A Novel Expert Classifier Approach to Pre-Screening Obstructive Sleep Apnea during Wakefulness

Cameron A. MacGregor, *Student Member, IEEE,* Zahra Moussavi, *Senior Member, IEEE*

Abstract— **Obstructive sleep apnea (OSA) is a widespread disorder that is cumbersome to diagnose using the gold-standard, overnight polysomnography (PSG). This paper highlights further development of our Awake-OSA method for predicting whether someone has severe sleep apnea using breath sounds recorded during wakefulness. We propose the use of an expert classification approach that consists of individual majority-voting classifiers. Each classifier is trained to distinguish one class of subject from all other classes. The outcomes of these classifiers are, in turn, combined using a truth matrix to determine the final outcome. Using the breath sound features of 249 subjects, the classifiers attempted to classify 180 subjects as either non-OSA (AHI less than 5) or severe-OSA (AHI greater than 30). 79% and 75% of OSA and non-OSA subjects, respectively, could be classified. Of those classified, the resultant testing sensitivity and specificity were found to be 78% and 86%, respectively. The consistency of the testing to training accuracies indicates the robustness and generalizability of using multiple expert classifiers on the dataset. This technique has the potential to be used in a doctor's office to rapidly and cheaply pre-screen for OSA, so that physicians may be better able to determine which patients are in need of overnight PSG.**

I. Introduction

Obstructive sleep apnea (OSA) is a prevalent disorder that affects at least 2% of women and 4% of men above the age of 30 [1]. It is believed that many people with OSA are undiagnosed [2]. If left undiagnosed and untreated, evidence suggests an increased risk of hypertension and traffic collisions [3]. The gold standard method of sleep apnea diagnosis is overnight polysomnography (PSG), which is costly, time-consuming, and has long wait times between subject referral and diagnosis. Furthermore, in the dataset examined in our previous studies [4, 5], the greatest proportion of subjects referred to overnight PSG did not have severe OSA that required treatment. Thus, a low-cost method of pre-screening subjects to identify those who could benefit most from overnight PSG would enable more efficient utilization of sleep clinic resources, thereby reducing healthcare costs. Since those with OSA tend to have a narrower upper airway during wakefulness [6], the effect of this anatomical difference on subjects' breath sounds might be exploited during wakefulness as a quick and inexpensive pre-screening tool; with such a tool it would be possible to determine whether the time and expense of an overnight PSG is justified. Studies conducted by members of our research group demonstrate the potential that breath sounds recorded during wakefulness have in screening subjects for OSA [4, 5]; we call the classification method that has arisen from this research "Awake-OSA". In this study, we collected additional data, modified our previous classification techniques [4], and found the cross-validated sensitivity and specificity of our classifier.

II. Methods

A. Data Collection

The data in this study consists of a dataset adopted from our previous studies [4, 5], and new data which has since been collected. All data were collected from consenting adult subjects who were referred to the Sleep Disorder Centre, Misericordia Hospital, Winnipeg, Canada, prior to undergoing overnight polysomnography (PSG). Data collection for this study was approved by the biomedical research ethics boards of both the University of Manitoba and Misericordia Hospital. The tracheal breath sounds of each subject were recorded using a Sony microphone (ECM-77B) embedded in a 6 mm diameter chamber with a 2 mm cylindrical space between the microphone and the skin, and placed over the suprasternal notch of the subject's neck. The chamber was attached to the skin with double-sided adhesive tape. The microphone and chamber were held in place with a soft neckband, which was fastened gently around subject's neck (Fig. 1). Sound signals were amplified with a gain of 200, band-pass filtered with cut-off frequencies of 0.05 Hz to 5 kHz, and sampled at a rate of 10240 Hz, all using Biopac (DA100C) amplifiers. The cut-off frequencies were chosen within the available filter options of the amplifier to remove DC and avoid aliasing.

The subjects were instructed to breathe deeply, following the hand gestures of the experimenter to ensure constant airflow, in 4 configurations: (i) nose and (ii) mouth breathing while upright, (iii) nose and (iv) mouth breathing while

Figure 1. The microphone and neckband used (left); how the microphone was mounted on the subject (right).

This study was supported in part by Natural Sciences and Engineering Research Council of Canada (NSERC) and also by Philips. The authors would like to thank Ehsan Shams and Davood Karimi for assisting with data collection. The assistance of the staff of the Sleep Disorder Centre, Misericordia Hospital, Winnipeg, MB, Canada in facilitating the collection of data used in this study is greatly appreciated.

Cameron MacGregor (ummacgrc@cc.umanitoba.ca) and Zahra Moussavi (mousavi@cc.umanitoba.ca) are with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada.

supine. All recordings started with the inspiratory phase. The apnea-hypopnea index (AHI), a measure of sleep apnea severity, as determined by a trained sleep technician at the hospital, was later collected from the sleep clinic.

Both datasets combined contain 308 subjects with usable sound data (Table 1). The subjects were assigned to three classes of OSA severity based on their general categorical diagnosis: 122 subjects (66 males, average age 50 ± 13 years) with AHI≤5 do not have OSA (G1); 69 subjects (44 males, average age 56 ± 12 years) with 10≤AHI≤25 have moderate OSA (G2); 58 subjects (47 males, average age 51 ± 11 years) with AHI≥30 have severe OSA (G3).

We used the G1, G2, and G3 subjects to train our expert classifiers, but tested the classifier using only the G1 and G3 subjects. In other words, in this pilot study, we only focus on two-class classifications, while the data from three classes are used for training.

B. Signal Conditioning

We manually sequestered the tracheal respiratory sounds into their inspiratory and expiratory phases. We always started recording at the inspiratory phase; however, we verified the respiratory phases by auditory and visual inspection of the spectrogram of the data. We listened to all individual breath sounds, and those which contain noise from an external source (eg. door slamming) or internal source (eg. thumps caused by mucus) were discarded.

TABLE I. NUMBER OF SUBJECTS WITH ANTHROPOMETRIC INFORMATION WITHIN EACH CLASSIFICATION GROUP.

AHI Range	# of Subjects	# of Males	Age* [years]	BMI* [kg/m^2]
AHI ≤ 5	122	66	50 ± 13	31 ± 7
10 ≤ AHI ≤ 25	69	44	56 ± 12	33 ± 6
AHI ≥ 30	58	47	51 ± 11	39 ± 8
Overall	249	157	52 ± 13	33 ± 8

*Age and BMI (body-mass index) are reported as mean ± standard deviation.

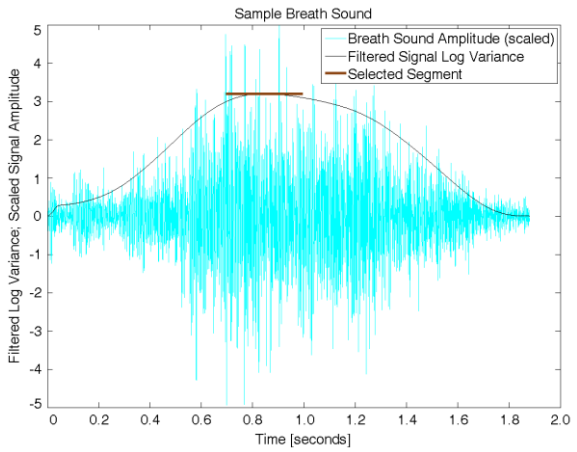

Figure 2. The selected breath segment shown in the context of the filtered log variance and filtered signal

In both the old and new datasets, 300 ms of audio was isolated from each usable breath corresponding to the upper 40% of respiratory flow [7] using the following steps (note that these steps are not used to modify the breath sound, they are only to select a 300 ms segment of the breath sound within the middle of the respiratory phase). (1) Band-pass filter the isolated breath sound using a second order Butterworth filter between 500 and 2500 Hz to prevent low or high frequency noise from influencing the log-variance plot generated in step 4. (2) Normalize the signal by zeroing the mean and dividing by the standard deviation. (3) Slide a hamming window across the whole signal. (4) At each new position, multiply each segment by the hamming window and calculate log of the variance of the resultant windowed signal. (5) Low-pass 2nd order Butterworth filter with cut-off frequency of 5 Hz was applied to capture the slow, broad changes and remove the quick, jittery changes in log variance. (6) The log-variance curve maximum was located, and a 300 ms rectangular window was centered about this point. The audio of the original signal outside the rectangular window was discarded. Figure 2 illustrates part of this process by showing, for a sample breath sound, the filtered and scaled breath sound amplitude, filtered log variance of the signal, and the selected segment all on the same plot. This process was applied to every breath sound segment to detect the 300 ms segment within the upper 40% area of the flow signal; note that since we did not record the flow signal, we adopted this routine to ensure we extract a segment of sound within the upper 40% of the flow.

Next, we calculated spectral and bispectral features from the original sound data within the 300 ms segments covering frequency ranges within 100 and 2600 Hz; the details of the features, specific sub-frequencies used, and equations used to calculate the features are described in [8]. Since we hypothesize that changes in upper airway anatomy affect the breath sounds, it is logical to assume that these changes might be detectable from the power spectra and bispectra of the breath sounds. We chose spectral features that characterize how the power is distributed over different frequency bands that are of interest and common in respiratory sounds analysis. The extracted spectral features were: signal power, relative signal power, spectral centroid, spectral bandwidth, spectral flatness, and crest factor. Spectral features, while appearing intuitively useful, may not contain enough desirable information for the classifier; this is because, like the power spectrum they are calculated from, they do not contain information on phase relations between different frequency components. Bispectral features, computed from the bispectra of the breath sounds, were included so that potential information within the phase relationships between frequency components might be considered. Bispectral features computed from the bispectrum may capture the tendency towards turbulent airflow in a narrower airway [9]. The extracted bispectral features are: bispectral invariant parameter, average magnitude, average power, normalized entropy, normalized squared entropy, sum of logarithmic amplitudes, first- and second-order moments of the logarithmic amplitudes, phase entropy, and median bifrequency. These features were calculated from the sound

signals recorded in the 4 different recording configurations. Thus, a vector of all feature values was generated for each 300 ms breath clip. For each subject, each feature value was averaged over all breaths of the same respiratory phase of the same recording configuration (e.g. nose breathing while supine). These average values were then used by the classifiers during the training and testing phases.

C. Classifier Design

A two-class "expert" classifier design is proposed which is intended to deal with classification of a very heterogeneous population, such as the breath sounds of people with and without OSA. Specifically, an "Expert-G1" classifier, which is trained to distinguish G1 from non-G1 (i.e. G2 and G3) subjects, and an "Expert-G3" classifier, which is similarly trained to distinguish G3 from non-G3 (i.e. G1 and G2) subjects, were developed and used to class all subjects in G1 and G3 using a truth matrix. Each classifier uses our previously developed majority voting strategy [4], where each sound feature used by the classifier votes for the group whose median feature value is closest to the subject's feature value. The outcome with the highest number of votes becomes the final outcome; a tie in votes is broken using the total sum of all absolute feature differences to group medians for each group.

The outcomes of Expert-G1 and Expert-G3 are combined to form the final classification of each subject, as shown in Table 2. When the classifier outcomes are not in conflict, the final classification is made according to the logical outcome. If, however, the classifier outcomes are in conflict, no classification of the subject is made. For example, the subjects that Expert-G1 claim as its own and that Expert-G3 do not claim as its own are classed as G1; the same logic applies for the G3 outcome. In contrast, the subjects that both Expert-G1 and Expert-G3 claim as their own are not classified. Therefore, after examination by the expert classifiers, the final outcome for all G1 (non-OSA) and G3 (OSA) subjects is either non-OSA, OSA, or unclassified.

D. Classifier Training and Cross-Validation

10-fold cross-validation was used, in tandem with a sequence of feature reduction techniques, to test the robustness of both the expert classifier and feature selection schemes; that is, the ability to generalize to new data. Subjects in groups G1, G2, and G3, representing a total of 249 subjects, were randomly ordered and partitioned into 10 equally sized folds. One fold, called the testing set, was set aside, while the remaining 90% of the dataset, called the training set, was used to select features, or "train" Expert-G1 and Expert-G3 to classify their own kinds of subject.

TABLE II. TRUTH MATRIX DESCRIBING HOW EXPERT CLASSIFIER OUTCOMES ARE COMBINED TO PRODUCE THE FINAL OUTCOME

		Possible Outcomes			
Expert-G1 classifier	***(G1) vs. (G2+G3)***	(G1)	(G2+G3)	(G1)	(G2+G3)
Expert-G3 classifier	***(G3) vs. (G1+G2)***	(G1+G2)	(G3)	(G3)	(G1+G2)
Final Outcome		G1	G3	Not classified	

The training process finds a set of features for each classifier that allows the classifier to most accurately class its own grouping of subjects within the particular fold. For Expert-G1, the training set was split into G1 and non-G1 (G2 + G3) subjects, thereby forming the two groups that Expert-G1 will be trained to distinguish between; this was likewise done for Expert-G3 by splitting the same training set into G3 and non-G3 (G1 + G2) subjects. The following steps were applied, in parallel, to the sub-groupings of subjects associated with the Expert-G1 and Expert-G3 classifiers. (1) A two-tailed Welch's *t*-test ($P \leq 0.01$) was first applied to select the significantly different sound features; Welch's *t*-test was used to address the potential unequal feature variances between groups of subjects. (2) Next, the minimum redundancy maximum relevancy algorithm [4] was used to rank the significant features. (3) The top 30 ranked features were passed to a floating search algorithm [4] that was used to select 1 to 15 features that together resulted in the highest classification accuracy for the training set. Accuracy was defined as the sensitivity of true positive identification of each individual group. (4) Expert-G1 and Expert-G3 were then used to classify all G1 and G3 subjects in both the training and testing sets based on the group feature medians of the training set. (5) Finally, the outcomes of Expert-G1 and Expert-G3 were combined as previously described for both the testing and training sets. These 5 steps were repeated for all 10 folds, thereby allowing all subjects to be omitted from the training set exactly once. In summary, the training process selects a set of features for each fold and classifier that can most accurately class the training subjects; the training process does not, however, select features to optimize the final accuracy that results from combining the Expert-G1 and Expert-G3 classifiers.

III. RESULTS AND DISCUSSION

A. Training Results

For each fold, combining the classifications made by Expert-G1 and Expert-G3 of the training subjects results in a specific number of subjects assigned to each of the final outcomes of Table 2. The number of training subjects assigned to each final outcome varies between folds. Thus, the training results are reported in Table 3 as the mean and standard deviation of the number of training subjects to be assigned to each final outcome over all 10 folds. On average, 87% ± 4% and 91% ± 4% of G1 and G3 subjects, respectively, could actually be classified; that is, the expert classifiers made a prediction that was logically consistent with one another. Of the subjects that could be classified, 87% ± 9% and 81% ± 8% of G1 and G3 subjects, respectively, were correctly classified.

B. Testing Results

Combining the classifications made by Expert-G1 and Expert-G3 of the testing subjects results in each G1 and G3 subject being examined by the classifiers exactly once. Thus, while the training results are reported as averages, the testing results are reported as absolute numbers. How G1 and G3 subjects were classified by combining the output of the

TABLE III. MEAN ASSIGNMENT COUNT ± STANDARD DEVIATION OF TRAINING SUBJECTS OVER ALL 10 FOLDS

			Predicted Class	
		# of Not Classified	*G1 (non-OSA)*	*G3 (OSA)*
True Class	*G1 (non-OSA)*	16.0 ± 5.3	81.3 ± 5.3	12.5 ± 2.8
	G3 (OSA)	5.2 ± 2.6	9.1 ± 1.5	37.9 ± 2.6

expert classifiers, over all 10 testing folds, is shown in Table 4. 75% and 79% of G1 and G3 subjects, respectively, could actually be classified. Of the subjects that could be classified, 86% and 78% of G1 and G3 subjects, respectively, were correctly classified.

C. Discussion

The testing accuracies are approximately equal to the training accuracies, with only a ~10% drop in the number of G1 and G3 subjects that could be classified, compared to the training results. This indicates that the classification approach is both robust and generalizable to data that the classifiers have not seen before.

Overall, in the dataset of a very heterogeneous population, such as breath sounds during wakefulness, the classic approaches of classification fail due to wide variability of the data and confounding factors; for example, not only does OSA affect the breath sound characteristics, but smoking, height, weight, age, etc. also affect the breath sounds; thus, usually the classic approach of classifications do not give a high accuracy. In contrast, we believe the expert classifier design, proposed in this paper, would result in reasonable accuracy in comparison to the classic approaches.

IV. CONCLUSION

This pilot study demonstrates a basic example of combining multiple, specifically-trained expert classifiers to determine a final outcome which is suitable for heterogeneous populations. We demonstrated that this classification technique, which uses the features of breath sounds recorded during wakefulness, has reasonable power to predict whether study subjects have severe OSA. The results are promising for three reasons. First, the majority (76%) of non-OSA and OSA testing subjects could be classified. Second, the testing subjects that could be classified were classified with favorable sensitivity (78%) and specificity (86%). Third, and finally, the consistent sensitivity and specificity between training and testing phases indicate the generalizability of the classification approach to data that were not part of the initial classifier training process. This indicates that our classification approach is reasonably generalizable to subjects that the classifiers have never encountered before. This generalizability supports our goal of developing the Awake-OSA method into a fast, non-invasive, painless, and economical method of pre-screening for OSA in a doctor's office. This would allow physicians to better determine which subjects are in need of the detailed information provided by overnight PSG. From both a subject and cost perspective, the potential benefits of Awake-OSA are preferable due to the uncomfortable, expensive, and time-consuming nature of overnight polysomnography.

TABLE IV. ABSOLUTE ASSIGNMENT COUNT OF TESTING SUBJECTS OVER ALL 10 FOLDS

			Predicted Class	
		# Not Classified	*G1 (non-OSA)*	*G3 (OSA)*
True Class	*G1 (non-OSA)*	31	79	12
	G3 (OSA)	12	10	36

A. Future Work

The next immediate step is to augment the decision matrix (Table 2) to take into account additional combinations of expert classifier outcomes. This would allow more subjects to have a final classification assigned to them. For example, when the outcomes of the individual expert classifiers are in conflict, the proportion of votes between classifiers could be compared to resolve the conflict.

Next, to better confirm the practical viability of combining multiple expert classifiers to determine the final predicted outcome, an in-depth analysis of how G2 (moderate-OSA) subjects are classed is required. Finally, subjects with AHI indices that excluded them from each of the three groups in this study (G1, G2, G3) should also be classed to see whether they are placed in one of their neighboring groups.

REFERENCES

[1] "Wait times for sleep apnea care in Ontario: A multidisciplinary assessment" Brian W Rotenberg, Charles F George, Kevin M Sullivan, Eric Wong, Can. Respir. J., vol. 17, no. 4, July/August 2010.

[2] T. Young et. al., "The occurrence of sleep-disordered breathing among middle-aged adults," in N. Engl. J. of Med., vol. 328, no. 17, pp. 1230-1235, 1993.

[3] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of Obstructive Sleep Apnea: A Population Health Perspective," in Am. J. Respir. Crit. Care. Med., vol. 165, pp. 1217-1239, 2002.

[4] D. Karimi, "Spectral and Bispectral Analysis of Awake Breathing Sounds for Obstructive Sleep Apnea Diagnosis," M.Sc. thesis, Dept. of Elec. and Comp. Eng., Univ. of Manitoba, Winnipeg, MB, 2012.

[5] A. Montazeri, E. Giannouli and Z. Moussavi, "Assessment of Obstructive Sleep Apnea by Respiratory Sound Analysis during Wakefulness," *J. Annals on Biomed. Eng.,* vol. 4, no.4, pp. 916-924, 2012.

[6] R. J. Schwab et al., "Identification of Upper Airway Anatomic Risk Factors for Obstructive Sleep Apnea with Volumetric Magnetic Resonance Imaging," *Am. J. Respir. Crit. Care. Med.*, vol. 168, no. 5, pp. 522-530, May 13, 2003.

[7] A. Yadollahi and Z. Moussavi, "Acoustical Flow Estimation: Review and Validation," *IEEE EMB Magazine*, vol. 26, no. 1, pp. 56-61, Jan. 2007.

[8] C. A. MacGregor, D. Karimi, A. Azarbarzin, Z. Moussavi, "Statistical Analysis of Tracheal Breath Sounds during Wakefulness for Screening Obstructive Sleep Apnea," *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annu. Int. Conf. of the IEEE*, Osaka, Japan, pp. 4549-4552, July 3-7, 2013.

[9] E. Shams, D. Karimi, Z. Moussavi, "Bispectral analysis of tracheal breath sounds for Obstructive Sleep Apnea," *Engineering in Medicine and Biology Society (EMBC), 2012 Annu. Int. Conf. of the IEEE*, San Diego, CA, pp. 37-40, Aug. 28 - Sept. 1, 2012.