CLASSIFICATION OF RESPIRATORY EFFORT AND DISORDERED BREATHING DURING SLEEP FROM AUDIO AND PULSE OXIMETRY SIGNALS

Brian R. Snider and Alexander Kain

Oregon Health & Science University
Center for Spoken Language Understanding
Portland, OR, USA
sniderb@ohsu.edu, kaina@ohsu.edu

ABSTRACT

Sleep-disordered breathing (SDB) is a highly prevalent condition associated with many adverse health problems. As the current means of diagnosis (polysomnography) is obtrusive and ill-suited for mass screening of the population, we explore a minimal-contact, automatic approach that uses acoustics-based methods in conjunction with pulse oximetry. We present a two-stage method for automatically classifying breathing sounds produced during sleep to track respiratory effort and predicting disordered breathing events using respiratory effort durations and oxygen desaturations. We compare our method for tracking respiratory effort and predicting disordered breathing with human expert event scoring. Our subject-independent method tracks respiratory effort with 87% accuracy and predicts disordered breathing events with 40–52% accuracy.

Index Terms— sleep apnea, breathing, polysomnography

1. INTRODUCTION AND BACKGROUND

Sleep-disordered breathing (SDB) is believed to be a widespread, under-diagnosed condition associated with many detrimental health problems [1, 2]. Young et al. describe the total burden of sleep-disordered breathing on the health system and society as "staggering" [3]. The current gold standard for diagnosis of sleep-disordered breathing is a sleep study, or polysomnography (PSG). This overnight procedure is obtrusive, requiring many sensors to be attached to the patient's body; moreover, it is time-consuming, expensive, and ill-suited for mass screening of the population.

Previous studies have explored acoustics-based approaches using high-quality audio recordings of sleep breathing sounds. Several studies focus on snore detection, as snoring is seen as a possible indicator for the most common form of SDB, obstructive sleep apnea (OSA) [4, 5], based on the hypothesis that snore signals carry relevant information about the state of the upper airways, especially the partial or full collapse thereof [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16].

Current work seeks to not simply supplement typical PSG sensors with less obtrusive alternatives, but to also explore automatic and computer-assisted manual SDB event scoring of clinical and home sleep recordings, with an eye toward "scoring as a service" rather than in-house scoring. Several recent studies address various facets: the performance of automated PSG scoring versus computer-assisted manual scoring [17]; the accuracy of automated scoring of home sleep recordings [18]; the efficacy of portable sleep testing [19]; inter-labeler agreement across sleep centers [20, 21]; and a comprehensive survey of the numerous "state of the art" methods for computer-assisted SDB event scoring [22].

In this study, we build on our previous work [23], using minimally-obtrusive sensors and an automatic, two-stage method for classifying breathing sounds to track respiratory effort and predicting disordered breathing events using respiratory effort durations and oxygen desaturations.

2. DATA

2.1. Data collection

As our approach relies on acoustic data not typically collected during full-night PSG, we created a corpus of PSG sensor data and time-aligned high-quality audio. We collected the data during routine clinical PSG at Oregon Health & Science University's sleep lab. Trained PSG technicians and clinicians scored each study per the AASM guidelines in effect at the time of the study [24]. A total of 15 adult subjects participated in the study.

We recorded uncompressed 16-bit audio at a sampling rate of 16 kHz using a highly directional microphone (Audio-Technica AT8035). The microphone was affixed to an articulated microphone stand in the subject's room and oriented toward the subject's head when in a supine position. Audio recordings were made in parallel with typical PSG sensor data during each overnight study. We time-aligned the separate, high-quality audio recordings with the low-quality audio captured by the PSG system's passive infrared video camera, thereby time-aligning the high-quality audio with the PSG sensor data.

We referred to technician annotations on the scored PSG studies to exclude audio recorded before each subject fell asleep or after he or she woke up to constrain our analysis to only those respiratory sounds made during actual sleep. We also excluded audio that was captured after remedial measures were taken (e. g. positive airway pressure was titrated or oxygen was administered), as these measures introduced additional airflow noise in the sleep environment near the subject's mouth and nose. Many subjects had very little time asleep before remedial measures were taken. Finally, we also excluded audio that contained air conditioner, fan, furnace, or television background noise. After considering these factors, only 4 of the 15 subjects had usable sleep breathing audio.

2.2. Manual respiratory effort labeling

For each subject, we identified four continuous regions of audio, each approximately four minutes in length. We selected these regions from various times during the night to cover possible differences due to stage of sleep, bed posture, varied breathing patterns and rates, and episodes of snoring. Additionally, we consulted the PSG data to ensure that a variety of apneic and typical events were present in the selected regions of audio.

We then listened to each region of audio while visually inspecting the corresponding spectrogram and applied respiratory effort labels. We labeled inhalation as either *breathing-in* (*Bi*) or *snoring-in* (*Si*), and exhalation as either *breathing-out* (*Bo*) or *snoring-out* (*So*). Finally, we labeled the remaining portions as *no-effort* (*N*).

Figure 1 illustrates a brief excerpt of an example region and the corresponding respiratory effort labels. Note that a single inhalation or exhalation may consist of more than one constituent type, such as a *breathing-in* that turns into a *snoring-in*. During manual labeling, we restricted a single inhalation or exhalation to include up to three constituent portions. For example, an inhalation may be as complex as *Bi-Si-Bi* (see Figure 1, 3.8–5.6 seconds), but not *Bi-Si-Bi-Si*. Only one of the subjects exhibited *So* events in the selected regions.

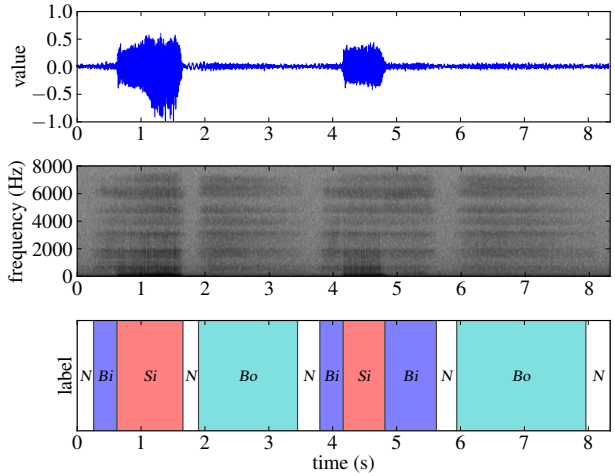

Fig. 1. Waveform, spectrogram, and respiratory effort labels for a brief excerpt of an example region of audio.

3. EXPERIMENT

Building on our previous work, we first used acoustic features extracted from high-quality audio to classify sleep breathing sounds into respiratory classes (Stage I, Section 3.1). Then, we used features extracted from the *output* of Stage I with additional features extracted from pulse oximetry data (in this case, peripheral capillary oxygen saturation, or SpO_2) for use in a new, second-stage classifier (Stage II, Section 3.2).

3.1. Stage I: Respiratory effort classification

3.1.1. Feature extraction

We used reflection coefficients from linear predictive coding (LPC), the highest-performing acoustic features from our previous work. We extracted these features from the audio waveform for each region using a frame length of 150 ms, zero overlap, and a Hanning analysis window. For each frame, we calculated 13 LPC coefficients and their first-order deltas.

3.1.2. Classifier

During initial exploration of our corpus, we noted that the intra-cycle no-effort portion was typically much shorter than the inter-cycle no-effort portion. We therefore created four discrete no-effort types that preserve the temporal relationship of the no-effort label with the surrounding respiratory effort labels. We use a hidden Markov model (HMM) to predict state sequences to capitalize on the sequential nature of a typical respiratory cycle. We assume *a priori* that respiratory effort states evident in the acoustic data can be learned and predicted by the HMM, much like phone states in speech recognition applications. Figure 2 illustrates the topology of the Stage I model. Note that each respiratory effort type consist of three states per label, while the no-effort types only consist of one state.

We observed many interesting respiratory cycle phenomena during manual labeling. For example, within a single inhalation, a breath-in may turn into a snore-in; likewise, during an exhalation, a snore-out may degrade into a breath-out. Additionally, we observed that an inhalation may be immediately followed by an exhalation, with no intermediate no-effort state (top arcs emanating from null states in Figure 2). Finally, we account for multiple short inhalation or exhalation attempts in rapid succession, separated by brief no-effort (*N*) states (bottom arcs). We observed this type of phenomenon during obstructive apnea events, when a subject tried repeatedly to breathe in with limited success. We designed our model to capture these various phenomena via learning the transition probabilities between states.

3.1.3. Automatic label remapping

A remapping algorithm is necessary to convert respiratory effort labels to the state names used by the model. Similar to our previous work, we divided individual inhalation and exhalation labels according to the same original rules: (1) if a label consists of one constituent, divide the label into three equal-duration states; (2) if a label consists of two constituents, divide the longer-duration portion into two equal-duration states, and assign the shorter-duration portion to a third state; and (3) if a label consists of three constituents, assign each portion to a single state, preserving the original durations.

We enhanced the remapping algorithm to convert no-effort (*N*) labels into the four discrete no-effort types: *N* between inhalations (*Nii*); *N* between inhalation and exhalation (*Nio*); *N* between exhalation and exhalation (*Noo*); and *N* between exhalation and inhalation (*Noi*).

3.1.4. Training and testing

We used a k-fold cross-validation scheme, separating the data into different training and testing sets. For each fold, we held out one subject's data for testing, using the remaining three subjects' data for training. Each training set contained 48 minutes of audio, with 16 minutes held out for testing. The held-out portion was cycled through all four folds, and the resulting fold's training and testing sets were used to train and test the Stage I classifier, respectively.

We calculated the start probabilities (π) and transition probabilities (A) using observed sequences from the training set. Next, using the state-labeled data, we grouped the frame-level feature vectors from the training set by state. For each state, we calculated the mean and covariance of the feature vectors for that state. We used these statistics to fit a Gaussian mixture model (GMM) for each state (with three mixture components and full covariance), to model the observation probabilities (B). Then, we initialized the HMM using the precomputed π-values and A and B matrices. Finally, we used the HMM to decode the test set using the Viterbi search algorithm. We recorded the predicted state sequences, mapping model state names back to respiratory effort labels, and then merging identical adjacent labels to enable direct comparison to the original, manually labeled sequences.

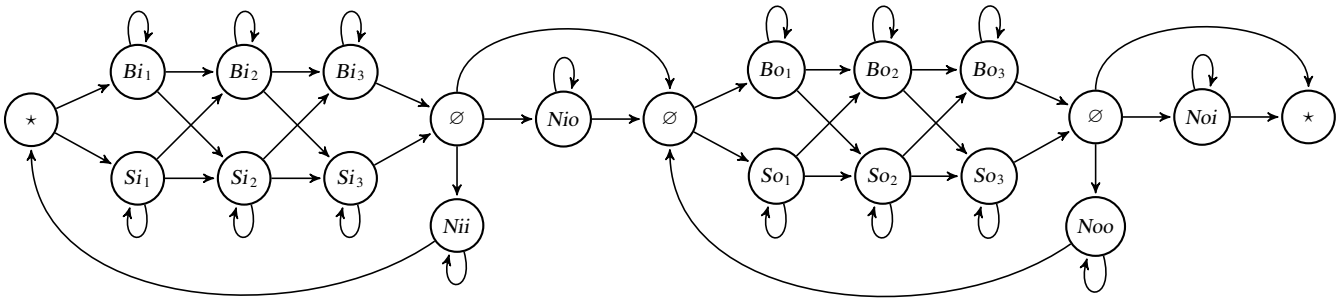

Fig. 2. HMM topology with three states per respiratory label type (*Bi*, *Bo*, *Si*, *So*) and one state per no-effort type (*Nii*, *Nio*, *Noo*, *Noi*). Stars ($\star$) denote the null state at the start of a respiratory cycle. Nulls ($\varnothing$) denote intermediate null states.

3.2. Stage II: Disordered breathing classification

3.2.1. Feature extraction

When scoring PSG studies, trained clinicians and PSG technicians evaluate respiratory effort, airflow, and SpO_2 data to identify disordered breathing events, looking for reduction in or cessation of breathing effort or airflow and a corresponding drop in blood oxygen saturation, in accordance with AASM guidelines [24]. With these criteria in mind, we created new feature vectors for Stage II by extracting respiratory effort features from the *output* of Stage I, incorporating additional SpO_2 features extracted from the PSG data.

During manual labeling of respiratory effort (Section 2.2), we noted changes in respiratory effort label duration during disordered breathing events such as hypopnea (*H*), obstructive apnea (*OA*), or central apnea (*CA*), when compared to typical breathing (*T*) of that same effort type. (Mixed apnea was possible but not present in our corpus, so it is omitted from further discussion.) Based on this discovery, we selected duration-related respiratory effort features. Using the predicted respiratory effort labels from Stage I, we extracted the duration of the current respiratory effort label to create a *one-hot duration* vector for each frame. In this design, only one of the eight possible effort labels can be "hot" (i. e. non-zero) per frame.

Figure 3 depicts respiratory effort label durations wholly contained within a given disordered breathing event. Red scatterplot markers indicate a significant difference in mean duration from typical breathing (*T*) for a given respiratory effort label ($p < 0.05$). Note the significant shortening of inhalation (*Bi*, *Si*), intra-cycle no-effort (*Nio*), and exhalation (*Bo*) labels and lengthening of inter-cycle no-effort (*Noi*) labels during hypopnea (*H*). Also note the mere presence of snoring is not necessarily indicative of disordered breathing, but a significant decrease in snore-in (*Si*) duration may indicate disordered breathing. Similarly, a significant decrease in intra-cycle no-effort (*Nio*) may also indicate disordered breathing. Instances of snoring (*So*) and no-effort (*Noo*) during exhalation were very rare in our corpus.

We then extracted SpO_2 features from the time-aligned PSG data. First, we estimated a single "baseline" SpO_2 value per subject by computing the 95th-percentile SpO_2 value from each full-night study. Next, we computed a *desaturation from baseline* value for each frame, where desaturation was defined as the baseline minus the observed SpO_2 value. Finally, we appended this desaturation value to the one-hot duration vector to form the feature vector for each frame. Figure 4 illustrates the Stage I respiratory effort labels, SpO_2 desaturation, corresponding Stage II feature vectors, and disordered breathing event labels for a brief excerpt of training data.

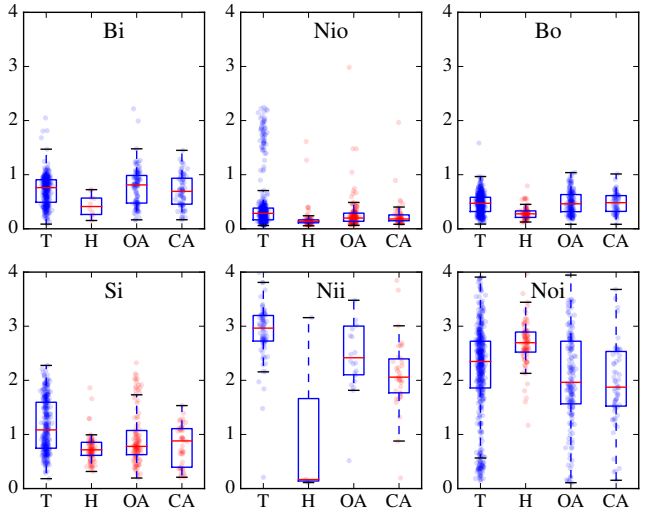

Fig. 3. Respiratory effort label (subplot title) durations (y-axis, in seconds) by disordered breathing type (x-axis). Red scatterplot markers indicate a significant difference in mean duration from typical breathing (*T*) of that same effort type ($p < 0.05$).

3.2.2. Classifier

We created a new, second-stage HMM to classify disordered breathing events during sleep. Figure 5 illustrates the topology of the Stage II model. In this stage, the possible states represent observed disordered breathing types: typical breathing (*T*), hypopnea (*H*), obstructive apnea (*OA*), and central apnea (*CA*).

3.2.3. Training and testing

As in Stage I, we used a k-fold cross-validation scheme, replacing the Stage I respiratory effort HMM topology with the Stage II disordered breathing topology, and using the Stage II duration and desaturation feature vectors. For each fold, we held out one subject's data for testing, using the remaining three subjects' data for training to initialize the HMM, then predicting the disordered breathing labels for the test set. During testing, we noted that the sparsity of the *So* and *Noo* effort types hindered Stage II classification accuracy; we then excluded those two types from the one-hot duration portion of the Stage II feature vectors.

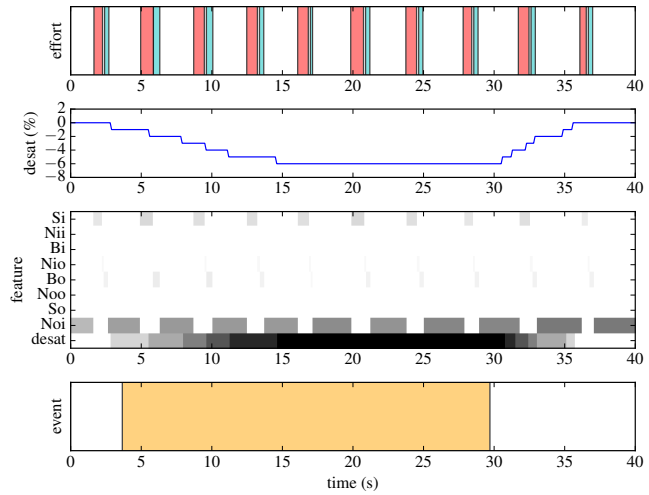

Fig. 4. Stage I respiratory effort labels (*Si*–*Bo* pattern), SpO_2 desaturation, corresponding Stage II duration and desaturation feature vector, and disordered breathing event labels (hypopnea, 3.5–29.5 s, surrounded by typical breathing) for a 40-second excerpt.

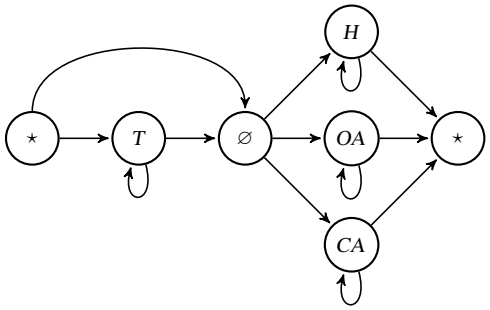

Fig. 5. Stage II HMM topology with one state per disordered breathing type (*H*, *OA*, *CA*) and one state for typical breathing (*T*).

4. RESULTS

4.1. Stage I

As in our previous work, we evaluated Stage I classifier accuracy at three levels of granularity: fine-, medium-, and coarse-grain accuracy. For fine-grain accuracy, we combined states of the same type into one event (e. g. predicted states Si_1, Si_2, and Si_3 were all merged into one *Si* event). The fine-grain accuracy was used to evaluate the basic accuracy of the classifier. For medium-grain accuracy, we combined in and out events of the same parent type into one category: *Bi* and *Bo* labels were both considered "breath" (*B*); *Si* and *So* were "snore" (*S*), and all four no-effort labels became "no-effort" (*N*). The medium-grain accuracy was used to evaluate the potential for identifying breaths and snores. Finally, for coarse-grain accuracy, we combined all breath and snore labels into one generic "effort" (*E*) label, to evaluate the potential for identifying effort versus no-effort in the breathing cycle.

Table 1 summarizes our Stage I accuracy results by label granularity. We observed very good tracking of respiratory effort at the coarse- and medium-grain levels, with the classifier generally predicting effort correctly but occasionally confusing loud breathing for snoring or very quiet breathing for no-effort. In the latter case, the predicted *N* labels exhibited high confusability with other *N* labels due to the classifier losing track of the inhalation–exhalation cycle.

	Label granularity		
	Fine	Medium	Coarse
Effort labels	*Bi*, *Bo*, *Si*, *So*	*B*, *S*	*E*
No-effort labels	*Nii*, *Nio*, *Noo*, *Noi*	*N*	*N*
Accuracy	0.57 (0.10)	0.75 (0.04)	0.87 (0.02)

Table 1. Stage I classifier results for mean fine-, medium-, and coarse-grain accuracy and standard deviation.

4.2. Stage II

We evaluated Stage II classifier accuracy in a similar manner as in Stage I, with two levels of granularity: fine- and coarse-grain accuracy. For fine-grain accuracy, we left events as is, allowing all four possible event labels: typical breathing (*T*), hypopnea (*H*), obstructive apnea (*OA*), and central apnea (*CA*). For coarse-grain accuracy, we combined all disordered events into one generic "disordered" label, to evaluate the potential for identifying typical breathing versus disordered breathing.

We ran four variations in Stage II: first using the manually-labeled respiratory effort when extracting the one-hot duration features, then using the Stage I-predicted respiratory effort; and a second pass considering only subjects with some degree of disordered breathing according to the scored full-night PSG. Table 2 summarizes our Stage II accuracy results by subject type, effort label source, and label granularity.

Subjects	Label source	Label granularity	
		Fine	Coarse
All	Manual	0.39 (0.22)	0.40 (0.23)
	Stage I	0.39 (0.22)	0.40 (0.22)
Disordered	Manual	0.50 (0.12)	0.52 (0.12)
	Stage I	0.50 (0.12)	0.51 (0.12)

Table 2. Stage II classifier results for mean fine- and coarse-grain accuracy and standard deviation, using manually-labeled ("Manual") and Stage I-predicted ("Stage I") respiratory effort.

5. DISCUSSION AND FUTURE WORK

Stage II classification accuracy, while generally poor, was relatively unchanged when moving from the manually-applied respiratory effort labels to the Stage I-predicted labels. Upon further exploration of the underlying audio and oximetry data, we discovered that many manually-labeled disordered events had minimal (i. e. 3–4%) desaturation or effort label duration. Despite attempting to adhere to AASM scoring criteria (3–4% desaturation from baseline with a reduction of effort lasting 10 seconds or more), we surmise that: (1) the included signals are insufficient to fully capture the notion of both respiratory effort *and* airflow; and (2) human experts use additional knowledge that is not codified in the AASM criteria, whereas our algorithm strictly adheres to the criteria.

We plan several enhancements in future work: first, to use effort and airflow data from PSG sensors directly, rather than a surrogate (high-quality audio); next, to create a substantially larger corpus to address sparsity issues; and finally, to leverage additional machinery (e. g. deep neural networks) to learn the features and patterns that trained human experts intuitively extract and recognize when evaluating PSG sensor data.

6. REFERENCES

[1] Terry Young, Paul E. Peppard, and D. J. Gottlieb, "Epidemiology of Obstructive Sleep Apnea: A Population Health Perspective," *American Journal of Respiratory and Critical Care Medicine*, vol. 165, no. 9, pp. 1217–1239, 2002.

[2] Vishesh Kapur, Kingman P. Strohl, Susan Redline, Conrad Iber, George O Connor, and Javier Nieto, "Underdiagnosis of Sleep Apnea Syndrome in U.S. Communities," *Sleep and Breathing*, vol. 6, no. 2, pp. 49–54, 2002.

[3] Terry Young, M. Palta, J. Dempsey, Paul E. Peppard, F. J. Nieto, and K. M. Hla, "Burden of Sleep Apnea: Rationale, Design, and Major Findings of the Wisconsin Sleep Cohort Study," *WMJ: Official Publication of the State Medical Society of Wisconsin*, vol. 108, no. 5, pp. 246–249, 2009.

[4] V. Hoffstein, *Snoring—The Principles and Practice of Sleep Medicine*, Saunders, Philadelphia, PA, 3rd edition, 2000.

[5] Dirk Pevernagie, Ronald M. Aarts, and Micheline De Meyer, "The acoustics of snoring," *Sleep Medicine Reviews*, vol. 14, no. 2, pp. 131–144, 2010.

[6] J. Sola-Soler, R. Jane, J. A. Fiz, and J. Morera, "Pitch analysis in snoring signals from simple snorers and patients with obstructive sleep apnea," in *Second Joint EMBS-BMES Conference 2002 24th Annual International Conference of the Engineering in Medicine and Biology Society. Annual Fall Meeting of the Biomedical Engineering Society*. 2002, pp. 1527–1528, IEEE.

[7] Udantha Ranjith Abeyratne, C. K. K. Patabandi, and Kathiravelu Puvanendran, "Pitch-jitter analysis of snoring sounds for the diagnosis of sleep apnea," in *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, 2001, pp. 2072–2075.

[8] T. H. Lee, Udantha Ranjith Abeyratne, Kathiravelu Puvanendran, and K. L. Goh, "Formant-structure and phase-coupling analysis of human snoring sounds for the detection of obstructive sleep apnea," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 3, 2000.

[9] J. A. Fiz, "Acoustic analysis of snoring sound in patients with simple snoring and obstructive sleep apnea," *European Respiratory Journal*, vol. 9, pp. 2365–2370, 1996.

[10] F. Dalmasso and R. Prota, "Snoring: analysis, measurement, clinical implications and applications," *European Respiratory Journal*, vol. 9, pp. 146–159, 1996.

[11] Udantha Ranjith Abeyratne, A. S. Wakwella, and Craig Hukins, "Pitch jump probability measures for the analysis of snoring sounds in apnea," *Physiological Measurement*, vol. 26, no. 5, pp. 779–798, 2005.

[12] W. D. Duckitt, S. K. Tuomi, and T. R. Niesler, "Automatic detection, segmentation and assessment of snoring from ambient acoustic data," *Physiological Measurement*, vol. 27, no. 10, pp. 1047–1056, 2006.

[13] Andrew Keong Ng, Tong San Koh, Eugene Baey, Teck Hock Lee, Udantha Ranjith Abeyratne, and Kathiravelu Puvanendran, "Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea?," *Sleep Medicine*, vol. 9, no. 8, pp. 894–898, 2008.

[14] Asela S. Karunajeewa, Udantha Ranjith Abeyratne, and Craig Hukins, "Silence-breathing-snore classification from snore-related sounds," *Physiological Measurement*, vol. 29, no. 2, pp. 227–243, 2008.

[15] M. Cavusoglu, M. Kamasak, O. Erogul, T. Ciloglu, Y. Serinagaoglu, and T. Akcam, "An efficient method for snore/nonsnore classification of sleep sounds," *Physiological Measurement*, vol. 28, no. 8, pp. 841–853, 2007.

[16] Ali Azarbarzin and Zahra M. K. Moussavi, "Automatic and unsupervised snore sound extraction from respiratory sound signals.," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 5, pp. 1156–1162, May 2011.

[17] A Malhotra, M Younes, ST Kuna, R Benca, CA Kushida, J Walsh, A Hanlon, B Staley, AI Pack, and GW Pien, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring," *SLEEP*, vol. 36, no. 4, pp. 573–582, 2013.

[18] R. Nisha Aurora, Rachel Swartz, and Naresh M. Punjabi, "Misclassification of OSA severity with automated scoring of home sleep recordings," *Chest*, vol. 147, no. 3, pp. 719–727, 2015.

[19] Douglas B. Kirsch, "Pro: Sliding into home: portable sleep testing is effective for diagnosis of obstructive sleep apnea," *Journal of Clinical Sleep Medicine*, vol. 9, no. 1, pp. 5–7, 2013.

[20] UJ Magalang, NH Chen, PA Cistulli, AC Fedson, T Gislason, D Hillman, T Penzel, R Tamisier, S Tufik, G Phillips, and AI Pack, "Agreement in the scoring of respiratory events and sleep among international sleep centers," *SLEEP*, vol. 36, no. 4, pp. 591–596, 2013.

[21] ST Kuna, R Benca, CA Kushida, J Walsh, M Younes, B Staley, A Hanlon, AI Pack, GW Pien, and A Malhotra, "Agreement in computer-assisted manual scoring of polysomnograms across sleep centers," *SLEEP*, vol. 36, no. 4, pp. 583–589, 2013.

[22] Diego Alvarez-Estevez and Vicente Moret-Bonillo, "Computer-Assisted Diagnosis of the Sleep Apnea-Hypopnea Syndrome: A Review," *Sleep Disorders*, vol. 2015, 2015.

[23] Brian R. Snider and Alexander Kain, "Automatic classification of breathing sounds during sleep," in *Proceedings of The 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 699–703, IEEE.

[24] Conrad Iber, Sonia Ancoli-Israel, Andrew Chesson, and Stuart F. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, American Academy of Sleep Medicine, 2007.