AUTOMATIC 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 sleep disorder associated with several anatomical abnormalities of the upper airway. Our hypothesis is that it is possible to distinguish between OSA and non-OSA subjects by analyzing particular speech signal properties using an automatic computerized system. The database for this research was constructed from 90 male subjects who were recorded reading a one-minute speech protocol immediately prior to a full polysomnography study; specific phonemes were isolated using closed group phoneme identification; seven independent Gaussian mixture models (GMM)-based classifiers were implemented for the task of OSA \ non-OSA classification; 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. Results of 91.66% specificity and 91.66% sensitivity were achieved using a leave one out procedure when the data was manually segmented. The system performances were somewhat decreased when the automatic segmentation was used, resulting in 83.33% specificity and 81.25% sensitivity.

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

I. INTRODUCTION

Obstructive sleep apnea (OSA) is a sleep disorder that is characterized by recurrent episodes of complete and/or partial cessations of breathing during sleep. These cessations of breathing (called apneas for complete cessations and hypopneas for partial cessations) are caused by the airway being sucked closed on inspiration while the patient sleeps. The primary defect that causes this disorder is a combination of two simultaneously existing conditions: anatomic factors that predispose the airway to collapse during inspiration and an insufficient neuromuscular compensation during sleep [1]. 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]. The severity of this syndrome is defined by the number of obstructive apnea and hypopnea events per hour of sleep (apnea hypopnea index – AHI). OSA can lead to numerous complications such as hypertension, cardiovascular disorders, and excessive daytime sleepiness [2]. Currently, OSA is usually diagnosed using overnight multi-channel polysomnography (PSG) recording which is considered the "gold standard" approach for the diagnosis of OSA. PSG is expensive, time consuming, and uncomfortable for the patient.

In [3], Davidson found that OSA is associated with several anatomical abnormalities of the upper airway that are unique to this disorder. It is well known that acoustic parameters of human speech are affected by the physiological properties of the vocal tract such as vocal tract structure and soft tissue characteristics. Therefore, it was suggested that acoustic speech parameters of an OSA patient may differ from those of a non-OSA subject [4].

Our hypothesis is that it is possible to distinguish between OSA and non-OSA subjects by analyzing particular speech signal properties using an automatic computerized system. In previous studies [5] [6], a similar goal was pursued; in both studies one classifier was trained on all speech segments using various acoustic features.

In this study we’ve designed a system that uses potential patients’ speech recordings to automatically diagnose OSA. The first stage of the diagnosing system is phoneme recognition; this stage automatically isolates the relevant segments (vowels and nasal phonemes) of the speech signal for the OSA classification stage that follows. In order to achieve this goal, an automatic phoneme recognition engine based on an eight class k-nearest neighbors (K-NN) classifier was developed. In the second stage (the OSA classification stage), we designed a system that fuses several Gaussian mixture model (GMM)-based classifiers, one for each of the voiced and nasal phonemes, using different acoustic features and model parameters for each classifier.

Our primary goal is to use this system for initial screening of potential OSA patients, thereby reducing the number of patients referred to sleep clinics for diagnosis.

II. METHODOLOGY

The proposed diagnostic system in this research was divided into two stages: automatic phoneme recognition and OSA – non OSA classification. A database, constructed for this research, was used for training and testing both stages of the system; separately and together as a whole system.
2.1 Data and experimental setup
The test population of this research was constructed from 87 male subjects who were referred to a sleep clinic by different doctors as “potential” OSA patients; all subjects underwent a full PSG examination, were diagnosed, and given an AHI by the clinic’s medical staff. Subjects’ age, AHI, and body mass index (BMI) are presented in table 1. Each subject was recorded using a digital audio recorder (Handy recorder “H4” by ZOOM) reading a one-minute text protocol in Hebrew, which had been designed by the researchers to emphasize certain elements of speech. The experiment protocol included elongated utterance of vowels (/a/, /i/, /u/, /e/, /o/) to be used later for vocal tract length normalization (VTLN); specific long sentences containing the mentioned vowels; short yes or no questions and a list of individual words. In order to avoid over-fitting, the data was divided into two separate databases: design and verification (validation).

2.2 Stage 1 - Automatic phoneme recognition
The main goal of this research is to diagnose OSA by analyzing patients’ speech signals – specifically, vowels and nasal phonemes. In order to do so, these parts of the speech signal must be isolated and identified. A simple context independent phoneme identification engine was designed for this purpose.

 usually, automatic phoneme recognition is the front end step of an automatic speech recognition system. Speech recognition based on phonemes may be preferable since it is free from vocabulary limitations. In our research we have no need for recognizing the context or the whole word; we only analyze the vowels and nasal phonemes regardless of the phoneme or word context. We chose to use a closed group phoneme identification approach. In order to train and test our phoneme recognition system, all speech signals were manually segmented to the following categories: 5 vowels (/a/, /i/, /u/, /e/, /o/), nasal phonemes (/m/, /n/) and all other segments of the signal (unvoiced and silence) as one category (total of 7 classes). A block diagram of the phoneme recognition engine (system) is presented in figure 1.

![Diagram of the phoneme recognition engine](image)

**Figure 1 - Block diagram of the phoneme recognition engine**

### Table 1 – The subjects' information

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Number of subjects</th>
<th>AHI average ± STD</th>
<th>Age average ± STD</th>
<th>BMI average ± STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-OSA</td>
<td>17</td>
<td>5.78 ± 2.58</td>
<td>46.23 ± 13.7</td>
<td>30.35 ± 5.24</td>
</tr>
<tr>
<td>OSA</td>
<td>70</td>
<td>25.74 ± 14.95</td>
<td>56.87 ± 13.04</td>
<td>30.35 ± 5.24</td>
</tr>
</tbody>
</table>

2.2.1 Training phase

**Vocal tract length normalization (VTLN):** Differences in vocal tract size among individual speakers contribute to the variability of speech waveforms [7]. Scaling of the frequency axis is the main effect of these differences.

This proposed system normalizes the frequency axis by estimating the scale factor for each speaker and resampling the spectrum in that factor.

The estimation of the scale factor ($\alpha$) was done as follows: each speech signal begins with elongated utterance of the 5 vowels (2 seconds utterance for each vowel) in a known order of appearance; the algorithm detects these phonemes and extracts the average frequency of the second formant from each phoneme, $F_i$ ($i$ is the phoneme index), which is then divided by the common frequency of this formant found in literature, $F_i^*$. These values were used to calculate the scale factor as shown in equation (1).

$$\alpha = \frac{1}{\sum_{i=1}^{5} F_i^*} \sum_{i=1}^{5} \frac{F_i}{F_i^*}$$

**Pre-processing and framing:** Each recorded speech signal underwent a pre-processing procedure of down-sampling (from 44.1 kHz to 16 kHz), DC removal, pre-emphasizing, and normalization; the signals were then framed into 30 msec frames with 50% overlap.

**Silence removal:** Silence was removed from the speech signal using a voice activity detector (VAD) based on [8].

**Feature extraction:** One feature vector containing 16 mel frequency cepstral coefficients (MFCC) was extracted from each normalized frame. The MFCCs are features commonly used for speech phoneme recognition tasks.

**Phoneme identification:** In previous studies [9][10] different classification methods were used for the recognition of speech and phonemes. The relatively small database of our study encouraged us to seek a simple classification method that will provide dependable results when trained with the given data. Two options were considered and examined: a parametric classifier - Gaussian mixture model (GMM) and a non-parametric classifier – the k-nearest neighbor (K-NN). The non-parametric approach produced more accurate results, therefore it will be discussed in this paper.

When a non-parametric approach is used, the density estimation is based on the genuine characteristics of the data, generally according to the information that is derived from the close neighborhood of the query point itself [9].

The k-NN algorithm is a method for classifying test observations based on the proximity to training observations in the feature space; the test observation is assigned to the class most common among its k nearest neighbors (majority vote). For the K-NN classifier no model training is required, since for each tested point we simply calculate the distance from all points in the feature space.
2.2.2 Testing phase

The tested speech signals were processed using the same steps as the training data. The classifier in the final training step was used for classifying each frame from the speech signal to one of the seven classes independently to the adjacent frames' identified labels. 

**Median filter:** Since we were only interested in phonemes that are longer than 60 msec (for quality purpose), the system filtered the label vector with a 7th order median filter eliminating all short term changes in the label vector, thereby smoothing the label vector.

The phoneme recognition stage produced a pre-processed framed speech signal, with a label vector labeling each frame as one of the vowels, nasal phoneme or unvoiced and silence.

2.3 Stage 2 – OSA / non-OSA classification

The goal of this stage is to distinguish between speech signals that were recorded by OSA patients and those recorded by non-OSA subjects.

**Pre-processing and feature extraction:** Each recorded speech signal underwent pre-processing and framing procedures in the previous stage (stage 1). One hundred and three different acoustic features were extracted from each frame. The extracted features can be divided into four groups: time domain features, spectral features, cepstrum domain features and hyper-nasal speech related features [11]. 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. These “long-term features” represent the stationary position of the vocal tract uttering different vowels [5].

**Feature selection and model estimation:** Seven GMM-based classifiers were implemented in this stage; one for each of the five vowels, one for the nasal phonemes, and one for “long-term features”.

The GMM is a well-known classifier that is defined as a parametric probability density function as represented as a weighted sum of Gaussian component densities [12].

\[
p(x|\omega) = \sum_{i=1}^{M} c_i g(x|\mu_i, \Sigma_i) \tag{2}
\]

Where \( \omega \) is a given class, \( x \) is a \( d \)-dimensional data vector (feature vector), \( c_i \) is the weight of the \( i \)-th Gaussian and \( g \) is the component Gaussian densities of the form:

\[
g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right) \tag{3}
\]

with mean vector \( \mu_i \) and covariance matrix \( \Sigma_i \).

Each phoneme-based classifier was trained separately on a different subset of features selected via a sequential forward floating selection algorithm (SFFS). 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 [13], getting 7 normalized scores \( \Lambda_i(x) (i = 1, \ldots, 7) \) – one for each classifier:

\[
\Lambda_i(x) = \frac{\frac{1}{2} \sum_{j=1}^{N_i} \log p(x_j|\omega_{HI}) - \frac{1}{2} \sum_{j=1}^{N_i} \log p(x_j|\omega_{LO}) - \mu_o}{\sigma_o} \tag{4}
\]

where \( p(x_j|\omega_{HI}) \) and \( p(x_j|\omega_{LO}) \) are the likelihood probabilities of the \( j \)-th feature vector \( x_j \), given the model for healthy subjects and for OSA patients, respectively. \( \mu_o \) and \( \sigma_o \) are the OSA population’s mean and variance, respectively, and \( N \) is the number of frames.

A fusion process was performed in order to combine all scores; the fusion process was founded on issuing different weight, \( w_i \) \((i = 1, \ldots, 7)\), to each score based on the significance of the classifiers’ results. The significance of each classifier was evaluated by conducting a leave one out (LOO) validation procedure over the design data. The scores were weighted in proportion to their significances; the total of all weights is set to be 1. During the training phase, one OSA/non-OSA decision threshold was calculated for all classifiers.

**Validation:** The validation data was used for evaluating the performances of the OSA classification system. Each subject was tested in a LOO process and scores were given for each model. The scores were then summed up using the previously calculated weight function:

\[
\Lambda^w(x) = \sum_{i=1}^{7} w_i \Lambda_i(x) \tag{5}
\]

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 AND DISCUSSION**

System performances were evaluated using the validation data. Initially, the two stages (sub systems) were tested separately and then the entire system was tested as whole.

3.1 Phoneme recognition engine

Figure 2 displays an example of the output of the phoneme recognition engine using the k-NN classifier.

![Figure 2 - Phoneme recognition output](image)

The results may also be shown as a confusion matrix presenting the recognition percentage for each phoneme, as seen in table 2.
The results presented in Table 2 indicate that the majority of the frames were indeed classified correctly. The vowels that were misclassified as unvoiced were (automatically) taken out from the testing phase of the second stage (OSA/non-OSA classification); although this misclassification causes the loss of data, it does not cause a performance degradation as much as misclassification of unvoiced as vowels.

### 3.2 OSA \ non-OSA Classification

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. The fusion procedure resulted with each of the 7 classifiers given a different relative weight.

In a previous study [11], the same database was used for diagnosing OSA while segmenting the signal manually (system A). Table 3 presents the results of system A compare to the results obtained in this current study (system B).

<table>
<thead>
<tr>
<th>True label Classification</th>
<th>/a/</th>
<th>/e/</th>
<th>/i/</th>
<th>/o/</th>
<th>/u/</th>
<th>/m/+/n/</th>
<th>Unvoiced and silence</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>70</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>/e/</td>
<td>7</td>
<td>73</td>
<td>15</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>/i/</td>
<td>1</td>
<td>8</td>
<td>67</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>/o/</td>
<td>6</td>
<td>1</td>
<td>72</td>
<td>13</td>
<td>4</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>/u/</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>61</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>/m/+/n/</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>8</td>
<td>71</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2 - Confusion matrix for k-NN classifier (in percentage)

The procedure of training different classifiers with different feature sets for each phoneme (system A) indeed produced satisfying 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 others.

When combining both stages of the system together to establish a fully automatic system, the performances were degraded from 91.66% specificity and sensitivity to 83.33% specificity and 81.25% sensitivity (as can be seen in table 3).

### V. SUMMERY AND CONCLUSIONS

In this study a fully automatic system for diagnosing OSA using speech signals was proposed. This system can assist the screening of potential OSA patients before sending them to sleep clinics.

### VI. REFERENCES


