Tracheal Opening Discrimination During Intubation Using Acoustic Features and Gaussian Mixture Model

Wei-Hao Chen1, Yu-Hsien Chiu2, Hsien-Chang Wang3, Yu-Wei Hung1, Hao-Po Su4, Kuo-Sheng Cheng1,6

1Department of Biomedical Engineering, National Cheng Kung University, Tainan 701, Taiwan, ROC
2Department of Healthcare Administration and Medical Informatics, Kaohsiung Medical University, Kaohsiung 807, Taiwan, ROC
3Department of Information Management, Chang Jung Christian University, Tainan 711, Taiwan, ROC
4Department of Anesthesiology, Yuan’s General Hospital, Kaohsiung 802, Taiwan, ROC

Abstract

Correct identification of glottic opening is crucial during endotracheal tube intubation for airway management. Direct glottis visualization by the physician is considered best standard practice, but it is dependent on conditions, skill, and experience. This study proposes an improved oxygen insufflation system that applies an acoustic modeling approach to discriminate glottic and non-glottic areas utilizing the acoustic response from a steady directional airflow at the hypopharynx. An electric stethoscope is placed at the suprasternal notch to record sound produced by insufflation during intubation. The Gaussian mixture model with mel-frequency cepstral coefficients (MFCCs) is used to determine differences between glottic and non-glottic areas. A dataset containing 56.567 seconds of non-glottic sound and 46.472 seconds of glottic sound was recorded from 8 anesthetized adults receiving intubation. Short-time analysis and several objective evaluations were performed to investigate system performance. The evaluation results show that the system achieved a high classification accuracy of 93.24%. The proposed approach outperformed a baseline linear discriminant analysis method with various configurations of linear prediction coding coefficients and MFCCs, and shows potential in improving glottic identification during endotracheal tube intubation.

Keywords: Airway management, Endotracheal intubation, Gaussian mixture model, Mel-frequency cepstral coefficients

1. Introduction

Endotracheal tube (ETT) intubation, where a plastic tube is introduced into the trachea to provide a conduit for mechanical ventilation, is a critical procedure in airway management. Delayed or unrecognized conditions caused by inappropriate ETT positioning result in poor ventilation and may lead to hypoxemia, brain damage, arrhythmia, cardiac arrest, and even death. Most intubations are performed under direct observation using conventional rigid and fiberoptic laryngoscopes [1]. The basic principle of these methods is to ensure proper tube placement by observing the tube pass through the vocal cords (tracheal opening, glottis). Intubation is subjective and highly dependent on practitioner experience [2,3]. Accordingly, developing more reliable and objective methods is important for verifying the tracheal opening during intubation for emergency airway management.

Related works have studied distinguishing inappropriate intubation [4,5]. Capnography is a reliable method for detecting tube position [6]. Auscultation of breath sounds over the bilateral chest and the epigastrium is commonly applied to ensure proper tube placement. However, these conventional methods focus on confirming tube position after intubation rather than ascertaining correctness during intubation. The American College of Emergency Physicians stated that verification of ETT intubation via ultrasound may be a possible future option [7]. Transtracheal scanning below the cricoid ring using ultrasound produces images of the ETT tube passing through the vocal chord and may offer objective evidence for identification during intubation [8,9]. The successful verification rate of real-time transtracheal ultrasound is 68%, which is similar to that for the auscultation method [10].

Moreover, viewable clearance during ETT intubation is an important issue in operation. Oxygen insufflation, where air is blown at 4–6 l/min from the tip of the fiberoptic endoscope, is used to keep secretions and hypopharyngeal tissue away from the lens for an unobstructed view [11,12]. Anatomically, the cartilaginous rings of the trachea prevent tracheal tube collapse, but the lack of supporting structure in the esophagus may allow
it to collapse. The structural differences between the tracheal and esophageal tubes result in different acoustic characteristics, which assist the practitioner to discriminate the tracheal opening during intubation. However, few studies have been conducted on such sound characteristics. Speech, produced from the human vocal tract, can be modeled as a time-variable filter of excited airflow. The mechanism for sound production varies according to the different types of speech. Robust features, such as mel-frequency cepstral coefficients (MFCCs) and linear prediction coding (LPC) coefficients, are applied for modeling and applications in speech recognition and synthesis \[13,14\]. In MFCC computation, the characteristics of human sound perception are mimicked by filter banks to transform speech signals into acoustic features. Specifically, MFCCs are used to estimate an approximation of the spectral envelope by means of several triangular band pass filters spaced linearly on the mel scale \[15\]. LPC is a speech production model that can simulate the shape characteristic of the vocal tract using an all-pole filter. LPC is used to analyze speech signals as a linear convolution of source and filter that is an approximation of the original spectrum. The LPC coefficients are computed to predict the speech signal using a linear function of previous samples \[14\].

In recent years, both LPC and cepstral analysis have been applied to analyze pathological laryngeal and lung sounds \[16,17\].

Motivated by the oxygen insufflation method and previous studies related to speech, a Gaussian mixture model (GMM) approach with MFCCs is proposed to determine differences between the sound signals produced by noisy airflow blown into glottic and non-glottic areas. The study also proposes an improved oxygen insufflation system based on a fiberoptic laryngoscope to enable reliable and robust tracheal sound classification for assisting intubation. The system can also serve as an alternative instrument for confirming tracheal tube placement. LPC features are compared to examine sound characteristics in both areas. A case study was performed to assess the performance of the proposed model-based approach.

2. Materials and methods

Correct intubation of the ETT should pass through the glottis, which is the inlet of the trachea at the hypopharynx. The basic idea is to identify the tracheal opening by the acoustic response of the introduced airflow to the hypopharynx. An oxygen insufflation system was modified to provide oxygen flow in the direction of the fiberoptic view. An electronic stethoscope system was developed for recording the sound produced by the insufflation flow for glottic and non-glottic areas. The glottis was confirmed under fiberoptic view in order to verify system efficacy.

2.1 Improved measurement instrumentation

A Norley’s device, which can continuously insufflate oxygen or air from the front of the bronchoscope during the fiberoptic intubation process, was assembled and modified \[18\]. A 1.1-mm single-orifice catheter was threaded through the working channel of a bronchoscope to protrude 0.1 cm beyond the tip. The proximal end of the catheter was connected to a flow-adjustable oxygen source. A 1-3 l/min airflow can pass through the catheter and out the tip of the bronchoscope.

Figure 1 depicts the framework of the proposed system for monitoring acoustic responses. The respiratory sound recording system includes a highly sensitive microphone within the stethoscope chestpiece, an audio acquisition card (Creative SoundBlaster X-Fi Go) with a preamplifier and an analog-to-digital converter (ADC), and custom signal analysis software. Based on the physiological and acoustic characteristics of respiratory sound, the hardware and associated software filters were designed with consideration of the linear gain and attenuation adjustment in the frequency range of 80 Hz to 8 kHz. The ADC utilized had a maximum sampling frequency of 10 kHz and 24-bit resolution. These specifications satisfy the requirements for collecting respiratory sound signals with minimum distortion. A software package was developed in C# with a user-friendly graphical user interface to aid in the recording, segmentation, tagging, and analysis of the sound signals.

2.2 Experimental protocol

The experimental protocol was approved by the institutional review board of Yuan’s General Hospital, Kaohsiung, Taiwan (NO. 20111006B) and written informed consent was obtained from each subject. The participants received general anesthesia for the operation. Under adequate anesthesia and neuromuscular blockade, the patient received fiberoptic intubation using the proposed device set at 2 l/min oxygen insufflation. The chestpiece of an electronic stethoscope was put on the suprasternal notch for recording. After the position of the fiberoptic tip was confirmed, at least 5 seconds of insufflation flow sounds were recorded for glottic and non-glottic areas. The clinician validated both glottic and non-glottic sounds produced from insufflated airflow in the hypopharynx using a stethoscope. The collected sounds were pre-amplified and pre-processed by band-pass filtering with low-frequency cutoff of 80 Hz and high-frequency cutoff of 8 kHz. The lower cutoff frequency diminishes the effects of heart noise and electronic interference.
2.3 Acoustic feature extraction

Short-time analysis was performed to split each sound recording into concatenated frames. To select representative sound features, previous studies on speech analysis used LPC and MFCCs, which resulted in good performance. This study uses LPC to examine vocal tract influence, and MFCCs to differentiate the spectral differences between the two areas. Figure 2 shows the block diagram for acoustic feature extraction and modeling. After feature analysis, linear discriminant analysis (LDA) was used as a baseline method for classifying glottic/non-glottic sounds for comparison to the GMM model.

$$E[e^2(j)] = r_{xx}(0) - \sum_{n=1}^{p} a_n r_{xx}(n)$$  \hspace{1cm} (3)

When the signal $e(j)$ is an uncorrelated zero-mean random process, all of the correlation functions have a recursive relationship among magnitudes at different lag times. The relationship for the autocovariance function is given as:

$$E[e(j)e(j-l)] = \begin{cases} \sigma^2_e & \text{if } j = l \\ 0 & \text{if } j \neq l \end{cases}$$  \hspace{1cm} (4)

where $\sigma^2_e$ is the variance of $e(j)$. The algorithm starts with a predictor of order zero and then calculates the coefficients of a predictor of order $i$ by using the coefficients of a predictor of order $i-1$, until the coefficients for the maximum order are calculated. This study uses the MATLAB software package (R2012a, The MathWorks Inc., Natick, MA, USA) to perform the calculation.

Another way to approximate the spectral envelope is to apply the filter bank processing of MFCCs. The mel-frequency cepstrum (MFC) is defined as the modified cepstrum of a windowed short-time signal derived from the fast Fourier transform of that signal [19]. Figure 3 shows the process of MFCC feature extraction. Each frame is multiplied by a Hamming function to reduce the discontinuity effect at the frame boundaries. The mel-frequency is proportional to the logarithm of the linear frequency, reflecting similar effects in human subjective hearing perception. The non-linear mel-frequency scale is defined as:

$$M(f) = 1125 \ln \left(1 + \frac{f}{700}\right)$$  \hspace{1cm} (5)

where $f$ is the frequency for a certain mel filter bank.

To compute MFCCs, a filter bank with a set of triangular band-pass filters computes the average spectrum to get the logarithm energy of each filter around each center frequency with increasing bandwidth. The discrete cosine transform is applied to the logarithm energy $E_n$ obtained from the triangular filter to derive the MFCCs as follows:

$$C_m = \sum_{n=1}^{F} \cos \left( m \cdot (n - 0.5) \cdot \frac{\pi}{F} \right) \cdot E_n \hspace{1cm} m = 1, 2, \ldots, M$$  \hspace{1cm} (6)

where $F$ is the number of triangular band-pass filters and $M$ is the number of mel-scale coefficients. Processing using an MFCC triangular filter bank results in higher spectral resolution at the lower frequency band, where the harmonic of the spectrum is removed and the spectral envelope is kept.
2.4 GMM-based classifier

The extracted MFCC features can be modeled for investigating the distributions of the sounds. The Gaussian process is the most widely applied of all probability models. The probability density functions (pdf) of many processes, such as speech, are non-Gaussian. A non-Gaussian pdf may be approximated by a weighted sum of a number of Gaussian densities of appropriate mean vectors and covariance matrices. Based on such characteristics, this study adopts a GMM with MFCC features as a binary classifier for modeling the glottic and non-glottic sounds. Let the parametric model $\Phi_i$ of class $i$ that models data sample $x$ have the conditional probability $P(\Phi_i|x)$. To choose class $\Phi_i$ with the greatest posterior probability, $k$ can be determined as:

$$ k = \text{argmax}_i P(\Phi_i|x) $$

(7)

Based on the class-conditional pdf and Bayesian rule, Eq. (4) becomes:

$$ k = \text{argmax}_i \left( \frac{P(x|\Phi_i)P(\Phi_i)}{P(x)} \right) = \text{argmax}_i P(x|\Phi_i)P(\Phi_i) $$

(8)

where the probability $P(x)$ is unrelated to the estimation of $k$ and can be ignored. In multivariate GMM, the pdf for observable data $x$ is the weighted sum of each Gaussian component and is represented as:

$$ P(x|\Phi) = \sum_{i=1}^{K} c_i P(x|\Phi_i) = \sum_{i=1}^{K} c_i N(x; \mu_i, \sigma_i) $$

(9)

where $0 \leq c_i \leq 1$, for $1 \leq k \leq K$, and $\sum_{i=1}^{K} c_i = 1$. To model a signal space with a $k$-mixture pdf, the expectation-maximization (EM) method can be applied for the estimation of the parameters of the Gaussian pdf models [20].

2.5. Statistical analysis

Several objective evaluations were performed to determine the suitable LPC order, MFCC combinations, and GMM. The statistical analysis approach consisted of:

1. Investigating the performance of the adopted acoustic features based on spectral differences, sensitivity, specificity, and classification error rate criteria under various configurations of LPC order and numbers of MFCCs with glottic and non-glottic sounds. Step-wise regression is used to select significant features. Features that can discriminate between glottic or non-glottic sounds are selected first. Independent features are chosen based on their incremental contribution over the features already selected.

2. Comparing the performance between the proposed GMM classification approach and the baseline LDA method using inside and outside test schemes under various numbers of mixture components in GMMs.

3. Determining the efficacy of the proposed approach via the receiver operating characteristic (ROC) curve. The area under the ROC (AUROC) is used to examine the feasibility of the proposed approach for glottic sound classification.

3. Results and discussion

Sounds were recorded from 8 adult subjects without respiratory diseases in Yuan’s General Hospital. All case studies resulted in successful ETT intubations. Neither discomfort nor complications related to the study were found in the follow-ups. All collected sounds were processed into frames and their LPC/MFCC features. The proposed GMM approach uses the EM algorithm to model the MFCC feature space with respect to the glottic and non-glottic sound frames. For evaluating the proposed approach, the $k$-fold cross validation technique was applied and all observations were used for both training and validation. Each observation was used for validation exactly once. In this study, the value of $k$ was set to 5. The data were divided into two subsets, one with 80% of the data (used for training) and the other with the remaining 20% of the data (used for validation).

3.1 Sound acquisition and preprocessing

For each subject, more than 5 seconds of sound signals were recorded from non-glottic and glottic locations. The sampling rate for the sound was 8 kHz with 16-bit resolution. For analysis, the sound signals were further segmented and tagged by an experienced clinician according to their corresponding locations. The total combined duration of the non-glottic sound recordings was 56.567 seconds and that of the glottic recordings was 46.472 seconds. Each recording was segmented into frames of 256 samples (32 ms) with 50% overlap with the following frame. There were a total of 3524 non-glottic frames and 2892 glottic frames for further feature extraction.

A comparison of the raw sound signals and the power spectra of glottic and non-glottic areas are shown in Fig. 4. Both signals show noise-like characteristics and little periodicity. There is no obvious harmonic effect observed from the power spectra. Both spectral envelopes are similar, but the energy attenuations of the glottic and non-glottic sounds have significant differences in the range of 200 Hz to 2 kHz. The major attenuation was within 200 - 1000 Hz, and the minor one was within 1000 - 1800 Hz. As previously described, the mel-frequency scale is suitable for modeling the characteristics. Its spectral resolution decreases with increasing frequency. The information in the lower frequency region is weighted by the mel-frequency scale to enable finer discrimination. The filter band processing is robust to noise and spectral estimation errors.

Figure 5 shows the developed software for analysis. The software was developed under the Microsoft Windows 7 operating system using the Microsoft Visual Studio .Net framework in the C# programming language. The software can select a source recording file and configure the frame size, frame overlap, LPC order, and combination of MFCCs. Furthermore, individual MFCCs can be picked to produce different outputs for further examination.
3.2 Determination of number of LPC and MFCC features

The number of LPC or MFCC features has a direct influence on the performance of sound classification. LPC order selection was based on a suitable prediction error rate. Figure 6 shows the evaluation results for both glottic and non-glottic sounds. The error rate of non-glottic sound converges when the LPC order is greater than 3; the glottic sound error rate shows an obvious elbow when the LPC order is 3. A large order resulted in inconsistency when modeling the sound. In determining computational efficiency, an order of 3 was selected and applied in the following evaluation. In addition, typically, only the first 12 cepstrum coefficients are used in speech recognition. Therefore, in this study, 12 cepstrum coefficients were used to parameterize and analyze the sound.

Table 1 shows the baseline evaluation results for the LDA method. Three parametric configurations, namely LPC with an order of 3, 12 MFCCs, and MFCCs after step-wise regression, were compared with/without data bootstrapping. Bootstrapping was performed 500 times. The MFCC and LPC features exhibit similar sensitivities, specificities, and accuracies with or without bootstrapping. This result suggests that the extracted features are sufficiently statistically sampled for representing the characteristics of the target sound. The comparison results show that MFCC features outperform LPC ones in all aspects of evaluation criteria. These results indicate that the characteristics of glottic and non-glottic sounds are well modeled using the MFCC approach, which is less affected by differences between individual vocal tracts. The step-wise regression technique degrades the overall performance because some MFCCs are screened out. Thus, all 12 MFCCs were used as the sound features for further GMM evaluation.

3.3. Evaluation of GMM-based classification

Figure 7 illustrates the architecture of the GMM-based classifier. Signal frames from the glottic or non-glottic groups were randomly divided into training and validation sets. The features of LPC and MFCCs were used to train the GMM by using the HMM toolkit (HTK ver. 2.2) [21]. Inside and outside tests on the collected datasets were conducted. Figure 8 shows the classification accuracy for the inside test for LPC (order = 3, with/without bootstrapping) and MFCC (order = 12, with/without bootstrapping) with GMM mixture numbers set to 1-20. The inside test used all sound files from the training subset for the model training and validation. MFCCs greatly outperform LPC with and without bootstrapping. A trade-off mixture number of 3 can be observed where the classification accuracy of GMM is about 93.24%. This accuracy is higher than that of conventional data-based LDA (86.35%).

Table 1. Comparison of three parametric configurations with classification rate obtained using LDA.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Without</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPC_3</td>
<td>74.432</td>
<td>53.562</td>
<td>65.025</td>
</tr>
<tr>
<td>MFCC_12</td>
<td>87.770</td>
<td>84.613</td>
<td>86.347</td>
</tr>
<tr>
<td>MFCC_StpReg</td>
<td>84.989</td>
<td>83.230</td>
<td>84.196</td>
</tr>
<tr>
<td><strong>With</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPC_3_Bstp</td>
<td>73.060</td>
<td>54.437</td>
<td>63.831</td>
</tr>
<tr>
<td>MFCC_12_Bstp</td>
<td>87.720</td>
<td>85.106</td>
<td>86.424</td>
</tr>
<tr>
<td>MFCC_StpReg_Bstp</td>
<td>85.194</td>
<td>83.563</td>
<td>84.386</td>
</tr>
</tbody>
</table>

**TABLE FOOTNOTE**
LPC_3: LPC with order 3
MFCC_12: 12 MFCCs
MFCC_StpReg: MFCCs selected by step-wise regression
LPC_3_Bstp: LPC with order 3 after bootstrapping
MFCC_12_Bstp: 12 MFCCs after bootstrapping
MFCC_StpReg_Bstp: MFCCs after bootstrapping and selected by step-wise regression

Figure 6. Selection of LPC order.
The outside test was conducted using the 5-fold cross validation scheme described above. The outside test was repeated for 5 rounds. In each round, 20% of the total data was randomly selected as the testing data. Figure 9 shows the results of the outside test as a function of GMM mixture number. The results are similar to those of the inside test, with MFCCs outperforming LPC, and the suitable GMM mixture number being 3. The classification accuracy of the GMM method is 92.44%.

In addition, the ROC curve of the evaluation (Fig. 10) for the proposed approach was drawn empirically based on sensitivity and specificity. For both glottic and non-glottic sounds, the ROC curve climbs sharply in the specificity range of 0 to 1. The AUROC for both cases is greater than 0.95, i.e., the sensitivity and specificity are both close to 1.0, which means that false positives and false negatives are both very low. This encouraging result shows that the proposed approach can efficiently discriminate between glottic and non-glottic sounds.

This pilot study investigated the characteristics of hypopharyngeal sound produced by insufflation oxygen. After neuromuscular blockade drugs are administered during intubation, the vocal tract remains unchanged and the LPC method shows low discrimination, as illustrated in Table 1. In contrast, MFCCs are suitable for modeling and differentiating the spectral differences between the two sounds. Despite the small number of cases, the performance of both LPC and MFCCs on the bootstrapped data shows no significant differences.

4. Conclusion

This study proposed an improved oxygen insufflation system and a GMM-based classification method with MFCC features for reliably and robustly identifying the tracheal opening during ETT intubation. A classification accuracy of 93.24% was obtained with 3 mixtures for GMM. Compared to the LPC results, the sound signals seem less related to the vocal tract characteristics, which means that the MFCC method may be independent of individual anatomic features. ROC also showed good performance in sensitivity and specificity. The proposed approach has potential in assisting glottis verification during ETT intubation. Further study with more data from many subjects and dynamic scanning for the glottic opening is needed. Future research may develop new measurement instrumentation tools for emergency airway management.

Acknowledgements

This paper was supported by the National Science Council of Taiwan under Grant NSC 102-2221-E- 037-003-MY3. We
also thank all the patients and staff for their participation in the study.

References