Wireless neck-surface accelerometer and microphone on flex circuit with application to noise-robust monitoring of Lombard speech

Daryush D. Mehta[1,2], Patrick C. Chwalek[2], Thomas F. Quatieri[2], Laura J. Brattain[2]

[1]Ctr. for Laryngeal Surgery and Voice Rehab., Mass. General Hospital, Boston, MA, USA
[2]Bioengineering Systems and Technologies, MIT Lincoln Laboratory, Lexington, MA, USA

mehta.daryush@mgh.harvard.edu, [patrick.chwalek, quatieri, brattainl]@ll.mit.edu

Abstract

Ambulatory monitoring of real-world voice characteristics and behavior has the potential to provide important assessment of voice and speech disorders and psychological and emotional state. In this paper, we report on the novel development of a lightweight, wireless voice monitor that synchronously records dual-channel data from an acoustic microphone and a neck-surface accelerometer embedded on a flex circuit. In this paper, Lombard speech effects were investigated in pilot data from four adult speakers with normal vocal function who read a phonetically balanced paragraph in the presence of different ambient acoustic noise levels. Whereas the signal-to-noise ratio (SNR) of the microphone signal decreased in the presence of increasing ambient noise level, the SNR of the accelerometer sensor remained high. Lombard speech properties were thus robustly computed from the accelerometer signal and observed in all four speakers who exhibited increases in average estimates of sound pressure level (+2.3 dB), fundamental frequency (+21.4 Hz), and cepstral peak prominence (+1.3 dB) from quiet to loud ambient conditions. Future work calls for ambulatory data collection in naturalistic environments, where the microphone acts as a sound level meter and the accelerometer functions as a noise-robust voicing sensor to assess voice disorders, neurological conditions, and cognitive load.

Index Terms: ambulatory voice monitoring, voice analysis, accelerometer, microphone, wearable technology

1. Introduction

Ambulatory monitoring of voice and speech characteristics has the potential to provide valuable data for the diagnosis, treatment, and prevention of voice and speech disorders, neurological conditions affecting speech production, and the overall assessment of one's psychological and emotional state. Although persistent monitoring of physiological signals (speech, body movement, heart rate, etc.) is becoming more ubiquitous, in particularly in the form of wearable sensors [1], most off-the-shelf devices do not support plug-and-play of sensors nor allow for full access to raw data streams that is critical for post-processing and algorithmic development.

In this paper, we report on a novel wireless voice monitor that uses flexible circuit technology and consists of synchronized acoustic and non-acoustic (accelerometer-based) sensors. A key clinical application area is determining whether individuals are speaking in a loud voice in a quiet setting or in reaction to a noisy environment, thus exhibiting the Lombard effect. The system captures voice-related features that are important for speaker identification, noise reduction, and, most notably, for exploiting non-acoustic vocal signatures in real-world environments to provide long-duration monitoring and real-time biofeedback. The device is typically positioned on the anterior neck surface just above the collarbone to capture both acoustic and non-acoustic vocal signatures. Since naturalistic environments make it challenging to estimate many important voice characteristics in noisy conditions, recordings of neck-surface vibration have been the subject of ongoing investigation due to its robustness to acoustic environmental noise, low profile, and lack of speech intelligibility (alleviating confidentiality concerns) [2]. However, microphone recordings continue to be desirable to capture the airborne acoustic signal that can be analyzed to quantify speech features (e.g., formants) and the environment.

The current study provides pilot data to validate the ability of the developed wireless wearable module to accurately capture environmental noise levels using an acoustic microphone and quantify Lombard speech properties through noise-robust estimates of sound pressure level (SPL), fundamental frequency (f_o), and overall voice quality using the non-acoustic accelerometer sensor.

2. Background

2.1. Acoustic and non-acoustic sensors

Acoustic and non-acoustic vocal sensors have been applied previously to robustly characterize speech in the presence of various types of background noise by fusing features from multiple sensors, including a bone conduction microphone, radar sensor, and contact microphones [3]. To date, ambulatory voice monitoring technologies have consisted of wired systems that have focused on the real-time computation of SPL and f_o. For example, data from the Ambulatory Phonation Monitor [4], NCVS voice dosimeter [5], and VoxLog monitor [6] yield frame-based estimates of SPL and f_o for voiced segments. In addition, the VocaLog monitoring device [7] records real-time estimates of SPL only. Recent work takes advantage of a smartphone platform to record raw sensor signals, develop novel voice parameters, and communicate with smartwatches [8]. These technologies have

---

DISTRIBUTION STATEMENT A. Approved for Public Release. Distribution Unlimited.
This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering. This work was also supported by the MIT McGovern Institute Neurotechnology Program.

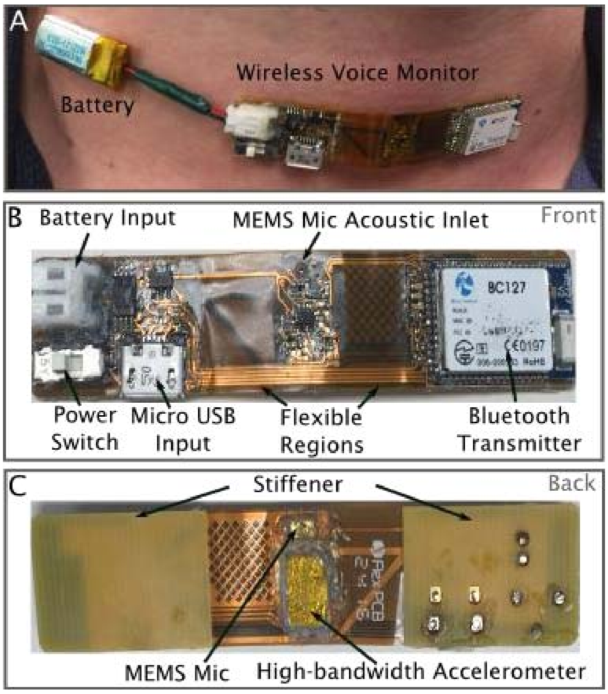

Figure 1: *Wireless voice monitor, showing (A) transmitting device on the anterior neck surface, (B) components on the front layer of the flexible circuit, and (C) components on the back circuit layer.*

allowed for the computation of parameters that have been associated with heavy voice use (increased talk time, inappropriate pitch and loudness, etc.) [9] in order to modify speaker behavior through real-time biofeedback.

All of the above technologies incorporate acoustic and/or non-acoustic sensors on the anterior neck surface that are wired to a central processing unit typically placed in a pocket or belt holster. The work reported here expands on knowledge gained from these previous studies to design a wireless module built on a flexible circuit substrate that conforms comfortably to the neck surface and aids in subject compliance and alleviates the cumbersome cable management needed for wired devices. An early prototype of the wireless module was initially reported by our group as part of a comprehensive multimodal system for animal behavior monitoring [10]; the module has since undergone significant improvements and optimization for human speech analysis for the current study.

2.2. The Lombard effect

The Lombard effect is an "involuntary vocal response by speakers to the presence of background noise" [11]. Lombard speech generally refers to several modifications to vocal characteristics (e.g., changes in loudness, pitch, and spectral tilt) in the context of elevated background noise levels. These effects have been observed in controlled laboratory settings and naturalistic settings that are known to induce Lombard speech [6, 12-15]. Characterizing background noise levels and Lombard speech in naturalistic environments is of critical importance to clinical voice assessment using ambulatory monitoring technology. Many common voice disorders are chronic conditions associated with inefficient patterns of vocal behavior referred to as vocal hyperfunction, and patients diagnosed with vocal hyperfunction may exacerbate their condition with Lombard speech effects such as loud voice production [13]. For example, knowing whether individuals are naturally speaking loudly in quiet environments or projecting their voice in a noisier context could help guide clinical treatment paradigms.

3. Materials and Methods

3.1. Hardware design

Figure 1 illustrates the wireless Bluetooth wearable transmitter module that is small, lightweight, and built on a flexible circuit board to conform easily to body surfaces. In addition to housing an acoustic microphone (MIC), the module has a piezoelectric accelerometer (ACC) for recording vocal vibration signals. Real-time wireless communication at minimal power consumption can be performed with either a Bluetooth-enabled smartphone or dedicated receiver module. Contrary to existing systems, the current system provides full control over hardware and software with access to the raw data streams, which can be subsequently processed. Most importantly, synchronized multimodal sensor recording on one unified wearable device greatly facilitates effort in characterizing speaker behaviors spanning voice and speech production and interaction with real environmental contexts.

Table 1 summarizes specifications of the wireless voice monitoring system. The circuit includes a single-axis, high-bandwidth ACC (BU-27135, Knowles Electronics, Itasca, IL) placed just above the collarbone and attached to the neck skin. An omnidirectional MEMS MIC (SPA2410LR5H-B, Knowles Electronics) is housed adjacent to the ACC for audio recording. Both sensors pass their output signals through a preamplifier to boost signals prior to being input individually to separate channels within a Bluetooth transceiver module. The preamplifiers allow for the individual tuning of gain settings for each sensor and additional noise filtering.

Sensors and active circuit components are powered by a single-cell, rechargeable, lithium-ion polymer battery that can be charged through a micro-USB input on the circuit. The micro-USB input also allows for communication to the Bluetooth module to modify firmware settings (e.g., gain settings) and troubleshoot. Additional features include electrostatic discharge protection, an on/off switch for the battery, status LEDs, and a logic switch that enables the `Bluetooth module to be fully functional when simultaneously powered via USB and battery.

The system contains a small receiver that is equipped with

Table 1: *System features and specifications for the wireless voice monitor recording two synchronized channels from accelerometer (ACC) and microphone (MIC) sensors.*

Feature	Specification
Sample rate	44.1 kHz (per channel)
Resolution	16 bits
Bandwidth	ACC: 0–5 kHz, MIC: 0–15 kHz
Power	50 mW (transmit), 18 mW (idle)
Battery life	Up to 8 hours (110 mAh)
Weight	4.0 g (12.5 g with battery)
Size	Monitor: 68 mm × 14.5 mm × 5 mm
	Receiver: 59 mm × 25 mm × 10 mm
Wireless version	Bluetooth 4.0

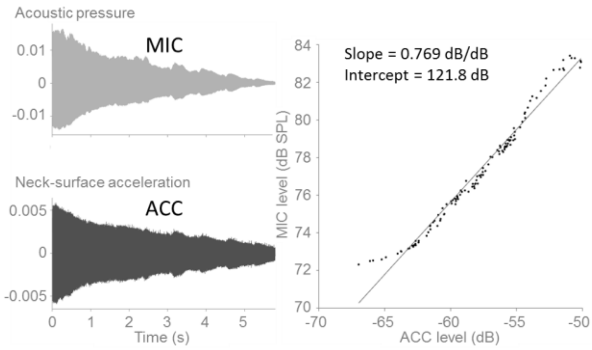

Figure 2: *Illustration of the participant-specific calibration of accelerometer (ACC) signal level to sound pressure level from microphone (MIC) signal using the loud-to-soft /a/ vowel task.*

the same Bluetooth module as the transmitter (BC127, BlueCreation, Cambridge, UK). The Advanced Audio Distribution Profile (A2DP) is utilized for stereo audio compression over Bluetooth. Play/Pause control is controlled using the Audio/Video Remote Control Profile (AVRCP). An onboard microcontroller receives the digital stereo data from the receiver over an Integrated Interchip Sound (I2S) serial bus interface. Signal streams received can be saved directly to a micro Secure Digital (SD) memory card or routed to a computer in real-time via a USB connection to the receiver.

3.2. Participants and speech tasks

Four adult participants (two male, two female) wore the wireless voice monitor inside an acoustically treated sound booth that contained loudspeakers that allowed for the simulation ambient acoustic stimuli at varying calibrated sound pressure levels. The flexible circuit was affixed using double-sided tape on the neck skin below the Adam's apple and above the collarbone. Each participant was instructed to stand in the middle of the sound booth and perform the following speech tasks: 1) produce an /a/ vowel starting at a loud intensity and gradually decreasing intensity to a soft level and 2) read aloud the first paragraph of the phonetically balanced Rainbow Passage [16]. This protocol was repeated in the presence of four different levels of the same background noise stimulus, which was an environmental recording of a helicopter with spinning rotors. Quiet, mild, moderate, and loud stimulus levels were produced at 26, 43, 54, and 66 dBA, respectively. The protocol and consent form were approved by the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology.

3.3. Data analysis

The MIC signal was calibrated to units of dB SPL using sustained vowels produced at multiple loudness levels [17], where the reference SPL was measured using a Class 2 sound level meter (CR:172B; Cirrus Research plc, Hunmanby, North Yorkshire, UK). For each participant, the ACC signal was also calibrated in terms of dB SPL units by comparing the synchronized MIC and ACC amplitudes in the loud-to-soft vowel task [18]. Figure 2 illustrates the linear regression equation mapping ACC amplitude (in arbitrary dB units) to dB SPL using the loudness sweep of the vowel task. Estimates of MIC- and ACC-derived SPL were thus obtained on a frame-by-frame basis (50 ms duration, 50% overlap) for voiced segments during the Rainbow Passage. The voice activity

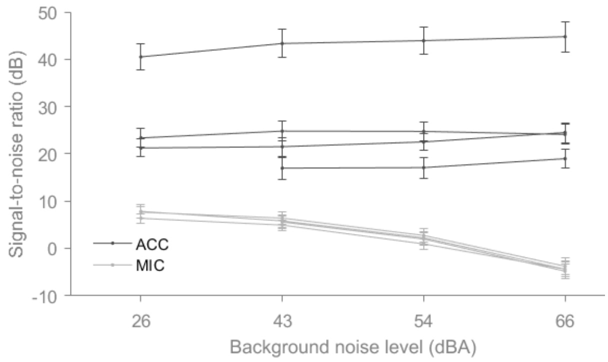

Figure 3: *Comparison of signal-to-noise ratio of the ACC and MIC signals for the Rainbow Passage produced by four speakers in context of four background noise levels. Error bars are ± 1 std. dev.*

detection algorithm followed previous work that incorporated ACC-based features of SPL, f_o, autocorrelation peak amplitude, and low-to-high spectral energy ratio to classify frames as voiced or unvoiced/silence [19]. These ACC-based voice activity decisions were then translated to the time-aligned MIC signal that was shifted earlier in time to compensate for the ACC-to-MIC acoustic propagation time. Voice activity decisions were not made directly from the MIC signal due to signal distortion at the louder stimulus levels.

Each voiced frame (50 ms duration, 50% overlap) in the MIC and ACC domains was analyzed to yield the following measures per participant. The signal-to-noise ratio (SNR) for each sensor was computed as the mean of the frame-level ratios of signal power to background noise level over all voiced frames of the Rainbow Passage. The f_o of each voiced frame was computed using the time-domain autocorrelation method in Praat (pitch floor = 60 Hz, pitch ceiling = 600 Hz, octave cost = 0.01, octave jump cost = 0.35) [20]. Overall voice quality was estimated using the cepstral peak prominence (CPP) measure, which was defined as the difference, in dB, between the magnitude of the highest peak in the power cepstrum and the noise floor for quefrencies greater than 2 ms [21]. Since MIC- and ACC-derived CPP have been previously shown to correlate highly when computed from sustained vowels [22], the current study extends the analysis to MIC-ACC CPP relationships during continuous speech production.

4. Results

Figure 3 illustrates the noise-robustness of the ACC signal relative to the MIC signal for the four speakers (one data point for a male speaker was not available for the 26 dBA noise level). As expected, the ACC-based SNR remained stable across all background noise levels when compared with the decreasing values for MIC-based SNR. This result, coupled with the ACC-MIC level mapping (Figure 2), suggests that estimates of voice SPL may be better obtained using the ACC signal as compared to the MIC signal in naturalistic environments that exhibit varying levels of background acoustic noise. In those scenarios, the participant-specific mapping of Figure 2 would need to be obtained in a quiet setting so that the mapping could then be applied to ACC signal levels in other, potentially noisier, settings. Figure 3 indicates that, although robust to acoustic noise, ACC-based SNR can vary according to anatomy (neck muscle tissue, etc.).

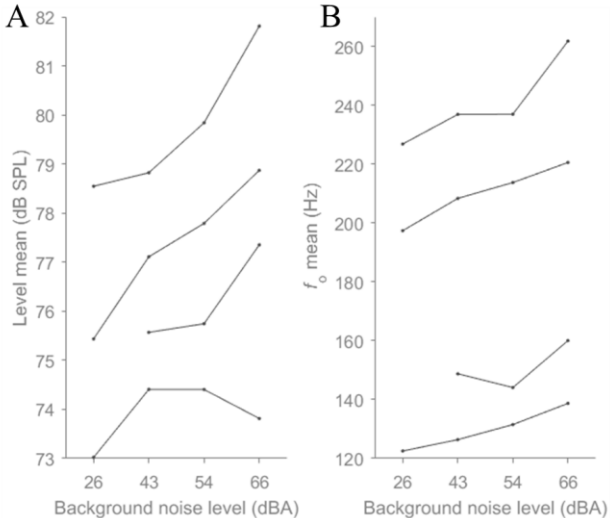

Figure 4: *Lombard speech effect exhibited by the four speakers as measured by ACC-based estimates (A) mean sound pressure level and (B) mean f_o for four increasing levels of background noise.*

Figure 4 shows that, overall, all four speakers exhibited Lombard speech effects related to ACC-based SPL and f_o features. Specifically, mean SPL and mean f_o measures across the Rainbow Passage increased with increasing background noise level. The mean increase in SPL from quiet to loud noise condition was 0.8 dB and 3.4 dB for the two female speakers, and 1.8 dB and 3.3 dB for the two male speakers. The increase in mean f_o from the quiet to loud noise condition was 34.9 Hz and 23.2 Hz for the two female speakers, and 11.3 Hz and 16.2 Hz for the two male speakers. Correlations were very high ($r > 0.99$, $p < 0.001$) when comparing f_o for the same frame times in the ACC and MIC signals, validating the expected high accuracy of ACC-based f_o estimation.

Figure 5 shows the ability of the MIC-based measure of CPP to be captured by the ACC signal in one of the female participants. Figure 5A exemplifies the frame-by-frame relationship between ACC- and MIC-based CPP. In the quietest noise condition, Pearson's correlation coefficient for this relationship was 0.65 and 0.73 for the two female speakers, and 0.56 and 0.50 for the two male speakers. Since MIC-based CPP is known to be affected by both voice-related breathiness and environmental acoustic noise, Figure 5B shows the advantage of ACC-based estimates of mean CPP across the Rainbow Passage to help act as noise-robust measures of overall voice quality. The four speakers increased their mean ACC-based CPP by 2.1, 1.7, 1.3, and 0.1 dB from their quietest to loudest condition, indicating a potential Lombard speech effect that is challenging to quantify accurately using the MIC signal alone due to ambient noise corruption (mean MIC-based CPP decreased by 5.3, 5.2, 4.4, and 6.4 dB for the four speakers).

5. Conclusion and discussion

In this paper, we presented our work developing and implementing a new wireless voice monitor that uses Bluetooth technology and a wearable, flexible circuit. Synchronized data streaming from both acoustic MIC and neck-surface ACC sensors makes it feasible to compute

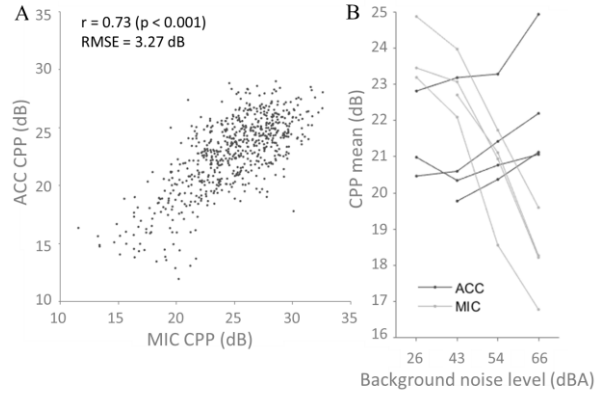

Figure 5: *Cepstral peak prominence (CPP) of Rainbow Passage. (A) Exemplary correlation between frame-level MIC- and ACC-based CPP in the quiet background condition; Pearson's r and root-mean-square error (RMSE) shown. (B) CPP for the four speakers at the four increasing levels of background noise, indicating that the Lombard effect can only be observed in the ACC signal.*

complementary acoustic and non-acoustic speech/voice features. The MIC also provides critical information related to environment noise levels that is important to collect in real-world conditions. The ACC sensor is more immune than the MIC to acoustic noise and can reveal noise-robust voice signatures, including voice SPL, voice quality, and vocal dose measures [9]. Characterizing Lombard speech—voice properties and a speaker's reaction to varying ambient noise levels—is enabled with synchronous ACC and MIC recording.

The ultimate use of the wireless voice monitor is the ambulatory tracking of everyday verbal communication as individual's go about their typical daily activities. Efforts to develop a custom wireless solution have been motivated by experience demonstrating that patient compliance improves when technology is easy to use and less cumbersome. Future device development can also take advantage of the modularity of the system to add additional sensors and real-time processing of voice features that can provide user biofeedback via mobile devices such as smartphones and smartwatches.

Future work calls for the study of additional features that have been associated with disordered voice production, including jitter/shimmer [22], glottal aerodynamic measures [19, 23], subglottal pressure [24], and vocal fry [25]. Computation of these features from both the ACC and MIC signals is warranted to better understand which voice-related features can be extracted accurately from the ACC waveform. Big data processing of daily vocal behavior in naturalistic conditions holds great potential to better understand the cause, progression, and treatment of various medical conditions. The work presented here on the development of a wireless voice monitor holds promise to enable the quantitative analysis of long-term, longitudinal voice data in real-world situations.

6. Acknowledgments

Special thanks to Prof. Robert Desimone, Prof. Guoping Feng, and Dr. Charles Jennings at the McGovern Institute for Brain Research at MIT, Dr. Rogier Landman at the MIT/Harvard Broad Institute, and Mr. Kerry Johnson, Mr. Tejash Patel, and Dr. Christopher Smalt at MIT Lincoln Laboratory for their generous support and help.

7. References

[1] Zheng, Y. L., Ding, X. R., Poon, C. C. Y., Lo, B. P. L., Zhang, H., Zhou, X. L., Yang, G. Z., Zhao, N., and Zhang, Y. T., "Unobtrusive sensing and wearable devices for health informatics," IEEE Trans. Biomed. Eng., 61(5):1538-1554, 2014.

[2] Zañartu, M., Ho, J. C., Kraman, S. S., Pasterkamp, H., Huber, J. E., and Wodicka, G. R., "Air-borne and tissue-borne sensitivities of bioacoustic sensors used on the skin surface," IEEE Trans. Biomed. Eng., 56(2):443-451, 2009.

[3] Quatieri, T. F., Brady, K., Messing, D., Campbell, J. P., Campbell, W. M., Brandstein, M. S., Weinstein, C. J., Tardelli, J. D., and Gatewood, P. D., "Exploiting nonacoustic sensors for speech encoding," IEEE Trans. Audio Speech Lang. Processing, 14(2):533-544, 2006.

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

[5] Popolo, P. S., Švec, J. G., and Titze, I. R., "Adaptation of a Pocket PC for use as a wearable voice dosimeter," J. Speech. Lang. Hear. Res., 48(4):780-791, 2005.

[6] Lindstrom, F., Waye, K. P., Södersten, M., McAllister, A., and Ternström, S., "Observations of the relationship between noise exposure and preschool teacher voice usage in day-care center environments," J. Voice, 25(2):166-172, 2011.

[7] Searl, J. and Dietsch, A., "Testing of the VocaLog vocal monitor," J. Voice, 28(4):523.e527-523.e537, 2014.

[8] Mehta, D. D., Zañartu, M., Feng, S. W., Cheyne II, H. A., and Hillman, R. E., "Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform," IEEE Trans. Biomed. Eng., 59(11):3090-3096, 2012.

[9] Titze, I. R. and Hunter, E. J., "Comparison of vocal vibration-dose measures for potential-damage risk criteria," J. Speech. Lang. Hear. Res., 58(5):1425-1439, 2015.

[10] Brattain, L. J., Landman, R., Johnson, K. A., Chwalek, P., Hyman, J., Sharma, J., Jennings, C., Desimone, R., Feng, G., and Quatieri, T. F., "A multimodal sensor system for automated marmoset behavioral analysis," Proceedings of IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2016.

[11] Zollinger, S. A. and Brumm, H., "The Lombard effect," Curr. Biol., 21(16):R614-R615, 2011.

[12] Bottalico, P., Graetzer, S., and Hunter, E. J., "Effect of training and level of external auditory feedback on the singing voice: Volume and quality," J. Voice, 30(4):434-442, 2016.

[13] Aronsson, C., Bohman, M., Ternström, S., and Södersten, M., "Loud voice during environmental noise exposure in patients with vocal nodules," Logopedics Phoniatrics Vocology, 32(2):60-70, 2007.

[14] Yiu, E. M. and Yip, P. P., "Effect of noise on vocal loudness and pitch in natural environments: An accelerometer (Ambulatory Phonation Monitor) study," J. Voice, 30(4):389-393, 2016.

[15] Hunter, E. J., Bottalico, P., Graetzer, S., Leishman, T. W., Berardi, M. L., Eyring, N. G., Jensen, Z. R., Rolins, M. K., and Whiting, J. K., "Teachers and teaching: Speech production accommodations due to changes in the acoustic environment," Energy Procedia, 78:3102-3107, 2015.

[16] Fairbanks, G., Voice and Articulation Drillbook, vol. 2. New York: Harper and Row, 1960.

[17] Maryn, Y. and Zarowski, A., "Calibration of clinical audio recording and analysis systems for sound intensity measurement," Am. J. Speech Lang. Pathol., 24(4):608-618, 2015.

[18] Švec, J. G., Titze, I. R., and Popolo, P. S., "Estimation of sound pressure levels of voiced speech from skin vibration of the neck," J. Acoust. Soc. Am., 117(3):1386-1394, 2005.

[19] Mehta, D. D., Van Stan, J. H., Zañartu, M., Ghassemi, M., Guttag, J. V., Espinoza, V. M., Cortés, J. P., Cheyne II, H. A., and Hillman, R. E., "Using ambulatory voice monitoring to investigate common voice disorders: Research update," Frontiers in Bioengineering and Biotechnology, 3(155):1-14, 2015.

[20] Boersma, P., "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," Proceedings of Institute of Phonetic Sciences, University of Amsterdam, 1993.

[21] Hillenbrand, J., Cleveland, R. A., and Erickson, R. L., "Acoustic correlates of breathy vocal quality," J. Speech Hear. Res., 37(4):769-778, 1994.

[22] Mehta, D., Van Stan, J., and Hillman, R., "Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer," IEEE/ACM Trans. Audio Speech Lang. Processing, 24(4):659-668, 2016.

[23] Zañartu, M., Ho, J. C., Mehta, D. D., Hillman, R. E., and Wodicka, G. R., "Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration," IEEE Trans. Audio Speech Lang. Processing, 21(9):1929-1939, 2013.

[24] Fryd, A. S., Van Stan, J. H., Hillman, R. E., and Mehta, D. D., "Estimating subglottal pressure from neck-surface acceleration during normal voice production," J. Speech. Lang. Hear. Res, 59:1335-1345, 2016.

[25] Kane, J., Drugman, T., and Gobl, C., "Improved automatic detection of creak," Comp. Speech Lang., 27(4):1028-1047, 2013.