Acoustic Breath Detection and Classification Modeling Respiratory Events

Bryce E. Hill, University of Utah Anesthesia Bioengineering Laboratory

***Abstract*— A continuation of research into modeling airway events of patients undergoing sedation is described. Sounds recorded at the trachea were recorded and separated by means of a threshold algorithm. The threshold was determined by the expectation maximization algorithm on filtered data. A comparison between the respiratory rate of the threshold algorithm and that of the direct airflow measure is done. Classification of the audio airway events is discussed using both Neural Networks and Polynomial Classifiers. Future work will be discussed**

I. Introduction

There are a number of ways to determine if a patient undergoing sedation is doing all right. One of the simplest, yet rarely used vital signs is simply listening. Long before the pulse-oximeter physicians and nurses would listen to their patients and not just their heart. Typically an anesthesiologist would tether himself to a patient by a precordial stethoscope and listen to the airway during every procedure. This is obviously tedious and can sometimes distract from more important observations. It has also become obsolete due to the vast number of monitors available today. Despite the tedious nature of this practice it provides data that cannot be provided by any commercially available monitor. Listening also alerts the physician earlier to airway problems than most other monitors.

Patients who are sedated commonly suffer from two major breathing complications. The first, respiratory depression happens when the subject becomes so relaxed from the analgesic as to stop the diaphragm muscles from trying to breath. The second, airway obstruction, happens when the anesthetic relaxes the muscles around the airway enough to collapse the airway. When snoring begins and at total obstruction no air is passed at all but the diaphragm muscles continue to try to push air.

It is proposed to build an electronic stethoscope which would listen to the patient's airway and determine the state of the airway and their respiratory rate. Such a device would be simple in nature and may be quite cheaply manufactured. This kind of monitor can also be easily converted to help in sleep studies involving sleep apnea, physical stress tests, and coma situations.

Monitoring the airway autonomously can be difficult because of the high amount of variability in sounds produced at the trachea. The proposed design involves several steps which can be seen in Fig. 1. The two steps which will be discussed will be the breath detection algorithm and the event classification algorithm.

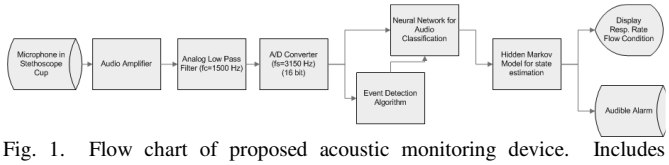

Fig. 1. Flow chart of proposed acoustic monitoring device. Includes hardware description and future designs.

Section II describes the data set and any preprocessing which has been done thus far. Section III describes the event segmentation problem and the current solution to finding the noise floor threshold values through the expectation maximization algorithm. Section IV describes the classification problem and the two possible solutions that need to be compared. Section V describes future work.

II. Data Set

Data was collected from 24 subjects for an IRB approved study. Each subject was sedated using a combination of remifentinal and propofol in incrementing dosages. During the sedation procedure a precordial stethoscope with a condenser microphone inside was placed on the trachea and audio data collected at 22050 Hz at 24 bit resolution through an electronic stethoscope cup placed on the trachea as shown in Fig. 2. Flow data was also recorded through a facemask with a pneumotachograph measuring flow(Cosmo +II respironics). This device measures the flow by differential pressure measurement. Chest and abdomen excursions were also measured using a respritrace device. Each subject was sedated three times for about two hours each providing a very large database for post-processing. Throughout the data collection many different audio events were captured such as vocalization, snoring, swallowing, pre-

Fig. 2. Precordial stethoscope with microphone attached by a double stick disk to the outside of the trachea.

obstruction sounds, and many other events not related to breathing.

Small portions of this data have been used which include events during snoring vocalization and normal breathing.

An example of the audio data with noise floor is shown in Fig. 3. The approximate threshold of the noise floor is shown as well.

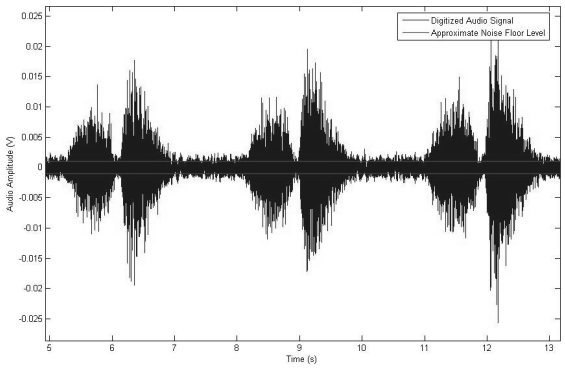

Fig. 3. Example of audio data with approximate noise floor. Shown are three breaths showing both inspiration and expiration.

The audio data was not time synchronized with the flow or chest excursion data which poses a problem for future comparison. Data in the sets used for this paper were visually time synchronized as the flow data relates very well to audio amplitude during normal breathing [1].

There are many characteristics in the audio which can be filtered off as they do not help in processing the breathing. In this case the heartbeat sound was removed spectrally as it has a frequency of 50 Hz and below. Strong heartbeats can be seen in higher bands as noise but this is only due to the microphone being shaken by the actual beat of the heart. The data can also be low pass filtered and decimated to remove noise and simplify the audio processing for this problem. In this case the stethoscope cup and tracheal tissue filter the audio at around 1300 Hz, thus the data was filtered digitally at 1300 Hz and the sample rate decimated by a factor of 7 to produce an audio rate of 3150 Hz. For the sake of comparison both audio sampled at 22050 and 3150 Hz will be used in different respects to provide a comparison and to determine if anything is lost in the filtering.

III. Event Segmentation

A tracheal audio event can be defined as any signal that rises above the noise floor and is sustained for at least 0.25 seconds. In general these sounds can be classified under four categories of clear breaths, vocalization, snoring, and events not related to breathing such as swallowing and external interference. It is important to segment each breath in order to classify it differently. Because the clear breaths are of such small amplitude in comparison with the noise floor and other audio events the noise floor cannot be arbitrarily defined.

The noise floor is caused by several factors. First there is electronic noise and shot noise in the electronic system which can be measured and quantified but may change over time due to EMI. The main source of the noise floor however is the ambient audible noise inherent in any audio recording. This noise is further filtered by the stethoscope cup and minimized in this way, but also added to by body noises such as heartbeat and circulatory sounds. These sounds change with every subject and placement of the stethoscope cup. It is not desired to eliminate the background noise as much as minimize it. It is also desired to be able to differentiate between the noise floor and tracheal events. In order to do this a threshold of the noise floor must be determined.

In order to determine the noise floor a few assumptions must be made. First it is important to filter off all spectrally non-stochastic signals such as the heartbeat. The heartbeat signal is typically in the frequency band below 50 Hz. This can be easily filtered off. After this has been removed it is assumed that the remaining signal is white Gaussian in nature. Statistically this can be determined by a QQ plot as shown in Fig. 4. A one to one relationship is desired to determine the Gaussian nature of the noise floor signal. In most cases after the heartbeat has been filtered the signal looks very Gaussian.

With the previous assumptions the Expectation Maximization algorithm [2] can be used to determine the standard deviation of the noise floor. The standard deviation can then be used as a threshold in determining if a signal is an event and

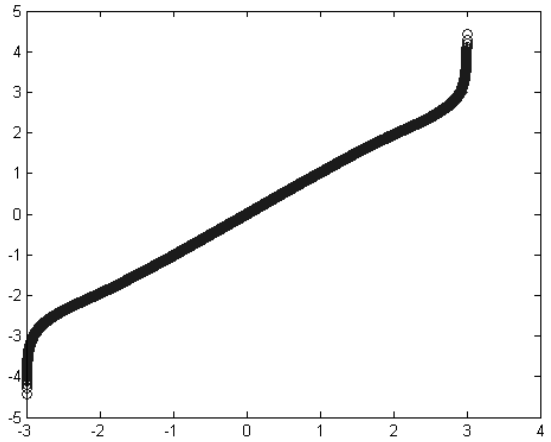

Fig. 4. QQ plot of noise floor after filtering of undesired components.

can then segment the audio. From further observations it was discovered that over longer periods of time (2 minutes or more) the audio histogram looks like a Gaussian and Laplace mixture algorithm as shown in Fig 5. The Gaussian signal seen in the center is the noise floor and has a probability density function (PDF) of $g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(\frac{|x|^2}{2\sigma^2})}$. The rest of this signal can be estimated as a Laplace distribution with PDF $f(x) = \frac{1}{2b} e^{-(\frac{|x|}{b})}$. Assuming this PDF for this rest of the signal is a broad assumption but can be done because of the predictability of the Gaussian distribution. Putting these equations together yields the equation $(x) = p \times g(x) + (1 - p) \times f(x)$ which is the mixture PDF model of this signal. There are three unknowns of σ, b, and p. The only value really that is needed is σ but in order to find it all three must be found by means of the Expectation Maximization algorithm.

In the Expectation Maximization algorithm an initial value of σ is estimated using the initial data.

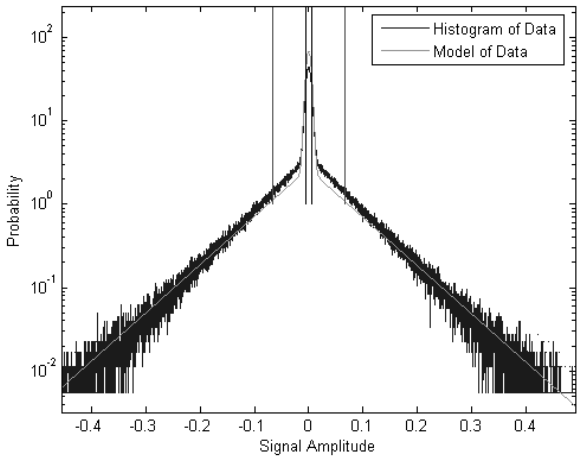

Fig. 5. Histogram of 10 minutes of audio data and model overlayed.

The input signal is then separated into groups that belong and do not belong to a Gaussian signal with standard deviation of σ. This is first done by assuming that no values of the input signal exist outside of three standard deviations. After this is done the values within three standard deviations are compared with a true Gaussian signal with standard deviation σ which is independent of the input signal. Those that fall within this comparison are kept and the standard deviation is taken of these remaining samples which replaces σ. This process is iterated until σ converges to the standard deviation of the noise floor. The number of samples which belong to the Gaussian signal divided by the total number of signals evaluates p and b can then be measured by taking the standard deviation of the samples that were not considered Gaussian. The model shown in Fig. 5 has been estimated using this algorithm and appears in this case to be a very good estimate of all of the parameters.

It may be that a Laplace PDF is not a good estimate for *f(x)*. This does not hurt this algorithm because σ is the real point of interest and it is only important that the standard deviation of the noise floor is less than the standard deviation of the detected signals.

In order to turn this threshold into something useful basic detection theory is used to determine the error rate of false alarms. The error rate of 10% was chosen which puts the absolute threshold to $1.2816 \times \sigma$. After this point some features of the events are needed such as the minimum length of an event and the maximum length of an event. The minimum length of 0.25 seconds was determined by observation of the lengths of breaths, snores and vocalizations. Using this simply states that if the signal does not exceed the threshold approximately 50% of the time (due to bipolar nature of audio signals) for 0.25 seconds it is not considered an event. Either it is too short or not loud enough. Otherwise this would be considered an event. There is no minimum to the time between events and currently no maximum set to the length an event can be. An example of the event markers can be seen in Fig. 6 where the audio envelope is scaled to the flow signal and time synchronized. It can bee seen that the audio markers are very close to the markers determined directly by the flow sensor. When comparing these to sets of data a strong correlation can be seen between that of the respiratory rate measured by the event algorithm and the respiratory rate measured directly as shown in Fig. 7.

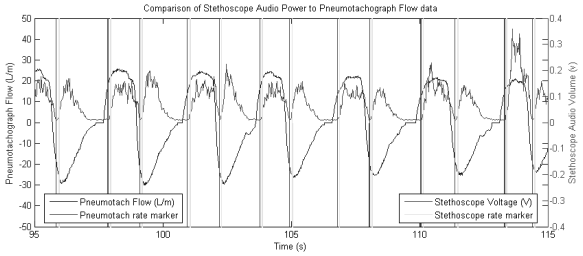

Fig. 6. Comparison of flow data overlayed with audio envelope with detected event markers for both. A strong correlation between all can be seen.

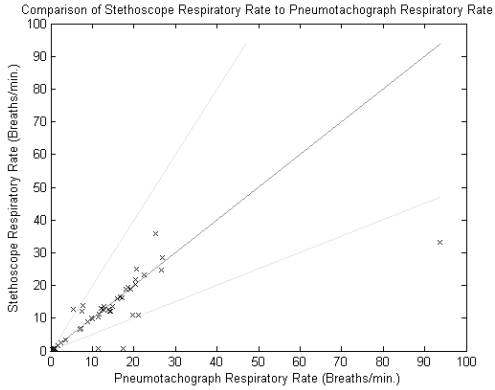

Fig. 7. Correlation of flow breath detection to audio breath detection. A strong positive correlation can be seen.

IV. CLASSIFICATION

The next step after segmentation is classification of tracheal events. There will only be four classes of events, namely clear breathing, snoring/pre-obstruction, vocalization, and events not related to breathing.

Clear breathing is a modulated white noise. For just one audio segment it appears to be Gaussian. It is louder than the noise floor and over long periods of time the modulation of the white noise makes it appear to have a Laplace distribution rather than a Gaussian. It's spectrum reaches up to about 1200 Hz and is very low in amplitude in comparison to other events.

Snoring/pre-obstruction is the hardest sound to define with features. It can look white or it can be harmonic in nature. It can be low or high amplitude. It is important to determine this signal due to its correlation to obstruction. At full obstruction no air passes the trachea and thus no sounds are produced, but just previous to that the airway is constricted and produces a large amount of noise. Because snoring and obstruction sounds typically only occur on inspiration the relative flow can be seen on the expiration of the breath.

Vocalization is harmonic, extremely loud and does not have a very predictable pattern at the trachea. It is important to determine vocalization patterns so that they do not false alarm one of the other classes.

Events not related to breathing can be anything including swallowing and disturbances at the stethoscope cup. This kind of signal is not predictable because of the number of different sources it can come. This also does not happen at any particular rate such as snoring or breathing which is an important feature.

In order to classify these events two methods have been considered. The first is a multi-layered perceptron Neural Network as described in [3]. The second is a polynomial classifier which is described in [4].

An artificial neural network is a structure which can have an arbitrary number of inputs and arbitrary number of outputs. The reason they are desirable is because of the ease of training them and the ability for complex neural networks to solve problems that are not a single decision boundary. Lippman etal. [3] explains in great detail the many abilities of several types of neural networks. In this case the multilayer perceptron was used because of its simple structure and complexity. In this case there was the input layer, the output layer and two hidden layers. Between each layer a sigmoid non-linearity of form $f(\alpha) = \frac{1}{1+e^{-(\alpha-\theta)}}$ [3].

During training backward error propagation is used which measures the eror at the output layer. The error is propagated through the weights to adjust the weights of the neural network. The error is scaled by a learning rate much like that of an adaptive filter. The only difference is the non-linearity and the multi-layered approach.

Because of the adaptive type algorithm present, the same set of data can be reused iteratively in order to better train the weights. The advantage of iteratively training the algorithm is that the network can learn from previous mistakes. The disadvantage of this method is that overtraining can occur which is when the network becomes specific to only the training data used. This creates problems for data which is slightly different from the training data.

The inputs into the neural network need special attention in order to reduce the number of iterations during training. In most cases it is important to scale the inputs by the mean and standard deviation of the training set for each input. The method would be to subtract the mean and divide by the standard deviation at each input. If the inputs are not scaled each weight has to be iteratively scaled and this scaling is done in increments of the learning rate which can be very small. It is more helpful if the input is scaled before training to expedite this process.

A polynomial classifier is a special type of polynomial filter. This classifier has an additional

polynomial filter for each class. Each filter is trained to have an average output of one for that class and zero for the average output of any other class. The features are fed into each filter and the one which is closest to one is the estimated class.

Similar to [4] the order of polynomial filter for this project was limited to 4 for all trials. The number of inputs varied from one set of features to another but just like the Neural Network, the same number of inputs are needed for each test signal in order for the weights to be matched to an input.

For each input vector $\boldsymbol{x} = [x_1\, x_2\, x_3\, \ldots x_N]$ a vector of polynomial values was constructed with $\boldsymbol{p(x)} = [1\, x_1\, x_2\, x_3\, \ldots\, x_N\, x_1^2\, x_1 \times x_2\, \ldots\, x_1 \times x_N \ldots x_1 \times x_2 \times \ldots x_N \ldots]^T$. It is easy to see with such an illustration that this can easily get out of hand. Thus both order and delays between multiplied inputs must be limited. Each coefficient in the vector ***p(x)*** must now have a filter coefficient which must be trained which will be called $\boldsymbol{w_{spk}}$. For each different sound to be identified a $\boldsymbol{w_{spk}}$ needs to be identified. In the end the input features are fed through each filter and the numerical average is taken as shown here.

$$s = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{w}_{spk}^T \boldsymbol{p}(\boldsymbol{x_i})$$

Training $\boldsymbol{w_{spk}}$ is also a simple process. In this case a matrix is made with each polynomial feature set $\boldsymbol{p(x)}$ as a row. There are as many rows as training data for that class of data so that the final matrix looks like:

$$M_{spk} = \begin{bmatrix} \boldsymbol{p(x_1)^T} \\ \ldots \\ \boldsymbol{p(x_N)^T} \end{bmatrix}$$

Similarly each class has its own "M" matrix but for a desired class every other "M" matrix will be considered an imposter data set or M_{imp}. A complete M matrix consists of

$$M = \begin{bmatrix} M_{spk} \\ M_{imp} \end{bmatrix}$$

where M_{imp} can be either the entire imposter training set or randomly selected vectors from several imposters. One other vector has to be defined which is $\boldsymbol{o}$. This is a vector made up of ones for the number of rows in the M_{spk} matrix and zeros for the number of rows in M_{imp}.

In the end the calculation of $\boldsymbol{w_{spk}}$ is simply solving for the end equation

$$\mathrm{M^T M}\, \boldsymbol{w_{spk}} = \mathrm{M^T}\boldsymbol{o}$$

Preliminary research in this area has shown that both the Polynomial Classifier and Neural Network perform relatively well against one another. With general features the Neural Network performed with an error rate of about 15% whereas the Polynomial filter performed at about 17%. The reason for the high error rate can be attributed to the limited training set used and the arbitrarily chosen feature sets. In the future features will have to be hand-picked in order to improve the difference between each class and make it easier for a signal to be classified.

V. Future Work

There are three major pieces of work to do before much progress can be made. The first is to time synchronize the audio data with that of the other data equipment over all the sets and segment the audio into shorter easier to handle sets. The synchronization requires a large amount of visual comparison over every data set.

The second major project is to have the flow, chest, and abdomen excursion data predict the state of the airway. This can be done by determining how hard the chest is working vs. how much flow is measured.

The final major project is to use the standard of the flow and chest data to train several classifiers and determine which to use and how which features work the best at distinguishing classes.

References

[1] V. Paul Harper, Hans Pasterkamp, Hiroshi Kiyokawa, George R. Wodicka,, "Modeling and Measurement of Flow effects on Tracheal Sounds" in IEEE Transaction on Biomedical Engineering, Vol 50, No 1, January 2003

[2] Moon,T.K., "The Expectation-Maximization Algorithm," IEEE Signal Processing Magazine, Nov 1996 Vol 13, Issue 6 pages 47-60

[3] Richard P. Lippman, "An Introduction to Computing with Neural Nets" in IEEE ASSP Magazine April 1987

[4] Campbell W M, Asselah K T, Broun C C, "Speaker Recgonition with Polynomial Classifiers" in IEEE Transactions on Speech and Audio Processing, Vol 10. NO. 4 May 2002, pp 205-212.