Original Paper

# Ambulatory Phonation Monitoring With Wireless Microphones Based on the Speech Energy Envelope: Algorithm Development and Validation

Chi-Te Wang[1,2,3], MD, PhD; Ji-Yan Han[4], BSc; Shih-Hau Fang[2,5], PhD; Ying-Hui Lai[4], PhD

[1]Department of Otolaryngology Head and Neck Surgery, Far Eastern Memorial Hospital, New Taipei City, Taiwan

[2]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan

[3]Department of Special Education, University of Taipei, Taipei, Taiwan

[4]Department of Biomedical Engineering, National Yang-Ming University, Taipei, Taiwan

[5]Ministry of Science and Technology Joint Research Center for Artificial Intelligence Technology and All Vista Healthcare, Taoyuan, Taiwan

**Corresponding Author:**
Ying-Hui Lai, PhD
Department of Biomedical Engineering
National Yang-Ming University
No155, Sec 2, Linong Street
Taipei, 112
Taiwan
Phone: 886 228267021
Fax: 886 228210847
Email: yh.lai@gm.ym.edu.tw

## Abstract

**Background:** Voice disorders mainly result from chronic overuse or abuse, particularly in occupational voice users such as teachers. Previous studies proposed a contact microphone attached to the anterior neck for ambulatory voice monitoring; however, the inconvenience associated with taping and wiring, along with the lack of real-time processing, has limited its clinical application.

**Objective:** This study aims to (1) propose an automatic speech detection system using wireless microphones for real-time ambulatory voice monitoring, (2) examine the detection accuracy under controlled environment and noisy conditions, and (3) report the results of the phonation ratio in practical scenarios.

**Methods:** We designed an adaptive threshold function to detect the presence of speech based on the energy envelope. We invited 10 teachers to participate in this study and tested the performance of the proposed automatic speech detection system regarding detection accuracy and phonation ratio. Moreover, we investigated whether the unsupervised noise reduction algorithm (ie, log minimum mean square error) can overcome the influence of environmental noise in the proposed system.

**Results:** The proposed system exhibited an average accuracy of speech detection of 89.9%, ranging from 81.0% (67,357/83,157 frames) to 95.0% (199,201/209,685 frames). Subsequent analyses revealed a phonation ratio between 44.0% (33,019/75,044 frames) and 78.0% (68,785/88,186 frames) during teaching sessions of 40-60 minutes; the durations of most of the phonation segments were less than 10 seconds. The presence of background noise reduced the accuracy of the automatic speech detection system, and an adjuvant noise reduction function could effectively improve the accuracy, especially under stable noise conditions.

**Conclusions:** This study demonstrated an average detection accuracy of 89.9% in the proposed automatic speech detection system with wireless microphones. The preliminary results for the phonation ratio were comparable to those of previous studies. Although the wireless microphones are susceptible to background noise, an additional noise reduction function can alleviate this limitation. These results indicate that the proposed system can be applied for ambulatory voice monitoring in occupational voice users.

***(JMIR Mhealth Uhealth 2020;8(12):e16746)*** doi: 10.2196/16746

**KEYWORDS**

voice disorder; speech envelope; phonation habits; background noise; noise reduction; adaptive threshold; dosimetry; phonotrauma

# Introduction

Human voice is produced via the periodic vibrations of vocal folds, driven by expiratory airflow. Cumulative voice loads and excessive vocal fold vibrations result in phonotraumatic injuries, such as vocal nodules and polyps [1]. The common symptoms of dysphonia (ie, phonation discomfort) include hoarseness, vocal fatigue, increased effort, and throat pain, which may limit the performance and long-term careers of occupational voice users [2]. Dysphonia also results in considerable financial losses for individuals and society [3,4]; the estimated annual cost associated with dysphonia is US $2.5 billion [5]. In addition, voice-related disorders significantly lower the quality of life in terms of physical functioning, general health, bodily pain, fatigue, and role limitation [6].

The most recognized risk for voice disorders is occupational voice overuse, commonly found in salespeople, industrial/factory workers, teachers, clergy, lecturers, and singers [6,7]. Among these occupations, the teaching profession has been significantly investigated by academic researchers [8-10]. In comparison to other occupations, teachers are more likely to report voice problems and the negative effects of dysphonia on their work performance [11]. Roy et al [2] reported that the prevalence of voice disorders was significantly higher in teachers (137/1243, 11.0%) in comparison to nonteachers (80/1288, 6.2%). The lifetime prevalence of dysphonia for teachers (717/1243, 57.7%) was also significantly higher than that for nonteachers (371/1288, 28.8%).

Voice therapy has been widely applied as the first-line treatment for voice disorders related to voice overuse or abuse [12,13]. By implementing multiple treatment strategies, voice therapy can effectively ameliorate the dysphonic symptoms, lower the phonation effort, and improve the voice quality [13]. However, one of the major challenges for voice therapy is the carryover of voicing techniques and habits taught during treatment sessions into their daily lives. To facilitate the maintenance of adequate phonation behavior, serial studies proposed the concept of ambulatory voice monitoring with promising results [14-16]. Most of these studies used a contact microphone or accelerometer attached to the anterior neck [17,18]. Subsequent studies demonstrated that this technology could significantly aid patients in controlling and tracking their vocal hyperfunction [19,20]. Another device for ambulatory voice monitoring was designed as a neck collar embedded with a contact microphone [21]. Although contact microphones and accelerometers can accurately detect phonation via the vibration of neck skin, the wiring and taping involved with these devices may cause discomfort in users. Furthermore, voice usage was mostly analyzed over a certain period with post hoc feedback [22], whereas real-time monitoring of the phonation ratio has not yet been reported.

To overcome the limitations in current devices, we propose a novel automatic speech detection system using a wireless microphone to capture acoustic signals from users, which can eliminate the discomfort associated with the wiring and taping of contact microphones. Our study hypothesizes that the speech energy envelope received via a wireless microphone can be used for ambulatory phonation monitoring. To examine this hypothesis, we designed this research with the following objectives: (1) to investigate the detection accuracy of speech, (2) to compare the measured phonation ratio and length of speech segments with those in existing literature, and (3) to examine the robustness of the noise reduction algorithm in simulated noisy conditions.

# Methods

## Overall Study Design

We proposed an automatic speech detection system using a wireless microphone for real-time ambulatory voice monitoring. We invited 10 teachers to participate in the pilot study. We designed an adaptive threshold (AT) function to detect the presence of speech based on the energy envelope. All participants were equipped with a wireless microphone during a teaching session (around 40-60 minutes) in a quiet classroom (background noise <55 dB sound pressure level [SPL]). We developed software for manually labeling the speech segments according to the time and frequency domains. We randomly selected 25 utterances (10 seconds each) from the recorded audio files to acquire the coefficients required for the AT function using a genetic algorithm (GA). Another 5 random utterances were used to test the accuracy of the automatic speech detection system using manually labeled data as the ground truth. We also mimicked scenarios of noisy backgrounds by mixing 4 different types of noise (at a signal-to-noise ratio [SNR] of 0, 3, and 5 dB) into the original recordings. An adjuvant noise reduction function using a log minimum mean square error (logMMSE) [23] algorithm was applied to counteract the influence on detection accuracy.

## Participants

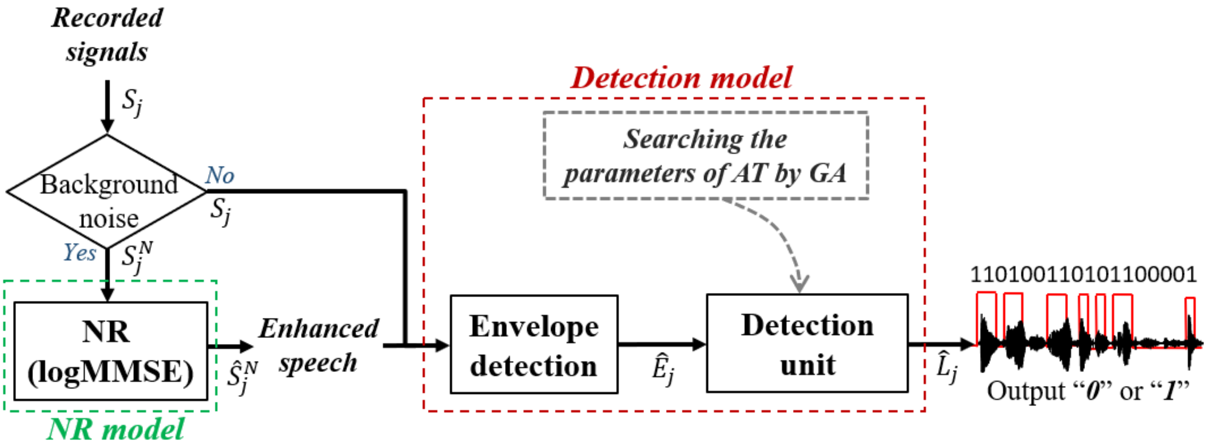
We invited 10 teachers to participate in this study. This study was conducted at Far Eastern Memorial Hospital and National Yang-Ming University. The study protocol was approved by the Research Ethics Review Committee of Far Eastern Memorial Hospital (FEMH 108019-E). For the first period of study, we recruited 5 teachers from April to June 2019; for the second period, we recruited another 5 teachers from February to April 2020. Each teacher was provided with a wireless microphone during a regular teaching session of 40-60 minutes. The average background noise level was controlled under 55 dB SPL, established as the controlled environment test condition.

## Automatic Speech Detection System

### Overview

Figure 1 illustrates the automatic speech detection system proposed in this study. The main corpus of this system is the detection model, which automatically divides acoustic signals into speech and nonspeech segments based on the energy envelope. We used a frame size of 32 milliseconds with a sampling rate of 16 kHz. Under simulated noisy conditions, the noise reduction model can be turned on to alleviate the effects from the background noise.

**Figure 1.** Proposed automatic speech detection system. AT: adaptive threshold; GA: genetic algorithm; logMMSE: logarithm minimum mean square error; NR: noise reduction.



### Detection Model

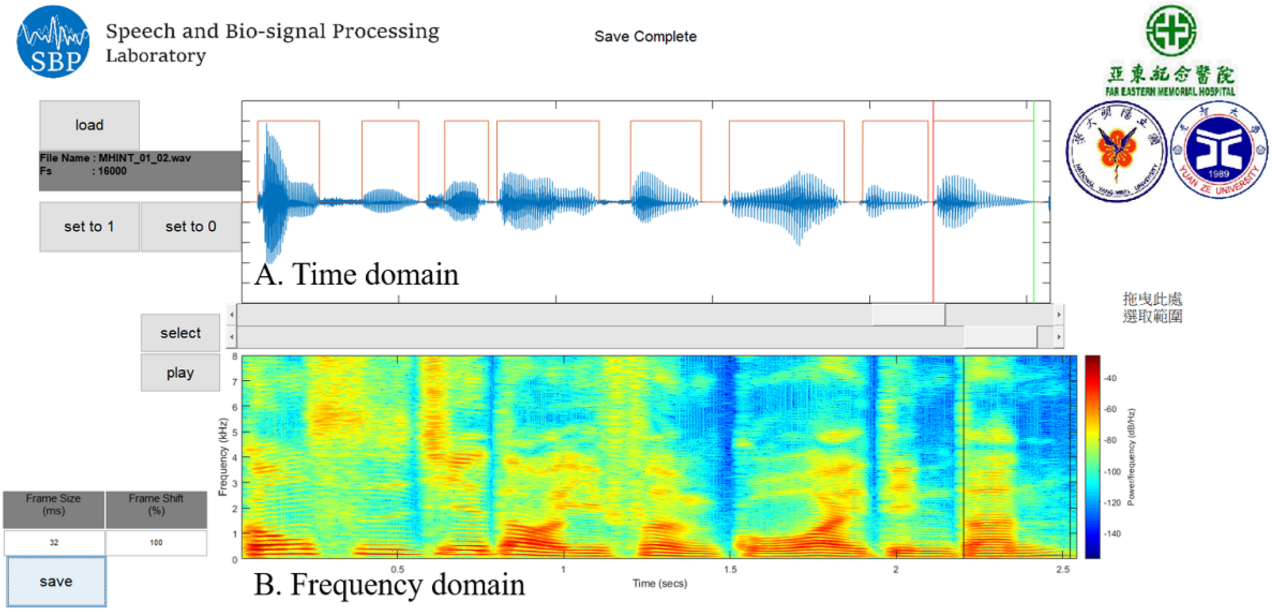
The signals ($S_j$) recorded from the wireless microphone were converted to envelope ($\hat{E}_j$) by an "envelope detection" unit; the power energy can be used in this unit. Then, the "detection unit" predicted whether the input frames were speech or nonspeech by comparing the value of $\hat{E}_j$ with that of the AT. The AT can be calculated using Equation (1); it is based on the energy of the input frame and 3 consecutive preceding frames.

$$\text{AT} = b + \sum_{i=0}^{3} a_i \hat{E}_{j-i} \quad \textbf{(1)}$$

Here, $a_i$ represents the coefficients of the energy envelope of the current frame ($i$=0) and 3 successive preceding frames ($i$=1 to 3). It should be noted that the 3 successive preceding frames are included in this equation according to the best performance observed in our pilot study. $\hat{E}_{j\text{-}i}$ represents the input acoustic energy features at the $j$-$i$ frame index and $b$ is the bias. When the value of $\hat{E}_j$ exceeded the threshold derived from the AT in Equation (1), the system generated "*1*" as the output, indicating that this frame was recognized as speech. However, if the value was lower than the threshold derived from the AT in Equation (1), the system generated "*0*" (nonspeech) as the output.

The 5 coefficients required to calculate the AT were defined by the following two steps: (1) manually labeling speech segments of the recoded audio files and (2) using a GA to search for these 5 coefficients. In the first step, we developed software (Figure 2) to manually label the speech segments according to their time and frequency domains. We applied the GA [24] to search for these 5 coefficients for the AT function based on 25 randomly selected utterances of 10 seconds each (details are provided in Multimedia Appendix 1). After acquiring the coefficients required for the AT function, another 5 random utterances were used to test the accuracy of the automatic speech detection system. Similarly, the manual labeling of the speech segments was considered as the ground truth. The overall accuracy of each subject was calculated based on each frame (ie, by dividing the predicted number of speech frames by the total number of speech frames labeled manually).

**Figure 2.** Overview of user interfaces in the proposed labeling tool. The selected speech segments are displayed as red brackets.


Speech and Bio-signal Processing Laboratory

Save Complete

File Name : MHINT_01_02.wav
Fs : 16000

Frame Size (ms): 32
Frame Shift (%): 100

### Noise Reduction Model

Because a wireless microphone is an air-conducted device that is susceptible to background noise, we performed additional experiments to examine the performance of the noise reduction model. We mimicked the presence of background noise by mixing the recorded speech signal ($S_j$) with 4 different common background noises (crowd cheering noise, sharp speech noise, street noise, and white noise, shown in Multimedia Appendix 2) at 3 SNR levels (0, 3, and 5 dB), denoted by $S_j^N$ ). The noisy signals ( $S_j^N$ ) were then processed by the logMMSE algorithm to obtain enhanced signals ( $\hat{S}_j^N$ ). Next, the $\hat{S}_j^N$ were sent into the detection model for energy and speech detection. We evaluated the performance of the noise reduction model by comparing the accuracy of speech detection under simulated noisy conditions with or without the noise reduction function.

### Measuring Phonation Ratio and Duration of Speech Segments

To examine the applicability of the automatic speech detection system, we calculated the phonation ratio of the participants as shown in Equation (2), which is a common approach used to analyze phonation habits and usage [19,25].

$$\text{Phonation ratio (\%)} = \frac{speech\ frames}{total\ frames} \times 100\% \quad (2)$$

In addition, we calculated the duration and distribution of the phonation and nonphonation segments in a similar manner as in previous literature [22].

## Results

Figure 3 displays the average recognition accuracy of the automatic speech detection system, which was 89.9% (frame-based) in the controlled environment, ranging from 81.0% (67,357/83,157) to 95.0% (199,201/209,685). Figures 4 and 5 illustrate the phonation ratio for the 10 teachers evaluated during the teaching session. On average, the phonation ratio ranged from 44.0% (33,019/75,044) to 78.0% (68,785/88,186). We also noted a drastic decrease in the phonation ratio in subject 5 at approximately 20 minutes (asterisk, Figure 4). After reviewing the recorded audio file, we observed that this teacher did not speak for a while because he left the podium to fetch chalk; this example further demonstrated the excellent sensitivity of the proposed automatic speech detection system in practical scenarios. Figures 6 and 7 illustrate the distribution of the speech and nonspeech segments in logarithmic scales. Analytical results showed that the durations of most of the speech and nonspeech segments were less than 10 seconds.

**Figure 3.** Average accuracy of the automatic speech detection system with respect to 10 teachers in a controlled environment.

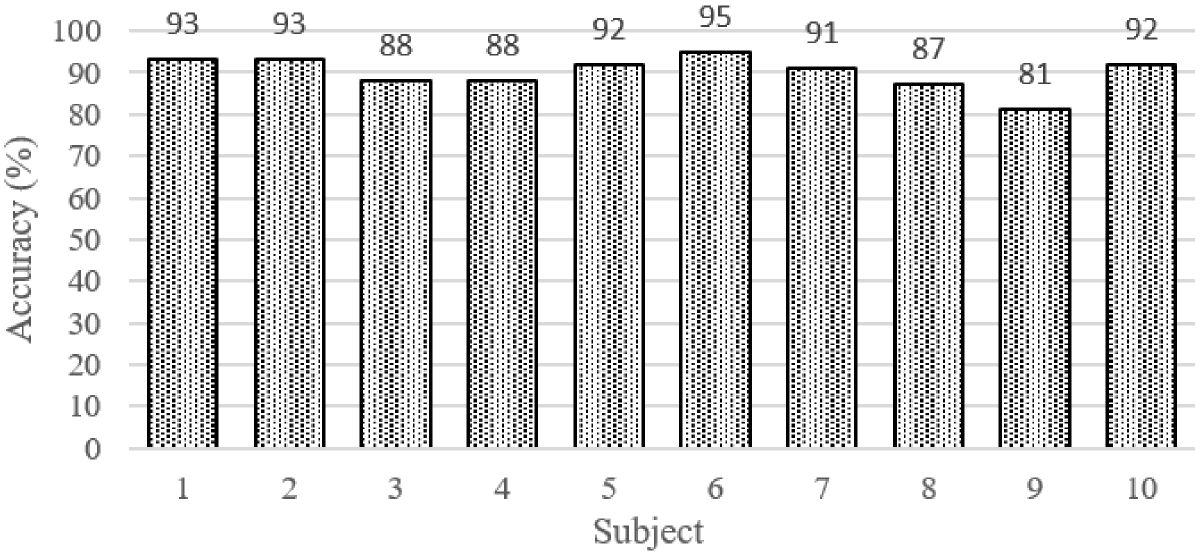


**Figure 4.** Phonation ratio over time for the 10 teachers (Subjects 1-6).

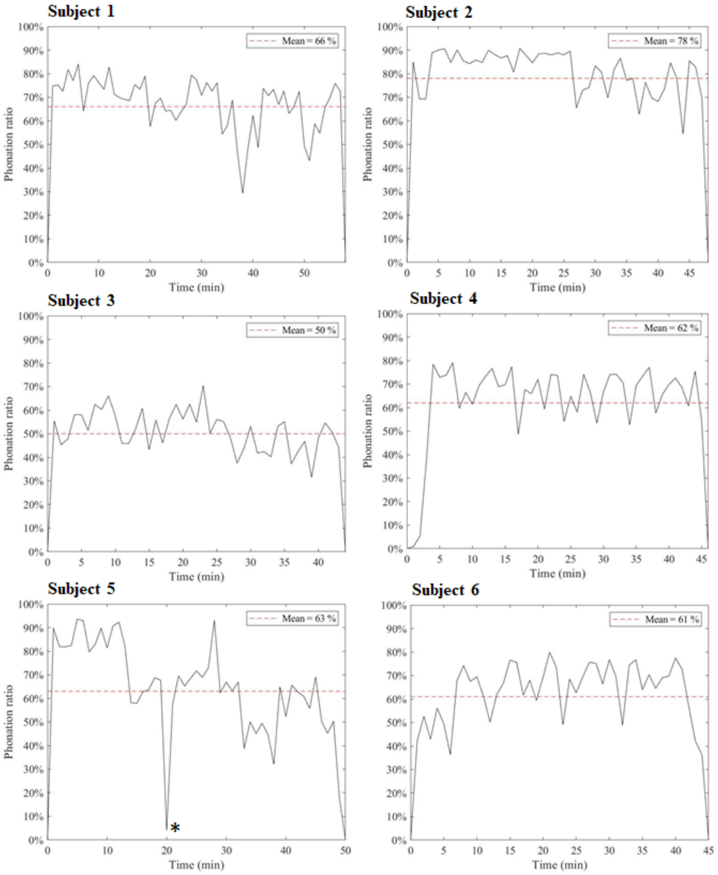


**Figure 5.** Phonation ratio over time for the 10 teachers (Subjects 7-10).

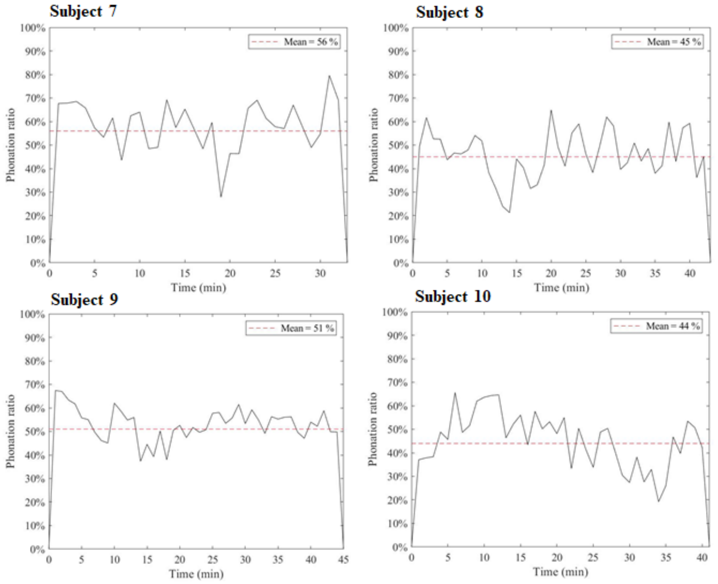


**Figure 6.** Speech and nonspeech segments measured by the automatic speech detection system. The x-axis indicates the length of the speech segments in a logarithmic scale, while the y-axis represents the occurrences during the recording period in a logarithmic scale (Subjects 1-6).

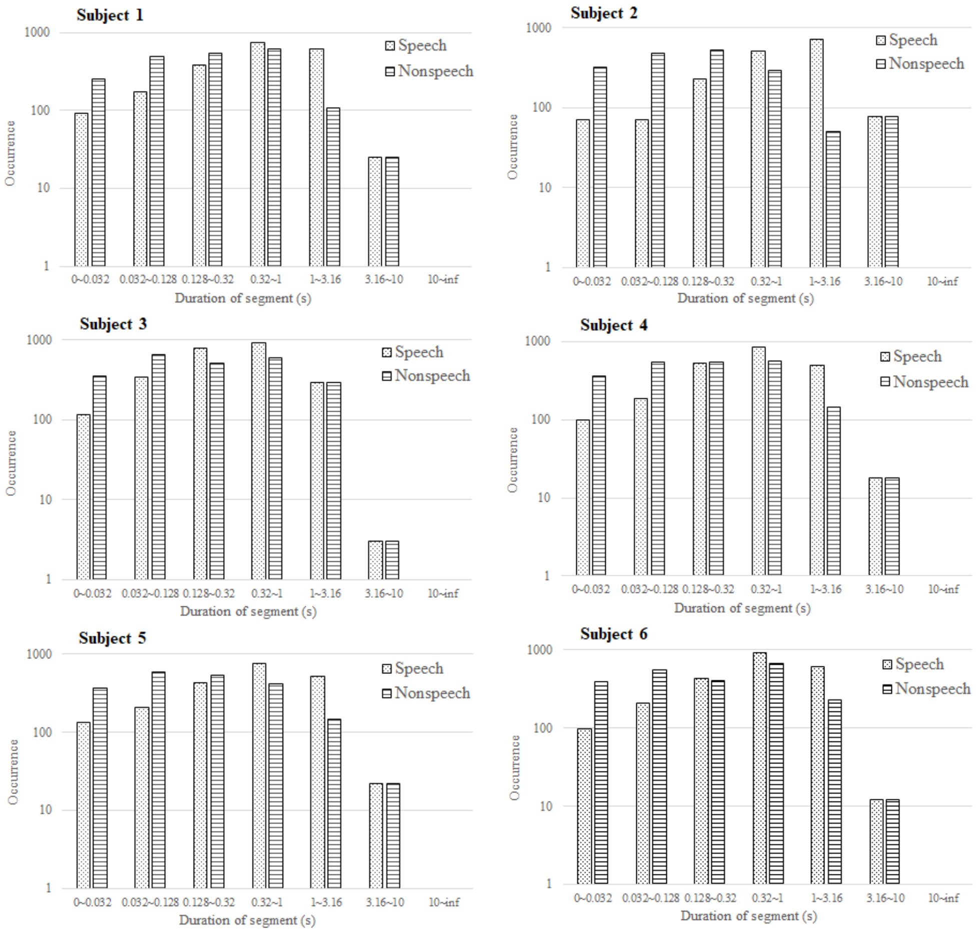
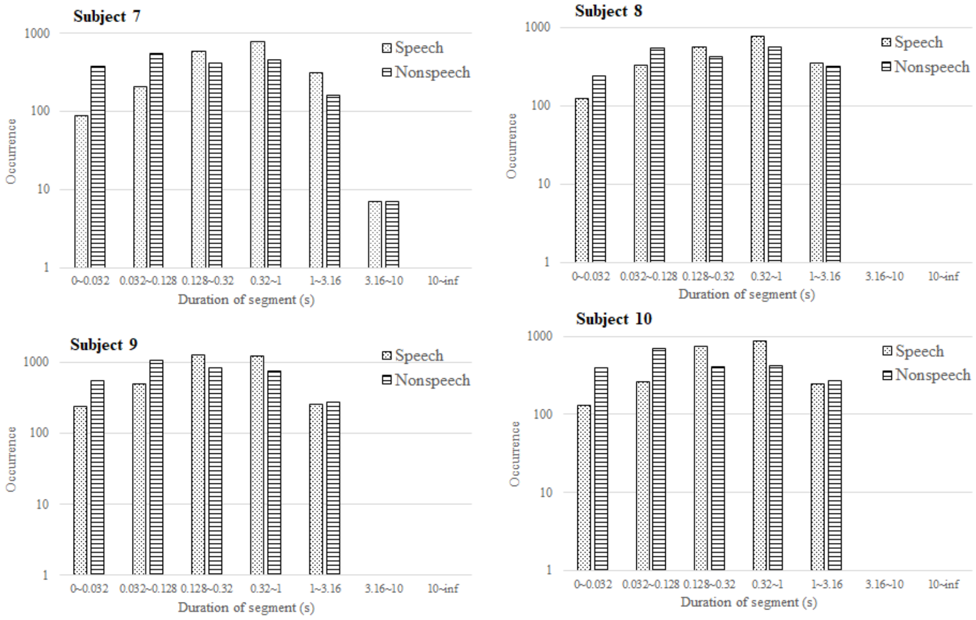


**Figure 7.** Speech and nonspeech segments measured by the automatic speech detection system. The x-axis indicates the length of the speech segments in a logarithmic scale, while the y-axis represents the occurrences during the recording period in a logarithmic scale (Subjects 7-10).



Figures 8 and 9 present a comparison of the same recordings under the controlled environment and simulated noisy conditions. We noticed that the detection accuracy dropped significantly under noisy conditions, indicating that the performance of the automatic speech detection system can be easily affected by the presence of noise without a noise reduction function. On average, the additional noise decreased the accuracy by approximately 33%, 36%, 34%, and 36% for 4 different types of noise: crowd cheering noise, sharp speech noise, street noise, and white noise, respectively.

Figure 10 presents an example of the relationship between the speech envelope and AT with and without the noise reduction function. Under the controlled environment, the AT was higher than the speech signal during nonspeech segments; in contrast, the energy envelope exceeded the AT in the presence of speech (Figure 10A). However, when the speech signal was contaminated by background noise, the overall energy exceeded the AT in both the speech and nonspeech segments (Figure 10B); thereby the proposed system may not be as effective in differentiating between speech and nonspeech signals. After enabling the logMMSE noise reduction function (Figure 10C), the AT could accurately detect the segments of speech versus nonspeech. On average, the additional noise reduction function yielded an average improvement of 4.9%, 27.5%, 19.3%, and 29.8% under the conditions of crowd cheering noise, sharp speech noise, street noise, and white noise, respectively. The detailed improvements are provided in Multimedia Appendix 3.

**Figure 8.** Accuracy of the automatic speech detection system with the presence of 2 different types of background noise (crowd cheer noise and speech sharp noise) at 3 SNR levels. The first bar of each graph indicates the accuracy of the original recording under the controlled environment. SNR: signal-to-noise ratio.

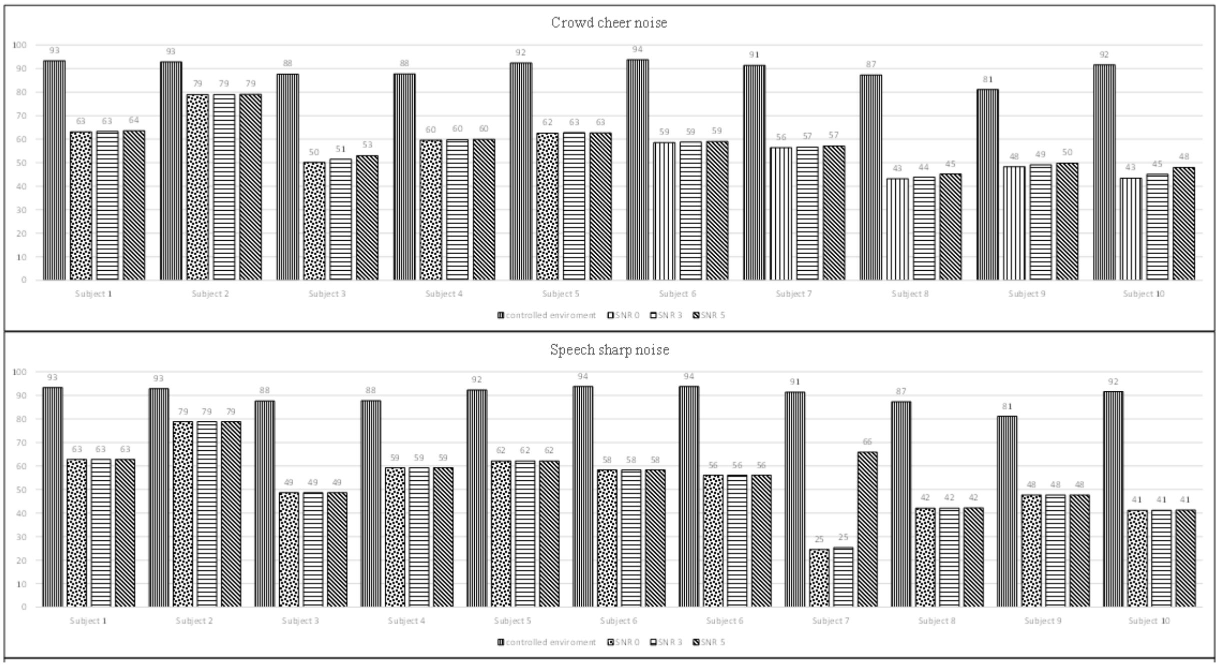


**Figure 9.** Accuracy of the automatic speech detection system with the presence of 2 different types of background noise (street noise and white noise) at 3 SNR levels. The first bar of each graph indicates the accuracy of the original recording under the controlled environment. SNR: signal-to-noise ratio.

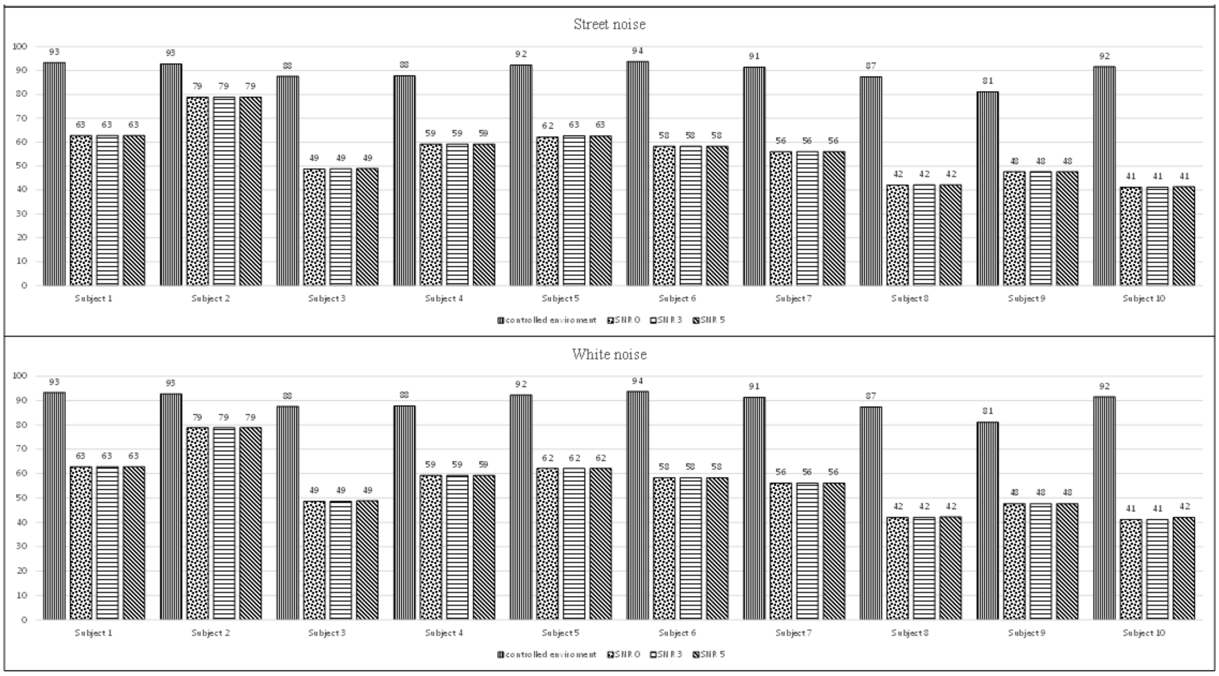
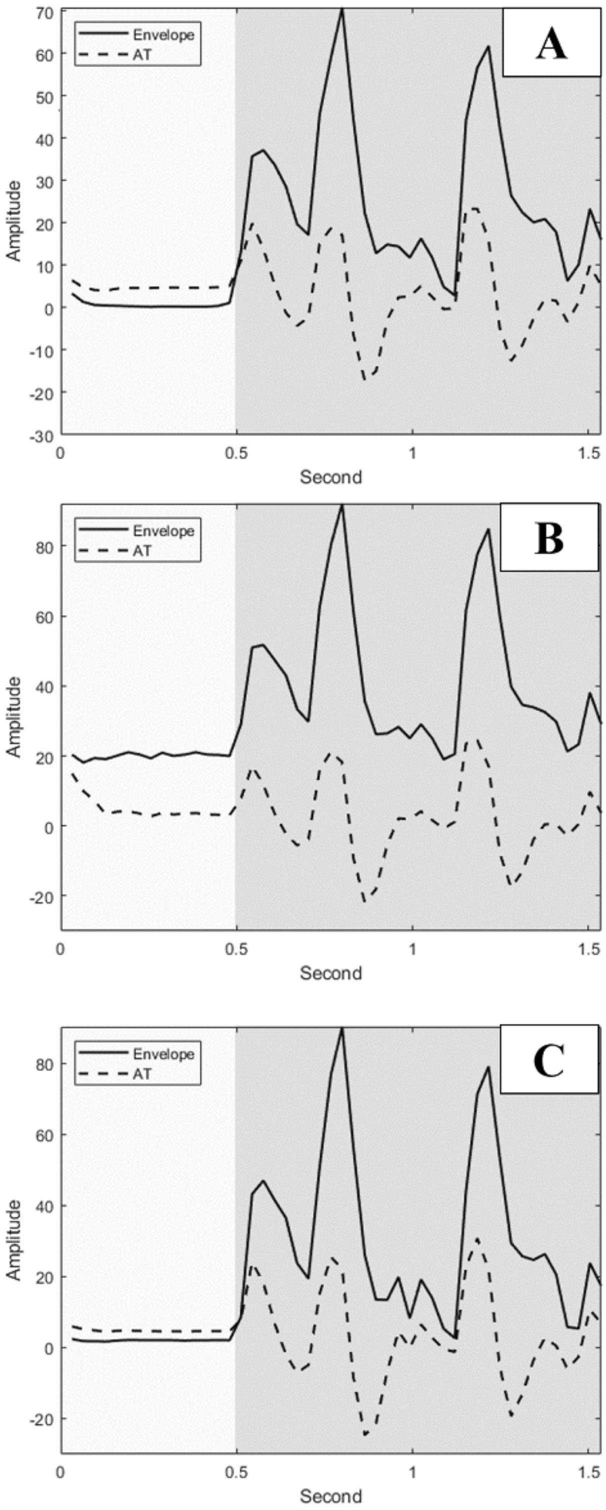


**Figure 10.** Example of the relationship between the speech envelope and AT in (A) the controlled environment, (B) noisy conditions without the noise reduction function, and (C) noisy conditions with the noise reduction function. Each part of the figure presents the same speech; the gray background denotes the speech segments, while the other areas indicate the nonspeech segments. AT: adaptive threshold.



## Discussion

### Principal Findings

In this study, we proposed an ambulatory phonation monitoring system with a wireless microphone. The results demonstrated that the proposed system can accurately differentiate between speech and nonspeech segments based on the energy envelope in a controlled environment. The implementation of an additional noise reduction function using a logMMSE algorithm can effectively reduce the impact of background noise. Preliminary results of the phonation ratio and the distribution of speech segments in 10 teachers were compatible with those in previous literature [18,19].

## Applicability and Accuracy of Automatic Speech Detection System

Most studies in the existing literature used a neck accelerometer to detect the vibrations of vocal folds via skin [18,25]. Although contact microphones can effectively suppress the effects of background noise [16], they may not always be convenient for the users, owing to the cumbersome wiring and taping. In contrast, the wireless microphone used in this study eliminated the discomfort associated with the wiring and taping of contact microphones. All the participants reported good tolerance using wireless microphone, without any physical discomfort during the teaching session.

Previous studies [22] applied predefined criteria to detect voice activity, such as the fundamental frequency during normal speaking (ie, 70 to 1000 Hz), SPL greater than 30 dB, and a low/high ratio of at least 22 dB. In this study, we specifically designed software to manually label the speech segments (Figure 2), which served as the ground truth for examining the detection accuracy of this novel system. Figure 3 demonstrates an average detection accuracy of 89.9% in the controlled environment, which established the applicability and reliability of the proposed system.

In comparison to previous studies [18,25], our results demonstrated a higher phonation ratio (range: 44.0%-78.0%; Figure 3) owing to the continual lecturing of the teachers in the classroom. Similarly, the durations of most of the speech segments were less than 10 seconds. We did not observe long durations of silence (nonspeech) in this study (Figures 6 and 7). In contrast, previous studies recorded the phonation ratio throughout the day (except sleeping) [18,25]; thus, longer silence periods were more likely to be documented.

## Benefits of Noise Reduction Function

Because wireless microphones are more susceptible to background noise, we examined the effectiveness of the additional noise reduction function by mixing 4 different types of background noise to simulate noisy conditions. Our results showed that the noise reduction function using the logMMSE algorithm can improve the detection accuracy by up to 45.8% (maximum) in stable noise conditions (eg, sharp speech noise and white noise) (Multimedia Appendix 3); however, logMMSE works less efficiently in competing voice signals (eg, crowd cheering noise), resulting in an improvement of approximately 5%, similar to previous literature [26]. Accordingly, other noise reduction approaches, such as deep learning [27], may be more robust for enhancing the automatic speech detection system in the future. Additionally, automatic gain control [28] can also be integrated into the system to normalize the input volume and improve the accuracy in cases where sudden changes are observed in the input volume.

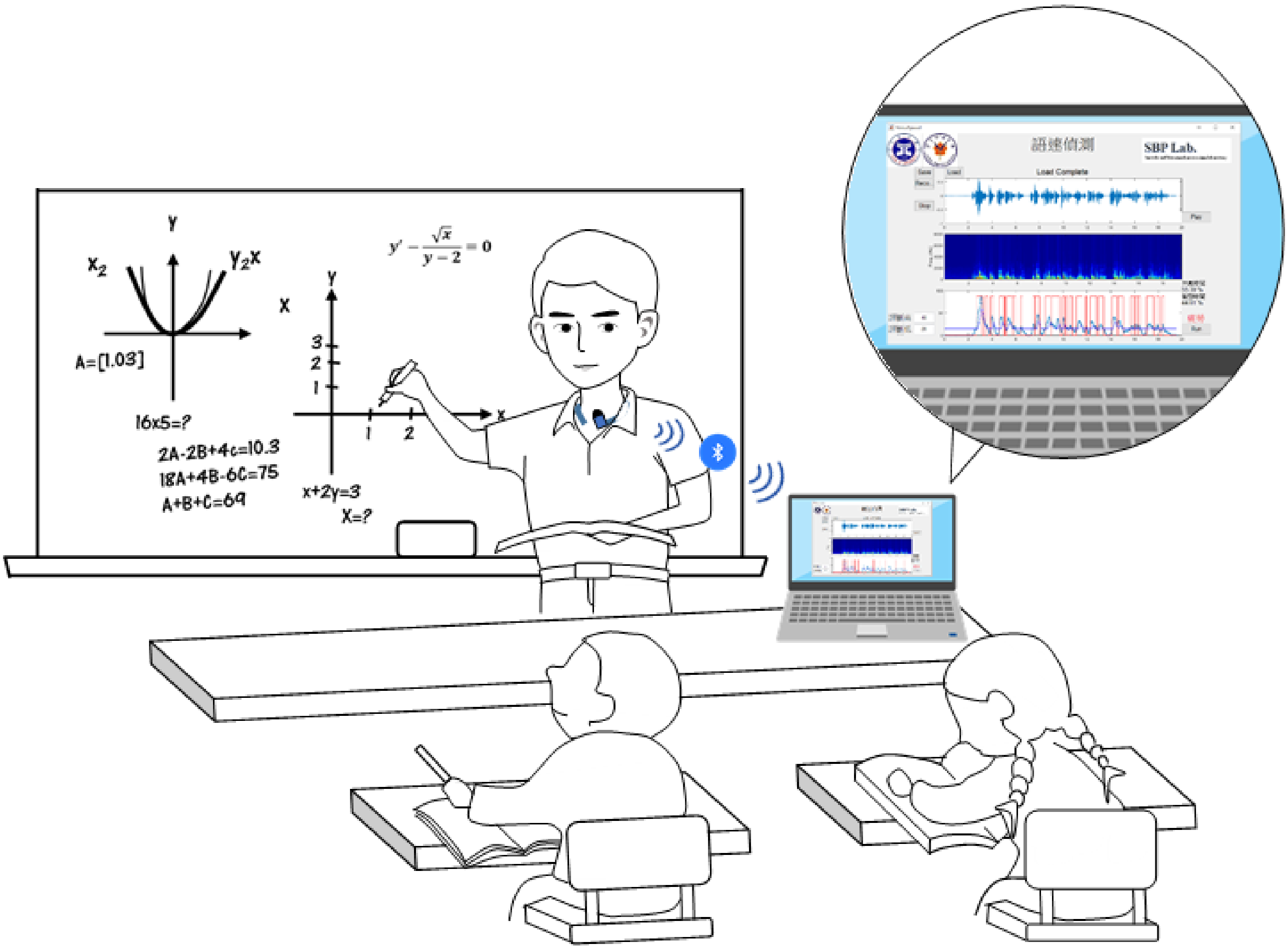
## Speaker Identification

The proposed system yielded comparable accuracy in most of the test conditions and an additional noise reduction function further improved the performance of the proposed system in noisy conditions. However, there is still room for improvement in some challenging conditions (eg, video sound or sudden increase in volume). We observed that subject 3 played a video clip with speech context during the class, and the loud speech from the video was misidentified as the speech of subject 3. For subject 4, several conversations took place between the teacher and students, which also caused the voice signals from the students to be misidentified as the speech of subject 4 and reduced the accuracy. One way to alleviate this inherent limitation of wireless microphones (ie, susceptibility to noise and competitive speakers) is the use of a microphone array with a beamforming algorithm that can fix (or adapt to adjust) the recorded position to distinguish between the speech of the speaker and background noise or other speakers. Another option to improve our system is implementing the speaker identification algorithm [29]; however, it requires significantly higher computing power to handle complex features (such as i-vector or x-vector [30,31]) using deep learning–based technology.

## Future Perspective

The study results suggest that the proposed automatic speech detection system with wireless microphone can be applied in practical scenarios to overcome the limitations of contact microphone for ambulatory phonation monitoring. The proposed system can be further implemented on personal laptops (or mobile phone devices) for daily use and timely feedback, as illustrated in Figure 11. By monitoring the baseline phonation ratio, doctors and speech language pathologists can prescribe a certain threshold of phonation ratio based on individual conditions. Upon exceeding this limit, an alarm signal (flash or sound) could be sent to the user to ensure that they take enough breaks; promising results are available with respect to this concept [32] but it requires further evidential support from ongoing studies.

**Figure 11.** Exemplary use of the automatic speech detection system by a teacher.



Although the proposed automatic speech detection system achieved 89.9% accuracy in this study for the proposed ambulatory phonation monitoring, it still has room for improvement. More recently, deep learning–based automatic speech recognition (ASR) [33] and natural language processing (NLP) [34,35] systems were proven to achieve higher speech recognition efficiency for conventional communication between human-machine applications (eg, Amazon Alexa, Google Home, and Apple Siri). These deep learning–based ASR and NLP systems could be applied in ambulatory phonation monitoring; however, some critical issues need to be addressed. For example, ASR and NLP technologies might violate the user's privacy because they recognize the context of the user's speech. In contrast, the automatic speech detection system of this study is energy-based; it will not directly access the content of speech and might be more acceptable to the users. In addition, ASR and NLP technologies require high computing power, especially when a deeper structure of the neural network is implanted to achieve higher speech recognition accuracy. A cloud-based ASR and NLP system could be effective in alleviating this limitation; however, the recorded speech data still needs to be uploaded to the server, which may lead to additional privacy and security issues. More recently, phonetic posteriorgram features obtained from the acoustic model of the ASR system was introduced for speech processing applications, and it has proven to achieve benefits in many tasks [36-38]. Following the success of phonetic posteriorgram, our future study could apply its features and deep learning technology to improve the performance of the current model.

Furthermore, this system can also be extended for detecting speech and communication disorders [39] (eg, Parkinson disease [40] and depression [41]). However, such work may require more sophisticated features of voice signals and computation techniques, such as the combination of the Mel frequency cepstral coefficients and deep neural networks, which was used in a previous study [42]. With the significant advancements in smartphones and smart home devices, the proposed automatic speech detection system can potentially be implemented in these devices to further decrease the clumsiness of any additional devices [43].

## Limitations

The first limitation of this study is the small number of participants (N=10). A larger cohort is required to obtain more robust evidence for the clinical use of automatic speech detection for ambulatory phonation monitoring. In addition, only teachers were recruited owing to the approved IRB protocol. Other occupations with high vocal demands (eg, salespeople and customer service representatives) will be included in the future to expand the potential use of the proposed system. Second, the proposed automatic speech detection system cannot precisely identify the speech of the speaker in the presence of loud competing background noise or other speakers. To overcome this issue, algorithms that require higher computing power, such as speaker identification or microphone array algorithms, could be used in future studies. Lastly, this automatic speech detection system requires manually labeling the recorded speech for model training. Considering the high accuracy achieved in this study, future research does not need to record the original voice content, so the confidentiality of the participants can be better protected.

## Conclusions

This study proposed an automatic speech detection system comprising a wireless microphone to receive the acoustic signals and an adaptive threshold for speech detection based on the energy envelope. The proposed system demonstrated a speech detection accuracy of 89.9%, and the analytical results for the phonation ratio and speech segments were comparable to those of previous research. Moreover, the use of an unsupervised noise reduction function (logMMSE) can improve the robustness of the proposed system in noisy conditions. These results imply that the proposed system can be a potential tool for ambulatory voice monitoring in occupational voice users.

## Acknowledgments

This study was supported by the Ministry of Science and Technology of Taiwan under the 108-2314-B-418-013 and 109-2218-E-010-004 projects and Far Eastern Memorial Hospital-National Yang-Ming University Joint Research (FEMH-NYMU108DN34).

## Conflicts of Interest

The authors CTW and YHL are the inventors of “Real-time monitor system of phonation,” (Taiwan Patent No. TW I626647).

## Multimedia Appendix 1

Using genetic algorithm to determine the parameters of the adaptive threshold.
[DOCX File , 395 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

Spectrogram of the background noises used in this study.
[DOCX File , 1301 KB-Multimedia Appendix 2]

## Multimedia Appendix 3

Mean improvements in recognition accuracies when using the noise reduction function for the proposed auto speech detection system under simulated noisy conditions.
[DOCX File , 26 KB-Multimedia Appendix 3]

## References

1. Titze IR, Martin DW. Principles of voice production. Saddle River City, NJ: Prentice Hall; 1994:169-190.
2. Roy N, Merrill RM, Thibeault S, Parsa RA, Gray SD, Smith EM. Prevalence of voice disorders in teachers and the general population. J Speech Lang Hear Res 2004 Apr;47(2):281-293. [doi: 10.1044/1092-4388(2004/023)] [Medline: 15157130]
3. Van Houtte E, Claeys S, Wuyts F, Van Lierde K. The impact of voice disorders among teachers: vocal complaints, treatment-seeking behavior, knowledge of vocal care, and voice-related absenteeism. J Voice 2011 Sep;25(5):570-575. [doi: 10.1016/j.jvoice.2010.04.008] [Medline: 20634042]
4. Cohen SM, Kim J, Roy N, Asche C, Courey M. The impact of laryngeal disorders on work-related dysfunction. Laryngoscope 2012 Jul;122(7):1589-1594. [doi: 10.1002/lary.23197] [Medline: 22549455]
5. Verdolini K, Ramig LO. Review: occupational risks for voice problems. Logoped Phoniatr Vocol 2001;26(1):37-46. [Medline: 11432413]
6. Titze IR, Lemke J, Montequin D. Populations in the U.S. workforce who rely on voice as a primary tool of trade: a preliminary report. J Voice 1997 Sep;11(3):254-259. [doi: 10.1016/s0892-1997(97)80002-1] [Medline: 9297668]
7. Williams NR. Occupational groups at risk of voice disorders: a review of the literature. Occup Med (Lond) 2003 Oct;53(7):456-460. [doi: 10.1093/occmed/kqg113] [Medline: 14581643]
8. Simberg S, Laine A, Sala E, Rönnemaa AM. Prevalence of voice disorders among future teachers. J Voice 2000 Jun;14(2):231-235. [doi: 10.1016/s0892-1997(00)80030-2] [Medline: 10875574]
9. Chen SH, Chiang SC, Chung YM, Hsiao LC, Hsiao TY. Risk factors and effects of voice problems for teachers. J Voice 2010 Mar;24(2):183-90, quiz 191. [doi: 10.1016/j.jvoice.2008.07.008] [Medline: 19481416]
10. Bermúdez de Alvear RM, Barón FJ, Martínez-Arquero AG. School teachers' vocal use, risk factors, and voice disorder prevalence: guidelines to detect teachers with current voice problems. Folia Phoniatr Logop 2011;63(4):209-215. [doi: 10.1159/000316310] [Medline: 20938203]
11. Smith E, Lemke J, Taylor M, Kirchner HL, Hoffman H. Frequency of voice problems among teachers and other occupations. J Voice 1998 Dec;12(4):480-488. [doi: 10.1016/s0892-1997(98)80057-x] [Medline: 9988035]

12. Stemple JC, Roy N, Klaben BK. Clinical voice pathology: Theory and management. San Diego, CA: Plural Publishing; 2018.
13. Chen SH, Hsiao T, Hsiao L, Chung Y, Chiang S. Outcome of resonant voice therapy for female teachers with voice disorders: perceptual, physiological, acoustic, aerodynamic, and functional measurements. J Voice 2007 Jul;21(4):415-425. [doi: 10.1016/j.jvoice.2006.02.001] [Medline: 16581227]
14. Ryu S, Komiyama S, Kannae S, Watanabe H. A newly devised speech accumulator. ORL J Otorhinolaryngol Relat Spec 1983;45(2):108-114. [doi: 10.1159/000275632] [Medline: 6341919]
15. Cheyne HA, Hanson HM, Genereux RP, Stevens KN, Hillman RE. Development and testing of a portable vocal accumulator. J Speech Lang Hear Res 2003 Dec;46(6):1457-1467. [doi: 10.1044/1092-4388(2003/113)] [Medline: 14700368]
16. Popolo PS, Svec JG, Titze IR. Adaptation of a Pocket PC for use as a wearable voice dosimeter. J Speech Lang Hear Res 2005 Aug;48(4):780-791. [doi: 10.1044/1092-4388(2005/054)] [Medline: 16378473]
17. Carroll T, Nix J, Hunter E, Emerich K, Titze I, Abaza M. Objective measurement of vocal fatigue in classical singers: a vocal dosimetry pilot study. Otolaryngol Head Neck Surg 2006 Oct;135(4):595-602 [FREE Full text] [doi: 10.1016/j.otohns.2006.06.1268] [Medline: 17011424]
18. Titze I, Hunter E, Svec JG. Voicing and silence periods in daily and weekly vocalizations of teachers. J Acoust Soc Am 2007 Jan;121(1):469-478 [FREE Full text] [doi: 10.1121/1.2390676] [Medline: 17297801]
19. Mehta DD, Zañartu M, Feng SW, Cheyne HA, Hillman RE. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. IEEE Trans Biomed Eng 2012 Nov;59(11):3090-3096 [FREE Full text] [doi: 10.1109/TBME.2012.2207896] [Medline: 22875236]
20. Remacle A, Morsomme D, Finck C. Comparison of vocal loading parameters in kindergarten and elementary school teachers. J Speech Lang Hear Res 2014 Apr 01;57(2):406-415. [doi: 10.1044/2013_JSLHR-S-12-0351] [Medline: 24129011]
21. Searl J, Dietsch AM. Tolerance of the VocaLog™ Vocal Monitor by Healthy Persons and Individuals With Parkinson Disease. J Voice 2015 Jul;29(4):518.e13-518.e20. [doi: 10.1016/j.jvoice.2014.09.011] [Medline: 25726068]
22. Van Stan JH, Mehta DD, Sternad D, Petit R, Hillman RE. Ambulatory Voice Biofeedback: Relative Frequency and Summary Feedback Effects on Performance and Retention of Reduced Vocal Intensity in the Daily Lives of Participants With Normal Voices. J Speech Lang Hear Res 2017 Apr 14;60(4):853-864 [FREE Full text] [doi: 10.1044/2016_JSLHR-S-16-0164] [Medline: 28329366]
23. Scalart P, Filho JV. Speech enhancement based on a priori signal to noise estimation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. 1996 Presented at: IEEE International Conference on Acoustics, Speech, and Signal Processing; 1996; Atlanta, GA p. 629-632. [doi: 10.1109/icassp.1996.543199]
24. John H H. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. In: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. Boston: A Bradford Book; 1992.
25. Van Stan JH, Mehta DD, Zeitels SM, Burns JA, Barbu AM, Hillman RE. 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 2015 Nov;124(11):864-874 [FREE Full text] [doi: 10.1177/0003489415589363] [Medline: 26024911]
26. Wang SS, Tsao Y, Wang HLS, Lai YH, Li LPH. A deep learning-based noise reduction approach to improve speech intelligibility for cochlear implant recipients in the presence of competing speech noise. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2017 Presented at: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC); 2017; Kuala Lumpur p. 808-812. [doi: 10.1109/apsipa.2017.8282144]
27. Xu Y, Du J, Huang Z, Dai LR, Lee CH. Multi-objective learning and mask-based post-processing for deep neural network-based speech enhancement. arXiv. 2017. URL: https://arxiv.org/abs/1703.07172 [accessed 2020-10-14]
28. Shan T, Kailath T. Adaptive algorithms with an automatic gain control feature. IEEE Trans. Circuits Syst 1988 Jan;35(1):122-127. [doi: 10.1109/31.1709]
29. Reynolds D, Rose R. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process 1995 Jan;3(1):72-83. [doi: 10.1109/89.365379]
30. Kanagasundaram A, Vogt R, Dean DB, Sridharan S, Mason MW. I-vector based speaker recognition on short utterances. 2011 Presented at: 12th Annual Conference of the International Speech Communication Association (ISCA); 2011; Florence, Italy p. 2341-2344 URL: https://eprints.qut.edu.au/46313/1/IS110023.PDF
31. David S, Garcia-Romero D, Sell G, Povey D, Khudanpur S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. NY, US: IEEE; 2018 Presented at: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 15-20 April 2018; Calgary, AB, Canada. [doi: 10.1109/ICASSP.2018.8461375]
32. Van Stan JH, Mehta DD, Petit RJ, Sternad D, Muise J, Burns JA, et al. Integration of Motor Learning Principles Into Real-Time Ambulatory Voice Biofeedback and Example Implementation Via a Clinical Case Study With Vocal Fold Nodules. Am J Speech Lang Pathol 2017 Feb 01;26(1):1-10 [FREE Full text] [doi: 10.1044/2016_AJSLP-15-0187] [Medline: 28124070]

33. Zhang W, Cui X, Finkler U, Kingsbury B, Saon G, Kung D. Distributed deep learning strategies for automatic speech recognition. 2019 Presented at: ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019; Brighton, United Kingdom p. 5706-5710. [doi: 10.1109/icassp.2019.8682888]
34. Kaufman DR, Sheehan B, Stetson P, Bhatt AR, Field AI, Patel C, et al. Natural Language Processing-Enabled and Conventional Data Capture Methods for Input to Electronic Health Records: A Comparative Usability Study. JMIR Med Inform 2016 Oct 28;4(4):e35 [FREE Full text] [doi: 10.2196/medinform.5544] [Medline: 27793791]
35. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
36. Tian X, Chng ES, Li H. A Speaker-Dependent WaveNet for Voice Conversion with Non-Parallel Dat. 2019 Presented at: Interspeech 2019; 2019; Graz, Austria p. 201-205. [doi: 10.21437/interspeech.2019-1514]
37. Sun L, Li K, Wang H, Kang S, Meng H. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. NY, US: IEEE; 2016 Presented at: IEEE International Conference on Multimedia and Expo (ICME); 2016; Seattle, WA. [doi: 10.1109/ICME.2016.7552917]
38. Chen CY, Zheng WZ, Wang SS, Tsao Y, Li PC, Li YH. Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-based Voice Conversion System. In: IEEE Interspeech. 2020 Presented at: IEEE Interspeech; 2020; Shanghai.
39. Furlong LM, Morris ME, Erickson S, Serry TA. Quality of Mobile Phone and Tablet Mobile Apps for Speech Sound Disorders: Protocol for an Evidence-Based Appraisal. JMIR Res Protoc 2016 Nov 29;5(4):e233 [FREE Full text] [doi: 10.2196/resprot.6505] [Medline: 27899341]
40. Erdogdu Sakar B, Serbes G, Sakar CO. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. PLoS One 2017;12(8):e0182428 [FREE Full text] [doi: 10.1371/journal.pone.0182428] [Medline: 28792979]
41. Taguchi T, Tachikawa H, Nemoto K, Suzuki M, Nagano T, Tachibana R, et al. Major depressive disorder discrimination using vocal acoustic features. J Affect Disord 2018 Jan 01;225:214-220. [doi: 10.1016/j.jad.2017.08.038] [Medline: 28841483]
42. Fang S, Tsao Y, Hsiao M, Chen J, Lai Y, Lin F, et al. Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach. J Voice 2019 Sep;33(5):634-641. [doi: 10.1016/j.jvoice.2018.02.003] [Medline: 29567049]
43. Chung AE, Griffin AC, Selezneva D, Gotz D. Health and Fitness Apps for Hands-Free Voice-Activated Assistants: Content Analysis. JMIR Mhealth Uhealth 2018 Sep 24;6(9):e174 [FREE Full text] [doi: 10.2196/mhealth.9705] [Medline: 30249581]

## Abbreviations

**ASR:** automatic speech recognition
**AT:** adaptive threshold
**GA:** genetic algorithm
**logMMSE:** log minimum mean square error
**NLP:** natural language processing
**SPL:** sound pressure level

*Edited by G Eysenbach; submitted 23.10.19; peer-reviewed by B Johnson, C He; comments to author 20.01.20; revised version received 03.06.20; accepted 03.10.20; published 03.12.20*

*Please cite as:*
*Wang CT, Han JY, Fang SH, Lai YH*
*Ambulatory Phonation Monitoring With Wireless Microphones Based on the Speech Energy Envelope: Algorithm Development and Validation*
*JMIR Mhealth Uhealth 2020;8(12):e16746*
*URL: https://mhealth.jmir.org/2020/12/e16746*
*doi: 10.2196/16746*
*PMID: 33270033*