DETECTION OF OBSTRUCTIVE SLEEP APNEA USING SPEECH SIGNAL ANALYSIS

O. Elisha[1], A. Tarasiuk [2], Y. Zigel [1]
[1]Department of Biomedical Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2] Sleep-Wake Disorders Unit, Soroka University Medical Center and Department of Physiology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel
orenelis@bgu.ac.il, tarasiuk@bgu.ac.il, yaniv@bgu.ac.il

***Abstract*: Obstructive sleep apnea (OSA) is a prevalent sleep related breathing disorder associated with several anatomical abnormalities of the upper airway. Acoustic parameters of human speech are influenced by properties of the vocal tract, which includes the upper airway. We hypothesize that it is possible to differentiate OSA patients from non-OSA (healthy) subjects by analyzing potential patients' speech signals. Using speaker recognition and signal processing techniques, we designed a system for classifying a given speech signal into one of the two groups. The database for this research was constructed from 92 subjects who were recorded reading a one-minute speech protocol immediately prior to a full polysomnography study; one hundred and three acoustic features were extracted from each signal; seven independent Gaussian mixture models (GMM)-based classifiers were implemented; a fusion process was designed to combine the scores of these classifiers and a validation procedure took place in order to examine the system's performance. Specificity and sensitivity of 91.66% and 91.66% were achieved for the male population; and 88.89% and 85.71% were achieved for female population, respectively. Such a system can be used as a tool for initial screening of potential OSA patients.**

***Keywords*: obstructive sleep apnea, speech signal processing, speaker recognition.**

I. INTRODUCTION

Obstructive sleep apnea (OSA) is a sleep disorder that is caused by obstruction of the upper airway. OSA severity is defined by the number of obstructive apnea and hypopnea events per hour of sleep (apnea hypopnea index – AHI). OSA affects approximately 5% of adults in the western population; a 2- to 3-fold greater risk for men compared to women has been reported [1]. OSA can lead to numerous complications such as hypertension, cardiovascular disorders, and excessive daytime sleepiness [2]. Currently, diagnosis of OSA is conducted in a sleep laboratory where a full polysomnography (PSG) study is performed. PSG is expensive, time consuming, and uncomfortable for the patient.

In earlier studies, researchers found that OSA is associated with several anatomical abnormalities of the upper airway that are unique to this disorder [3]. Acoustic parameters of human speech are affected by the physiological properties of the vocal tract (which includes the upper airway) such as vocal tract structure and soft tissue characteristics. Therefore, it was suggested [4] that acoustic speech parameters of an OSA patient may differ from those of a non-OSA subject (speaker). Our hypothesis is that speech signal properties of OSA patients will be different than those of control (non-OSA) subjects, and that we are able to distinguish between the two groups using a computer-based system that will analyze the subject's voice. The influence of OSA on speech is not yet fully understood but some researchers have tried to classify OSA subjects using speech signals [5] [6]; in both studies one classifier was trained on all speech segments using various acoustic features.

In this study we designed a system that fuses several Gaussian mixture model (GMM)-based classifiers, one for each of the voiced phonemes, using different acoustic features and model parameters. Our primary goal is to use this set of classifiers for initial screening of potential OSA patients that will assist in reducing the number of patients referred to sleep clinics for diagnosis. Our secondary goal is to improve our understanding of the effect of the disorder on speech including investigating the hyper-nasalization degree of the speech signals.

II. METHODS

A. Experiment setup

The test population of this research was constructed from 60 male subjects and 32 female subjects; subjects' age, AHI, and body mass index (BMI) are presented in Table 1. All subjects are patients who were referred to a sleep clinic by different doctors as "potential" OSA patients. All subjects underwent full PSG examination, were diagnosed, and given an AHI by the clinic's medical staff. Each subject was recorded using a digital audio recorder (Handy recorder "H4" by ZOOM) reading a one-minute text protocol in Hebrew, designed by the researchers to emphasize certain elements of speech. In

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

order to avoid over-fitting, the speech data was then divided into two separate databases: design and verification (validation).

Table 1 – The subjects' information

Diagnosis	*Number of subjects*	*AHI average ± Std*	*Age average ± Std*	*BMI average ± Std*
Male				
Healthy	12	4.83 ± 1.79	45.55 ± 13.6	27.66 ± 4.07
OSA	48	28.26 ± 20.17	56.58 ± 13.18	31.2 ± 5.7
Female				
Healthy	14	3.44 ± 2.28	47.23 ± 13.87	28.35 ± 6.73
OSA	18	24.44 ± 17.33	58.65 ± 10.59	33.44 ± 6.07

B. Pre-processing and feature extraction

Each recorded speech signal underwent a pre-processing procedure of down-sampling (to 16 kHz), DC removal, pre-emphasizing, and normalization; followed by manual segmentation of the signal in order to isolate specific phonemes. Using the signals from the vowels (/a/, /e/, /i/, /o/, and /u/) and nasal phonemes (/n/, /m/) alone, the signals were further framed into 30 msec frames. One hundred and three different acoustic features were extracted from each frame. The extracted features can be divided into four groups: time domain features, such as energy, pitch, jitter, and shimmer; spectral features, such as linear predictive coding coefficients (LPC) and their first and second derivatives, formant location and bandwidth, auto regressive moving average (ARMA) coefficients, and other potentially relevant spectral features; cepstrum domain features such as mel-frequency cepstral coefficients (MFCC) and their derivatives; and features for detection of hyper-nasal speech, which will be further elaborated later.

In addition to these "short term features" that were extracted from each frame, another set of features was computed as statistics of some of the short-term features through the entire speech signal, such as average of harmonic to noise ratio and average distance between formants. These "long-term features" represent the stationary position of the vocal tract uttering different vowels [5].

C. Abnormal nasalization degree detection

In [7], the researchers suggested that OSA patients demonstrate an abnormal nasalization degree in their speech. This abnormality is usually caused by a defective velopharyngeal mechanism [8] that may be associated with OSA. Hyper-nasal speech is characterized by amplitude reduction of the first formant, presence of zeros in the spectrum due to coupling of nasal and oral cavities, presence of reinforced harmonics resulting from the sound resonance in the nasal cavity, and a shift of formants [9]. In order to differentiate OSA patients from non-OSA subjects we added three features to the feature extraction process for estimation of hyper-nasalization degree of each given frame.

The first feature is based on a nonlinear operator called Teager energy operator (TEO) [8].

$$\psi\{s[n]\} = s^2[n] - s[n-1]s[n+1] \quad (1)$$

where $s[n]$ is the speech signal in time domain. The TEO can be shown to be sensitive to multi-component signals (such as hyper-nasal speech signals). The extraction of this feature was implemented as follows: each signal was filtered once with a BPF around the first formant and once with LPF, which was set to remove the frequencies that are higher than those of the first formant. TEO was extracted from both signals, and cross correlation between the two outputs was calculated. The assumption is that if there is only one component in the signal (no nasal harmonic near the first formant) the signals will be similar, but in the case of hyper-nasalized speech, the signals will be different.

The second feature proposed is based on using high and low order LPC [9]; where in case of hyper-nasal speech, there will be a large difference between the spectra obtained from these two sets of coefficients. The distance between the LPC sets was calculated by calculating the real LP cepstrum $c(k)$ and finding the geometric distance between the two sets using (2).

$$d = \sum_{k=0}^{\infty}[c_H(k) - c_L(k)]^2 \quad (2)$$

where $c_H(k)$ and $c_L(k)$ are high and low order LP cepstral sequences, respectively.

The third feature is set to detect the spectral flattening associated with hyper-nasalization of a given speech signal. Power spectral density (PSD) was estimated for each frame using Welch's method and standard deviation (STD) was calculated on the PSD between 300 Hz to 2000 Hz [10].

These features were added to previously described features that discriminate between normal and abnormal nasalization degree of speech, such as first formant location and bandwidth, distance between first and second formants, and ARMA coefficient.

D. Feature selection and model estimation

Seven GMM-based classifiers were implemented; one for each of the five vowels, one for the nasal phonemes, and one for "long-term features". Each phoneme-based classifier was trained separately on a different subset of features selected via a sequential forward floating

selection algorithm (SFFS). The most discriminative features for each model were chosen to maximize the performance of the classifier. After designing all seven phoneme-based classifiers and calculating the parameters for an OSA model and a healthy model for each classifier, each subject (of the design data) was tested over all models and scored using log-likelihood ratio and Z normalization [11], getting 7 normalized scores $\Lambda_i(\mathbf{x})$ $(i = 1, \dots, 7)$ – one for each classifier:

$$\Lambda_i(\mathbf{x}) = \frac{\frac{1}{N}\sum_{j=1}^{N} log\left(p(\mathbf{x}_j|\omega_{Hi})\right) - \frac{1}{N}\sum_{j=1}^{N} log\left(p(\mathbf{x}_j|\omega_{Oi})\right) - \mu_O}{\sigma_O} \quad i = 1, \dots, 7 \tag{3}$$

where $p(\mathbf{x}_j|\omega_{Hi})$ and $p(\mathbf{x}_j|\omega_{Oi})$ are the likelihood probabilities of the *j*th feature vector $\mathbf{x}_j$, given the model for healthy subjects and for OSA patients, respectively. μ_O and σ_O are the OSA population's mean and variance, respectively, and *N* is the number of frames.

The significance of each classifier was evaluated by conducting a leave one out (LOO) validation procedure on the design data. A fusion process was performed in order to combine all scores; the fusion process was found on issuing different weight, w_i $(i = 1, \dots, 7)$, to each score based on the significance of the classifiers' results. Classifiers that resulted in total significance of 60% or less were taken out of the final score and the remaining scores were weighted in proportion to their significances; the total of all weights is set to be 1. During the training phase, a threshold was calculated for all classifiers.

E. Validation

A validation procedure took place using the validation data; each subject was tested in a leave one out process, scores were given to the subject for each model, and summed using the previously calculated weight function:

$$\Lambda^w(\mathbf{x}) = \sum_{i=1}^{7} w_i \Lambda_i(\mathbf{x}) \tag{4}$$

The weighted score and the previously calculated threshold were used to decide whether to label each subject as OSA or non-OSA (healthy).

III. Results

Using the design database, the feature selection procedure resulted in a different set of selected features for each classifier; moreover, a different order of GMM was proven more efficient for each different phoneme.

In a recent study conducted in our lab [5], an identical database was used to achieve the same purpose of differentiating OSA from non-OSA (healthy) patients, using a **single** GMM classifier (baseline system, C) for all speech frames; an 8th order GMM model was implemented on a 5-dimension feature space for males and 4th order GMM model was implemented on a different 5-dimension feature space for females. The features in baseline system (C) were selected using the same SFFS procedure and out of the same 100 features described previously in section II, but without the "hyper-nasal" features. In order to evaluate the efficiency of our method of training 7 phoneme-based classifiers separately, and the effect each phoneme has on the final score, we examined our system using the same 5 features selected in [5] for each of the seven classifiers (system B). The results of all 3 systems are presented in Table 2.

Adding the three hyper-nasal speech detection features to the model further improved our result. System A was retrained using all 103 features; results are presented in Table 3.

Table 3 – Results for system A with hyper nasal detection features

Male		
	classified as O	classified as H
true label O	**91.66%**	8.33%
true label H	8.33%	**91.66%**
Female		
	classified as O	classified as H
true label O	**88.89%**	11.22%
true label H	14.29%	**85.71%**

The results of each phoneme-based classifier were fused with the weight function calculated with the design data; this function is presented in Table 4.

Table 2 – Results of 3 different systems (O-OSA, H-healthy)

	System A: 7 phoneme-based classifiers, separate feature selection procedure		**System B:** 7 classifiers, same features		**System C** (baseline system): 1 classifier for all phonemes	
	Male					
	classified as O	classified as H	classified as O	classified as H	classified as O	classified as H
true label O	**85.42%**	14.58%	83%	17%	83%	17%
true label H	16.66%	**83.33%**	33.33%	66.66%	21%	79%
Female						
	classified as O	classified as H	classified as O	classified as H	classified as O	classified as H
true label O	**83.33%**	16.66%	77.77%	22.23%	86%	14%
true label H	14.29%	**85.71%**	21.43%	78.57%	16%	84%

Table 4 – Weight function for each gender

	/a/	/e/	/i/	/o/	/u/	/m/+/n/	Long term
Male	0.16	0.05	0.00	0.00	0.16	0.11	0.52
Female	0.02	0.00	0.00	0.28	0.00	0.70	0.00

IV. Discussion and Conclusion

From Table 2 one can see that the proposed system (A), which offers an optimal feature set for each phoneme and a fusion between phoneme-based classifiers, is superior to the other compared systems (B and C).

For comparison, the results presented in [5] (system C) are 83% specificity and 79% sensitivity (for males). Implementing the same optimal 5 features of system C on the phoneme-based system (B) caused performance degradation to 83% specificity and 66% sensitivity, implying that those five features are not the optimal features for each phoneme.

Adding the hyper-nasal speech detection features to the model further improved the results, increasing specificity and sensitivity to 91.66% and 91.66% for male subjects, and 88.89% and 85.71% for female subjects. These improvements imply a difference in the nasalization properties between OSA and non-OSA groups. In order to further examine this potential discriminating property we trained our system (system A) using only 7 features: 3 hyper nasal features and first and second formants' location and bandwidth. Classification results of 70.8% specificity and 75% sensitivity were achieved, reinforcing the assumption of hyper nasalization in OSA patients' speech.

The procedure of training different classifiers with different feature sets for each phoneme (system A) indeed improved the results; moreover, the weight function and the results of each model led us to conclude that some phonemes (such as /a/ and nasal phonemes) carry more distinguishing information than other phonemes between OSA subjects and healthy subjects.

From the results of this research, it appears that initial screening of potential OSA patients using speech signals is indeed possible.

V. References

[1] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of obstructive sleep apnea," *American Journal of Respiratory and Critical Care Medicine*, vol. 165, no. 9, pp. 1217-1239, 2002.

[2] N. J. Douglas, *Harrison's Principles of Internal Medicine*, 17th ed., New York: McGraw-Hill Medical, 2008.

[3] T. M. Davidson, "The great leap forward: the anatomic basis for the acquisition of speech and obstructive sleep apnea," *Sleep Medicine*, vol. 4, no. 3, pp. 185-194, 2003.

[4] T. M. Davidson and J. Sedgh, "The anatomic basis for the acquisition of speech and obstructive sleep apnea: evidence from cephalometric analysis supports the great leap forward hypothesis," *Sleep Medicine*, vol. 6, no. 6, pp. 497-505, 2005.

[5] E. Goldshtein, A. Tarasiuk, and Y. Zigel, "Automatic detection of obstructive sleep apnea using speech signals," *IEEE Trans. on Biomedical Eng*., Vol. 58, No. 5, pp. 1373-82, 2011.

[6] R. F. Pozo, J. L. B. Murillo, L. H. Gómez, E. L. Gonzalo, J. A. Ramírez, and D. T. Toledano, "Assessment of severe apnea through voice analysis, automatic speech, and speaker recognition techniques," *EURASIP Journal on Advances in Signal Processing*, doi:10.1155/2009/982531, 2009.

[7] A.W. Fox, P.K. Monoson and C.D. Morgan, "Speech dysfunction of obstructive sleep apnea. a discriminant analysis of its descriptors", *Chest*, vol. 96 no. 3 pp. 589-595, September 1989.

[8] D.A. Cairns, J.H.L. Hansen, J.F. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear Teager energy operator," *Spoken Language,*. ICSLP 96. Proceedings., Fourth International Conference, vol.2, pp.780-783, Oct 1996.

[9] D.K. Rah, Y.I. Ko, and C. Lee, "A noninvasive estimation of hypernasality using a linear predictive model", *Annals of Biomedical Engineering*, Springer Netherlands, vol. 29, no. 7, pp. 587-594, 2001.

[10] T. Pruthi, and C. Y. Espy-wilson, "Acoustic parameters for the automatic detection of vowel nasalization," *in INTERSPEECH2007*, Antwer, Belgium, August 2007.

[11] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems", *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.