Incorporating Speech Activity Information into Intelligent E-learning Systems

Yücel Ugurlu

Department of Creative Informatics, The University of Tokyo,
7-3-1 Hongo, Bukyo-ku, Tokyo, Japan

E-mail: yucel.ugurlu@ni.com

Abstract

We propose a new approach for e-learning systems that incorporates speech activity information to help build intelligent learning systems. The proposed algorithm is based on the pitch frequency, power amplitude, and duration of a voice activity signal. The approach was tested on users for an e-learning duration of 60 min. The robustness of the proposed algorithm was measured using various sound signals (speech, noise, and emotions) obtained from varying sources.

Keywords: human-computer interaction, e-learning, graphical programming, intelligent systems

1. Introduction

The emergence of new technologies and the Internet has resulted in one of the major revolutions in education in the form of e-learning. Today, e-learning is a popular alternative for students, engineers, and executives for training and development because of its fully customizable and user-oriented framework⁽¹⁾.

Next-generation e-learning systems should be sufficiently smart to be able to adapt to a student's learning style and assure high standards of accessibility and usability, in order to make learner's interaction with the system as natural and intuitive as possible. The primary objective of our work is to include, in addition to traditional interactions, computer-interaction data such as videos and audio signals in e-learning systems. Furthermore, we expect to understand user behaviors to build more intelligent and collaborative learning environments^(2, 3).

The rest of this paper is organized as follows: in section 2, the proposed speech activity detection approach

is explained; in section 3, the framework of the intelligent e-learning system is described; in section 4, the approach is tested on an e-learning portal and results are assessed; in section 5, future work is considered.

2. Speech Activity Detection

Speech activity detection is a technique used in speech processing in which the presence or absence of human speech is detected. It is an important component of speech processing techniques such as speech enhancement, speech coding, and automatic speech recognition. Various speech activity algorithms have been developed that provide varying features and involve trade-offs between latency, sensitivity, accuracy, and computational cost in communication systems⁽⁴⁻⁶⁾.

However, using speech activity detection to formulate human-computer interaction (HCI) data such as sleep, speech, silence, or emotions is a new concept in e-learning systems. In this research, speech or silence detection information is embedded in the e-learning system. An appropriate pitch frequency and high power of a voice signal implies that the student is speaking to others. In general, e-learning systems are passive and a voice signal is not required to interact with computers. Therefore, similar to traditional teaching classes, user voice activity can be utilized to evaluate the user interests.

Sounds produced by humans are created from the vibration of vocal chords that interrupts the flow of air and produces a frequency ranging approximately between 50 and 500 Hz. The voiced speech of a typical adult male will have a fundamental frequency between 85 and 180 Hz, and that of a typical adult female will be between 165 and 255 Hz^(7, 8).

Fig. 1 Block diagram of speech activity detection

The basic principle of speech activity detection is that it extracts measured features or quantities from the input signal and then compares these values with thresholds usually extracted from noise-only periods. A block diagram of the proposed speech activity detection algorithm is shown in Fig. 1.

First, the voice signal $x(k)$ is acquired according to Shannon's sampling theorem, which states that a continuous signal must be discretely sampled with at least twice the frequency (Nyquist frequency) of the highest frequency present in the signal⁽⁹⁾.

Next, the speech signal is characterized by a sequence of peaks that occur periodically at the fundamental frequency of the speech signal. In contrast, during unvoiced intervals, the peaks are relatively smaller and do not occur in any discernible pattern. Thus, the maximum peak amplitude during an analysis interval can be used to determine the amplitude of the signal in a simple manner and help distinguish between voiced and unvoiced speech segments.

A pitch detection algorithm used to estimate the pitch or fundamental frequency of speech or a musical note. This

can be done in the time or frequency domain, or in both the domains. Fourier transform, which expresses a signal in terms of the frequencies of the waves that constitute that signal, is an efficient approach for understanding the characteristics of speech signals. The Fourier transform of a signal is represented as

$$X(n) = \frac{1}{N} \sum_{k=0}^{N-1} x(k) e^{-\frac{j2\pi kn}{N}} \quad \text{for } n = 0, N-1 \quad (1)$$

in which $x(k)$ is the amplitude at time sample k , N is the number of samples or frequency points of interest, n is a frequency value from 0 to $N-1$, and $X(n)$ is the spectral-domain representation of $x(k)$.

Fourier transform generates a power spectrum of the input signal and yields the main frequency component and amplitude value. User voice activity can be detected using

$$v_t = \begin{cases} 1 & 500 > f \geq 85 \text{ and } A \geq 0.001, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

where t is the time in seconds, f is the fundamental frequency, and A is the power amplitude. For $v_t = 1$, the fundamental frequency (f) and amplitude (A) of the signal matches the range of the speech signal.

Eq. (2) can be used to detect all kinds of sound activities including voice, noise, and emotions. However, to obtain a more robust system, normal speech signals should be differentiated from noise and emotion signals. Therefore, the type of speech signal can be identified using

$$s_t = \begin{cases} 1 & \sum_{t=1}^{60} v_t > 10, \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where t is the duration of the speech signal.

We assume that a user's emotions usually last less than 10 s. In this case, there is no significant speech activity present and $s_t = 0$. However, it is considered a speaking activity ($s_t = 1$) if the duration of the speech signal exceeds the 10 s threshold time. This threshold value was selected experimentally and works well in most situations.

3. Intelligent E-learning Systems

Intelligent e-learning is a new learning and teaching medium that uses an intelligent tutoring system such that online learning can adapt to a student's level of knowledge.

Intelligent systems are designed to simulate human reasoning and learning, thus reducing the need for human intervention in the application process. Intelligent e-learning provides students with customized educational content and the unique feedback that they need and when they need it ⁽¹⁰⁾. In this study, we introduce an approach based on intelligent techniques, which in turn are based on the user activity data, and build an adaptive framework.

However, utilizing only the user access information and duration is not enough to build intelligent and adaptive systems. The relationship between e-learning content and user reaction should be measured and used for further analysis and evaluation. For this purpose, the integration of e-learning access and voice activity information with the system should be implemented to build better systems.

An intelligent e-learning system records HCI data and uses this information to determine a user's interest, which is then used to recommend topics or to improve contents. Therefore, an intelligent e-learning system should incorporate the behavior of users, using image and signal processing ⁽¹¹⁾; typical examples are tracking face and head postures and detecting voice activities and emotions. An integrated voice activity signal detection component in an e-learning system is an essential source of information, because it is an easy and common setup even for homes, schools, or offices.

In this study, a speech activity detection approach is presented to build intelligent e-learning systems. The general system schema is illustrated in Fig. 2. As shown, speech activity detection provides additional information about the user's behavior and the actual usage of the e-learning system, and these additional sources of information can be easily integrated into the final evaluation process.

Fig. 2. An intelligent e-learning system involving speech activity detection

E-learning access history can be represented numerically using

$$h_t = \begin{cases} 1 & \text{access confirmed,} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where t is the time in minutes, and h_t equals 1 if user access is confirmed.

In the previous section, we introduced a method to detect whether the speaking information is significant, using Eq. (3). A long speech indicates little interest in the topic being taught. These topics can be identified using the following equation:

$$d_t = h_t - s_t \quad (5)$$

When $d_t = 1$, there is a good match between the use of e-learning and voice activity detection (silence), but when $d_t = 0$, the user possibly has little interest in the presented content or has lost motivation for some reason.

Eq. (5) indicates the level of HCI in terms of the corresponding durations. However, actual e-learning topics should be identified using the time information. E-learning systems includes a series of topics and the accessed topics can be represented as

$$T = (T_1, T_2, T_3, \dots, T_n) \quad (6)$$

where n is the number of completed topics for a user. Similarly, the corresponding HCI data is represented as

$$Z = (Z_1, Z_2, Z_3, \dots, Z_n) \quad (7)$$

The interest level for each individual topic can be calculated as follows:

$$Z_i = \frac{\sum_{t=1}^m d_t}{m} \quad (8)$$

where i is the topic number and m is the number of samples taken in a single topic.

If there is no HCI in certain topics, these topics can be displayed as incomplete or keywords of these topics can be utilized to build adaptive systems. Each topic will be recorded in the system according to its completion rate and the student interest level information. This information can be used to build an intelligent e-learning system. If $Z_i \geq 0.8$, there is a good HCI between the user and the e-learning system because of silence (no speaking activity). This threshold value was selected experimentally, but it may differ for different learning environments, courses, and

cultures. For $Z_i < 0.8$, the e-learning system will provide some adaptive feedback to the user when the speech activity exceeds the threshold level.

In conclusion, users learn at varying rates, with different levels of knowledge, and with different levels of understanding. In traditional courses, teachers regularly adjust the homework based on their students' performance. Similarly, intelligent e-learning systems can provide feedback based on the user's interaction and behavior in order to encourage better performance.

4. Experimental Results

To measure the performance of the proposed approach, we tested it on users of National Instruments' e-learning portal. This portal consists of more than 200 topics and 30 h of learning time, all pertaining to teaching graphical programming in LabVIEW to engineering students. The e-learning content includes various single-topic videos, each of a 5 to 10 min duration, readings, and short quizzes. A typical learning environment of LabVIEW graphical programming is shown in Fig. 3. In practice, a unique user ID and password are provided to access all e-learning functions. Since users are logged into the system using their user IDs, it is easy to acquire their access history, as formalized in Eq. (4).

In our speech activity detection system for LabVIEW, a standard sound card and microphone was used to acquire the voice signals⁽¹²⁾. LabVIEW includes simple virtual instruments (VIs) for analog input and output using the sound card built into many PCs. This is convenient for laptops because the sound card and microphone are usually built-in.

Human speech signal contains a significant amount of

Fig. 3 E-learning environment, which explains a typical data acquisition method, for LabVIEW

Fig. 4 Front panel of LabVIEW-based speech activity detection program

Fig. 5 Block diagram of LabVIEW code

energy (under 2.5 KHz). Therefore, we took a 44 KHz sampling rate of the speech signal, and a 16-bit mono using the Acquire Sound Express VI in LabVIEW. Next, the Tone Measurement Express VI was used to calculate the frequency and amplitude of the input signal; this system finds the tone with the highest amplitude in the signal and calculates the amplitude and frequency. Then, the amplitude and frequency information was used to determine if any user speech activities existed during the e-learning time. The front panel and block diagram of the LabVIEW program is shown in Figs. 4 and 5, respectively.

In our experimental setup, we collected voice samples for every second of the 60 min duration of e-learning for a

Fig. 6 Power and frequency plot of user data

Fig. 7 User voice activity for the entire learning duration

Fig. 8 Detection of speech activity signal

single user. The voice activity signals were analyzed based on the proposed algorithm in section 2. Subsequently, the user's accessed topics along with the corresponding timestamp information were also analyzed. Fig. 6 shows the power amplitude and frequency of the acquired signals. For

simplicity, we calculated the maximum amplitude and frequency values of the voice signal for each minute of the 60 min experiment duration.

Fig. 7 shows the voice activity signal based on the proposed method. As shown, a user engaged in intensive speech between the 38th and 48th min. The frequency and power amplitude values fall within the detection interval. Therefore, speech signal activities were successfully detected as shown in Fig. 8.

During the 60 min duration of e-learning, seven different topics were covered by the user. According to the experimental results, we found that the HCI was sufficiently good for topics 1, 2, 3, 4, and 7. Thus, for these cases Z_i equals 1. However, poor HCI was detected for topics 5 and 6 because of the significant amount of speech. Thus, for these cases Z_i equals 0.6 and 0.4, respectively. Therefore, the user interaction history can be used to build an adaptive and intelligent system. The aim of this analysis is to instruct the user to re-visit certain topics characterized by poor HCI data. Table 1 summarizes the speech activity detection results for corresponding intelligent feedback.

Table 1. Speech activity detection for corresponding e-learning topics

Topics Visited	Speech (s_t)	HCI (Z_i)	Revisit Required
Topic 1	0	1	No
Topic 2	0	1	No
Topic 3	0	1	No
Topic 4	0	1	No
Topic 5	1	0.6	Yes
Topic 6	1	0.4	Yes
Topic 7	0	1	No

Furthermore, in order to measure the robustness of the proposed approach, we tested the algorithm for various sound signals such as those produced from crying children, music, a ringing phone, a television, and human emotions. Table 2 summarized the analysis results of these signals based on the frequency, amplitude, and duration, as well as how they are identified (speech or silence) using Eq. (3).

The pitch frequency was higher than regular speech signals in the cases of crying children, music, and a ringing phone, whereas a news broadcast on the television exhibited a low-frequency pitch and power. Moreover, the duration of emotion-related signals was shorter than that of speech-related signals.

Table 2. Speech detection for various sound signals

Voice Type	Frequency (Hz)	Amplitude (Volt)	Time (sec)	Speech (s_t)
Men (speaking)	158	0.0164	30	1
Women (speaking)	467	0.0176	44	1
Children (crying)	1124	0.1045	65	0
Music (waka waka)	2062	0.0202	180	0
Phone ringing (classis)	1583	0.0015	20	0
Television (news)	50	0.0012	60	0
Emotion 1	210	0.0151	2	0
Emotion 2	126	0.010	3	0
Emotion 3	179	0.0126	5	0

5. Conclusion

This paper proposed a new approach that incorporates speech activity information into an e-learning system. To that end, HCI was acquired which represented the student's behavior or interest during e-learning.

The proposed approach is cheap, robust, and very practical to implement in most e-learning systems. The experimental results show that speech activity information provides us with additional information on HCI data and helps us determine the student's behavior during e-learning. In addition, we tested our approach in different voice activity scenarios such as emotions, sounds from varying sources, and noise, and found that the algorithm is able to detect user speech accurately.

However, speech information is not enough to completely understand user behavior. Therefore, we require the complete set of HCI data. Our future study will be dedicated towards integrating visual and acoustic information to better acquire HCI information and build intelligent and adaptive e-learning systems.

Acknowledgment

This work was supported by National Instruments Japan Corporation.

References

- (1) Ruth C. Clark and Richard E. Mayer : e-Learning and the Science of Instruction, Pfeiffer, 2007.
- (2) Thanasis Daradoumis, Santi Caballé, Joan Manuel Marquès, and Fatos Xhafa : Intelligent Collaborative

e-Learning Systems and Applications (Studies in Computational Intelligence), Springer, 2010.

- (3) Abdul-Rahim Ahmad, Otman Basir, and Khaled Hassanein : "Adaptive user interfaces for intelligent e-learning: Issues and trends", International Conference on Electronic Business, Beijing, pp. 923–934, 2004.
- (4) Javier Ramírez, José C. Segura, Juan M. Górriz, and Luz García : "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 8, pp. 2177–2189, 2007.
- (5) Rabiner Lawrence and Juang Bing-Hwang : "Fundamentals of Speech Recognitions", Prentice Hall New Jersey, 1993.
- (6) Dongwen Ying, Yonghong Yan, Jianwu Dang, and F. K. Soong: "Voice activity detection based on an unsupervised learning framework", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No.8, pp. 2624–2633, 2011.
- (7) Ingo R. Titze : Principles of Voice Production, Prentice Hall, pp. 188, 1994.
- (8) Ronald J. Baken and Robert F. Orlikoff : Clinical Measurement of Speech and Voice, London, Taylor and Francis Ltd., 1987.
- (9) Nasser Kehtarnavaz and Sidharth Mahotra : Digital Signal Processing Laboratory, BrownWalker Press, Boca Raton, Florida, 2010.
- (10) Valerie Shute and Brendon Towle : "Adaptive e-learning", Educational Psychologist, Vol. 38, No. 2, pp. 105–112, 2003.
- (11) Yücel Uğurlu : "Smart e-learning: Enhancement of human-computer interactions using head posture images", International Journal of Engineering Education, Vol. 29, No. 3, pp. 568–577, 2013.
- (12) Robert King : Introduction to Data Acquisition with LabView, McGraw-Hill Science/Engineering/Math, 2012.