Medical Engineering & Physics 32 (2010) 668–672

Contents lists available at ScienceDirect

Medical Engineering & Physics

journal homepage: www.elsevier.com/locate/medengphy

Technical note

Vocalization removal for improved automatic segmentation of dual-axis swallowing accelerometry signals

Ervin Sejdić[a], Tiago H. Falk[a], Catriona M. Steele[b], Tom Chau[a,*]

[a] Bloorview Research Institute, Bloorview Kids Rehab and the Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada
[b] Toronto Rehabilitation Institute and the Department of Speech-Language Pathology, University of Toronto, Toronto, Ontario, Canada

ARTICLE INFO

Article history:
Received 16 February 2010
Received in revised form 6 April 2010
Accepted 7 April 2010

Keywords:
Speech removal
Cough removal
Dysphagia
Dual-axis swallowing accelerometry signals
Signal processing

ABSTRACT

Automatic segmentation of dual-axis swallowing accelerometry signals can be severely affected by strong vocalizations. In this paper, a method based on periodicity detection is proposed to detect and remove such vocalizations. Periodic signal components are detected using conventional speech processing techniques and information from both axes are combined to improve vocalization detection accuracy. Experiments with 408 healthy subjects performing dry, wet, and wet chin tuck swallows show that the proposed method attains an average 95.3% sensitivity and 96.3% specificity. When applied in conjunction with an automatic segmentation algorithm, it is observed that segmentation accuracy improves by approximately 55%. These results encourage further development of medical devices for the detection of swallowing difficulties.

1. Introduction

Swallowing accelerometry, a potentially informative approach to bedside dysphagia screening, requires minimally invasive measurements that require only the superficial attachment of a sensor anterior to the thyroid notch [1–5]. To automatically analyze swallowing accelerometry signals, a critical first step is the demarcation of individual swallows within an extended recording of vibrations collected from the neck. For that purpose, several algorithms have been proposed in recent years (e.g. [6–8]). Given the two-dimensional movement of the hyoid and larynx during swallowing [9,10], these techniques exploited both anterior–posterior (A–P) and superior–inferior (S–I) vibrations simultaneously in order to achieve accurate results. The main goal of automatic segmentation is to accurately isolate the signal segments corresponding to the physiological phenomena under consideration (i.e. swallowing), while disregarding undesired signal contributions from vocalizations and motion artifacts.

Previous contributions in the field have observed that vocalizations (speech or cough) can severely contaminate swallowing accelerometry signals (e.g. [6,11]). Such vocalizations can be classified as voluntary or involuntary. Voluntary vocalizations are usually associated with speech. In clinical settings, patients are often cued to vocalize after swallowing in order to check for potential signs of aspiration [12]. Involuntary vocalizations, in turn, are associated with coughing, which can occur immediately after swallowing. Generally, coughing is a physiological response to aspiration [13]. Despite the fact that different vocalizations can be indicative of swallowing difficulties [14], their presence can mask true swallows and hamper automatic analysis of the signals, such as, automatic swallow demarcation [6]. Therefore, there is a growing need for the automatic removal of vocalizations from swallowing accelerometry signals. The goal of this paper is to develop such a system for improved segmentation of swallowing signals.

The remainder of this paper is organized as follows: Section 2 outlines the experimental protocols as well as the proposed system. Section 3 presents experimental results and conclusions are drawn in Section 4.

2. Methodology

2.1. Experimental protocol

We recruited four hundred and eight neurologically healthy adult (aged 18–65) participants with no history of swallowing disorders from a local science center. The research protocol was approved by the Toronto Rehabilitation Institute and Bloorview Kids Rehab, both located in Toronto, Ontario, Canada. All participants provided written consent. We collected swallowing accelerometry signals using a dual-axis accelerometer (ADXL322, Analog Devices) attached to the participant's neck (anterior to the cricoid cartilage) as shown in Fig. 1. The axes of acceleration were aligned to the anatomical anterior–posterior (A–P) and

* Corresponding author. Tel.: +1 416 425 6620x3515.
E-mail addresses: esejdic@ieee.org (E. Sejdić), tiago.falk@ieee.org (T.H. Falk), steele.catriona@torontorehab.on.ca (C.M. Steele), tom.chau@utoronto.ca (T. Chau).

1350-4533/$ – see front matter © 2010 IPEM. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.medengphy.2010.04.008

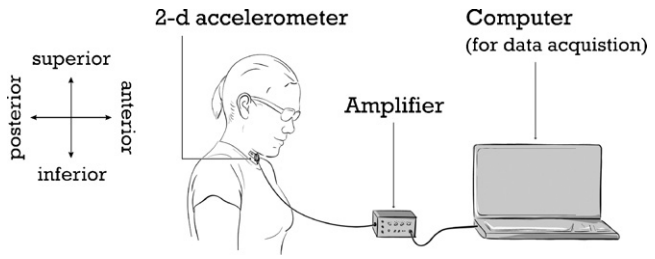

Fig. 1. Experimental setup.

superior–inferior (S–I) directions. As in previous dual-axis swallowing accelerometry studies (e.g. [7,15]), data were band-pass filtered using hardware with a pass band of 0.1–3000 Hz and sampled at 10 kHz using a custom LabVIEW program running on a laptop computer.

During data collection, participants were cued to perform three types of swallows. Initially, each participant performed five saliva swallows (so called dry swallows) with a brief rest interval between swallows to allow for saliva production. Next, the participant completed five water swallows by cup with their chin in the natural position (so called wet swallows) and five water swallows in the chin-tucked position (so called wet chin tuck swallows). The entire data collection session lasted 15 min per participant. The participants were instructed not to vocalize. Nonetheless, approximately one quarter of all recordings contained either voluntary (e.g. speech) or involuntary (e.g. coughs) vocalizations. The voluntary vocalizations occurred because some participants did not comply with the instructions to remain silent.

2.2. Proposed approach

A typical swallowing accelerometry signal, $x(n)$, can be expressed as follows:

$$x(n) = \phi(n) + \varepsilon(n), \tag{1}$$

where $0 \leq n \leq N-1$ and N represents the length of the signal, $\phi(n)$ is a signal asscociated with swallowing activities of a person, and the observed noise, $\varepsilon(n)$, is assumed to be additive. No assumptions are made about the probability distribution function of noise. As pointed out in previous contributions (e.g. [6]), vocalizations (speech or cough) can severely alter the amplitudes of dual-axis swallowing accelerometry signals and hence confound any subsequent data processing. Therefore, when vocalizations are present, the recorded swallowing accelerometry signal, $\chi(n)$, can be represented as

$$\chi(n) = \upsilon(n) + x(n), \tag{2}$$

where $\upsilon(n)$ is a signal associated with vocalization. Due to physiological reasons, vocalization and swallowing cannot occur simultaneously [13]. In order to reliably detect vocalization instances, a conventional speech processing tool, namely fundamental frequency estimation, or pitch tracking, is explored.

Speech sounds are produced by forced air from the lungs as it passes between the vocal cords in the larynx at the base of the throat. The vocal cords vibrate periodically and create voiced sounds characterized by the fundamental frequency of the vibration, termed the pitch frequency [16]. Typical fundamental frequencies for the human voice range from 85 to 155 Hz for the adult male, 165 to 255 Hz for the adult female and 208 to 410 Hz for an infant or child younger than 10-years of age [16]. Coughing is a defence mechanism to clear the airway of inhaled foreign bodies [17]. Additionally, it enhances mucociliary clearance in cases of impaired ciliary function and excessive mucus production [18]. It is well known that coughing is presented by a sudden expulsion of air which is accompanied by a typical sound. These sounds have durations which can last from 0.4 to 1 s [17,19]. Frequency analysis of coughs have revealed periodicity with the fundamental frequency ranges from approximately 50 Hz to as high as 600 Hz (e.g. [17,19]). The plots in Fig. 2(a)–(f) illustrate the manifestation of speech and coughs in cervical vibration signals. Fig. 2(a) and (b) depict a sample 0.8 s speech vocalization in the A–P direction and a 40 ms zoomed-in portion, respectively. Similarly, Fig. 2(c) and (d) depict a sample 0.4 s duration cough and a 40 ms zoomed-in portion, respectively.

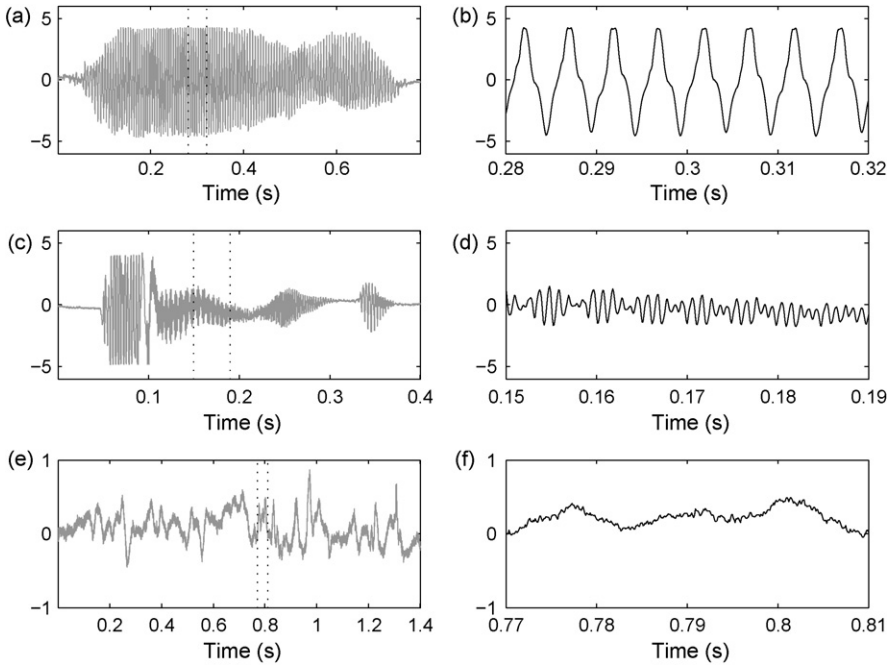

Fig. 2. Comparison of various vibration signals: subplot pairs (a and b), (c and d) and (e and f) depict sample vibrations and zoomed-in portions for speech, cough and swallows in the A–P direction, respectively. The dotted vertical lines in (a), (c) and (e) represent the locations of zoomed-in portions shown in (b), (d) and (f), respectively.

As observed, the two signals exhibit high periodicity. On the other hand, swallowing signals are shown to exhibit aperiodic behaviour, as illustrated by the plots in Fig. 2(e) and (f). It is postulated that swallowing vibration signals are primarily due to a mechanical phenomenon, i.e. motion of the hyolaryngeal structure [11], and hence, do not exhibit periodic behaviour. Furthermore, a frequency analysis of accelerometry signals containing speech and coughs reveals frequency components with significant power even below 50 Hz, which overlap with the frequency band containing swallows [11].

Pitch tracking algorithms have been used by the speech processing community for decades and serve to detect periodicities and measure fundamental frequencies in real time. Here, the state-of-the-art in pitch tracking, namely the robust algorithm for pitch tracking (RAPT) is explored [20]. Initially, the algorithm detects location and duration of periodic components independently on each axis. This information is then combined and periodic components due to vocalizations are removed from the recording. To summarize, the proposed algorithm is defined via the following steps:

1. Using RAPT, the locations and durations of periodic components, if any, are detected from each axis.
2. Two indicator sequences, $I_{A-P}(n)$ and $I_{S-I}(n)$, denote the locations and durations of possible vocalizations on each axis.

$$I_{A-P}(n) = \begin{cases} 1 & \text{if vocalization detected in the A–P direction} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$I_{S-I}(n) = \begin{cases} 1 & \text{if vocalization detected in the S–I direction} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $0 \le n \le N-1$, and as before, N is the length of the signal.

3. A combined indicator sequence, $I(n)$, is formed by computing the logical conjunction between $I_{A-P}(n)$ and $I_{S-I}(n)$:

$$I(n) = I_{A-P}(n) \Lambda I_{S-I}(n) \quad (5)$$

4. Using the combined indicator sequence, the accelerometry signal at the time instant n is removed if $I(n) = 1$.

If vocalizations are present in the recorded signals, the length of the processed signals will be shorter than the original; otherwise, signal length remains unchanged. The essential step in the proposed method is the formation of the combined indicator sequence. As will be shown in Section 3, by combining information from each axis separately, false positives (e.g. very short-duration periodicities during swallows or noise components) are avoided, since vibrations associated with vocalizations are strongly reflected in both axes. False negatives (undetected vocalizations) are seldom observed.

2.3. Algorithm evaluation

As in [15], each recorded signal was inspected visually and auditorily by two independent raters. Their consensus ratings provided the swallow and vocalization labels against which the proposed algorithm was evaluated. In order to apply RAPT, various parameters had to be set. For most of the coefficients, recommendations made in [20] were followed. Here, we only list coefficients whose values were customized to the present application: the duration of the analysis frame was set to 0.035 s; the duration of the correlation window size was set to 0.01 s; and the minimum acceptable peak value in the normalized cross-correlation function was set to 0.2. The reader is referred to [20] for more details regarding the RAPT algorithm.

To test the robustness of RAPT, performance metrics sensitivity and specificity were used with the following parameters:

- *True positive* (*TP*)—the number of correctly identified vocalizations;
- *False positive* (*FP*)—the number of incorrectly identified vocalization-free segments as vocalizations;

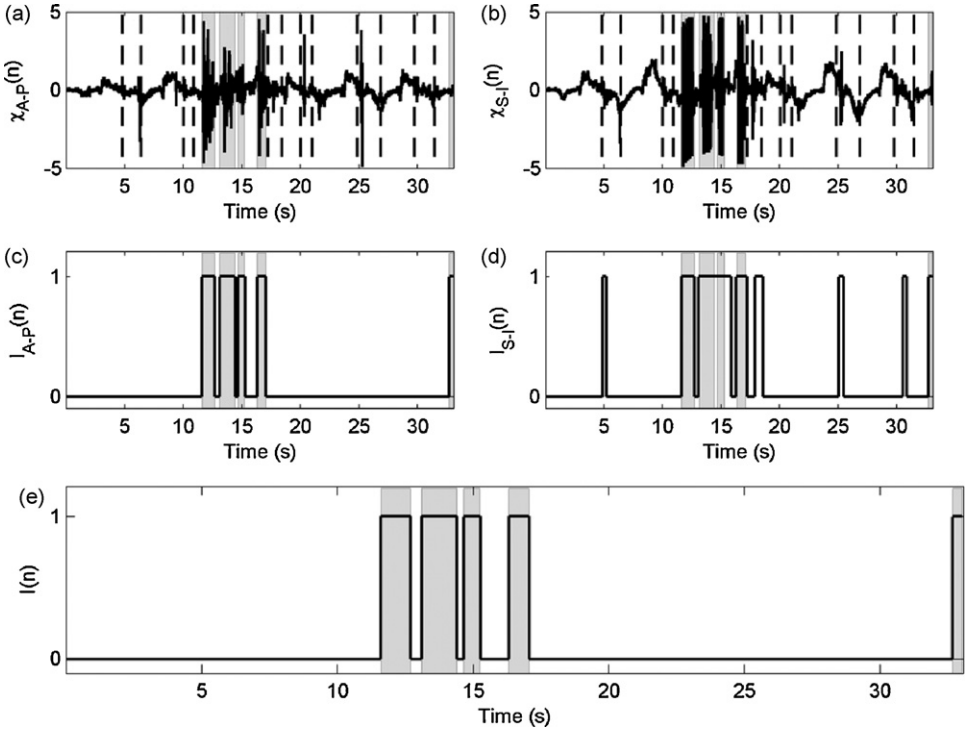

Fig. 3. Sample wet chin tuck swallowing recordings in the A–P direction (a) and the S–I direction (b) with vocalization present. The resulting indicator sequences: $I_{A-P}(n)$ (c), $I_{S-I}(n)$ (d), $I(n)$ (e).

Table 1
Accuracy analysis of the proposed approach. NRV = number of recordings containing vocalizations; TNV = total number of vocalizations; NDV = number of detected vocalizations; SENS = sensitivity; SPEC = specificity.

Swallowing type	NRV	TNV	NDV	SENS (%)	SPEC (%)
Dry swallows	147	365	350	96.2	99.8
Wet swallows	86	147	136	92.5	97.1
Wet chin tuck	113	245	233	95.5	95.3
Overall	346	757	719	95.3	96.3

- *True negative* (*TN*)—the number of correctly identified swallows;
- *False negative* (*FN*)—the number of missed vocalizations.

It is important to emphasize that for a vocalization to be counted as a TP, the proposed algorithm must capture more than 90% of the vocalization duration; otherwise, the vocalization is counted as a FN. Using these metrics we calculated the sensitivity ($= TP/(TP + FN)$) and specificity ($= TN/(TN + FP)$).

Next, we examined the effect of removing vocalizations on segmentation accuracy. Initially, recordings containing vocalization were segmented using the procedure described in [6]. For each recording, we denoted the number of swallows present, the number of correctly segmented swallows (CSS) and the number of incorrectly identified swallows (IIS). A swallow was considered correctly identified only when more than 90% of the swallow duration was captured by the segmentation process. As the second step of this analysis, the recordings were pre-processed using the proposed approach and segmented as outlined in [6]. CSS and IIS were subsequently computed and compared to values obtained without vocalization removal.

3. Results and discussion

Table 1 reports sensitivity and specificity performance metrics for the proposed vocalization removal technique for the three different swallow types. Over 95% sensitivity and 96% specificity were obtained on average.

Fig. 3(a)–(e) serve to further illustrate the performance of the proposed method. Fig. 3(a) and (b) depict a swallowing recording with vocalization present in the A–P and S–I directions, respectively. Vertical dashed lines indicated the boundaries of swallows, while the shaded region represents the location of vocalization. Fig. 3(c)–(e) further depict the indicator sequences $I_{A-P}(n)$, $I_{S-I}(n)$ and $I(n)$ computed by the algorithm, respectively. As observed, all vocalizations are correctly detected both in the A–P and S–I directions. Due to short-duration periodicities observed in the swallows in the S–I directions, four swallow components were erroneously classified as vocalizations (at approximately 17, 23, 25 and 31 s). Such errors are removed once the combined indicator sequence, $I(n)$, is used. We should also point out that the main reason for the occurrence of false positives is the fact that the vibrations associated with swallowing can become partly periodic in some cases (e.g. a person gulps while swallowing).

Table 2 summarizes the gains obtained in automatic swallowing segmentation once vocalization removal is applied for the 346 recordings containing vocalizations. To remain within the scope of the current manuscript, we did not process recordings without vocalizations. As observed, for dry swallows, the percentage of correctly classified swallows increased from 44% (without vocalization removal) to over 91% with the proposed method. The increase in the accuracy for wet and wet chin tuck swallows was more moderate; such a behaviour is expected since other factors can confound segmentation (e.g. head movements are pronounced during wet and wet chin tuck swallows). On average, a gain in CSS of approximately 55% and a reduction in IIS of over 65% was obtained over the three swallow types.

To further illustrate the gains obtained with vocalization removal, Fig. 4 depicts swallowing signals with a vocalization in the A–P and S–I directions (subplots (a) and (b), respectively), and the segmentation sequence in subplot (c), obtained using the approach in [6]. As before, the vertical dashed lines denote true locations of

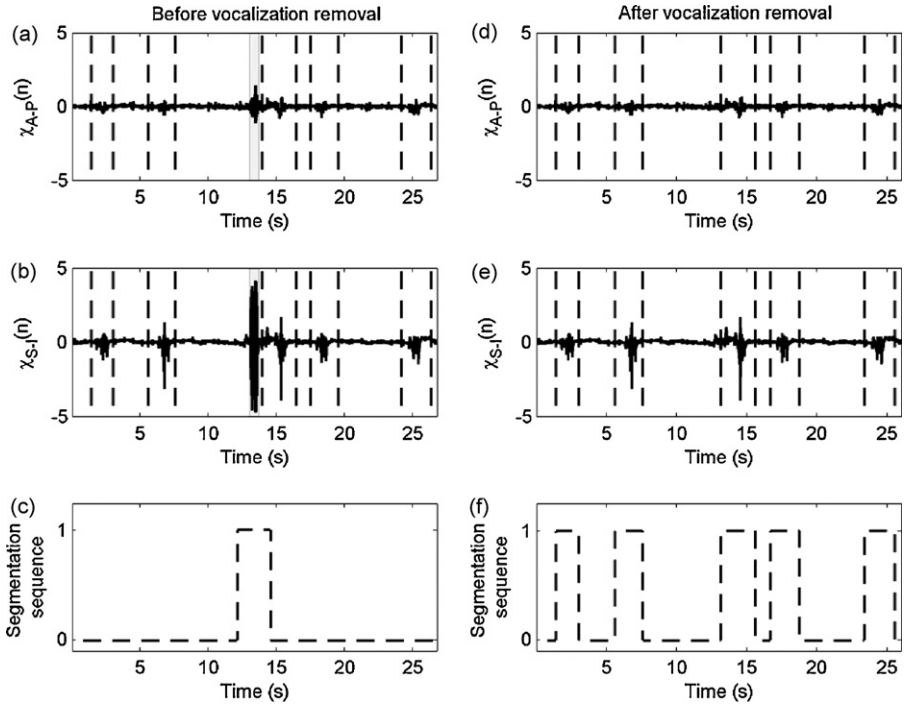

Fig. 4. Sample dry swallowing recordings in the A–P direction (a) and the S–I direction (b) with vocalization present. The segmentation sequence (c) when applied to signals with vocalization present. The recordings in the A–P direction (d) and the S–I direction (e) with vocalization removed. The segmentation sequence (f) when applied to signals with vocalization removed.

Table 2
Accuracy of the segmentation algorithm before and after the proposed method. TNS = total number of swallows; CSS = number of correctly segmented swallows; %CSS = percentage of orrectly segmented swallows; IIS = number of incorrectly identified swallows; %IIS = percentage of incorrectly identified swallows.

Swallowing type	TNS	Without vocalization removal				With vocalization removal			
		CSS	%CSS	IIS	%IIS	CSS	%CSS	IIS	%IIS
Dry swallows	718	317	44.1	225	30.3	657	91.5	65	9.05
Wet swallows	424	296	69.8	93	21.9	382	90.1	32	7.55
Wet chin tuck	565	355	62.8	143	25.3	467	82.7	67	11.8
Overall	1707	968	56.7	461	27.0	1506	88.2	164	9.61

swallows, while the shaded vertical bars indicated the location of vocalizations. Clearly, the segmentation sequence in (c) missed the majority of swallows. In contrast, the same signals after vocalization removal are depicted in subplots (d) and (e). Note that the vocalization activity previously highlighted by the shaded bar is now absent. The corresponding segmentation sequence obtained again using [6], is shown in subplot (f). Evidently, upon vocalization removal, the segmentation algorithm positively identified all five swallows present in the recording.

It should be pointed out that due to the structure of the RAPT algorithm, the proposed scheme can be computationally expensive (the analysis was conducted using MATLAB on a 2.5 GHz PC with 4 GB of RAM). Nevertheless, we believe that through code optimizations and possible C implementation, the speed of the proposed algorithm can be significantly enhanced. Also, a large swallow punctuated with an audible vocalization (e.g. a gulp) can potentially introduce periodic components near the end of a swallow. In those situations, we risk removing parts of swallows as well.

4. Conclusion

In this paper, an algorithm based on fundamental frequency tracking was proposed for the removal of vocalization disturbances from dual-axis swallowing accelerometry signals. The algorithm was designed specifically to alleviate the effects of vocalization on the automatic segmentation process. Experimental results show that, on average, vocalizations are detected and removed with 95.3% sensitivity and 96.3% specificity. When the proposed method is used for preprocessing of dual-axis swallowing accelerometry signals, it is observed that segmentation accuracy increases by an average 55% and errors are reduced by over 65%.

Acknowledgments

This research was funded in part by the Ontario Centres of Excellence, the Health Technology Exchange, the Toronto Rehabilitation Institute, Bloorview Kids Rehab, and the Canada Research Chairs Program.

Conflict of interest

None.

References

[1] Reddy NP, Katakam A, Gupta V, Unnikrishnan R, Narayanan J, Canilang EP. Measurements of acceleration during videofluorographic evaluation of dysphagic patients. Medical Engineering and Physics 2000;22(July (6)):405–12.
[2] Das A, Reddy NP, Narayanan J. Hybrid fuzzy logic committee neural networks for recognition of swallow acceleration signals. Computer Methods and Programs in Biomedicine 2001;64(February (2)):87–99.
[3] Chau T, Chau D, Casas M, Berall G, Kenny DJ. Investigating the stationarity of paediatric aspiration signals. IEEE Transactions on Neural Systems and Rehabilitation Engineering 2005;13(March (1)):99–105.
[4] Lee J, Blain S, Casas M, Kenny D, Berall G, Chau T. A radial basis classifier for the automatic detection of aspiration in children with dysphagia. Journal of NeuroEngineering and Rehabilitation 2006;3(July (14)):17.
[5] Sejdić E, Komisar V, Steele CM, Chau T. Baseline characteristics of dual-axis swallowing accelerometry signals. Annals of Biomedical Engineering 2010;38(March (3)):1048–59.
[6] Sejdić E, Steele CM, Chau T. Segmentation of dual-axis swallowing accelerometry signals in healthy subjects with analysis of anthropometric effects on duration of swallowing activities. IEEE Transactions on Biomedical Engineering 2009;56(April (4)):1090–7.
[7] Lee J, Steele CM, Chau T. Swallow segmentation with artificial neural networks and multi-sensor fusion. Medical Engineering and Physics 2009;31(November (9)):1049–55.
[8] S. Damouras, E. Sejdić, C.M. Steele, T. Chau, An on-line swallow detection algorithm based on the quadratic variation of dual-axis accelerometry, IEEE Transactions on Signal Processing; accepted for publication.
[9] Kim Y, McCullough GH. Maximum hyoid displacement in normal swallowing. Dysphagia 2008;23(September (3)):274–9.
[10] Ishida R, Palmer JB, Hiiemae KM. Hyoid motion during swallowing: factors affecting forward and upward displacement. Dysphagia 2002;17(December (4)):262–72.
[11] Lee J, Steele CM, Chau T. Time and time-frequency characterization of dual-axis swallowing accelerometry signals. Physiological Measurement 2008;29(September (9)):1105–20.
[12] Warms T, Richards J. 'Wet voice' as a predictor of penetration and aspiration in oropharyngeal dysphagia. Dysphagia 2000;15(March (2)):84–8.
[13] Smith CH, Logemann JA, Colangelo LA, Rademaker AW, Paulosk BR. Incidence and patient characteristics associated with silent aspiration in the acute care setting. Dysphagia 1999;14(January (1)):1–7.
[14] Marik PE, Kaplan D. Aspiration pneumonia and dysphagia in the elderly. Chest 2003;124(July (1)):328–36.
[15] F. Hanna, S.M. Molfenter, R.E. Cliffe, T. Chau, C.M. Steele, Anthropometric and demographic correlates of dual-axis swallowing accelerometry signal characteristics: a canonical correlation analysis, Dysphagia; accepted for publication.
[16] Baken RJ, Orlikoff RF. Clinical measurement of speech and voice. 2nd ed. San Diego, USA: Singular Publishing Group; 2000.
[17] Korpáš J, Sadloňová J, Vrabec M. Analysis of the cough sound: an overview. Pulmonary Pharmacology 1996;9(October (5–6)):261–8.
[18] Piirilä P, Sovijärvi AR. Objective assessment of cough. European Respiratory Journal 1995;8(November (11)):1949–56.
[19] Van Hirtum A, Berckmans D. Assessing the sound of cough towards vocality. Medical Engineering and Physics 2002;24(September/October (7–8)):535–40.
[20] Talkin D. Speech coding and synthesis. In: A robust algorithm for pitch tracking (RAPT)). Amsterdam, Netherlands: Elsevier; 1995. p. 495–518.