset | |
|---|---|---|---|---|---|
| | | $R^2$ | CER | $R^2$ | CER |
| 4 | *SDRseg* | 0.12 | 36.6% | 0.17 | 29.3% |
| +8 | *+a1h1max800_Diff* | 0.23 | 31.7% | 0.20 | 36.6% |
| +2 | *+HNR* | 0.31 | 26.8% | 0.21 | 31.7% |
| +3 | *+Jitter* | 0.43 | 22.0% | 0.32 | 22.0% |
| +1 | *+F3-F2_i* | 0.46 | 17.1% | 0.37 | 24.4% |
| +12 | *+Duration4* | 0.48 | 22.0% | 0.44 | **14.6%** |
| +13 | *+PercSil1* | 0.50 | 22.0% | 0.45 | **14.6%** |
| +9 | *+Duration1* | 0.51 | 17.1% | **0.46** | **14.6%** |
| +14 | *+PercSil2* | 0.53 | 22.0% | 0.43 | 22.0% |
| +10 | *+Duration2* | 0.57 | 17.1% | 0.39 | 24.4% |
| +16 | *+PercSil4* | 0.59 | **12.2%** | 0.27 | 29.3% |
| +15 | *+PercSil3* | 0.59 | **12.2%** | 0.24 | 29.3% |
| +11 | *+Duration3* | 0.60 | **12.2%** | 0.25 | 24.4% |
| +5 | *+Shimmer* | 0.60 | **12.2%** | 0.26 | 26.8% |
| +6 | *+a1h1max800_Nasal* | 0.60 | **12.2%** | 0.26 | 26.8% |
| +7 | *+a1h1max800_Non-nasal* | **0.61** | **12.2%** | 0.24 | 26.8% |

Table 7
CER on the training dataset using a leave-one-out procedure for LDA and MLR models using all features and the 6–9-best features found in the incremental feature analysis. Best outcomes are highlighted for LDA and MLR models.

| | LDA model | MLR model |
|---|---|---|
| 6-best features | 18.3% | 19.5% |
| 7-best features | 20.7% | 22.0% |
| 8-best features | **17.1%** | **18.3%** |
| 9-best features | 18.3% | **18.3%** |
| All features | 23.2% | 22.0% |

features themselves do not provide additional valuable information) or to the limitations of the size of the corpus (we have 41 subjects in the *train_dev* subset and the models require estimating one parameter for each of the features considered, plus a bias term, so having 7 features requires estimating 8 parameters with 41 subjects). In the future it would be very interesting to expand these experiments to determine the reason for this overfitting.

In any case, what we can conclude with certainty is that, with this amount of material and with these features, the best set of features is the one containing the first 6–9 features, which include, in order: *SDRseg*, *a1h1max800_Diff*, *HNR*, *Jitter*, *F3-F2_i*, *Duration4*, *PercSil1*, *Duration1* and *PercSil2*. This set of features includes: classical measurements used in the characterization of pathological voices (such as *SDRseg*, *HNR* and *Jitter*); features already proposed by the authors for the specific problem of apnoea detection (Fernández et al., 2009), *F3-F2_i;* and other features not used previously in the characterization of apnoea voices such as *Duration1, Duration4, PercSil1, PercSil2* and *a1h1max800_Diff*, being this last measure based on previous perceptual findings from other authors.

### 4.3. Classification and diagnosis performance

As a result of the incremental feature combination analysis in the previous section, we have obtained two classifiers (MLR and LDA) combining all the features or a subset of best 6–9 features. In this section, we firstly analyze the performance when using the two classification models (MLR and LDA) with all the features or the best 6–9 features on the training database in terms of classification performance, using CER.

Then, in a second and probably most important analysis, diagnosis performance, we evaluate figures of merit traditionally employed in medical diagnosis such as *sensitivity*, *specificity*, *positive predictive value*, *negative predictive value* and *F1-measure*.

Given that the division of the training database into two equal-size subsets limits very much the amount of training examples (only 41 subjects), which could possibly jeopardize the accuracy achieved, we have performed these experiments using a leave-one-out procedure, that is repeatedly using all the subjects in the training database except one to train the models and perform a test on the excluded subject.

It should also be noted that, as the training database was used in the feature selection process, results in this section can be considered as a resubstitution upper bound that will be later compared with those obtained on a separate test database (Sections 4.4 and 4.5). Another limitation for the training database is the significant differences in age and BMI between OSA and Control groups, this issue will be also discussed in section 4.5.

#### 4.3.1. Classification performance

In practice, if our system is going to be used in diagnosis support it should make a decision and classify the voice as corresponding to an apnoea patient or a healthy subject. For this reason, we have performed this decision based on a threshold trained in the case of LDA (the threshold is trained on the 81 remaining subjects of corpus in the leave-one out testing protocol) and a fixed threshold in the case of MLR (0.5, corresponding to the middle point between the 0 for healthy subjects and 1 for apnoea patients).

Classification error rates (CER) are presented in Table 7. Results show that by combining all features with the LDA model and performing a leave-one-out experiment on the training database we attain a reasonably good classification performance, with a CER of 23.2%. This result is slightly better with the MLR model (22.0%). If we limit our models to the set of 6–9 best features found in Section 4.2.2, models attain better classification performance: LDA reaches a

Table 8
Sensitivity, Specificity, Positive and Negative Predictive Value and F1-measure of our voice-based approach. Results obtained on the training database using a leave-one-out approach with the LDA model and the 8 best selected features.

| Sensitivity | Specificity | Positive Predictive Value | Negative Predictive Value | F1 measure |
|---|---|---|---|---|
| 78.6% (33/42) | 87.5% (35/40) | 86.8% (33/38) | 79.6% (35/44) | 82.5% |

CER of 17.1%, while MLR achieves a CER of 18.3%. These best results are achieved with the 8 best features (8 or 9 in the case of the MLR model). This gives an idea of the importance of feature selection in this context.

In general we can conclude that using all features produces worse results than using only the best 6–9 features and that the LDA model produces slightly better results than the MLR model in this case. Best performance is achieved using 8 features.

#### 4.3.2. Diagnosis performance

Previous results have been presented in the typical form for classification and verification problems. However, the goal of the selected features and proposed methods is to help in the OSA screening. Therefore in this section we evaluate the best of the proposed methods (the combination of the 8-best features with the LDA model) using the criteria commonly used in clinical diagnosis (Knottnerus and Buttinx, 2008):

- *Sensitivity*: ratio of correctly classified OSA-suffering speakers (true positives) to total number of speakers actually diagnosed with OSA.
- *Specificity*: ratio of correctly classified healthy subjects (true negatives) to total number of speakers diagnosed as not suffering from OSA.
- *Positive Predictive Value* (PPV): ratio of true positives to total number of patients classified (correctly or incorrectly) as having an OSA voice.
- *Negative Predictive Value* (NPV): ratio of true negatives to total number of patients classified (correctly or incorrectly) as not having an OSA voice.
- *F1-measure*: a single measure of performance obtained as the harmonic mean of positive predictive value (precision) and sensitivity (recall). F1-measure = $2 \times (\text{PPV} \times \text{sensitivity})/(\text{PPV} + \text{sensitivity})$.

Results in Table 8 are very encouraging and seem to indicate that our approach could be useful for performing fast and convenient screening for OSA. Comparing these results with our previous work (Fernández et al., 2009) where we reported a sensitivity value of 77% and specificity value of 85%, the obtained values here are slightly better using a completely different approach, which leads us to believe we could achieve better results by combining the two different approaches.

### 4.4. Analysis of diagnosis performance on a separate test set

Up to this point we have used the training database to analyze and select features and to train and test models. We will now take the best of the proposed methods (8 best features selected and LDA model) and we will train a model on the whole training database in order to evaluate it on a different test dataset (test database) that has not been used so far. Diagnosis performance results obtained on this independent test set are shown in Table 9.

Results obtained on the test database are only slightly worse in terms of performance than those obtained on the training database (Table 8). Features were selected using results of the training database; therefore, we could expect some degradation when moving to a different database. It is also noteworthy to highlight that the trade-off between sensitivity and specificity has changed considerably between Tables 8 and 9. This means that the calibration (or threshold setting) of the method proposed for apnoea screening would probably require further research.

Table 9
Sensitivity, specificity, positive and negative predictive value and F1-measure of our voice-based approach. Results obtained on the test database with the LDA model and the 8 best selected features.

| Sensitivity | Specificity | Positive predictive value | Negative predictive value | F1 measure |
|---|---|---|---|---|
| 85.0% (17/20) | 75.0% (15/20) | 77.3% (17/22) | 83.3% (15/18) | 81.0% |

Table 10
Sensitivity, specificity, positive and negative predictive value and F1-measure on the training and test database with our voice-based method or only Age and BMI.

| Database | Method | Sensitivity | Specificity | Positive predictive value | Negative predictive value | F1-measure |
|---|---|---|---|---|---|---|
| TRAIN | Voice-based | 78.6% | 87.5% | 86.8% | 79.6% | 82.5% |
| | BMI + age | 78.9% | 77.5% | 76.9% | 79.5% | 77.9% |
| TEST | Voice-based | 85.0% | 75.0% | 77.3% | 83.3% | 81.0% |
| | BMI + age | 65.0% | 55.0% | 59.1% | 61.1% | 61.9% |

### 4.5. Comparison to screening using only BMI and age

Our voice-based screening method seems to be comparable in accuracy to other OSA screening strategies that employ clinical prediction models using anthropometric characteristics (like BMI and age) and epidemiological parameters (Gurubhagavatula et al., 2004; Friedman et al., 2010). However, although these approaches report similar sensitivity values (around 80%), they generally present appreciably worse specificity values (around 70%).

To study our particular case, we compared the results in Tables 8 and 9 (using the selected LDA model and 8 features) with those obtained with a model that does not contain information from speech but only uses age and BMI data. On the training database accuracy, results shown in Table 10 using only the variables mentioned, are lower than when using our voice-based approach. The only exceptions are the results for the Sensitivity and Negative Predictive Value, which are similar. In terms of the F1-measure the difference between our proposed approach and BMI and age alone is 4.6% in absolute terms.

On the test database, both methods suffer some degradation. However our voice-based features method only suffers an absolute reduction of 1.5% in terms of F1-measure while the absolute reduction reaches a 16% for the method based on age and BMI. The reason for the important degradation in performance when using only age and BMI is that the test database was carefully designed to avoid statistically significant differences in terms of both factors between the two groups. Thus this low performance on this database was expected. The main conclusion from Table 10 is that the method based on voice features continues working almost as well as on the training dataset even when there are not statistically significant differences between the two groups of speakers in terms of age and BMI. This clearly indicates that our method provide information related to apnoea not only through their relation to age and BMI.

We have also delved into the study of the correlations between the 8 selected voice features and the age and BMI variables. Table 11 shows the correlation coefficients and their associated $p$-values for both training and test databases. The first noticeable difference between the two databases is that in the training database there is a number of statistically significant correlations, while in the testing database no correlation is close to be statistically significant at the 95% confidence value. The main difference between the two databases is that in the training database there are statistically significant differences in terms of BMI and age between the Apnoea and Control group, while the test database was

Table 11
Correlation coefficient and p-values between speech features and Age and BMI. Statistically significant correlations are highlighted.

| | TRAIN database | | | | TEST database | | | |
|---|---|---|---|---|---|---|---|---|
| | Correlation coefficient | | $p$-Value | | Correlation coefficient | | $p$-Value | |
| FEATURE | AGE | BMI | AGE | BMI | AGE | BMI | AGE | BMI |
| *SDRseg* | −0.21 | **−0.26** | 0.06 | **0.02*** | −0.08 | −0.23 | 0.60 | 0.14 |
| *a1h1max800_Diff* | −0.02 | 0.04 | 0.87 | 0.72 | −0.11 | −0.06 | 0.47 | 0.70 |
| *HNR* | −0.16 | −0.02 | 0.16 | 0.89 | −0.09 | −0.02 | 0.57 | 0.90 |
| *Jitter* | 0.16 | 0.15 | 0.16 | 0.20 | −0.11 | 0.10 | 0.48 | 0.51 |
| *F3-F2_i* | 0.22 | **0.28** | 0.06 | **0.01*** | −0.05 | 0.11 | 0.72 | 0.48 |
| *Duration4* | **0.35** | **0.28** | **0.002*** | **0.01*** | 0.13 | −0.1 | 0.40 | 0.54 |
| *PercSil1* | **0.23** | 0.07 | **0.04*** | 0.56 | 0.04 | −0.06 | 0.77 | 0.71 |
| *Duration1* | **0.34** | 0.10 | **0.003*** | 0.37 | 0.076 | −0.14 | 0.64 | 0.37 |

* Statistically significant differences found at the 95% confidence level.

designed to make those differences not statistically significant. The absence of statistically significant correlations in this second test seems to indicate that the correlation found for some speech features in the training database could be due to their correlation with the apnoea condition and the correlation of this condition with BMI and age.

## 5. Conclusions and discussion

This article has studied a set of 16 speech features that could be related to obstructive sleep apnoea. Some of these features are traditionally used in the detection of voice pathologies (such as HNR, Jitter, etc.), while others have been taken from previous research on voice characteristics of OSA patients or have been proposed by the authors from the analysis of the speech apnoea databases collected for this work. From our results on discriminative power of individual OSA-related voice features, we can highlight that we have found among the most discriminative features some already proposed by authors in previous works (i.e. the difference between the third and second formants for the/i/phoneme) and some first time implemented in this article, although already hypothesized in related works (i.e. the difference in nasality measured between vowels in nasal and non-nasal contexts).

We have also assessed OSA detection based on a selected subset of 8-best speech features and a LDA model. Classification results reached an overall performance of over 80%, similar to other OSA screening strategies based on clinical data such as BMI and age (Gurubhagavatula et al., 2004). Given that speech features could be related with BMI and age, which are known to be related to apnoea, classification results using the selected speech features have been compared to those obtained using only BMI and age. Results using the voice-based approach on a test database with no statistical differences in BMI and age distributions, and showing no significant correlations between speech features and both factors, have provided a F1-measure value 20% higher in absolute terms compared to using only BMI and age factors. This result can lead us to conclude that the proposed voice measures have the potential to improve OSA detection on segments of population where age and BMI have no discriminative power.

At this point it must be noted that, in addition to BMI and age, there are many other clinical data and symptoms in OSA patients, such as snoring, daytime sleepiness, smoking history, etc. that can influence their voice characteristics. The study of these dependencies is beyond the scope of this work, as we actually do not have complete clinical data for the patients in the databases, but it will be a line for future research. Nevertheless, in some cases, these studies could be faced through the articulation of complementary research lines under the speech analysis field; for example it will be worth considering how our speech features are correlated to the phonetic features already related to sleepiness in the recent research presented by Krajewski et al. (2012).

Despite these limitations, we consider that results in this work are encouraging and clearly show that the distinctive apnoea traits identified by our incremental feature selection method can be combined to provide reasonable diagnosis performance for screening purposes. Furthermore, such promising outcomes were obtained with just 8 speech features and with very simple and basic classification models. In previous works (Fernández et al., 2009) we reported results slightly worse than those presented here, but following a completely different approach. We expect to achieve even better results by combining these models with those previously proposed models.

Finally, further experimentation with a larger population of control and OSA subjects (we are involved in a continuous process of recording speech from subjects attending the Respiratory Sleep Disorders Unit) would be also desirable both to deepen our understanding on the analyzed and new speech features.

## Acknowledgments

The activities described in this article were funded by the Spanish Ministry of Economy and Competitivity as part of the TEC2012-37585-C02 (CMC-$V^2$) project. We also thanks to José D. Alcázar-Ramírez, MD, for his valuable support for collecting the speech databases used in this study.

## References

Ahmadi, N., Chung, S.A., Gibbs, A., Shapiro, C.M., 2008. The Berlin questionnaire for sleep apnea in a sleep clinic population: relationship to polysomnographic measurement of respiratory disturbance. Sleep Breath 12, 39L 45.

Bettens, F., Grenez, F., Schoentgen, J., 2005. Estimation of vocal dysperiodicities in disordered connected speech by means of distant-sample bidirectional linear predictive analysis. J. Acoust. Soc. Am. 117, 328–337.

Blanco, J.L., Fernández, R., Pardo, D., Siguenza, A., Hernández, L.A., Alcázar, J., 2009. Analyzing GMMs to characterize resonance anomalies in speakers suffering from apnoea. In: 10th Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 1459–1462.

Blanco, J.L., Fernández, R., Torre, D., Caminero F.J., López, E., 2011. Analyzing training dependencies and posterior fusion in discriminative classification of apnea patients based on sustained and connected speech. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3033–3036.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the Institute of Phonetic Sciences, vol. 17, pp. 97–110.

Boersma, P., Weenink, D., 2006. Praat: Doing Phonetics By Computer (version 4.5.01), Available in http://www.praat.org. (accessed 2.4.2013).

Boyd, J.H., Petrof, B.J., Hamid, Q., Fraser, R., Kimoff, R.J., 2004. Upper airway muscle vinflammation and denervation changes in obstructive sleep apnea. Am. J. Respir. Crit. Care Med. 170, 541–546.

Davidson, T.M., 2003. The Great Leap Forward: the anatomic evolution of obstructive sleep apnea. Sleep Medicine 4, 185–194.

Draper, N.R., Smith, H., 1998. Applied Regression Analysis, 3rd ed. Wiley-Interscience, Hoboken, NJ.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, 2nd ed. Wiley-Interscience, NY.

Fernández, R., Hernández, L.A., López, E., Alcázar, J., Portillo, G., Toledano, D.T., 2008. Design of a multimodal database for research on automatic detection of severe apnoea cases. In: Proceedings of 6th Language Resources and Evaluation Conference. LREC, Marrakech.

Fernández, R., Blanco, J.L., Hernández, L., López Gonzalo, E., Alcázar, J., Toledano, D., 2009. Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. EURASIP Journal on Advances in Signal Processing 2009 (Article ID 982531).

Fiz, J.A., Morera, J., Abad, J., Belsulnces, A., Haro, M., Fiz, J.I., Jane, R., Caminal, P., Rodenstein, D., 1993. Acoustic analysis of vowel emission in obstructive sleep apnea. Chest J. 104, 1093–1096.

Fox, A.W., Monoson, P.K., 1989. Speech dysfunction of obstructive sleep apnea. A discriminant analysis of its descriptors. Chest J. 96 (3), 589–595.

Friedman, M., Wilson, M.N., Pulver, T., Pandya, H., Joseph, N.J., Lin, H.C., Chang, H.W., 2010. Screening for obstructive sleep apnea/hypopnea syndrome: subjective and objective factors. Otolaryngol. Head Neck Surg. 142, 531–535.

Gurubhagavatula, I., Maislin, G., Nkwuo, J.E., Pack, A.I., 2004. Occupational screening for obstructive sleep apnea in commercial drivers. Am. J. Respir. Crit. Care Med. Vol170, 371–376.

Hidalgo, A., Quilis, M., 2002. Fonética y fonología españolas. Editorial Tirant blanch.

Knottnerus, J.A., Buttinx, F., 2008. The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research. Wiley, London, England.

Krajewski, J., Schnieder, S., Sommer, D., Batliner, A., Schuller, B., 2012. Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. Neurocomputing 84, 65–75.

Maltais, F., Carrier, G., Cormier, Y., Sériès, F., 1991. Cephalometric measurements in snorers, non-snorers, and patients with sleep apnoea. Thorax 46 (6), 419–423.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18, 50–60.

Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.B., Naude, C., 1993. ALBAYZIN Speech Database: Design of the Phonetic Corpus. In: Proceedings of Eurospecch 93, vol. 1, Berlin, Germany, pp. 175–178, 21–23.

Morgenthaler, T.I., Kagramanov, V., Hanak, V., Decker, P.A., 2006. Complex sleep apnea syndrome: is it a unique clinical syndrome? Sleep 29 (9), 1203–1209.

Parsa, V., Jamieson, D.G., 2001. Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. J. Speech, Lang. Hear. Res. 44 (2), 327–339.

Payne, R.J., Kost, K.M., Frenkiel, S., Zeitouni, A.G., Sejean, G., Sweet, R.C., Naor, N., Hernandez, L., Kimoff, R.J., 2006. Laryngeal inflammation assessed using the reflux finding score in obstructive sleep apnea. Otolaryngol Head Neck Surg 134 (5), 836–842.

Pruthi, T., 2007. Analysis, vocal-tract modeling and automatic detection of vowel nasalization. Doctor Thesis at the University of Maryland.

Puertas, F.J., Pin, G., María, J.M., Durán, J., 2005. Documento de consenso Nacional sobre el síndrome de Apneas-hipopneas del sueño (SAHS). Grupo Español De Sueño (GES). Arch Bronconeumol 41 (4), 1–110.

Ramachandran, S.K., Josephs, L.A., 2009. A meta-analysis of clinical screening tests for obstructive sleep apnea. Anesthesiology 110 (4), 928–939.

Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., 2003. The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4, pp. p784–p787.

Robb, M., Yates, J., Morgan, E., 1997. Vocal tract resonance characteristics of adults with obstructive sleep apnea. Acta Otolaryngol. 117, 760–763.

Ryan, C.M., Bradley, T.D., 2005. Pathogenesis of obstructive sleep apnea. J. Appl Physiol. 99 (December (6)), 2440–2450.

Teculescu, D., 1998. Can snoring induce or worsen obstructive sleep apnea? Med. Hypotheses 50 (2), 125–129.

Toledano, D.T., Hernández-Gómez, L., Villarrubia-Grande, L., 2003. Automatic phonetic segmentation. IEEE Trans. Speech Audio Process. 11 (6), 617–625, ISSN 1063-6670.

Vittinghoff, E., Glidden, D.V., Shiboski, S.C., McCulloch, C.E., ISBN 0-387-20275-7 2004. Regression methods in biostatistics. Linear, logistic, survival, and repeated measures models. In: Statistics for Biology and Health., pp. 43–44.

Young, S., 2002. The HTK Book (for HTK Version 3.2). First published December 1995. Revised for HTK Version 3.2 December.

Zhang, Y., Jiang, J.J., 2008. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. J. Voice 22 (1), 1–9, 119, 147.