Automatic speech and singing classification in ambulatory recordings for normal and disordered voices

Andrew J. Ortiz,[a] Laura E. Toles,[b] and Katherine L. Marks[b]
Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
aortiz15@mgh.harvard.edu, ltoles@mghihp.edu, kmarks@mghihp.edu

Silvia Capobianco,[c]
Universita di Pavia, Pavia, Italy
silvia.capobianco01@universitadipavia.it

Daryush D. Mehta,[c] Robert E. Hillman,[c] and Jarrad H. Van Stan[c]
Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
mehta.daryush@mgh.harvard.edu, hillman.robert@mgh.harvard.edu, jvanstan@mgh.harvard.edu

Abstract: Ambulatory voice monitoring is a promising tool for investigating phonotraumatic vocal hyperfunction (PVH), associated with the development of vocal fold lesions. Since many patients with PVH are professional vocalists, a classifier was developed to better understand phonatory mechanisms during speech and singing. Twenty singers with PVH and 20 matched healthy controls were monitored with a neck-surface accelerometer–based ambulatory voice monitor. An expert-labeled ground truth data set was used to train a logistic regression on 15 subject-pairs with fundamental frequency and autocorrelation peak amplitude as input features. Overall classification accuracy of 94.2% was achieved on the held-out test set.

[DDO]
Date Received: March 2, 2019 **Date Accepted:** June 17, 2019

1. Introduction

Phonotraumatic vocal hyperfunction (PVH) is defined as excessive and/or imbalanced muscular forces during phonation in the presence of benign lesions on the medial (contact) surfaces of the vocal folds (e.g., vocal fold nodules, polyps) (Mehta *et al.*, 2015). These lesions are assumed to be caused by, or associated with, pathological vocal behaviors in the patient's daily life; however, the role of habitual vocal behavior in voice-use-related disorders is not well understood. Currently, speech-language pathologists must rely on the patient's self-report concerning their own behavior outside of the clinic; which has been shown to be unreliable (Mehta *et al.*, 2016). Therefore, researchers have been developing ambulatory voice monitoring technology to objectively characterize habitual voice use outside of the clinic and to better understand the relationship between vocal behaviors and behaviorally based voice disorders (Titze and Hunter, 2015; Van Stan *et al.*, 2014). Most ambulatory voice monitoring technology uses a neck-placed miniature accelerometer to record voicing because an accelerometer is robust to environmental noises, speech is not recorded in the raw signal (confidentiality is maintained), speech and singing in the surrounding environment are not recorded, and an accelerometer can be worn underneath clothing such as a scarf or collar [Popolo (2005); for a review, see Hillman *et al.* (2006)].

Previous studies using data from weeklong, ambulatory recordings have failed to find differences in average measures of voice use between groups of patients with PVH and healthy matched controls (age, sex, and occupation) (Van Stan *et al.*, 2015).

[a]Author to whom correspondence should be addressed.
[b]Also at: MGH Institute of Health Professions, Boston, MA 02129, USA.
[c]Also at: Department of Surgery, Massachusetts General Hospital–Harvard Medical School, Boston, MA 02115, USA.

A majority of the subjects in these studies were singers, which is not surprising since there is an elevated risk of phonotrauma in individuals who sing professionally (Titze *et al.*, 1997). However, this led to the question of whether better differentiation between the pathological and normal groups could be attained by separately examining speech and singing; i.e., does combining speech and singing (as was done in these studies) mask differences that would be revealed by examining the two modes of phonation separately? Because the amount of data is so large (on average, 80 h of recording time per subject), there would need to be a method for automatically separating speech from singing in the ambulatory recordings.

An ongoing area of study in the field of music information retrieval is singing voice detection, in which segments of music recordings containing singing are identified and/or extracted (Humphrey *et al.*, 2019). Previous work has also focused on automatic discrimination of speech and singing from monophonic acoustic recordings using both mel-frequency cepstral coefficients (MFCCs) and pitch information (Tsai and Ma, 2014). However, to our knowledge, automatic classification of singing and speech from ambulatory neck-surface acceleration data has yet to be accomplished. Although singers could manually indicate singing times, computational analyses of a wearable sensor allow the user to be monitored passively without interrupting their natural daily behavior. Therefore, the purpose of this study is to develop and test a computationally efficient classification algorithm based on a simple decision tree whose nodes utilize a logistic regression and phrase-based reassignment step.

2. Methods

Forty female subjects were included in the analysis who self-identified as professional vocalists, college students majoring in vocal performance, or amateur singers with a significant background in musical performance. Twenty of the subjects were diagnosed with vocal fold nodules and were recruited through sequential convenience sampling at the Massachusetts General Hospital–Center for Laryngeal Surgery and Vocal Rehabilitation (MGH Voice Center). Snowball sampling was used to recruit the remaining twenty (vocally normal) control subjects, who were each matched to a corresponding patient according to approximate age (Mean: 22.6 years, SD: 6.7 years) and singing genre. Only female participants were selected for this study to provide a homogeneous sample of a group that has a significantly higher incidence of vocal fold nodules (Herrington-Hall *et al.*, 1988) and comparable values of fundamental frequency (f_o). Diagnoses were based on a comprehensive team evaluation (laryngologist and speech-language pathologist) at the MGH Voice Center. The normal vocal status of all control participants was verified via interview and videostroboscopic imaging of the larynx.

Figure 1 shows the smartphone-based ambulatory voice monitor (Mehta *et al.*, 2012), which incorporated a high-bandwidth accelerometer (BU-27135; Knowles Corp., Itasca, IL) positioned on the anterior neck surface to assess the voice use of each participant for one week. Each week of subject data typically contains approximately 12 h per day of ambulatory data for each of the 7 recording days, amounting to an average of 8 h of voiced data per subject-week.

2.1 Signal analysis

The daily recordings (raw neck-surface acceleration waveforms) for all subjects were divided into non-overlapping 50-ms frames, and each frame was considered voiced if four features passed the following thresholds: (a) vocal sound pressure level was greater than

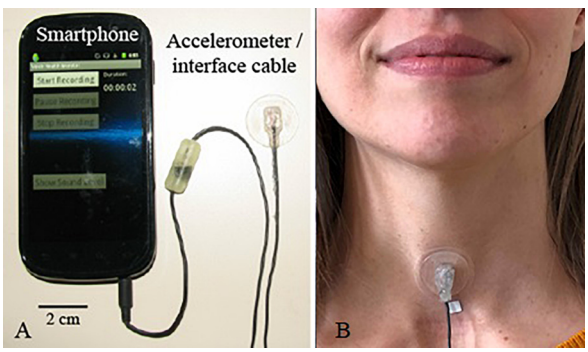

Fig. 1. (Color online) Ambulatory voice monitor: (A) Smartphone, accelerometer sensor, and interface cable with circuit encased in epoxy and (B) wired accelerometer mounted on a silicone pad affixed to the anterior neck surface midway between the thyroid prominence and the suprasternal notch.

45 dB sound pressure level (SPL) (the accelerometer was calibrated relative to dB SPL at 15 cm from the lips using a linear regression over a loudness glide) (Mehta *et al.*, 2012), (b) the first (non zero-lag) peak in the normalized autocorrelation exceeded an amplitude threshold of 0.6, (c) f_o (reciprocal of the time lag of that autocorrelation peak) was between 70 and 1000 Hz, and (d) the ratio of low- to high-frequency energy (boundary frequency of 2000 Hz) exceeded 22 dB. When a 50-ms frame was considered voiced, all of the four aforementioned features were saved and all f_o data were transformed into semitones, where the reference frequency was the individual subject's weekly f_o mode. Contiguous voiced and unvoiced frames were subsequently grouped into phrase groups if unvoiced intervals between successive voiced segments were less than 0.5 s in duration (Mehta *et al.*, 2015).

2.2 Expert labeling

A two-step process was used to identify time periods of pure speech and singing to serve as ground truth. First, one experimenter with a professional singing background listened to the ambulatory recordings and extracted approximately two minutes of singing and two minutes of speech (~4800 voiced frames) from the weeklong recordings of each subject. The following general guidelines were followed to select segments of singing and speech from each subject's weeklong recordings. Using a custom graphical user interface, the experimenter could visualize the percentage of voicing over sliding five-minute windows during each daily recording. A high percentage of voicing was indicative of regions containing either heavy voice use or singing. The interface then allowed the experimenter to zoom in on the high-voice-use segments and listen to an audio playback of the signal. Even though the accelerometer recordings are unintelligible, enough information is retained (e.g., pitch characteristics) to differentiate singing and speech based on listening to the signals. The goal was to extract approximately two minutes of contiguous singing (uninterrupted by speech) and two minutes of contiguous speech for each subject. Figure 2 displays example accelerometer segments containing singing and speech segments for a healthy control subject (see Mm. 1 and Mm. 2 for longer media files).

Mm. 1. Example of a singing segment from the ambulatory voice recording of a subject with no history of voice disorders. The signal is from an accelerometer sensor placed on the anterior neck surface and is sampled at 11025 Hz. This is a file of type "wav" (4473 KB).

Mm. 2. Example of a speech segment from the ambulatory voice recording of a subject with no history of voice disorders. The signal is from an accelerometer sensor placed on the anterior neck surface and is sampled at 11025 Hz. This is a file of type "wav" (8642 KB).

In the second step, two speech-language pathologists specializing in voice disorders and who had professional singing experience independently listened to all

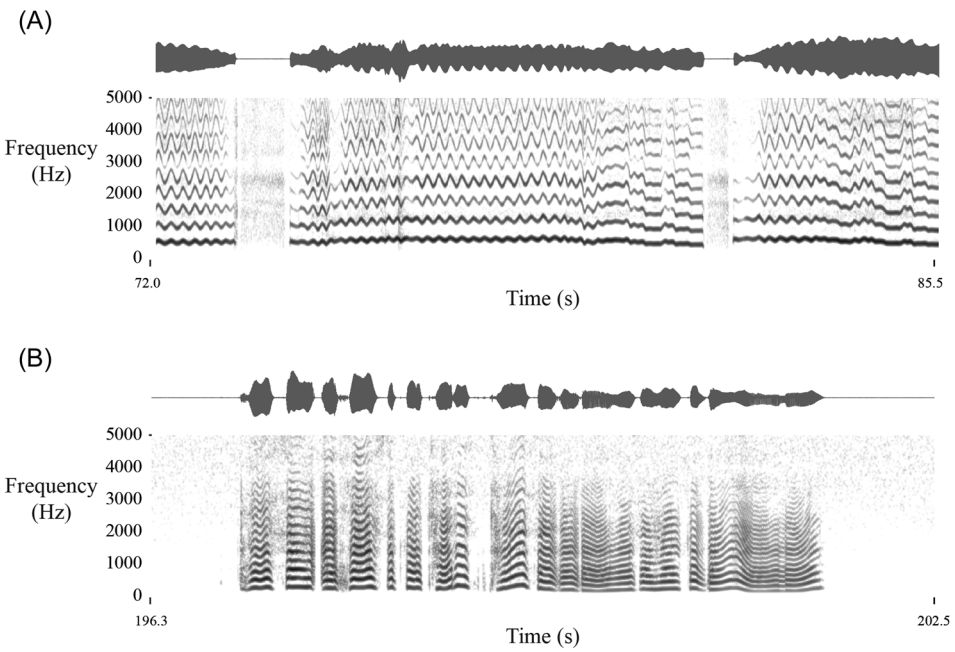

Fig. 2. Example waveforms and spectrograms for accelerometer segments from a healthy control subject containing (A) singing and (B) speech. See Mm. 1 and Mm. 2 for longer media files.

extracted voice samples. If one of the two independent raters identified any voicing as not obviously speech or singing, the vocalization was deleted. Furthermore, all obvious instances/segments of coughing, throat clearing, burping, laughing, crying, and audible swallowing were manually removed.

2.3 Classification algorithm

The first step of the classification algorithm used a logistic regression on two singing-related features computed for each 50-ms voiced frame. The two features were (1) the first non–zero-lag autocorrelation peak amplitude (normalized by the zero-lag amplitude) and (2) f_o in semitones (with reference to each subject's weeklong f_o mode). These two features were hypothesized to capture singing characteristics related to enhanced periodicity/resonance and elevated pitch. Each voiced frame was initially classified as singing if the predicted probability of the logistic regression was greater than 0.5; the voiced frame was classified as speech if the predicted probability was less than 0.5. The logistic regression was trained on a training set of 15 patient-control pairs (30 subjects). Figure 3 shows the distributions of normalized autocorrelation peak and for each class in the training data.

The second step of the classification algorithm consisted of a reassignment of labels for each voiced frame based on a majority rule within each phrase group. Thus, all the voiced frames within each phrase group were re-classified as singing or speech according to the percentage of voiced frames initially classified as singing or speech, respectively, within the phrase group. To determine the appropriate singing percentage threshold on the training set, a cost function was applied that penalized false positives and false negatives with weights of 0.75 and 0.25, respectively (future work could explore alternative weighting schemes). The optimal point on the receiver operating characteristic (ROC) curve for this cost function was calculated using the *perfcurve* function in MATLAB 2017a (The MathWorks, Inc., Natick, MA). The imbalanced penalization was desired for when the algorithm would be used in practice. Since, in daily life, subjects typically spend significantly less time singing than speaking, speech frames misclassified as singing could have severe confounding effects when analyzing long-term, ambulatory data. Therefore, a classifier that weighted the false positive and false negative rates equally was not desired.

The resulting logistic regression equation and singing percentage threshold for phrase-based label reassignment were then applied on a held-out test set of the remaining five patient-control pairs (ten subjects).

3. Results

The first classification step on the training set resulted in the following frame-based logistic regression:

$$y = \frac{1}{1 + e^{-(14.42*\mathrm{NP}+0.31*f_o-14.2)}}, \tag{1}$$

where NP is the normalized autocorrelation peak amplitude and f_o is the fundamental frequency (in semitones) for each voiced frame. For the second, phrase-based frame reassignment step, the optimal point on the initial ROC curve using the weighted cost function resulted in a singing percentage threshold of 57%; i.e., voiced frames within

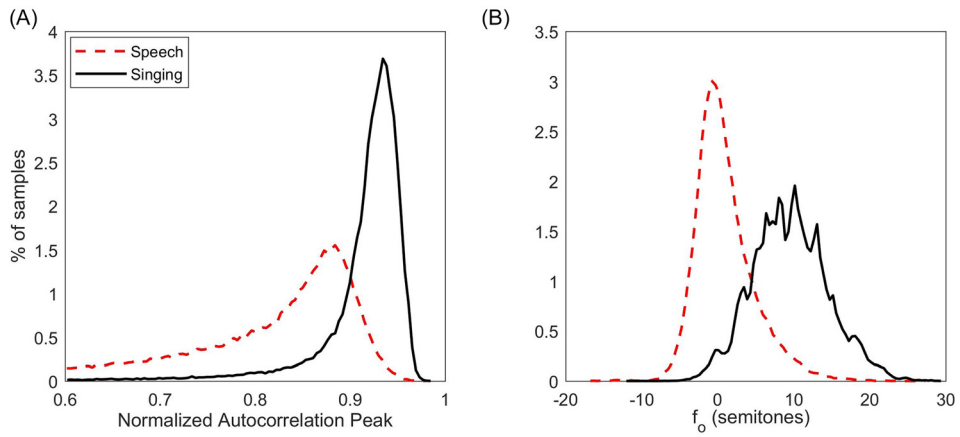

Fig. 3. (Color online) Distributions of (A) normalized autocorrelation peak and (B) f_o for ground-truth labeled singing and speech in training set.

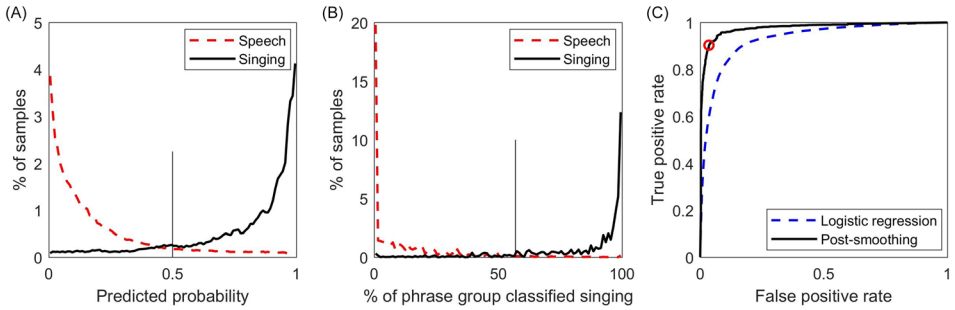

Fig. 4. (Color online) (A) Predicted probability plot for the first logistic regression step on the training set, with 0.5 cutoff value indicated (vertical line). (B) Percentage of frames within a phrase group classified as singing by logistic regression, with optimal 57% threshold indicated (vertical line). (C) Receiver operating characteristic curves for the first logistic regression step and second phrase-group reassignment step, with operating point corresponding to the 57% threshold indicated (open circle).

an entire phrase were all re-classified as singing if at least 57% of the voiced frames were initially classified as singing, otherwise the frames were re-classified as speech).

Figure 4 shows the predicted probability of the logistic regression, the distribution of frames per phrase group classified as singing, and the ROC curves for both the logistic regression and reassignment steps.

Table 1 provides a summary of the results in terms of confusion matrices. For the training set, the first classification step (top panel of Table 1) resulted in a frame-based overall accuracy of 86.5%, sensitivity of 85.4% (correct classification of singing), and specificity of 87.5% (correct classification of speech). The second classification step (middle panel of Table 1) applied a 57% singing percentage threshold for phrase groups. This reassignment step resulted in a significantly improved overall accuracy of 93.3%, sensitivity of 90.3%, and specificity of 96.4%. Finally, applying both steps to the held-out test set (bottom panel of Table 1) resulted in an overall accuracy of 94.2%, sensitivity of 93.5%, and specificity of 95.0%.

4. Discussion

The results of the singing detector developed in this study are encouraging and provide evidence that an automated algorithm can perform well on ambulatory voice monitoring data. It is acknowledged that the algorithm was optimized on a limited subset of all the data available from study subjects. The experimental paradigm necessarily consisted of the manual selection of singing and speech segments from thousands of hours of recorded data to balance the need for a ground-truth database and the labor-intensive nature of the expert-labeling process. A more comprehensive analysis is outside of the scope of the current methodologically oriented study. Additional classifier architectures were also considered, including support vector machines (SVMs) and neural networks (NNs), but were not pursued due to the strong performance of the reported logistic regression. An effective logistic regression classifier with such a minimal feature set is also well-suited for a clinical research setting, where human-interpretability of classification decision rules may be desired.

As is often the case with signal processing, algorithms (e.g., for voice activity detection) are created with parameter settings that can be modified to suit given hardware configurations and experimental contexts. The particular parameter settings

Table 1. Singing/speech confusion matrices showing the percentage of correctly classified frames (bold) for the training set (before and after applying a phrase group-based reassignment of predictions based on a majority rule) and test set. Percentages are based on the total number of voiced frames in the data set. Therefore, perfect classification would be 50% per class instead of 100%.

	Logistic regression (training set)		Post-reassignment (training set)		Post-reassignment (test set)	
Expert label	Speech	Singing	Speech	Singing	Speech	Singing
Speech frames (% total frames)	**88 304 (43.2%)**	12 572 (6.2%)	**97 218 (47.6%)**	3 658 (1.8%)	**23 596 (45.6%)**	1 253 (2.4%)
Singing frames (% total frames)	15 103 (7.4%)	**88 318 (43.2%)**	10 011 (4.9%)	**93 410 (45.7%)**	1 759 (3.4%)	**25 175 (48.6%)**

chosen for voice activity detection are tuned toward the detection of nearly periodic vocalizations because of the ambulatory nature of the neck-skin acceleration recordings; i.e., to maximize identification of voiced analysis frames and minimize misclassification of noise or non-voicing as "voiced." Since voiced segments exhibiting high levels of dysphonia or singing growls/screams cannot be identified in ambulatory recordings without significantly increasing the amount of misclassifications, including phonation with higher degrees of aperiodicity will require different voice activity detection settings and a different type of singing detector. Results warrant future analyses of repeatability and robustness of the singing detector across other ambulatory voice monitoring devices, singing styles with high degrees of aperiodicity, and longer time periods of unstructured ambulatory recordings.

5. Conclusion

The present classification scheme results in highly accurate detection of singing from ambulatory neck-surface acceleration recordings, providing a promising clinical and research tool. Although extracting estimates of normalized autocorrelation peak and fundamental frequency at the frame-based level performed well initially, applying a phrase-based reassignment based on majority rule yielded a significant improvement in singing and speech classification. Future work can apply the singing detector to determine if differences in average voice measures and vocal load between patients with PVH and matched healthy controls can be detected by separately examining singing (especially given different singing genres) and speech segments. The singing detector may also have uses in real-time biofeedback applications where the goal might be to target vocal behaviors that are specifically associated with either singing or speech.

Acknowledgments

This work was supported by the Voice Health Institute and the National Institutes of Health (NIH), National Institute on Deafness and Other Communication Disorders (NIDCD) under Grant Nos. R33 DC011588 and P50 DC015446. The article's contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. Additional support was granted to S.C. by means of an Armenise-Ghislieri Summer Fellowship.

References and links

Herrington-Hall, B. L., Lee, L., Stemple, J. C., Niemi, K. R., and McHone, M. M. (**1988**). "Description of laryngeal pathologies by age, sex, and occupation in a treatment-seeking sample," J. Speech Hear. Disord. **53**(1), 57–64.

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), 795–801.

Humphrey, E. J., Reddy, S., Seetharaman, P., Kumar, A., Bittner, R., Demetriou, A., Gulati, S., Jansson, A., Jehan, T., Lehner, B., Krupse, A., and Yang, L. (**2019**). "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," IEEE Sign. Process. Mag. **36**, 82–94.

Mehta, D. D., Cheyne, H. A., Wehner, A., Heaton, J. T., and Hillman, R. E. (**2016**). "Accuracy of self-reported estimates of daily voice use in adults with normal and disordered voices," Am. J. Speech-Lang. Pathol. **25**(4), 634–641.

Mehta, D. D., Van Stan, J. H., Zañartu, M., Ghassemi, M., Guttag, J. V., Espinoza, V. M., Cortés J. P., Cheyne, H. A., 2nd, and Hillman, R. E. (**2015**). "Using ambulatory voice monitoring to investigate common voice disorders: Research update," Front. Bioeng. Biotechnol. **3**, 155.

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

Popolo, P. S., Svec, J. G., and Titze, I. R. (**2005**). "Adaptation of a pocket PC for use as a wearable voice dosimeter," J. Speech Lang. Hear. Res. **48**(4), 780–791.

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

Titze, I. R., Lemke, J., and Montequin, D. (**1997**). "Populations in the U.S. workforce who rely on voice as a primary tool of trade: A preliminary report," J. Voice **11**(3), 254–259.

Tsai, W.-H., and Ma, C.-H. (**2014**). "Speech and singing discrimination for audio data indexing," in *2014 IEEE International Congress on Big Data*.

Van Stan, J. H., Gustafsson, J., Schalling, E., and Hillman, R. E. (**2014**). "Direct comparison of three commercially available devices for voice ambulatory monitoring and biofeedback," Perspect. Voice Voice Disord. **24**(2), 80–86.

Van Stan, J. H., Mehta, D. D., Zeitels, S. M., Burns, J. A., Barbu, A. M., and Hillman, R. E. (**2015**). "Average ambulatory measures of sound pressure level, fundamental frequency, and vocal dose do not differ between adult females with phonotraumatic lesions and matched control subjects," Ann. Otol. Rhinol. Laryngol. **124**(11), 864–874.