Analyzing training dependencies and posterior fusion in discriminant classification of apnea patients based on sustained and connected speech

José Luis Blanco[1], *Rubén Fernández*[1], *Doroteo Torre*[2], *F. Javier Caminero*[3], *Eduardo López*[1]

[1] Signal Processing Applications Group (GAPS), Universidad Politécnica de Madrid, Spain
[2] Biometric Recognition Group (ATVS), Universidad Autónoma de Madrid, Spain
[3] Telefónica R&D, Spain

{jlblanco,ruben,eduardo}@gaps.ssr.upm.es, doroteo.torre@uam.es, fjcg@tid.es

Abstract

We present a novel approach using both sustained vowels and connected speech, to detect obstructive sleep apnea (OSA) cases within a homogeneous group of speakers. The proposed scheme is based on state-of-the-art GMM-based classifiers, and acknowledges specifically the way in which acoustic models are trained on standard databases, as well as the complexity of the resulting models and their adaptation to specific data. Our experimental database contains a suitable number of utterances and sustained speech from healthy (i.e control) and OSA Spanish speakers. Finally, a 25.1% relative reduction in classification error is achieved when fusing continuous and sustained speech classifiers.

Index Terms: obstructive sleep apnea (OSA), gaussian mixture models (GMMs), background model (BM), classifier fusion.

1. Introduction

Obstructive sleep apnea (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, *Apnea Hypopnea Index*, which represents the number of apneas and hypoapneas per hour of sleep) and it is usually associated with loud snoring and increased daytime sleepiness.

Since the upper airways are affected by OSA, it seems reasonable to consider whether there are any particular patterns in 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 in which skilled judges were asked to compare voices of apnea patients with those of a control group (referred to as ‘healthy’ subjects). As a result of their work several evidences of OSA disease were addressed, 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 rather than when no reference speech was provided, and therefore their discriminating power might be relevant enough to achieve an early diagnose of severe obstructive sleep apnea.

Working out a set of specific traces within speech signals which can be related to severe apnea cases requires an special effort to design and collect a consistent database which can meet our requirements, including records from speakers suffering from OSA and healthy ones, outlining a pair of reasonably homogeneous groups of speakers; those should be collected in the same acoustic conditions and over a specifically designed speech corpus. The design of the corpus, following some phonetic and linguistic criteria derived from the previous work of Fox and Monoson [2], as long as some details on the preliminary study developed to validate the designing and recording are all described in [3].

Despite the sparse literature focusing on the acoustic analysis of OSA speakers, there are a number of relevant references which have tackled this topic. For instance, interested readers can 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 the coupling and de-coupling of 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 also do, on apnea and vowel sounds. However, while they consider direct inspection on the spectral representation of the collected data, in the present contribution 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. Several peculiarities have to be taken into account when considering apnea detection through automatic speech processing. Unlike most research in pathological voices analysis, when looking for specific patterns in sleep apnea speakers’ utterances, there is no common agreement on whether continuous (connected) or sustained speech would be the best choice. In this work we will try to bring some light into this problem by considering three different sources of information: continuous speech (full phrases), voiced sounds extracted from continuous speech (namely vowels and semi-vowels) and sustained vowels (due to the limitations imposed by the available database we will restrict to sustained vowel /a/). Regarding the last two, some aspects have already been tackled on our previous work [7], as well as in [8]. However, these mainly focused on the differences among vowel sounds, the way in which those could be combined for OSA detection, and the peculiarities observed in vowel sounds within certain phonetic contexts.

Our baseline system for OSA detection relies on speaker recognition technology (as is extensively described in [6]) and can be roughly sketched as a GMM-based binary discriminant classifier (OSA vs. healthy –control– speakers) performing on top of a conventional MFCC parameterization of the entire acoustic space. Even though, several aspects should be taken into account in the training and adaptation procedures introduced to estimate accurate statistical GMM models. Those will be addressed in this work, as long as their influence on the final estimated models in terms of the deviation in the final classification accuracy from our baseline system.

Finally, in this contribution we will specifically evaluate the discriminative power and accuracy of a set of single classifiers based on the three sources of information we have enumerated: continuous speech, voiced sounds extracted from continuous speech and sustained vowels. All of them were built following the same criteria and methodology, though due to their intrinsic differences, as well as the particular characteristics of the OSA classification problem, different results were obtained for each of them. These intrinsic differences encouraged us to take a step further and consider the posterior fusion of the single classifiers in order to improve the accuracy of the current baseline system.

The remainder of this paper is organized as follows. In Section 2 we describe the methodological issues and experimental setup we will be using to train and evaluate the binary classifiers (OSA vs. Control) described in this contribution. Special attention is devoted to describe the influence of the training data and the adaptation techniques on the resulting classifiers. Later, in Section 3, the trained classifiers are presented, including the results from the estimation of their accuracy in detecting OSA. Additionally, in subsection 3.1 we consider the fusion of these classifiers in order to improve our baseline system by combining the information embedded in both connected and sustained speech. Finally, in Section 4, we summarize our conclusions by reviewing our OSA classification results and discussing the influence of the training data on these classification rates.

2. Method and experimental setup

In order to automatically detect severe apnea cases from the information embedded within speech signals, and out of a reasonably homogeneous group of speakers, we require a database which contains records from both Control (i.e. healthy) and OSA speakers. The required data was taken from the previously mentioned database for OSA which we have collected [3]. This database has been designed to cover relevant linguistic and phonetic contexts in which physiological OSA-related peculiarities could have a greater impact. These include:

1) Voiced sounds affected by certain preceding phonemes that have their primary locus of articulation near the back of the oral cavity, anatomical region that has been seen to display physical anomalies in OSA speakers.
2) Continuous voiced sounds to compute irregular phonation patterns related to muscular fatigue in apnea patients.
3) Vowels in different linguistic contexts to measure, for instance, how nasalization varied from nasal to non-nasal contexts
4) Sustained vowel /a/ instances for every speaker, four repetitions each.

There is no common agreement neither on the best choice for the features to be extracted from speech signals to achieve the best possible automatic classification system, nor on the sounds (vowels or not, sustained or connected speech) to be considered. Therefore, in this contribution we will be facing this later problem through the design of individual classifiers which can focus on certain continuous and sustained patterns, and the posterior fusion of those to improve the overall accuracy.

2.1. OSA acoustic modeling using GMMs

Gaussian Mixture Models (GMMs) are effective and efficient modelling techniques suitable for sparse speech data in Automatic Speaker Recognition systems [9]. In automatic OSA detection we will be using this same approach, as we will also be restricted to a binary classification problem. Additionally, there are several other similarities between the apnea detection problem and the speaker recognition one. Due to the fact that the amount of data in both scenarios happens to be insufficient to develop a Maximum Likelihood (ML) approach for the training, it appears to be quite common that statistical models have to be adapted from a universal background model (i.e. UBM), which is later adapted to derive a more specific one. The *Maximum A Posteriori* (MAP) adaptation algorithm is one of those techniques which is frequently used when the amount of data is big enough to guarantee the convergence of the algorithm to the desired model. On the other hand, if the available data is sparse, this technique will not diminish the accuracy of the adapted models. Finally, we should address that the BECARS open source tool [10] was used to train and adapt the statistical models for the OSA and control speakers' acoustic spaces.

But whatever the chosen adaptation technique is, the UBM has to be modelled, and the feature vectors describing the acoustic space have to be calculated.

2.2. Feature extraction

Every utterance in our database was processed using short-time analysis with a 20ms 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 (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.

We acknowledge that an optimized discriminative feature selection algorithm (e.g. LDA, PLDA, etc.) might become extremely useful to improve our description of the acoustic space, and subsequently enhance accuracy rates. However, as the acoustic characteristics of apnea speakers are still somewhat unclear, we have chosen to use a standard parameterization and look forward to a more specific representation for this particular field.

2.3. Training a suitable UBM

The trained UBM should be as specific but close to the final model as possible, in order to guarantee a quick and consistent convergence of the modelling. Thus, the UBM training process should be developed on a database general enough to represent the global acoustic space of a broad set of speakers. The influence of the prior distribution of the parameters (i.e. the trained UBM) on the adapted models is a well-known effect which could dramatically affect the estimation of the statistical models, and therefore compromise the accuracy of the overall classification scheme. Although there are several opportunities to design an statistical severe apnea detector based on the kind of information we have collected (connected and sustained speech), each of those requires a specific UBM to be trained. The influence of the initial priors suggests that this prior model should be as close to the final model as possible, as well as not to rely on the adaptation step (see subsection 2.4). Therefore, a different model should be trained for connected and sustained speech.

For the time being, most of our effort on apnea detection has been put into connected speech, which is actually our baseline system and requires a UBM built from continuous speech frames. Moreover the influence the GMMs model complexity (i.e number of gaussians) has on the final classification results was not considered, though it traditionally has a direct impact on the accuracy rates.

Besides, vowel sounds, and particularly sustained ones, are at the core of any recent pathological voice detection system. Meanwhile, for OSA classification our previous work on sustained vowels [7] undermined our perception of their discriminative power. Nevertheless, these results were obtained by adapting both OSA and Control GMMs to OSA and Control sustained speech in our evaluation database using an initial UBM trained with connected speech (the Albayzin database [15] was used: a reference speech database in Spanish for research on automatic speech recognition). Therefore the discriminative power of sustained sounds could be negatively affected by this starting point.

So far, in this contribution, additionally to the analysis on the influence of the GMM complexity, looking for a more accurate modelling of sustained sounds we have tested four different UBMs trained from four different datasets:

1. Three datasets coming from the phonetically balanced corpus in Albayzin:
   - Spanish vowel sounds in phonetically balanced sentences;
   - only sounds corresponding to all /a/ vowel instances;
   - instances of /a/ sounds appearing in minimum phonetic dependence contexts (i.e. surrounded by voiceless plosives or silences).
2. And a fourth dataset extracted from the Childers' [16], which includes sustained sounds from vowel /a/ in the Childers control group.

As it was pointed out before, the amount of available data to train each UBM, as well as the GMM complexity, will condition the results from the training, and therefore should be analyzed. Additionally, it has also to be noted that due to the fact that the utterances from the Childers database were recorded at a 10kHz sample rate, we were forced to resample them, up to 16kHz, in order to build a reliable UBM. The alternative (i.e. down sampling the OSA database records) was immediately rejected, as it will force us to neglect the information embedded in the high frequencies of the speech spectrum, which could be relevant [12] for classification purposes.

In summary, we can say that in our experimentation setup we have developed several UBM models considering the following criteria: parametrical models complexity (number of gaussians in the mixture), amount of data and its specificity.

2.4. Key aspects to UBM adaptation

As a binary discriminative classification problem, for each construction two models are derived from the prior UBM: the OSA-group and the control-group models. Therefore the second step in our modeling procedure shall perform a transformation of the models regarding the available data and certain adaptation criteria.

In this contribution we have chosen to use an iterative implementation of the MAP (*Maximum A Posteriori*) algorithm, which is widely used for compensating differences between sample characteristics when the amount of data is large enough. Though we acknowledge that our OSA database is short for a ML (*Maximum Likelihood*) training, it is still big enough for this adaptation technique, particularly for its iterative implementation, according to [11]. Moreover, the number of steps in the iterative process of the algorithm was left unbounded but for the basic posterior likelihood convergence criterion [9]. Though some authors have discussed the implications of this decision, we have observed that the number of iterations is quite reasonable (≤15).

Despite the previous considerations on the Childers dataset, the adaptation procedure following the UBM training is meant to be able to cope with this situation and introduce high-frequency information into the final GMMs [13].

Finally, it should be noted that a conventional *leave-1-out* cross-validation scheme will be performed to guarantee that the tests are fair and the results are significant enough.

3. Binary discriminant OSA classifiers

In Table 1 our five different test-cases are summarized. The first one is our baseline system, which was trained and tested on continuous speech. For the rest of them, GMMs were adapted and tested on sustained speech from our OSA database, but UBMs were trained from different datasets and acoustic units. Thus, more specific UBMs, as for example in test-case number 5, using Childers', can reduce the scope of the UBM modeling to an acoustic space which is closer to the final OSA and Control GMMs. Also relevant is the size of the available data for UBM training shown in the Table.

Table 1. *Details on the training, adaptation and testing of the designed GMM-based classifiers.*

	UBM database	**UBM dataset**	**Training Data** (min)	**Adp.&Test Dataset** (leave-1-out)
1	Albayzin	continuous speech	24,90	OSA continuous
2	Albayzin	extracted vowels	11,85	OSA sustained vowels
3	Albayzin	extracted /a/s	3,95	
4	Albayzin	particular extracted /a/s	1,92	
5	Childers	sustained vowels	2,82	

In order to analyze the influence of all these facts in the final system, we have chosen to consider the final Equal Error Rate (EER) as a figure of merit on the goodness of the whole training (train & adapt) scheme. Table 2 summarizes the results (EER) we obtained for the five test-cases for different number of gaussians. Due to the significant differences in the amount of training and adaptation data for test-cases 2 to 5 (adapted on sustained vowels), we were forced to limit the complexity of these models as beyond 16-component mixtures the training was extremely poor, and the adaptation step could not change this fact and make any improvements. In any case, this number of gaussians is reasonable enough to model the complexity of the acoustic space for vowel /a/, and therefore this limitation does not reduce the extent of our results.

Table 2. *Classification results (EER) for each configuration*

	Number of components in the mixture						
	256	128	64	32	16	8	4
1	23.1	25.0	25.1	25.0	27.7	29.2	30.8
2	-	-	-	-	45.6	46.9	46.9
3	-	-	-	-	46.3	45.6	46.3
4	-	-	-	-	43.1	47.5	45.6
5	-	-	-	-	39.4	41.3	46.3

3.1. Combining single classifiers

Trying to improve the results of our baseline system (test-case 1 in Table 2), and considering that continuous and sustained classifiers can represent uncorrelated information, we combined them into a single classifier.

For this purpose, we chose to use a simple linear fusion scheme based on [14]. The basic idea behind it is to estimate a linear combination for the classification scores: beginning from the best single classifier, it iteratively introduces the scores from the best remaining classifier and estimates a pair of coefficients so that the linear combination of the scores results in a better classification result. Nevertheless, we couldn't use a mutual information measure as the dataset was simply too small to have an accurate estimate. Instead, the minimum EER criterion was used, based on the results from

Table 2 and following a leave-1-out cross-validation scheme for the entire combination process. The following pseudo-code summarizes the algorithm for a set of classifiers:

```
$C_i \leftarrow$ sort ($C_i$, ascending EER)
$C_{fused} \leftarrow C_1$
for $i \leftarrow 2$, $i < $ numel($C$), $i \leftarrow i+1$ )
    $C^*_{fused} \leftarrow$ normalize( $C_{fused}$ )
    $C^*_i \leftarrow$ normalize( $C_i$ )
    $(\alpha, \beta) \leftarrow$ argmin( EER( $\alpha \cdot C^*_{fused} + \beta \cdot C^*_i$ ) )
    $C_{fused} \leftarrow \alpha \cdot C^*_{fused} + \beta \cdot C^*_i$
end
```

The results from the combination of the best sustained speech classifier (Childers-UBM, 16 components), with several configurations of the continuous speech classifier (16 to 256 gaussians, case 1 in Table 2), are shown on Table 3.

Table 3. *EER results for the classifiers linear fusion*

	Components in Continuous Speech Classifier				
	256	128	64	32	16
Sustained Speech 16-GMM	17.3	18.4	18.6	19.3	21.1

4. Conclusions

In this contribution we have highlighted the influence training data has on the final GMM models currently being used for OSA classification. The limitations of the available database and the complexity of the OSA acoustic space encouraged us to use a conventional two-step modeling scheme, introducing a universal background model (UBM) in order to improve the modeling, just as it is frequently done in Speaker Recognition. However, this scheme has several limitations related to the amount of available data. Therefore, the complexity of the models should be kept as low as possible to enhance the final detection system.

According to the results shown in Table 2, the GMM-classifier for connected speech (1) seems to perform much better than the sustained speech ones (2 to 5). Taking a deeper insight, there seems to be a tricky balance between the amount of data available, the complexity of the models trained, and the proximity of the prior distribution to the posterior one, which must be taken into account when building sustained speech acoustic models for OSA speakers (2 to 4). Nevertheless, the best results were obtained with a UBM trained on Childers DB (5), which is, from the two closest models (4 and 5 are context independent /a/ sounds), the one with the largest dataset.

Regarding the differences in the discriminative power for severe apnea cases detection, it seems that, as long as standard MFCC parameterization is the only one we have used, there is a huge difference between continuous and sustained speech. However, this is just one among many possible features that could be used for speech analysis and comparison, and thus, we refuse to say that there is little information in sustained vowels for OSA classification, while we currently work on this particular topic.

Moreover, the combination of sustained and connected speech has been proved to be extremely useful to improve the classification rates. Due to the fact that the information embedded in both classifiers is highly uncorrelated, a remarkable improvement has been achieved, up to a 25.1% relative reduction in EER. We are quite enthusiastic on these results, though some improvement is expected when additional features are included into the acoustic space description.

5. Acknowledgements

The activities described in this paper were funded by the Spanish Ministry of Science and Innovation as part of the TEC2009-14719-C02-02 (PriorSpeech) project. The authors will like to thank MD José Alcázar and Prof. Luis A. Hernández for their support.

6. References

[1] Puertas, F.J., Pin, G., María, J.M., & Durán, J, "Documento de consenso Nacional sobre el síndrome de Apneas-hipopneas del sueño". Grupo Español De Sueño, 2005.

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

[3] Fernandez R., Hernández L. A., López E., Alcázar J., Portillo G., & Toledano D. T. "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, 2008.

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

[5] Fiz, J.A., Morera, J., Abad, J., Belsulnces, A., Haro, M., Fiz, J.I., Jane, R., Caminal, P., & Rodenstein, D., "Acoustic analysis of vowel emission in obstructive sleep apnea". In Chest Journal; 104: 1093 – 1096, 1993.

[6] Fernández, R., Blanco, J.L., Hernández, L., López, E., Alcázar, J. and Torre, D., "Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker Recognition Techniques," EURASIP Journal on Advances in Signal Processing, vol. 2009, Article ID 982531, 11 pages, 2009.

[7] Blanco, J.L., Fernández, R., Díaz, D., Hernández, L.A., López, E. and Torre, D., "Apnoea voice characterization throught vowel sounds analysis using generative Gaussian Mixture Models". Proceedings of the 3rd Advanced Voice Function Assessment international Workshop. May 2009.

[8] Blanco, J.L., Fernández, R., López, E. and Hernández, L.A., "Exploring differences between phonetic classes in Sleep Apnoea Syndrome Patients using automatic speech processing techniques". To be published in the Journal of the International Phonetic Association. 2011.

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

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

[11] Pelecanos, J., Vogt, R. and Sridharan, S., "A Study on Standard and Iterative MAP Adaptation for Speaker Recognition". Proceedings of the 9th Australian International Conference on Speech Science & Technology. December 2002.

[12] Alonso, J.B., de Leon, J., Alonso, I. and Ferrer, M.A., "Automatic Detection of Pathologies in the Voice by HOS Based Parameters". EURASIP Journal on Applied Signal Processing 2001:4, 275-284, 2001.

[13] Morales, N., Toledano, D.T., Hansen, J.H.L., Colás, J. and Garrido, J., "Statistical class-based MFCC enhancement of filtered and band-limited speech for robust ASR". Proceedings Interspeech'05, pp. 2629-2632. September 2005.

[14] Al-Ani, A., Deriche, M., and Chebil, J., "A new mutual information based measure for feature selection". Intelligent Data Analysis, 7(1), pp 43-57, 2003.

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

[16] Childers, D.G., "Speech Processing and Synthesis Toolboxes". John Wiley & Sons, 2000.