Sound Waves Gesture Recognition for Human-Computer Interaction

Nguyen Dang Binh(✉)

Hue University of Sciences, 77 Nguyen Hue, Hue city, Vietnam
ndbinh@hueuni.edu.vn

Abstract. Gestures, a natural language of humans, provide an intuitive and effortless interface for communication with the computers. However, the achievements do not satisfy researcher's demands because of the complexity and instability of human gestures. We propose a new method to recognize gestures from sound waves. The main contribution of this paper is to recognize gestures based on the analysis of short-time Fourier transforms (STFT) using the Doppler effect to sense gestures. To do this, we generate an inaudible tone, which gets frequency-shifted when it reflects off moving objects like the hand. We measure this shift with the microphone to infer various gestures. Experimenting method and evaluating results by using the hand gestures of many different people to browse applications such as website, document and images in the browser on the computers in the classroom and library environment for accurate results. In addition, we describe the phenomena and recognition algorithm, demonstrate a variety of gestures, and present an informal evaluation on the robustness of this approach across Laptop device and people.

Keywords: In-air gesture sensing · Doppler effect · Interaction technique

1 Introduction

Gesture recognition is very useful for automation. Gesture is becoming a common means, as technology trends in the management and control interfaces. Gesture recognition based on the variation in sound waves frequency domain approach is a sound wave sensors utilize computer speakers and microphone. It has the advantage of not being affected by light, as the technical language recognition in images, video, voice. For example, the Microsoft's SoundWave and University of Washington [7] or Acoustic Doppler Sonar (ADS) [2]. Currently, this problem is continuing to develop applications, e.g., a gesture recognition system that leverages wireless signals to enable whole-home sensing and recognition of human gestures [6] and bringing gesture to all device [1]. Basically, the gesture recognition systems using more than one characteristic, using machine learning models Hidden markov models, Support vector machine to recognize gestures are really complex and restrict much of the processing speed of the system. We present a method of recognizing the gesture by dividing the energy levels of short-time Fourier transforms and using Doppler effect to recognize gestures. This method conducted discrete signal on the frequency domain with time into the signal frame of

P.C. Vinh and V. Alagar (Eds.): ICCASA 2015, LNICST 165, pp. 41–50, 2016.
DOI: 10.1007/978-3-319-29236-6_5

equal length and continue dividing the energy level briefly on each frame signal sequentially. Also, the analysis of energy levels in a short time each signal frame allows detection of noise and remove signal preprocessing steps for reliable results and somewhat reduce the fees charged math. Energy function in short-time been researched much in image recognition, speech recognition [5, 8]. We are inspired by the Doppler effect, the effect of Doppler is used to detect sound waves, voice, gestures [2–4, 7] (Fig. 1).

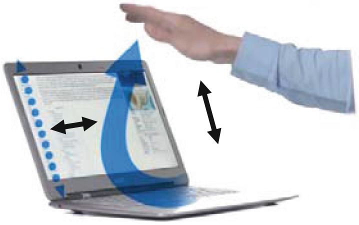

Fig. 1. An illustration of hand gesture recognition via sound waves used to control applications on laptop.

The paper is organized as follows. In Sect. 2 we review the main ideas, which build the sound waves gestures recognition system. A method that combines the Doppler effect is based on the division of power levels short-time Fourier transforms to recognize gestures. In Sect. 3 presents empirical evaluation results in sound waves gesture recognition. Finally, we summarize and conclude the paper in Sect. 4.

2 Sound Waves Gestures Recognition System

The system of sound waves gestures recognition and controls often include: signal sound waves acquisition, features extraction, finally classify, recognize and control interfaces. General diagram of gesture recognition system is shown on Fig. 2 sound waves uses existing speakers on commodity devices to generate tones between 18–22 kHz, which are inaudible. We then use the existing microphones on these same devices to pick up the reflected signal and estimate motion and gesture through the observed frequency shifts.

Sound Waves Acquisition: Use the built-in microphone on your computer to capture sound wave signal of gesture Doppler effect to change when it has changeable of the location of the two sources of waves.

Features Extraction: Typical short time energy is extracted Windowing choose soon step of Fourier transforms in a short time (STFT). Sound waves gestures and noise classification: positive and negative energy based on the distribution of energy levels in the short time-frequency domain to classify gestures and noise Doppler effect.

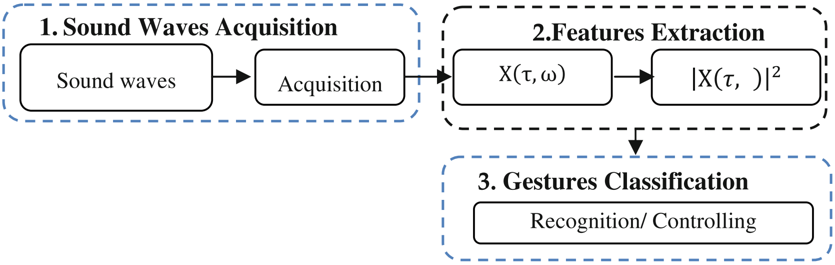

Fig. 2. The diagram of the basic stages of gestures recognition systems and controlling.

Gesture Recognition and Controlling: The user selected control functions, through the state classification system gestures to recognize gestures that execute commands from the controlling interface application browsers use.

2.1 Sound Waves Acquisition Based on Doppler Effect

In our approach, the computer speakers act as a broadcast source, microphone as receiver. Speakers will continuously emit a signal whose frequency from 18 kHz–22 kHz constant (adjusted by the user), although the sound waves can operate outside our scope but consistent over the range from 18 kHz–22 kHz because matching most hardware devices on the computer and we do not need higher frequencies to sense gestures in the air [2, 7, 10]. Then, use the built-in microphone on the device to capture and digitize signals through the recording (sampling frequency is 44.1 kHz) signal observed Doppler principle. It is a combination of two separate effects induced by two sound wave sources (hand moves and speakers), frequency will increase as the hand observer moving closer to the computer and will decrease when the hand moves away. The principle is the Doppler frequency shift of sound waves obtained in microphone (f_r) upon the relative shift of position with hands in the air compared with the computer speakers. The relationship between observed frequency f_r and emitted frequency f_t is given by

$$f_r = f_t \cdot \left(\frac{c+v}{c-v}\right)$$

where f_r is perceived frequency at microphone; f_t is original frequency from speaker; **c** is the velocity of sound waves in the air (speed of sound in air) and ***v*** is the velocity of hand in air; if hand is moving towards the source then positive

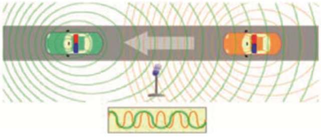

Fig. 3. Description of the Doppler effect is obtained at the microphone for two waves approach each sources.

(and negative in the other direction). Any motion nearby computer (about 1 m depending on speed), integrated microphone in your computer will obtain the frequency shift of the reflected Doppler effect (Fig. 3).

2.2 Features Extraction

In fact, the gesture moves continuously in a certain period of time and often not consistent between the times the performance (in speed, direction and time travel) and depending on the user (Campaign airborne sound transmission speed of 343 m/s, speed hand gestures about 0.25 m/s–4 m/s) [10]. Another factor, the acquisition of sound waves from the moving gesture greatly affected by noise (interference) available or random existence in their surroundings (as a hardware device or operating emitted program that uses active sound created). Therefore, the signal processing and extracting features selected to recognize what is a gesture or environmental noise in the frequency domain is really complex.

Windows of Signal: Signal in the short period of time can see the signal is relatively stable and unchanged over time. For a signal of gesture, this can be done by windowing of a signal x(n) into an unbroken chain of sequential x(t) window, t = 1,2,… T call is the signal frame. The selection window Hamming [7], for discrete signals into the signal frame (with sample points in 2048) we considered suitable for the energy spectrum will be concentrated in the middle of the frame signal:

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right) & \text{with } 0 \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

where n is the number of samples on a window (n is an even number), N is the number of signal frames. Featured short time energy. Energy shortly is determined by calculating the average of the total area of the sample (sample) single in each frame. With a window ends at the mth sample, short time energy function E(m) (Fig. 4):

$$E(m) = \sum_{-\infty}^{\infty} [x(n)\, w(m-n)]^2$$

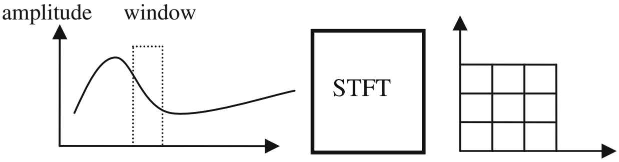

Fig. 4. Description of STFT.

Frequency signals emitted from the hand gestures or environmental noise signal is not stopped (non-stationary signals) means a periodic signal over time. Energy analysis shortly after each frame signal is unstable, or no detailed process of moving gesture or noise already exist or appear at random. So we continue discrete signal on each frame (framing) using short-time Fourier transforms.

$$\mathrm{STFT}\,\{x(t)\}\,(\tau,\omega) \equiv X\,(\tau,\omega) = \sum_{-\infty}^{\infty} x\,(t)\,\omega(t-\tau)e^{-j\omega t}$$

$$\mathrm{spectrogram}\,\{x(t)\}\,(\tau,\omega) \equiv |X\,(\tau,\omega)|^2 \equiv |\mathrm{STFT}\,(\tau,\omega)|^2$$

where, x(t) is the signal to change at time t $(1 \le t \le \tau)$, $X\,(\tau,\omega)$ represents the phase and amplitude of the signal in time and frequency. Spectral energy distribution function is the result of the transformation process STFT, featured short time energy is calculated after each step sliding window (windowing) of STFT:

$$\mathrm{Energy} = \sum_{t_i}^{t_{i+1}} x^2(t) = x^2(t_i) + x^2(t_{i+1})$$

in which energy is featured in short time, $x^2(t_i)$ is the energy spectrum signals in frame $t_i\,(1 \le i \le N)$ (Fig. 5).

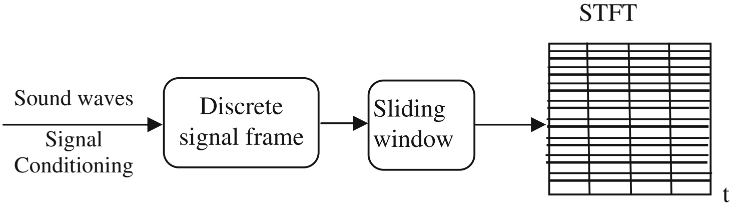

Fig. 5. Describe the process selected feature extraction.

2.3 Gestures Classification

In this paper, we limit the waving gesture (palms facing the computer, limit the distance in 1 m depending on the speed of movement of the hand than the computer and audio hardware of the computer) move from top-down "up to down", from right to left "right to left" is called the "state" move closer or move away from the computer including gestures moving from left to right "left to right", from bottom to top "down to up". The moving gesture is shown clearly in Fig. 6.

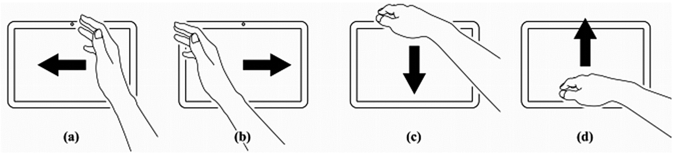

Fig. 6. The hand gestures. (a) and (b) is approaching "Coming". (b) and (d) is carried away "Leaving".

The idea of gestures classification is based on the transformation of the energy of the positive or negative on the frequency domain (oy axis is frequency-domain, time-domain is the ox axis) through the distribution of energy levels Shortly after wave change on every frame STFT signal Doppler effect principle. When there is not any movement, short time energy to wave signal emitted from the speakers (whose frequency is cf) called threshold (ethrd) will remain unchanged. When a gesture is moving towards or away from it a short time energy component increased or decreased distribution around the threshold of the Doppler principle. We call the energy increase in the short period of time is positive energy (pose) similar to the energy reduction (nege) is negative energy (this is why we calculated the energy function for the amount of the short time period after the sliding window step change STFT).

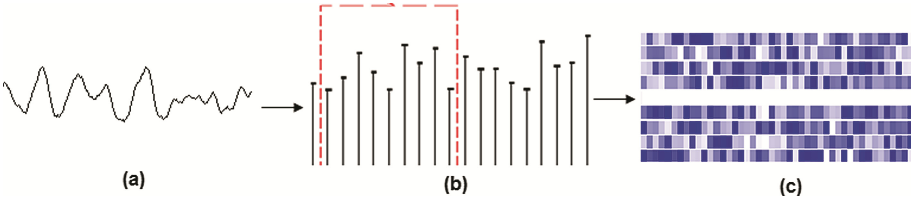

Fig. 7. Distribution of share positive/negative energy. (a) the signal of gesture has been digitized. (b) The modified signal STFT using Hamming window. (c) The distribution of the energy of negative/positive energy around the threshold when motions are shown colored boxes with value increases from white to dark blue (Color figure online).

The analysis of the energy positive or negative in scope (range) of the near threshold energy level (cf – oy ≤ range ≤ cf + range) is not specific to the "state" moving gesture hand because it changes so fast around the threshold. In this case, may be due to the noise of the hardware or the program that uses sound triggers. We are interested in the

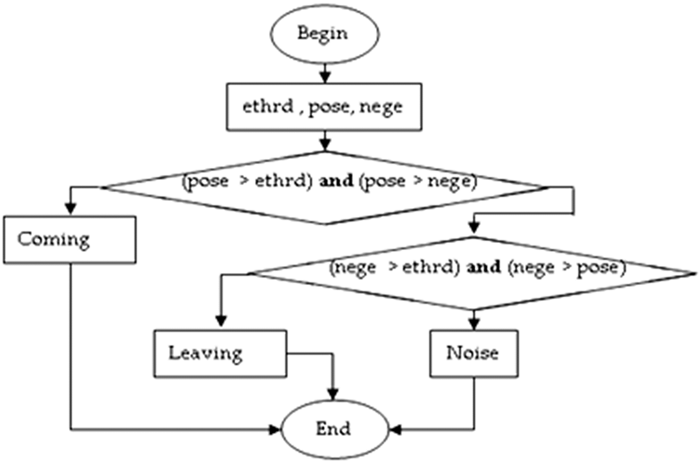
Begin
ethrd , pose, nege
(pose > ethrd) and (pose > nege)
Coming
(nege > ethrd) and (nege > pose)
Leaving
Noise
End

Fig. 8. Flowchart gestures classifier algorithm and noise.

range (oy < cf – range and oy > cf + range) to change the energy section briefly relatively clear than the threshold to classify gestures into advanced classes closer “Coming” or move away from the “Leaving”. Similarly, in a range (extrange) larger (oy < cf – extrange and oy > cf + extrange) existence of the energy available to a positive or negative than the threshold, then it is sure to be noise “Noise “which is not the state of motion of hand gestures, because it is not fast enough to spread the energy distribution around the threshold (Figs. 7 and 8).

Algorithm 1: Classification algorithm for gesture and noise

```
Input: Feature vectors are extracted from short time energy function of the sound
       waves.
Output: one of 5 Classes gesture:
        "Coming" = {go down; translated into left},
        "Leaving" = {go up on; translated to right},
        "Noise" = {other cases}.
Methods:
1. Browse the energy levels in the frequency domain (oy <cf - range and oy> cf +
range) to calculate the input parameters (ethrd, pose, nege) by the following formula:
   $\text{Energy} = \sum_{t_i}^{t_{i+1}} x^2(t) = x^2(t_i) + x^2(t_{i+1})$
2. Comparison of the parameters are calculated in step 1 with the threshold to classify
gestures to the state of motion "Coming", "Leaving", or "Noise".
3. End.
```

2.4 Gesture Recognition and Control Application Interface

We built a system to recognition and control the selection gesture function using virtual keys to browse applications include: Browse the document horizontally, browsing the document vertically, transforming a document page, scroll up or scroll down the document page by user selection options virtual keys in Fig. 9. When a moving gesture of waving at a computer, the system will classify it into class gesture “Coming” (approached), including two gestures: “shift to the left” and “go down” (the celebration only (a) and (c) in Fig. 6) or belonging to the class “Leaving” (moved away) include gesture “translated to” and “go down” (the gesture (b) and (d) in Fig. 9).

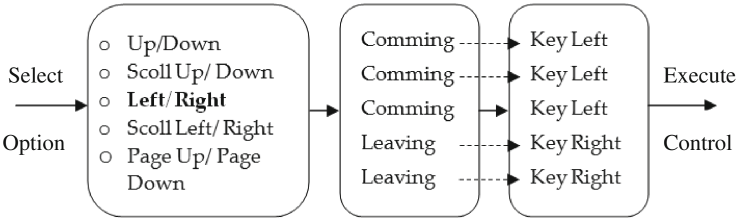

Fig. 9. Select functional diagram of the control system.

Through the selection of options to browse virtual key document by the gesture was introduced, the system will accurately recognize either gesture in each class from which execute control commands instead of interact directly with the computer keyboard or mouse.

3 Experiment and Results

In this section we focus on the construction of the experimental program identifiable sound waves hand gestures to control a laptop computer. With applications built, I show the results, evaluate the effectiveness and applicability of the method was developed. Since then outlined the limitations of the method and the innovative direction to develop better applications.

3.1 Sound Waves Gesture Recognition System

The aim to build the application as a motion recognition system of the laptop based on sounder speaker and microphone. That's touch less sensor can recognize movement of the hand. Users can use gestures to control programs like Flip Slide PPT, moving in the photo browser, browse PDF documents, Word, Excel, or surf websites … This app is similar to sound waves Microsoft is researching how to handle it but completely different. The system is operating on sound waves at 18 kHz–22 kHz frequency and can be adjusted in the user interface. Some techniques are built in the programming make the results reliable recognition, including short time Fourier transforms, Doppler effect, gesture recognition algorithm, the human voice and recognize ambient noise, as well as enhanced recognition algorithms strong identity.

3.2 Environmental Construction and Operation of the Application

Application is installed on Microsoft Visual C++ environment, so it may or fine on computers using Windows operating systems understand. Application allows to run on

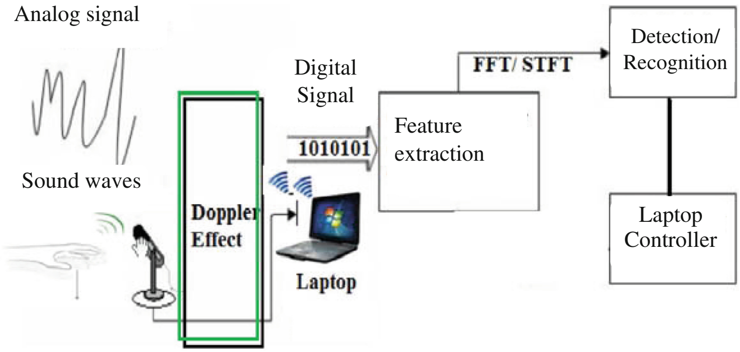

Fig. 10. Our sound waves gesture recognition system for controlling applications on Laptop.

any laptop computers have integrated at least a microphone and speakers. The application uses sound waves to the use of the day or night absolutely not affect other applications that use the camera identification. The application uses the built-in microphone and speakers on the laptop computer so the gap between users and computers in applications is only 0.7 m, the closer the distance, the higher the accuracy (Fig. 10).

3.3 Experimental Results

We conduct empirical methods in place (environment) calm, considered easy to recognize gestures such as a library or a quiet place. In these locations, usually at survival signals emitted sound waves or less random noise existed available to affect the classification process and recognize gestures. At the same time, we also experimentally in noisy places such as classroom, locations are often available randomly or other source of sound waves that are not only caused by. Through noisy places or quiet, we evaluate the effectiveness of methods for environmental use. Also, in both quiet environments (library) and noisy (classrooms) for many different users to assess the impact (in terms of speed, direction of movement) of the election different only for the method. To review the stability and effectiveness of the methods implemented, we use a number of different types of computers. The computers operate in different modes but (allowing applications to run simultaneously in multiple time difference) to control a number of applications such as web surfing, browse PDF documents, PPT slides, or browse Photos in the browser application using four hand gestures were introduced instead of using direct or mouse navigation keys on the keyboard. By aggregating data in the evaluation method mentioned above, we found that the method can recognize gestures resulting noise and very reliable. The experimental results are we averaged in each gestures shown in Table 1.

Table 1. The average percentage of four gesture recognition.

Environment experiment	Results percentage ratio of gestures recognition controller			
	Control move up	Control go down	Control shift to the left	Control shift to the right
In the classroom	80.7 %	74.6 %	82.4 %	73.5 %
In the library	89.3 %	77.7 %	88.1 %	79.9 %

Through the development, implementation and evaluation methods, we found that the analysis of the short-time energy levels through the energy of positive/negative reflects well the movement of the hand gestures on the frequency domain with time. Beside eliminate the impact of environmental noise interference. The method focuses on a specific analysis of energy that no combination with many other features, it kind of makes findings is somewhat restricted compared with other methods. In addition, the energy spectrum analysis only considered the basic hand gestures through two states approaching and moving away. The more complex gestures have not been considered in the energy spectral features shortly.

4 Conclusion

We have presented a novel sound waves gestures recognition system. The basic idea is to recognize gestures based on the analysis of short-time Fourier transforms (STFT). Our method that combines the Doppler effect and the division of power levels short-time Fourier transforms on the frequency domain to recognize gestures. The method is based on a single feature to recognize gestures and noise removal of ambient devices make cost in terms of computation and processing improved somewhat. Also, the method also shows the simplicity and ease application deployment for leverage spacious sound hardware often built on the device from which the cost price is minimized. Next time, we will research and development towards detection and gesture control completely automatic no longer depend on the selection of the control functions of the user. Using machine learning models (HMM, SVM) based on the energy characteristic short time to train and discovered many more gestures. The use of filters and sound waves combine this method with other methods of detecting gestures from images through the camera in a system to improve the accuracy and detect multiple complex gestures.

References

1. Bryce, K., Vamsi, T., Shym, G.: Bringing gesture recognition to all devices. In: NSDI, April 2014
2. Kalgaonkar, K., Raj, B.: Ultrasonic doppler sensor for speaker recognition, acoustics, speech and signal processing. In: ICASSP (2008)
3. Kalgaonkar, K., Raj, B.: Ultrasonic doppler sensor for voice activity detection. IEEE Sig. Process. Lett. **10**, 754–757 (2007)
4. Kalgaonkar, K., Raj, B.: One-handed gesture recognition using ultrasonic doppler sonar. In: Proceedings of IEEE Acoustics, Speech and Signal Processing (2009)
5. Liu, D., Tian, J., Yang, B., Sun, J.: Time-frequency analysis based motor fault detection using deconvolutive STFT spectrogram. J. Convergence Inf. Technol. (JCIT) **7**, 1–6 (2012)
6. Pu, Q., Gupta, S., Gollakota, S., Patel, S.: Whole-home gesture recognition using wireless signals. In: Proceedings of the 19th Annual International Conference on Mobile Computing and Networking (2013)
7. Pu, Q., Gupta, S., Gollakota, S., Patel, S.: Soundwave: using the doppler effect to sense gestures. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems (2012)
8. Chikkerur, S., Govindaraju, V., Cartwright, A.N.: Fingerprint image enhancement using STFT analysis. In: Singh, S., Singh, M., Apte, C., Perner, P. (eds.) ICAPR 2005. LNCS, vol. 3687, pp. 20–29. Springer, Heidelberg (2005)
9. Tarzia, S.P., Dick, R.P., Dinda, P.A., Memik, G.: Sonar-based measurement of user presence and attention. In: Proceedings of ACM UbiComp (2009)
10. Xin, L., Kang, L., Dong, C.L.: A sound-based gesture recognition technology designed for mobile platform. J. Inf. Comput. Sci. **12**, 985–991 (2015)