APNOEA VOICE CHARACTERIZATION THROUGH VOWEL SOUNDS ANALYSIS USING GENERATIVE GAUSSIAN MIXTURE MODELS

J.L. Blanco[1], R. Fernández[1], D. Díaz[1], L. A. Hernández[1], E. López[1], D. Torre-Toledano [2]
[1]Departmento de Señales, Sistemas y Radiocomunicaciones, Universidad Politécnica de Madrid,
Avda. Complutense 30, 28040 Madrid, Spain
[2] ATVS Biometric Recognition Group. Universidad Autónoma de Madrid, Spain
e-mail: jlblanco@gaps.ssr.upm.es

Abstract: **In this article we present a research scheme which aims to characterize apnoea speech through vowel analysis using GMMs. Working with a carefully designed speech database of healthy and OSA Spanish speakers, we use generative statistical modelling techniques on different vowel sounds in specific linguistic context in order to search discriminant apnoea voice characteristics. Three main experiments were developed focused on comparing continuous speech with sustained vowels and on analyzing how our modelling techniques discriminate between apnoea and healthy subjects into different vowel sounds. We also study abnormal nasalization in OSA patients by considering vowels in nasal and non-nasal constexts. Results on apnoea corpus have confirmed that there are indeed significant differences between apnoea and control group speakers in specific vowel sounds, and that statistical modeling techniques are able to describe the discriminative information.**
Keywords : Obstructive Sleep Apnoea (OSA), Gaussian Mixture Models (GMMs).

I. INTRODUCTION

Obstructive sleep apnoea (OSA) is a highly prevalent disease [1], affecting an estimated 2-4% of male population between the ages of 30 and 60 years. It is characterized by recurring episodes of sleep-related collapse of the upper airway at the level of the pharynx (AHI > 15, *Apnoea Hypopnoea Index*, which represents the number of apnoeas and hypoapnoeas per hour of sleep) and it is usually associated with loud snoring and increased daytime sleepiness.

Since upper airways are affected by OSA disease, it seems adequate to consider whether there are any particular patterns on speech signals which could be related with OSA. Evidences on this hypothesis have been provided in a few remarkable references. Though, most of the more valuable information in this area can be found in Fox and Monoson's work [2], a perceptual study with skilled judges comparing the voices of apnoea patients with those of a control group (referred to as "healthy" subjects). As a result they provide some evidences of OSA disease which were observed during the study they conducted, such as abnormal resonances (hyponasality and hypernasality), and both articulation (due to a probable velopharyngeal dysfunction) and phonation anomalies. Those anomalies become clearer when contrasting OSA speakers with those from the control group, and therefore discriminating power of those factors might relevant enough for an early diagnose of severe obstructive sleep apnoea.

Working out those particular traces within speech signals requires an special effort in order to design and collect a consistent database which can meet our requirements for speech data from both speakers suffering from OSA and healthy ones, collected in the same conditions and over a specifically designed speech corpus. The design, following some phonetic and linguistic criteria derived from the previous work of Fox and Monoson [2], as long as some data from the preliminary database are described in [3].

Other relevant references deepen into some particular aspects about acoustic analysis of OSA speakers. For instance, interesting readers will find in [4] an excellent work on vocal tract resonances of OSA adults from a physiological point of view. This matches perfectly with some conclusions in [2], where an inappropriate nasal resonance related to coupling and de-coupling both nasal and oral cavities had been identified. The work condensed in Fiz et al. [5] is also a good reference work, as they focus, as we do on both apnoea disease and vowel sounds. However, while they consider direct inspection on the spectral representation of the collected data, we will be applying generative statistical modelling techniques to describe the acoustic space, in a similar way to that being used in speech and speaker recognition systems. A generative statistical model can be trained to describe the acoustic space of a particular sound, speaker or voice, so further recognition or classification tasks can be based on the likelihood that a given unknown sound or utterance was generated by the trained model.

As we have already suggested, in this contribution we will focus on characterizing apnoea voice by means of vowel sounds. Differences between vowels' acoustic spaces for patients suffering from apnoea and a control group of healthy people will be the realm in which we will work, applying generative statistical modelling based on Gaussian Mixture Models (GMMs). Excellent classification rates have been achieved by modelling short-time speech spectrum information with cepstral coefficients and using statistical pattern classification techniques such as GMMs. Hence we will apply this same scheme into our problem, hoping they

will fit into our problem, reflecting the discriminative patterns within OSA speakers.

The remainder of this paper is organized as follows. In section II we present the methodology and experimental setup for our study. Later, in section III results for three different experiments are presented. First we compare the results to be expected when working with vowel's extracted from continuous speech and with sustained vowels. Second results are drawn on comparing sounds from different vowels. And finally we analyze vowels in different phonetic contexts and their connection between those and OSA disease.

II. METHOD

In this contribution we are focusing on analyzing vowel sounds in order to characterize OSA speakers. Vowel sounds have several desirable properties for applying speech processing techniques such as, pseudo-stationarity (quite useful when considering pitch estimation or other measures deriving from this), harmonic structure due to the excitation of the vocal folds, or greater energy, as an overall average, especially when compared with consonants.

All the required data was extracted from the previously mentioned database we have collected [3], as to our knowledge there wasn't any available resource which we could use for this specific task. As we pointed out in the introduction, the database has been designed to expect to cover relevant linguistic/phonetic contexts in which physiological OSA-related peculiarities could have a greater impact. This includes:

- Voiced sounds affected by certain preceding phonemes that have their primary locus of articulation near the back of the oral cavity, anatomical region has been seen to display physical anomalies in OSA speakers.
- Continuous voiced sounds to compute irregular phonation patterns related to muscular fatigue in apnoea patients.
- Vowels in different linguistic contexts to measure, for instance, how nasalization varied from nasal to non-nasal contexts

Every sentence in our speech database was processed using short-time analysis with a 20 ms time frame and a 10 ms delay between frames, which gives a 50% overlap. Each of the windows analyzed will later be presented in the form of a training vector for our statistical models (both HMMs and GMMs). For the task of acoustical space modelling we chose to use 39 standard components: 12 Mel Frecuency Cepstral Coefficients (MFCCs), plus energy, extended with their speed (delta) and acceleration (delta-delta) components, assuming an optimized representation may bring better results, although this will require a specific adaptation of the recognition techniques to be applied, which is out of our goals for this work. The vectors resulting from this front-end process are put together into training sets for statistical modelling. This grouping task can be carried out following a variety of criteria depending on the experiment we are interested in or the phonetic classes we need to model.

Since we needed to extract vowel sounds from phrases in the database and to consider specific acoustical features and phonetic contexts as we will see, we first performed a phonetic segmentation of every utterance in the database. This allows combining speech frames from different phonetic contexts for each sound in order to generate a global model, or classifying data by keeping them in separate training sets. For each sentence in the speech database, automatic phonetic segmentation was carried out using the open-source HTK tool [6]. A full set of 24 context-independent phonetic Hidden Markov Models (HMMs) was trained on a manually phonetically tagged subcorpus of Albayzin database [7]. As our speech apnoea database includes the transcription of all the utterances, forced segmentation was used to align a phonetic transcription using the 3-state context-independent HMMs; optional silences between words were allowed to model optional pauses in each sentence. Using automatic forced alignment avoids the need for costly annotation of the data set by hand. It also guarantees good quality segmentation, which is crucial if we are to distinguish phonemes and phonetic contexts.

After phonetic segmentation, statistical pattern recognition can be applied to classify, study or compare voices for specific speech segments. In our case we decided to train a universal background GMM model (UBM) from phonetically balanced utterances taken from the Albayzin database [7], and use MAP (*Maximum a Posteriori*) adaptation to derive the specific GMMs for the different classes to be trained [8]. This technique increases the robustness of the models especially when sparse speech material is available. Only the means were adapted, as is classically done in speaker verification.

For the experiments discussed below, both processes, generation of the UBM and MAP adaptation to train the apnoea and the control group GMM models, were developed with the BECARS open source tool [9]. For testing purposes, and in order to increase the number of tests and thus to improve the statistical relevance of our results, the standard leave-one-out testing protocol was used. This protocol consists in discarding one sample speaker from the experimental database to train the classifier with the remaining samples. Then the excluded sample is used as the test data. This scheme is repeated until a sufficient number of tests have been performed.

III. RESULTS

Three main experiments were developed in order to analyze vowel sounds in speakers suffering from obstructive sleep apnoea. Each of them is focused on a particular aspect, regarding the approach we have already described, based on statistical modelling on the cepstral domain, and the data we have at our disposal.

A. Sustained Vowels vs. Continuous Speech

Regarding the remarkable properties of vowel sounds, a first step to be taken is to decide whether to follow the conventional approach for pathological voices analysis, and focus on sustained vowel sounds, or to steer by the previous literature on apnoea, and try to analyze vowel sounds

extracted from continuous speech. Though we decided to apply statistical modelling onto acoustic space description, it seemed reasonable to compare results using collected data from both sustained vowels and continuous speech in order to make a good decision.

With the experimental setup previously described, we trained subjects with apnoea and a control group, on the one hand with /a/ sections of continuous speech, and on the other with isolated sustained /a/s. The results obtained are condensed in these DET plots.

We compared both methods and found that using the vowels extracted from the continuous speech indeed did work better than the sustained /a/s (we got better classification error rates even though less samples of voice were used in the analysis). The DET plots we obtained are the following.

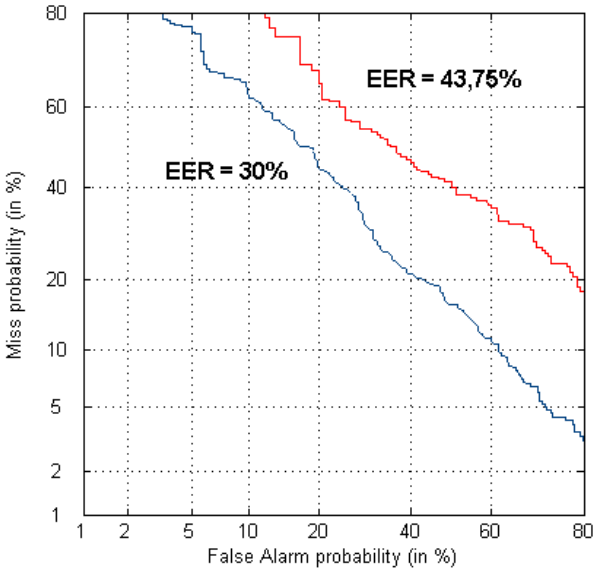

Fig 1. DET curves obtained for: /a/s in continuous speech extracted from the phrases (EER 43,75%), sustained /a/s (30%)

B. Comparing vowel sounds

Once we have chosen to continue our research focusing on vowel sounds extracted from continuous speech, it is worth to analyze how our modelling techniques apply into different sounds.

Following the same methodology, we extracted data for each vowel, regardless of their phonetic or linguistic context. Data was grouped in various training sets considering elementary classes (patients suffering from OSA and control group), and vowel sounds. Likelihood test was carried out in order to evaluate the potential for detecting apnoea, so same approach was followed, considering phonemes separately.

Finally, we have combined scores from the five previous classifiers, each corresponding to a Spanish vowel, in order to build a common one. This allows us to compare previous results from characteristic's fusion and the ones from the classifiers fusion.

We decided to plot the different values for the EER obtained as this value seems relevant enough to appreciate differences on the results for our modelling, and considering that further information contained on a DET plot will be related with other aspects of the procedure, not just the discrimination capacities of the training data.

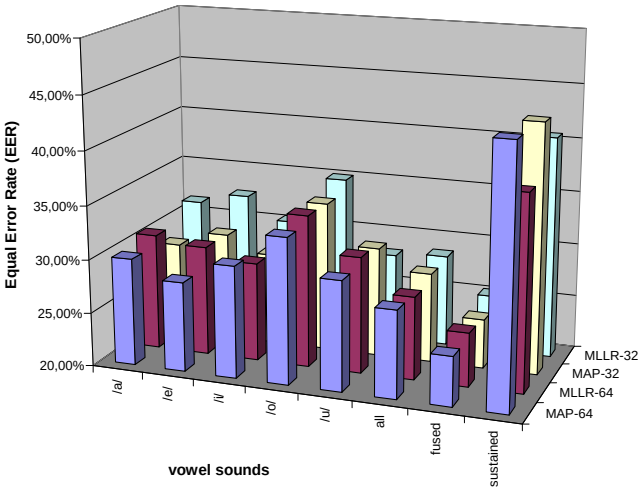

Fig 2. Equal Error Rate (EER) values obtained when: training separately vowel sounds taken from the phrases, training all together, uniformly combining previous vowels classifiers, when applied to sustained vowel /a/, with various adaptation algorithms (MAP and MLLR) and different parameters (64 and 32 gaussians in the mixtures).

C. Exploiting vowels' phonetic context

As it is discussed in [3] the provided database does account for various phonetic and linguistic contexts which seem to be relevant when analysing apnoea speakers, and therefore, useful to describe them. In this work we will focus on nasalization, which has been related with OSA as a natural phenomenon caused by a velopharyngeal dysfunction. Those abnormal resonances described in Fox and Monoson's work [2] could be perceived as a form of either hyponasality or hypernasality. The former is said to occur when no nasalization is produced when the sound should be nasal. Hypernasality is nasalization during the production of non-nasal (voiced oral) sounds. In other words, OSA sufferers will nasalize when they are not expected to, and vice versa. As a consequence we will expect statistical models (GMMs) trained with such training data to show smaller differences when comparing models for vowel sounds in nasal and non-nasal contexts. This may be tested in various different ways, for instance by considering the log-likelihood test previously used, but as what we are interested is in directly comparing models and measuring some sort of distance between them we decided to take a different approach.

Distances between statistical distributions are a controversial topic. In fact there is much to be discussed for any choice we could do between the various available distances. Actually we decided to take the fast approximation to *Kullback-Leibler (KL) divergence* for gaussian mixture models [10], as this distance is commonly used in automatic speaker recognition to define cohorts or groups of speakers producing similar sounds. This measure, though an upper bound to the *KL-divergence*, is fair enough to test our hypothesis of similarity between models for vowels in nasal and non-nasal contexts. The following plot (Fig. 3) shows the results we obtained when applying

statistical modelling to all vowels, and measuring distances between for both contexts.

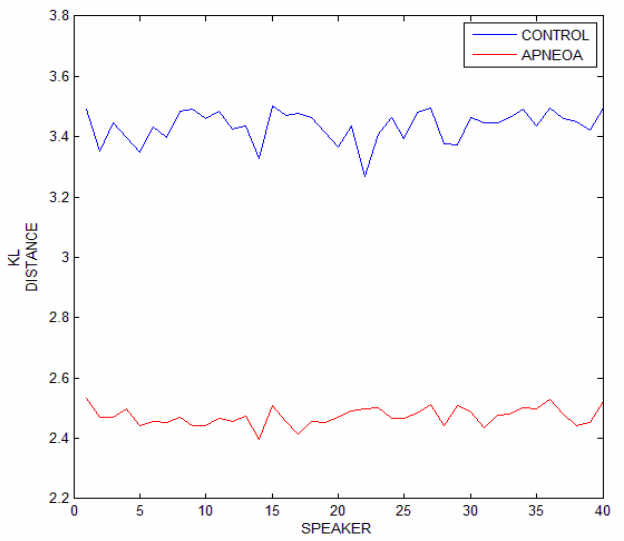

Fig 3. *Kullback-Leibler divergence between gaussian mixture models for vowels in nasal and non-nasal contexts*

IV. DISCUSSION & CONCLUSIONS

Going back through the presented results, it seems that our GMM approach discriminates both groups of speakers better when used with vowels in continuous speech that when sustained vowels are analysed. This could be an indication that there may be some kind of discriminating information present in continuous speech but absent in sustained vowels which our statistical modelling approach takes advantage of.

When we latter applied GMM modelling separately to each vowel by extracting the data from the phrases, we see EER values are in the same range, although with vowel /o/ seems to work worse than with the others. There is no clear reason to justify such result, but we guarantee that training has been done in the same way as with other vowels and that the amount of the data used for training was quite similar for all the experiments. On the other hand, fused information from all vowels sounds extracted has better results, as it is reasonable when combining different information into a single classifier.

Regarding the previous results on the fused characteristic's space, *posterior* uniform combination of independently trained classifiers has proved to have better results. This may be explained on the basis that characteristics fusion leads, in our case, to one GMM distribution which does not constraint each vowel to its own space. Meanwhile, separately trained classifiers do model each vowel's own acoustic region, and this prevents compound sounds, or outer sounds which may be assimilated to a combination of vowel sounds, from being wrongly classified.

Adaptation of the GMM distribution was carried out by considering not just MAP algorithm, but also MLLR (*Maximum Likelihood Linear Regression*). As it is shown in Fig. 2 results in both cases are rather similar, so we suspect the amount of data being used is just enough for the number of parameters/gaussians in the mixture, to be estimated.

Finally, though the distance measure we have used may require an extent discussion, we believe the fast approximation to the *Kullback-Leibler Divergence* has a good trade between computation rate and quality of the information it provides. Based on the results it seems to be a noticeable difference in the distances between models for nasal and non-nasal contexts, which fits perfectly with the conclusions addressed in [2].

Drawing all together, we are quite enthusiastic on our results as we have shed some light on the potential of vowel sounds analysis for apnoea characterization and have some promising results. We hope applying these findings would improve the performance of the automatic apnoea diagnosis system using speech processing algorithms on continuous speech, though still much work has to be done in describing acoustic space anomalies of patients suffering from OSA.

REFERENCES

[1] 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).

[2] Fox, A.W., & Monoson, P.K. (1989). "Speech dysfunction of obstructive sleep apnea. A discriminant analysis of its descriptors". In Chest Journal; 96(3): 589-595.

[3] Fernandez 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.

[4] Robb M., Yates J., and Morgan E. (1997) "Vocal Tract Resonance Characteristics of Adults with Obstructive Sleep Apnea" *Acta Otolaryngologica*, 117, 760--763.

[5] 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". In *Chest Journal*; 104: 1093 – 1096.

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

[7] 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. Berlin, Germany, 21-23. Vol. 1 pp. 175-178.

[8] Reynolds, D.A., Quatieri, T.F., & Dunn, R.B. (2000). "Speaker verification using adapted gaussian mixture models". In Digital Signal Processing 10: 19-41

[9] Blouet, R., Mokbel, C., Mokbel, H., Sanchez Soto, E., Chollet, G., & Greige, H. (2004). BECARS: a Free Software for Speaker Verification. In Proceedings of The Speaker and Language Recognition Workshop, ODYSSEY, pp 145-148.

[10] Do, M. N., (2003) "Fast approximation of Kullback-Leibler distance for dependence trees and Hidden Markov Models". IEEE Signal Processing Letter 10, 115-118.