Hidden Markov Model Based Respiratory Sound Classification

N. Jakovljević and T. Lončar-Turukalo

Abstract

This paper presents a method based on hidden Markov models in combination with Gaussian mixture models for classification of respiratory sounds into normal, wheeze and crackle classes. Input features are mel-frequency cepstral coefficients extracted in the range between 50 Hz and 2000 Hz in combination with their first derivatives. The audio files are preprocessed to remove noise using spectral subtraction. Our best score achieved in the official ICHBI Challenge second evaluation phase is 39.56.

Keywords

Respiratory sounds • Crackles • Wheezes • Hidden Markov models • Spectral subtraction

Introduction

Auscultation is a common, fast and noninvasive way to diagnose patients with lung diseases. Respiratory sounds according to their acoustic properties can be classified into normal and abnormal [1, 2]. Frequency content of normal respiratory sounds depends on stethoscope position and does not contain tonal (musical) components [2]. For example, lung or vesicular sounds are dominated by frequencies below 100 Hz, whereas in the tracheal sounds frequencies from 100 to 1500 Hz are more distinctive. Abnormal sounds consist of both normal and adventitious respiratory sound. Adventitious crackle sounds are discontinuous, nontonal lung sounds with a duration of less than 20 ms [2]. They are normally heard during inspiration and sometimes during expiration [2]. The crackle sounds' frequency range is 60–2000 Hz, with their major contribution below 1200 Hz [2]. Wheezes are continuous tonal lung sounds with the dominant frequency above 400 Hz, and with a duration longer than 100 ms [2].

The most comprehensive evaluation of different classification algorithms over healthy and asthmatic respiratory sound databases is presented in [3]. The best performance in [3] is obtained by the model based on Gaussian mixture models (GMM) in combination with mel-frequency cepstral coefficients (MFCCs). For these reasons this model has been selected as the baseline model. The functionality of this model has been enriched with the information about the frame position in a sequence, leading to hidden Markov model (HMM) instead of GMM. As hidden Markov models were the backbone in automatic speech recognition for many years [4], theoretical foundations have been developed, and many practical considerations are well defined. A respiration cycle varies in duration and acoustical content, just as in speech, which suggests that HMM is an appropriate tool to model it.

Methods

Preprocessing

The dataset contains audio recordings sampled at 44.1 kHz and 4 kHz. Even though a majority of the recordings is

N. Jakovljević (✉) · T. Lončar-Turukalo
Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia
e-mail: jakovnik@uns.ac.rs

N. Maglaveras et al. (eds.), *Precision Medicine Powered by pHealth and Connected Health*, IFMBE Proceedings 66, https://doi.org/10.1007/978-981-10-7419-6_7

sampled at 44.1 kHz, downsampling to 4 kHz is performed as the frequency content of both wheeze and crackle is in the range of 60–2000 Hz [2]. An additional benefit is a significant reduction in computational complexity of feature extraction.

To remove sounds caused by heartbeats, the signal components at low frequencies have to be suppressed. We have evaluated the performance of two different filters. The first one is the low order bandpass filter with the transfer function:

$$H_1(z) = \frac{1 - z^{-2}}{1 - 0.9z^{-2}} \tag{1}$$

The additional benefit of this filter are the reduced effects of sudden changes in signal which can appear at the edges of clipped segments if only a high pass filter was applied.

The second filter is the high pass finite impulse response filter with cutoff frequency f_c = 100 Hz and constant group delay τ_g = 1024 samples obtained by Hann window function. In this way components at frequencies below 96 Hz are attenuated by at least 54 dB, i.e. heartbeat sound is suppressed more than in the case of the first filter.

Noise Suppression

Many sound files in the dataset contain stationary noise, thus the following step in this algorithm is noise suppression. The implemented noise suppression is based on spectral subtraction [5], which is performed on the signal which is segmented into 30 ms long frames shifted by 15 ms using Hann window function. For each frame discrete Fourier transform (DFT) is performed and each magnitude spectrum is decreased by the estimated noise magnitude spectrum, i.e.:

$$|X_d(k,t)| = |X(k,t)| - |D(k)| \tag{2}$$

where $|X(k,\ t)|$, $|D(k)|$ and $|X_d(k,\ t)|$ are the magnitude spectra of the original signal, the noise, and the denoised signal at time t respectively, where k denotes the frequency bin. The noise magnitude spectrum $|D(k)|$ is estimated as the mean value of $|X(k,\ t)|$ over 1% of the frames with minimum energy in the audio signal, excluding invalid frames with zero energy.

The problem of the negative values of $|X_d(k)|$ has been solved using two approaches. The first approach, referred to as SS1, sets the negative magnitude values to 1% of $|X(k,t)|$, i.e.:

$$|X_d(k,t)| = \begin{cases} |X(k,t)| - |D(k)| & |X(k,t)| > |D(k)| \\ 0.01 \cdot |X(k,t)| & \text{else} \end{cases} \tag{3}$$

The second approach, referred to as SS2, additionally reduces the musical noise level introduced by magnitude spectrum subtraction. The musical noise is caused by sudden drops of magnitude at a certain frequency bin in successive frames. Relying on the assumption that breath sound should be dominant in the signal, for each k the estimated noise level $|D(k)|$ has been iteratively reduced by 10%, until in at least 60% of frames $|X(k,\ t)| > |D(k)|$ is fulfilled. The denoised magnitude spectrum is obtained by:

$$|X_d(k,t)| = \begin{cases} |X(k,t)| - |D(k)| & |X(k,t)| > |D(k)| \\ |X(k,t)|^2 & \text{else} \end{cases} \tag{4}$$

where instead of linear scaling of critical components, quadrature scaling is introduced, further suppressing small magnitudes in $|X_d(k,\ t)|$. It should be noted that $|X(k,\ t)|$ has to be range normalized to accommodate quadrature scaling.

To suppress sudden drops of magnitude, $|X(k,\ t)|$ is monitored in 5 successive frames. If $|X(k,\ t)| < |D(k)|$ in at least 3 of 5 adjacent frames, the frequency bin is marked as noise. An entire frame is considered as corrupted by noise and set to zero ($|X_d(k,t)| = 0$, for each k) if more than 70% of the bins are marked as noise.

In the synthesis step, the phase spectrum is approximated with the phase spectrum of the noisy signal, thus the spectrum of denoised signal is:

$$X_d(k,t) = |X_d(k,t)|e^{j\ \arg\{X(k,t)\}} \tag{5}$$

and the reconstructed signal is the sum of overlapping segments obtained by inverse DFT of $X_d(k,t)$.

Feature Extraction

The MFCCs are estimated every 10 ms using 30 ms long windows. The frequency range [50, 2000 Hz] is divided into 16 equal-width overlapped channels in mel-frequency domain. The discrete cosine transform is performed on the logarithm of 16 energy coefficients calculated for each channel.

$$C_n = \sum_{k=1}^{16} \log(E(k)) \cos\left(\frac{n\pi}{16}\left(k - \frac{1}{2}\right)\right) \tag{6}$$

for n = 0, 1, … 15, where C_n is the nth MFCC and $E(k)$ is the energy at the kth channel. The coefficient C_0, which represents signal energy in the selected frequency band, is discarded from further steps, since in some signals it significantly correlates with heartbeat sound.

The cepstral mean and variance normalization per record is applied to remove variations caused by the remaining noise and it is defined by:

$$\bar{C}_n(t) = \frac{C_n(t) - \bar{C}_n}{S_n} \tag{7}$$

where:

$$\bar{C}_n = \frac{1}{T}\sum_{t=1}^{T} C_n(t) \tag{8}$$

$$S_n = \frac{1}{T}\sum_{t=1}^{T} (C_n(t) - \bar{C}_n)^2 \tag{9}$$

and T is the duration of signal in frames.

Additionally, to track feature dynamics and to decorrelate successive feature vectors, first time derivatives of MFCCs are introduced, increasing the cardinality of the feature vector to $d = 30$.

Modeling

By visual inspection we have found that the same sound class varies in acoustic content depending on recording location, thus a respiration cycle for each location (trachea, anterior left/right, lateral left/right, posterior left/right) and sound class (normal, crackle, wheeze, and both crackle and wheeze) is represented as a sequential HMM with S states (see Fig. 1).

An HMM is described by its initial state probabilities ($\mathbf{\Pi}$), state transition matrix ($\mathbf{A}$), and emitting probability density function for each state (b_s). A state emitting probability density function (pdf) for a given d-dimensional observation $\mathbf{o}$ is defined by:

$$b_s(\mathbf{o}) = \sum_{i=1}^{M} w_i \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{o}-\boldsymbol{\mu}_i)} \tag{10}$$

where w_i, $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ are weight, mean and covariance matrix of the i-th mixture component, respectively. Although each state can have a different number of mixture components, it is common that the number is the same for all states.

In case of sequential model only one state can be the first one, so in the vector $\mathbf{\Pi}$ only one value is equal to 1 and the others are 0, and each row in the state transition matrix $\mathbf{A}$ contains at most 2 nonzero elements.

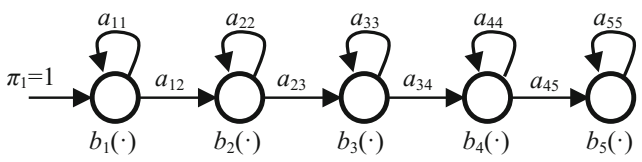

Fig. 1 Sequential HMM with $S = 5$ states

The standard criterion for HMM parameter estimation is the maximization of the likelihood that the models will generate the training sequence [4]. The optimization is usually performed using expectation maximization algorithm (Baum-Welch estimation). For an efficient estimation procedure, the initial values of model parameters should be carefully set. In this study, the initial parameters were obtained by the time equidistant partition of the observation sequence between states, and for each state the sample mean $\boldsymbol{\mu}_s$ and the covariance matrix $\mathbf{\Sigma}_s$ were calculated. In case of several mixture components per state, means ($\boldsymbol{\mu}_i$) were obtained by random sampling from normal distribution $N(\boldsymbol{\mu}_s, \mathbf{\Sigma}_s)$, and covariance matrices ($\mathbf{\Sigma}_i$) by assigning the corresponding sample covariance matrix ($\mathbf{\Sigma}_i = \mathbf{\Sigma}_s$). The initial transition probabilities (Fig. 1) were set to 0.5, with stay probability corresponding to the last HMM state, except for a_{SS}, which was initialized to 1.

The existing model parameters are used to calculate probabilities that the model will be in the state s at time t and will generate the observation ($\mathbf{o}_t$) using the m-th mixture component. These probabilities are used to update the values of the transition probabilities, means and covariance matrices of the model. In our experiments these parameters converged in 6–12 iterations.

During the test phase, an unknown observation sequence, denoted $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots \mathbf{o}_T]$, is aligned with all HMMs (λ_c), and the classification decision is based on the maximum likelihood criterion, i.e.

$$\hat{c} = \arg \max_{1 \le c \le C} p(\mathbf{O}|\lambda_c) \tag{11}$$

$$p(\mathbf{O}|\lambda_c) = \pi_1 b_1(\mathbf{o}_1) \sum_{s(2),\ldots,s(T)} a_{1s(2)} \prod_{t=2}^{T} b_{s(t)}(\mathbf{o}_t) a_{s(t)s(t+1)} \tag{12}$$

where $s(t)$ represents the state at time t, and C the number of classes. Having in mind computational complexity, the log probabilities are used instead of probabilities themselves.

Database

For training and evaluation, the official ICBHI Challenge respiratory sound database released in September 2017 was used [6]. The details on data acquisition and ethical considerations are provided [6]. The number of attempts for the official scoring was limited, therefore many of experiments were evaluated only on a validation set. The official training set was divided into 10 folds. The validation set in each fold contains at least one sound class for every possible recording

Table 1 Sensitivity (Se), specificity (Sp), and score evaluated on the validation set, and score on the official test for different preprocessing procedures (PP), the number of states (S), the number of mixture components per state (M), and covariance matrix type (CMT)

PP	S	M	CMT	Validation set			Official
				Se	Sp	Score	Score
T2	1	4	full	0.4381	0.4533	44.57	n/a
T2	1	8	full	0.4252	0.5136	46.94	n/a
T2	1	16	full	0.3517	0.6115	48.16	n/a
T2	1	32	full	0.2089	0.7671	48.80	n/a
T2	1	64	full	0.0917	0.8702	48.09	n/a
T1	5	1	full	0.4093	0.5326	47.09	39.32
T1	5	1	diag.	0.4079	0.4091	40.85	39.02
T1	6	1	full	0.4232	0.5669	49.50	39.37
T2	6	1	full	0.4102	0.5267	46.85	36.98
Class. ensemble				n/a	n/a	n/a	39.56

location. All respiratory cycle instances from an audio file were in the same (train/validation) set.

Evaluation Criterion

The performances of classifiers were evaluated using officially proposed scores [7] i.e. sensitivity (*Se*), specificity (*Sp*), and overall score, compactly written as:

$$Se = \frac{C_c + C_w + C_b}{T_c + T_w + T_b},\ Sp = \frac{C_n}{T_n},\ Score = \frac{Se + Sp}{2} 100\% \tag{13}$$

where C_i and T_i ($i = c,\ w,\ b$) are the number of correctly recognized instances of class i, and the total number of instances of class i in the test (or validation) set, respectively. Indices c, w, b, and n stand for classes: crackle, wheeze, both crackle and wheeze, and normal, respectively.

Results and Discussion

The selected results are summarized in Table 1. The classifiers differ by the preprocessing procedure, the number of states and mixture components per state and the type of the covariance matrix. In the first preprocessing procedure (T1), proposed in the first phase of ICBHI Challenge, the input signal is filtered through the bandpass filter $H_1(z)$ and noise suppression is based on the SS1 method. The second preprocessing procedure (T2) includes downsampling to 4 kHz, filtering by the high pass FIR filter and noise suppression based on SS2. It should be noted that the features are extracted in the frequency range [50, 2000 Hz] independently of the preprocessing procedure. Our initial experiments for the simpler models on reduced dataset have shown that there is no significant difference between these preprocessing procedures, but a difference has been noted on the extended dataset (see last two rows in Table 1).

The baseline system based on GMM has shown slightly inferior performance to the HMM based systems. It can be noted that with the increasing number of mixture components the overall score is improving, as the result of higher specificity. However, sensitivity is decreasing, indicating that the classifier could not resolve adventitious sound types.

Introducing HMM, i.e. taking into consideration the position of the frame in a sequence, increases the accuracy of the model without a significant increase of its complexity.

As the used features are correlated, modeling data with full covariance matrix increases the overall score by increasing the specificity, without degradation in sensitivity (Table 1, rows 6 and 7). The difference of the scores obtained on the validation set (6.24) is higher than the difference of the official test set scores (0.30).

The overall discrepancies of scores obtained in cross-validation using the publicly available dataset, and the official test set (Table 1, columns 7 and 8) are noticeable. One plausible reason for the score discrepancies might be the correlation of the recordings in the publicly available dataset (recordings from the same subject might be present in both training and validation set), whereas the test set strictly comprises a disjunct set of subjects [7].

To increase the overall score, we have tried with an ensemble of classifiers trained over the 10 different folds. All classifiers which had the same model complexity (28 models with 5 states and 1 Gaussian per state) were trained with a single learning method. The final decision was made by simple majority voting by the classifiers. This approach has achieved our best official score of 39.56, that represents a minor increase in the score (0.24) at the expense of 10 times greater computational complexity.

The presented results are modest in comparison with the results published in [1, 3, 8], where both less extensive databases and a smaller number of the adventitious sound classes are used. There are several challenging issues regarding the database used in this study: different types of noise, multiple recording locations, and small numbers of samples for different classes.

Conclusions

This study shows that MFCCs in combination with HMM can be used for classification or respiratory sounds into 4 categories: normal, crackle, wheeze, and both crackle and wheeze. The performances of the examined classifiers are modest because they were evaluated on real data under varying levels of different types of real noise. We assume that advanced noise suppression techniques can improve the overall score.

Acknowledgements This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, TR 32035 and TR 32040. We acknowledge the support of the COST Action ENJECT TD1405 in the form of ITC grant awarded to the first author.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Reichert S, Gass R, Brandt C, Andres E (2008) Analysis of respiratory sounds: state of the art. Clin Med Circ Respirat Pulm Med 2:45–58
2. Sarkar M, Madabhavi I, Niranjan N, Dogra M (2015) Auscultation of the respiratory system. Ann Thorac Med 10(3):158–168. https://doi.org/10.4103/1817-1737.160831
3. Bahoura M (2009) Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes. Comp Bio Med 39(9):824–843
4. Gales M, Young S (2008) The application of hidden Markov models in speech recognition. Found. Trends Signal Process 1 (3):195–304. https://doi.org/10.1561/2000000004
5. Berouti M, Schwartz M, Makhoul J (1979) Enhancement of speech corrupted by acoustic noise. In: Proceedings of. IEEE international conference on acoustics, speech and signal processing, pp 208–211. https://doi.org/10.1109/ICASSP.1979.1170788
6. Rocha BM, Filos D, Mendes L et al (2017) A respiratory sound database for the development of automated classification. In: Proceedings of international conference on biomedical and health informatics. Thessaloniki, Greece (in press)
7. ICBHI Challenge. https://bhichallenge.med.auth.gr/rules
8. Kochetov K, Putin E, Azizov S, Skorbogatov I, Filchenkov A (2017) Wheeze detection using convolutional neural networks. In: Proceedings of EPIA conference on artifical intelligence. Porto, Portugal, pp 162–173. https://doi.org/10.1007/978-3-319-65340-2