Proceedings of the 2020 Design of Medical Devices Conference
DMD2020
April 6, 7-9, 2020, Minneapolis, MN, USA

DMD2020-9081

REAL-TIME VOICE ACTIVITY DETECTION USING NECK-MOUNTED ACCELEROMETERS FOR CONTROLLING A WEARABLE VIBRATION DEVICE TO TREAT SPEECH IMPAIRMENT

Saurav Dubey[1], Arash Mahnan, Jürgen Konczak
Human Sensorimotor Control Laboratory, School of Kinesiology, University of Minnesota
Minneapolis, Minnesota, United States of America

ABSTRACT

Speech analysis using microphones can be problematic for Voice Activity Detection (VAD) in the presence of background noise. This study explored the use of wearable accelerometers instead of microphones. We assessed if accelerometers placed on the neck can be part of a VAD system embedded in a wearable collar-like device that delivers vibro-tactile stimulation (VTS) to the larynx during speech as a therapy for patients with the voice disorder spasmodic dysphonia. Specifically, we aimed to a) find the ideal location for placing accelerometers to the neck, and b) develop a VAD algorithm that detects the onset and offset of speech. Six healthy adult participants (M/F = 3/3, age = 26 (5.1)) vocalized 20 sample sentences with and without VTS at three neck locations: 1) thyroid cartilage, 2) sterno-cleidomastoid, and 3) posterior neck above C7. Based on time-synchronized acceleration and audio signals, VAD algorithm identified the Number of Onsets of Speech and Total Time Voiced. The thyroid cartilage attachment location had over 90% accuracy detecting speech in both measures. The average accuracy of the sternocleidomastoid and C7 locations were below 75% and 15% respectively. VAD accuracy decreased with the presence of VTS trials at all locations. We conclude that accelerometer signals due to tissue motion at thyroid cartilage are most suitable for real-time VAD. These findings support the feasibility of accelerometer-based voice detection for the use in medical devices that target speech and voice disorders.

Keywords: *Speech, vibration, voice disorder*

NOMENCLATURE

VAD	Voice Activity Detection
SD	Spasmodic Dysphonia
VTS	Vibro-tactile Stimulation
dB	Decibel

INTRODUCTION

Many current Voice Activity Detection (VAD) devices utilize microphones. However, the reliability of these systems decreases in environments where other voices or environmental noise obscure targeted speech. Interest in accelerometers to detect voice activity, especially in noisy environments, has increased to address the shortcomings of using microphones. Although binaural microphone recording is a commonly accepted method to reduce recorded background noise in low signal-to-noise ratio environments, a neck-attached accelerometer can capture voice signals significantly more accurate [1]. Even in the presence of 67.5 dB background noise, compared to the 40 dB volume tested by Lindstrom et. al. [1], a neck-attached accelerometer was resistant to noise while recording laryngeal vibrations [2]. Besides these studies that were conducted in controlled laboratory settings, accelerometers have successfully been employed in workplace environments with random background noise to track speech activity throughout a day [3].

While most speech analysis research focuses on populations with healthy speech production, some devices, like the Ambulatory Phonation Monitor [4], have also been tested with patients of spasmodic dysphonia (SD), a disorder where involuntary spasms of the laryngeal muscles decrease speech quality [5]. Issues with microphone recordings of healthy voices are exacerbated in these dysphonic patients due to the strained speech produced, often at a much lower volume than is normal for the healthy population. As dysphonia severity increases, the effectiveness of microphone-based measures decrease [6]. Accelerometers, however, are able to capture acceleration data from speech even in most severe cases of dysphonia [4].

The current study proposes the use of an accelerometer to measure neck surface vibrations as a mean for detecting speech in real-time. The work is part of a larger project with the goal of developing a wearable collar that delivers vibration therapy to

[1] Contact author: dubey019@umn.edu.

SD patients as they speak [7]. A persistent improvement in voice quality has been demonstrated in SD patients following the application of vibro-tactile stimulation (VTS) to the laryngeal muscles [8]. While existing devices are able to utilize either amplitude or frequency based algorithms for VAD, none have done so with concurrent VTS in the vicinity of the accelerometer as this study proposes. The introduction of VTS adds noise to the accelerometer's recording, which could interfere with the existing VAD techniques. The aims of this study are to (1) determine the anatomical location for accelerometer placement that provides a signal that most reliably distinguishes between speech and non-speech during laryngeal VTS and (2) to develop a signal processing algorithm for VAD for real-time implementation with VTS during speech activity. The attachment location must be within the region of a neck collar circumscribing the thyroid cartilage.

METHODS

2.1 Participants

Voice and acceleration data were collected from 6 healthy adult participants (M/F = 3/3, 26 ± 5.1 years). Healthy was defined as lacking any self-identified neurological, movement, or speech disorder. The experiment protocol was approved by the Institutional Research Board (IRB) of the University of Minnesota. All the participants signed a consent letter prior to attending the experiment.

2.2 Instrumentation

Audio data were recorded at 44100 Hz using an ECM-88B Electret Condenser Microphone (Sony Corporation, Tokyo, Japan) connected to a MixPre-6 microphone preamplifier. The preamplifier was connected to a computer using Audacity [9] to record the signal. The accelerometer used was a BU-27135-000, a single-axis accelerometer (Knowles Electronics LLC, Itasca, IL, United States). The acceleration data were collected and recorded directly to an SD card by an Arduino Uno R3 (BCMI, Italy) at 1000 Hz. This sampling frequency accounts for the average fundamental frequency of vocalized vowels in males and females of up to 400 Hz [10]. The accelerometer was connected to the Arduino Uno using three insulated multiple strand wires.

Acceleration and audio recordings were time-synchronized in order to compare accuracy of the VAD algorithm between each recording. Time synchronization was implemented by producing a 1000 Hz beep for 250 milliseconds at the start and end of accelerometer signal recording that was recorded in the audio signal. The audio signal was then trimmed to the start of the first and last beep. Laryngeal VTS was provided by 2 Pico Vibe™ 9-millimeter vibrators (Precision MicrodrivesTM, Model 307 – 100). The vibrators operated at a stimulation frequency of 100 Hz, supplied with 1.1 volts from an adjustable voltage power block. The accelerometer and vibrators were attached to the skin of the neck using double sided tape and one piece of Blenderm tape (3M, Maplewood, MN, United States) on top of each device to secure them to the skin (see FIGURE 1).

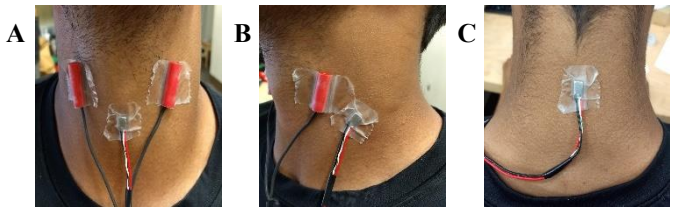

FIGURE 1: Attachment of vibrators, in red, to the laryngeal muscles lateral to the thyroid cartilage and accelerometer attachment locations: (A) thyroid cartilage, (B) sternocleidomastoid lateral to the thyroid cartilage, and (C) 2.5 centimeters above C7.

2.3 Procedure

The accelerometer was attached to three locations: thyroid cartilage below the thyroid notch, sternocleidomastoid in line with the thyroid notch, and 2.5 centimeters above C7 on the back of the neck. Vibrators were attached bilaterally over the laryngeal area, lateral to the thyroid cartilage (see FIGURE 1A).

The experimental protocol consisted of speaking 20 sample sentences in 6 different conditions of 2 variables: accelerometer attachment location and application of VTS (see FIGURE 2).

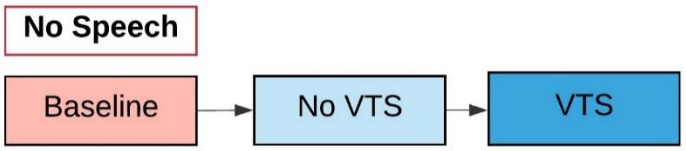

FIGURE 2: Experimental protocol at each accelerometer attachment location. Data were collected beginning with a baseline trial (5000 ms), followed by separate trials with and without the application of VTS.

Speech data were collected with and without VTS application at all accelerometer attachment locations. A baseline trial for 5000 milliseconds with no speech was recorded at each accelerometer location to determine a threshold for the VAD algorithm (see FIGURE 2).

2.4 Signal Processing

Audio data were exported from Audacity (The Audacity Team, Pittsburgh, Pennsylvania, U.S.A). Acceleration data were converted from binary to comma-separated value files by the Arduino and then exported from the SD card. Subsequent data analysis and signal processing was conducted in MATLAB R2018b (The MathWorks Inc., Natick, Massachusetts, U.S.A.).

Detecting voice activity from acceleration data requires signal processing to filter the data and set thresholds for speech. A basic, computationally efficient measure to characterize a signal for VAD is average power in the time and frequency domain. Average power, P, in the time domain of a discrete-time signal can be calculated as the average of the sum of each squared value, N, in the signal:

$$P = \frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2 \quad (1)$$

This calculation can also be conducted similarly in the frequency domain to find spectral energy density, power spectral density,

and average spectral power. The spectral energy density, $E_s(f)$ can be calculated using the Fourier transform, $X(f)$:

$$E_s(f) = |X(f)|^2 \quad (2)$$

The power spectral density of the signal, P_{xx}, where Δt represents the sampling interval is:

$$P_{xx}(f) = \frac{\Delta t}{N} |\sum_{n=0}^{N-1} X(f)_n|^2 \quad (3)$$

The average spectral power of this signal in the frequency domain, P_f, can then be calculated by integrating the power spectral density over all frequencies in the signal, where f_s is the sampling frequency:

$$P_f = \int_{-f_s/2}^{f_s/2} P_{xx}(f)\, df \quad (4)$$

Parseval's Theorem states the energy of the signal in the time domain is equal to the summation of all frequency components of the spectral energy density of the signal [11]. Thus, the calculated average power of a window of samples in the time and frequency domain would be equal as well.

Non-VTS and VTS trials were treated with different filters. Non-VTS trials were used as evidence for the feasibility of using an accelerometer at any of the chosen regions around the neck for generally applicable speech detection. VTS trials targeted the specific application of the delivery of VTS using a collar [7]. Thus, filters to remove the frequency and noise associated with VTS and the input current to the vibrators were only applied in trials with VTS.

Non-VTS trials were treated with a band-pass filter from 80-400 Hz to remove movement artifacts and high frequency noise from the accelerometer. VTS trials were treated with a band pass filter from 110-400 Hz. The increase from 80 Hz to 110 Hz on the lower stop band accounts for the frequency of the vibrators that varied from 99 Hz to 109 Hz across participants. This slight variance in frequency of the vibrators can be attributed to attenuation of the signal due to anatomical differences of the neck region between participants and minor variability of the input voltage provided to the vibrators. The VTS trials were then treated with three band-stop filters to remove the 2nd, 3rd, and 4th harmonic of the 60 Hz noise caused by the current to the vibrators. Only FIR filters were used to avoid nonlinear phase distortions. These filters were applied using the MATLAB function *filtfilt*, which produces zero-phase distortion, however, it does introduce a constant magnitude distortion. This distortion was addressed in the signal processing workflow by applying a scale factor to adjust threshold magnitude. Because there is a magnitude distortion and the possibility of phase distortion when using other filters, time domain calculations on VTS trials after filtering were excluded.

2.5 VAD Algorithm

To differentiate speech acceleration signals from no activity, a baseline trial in which the participants were asked to not speak and remain stationary for each location was used to calculate a threshold. A 5000-millisecond baseline trial was recorded and divided into 50-millisecond subintervals. The average power in time and frequency domain were calculated for each subinterval, then averaged across all subintervals which is then defined as the threshold for speech activity (see FIGURE 3).

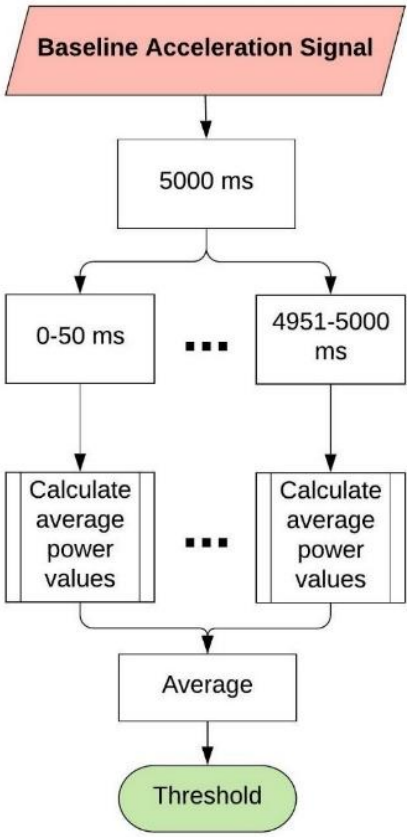

FIGURE 3: Threshold calculation. A 5000-millisecond interval of the baseline trial was divided into 100 subintervals of 50 milliseconds. The average power values for each of these subintervals were calculated. These average power values were then averaged together to calculate the threshold for speech activity to be used by the VAD algorithm (see Figure 5)

The VAD algorithm divided the complete signal into intervals that were filtered and analyzed sequentially, which resembles how the algorithm would intake real-time data. The interval length is modular in the algorithm, so the accuracy of the VAD algorithm was tested at multiple interval lengths to determine the ideal value of the interval length. Percentage accuracy was defined as the absolute percent error of the acceleration data trial subtracted from 100%. The result of the VAD algorithm run on the audio is considered as the reference value for this comparison. After each interval was filtered and analyzed, the complete filtered signal was returned for analysis. The VAD algorithm output was an integer vector with values

greater than zero indicating voiced frames. The number of onsets of speech activity and total time marked as voiced was calculated. The same measures were extracted from the audio and acceleration data and compared, using audio as the accepted value, to determine accuracy. The number of onsets of speech activity and total time voiced calculated, for both accelerometer and audio signals, were the main measures for VAD algorithm accuracy. Percentage accuracy for accelerometer data were calculated as the absolute percent error, in comparison to the audio data, subtracted from 100%. (see FIGURE 4).

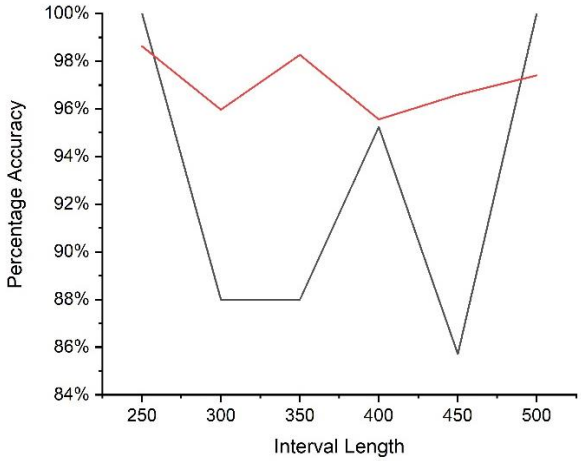

FIGURE 4: Accuracy of VAD Algorithm in comparison to audio data at multiple interval lengths. Black line is number of onsets and red line is total time voiced. Percentage Accuracy was calculated as absolute percent error subtracted from 100%. 250 milliseconds was chosen as the lowest tested interval length to accommodate the 240 sample minimum for using the MATLAB *filtfilt* function. The percentage accuracy in total time voiced is whithin 2% change for different interval length.

As shown in Figure 4, 250 milliseconds interval was chosen as the shortest interval with a similarly high accuracy to other intervals tested. The length of this interval is the expected duration for the delay that will be introduced in the real-time implementation of the algorithm. The intervals were then divided into 50 millisecond subintervals and the average power of each subinterval was calculated in both time and frequency domains. These two values were then compared to the threshold, and if any two contiguous subintervals were greater than the threshold, the entire interval (250 ms) was considered voiced (see FIGURE 5).

Due to the magnitude shift of the final signal caused by the filtering of the VTS trials, the threshold needed to be scaled up. Multiple scale factors were tested to find the scale factor with the highest accuracy. This was determined by comparing audio and acceleration data for the number of voiced onsets detected by the VAD algorithm. This scale factor was calibrated to each trial to maximize agreement of number of onsets between audio and acceleration data.

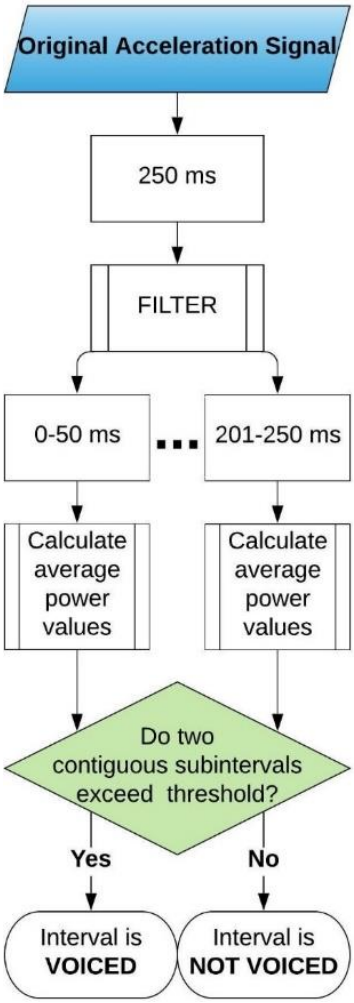

FIGURE 5: VAD Algorithm: The acceleration signal was analyzed in 250 millisecond intervals, where each interval was divided into 5 subintervals of 50 milliseconds. If two contiguous subintervals had an average power greater than the threshold, the entire interval was considered voiced.

RESULTS

When comparing the different accelerometer attachment locations in non-VTS trials, the acceleration signals decrease in overall amplitude from the thyroid cartilage to C7 positions. More importantly, the amplitude difference between speech and non-speech signals also decreases (see FIGURE 6). This decrease in differentiability of speech and non-speech acceleration signals is reflected by accuracy measurements. In non-VTS trials, the thyroid cartilage, on average, has greater than 90% accuracy in both number of onsets and total time voiced (see TABLE 1). In comparison to sternocleidomastoid and C7 positions, the thyroid cartilage has greater accuracy and lower variability, as seen in standard deviation comparison (see TABLE 1). The C7 position has 12% or lower accuracy in all measures and conditions.

Comparing non-VTS to VTS trials revealed that accuracy for total time voiced decreased at both the thyroid cartilage and sternocleidomastoid positions during VTS (see TABLE 1). The

percentage accuracy values for number of onsets slightly increased when adding VTS because the scale factor was adjusted to match this metric in the audio data. This also explains the lower accuracy in VTS trials for the total time voiced.

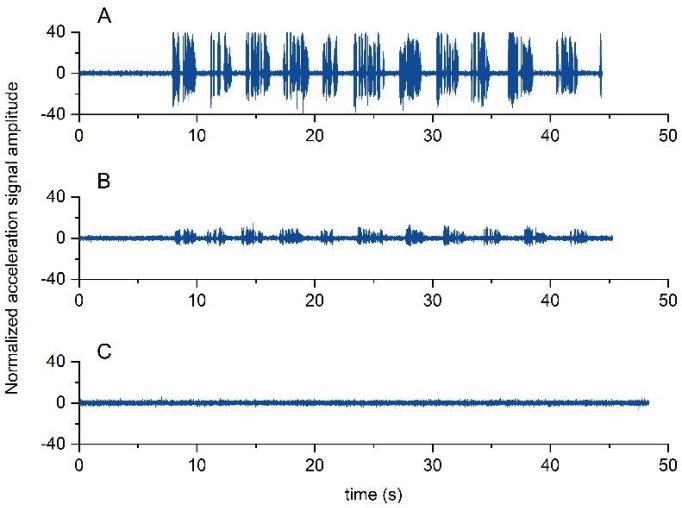

FIGURE 6: Comparison of acceleration signals, at each of three accelerometer attachment locations: A) thyroid cartilage, B) sternocleidomastoid, and C) C7. Each signal is a different trial without VTS. Y-axis scale is intentionally kept constant to emphasize amplitude differences between each signal.

TABLE 1: Mean percentage accuracy of VAD algorithm for acceleration data compared to audio data including standard error. Attachment locations: TC = Thyroid Cartilage; and Scm = Sternocleidomastoid.

Location	Number of Onsets	Total Time Voiced
No VTS, TC	91.67% ± 2.41%	94.89% ± 1.58%
No VTS, Scm	61.42% ± 12.39%	51.05% ± 12.39%
No VTS, C7	12.47% ± 6.70%	5.22% ± 2.92%
VTS, TC	94.88% ± 2.13%	62.03% ± 7.96%
VTS, Scm	73.31% ± 7.97%	52.17% ± 12.00%
VTS, C7	12.28% ± 7.88%	4.66% ± 2.90%

DISCUSSION

3.1 Attachment Location

The first aim of this study was to determine which of three accelerometer attachment locations would provide the highest accuracy signal for the VAD algorithm from neck surface vibrations amidst VTS applied to the laryngeal muscles. Based on the results of this study, the thyroid cartilage is the location that provides the acceleration signal with the highest accuracy when compared to the sternocleidomastoid and C7 positions. This can be attributed to the attenuation of the neck surface vibrations caused by speech as the distance from the larynx increases [12, 13]. The attenuation of signal amplitude and similarity of speech and non-speech signals causes the VAD algorithm to decrease in accuracy.

3.2 VAD Algorithm

The second aim of this study was to develop a VAD algorithm robust to VTS. The VAD algorithm shows high accuracy in the absence of VTS when provided with a strong signal at the thyroid cartilage. Unfortunately, the accuracy of the algorithm in calculating total time voiced at the thyroid cartilage decreases to roughly 60%, approaching the inaccuracy of the sterno-cleidomastoid location, when introducing VTS (see TABLE 1). The cause of this decrease in accuracy at the thyroid cartilage was the noise introduced to the acceleration signal by the VTS and the current to the vibrators. Adding VTS increased the overall amplitude of the signal, making it difficult to differentiate speech from non-speech activity.

While VTS introduced noise to the acceleration signals, the frequency of the vibrators themselves was easily filtered out. The noise due to the alternating current that powered the vibrators was more difficult to remove from the signal. The 60 Hz alternating current noise varied slightly as the vibrators' frequency was not perfectly consistent. This made filtering the 2nd, 3rd, and 4th harmonics, all of which had large energies relative to the speech signal, difficult. The VAD algorithm may not need to account for this alternating current noise, however, as the real-time application of this system in a collar would be powered by a direct current contained within the device, potentially eliminating the harmonic noise frequencies altogether. In the case of direct current, the VAD algorithm should be robust to the noise caused by introducing VTS and function at an overall higher accuracy than recorded in this study.

3.3 Implications for the Wearable Device

This study was conducted as part of a larger project with the goal of creating a wearable collar that delivers VTS therapy to SD patients as they speak . The findings of this study have important implications for the implementation of VAD in the collar. Regarding the location of the accelerometer attachment in the collar, the thyroid cartilage is the only viable location from those that has been tested. Even with VTS, the thyroid cartilage produced an acceleration signal amplitude from speech that was large enough to be distinguished from non-voice activity. The sternocleidomastoid was far too variable across participants to be feasible in a collar intended for a large population. C7 could be expected not to produce a usable signal for VAD in most SD patients. With a refined VAD algorithm, neck-surface vibrations from the thyroid cartilage should be effective in providing an accurate representation of voice activity.

The results of the VAD algorithm used in this study provide valuable insight for its use in the actual device. The modular aspects of the algorithm were the length of intervals and the scale factor used to account for magnitude distortion of acceleration signals after filtering. While modular aspects of the algorithm allow for flexibility in adjusting the accuracy of VAD, they require individual calibration to set at accurate values. Fortunately, both modular aspects could potentially be unnecessary in the collar. The need for a scale factor could be

removed if advanced signal processing, such as an adaptive filter, was implemented. The 250-millisecond interval length used in this study could also be maintained in the collar due to high accuracy of the VAD algorithm at this interval (see FIGURE 4). Based on these results, the VAD algorithm, if improved in signal processing, could easily be implemented in real-time for the wearable VTS collar.

CONCLUSION

This study showed that the thyroid cartilage was the best location for recording neck surface vibrations due to voicing. The VAD algorithm functioned with high accuracy at the thyroid cartilage without VTS, but in accuracy as measured *total time voiced* decreases upon the introduction of VTS. Fortunately, there is the potential to increase the accuracy of this algorithm by applying more sophisticated signal processing and use of a direct current in the collar. Application of adaptive filtering that adjusts to noise currently present in the environment and the user's characteristics would increase the accuracy of the VAD algorithm. This study aimed to determine the feasibility of utilizing an accelerometer for VAD so such sophisticated signal processing methods were not tested.

Further research should evaluate variations in voice such as changes in volume and patients with voice disorders. These new variables will require improvement of the VAD algorithm to accommodate a wider variety of speech signals.

REFERENCES

[1] Lindstrom, F., Ren, K., Li, H., and Waye, K. P., 2009, "Comparison of two methods of voice activity detection in field studies," J Speech Lang Hear Res, 52(6), pp. 1658-1663.

[2] Yiu, E. M., and Yip, P. P., 2016, "Effect of noise on vocal loudness and pitch in natural environments: an accelerometer (ambulatory phonation monitor) study," J Voice, 30(4), pp. 389-393.

[3] Matic, A., Osmani, V., and Mayora, O., "Speech activity detection using accelerometer," Proc. 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2112-2115.

[4] Cheyne, H. A., Hanson, H. M., Genereux, R. P., Stevens, K. N., and Hillman, R. E., 2003, "Development and testing of a portable vocal accumulator," J Speech Lang Hear Res, 46(6), pp. 1457-1467.

[5] Ludlow, C. L., Naunton, R. F., Sedory, S. E., Schulz, G. M., and Hallett, M., 1988, "Effects of botulinum toxin injections on speech in adductor spasmodic dysphonia," Neurology, 38(8), pp. 1220-1225.

[6] Hillman, R. E., Heaton, J. T., Masaki, A., Zeitels, S. M., and Cheyne, H. A., 2006, "Ambulatory monitoring of disordered voices," Ann Otol Rhinol Laryngol, 115(11), pp. 795-801.

[7] Mahnan, A., Konczak, J., and Faraji, S. A., "Wearable non-invasive neuromodulation device for the symptomatic treatment of the voice disorder spasmodic dysphonia," Proc. 2019 Design of Medical Devices Conference.

[8] Khosravani, S., Mahnan, A., Yeh, I. L., Aman, J. E., Watson, P. J., Zhang, Y., Goding, G., and Konczak, J., 2019, "Laryngeal vibration as a non-invasive neuromodulation therapy for spasmodic dysphonia," Scientific Reports, 9.

[9] Audacity Team, 2019, "Audacity(R): free audio editor and recorder."

[10] Stevens, K. N., 2000, Acoustic phonetics, MIT press, Cambridge, MA, pp. 261-266.

[11] Schafer, R. W., and Oppenheim, A. V., 2010, Discrete-time signal processing, Pearson, Upper Saddle River, NJ.

[12] Moser, H. M., and Oyer, H. J., 1958, "Relative intensities of sounds at various anatomical locations of the head and neck during phonation of the vowels," J Acoust Soc Am, 30(4), pp. 275-277.

[13] Munger, J. B., and Thomson, S. L., 2008, "Frequency response of the skin on the head and neck during production of selected speech sounds," J Acoust Soc Am, 124(6), pp. 4001-4012.

