LiveEar: An Efficient and Easy-to-use Liveness Detection System for Voice Assistants

Ling Yue†, Chenhong Cao, Yufeng Li, Jiantao Li, Qi Liu

1 School of Computer Engineering and Science, Shanghai university, Shanghai, 200000, China
* e-mail: YueLing@shu.edu.cn

Abstract. Voice assistants, such as Amazon Alexa, Apple Siri and Tmall Genie, using voice biometrics for the identity authentication, are becoming pervasive in our daily lives. However, voice assistants are vulnerable to reply attack due to the open nature of voice-input channels. An attacker can record the voice commands of victims and replay them to spoof voice assistants. Existing liveness detection approaches are mostly based on machine learning methods, which are expensive and complex. Recently, several approaches are proposed to leverage the human specific voice features or the distinctness of voice played by loudspeaker. However, they require the users and the voice assistant to be in a fixed position and at a very close distance, which is not user-friendly in practice. This paper proposes LiveEar, an efficient and easy-to-use liveness detection system for voice assistant. LiveEar utilizes the differences in phoneme positions between live-human voices and voices replayed through loudspeakers. Specifically, it calculates the time-difference-of-arrival (TDoA) in a sequence of phoneme sounds to the microphone on the voice assistant. Then, an SVM-based classification model is trained with the extracted TDoA features. This paper implements a prototype of LiveEar and evaluates its performance using real-world data. Results show that LiveEar achieves high detection accuracy in various flexible positions, with negligible runtime overhead.

1. Introduction
The proliferation of voice assistants enables more human-friendly interactions supporting identity authentication, online shopping, mobile financial services, and so on [19,20]. However, such security-critical and privacy-critical applications make them high-value targets for attackers. A major threat to the security of voice assistants is the replay attack due to the open nature of voice input channels. An attacker could easily record the voice commands of victims and replay them to spoof voice assistants. To mitigate such spoofing attacks, a set of deep learning-based approaches [15,16] have been proposed. However, employing deep learning models such as LCNN, CNN and RNN is computationally expensive and complex [15]. Recently, several approaches are proposed to leverage the human specific voice features or the distinctness of the voice played by loudspeaker. VoiceGesture [6] extracted lip movement using Doppler effect. However, the extent of articulatory movements would affect the accuracy of the results. VoicePop [11] utilized the characteristics that people speak with breathing sound while the machine did not have to identify replay attacks. However, the distance between the mouth and the microphone must be very close during authentication to capture the breathing sound. Although effective, these methods need to be held or moved in a specific way, and they are limited by the range of motion. In this paper, the method we propose is mainly for intelligent voice assistant, which have at least four microphones around it. The main contributions of this paper can be summarized as follows:
Our solution leverages the microphones on the voice assistant and it is easy-to-use as it can be easily implemented on existing voice assistant as a software or plug-in.

Our system uses SVM algorithm to extract the TDoA features of humans and machines, which can effectively distinguish humans from machines.

We develop LiveEar, a liveness detection system using voice assistant. Extensive evaluation results show that our system is highly effective in detecting replay attack.

The rest of this paper is organized as follows. We first show the preliminary in Section 2, followed by the system design in Section 3. The performance evaluation of the system is presented in Section 4. Finally, we review related work and make a conclusion in Section 5 and Section 6.

2. PRELIMINARY

2.1. System and Attack Model

According to [4], voice authentication systems can be classified into two types. The first type is text-dependent, requiring the voice spoken by the user for voice authentication remain unchanged during registration and verification. The second type, text-dependent, is prevailing and performs well. Our research belongs to the second one.

For the attack model, it is mainly divided into two categories: imitation attack and replay attack. The imitation attack can be solved by automatic speaker verification (ASV) system [1]. Replay attack is to use the speaker to play recorded voice. In this work, we assume that the voice sample can be easily obtained by the attacker from the victim.

2.2. Phonemes

Phonemes are the smallest units of speech and are generally marked with the International Phonetic Alphabet (IPA). The analysis of phonemes is generally described according to the pronunciation action, and a pronunciation action forms a phoneme. Moreover, the pronunciation of different phonemes is different.

In general, we classify phonemes into two categories: vowels and consonants. Different vowels can usually be distinguished by the position of the tongue. As shown in Figure 1, the highest part of the tongue moves to the front of the mouth and arches slightly to produce different vowels [4]. Unlike vowels, consonants are mainly distinguished by the manner of articulation and the place of articulation.

![Figure 1. Tongue positions of English vowels within the oral cavity.](image)

![Figure 2. The flow of our liveness detection system.](image)
3. SYSTEM DESIGN

3.1. Overview
The key idea of LiveEar is to extract features from the TDoA values of humans and machines. Then a model is trained to distinguish humans from machines. For live users, due to the different physical positions of phonemes in the human voice channel, most phoneme sounds have measurable TDoA differences to the four microphones of the voice assistant, which is almost non-existent for replay voice. For each passphrase, we can get several sets of TDoA values of a series of phonemes. First, we choose one of the four microphones of the voice assistant as the reference microphone, and then measure the TDoA of the phonemes arriving at the other three microphones and the basic microphone. Then, a large number of TDoAs are inputted into the SVM algorithm to train the model. After getting the model, we only need to input a series of TDoAs of phonemes to deduce whether it is a replay attack or not in each authentication.

Our system does not require the user to place the voice assistant close to the mouth. We have detected that the value of TDoA is not conducive to distinguishing people from machines when the distance exceeds 30 cm, and this distance is not conducive to other people’s replay attacks on the authentication system.

3.2. System Flow
Figure 2 describes the system architecture of the system, which consists of the following parts: phoneme segmentation, TDoA calculation, and classification. We should extract the phonemes from the speech samples firstly. In particular, we use the MAUS [14] for phoneme segmentation. In TDoA calculation phase, we calculate the time difference from the phoneme to different microphones around the voice assistant. In the classification phase of the result, we chose the SVM method in machine learning.

3.3. Phoneme Segmentation
We use MAUS as the main method of phoneme segmentation and labelling [14]. First of all, we need to recognize the words of the passphrase, which can be achieved by using the relatively mature technology of automatic speech recognition. Secondly, the recognized words are converted into the expected pronunciation according to the standard pronunciation model. The generated standard pronunciation, together with millions of possible user accents, produces a probability map that includes all possible hypotheses and corresponding probabilities. Finally, the system uses the “hidden Markov model” (HMM) to search the path of the most probable speech unit in the probability map. The result is the segmented phoneme unit.

Figure 3. Illustration of phoneme localization using the word of ‘voice’.

3.4. TDoA Calculation
There are already a variety of time difference estimation algorithms. In previous works [4], they chose the phase-transform weighted generalized cross-correlation method (GCC-PHAT) to calculate the TDoA of each phoneme. However, in the implementation process, we found that the use of GCC-PHAT for the phoneme TDoA calculation has the problems and challenge of low accuracy. This is due to the presence of noise in the environment (environmental white noise or accidental sound) and the non-periodic nature of the sound signal, which significantly reduces the performance of the GCC-PHAT. To
solve this problem, before performing TDoA calculation, we windowed the sound data and found that it can effectively improve the accuracy of the results. After comparing various window functions, we chose the Hanning window to process the sound signal. The Hanning window can use side lobes to cancel each other and eliminate high-frequency interference and energy exposure.

After windowing the speech signal, we use GCC-PHAT to calculate TDoA for each phoneme. We use a voice assistant with four microphones as shown in Figure 3. We choose the second microphone as the reference microphone. In order to maximize the calculated TDoA value, we are required the mouth position in line with Mic2 and Mic0 when training the model and authentication.

3.5. Classification

In the classification judgment of the result, we use the classification algorithm in machine learning: support vector machine algorithm (SVM). Considering that the difference between the test data obtained by human voice and machine voice is very small, we decided to use two-class SVM algorithms to judge the classification of human voice and machine voice. After collecting the TDoA of different phonemes reaching the same microphone, we will use the SVM algorithm to train the data, and get the training model that we need to identify human voice and machine voice.

4. EVALUATION

4.1. Experiment Setup and Methodology

In this section, we evaluate the performance of our system under replay attacks. We also evaluate the robustness of the LiveEar at different distances, data set size, and position. We evaluate the system with Raspberry Pi 4B and Respeaker 4. Respeaker 4 is a four-microphone expansion board for Raspberry Pi, suitable for AI and voice applications. Specifically, we use Respeaker 4 to collect voice samples and process them with Raspberry Pi.

To register in the system, we choose a word as a passphrase and speak the passphrase several times in the distances of 10 cm, 20 cm, 30 cm, 40 cm, 50 cm respectively to calculate the TDoA of the real user. In addition, we need to get the machine's voice. We use a Meizu 16s pro as a speaker to play the same volunteer’s recorded sound, and the sound also carried out several times in each distance. For the authentication process, the user only needs to speak the passphrase once.

After the actual test, we found that the microphone data sampling range is very small when the speaker is more than 50cm away from the microphone, even if we adjust the volume to the maximum range. This range is difficult to distinguish by the eye. The running time of the phoneme cutting algorithm is significantly increased, and the subsequent TDoA calculation results are also very close. In our experiment, in order to verify the accuracy and practicability of the algorithm, we only consider the case that the distance value is within 50 cm. Since we choose Mic2 as the reference microphone, the TDoA value is the largest when Mic0, Mic2 and the sound source are in a straight line, but this does not represent the sound source remains the same position.

To evaluate the performance of our system, we define the metrics as follows: False Accept Rate (FAR) is defined as the rate at which the system refers to an attacker as a live user to wrongly accept. False Reject Rate (FRR) is defined as the rate at which the system refers to a live user as an attacker to wrongly reject. Accuracy is defined as the rate that the system can accurately identify a live user and reject the attacker.

4.2. Impact of sound source to microphone distance

As one user may be in different distances to perform online authentication, we evaluate how our system behaves at different distances between sound source and microphone. We guess that the farther the distance from the sound source to the microphone, the smaller the TDoA between different phonemes and the microphone. Figure 4 shows our experimental results. According to the results, we can see that when the distance between the sound source and the microphone is within 30 cm, the accuracy of the experiment can be over 95%, and when the distance between the sound source and the microphone is
within the range from 30 cm to 50 cm, the accuracy of the experiment can reach more than 70%. From the above data, we can see that our system has the best performance when the distance between the sound source and the microphone is less than or equal to 30 cm, which can effectively resist replay attacks. Moreover, we compare our system with the VoiceLive [4]. Figure 5 illustrates the accuracy comparison of these two systems. We observe that the two systems achieve very high accuracy when the distance is less than 30 cm. It is hard to capture the TDoA dynamic of a passphrase when the distance is over 30 cm.

4.3. Impact of dataset size
For the training model, extremely large data set will result in overfitting, and extremely small data set will cause under-fitting. Therefore, it is necessary to study how the size of the data set effects the system’s performance. Figure 6 shows the relationship between the authentication accuracy of the system and the size of the data set. From the result, we can see that when the data set is 150, the Accuracy is the highest. The result indicates that the runtime overhead of our system is small as it can achieve good results with a small data set.

4.4. Effect of different position
To prove that our system can achieve high detection accuracy in various flexible positions, we evaluate LiveEar’s performance when there exists displacement of the sound source between the position for registration and authentication. Specifically, we will allow the volunteer to move positions to varying degrees when performing authentication, i.e., 0 cm, 2 cm and 5 cm away from the microphone’s position that he/she enrolled in the system. Such displacements occur on the axis of X (left and right) and Z(up), the Y axis is defined as the distance between the sound source and the microphone. Figure 7 shows our system is insensitive to the displacements on X axis and Z axis when the displacement is less than 5 cm. The reason is as follows: the feature of the TDoA value may be change when the displacement is more than 5 cm. The results indicate that our system is efficient in various positions within the specified area. Figure 8 illustrates the comparison of accuracy of two systems under different degrees of phone
displacement. We can see that the VoiceLive is sensitive to the Y axis while the LiveEar is not sensitive to the small movements of the phone in any direction.

5. RELATED WORK
More and more applications use voice biometric technology for identity authentication, and recent research shows that voice biometric technology is vulnerable to spoofing attacks. There are two types of spoofing attacks, one is human-based spoofing attack, and the other is replay attack. Therefore, the key to solving the latter attack is to find out the difference between human voice and machine voice. For example, the VoiceLive [4] system proposed to measure the time difference of arrival (TDoA) to a smartphone's two microphones. Different phonemes have different pronunciation positions in the human mouth, and the machine does not have this feature. Chen et al. [5] found that the loudspeaker generates a magnetic field when it is working, but it does not when people speak, so as to distinguish between people and machines. Wang et al. [11] proposed to identify live users by detecting whether there is pop noise, because people speak with pop noise, but the machine does not. The above methods are mainly used to solve machine-based spoofing attacks. The next method can detect both machine-based spoofing attacks and human-based spoofing attacks.

Figure 6. FAR, FRR and Accuracy of different dataset.

Figure 7. FAR, FRR and Accuracy under different degree of phone displacement.

Figure 8. Comparison of accuracy of two systems under different degree of phone displacement.
Zhang et al. [6] proposed to use the Doppler effect to roughly extract the movement of the lips, which can be used to distinguish people from machines because the movement of the lips is multi-dimensional, and the movement of the speaker is one-dimensional; it can also be used to distinguish between different people because lip movement is different when different people talk. However, Wu et al. [8] proposed the method based on the fusion of lip movement and speech features, which can get more accurate lip movement by measuring the change of phase shift. And Yan et al. [9] proposed CaField, which exploits the difference in sound fields between authentic users and spoofing attackers while achieving a balance of security and usability, and it supports continuous authentication of the voice input. In this work, what we have proposed is mainly used to solve machine-based spoofing attacks on voice assistants. Our system is easy-to-use as it is not restricted by position in the specified direction and it is efficient to distinguish between humans and machines.

6. CONCLUSION
In this work, we developed a liveness detection system for voice authentication that does not require extra hardware equipment except the microphone on the voice assistant. Our system LiveEar performs liveness detection by extracting the feature of the TDoA value of humans and machines. It distinguishes a live user from a replay attack by training a binary classification model based on the difference between the TDoA changes of the user and the replay attack. Extensive experiments confirmed that our system is reliable and effective to distinguish a live user and a replay attack under diversified environments.

Acknowledgments
This work was supported by the Development Program in Key Areas of Guangdong Province (No. 2018B010113001), National Science Foundation of China (No. 62002213), and the Shanghai Sailing Program (No.20YF1413700).

References
[1] T. B. Amin, J. S. German, and P. Marziliano.: Detecting voice disguise from speech variability: Analysis of three glottal and vocal tract measures. The Journal of the Acoustical Society of America, 134(5):4068–4068(2013).
[2] J. P. Olive, A. Greenwood, and J. Coleman. Marziliano: Acoustics of American English speech: a dynamic approach. Language & Speech, 38(1):115-118(1995).
[3] P.Ladefoged.: A course in phonetics. Hardcourt Brace Jovanovich Inc. NY,2014.
[4] Zhang L, Tan S, Yang J, et al.: VoiceLive: A Phoneme Localization based Liveness Detection for Voice Authentication on Smartphones[C]. In: computer and communications security, pp.1080-1091(2016).
[5] Si Chen, Kui Ren, Sixu Piao, et al.: You Can Hear But You Cannot Steal: Defending Against Voice Impersonation Attacks on Smartphones[C]. In: IEEE International Conference on Distributed Computing Systems. IEEE Computer Society, 2017.
[6] Zhang L, Tan S, Yang J.: Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication[C]. In: computer and communications security, pp. 57-71(2017).
[7] Feng H, Fawaz K, Shin K G, et al.: Continuous Authentication for Voice Assistants[C]. In: acm/ieee international conference on mobile computing and networking, pp. 343-355(2017).
[8] Wu L, Yang J, Zhou M, et al.: LVID: A Multimodal Biometrics Authentication System on Smartphones[J]. In: IEEE Transactions on Information Forensics and Security, pp. 1572-1585(2020).
[9] Yan C, Long Y, Ji X, et al.: The Catcher in the Field: A Fieldprint based Spoofing Detection for Text-Independent Speaker Verification[C]. In: computer and communications security, pp. 1215-1229(2019).
[10] Lu L, Yu J, Chen Y, et al.: Lip Reading-Based User Authentication Through Acoustic Sensing on Smartphones[J]. In: IEEE ACM Transactions on Networking. 27(1): 447-460(2019).
[11] Wang Q, Lin X, Zhou M, et al.: VoicePop: A Pop Noise based Anti-spoofing System for Voice Authentication on Smartphones[C]. In: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications. IEEE, 2019.

[12] Artur Janicki, Federico Alegre, Nicholas Evans, et al.: An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks. In: Security and Communication Networks 9, 15 (2016), 3030–3044.

[13] W. Shang, M. Stevenson.: Score normalization in playback attack detection. In: IEEE ICASSP, 2010.

[14] Kisler T, Schiel F, Sloetjes H.: Signal processing via web services: The use case WebMAUS. In: Digital Humanities (DH). 2012.

[15] Kinnunen, Tomi, et al.: The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In: conference of the international speech communication association. 2-6(2017).

[16] Lavrentyeva G, Novoselov S, Malykh E, et al.: Audio Replay Attack Detection with Deep Learning Frameworks.. In: conference of the international speech communication association. 82-86(2017).

[17] Wechat voiceprint: http://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password.

[18] Saypay technologies, http://saypaytechnologies.com.

[19] Hsbc offers voice biometric, http://www.bbc.com/news/business-35609833.